PATTERNS IN THE DIVERSITY AND DISTRIBUTION OF FLOWERING PLANT GENERA

VOLUME I

Neil Alistair Brummitt

A thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

2004

DECLARATION

I declare that all the work presented in this thesis has been undertaken by me, unless duly acknowledged, and has not been submitted for any prior degree or qualification, and that this thesis has been written by me.

Signed Neil Alistair Brummitt

17th May 2005

ABSTRACT

Regional distributions of all vascular plant genera have been compiled from herbarium specimens at the Royal Botanic Gardens, Kew, and this data has then been analysed for large-scale patterns in the diversity and distribution of flowering plants, at both genus and family levels. A strong latitudinal gradient in diversity is apparent at family, genus and species levels, though while western South America is most diverse at species and genus levels, it is the SW. Pacific which is most diverse at family level. However, the number of families and genera per region is very strongly correlated, irrespective of the region. There is a very strong relationship between area and both family and genus diversity, though not for numbers of endemic genera. Analysing floristic similarity between different regions of the world reveals very strongly supported continental groups, since most genera are confined to particular continents, although the latitudinal difference between regions is a better predictor of floristic similarity than is simply distance between regions. Latitudinal range-size for genera increases towards the equator, although taxon-size in general decreases with increasing latitudinal range-size. For both families and genera, the range-size frequency distribution is highly skewed towards small range sizes (more so for genera than families), which account for the majority of taxa. Distribution patterns show strong regional clustering, with almost 40% of genera single-region endemics, and approximately 20% of world distribution patterns accounting for about 80% of total angiosperm genus diversity. Analysis of these distribution patterns reveals a strong correlation between diversity and the number of floristic elements, which intersect to form the diversity of a region. In general, though with many exceptions, there is a correlation between recency of evolutionary origin and the size (number of taxa) and spread (size of distribution) of flowering plant families. However, while a phylogenetic perspective becomes essential for addressing within-family patterns of distribution, it is argued that over the whole clade of flowering plants the resulting patterns of diversity are constrained more by large-scale ecological processes.

CONTENTS

VOLUME I

VOLUME II

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This PhD project was generously funded by the Bernard Sunley Charitable Trust, to whom I remain extremely grateful. My supervisors, Dr. Eimear Nic Lughadha (Royal Botanic Gardens, Kew) and Professor Peter Furley and Dr. William Mackaness (both School of Geosciences, University of Edinburgh) always provided help and support as requested or needed, remained patient during long periods of silence, and still reviewed and commented on drafts of this thesis promptly. Thank you for this. Any remaining errors are my own.

Justin Moat (RBG, Kew) deserves special mention for providing invaluable practical and technical advice with GIS and also general computing support – much of this would not have been achieved without his help; thanks also to Sally Hinchcliffe (RBG, Kew) for occasional advice with database management issues. Rafaël Govaerts (RBG, Kew) very kindly provided the figures from which I extrapolated species counts for each region. I am also grateful to Kate Hardwick (RBG Kew) for checking all the references, and spotting many mistakes in the process. Few members of staff of the Herbarium of the Royal Botanic Gardens, Kew have not been asked for advice or assistance at one time or another when I was compiling the data; all of them gave their time willingly and I thank them all. In addition I have also benefited from discussions with Dr. Simon Mayo and Dr. David Simpson (both RBG Kew) and Professor Dave Unwin (University College, London). Dr. Paul Smith and Stuart Cable of the Millennium Seed Bank generously granted me time over the last few weeks of this project to make sure this thesis was finally finished; I remain very grateful to them for this, otherwise it would not have got done.

Lastly I would like to thank my family and friends for their continued tolerance, and lasting moral and emotional support for the duration of this project – it made a big difference.

CHAPTER 1

INTRODUCTION

1.1 Introduction to the thesis

The research presented in this thesis is based on work undertaken at the Royal Botanic Gardens, Kew. There, a compendium of global distributions of all vascular plant genera (flowering plants, conifers and other gymnosperms, and ferns and their allies) was compiled by me. From compiling these data it was particularly apparent that some areas of the world had many more genera than did other areas, and also that the same distribution patterns occurred again and again in unrelated groups of plants. How did these plants come to be distributed where they are now? Why are there more plant genera in some areas than in others? How similar are different areas and different plant distributions to each other, and how can this best be studied? This thesis tries to address such questions using these distribution data for all angiosperm genera. Inspiration came from a ground-breaking work on the objective analysis of distribution patterns, E.H. Rapoport's *Areography* (Rapoport, 1982). Rapoport (1982) perfectly encapsulated the frustration of trying to analyse quantitatively entities so intangible as plant distributions:

> *"geographical areas of distribution are the Chinese-lantern shadows produced by the different taxa on the continental screen: it is like measuring, weighing and studying the behaviour of ghosts."*
>
> E.H. Rapoport, 1982, p.1.

In spite of these intrinsic difficulties, this thesis presents several comprehensive analyses of patterns of taxonomic diversity around the world, and also of the patterns of distribution in different groups of plants which underpin the diversity patterns. By 'patterns of diversity', I mean taxonomic richness, or counts of all taxa (of the same rank) naturally occurring within an area, compared between separate areas. 'Patterns of distribution' therefore refers to the comparison of all native occurrences for taxa or collections of taxa, throughout different areas. The different chapters of this thesis analyse different aspects of plant diversity and distribution patterns, each using separate methods appropriate to the questions addressed in that chapter. The chapters thus stand on their own to some extent, with the relevant methods introduced in that chapter, each of them representing separate pieces of a larger puzzle but together forming a more comprehensive picture which will hopefully be clear with the last piece of the puzzle (the last chapter of the thesis) in place.

Chapter 1 offers an introduction to the thesis and the database of distribution data on which the research is based, outlines the practical and philosophical approach taken with both the geographical and the taxonomic units of analysis, and briefly reviews the classical and contemporary literature that has shaped the modern field of biogeography. Chapter 2 briefly outlines the data and methods used for the analyses presented in this thesis; given the importance of multivariate statistics for this thesis, multivariate techniques relevant to the analyses undertaken here are reviewed in more detail. Chapter 3 presents the general distribution of the numbers of taxa at each taxonomic level in different regions around the globe, analyses these patterns against the sizes of regions and the distances between regions, and discusses biogeographical hypotheses which may account for these patterns in diversity. Subsequent chapters then investigate these patterns in greater depth, each chapter extending the analysis of a particular factor. Chapter 4 focuses on the latitudinal gradient of diversity, and examines the possible role of the range sizes of different taxa, and the geographical distribution of the variation in those range sizes in relation to the amount of available land area in those regions, in determining the richness of tropical regions.

The next two chapters each present different multivariate statistical analyses of the data, the first analysing biogeographic relationships between different regions of the world, then the second analysing biogeographic relationships between the different families and genera of the world. Chapter 5 analyses the degree of floristic relationship between different regions of the world and discusses these results in the light of existing global floristic classifications. Chapter 6 uses multivariate techniques to analyse similarities between the distribution patterns of families and genera and presents comprehensive classifications of both global family and global genus distributions, and then also examines the relationship between the diversity of genera within a region and the distributions of those genera outside of that region. Chapter 7 analyses the different factors that contribute to the richness of a region in the light of the floristic composition of that region, and argues that results from this genus-level analysis may also hold true for patterns of species diversity – for which there are not such comprehensive data. Therefore there is not a single, uniform analysis undertaken or a single hypothesis which is proposed and then tested but rather a diversity of approaches is used; in essence the whole argument behind this thesis is that by analysing each biogeographic pattern separately one then misses the 'big picture'.

1.2 Compiling distributions for all genera of vascular plants

1.2.1 The Distributions of Vascular Plant Families & Genera database

At the Royal Botanic Gardens, Kew (hereafter RBG Kew), a taxonomic database of all genera of vascular plants has been compiled, and has since been continuously maintained following its initial publication over a decade ago (Brummitt, 1992). To this database have now been added global distributions for all taxonomically-accepted genera; currently, there are 14,724 accepted genera in the database. These distributions have been scored by each of the 52 Level 2 geo-political regions of the Taxonomic Databases Working Group (TDWG) world geographical scheme for recording plant distributions (Brummitt, 2001; see Figure 1.1 and also Appendix 1). This is an international data standard used in taxonomic databases throughout the world for recording standardised distributions of plants. As well as being scored by each region, the distribution of each genus has also been recorded as a verbal text string. The database contains a total of 70,550 individual distribution records. Distributions were compiled into a copy of the database installed on a laptop computer, primarily from records of herbarium specimens held at Kew, and these have been supplemented with additional literature records from an extensive search through standard regional Floras (reviewed by Frodin, 2001), from major taxonomic monographs, and from innumerable revisions of individual genera in the taxonomic literature. The great majority of distribution records were supplied by herbarium specimens; of 70,550 distribution records, 2,560 (3.63%) are from literature sources only, while very many individual distribution records that were represented either by only a few or by doubtfully-determined specimens were further corroborated with literature records.

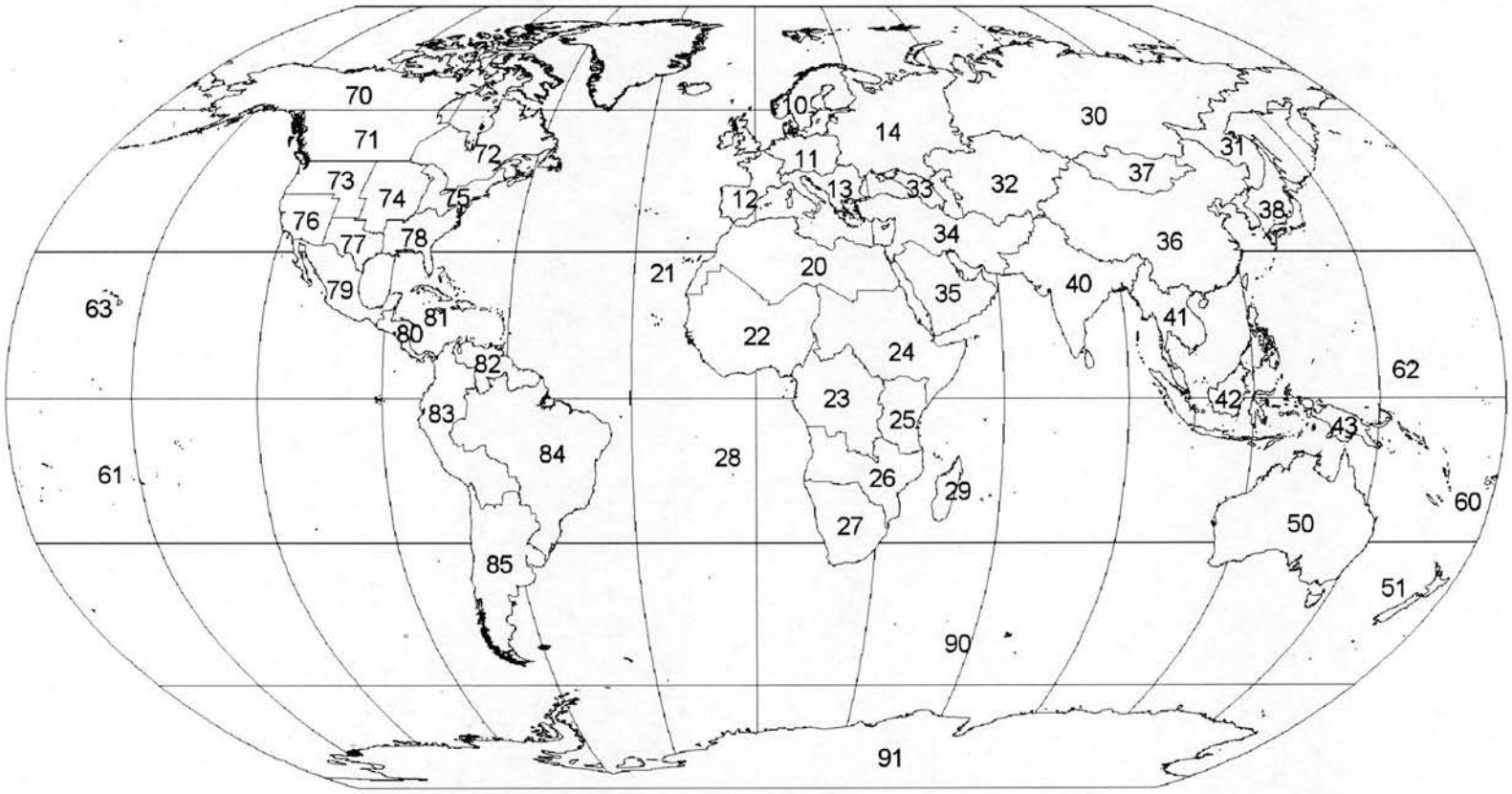


Figure 1.1 Map of the 52 Level 2 Regions of the TDWG World Geographical Scheme for recording plant distributions (Brummitt, 2001). Names for regions are given in Table 1.1, along with the numeric codes used in Figure 1.1; details of geographic composition are given in Appendix 1.

1.2.2 The design of the database

The simple structure of the database is shown in Figure 1.2. Since each genus (held in table Genus) can be present in many different regions, and each region (held in table Geog) can contain many individual genera, this many-many relationship is accommodated within a link table (table GeogGenLink), which stores all of the actual distribution records (70,550 records). Three criteria are recorded for each distribution record: the Voucher status, the Native status and the Extinct status (see Figure 1.2). The Voucher status is most often a verified herbarium specimen held at Kew, but in very rare cases only material preserved in spirit ('spirit') or only separate fruit ('carpological') specimens were present; in a small minority (3.63%) of cases no specimen was present at Kew, and these records were recorded as 'Literature record only'. The Native status was most commonly 'native', but could also be 'introduced', 'doubtfully native' or even, in cases where there were only poorly-identified specimens, 'doubtfully present'.. Specimens were assumed to be native except where the following situations were encountered: a sudden decline in the quantity of collections between areas (introduced taxa are often much less frequently collected); the same species name found in widely differing areas; a large disjunction in the range of a genus. In these situations the status of distribution records was carefully investigated from literature sources. Occasional non-native records were also added only from literature records, but as a rule the provenance of non-native distribution records also relied primarily on herbarium specimens – a systematic literature search for all non-native records was not undertaken. There are 6,793 non-native records in total, together making up 9.63% of the total number of all distribution records. The extinct status simply records whether or not a particular genus recorded from a particular place is subsequently known to have gone extinct there; putative extinctions proved extremely difficult to confirm with any certainty, however, and therefore these represent only very few records.

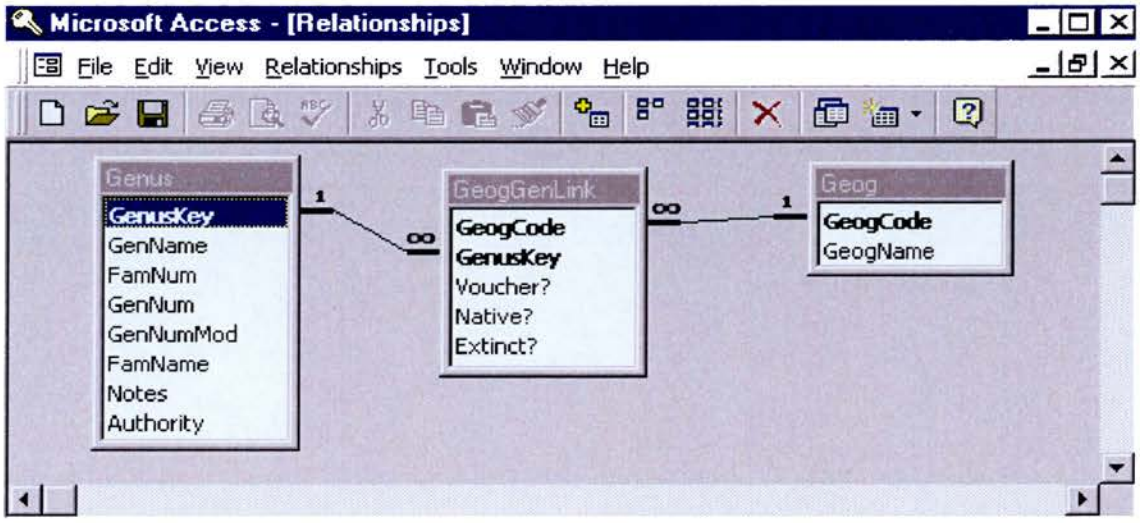


Figure 1.2 Structure of the Distributions of Vascular Plant database and relationships between the tables.

The RBG Kew Herbarium, perhaps uniquely in the world, is judged to be comprehensive enough and well-curated enough to give accurate distributions at this taxonomic and geographic scale, since a single reliably-identified specimen from a particular region is sufficient to stand as a distribution record. Inevitably there is some geographical bias in the numbers of specimens held in the Herbarium (it is better represented in the former tropical British colonies and other areas of more recent activity by RBG Kew, and has less duplication of specimens from temperate regions). However, the collections did prove to be remarkably comprehensive at the global scale. Over 97% of all accepted genera are represented in the Herbarium by at least one specimen, and over 96% of all distribution records were supplied by specimen data. The geographical bias in the collections as a whole therefore will have little impact on the accuracy of the distribution records because the more abundant representation from tropical regions does not count for any 'more' in a presence-absence matrix than does a single specimen from more poorly represented regions. Put another way, the distributions of all genera could conceivably be represented to the same standard of accuracy by a herbarium which contains only 67,990 specimens (the total number of distribution records from herbarium specimens in the data set used here), assuming this hypothetical herbarium could be selected so that there was no redundancy in genus representation (i.e. there was only one specimen of each genus from each region). However, the RBG Kew Herbarium contains some 7 million specimens, the vast majority of which must therefore be additional representation of any one genus from a region. Much of the time during this study has been spent in compiling, editing and subsequently maintaining this database, updating it as needed on an almost-weekly basis, adding new genera and revising existing taxonomy in line with major new publications. Additional new generic names, which are published at the rate of about 110 per year (R.A. Davies, pers. comm.), are entered either as new accepted genera or as synonyms largely on the advice of herbarium staff at Kew. Distributions of genera are then modified as necessary in the light of these taxonomic changes, and compiled again from re-curated herbarium specimens and/or new publications as appropriate.

Collectively, the 52 regions are an amalgam of multiple small countries, or groups of islands, or multiple states within single countries, or single countries in their entirety (see Appendix 1) – all are defined by hard political boundaries, and consequently none are of exactly equivalent size (see Table 1.1) or shape (see Figure 1.1). One of the possible outputs from this database is therefore a nominal presence/absence matrix of 52 regions *x* 14,724 genera. It is on these data that the work presented in this thesis is based, although the analysis is focused exclusively on angiosperm taxa (14,304 genera), since angiosperms represent a large and well supported monophyletic crown group (Soltis *et al.*, 1999; Savolainen *et al.*, 2000), increasing in number of species (Niklas, 1988; Niklas & Tiffney, 1994) and for the last 60 Myr dominating terrestrial ecosystems and comprising the majority of plant species (Lidgard & Crane, 1990). Pteridophytes and gymnosperms are both formerly more diverse groups now restricted in

size, and many pteridophyte genera in particular show highly scattered and irregular distributions when compared with angiosperm genera, presumably due to their spores being easily wind-dispersed (Tryon & Lugardon, 1990), which makes the interpretation of the kind of analyses presented here considerably more difficult. Also, although the introduction and particularly the spread of non-native taxa is an important area of study (Cronk & Fuller, 2001), only native occurrences of angiosperm taxa have been considered.

The geographical resolution of the regions used here is extremely coarse, yet the database nevertheless represents a rich source of biogeographic data. Many detailed local and regional Floras have been written (see Frodin, 2001, for a comprehensive survey), and many detailed distribution records compiled and biogeographic studies performed for particular taxa or areas, but there is no comparable data source on a global scale for plants. In addition to producing the matrix of scored distribution records, for many genera with small distributions these are augmented by the text string. For example, the newly-described genus *Guinetia* L.Rico & M.Sousa (Rico Arce *et al.*, 1999) is known only from western Oaxaca in Mexico; though within the TDWG Level 2 regions it can only be scored as endemic to Mexico, the textual notes record its more local distribution. For widespread taxa, however, distributions which might be summarised as just 'pantropical' or 'throughout north temperate regions' can be more accurately recorded by the particular set of regions in which they occur, since there is not necessarily a single 'pantropical' or 'north temperate' distribution (see Chapter 6). Though previous global studies of plant biogeography have been attempted (e.g. Good, 1974; Raven & Axelrod, 1974; Takhtajan, 1986) this database is regarded as more comprehensive than the data on which those studies have been based, and the main strength of this database is the comprehensive treatment of every genus in a standardised way.

TDWG Region		Area (km^2)	Latitude°
10	Northern Europe	1620215	49.91
11	Middle Europe	1080825	45.75
12	Southwestern Europe	1158137	35.91
13	Southeastern Europe	1056193	34.81
14	East Europe	4690405	44.38
20	Northern Africa	5713373	18.98
21	Macaronesia	13998	14.81
22	West Tropical Africa	6063089	4.27
23	West-Central Tropical Africa	4120263	0
24	Northeast Tropical Africa	5690458	0
25	East Tropical Africa	1773068	0
26	South Tropical Africa	3298998	-5.85
27	Southern Africa	2675954	-16.95
28	Middle Atlantic Ocean	232	-7.88
29	Western Indian Ocean	602411	-4.28
30	Siberia	9855164	49.08
31	Russian Far East	3063473	42.29
32	Central Asia	3974323	35.14
33	Caucasus	438212	38.39
34	Western Asia	3845970	25.07
35	Arabian Peninsula	2789669	12.59
36	China	9268483	18.17
37	Mongolia	1558842	41.58
38	Eastern Asia	628999	21.93
40	Indian Subcontinent	4423703	5.94
41	Indo-China	1932370	5.63
42	Malesia	2128984	0
43	Papuasia	906147	-0.01
50	Australia	7704687	-10.05
51	New Zealand	268760	-29.22
60	Southwestern Pacific	57339	0
61	South-Central Pacific	4081	0
62	Northwestern Pacific	2640	5.26
63	North-Central Pacific	16920	16.72
70	Subarctic America	7526750	51.21
71	Western Canada	2899891	48.30
72	Eastern Canada	3111341	41.68
73	Northwestern U.S.A.	1544658	37.00
74	North-Central U.S.A.	1842281	33.65
75	Northeastern U.S.A.	968634	37.20
76	Southwestern U.S.A.	1207900	31.33
77	South-Central U.S.A.	1000684	25.84
78	Southeastern U.S.A.	1371180	24.54
79	Mexico	1961910	14.55
80	Central America	519581	3.97
81	Caribbean	235202	10.04
82	Northern South America	1355352	0.65
83	Western South America	3783883	0
84	Brazil	8506150	0
85	Southern South America	4104401	-17.51
90	Subantarctic Islands	24542	-37.05
91	Antarctic Continent	12093000	-60.75

Table 1.1 The 52 Level 2 TDWG regions used in this thesis and their numeric codes; the area (km^2) and latitude (minimum distance from the equator in decimal degrees; negative values indicate the Southern Hemisphere) are also given for each region. A full breakdown for each region is given in Appendix 1.

1.2.3 Drawbacks with the database design

Three obvious criticisms can be levelled at the design of the database and which may be thought detrimental to the analyses in this thesis. Firstly, the regions by which distributions have been scored are very large compared to the actual distributions of most taxa – the nine most common distributions are all single-region endemics (see Chapter 3). Secondly, the regions have geo-political boundaries, which generally do not follow any recognised biogeographical, ecological or climatic pattern. Thirdly, the regions vary greatly in size, from just over 200 square kilometres (Region 28, Middle Atlantic Ocean – the islands of St. Helena and Ascension) to well over 12 million square kilometres (Region 91, Antarctic Continent), complicating comparisons between them (see Chapter 3). With regard to the first criticism, however, it was often difficult enough to establish from the scattered botanical literature whether or not a particular genus (in its modern conception) was actually present within a particular region, and was actually indigenous there; the problems with compiling, and maintaining, more detailed geographical information for many poorly-known taxa would have been very much greater.

The second of these drawbacks can be regarded as a strength of the database. The aim is first to establish what actually are the distribution patterns; any biogeographic, ecological or climatic interpretations of the data must come secondarily, without this having influenced the compilation of the data. If data are scored by pre-determined biomes or floristic regions, any subsequent interpretation of the validity of those regions is then hopelessly circular. The boundary between biomes such as the Amazonian forest and the Brazilian *cerrado*, for example, is not well defined (Furley *et al.*, 1992), and accurate information about the habitat in which herbarium specimens have been collected is not routinely noted on the specimen labels. One cannot compare floristic diversity of the *cerrado* with that of Amazonia, for example, if one has defined these areas poorly to begin with, and furthermore many 'Amazonian' taxa penetrate the '*cerrado*' (Oliveira-Filho & Ratter, 1995). Geo-political boundaries, for all their arbitrariness, are much more sharply defined – a plant either is growing in Brazil or it is not.

If one is to score distributions without *a priori* bias, there are two possible methods: using geo-political boundaries or using uniformly sized grid-squares. It could be argued that using uniform grid-squares, as in the WORLDMAP program (Williams, 1994), would have been the most appropriate method. However, in a historical herbarium such as Kew – probably the only such in the world where this project could even have been undertaken – the majority of early collections have such scanty locality information that it would have often been impossible or at best extremely time-consuming to have assigned each specimen to individual grid-squares. Only a small proportion of even recently-collected specimens routinely include latitude and longitude details (though this is now changing with the increasingly-standard use of GPS), but they can at least be expected to include the country of origin. It was

decided at the outset of this project that distributions should be compiled primarily from specimen records – literature records are simply not comprehensive enough to record distributions in a systematic way, whereas the Herbarium at Kew houses at least some material of more than 97% of accepted genera. This decision was borne out in compiling the data, when it was more often additional specimen records which augmented distributions quoted in the literature, and not *vice versa*.

1.3 Extracting, manipulating and analysing data

1.3.1 The data repository: Sybase

Once a database of distributions had been compiled and provisionally edited in Microsoft Access, this was uploaded onto the RBG Kew corporate server and subsequently edited and maintained as a database within the database management system Sybase, through a separate management client application. Sybase is a much more robust system for a multi-user environment; however, because of the lack of a freely-customisable querying interface for Sybase, querying for subsets of these data would have become cumbersome and time-consuming and required Visual Basic programming skills. Therefore, the original Microsoft Access database was retained, with the data tables in Figure 1.1 linked externally to the Sybase database. This meant that the data were read from the Sybase server but through Microsoft Access, and all querying could be done with simple Access queries being written as required, and saved within the Access database. Running newly-written or pre-saved queries therefore always retrieves the up to date records from the Sybase database.

1.3.2 Querying for data with Microsoft Access

Microsoft Access Query Design View allows the user to specify fields and criteria in order to retrieve certain groups or subsets of records from the database without affecting the integrity of the database as a whole. Queries can then be saved individually and re-run whenever occasion demands. For example, Figure 1.3 below shows the query which will extract numbers of genera of angiosperms for each TDWG Level 2 Region. Many dozens of such queries were written and saved within the Access database. Analysis then proceeded with the results from each of these queries, and the queries and analyses then re-run and results updated prior to submitting this thesis. All the numerous other queries in this thesis for extracting sets of data for analysis were written in a similar way to Figure 1.3, sometimes with secondary or tertiary queries themselves built on earlier queries in order to extract the relevant information for a given analysis.

In Figure 1.3, the GeogCode field in the Geog table lists each of the TDWG Level 2 Regions. The field GenusKey from the GeogGenLink table contains the unique identifier for each genus in the database, and the additional expression Count simply returns the total number of genus keys (= number of genera) for each region. The criterion "N" for the Native field in the GeogGenLink table restricts the query to distribution records where a genus is known to be native to that region. The expression 'Where <201' for the FamNum field in the Genus table only returns distributions where the number of the associated family is less than 201, which, following the numbering system for vascular plant families adopted in the RBG Kew Herbarium (which is modified from system of Bentham & Hooker), will exclude all gymnosperm and pteridophyte families (which are assigned family numbers above 200). Given that this query is to list each TDWG Region and the number of genera within it, it is not necessary to display these last two fields.

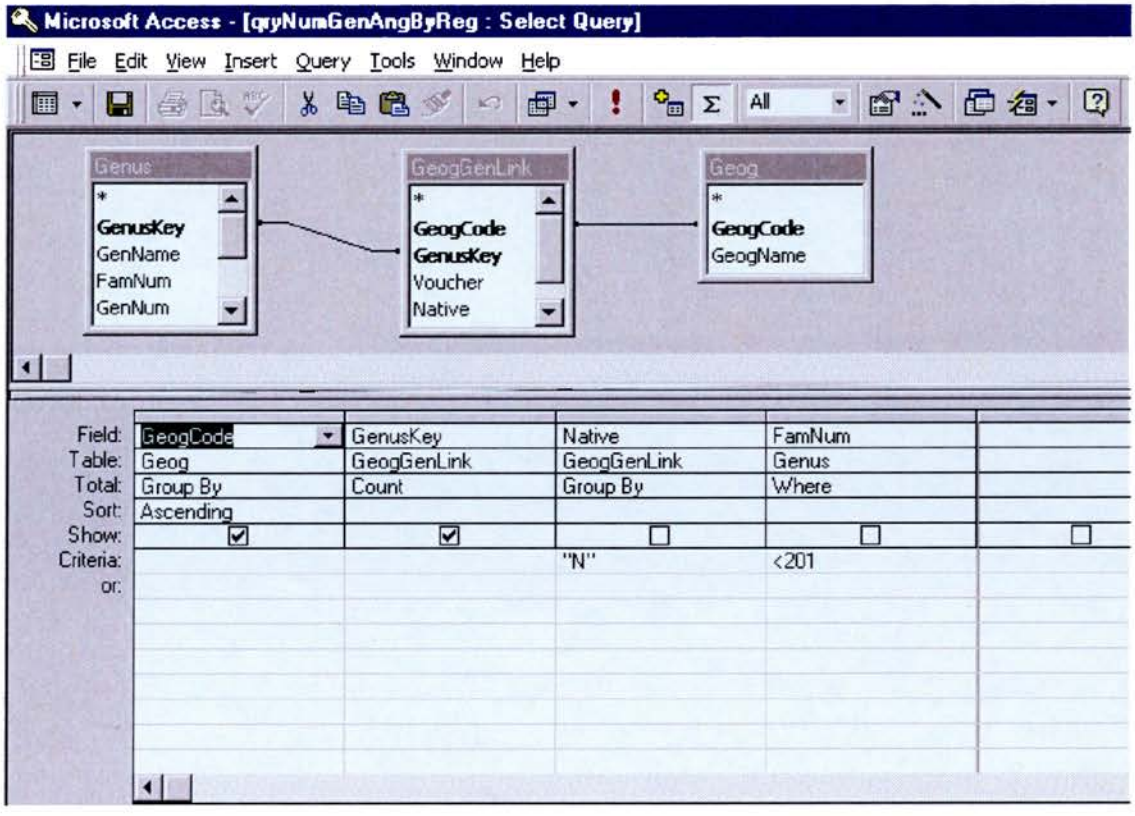


Figure 1.3 A select query in Microsoft Access to list numbers of genera of angiosperms in each TDWG Level 2 Region.

1.3.3 Linking queries to Microsoft Excel and ArcView GIS for analysis

The advantage of commonly available software packages is their inter-operability: they can all 'talk' to each other. Queries written in Microsoft Access were read in Microsoft Excel through Microsoft Query and in ArcView GIS via SQL Connect, both of which preserve the link to the 'live' Sybase data through Microsoft Access. It was hoped to be able to perform all the analyses with 'live' data and so have the results easily updated as the database was continually edited and maintained. For simple queries such as in Figure 1.3 this was the case. Unfortunately, however, multivariate statistical analyses are too complex and time-consuming to keep 'live' and packages such as NTSYS and PC-Ord, which were used extensively for this thesis, are not sufficiently commonly used to be inter-operable with other software, and in these cases data had to be exported from Microsoft Access, re-formatted in Microsoft Excel, and then imported into the relevant program for analysis. Results of analyses then often had to be re-imported into Microsoft Excel to format tables and produce graphs producing the thesis in Microsoft Word. Having to continually query, re-format and export data from one computer package to another, often several times within a single analysis, was a very frustrating aspect of all of this work. Figure 1.4 below shows the schematic relationships between the different software packages and their respective functions.

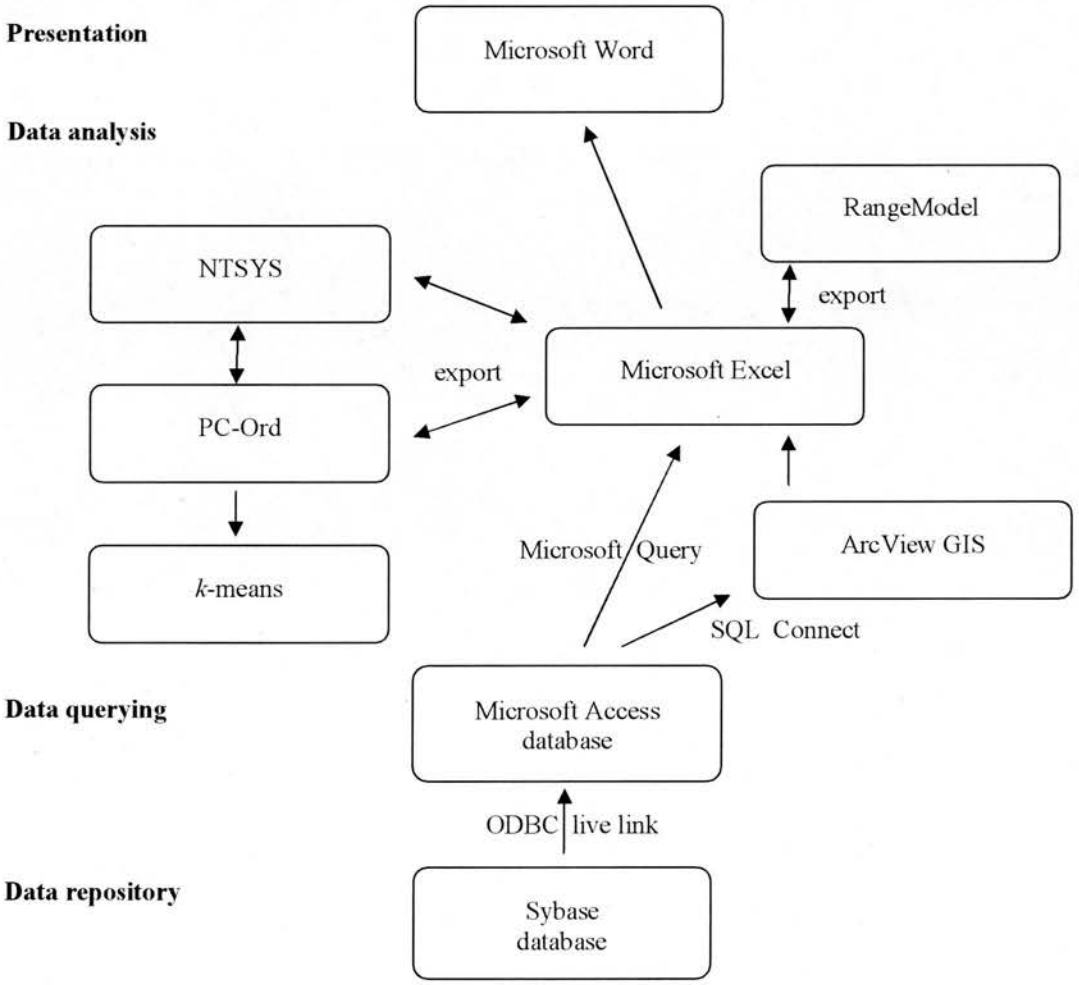


Figure 1.4 Schematic view of the flow of data between different software packages.

1.4 What is a genus?

Given that this thesis is primarily a study of worldwide distribution patterns of angiosperm genera, it is instructive to first spend some time considering the concept of a 'genus', and the historical and current approaches to identifying and delimiting genera. Much effort has been, and continues to be, spent on the philosophical question of what constitutes a species (e.g. Ereshefsky, 1992), but far less attention has been spent on the question of what does a genus, if anything, really represent.

1.4.1 A genus is a nomenclatural category ...

Linnaeus owes his unique place in the history of the development of biology to an innovation in the convention of naming upon which he did not set great store, and which was certainly not the primary intent of his work – the consistent use of binomials consisting of linked genus names and specific epithets to identify all species then known (Stearn, 1957). Originally, binomial names were included as marginal annotations and intended as merely a shorthand reference system to complement the plurinomial system of *nomina specifica legitima* that concisely expressed the diagnostic characters of a species (Stearn, 1957). The rapid adoption of the Linnaean binomial system was a consequence of its great practical utility. The binomial *Physalis angulata* L. says less about the species than does the phrase-name *Physalis annua ramosissima, ramis angulosis glabris, foliis dentato-serratis*, but it is much easier to write and remember. A genus for Linnaeus constituted a particular level of distinctness, a collection of species differing in consistent but minor ways from each other, but differing in more fundamental ways from other such genera.

Consistent application of the binomial system, however, demanded that even very distinct species should be named the same way, when for monotypic genera there is no difference in composition between the genus taxon or the species taxon. Linnaeus thus created two logically distinct ranks for what had previously been a single concept. For the monotypic genus *Imperatoria* L., for example, a specific name had previously been considered to be redundant and illogical. Linnaeus clearly implied the distinction of generic and specific characters and paved the way for the later development by others of a hierarchical system of classification in which ranks are inter-nesting (species within genera, genera within families, etc.) but taxa are mutually exclusive at any one taxonomic rank (Brummitt, 1997; Knox, 1998). However, the concept of the *genus*, literally just meaning a 'kind', can be traced back at least to the writings of Pliny, who referred to *anagallis mas* and *anagallis femina*, and can also be found in modern-day folk taxonomies, for example the Malay names *jambu ayer*, *jambu bol* and *jambu chili* for different species of

the genus *Syzygium* L. (Steam, 1957). Linnaeus' concept of the genus was influenced by the work of previous botanists such as Tournefort (Steam, 1957), but he makes the following comment in his *Genera Plantarum* of 1737:

> "*Accordingly genera are as many as there are common close attributes of distinct species according to which they were created in the beginning; revelation, discovery and observation confirm this. Hence* all genera and species are natural."
>
> Linnaeus, 1737a; italics in the original

Linnaeus believed species had been created by God and were characterised by an Aristotelian 'essence', fixed and unchanging:

> "*... there are two kinds of difference between plants: one a true difference, the diversity produced by the all-wise hand of the Creator* [the 'essence']*; but the other, variation in the outside shell, the work of nature in a sportive mood.*"
>
> Linnaeus, 1737b

1.4.2 ... and also an evolutionary unit?

By the time of Darwin, however, the creationist view of the world upon which 18th century science had been based was under threat. Darwin's theory of evolution broke with the essentialist tradition in two major ways: firstly, species were anything but fixed and unchanging; secondly, accidental variation ("*the work of nature in a sportive mood*") was the very means by which natural selection operated — rather than being an obstacle to a true understanding of the Creation, variation instead became the key to understanding the process of evolution. Evolution became the underlying unifying principle of taxonomy which would result in a natural phylogenetic classification, reflecting the patterns of descent of species. But if taxa evolved through time into other taxa, how could they be defined by 'essential' properties? The only basis for natural classification is evolutionary theory, but, according to evolutionary theory, species were continually subject to natural selection and thus evolved gradually. If species evolved gradually they cannot be defined by means of a single fixed, unchanging property or set of properties (their 'essence'). If taxa could not be defined by essences they cannot be natural kinds, and classification then becomes completely arbitrary. Belief in evolution thus confronted taxonomists with a dilemma: if they accepted evolutionary theory, they had to give up hope of ever having real taxa; but if they wished to retain real

taxa, they had to give up any hope of ever having a natural classification (Hull, 1965). The sense of frustration is captured in the following quote from Darwin:

> "*We shall have to treat species in the same manner as naturalists treat genera, who admit that genera are merely artificial combinations made for convenience*"
>
> Darwin, 1859: page 447

Since the time of Linnaeus, therefore, the concept of the genus being a real, natural, fixed element of nature had given way to an acceptance of taxonomy, as to some extent, containing an inevitable element of subjectivity. Darwin's theory of evolution denied the existence of species' essences, but saw species and higher taxa as essence-less classes. However, the view of species as classes of organisms is itself contrary to belief in evolution (Hull, 1965; Ghiselin, 1974; Nelson *et al.*, 2003). If species are classes then each class must be defined by the possession of a unique character or combination of characters, and therefore any organism possessing those characters must by definition belong to that class. If, however unlikely it may be, the character or characters defining a species were to be evolved independently by an unrelated organism, that organism would also, under the definition of that class, have to be included within that species. Furthermore, if the characters defining the species were to change over time, through evolution, then a new species would have to be recognised, with a new set of characters.

1.4.3 Taxa are 'individuals'

Belief in evolution thus seemed to deny the reality of taxa. The solution to this impasse came with the insight that species are not natural kinds, or *classes*, but instead are *individuals* (Ghiselin, 1974; Hull, 1978; Mayr, 1988). In logic, individuals are historical entities with definite beginnings and ends ('spatio-temporally bounded') which show internal cohesion and continuity. Species are bounded in space and time by speciation events (and ultimately, by extinction events). This also helped to clarify the distinction between the taxonomic category and the constituent taxa: taxa are individuals; the taxonomic categories to which they are assigned are classes (Mayr, 1976). If taxa are viewed as individuals, taxon definition is no longer based purely on characters but explicitly on their evolutionary history – on phylogenetic relationships. Taxa higher than the rank of species, such as genera and families, are also seen as historical individuals, provided that they stem from a single common ancestor and include all the descendants of that ancestor (monophyletic *sensu* Hennig, 1966) (Cracraft, 1987; Ereshefsky, 1991). In the logical sense of having a definite origin, therefore, genera and other higher taxa are 'real', provided they are determined as being monophyletic (clades). This reasoning has given additional credence to the

use of cladistic methodology not just for determining taxon relationships but also for constructing phylogenetic classifications.

1.4.4 Testing the reality of genera

The relative numbers of higher taxa of different sizes have also been used to help justify the structure of the Linnean hierarchy. A frequency distribution of taxonomic categories is always strongly skewed: there are many monotypic taxa, and very few, very large taxa (Clayton, 1972, 1974; Cronk, 1989; Scotland & Sanderson, 2004; see also Chapter 3). Plotting this frequency distribution results in a characteristic 'hollow curve', with a modal value for taxon size of 1. This is observed at successive taxonomic levels (genera most frequently contain only a single species, families most frequently contain only a single genus, etc.) (Clayton, 1972, 1974; Cronk, 1989) and has prompted some biologists to claim that taxonomic systems show a 'fractal' pattern of taxon-size distribution, with self-similarity repeated at different scales (Dial & Marzluff, 1989; Burlando, 1990; Minelli *et al.*, 1991). By drawing analogies between taxon-size distribution and natural fractal patterns of biological growth and development (for example, the branching pattern of a tree), these authors have expressed their belief in the 'reality' of the inter-nesting hierarchy of traditional Linnaean nomenclature. Such strongly-skewed frequency distributions are shown not just by taxon size but also by spatial taxon distribution (many narrowly distributed taxa and few very widely distributed taxa) (Colwell & Lees, 2000; Gaston, 2003; see also Chapter 3) and temporal taxon distribution (many short-lived taxa and few very long-lived taxa) (Rosenzweig, 1995), further suggesting that the hollow curve is actually a real phenomenon and not just an artefact of the ways in which taxonomists have erected taxa (but see also Scotland & Sanderson, 2004).

The occurrence of the hollow curve on both spatial and temporal scales suggests a correlation between the two, which might be caused by a frequency-directed evolutionary process. Bateman & DiMichele (1994) have proposed the existence of widespread, long-living 'stem species', each of which, through chance macro-mutations in regulatory developmental genes, is the progenitor of many sequential, reproductively isolated, short-lived, localised daughter species. These daughter species would account for the greater numbers of taxa over short spatial and temporal scales that result in the hollow-curve. However, the developmental basis for this possible mechanism has not yet been established and this 'supra-Darwinian' model must be regarded as tentative. Studies of the evolutionary divergence between developmental genes suggest that duplication of regulatory genes or their associated regulatory elements, which allow subsequent independent evolution of one of the duplicates in a new developmental pathway, may be a significant factor behind large-scale evolutionary change between lineages. Most of this research has to date sought to explain macro-evolutionary patterns in 'deep time' occurring in stem lineages at the

base of major clades, for example between different animal phyla (e.g. Erwin, 2000; Shubin & Marshall, 2000, and references therein), rather than between closely-related species. Stern (1998), however, found that small changes in the activity pattern of the *Drosophila* homeobox gene *Ubx* correlate with differences between species in that genus; and a recent study of the origin of maize (Doebley & Wang, 1997) suggests that its evolution from ancestral wild teosinte may involve changes in as few as five genetic loci, though this may be a function of the strong artificial selection in crop plant origins (Shubin & Marshall, 2000).

1.4.5 Delimiting natural taxa

The hollow curve frequency distribution suggests that the prevalence of monotypic taxa is a natural phenomenon. However, in spite of these hollow-curves at both spatial and temporal scales, it is not known whether it is the same taxa which are both widespread and long-lived. Since the hollow-curve distribution has been found at all taxonomic levels, a mechanism to account for this pattern must equally apply at all taxonomic levels. Also, the frequency of monotypic genera detected has been found to be above that expected from models of hollow curve frequency distributions (Clayton, 1983; Dial & Marzluff, 1989; Cardillo *et al.*, 2003). Explanations suggesting rapid adaptive radiation in large taxa together with prevalent extinction leaving behind numerous monotypic taxa have been proposed to account for this discrepancy (Dial & Marzluff, 1989), although Clayton (1983) suggested that the ready creation of monotypic taxa for distinctive species by 'tidy-minded' taxonomists has inflated the number beyond its 'natural' level. The desire to have easily identifiable and morphologically coherent higher taxonomic units leads many taxonomists to exclude species which destroy the unity of higher taxa, and so creates an abundance of monotypic taxa, although these excluded taxa may nonetheless make other taxa paraphyletic. Scotland & Sanderson (2004) proposed a model which generates even greater numbers of monotypic taxa and also larger big taxa than are found in actual hollow-curves in nature simply through random character change across a phylogenetic tree, but they similarly suggested that this discrepancy has a taxonomic rather than a biological cause: taxonomists tend to neglect both monotypic genera and large, taxonomically difficult taxa, so in fact under-inflating the 'true' figures.

The standard methodology for delimiting natural taxa is to use phylogenetic systematics, based on the principle of monophyletic taxa that include all descendents from a single common ancestor (Hennig, 1966). However, although objectively-delimited taxa can be identified in this way, this does not solve the question of the rank at which to assign taxonomic status. Given that hollow-curve frequency distributions show that monotypic taxa are the most prevalent, when should a monophyletic lineage be recognised as a monotypic genus, or when might this perhaps be worthy of monotypic family status?

Backlund and Bremer (1998) tackled the problem of monotypic taxa at the family level with regard to the doubtfully-accepted family Triplostegiaceae Airy Shaw, which has at various times been included in either the Valerianaceae or the Dipsacaceae. Backlund and Bremer (1998) argue that monotypic taxa should generally be avoided since they do not provide any phylogenetic information for the constituent species and create 'redundancy' in classification. Put another way, it is more meaningful for an evolutionary understanding to know what a species is related to than it is to know how distinctive it is. Governed by a desire to understand evolutionary relationships and processes, rather than just the Linnaean ideal of cataloguing and documenting biological diversity, Backlund and Bremer (1998) thus emphasize monophyly as the primary principle of classification (see also Angiosperm Phylogeny Group, 1998, 2003). They also admit secondary principles, which are:

1. Maximising stability of the classification
2. Maximising phylogenetic information (minimising redundancy)
3. Maximising support for monophyly
4. Maximising ease of identification

This application of the criterion of monophyly demonstrates the situation described by Scotland & Sanderson (2004) of a reduction in monophyletic taxa from 'natural' levels. Scotland & Sanderson (2004), however, acknowledge that even if taxa can be objectively delimited there are not really any objective criteria for ranking higher taxa, and others admit that 'discussion as to whether a widely-accepted monophyletic group should be a superorder, order, suborder or family is largely vacuous because this will always be an arbitrary decision' (Angiosperm Phylogeny Group, 1998). This creates the problem, when comparing taxa across clades, of deciding whether or not the same rank has been applied consistently between lineages, i.e. is a genus in one group equivalent to a genus in another group. One possible method for comparing like clades with like might be node height (number of nodes from the base of the tree; Barraclough *et al.*, 1999); however, where the sampling for a group is incomplete the topology of the cladogram (and hence node height) will depend on the taxa represented within the study (Graybeal, 1998; Hillis, 1998; but see also Poe, 1998). As yet, no consensus has been reached on criteria to be used for ranking clades and so it is impossible to say if all genera really represent equivalent evolutionary units.

1.4.6 Importance of generic placement

The assignment of a species to a genus remains of fundamental importance to biological classification simply because of the universal usage of Linnaean binomial nomenclature of genus names

with species epithets. The special importance of the binomial nomenclatural ranks is captured in the following quotation from Davis and Heywood:

> "*When in doubt whether to accord generic rank to a group, there is much to be said for the* subgenus *as a suitable category; it draws attention to the group in the classification and at the same time allows people to use the old binomial.*"
>
> Davis & Heywood, 1963, page 106; italics in the original

The taxonomic hierarchy of inter-nesting ranks pre-dates the development of evolutionary theory, with the genus and species of Linnaeus' short-hand binomial nomenclature subsequently being augmented by Jussieu's concept of plant families, and later orders, classes, phyla and kingdoms. Since it was conceived under a non-evolutionary world-view, taxa at each rank were originally intended as purely phenetic concepts, assumed to somehow constitute an equivalent level of distinctness from each other; evolutionary interpretations came later. For example, the family Rubiaceae was created for species with opposite entire leaves, interpetiolar stipules, tubular corollas and inferior fruits well before these species were known to be an evolutionary lineage that shared these characters due to their descent from a common ancestor. The Linnean hierarchy is thus seen now as an imperfect way of representing what we have since learned are patterns of shared ancestry – the evidence of evolutionary relationships between taxa. The development of phylogenetic systematics (Hennig, 1966) offered the chance to rigorously uncover these evolutionary relationships and construct more objective classifications of taxa. To date, however, there is still no consensus on what criteria to use to assign a particular rank to a clade, and as yet not all genera have been assessed phylogenetically.

That different genera can be treated as somehow 'equivalent' to each other is therefore a central assumption in this thesis. However, it is impossible to really say whether or not all the genera used in these analyses may actually be 'equivalent' to each other in evolutionary status. The argument used for this thesis is that the 'hollow curve' distribution, which is repeated at different taxonomic levels, is a representation of a fractal-like underlying phylogenetic structure (Minelli *et al.* 1990). Each genus is assumed to represent a portion of a larger phylogenetic tree, with an independent evolutionary and biogeographic history. It still remains a somewhat arbitrary decision as to which node represents a taxon worthy of formal recognition (although of course that node may well mark the position of a readily-observable morphological synapomorphy which would serve as a character for generic recognition). At any particular level, therefore, although the status of individual taxa may change, the overall frequency distribution of taxa should not. That is, some genera of uncertain status may be sunk, but this will be compensated by other genera becoming newly-recognised. For the analyses presented here, therefore, this uncertainty in the status of genera may not matter (see Chapter 3).

1.5 Plant distribution patterns

The most striking aspect about the distributions of plant species on the earth is their non-random pattern. As one distinguished botanist put it:

'*Les plantes ne sont pas jetées au hasard sur la terre.*' (Germain de Saint-Pierre)

Of course, this begs the question 'Why not?'. The study of biogeography involves the identification and explanation of the distributions of organisms. Patterns are not only non-random, but often will be coincident: different species of plants are found in the same places (Good, 1974; Stott, 1981). While no two distributions are ever exactly the same, at the scale of individual plants, this fact is significant because it suggests a common cause behind the distribution pattern. The comparison of similar distribution patterns in unrelated taxa is the essence of biogeography (Croizat, 1964; Humphries and Parenti, 1999); otherwise it becomes little more than an endless list of disparate facts, 'a science of the rare, the mysterious and the miraculous' (Nelson, 1978), and no meaningful generalisations can be made.

Distributions can be recorded in different ways. All plant species have particular ecological requirements, and most are confined to certain types of habitat. This is not the same as their geographical distribution, however. For example, the genus *Rhabdodendron* Gilg. & Pilg. (Rhabdodendraceae) is only found in the evergreen rain forests of central Amazonia; but rain forests occur more widely in South America than just in central Amazonia, and are also found in SE. Asia and in west and central Africa, whereas *Rhabdodendron* is not. The distinction between ecological and geographical factors in plant distributions was first made by De Candolle in 1820, with his concepts of *stations* (habitats) and *habitations* (distributions). De Candolle commented that 'the confusion of these two classes of ideas is one of the causes that have most retarded the science, and that have prevented it from acquiring exactitude.' The modern situation is completely the reverse: there is almost total distinction between ecological biogeographers on the one hand and historical biogeographers on the other (Nelson & Platnick, 1981; Humphries & Parenti, 1999), probably to the detriment of both (Myers & Giller, 1988).

The fundamental question of biogeography – *why* is *what* found *where*? – therefore relies on the prior identification of two distinct units of analysis: the *what*, or the particular taxon involved; and the *where*, or the area in which it is found. Essentially, there are two possible explanations: either the organisms moved and the areas remained the same, or the organisms remained in place but the areas moved. These two contrasting explanations go by the names 'dispersal' and 'vicariance', respectively.

Although obviously two sides of the same coin – for example, both the organisms and the areas might have moved over time – this distinction polarised debate in biogeography for 100 years. Two good friends, Darwin and Hooker, were on opposite sides of it. Darwin believed that:

> *"The endurance of each species and group of species is continuous in time; ... so in space, it certainly is the general rule that the area inhabited by a single species or by a group of species is continuous and the exceptions, which are not rare ... be accounted for by former migrations under different circumstances, or through occasional means of transport, or by the species having become extinct in the intermediate tracts ..."*
>
> Darwin, 1859, page 409

Hooker, travelling on the *Erebus* and *Terror* Antarctic voyage, was more impressed by the floristic similarities between the widely separate land masses of the southern temperate regions. Invoking Lyell's principle that each species must have arisen at only one point on the globe, Hooker (1853) argued that the areas visited by the Antarctic voyage must have, in the past, formed a single land mass occupying a continent larger than that of the Antarctic Ocean. Though there was no evidence for it, Hooker thus favoured a vicariant explanation: there had been former land connections between New Zealand and South America, for example, which explained their floristic similarities.

Alfred Russell Wallace, the co-originator of the theory of natural selection who went on to make important studies of animal distribution patterns (e.g. Wallace, 1876), like Darwin also believed in the importance of dispersal:

> *"The biological causes* [of distributions] *are mainly of two kinds — firstly, the constant tendency of all organisms to increase in numbers and to occupy a wider area, and their various powers of dispersion and migration through which, when unchecked, they are enabled to spread widely over the globe; and secondly, those laws of evolution and extinction which determine the manner in which groups of organisms arise and grow, reach their maximum, and then dwindle away, often breaking up into separate portions which may survive in remote regions."*
>
> Wallace, 1880, pp. 531-532

1.5.1 Factors influencing plant distributions

Factors influencing the distribution of plants are many and varied. They have been identified and classified by numerous authors going back to Schimper (1903). Though they differ in some details, they are broadly comparable; one may as well start with Good (1974):

1. Place and time of origin.
2. Distribution of climatic values (temperature, rainfall, light, wind):
    a. in the present.
    b. in the past.
3. Distribution of edaphic values (physical, chemical, physiographic):
    a. in the present.
    b. in the past.
4. Potentialities for dispersal.
5. Configuration of land and sea:
    a. in the present.
    b. in the past.
6. Influences exerted by other plants:
    a. direct competition.
    b. indirect influences.
7. Human influences.

The above list would seem to account for all possible scenarios, and no biogeographer would say that any one of these factors was never important. However, the central issue for both Darwin and Wallace was that, since different species had arisen as a result of different selection pressures and thus had different ecological tolerances (Good's 'climatic' and 'edaphic' values) and dispersal capabilities, and since they thought that dispersal occurred as migration, every group of organisms would then have its own distributional history. Thus there was not thought to be a single biogeographic pattern to reconstruct, only a multitude of individual histories of dispersal (Humphries and Parenti, 1999). Any similarities in species composition between separate areas of the globe, or similarities in the distributions between different taxa, were interpreted as merely an artefact of coincidental but independent migrations. This philosophical stance, which assumes *a priori* that each individual taxon has a unique biogeographical explanation and thus effectively precludes the discovery of more general biogeographical hypotheses, has continued until the present (e.g. Raven & Axelrod, 1974; Cox & Moore, 1993). The search for common explanations to shared, repeating patterns, however, is the essence of most other branches of science. In the search for common explanations to repeating patterns of distribution, therefore, it is useful to consider the similarities and differences between general types of plant distribution.

1.6 Types of plant distribution

The complexity of plant distribution patterns can be usefully conceptualised in terms of three possible extremes of distribution. Any taxon can be either: a) found only in a very small area (an endemic taxon); b) found throughout a very large area (a widespread taxon); or c) found in more than one area (a disjunct taxon) (Stott, 1981). Distribution patterns show characteristic frequency distributions: the famous 'hollow curve' on which Willis placed so much importance (e.g. Willis, 1922; see also Colwell & Lees, 2000; Gaston 2003). The majority of species have localised distributions, found only over a small area; only a few species are very widespread. Overall, however, there is no obvious discontinuity in range-size which would justify using the contrasting terms 'endemic' and 'widespread', while gaps in the range of a taxon may cause that distribution to be described as 'disjunct', depending on the distance between the disjunct populations and the overall range-size of the taxon. The above trichotomy is obviously a simplification, therefore, and is obviously also a product of the scale at which these questions are framed. A genus such as *Caesalpinia* L. would be described as 'pantropical' as it occurs throughout tropical (and into warm temperate) regions. It is very widespread at a global scale. However, tropical regions are themselves not contiguous, so *Caesalpinia* may also be described as disjunct between the different areas of the tropics (c.f. Thorne, 1972). It may also be described as endemic to tropical and warm temperate regions. So a single taxon may be widespread, yet also disjunct, and still endemic. To explain 'why' organisms are found in particular places, what is important is not so much the idea of where they are as that they are there and not elsewhere. However, the first stage in the analysis of plant distribution patterns is to determine exactly 'what' is actually 'where'.

1.6.1 Endemic taxa

An endemic taxon is one which is confined to a particular place. Traditionally in biogeography the term is applied to taxa with restricted ranges at the regional, continental or world scales (Stott, 1981). For example, the Sierra de la Neblina on the borders of southern Venezuela and northwestern Brazil has several genera endemic to this one mountain (e.g. *Neblinaea* Maguire & Wurdack (Compositae), *Neblinanthera* Wurdack (Melastomataceae), *Neblinathamnus* Steyerm. (Rubiaceae)). The genera *Heliamphora* Benth. (Sarraceniaceae) and *Saccifolium* Maguire & J.M.Pires (Saccifoliaceae) are more broadly endemic to the Guayana Highland region, while genera such as *Bonnetia* Mart. (Theaceae) and *Rhabdodendron* (Rhabdodendraceae) are more broadly endemic still within tropical South America. The importance of applying the term 'endemic' at the appropriate scale is obvious. At the broadest of scales, all organisms are endemic to the earth.

Endemism is a phenomenon with many complex causes, but it is generally conceptualised in terms of two possible types of endemic: neoendemics and palaeoendemics. Neoendemics (progressive or secondary endemics, or autochthonous taxa), are 'new' taxa which have evolved in a particular area from which they cannot or have not yet spread; palaeoendemics (relicts or epibiotics) are 'old' taxa which were formerly more widespread but which are now much more narrowly distributed (Stott, 1981). All species must, in effect, start as neoendemics and finish as palaeoendemics (Richardson, 1978). Taxa which remain with a restricted distribution throughout their existence are termed holoendemic, though the majority of species will undergo the following distribution series (after Richardson, 1978):

origin; expansion; stabilization; diversification; migration/fragmentation;
contraction; relictual phase; extinction

Patterns of generic endemism are studied in Chapter 3.

1.6.2 Floristic regions and areas of endemism

An important concept in the study of plant distribution patterns is that of the 'floristic region'. It is analogous to the 'faunal region' of zoology, formulated by Wallace (1876) based on the work of Sclater (1858). However, the ideas upon which Wallace was drawing were much older, going back to Buffon (1776) and Humboldt (Humboldt and Bonpland, 1805). 'Buffon's Law', based on an observation that no mammal species was common to both South America and Africa, was generalised as the tendency that different, widely separated areas contained different organisms. Humboldt and Bonpland (1805) confirmed that, with plant distributions, certain areas both had characteristic flora (that is, they were without many species common to other areas), but also for these areas there were many other species which revealed floristic similarities between these and other areas. Important in this work is the implicit assumption that the history of organisms and the history of the earth go together.

> "... *geology bases itself on the analogous structure of coastlines, on the similarity of animals inhabiting them and on ocean surroundings. Plant geography furnishes most important material for this kind of research. It can, up to a certain point, determine the islands which, at one time united, have become separated from one another;*"
>
> Humboldt & Bonpland, 1805, page 19

De Candolle (1820) broadened the application of Buffon's Law and proposed a list of 20 'botanical regions', many similar in concept to those later delimited by Engler (Engler & Diels, 1936), Good (1974) and Takhtajan (1986). For De Candolle, dispersal of organisms only accounted for widespread or cosmopolitan species – exceptions to Buffon's Law (Nelson, 1978). Buffon was an early French evolutionist, like Lamarck; though he disagreed with the Lamarckian evolutionary mechanism, and

had no better suggestion himself, citing only un-named 'external circumstances' (analogous to Darwinian natural selection). Species occurred in particular floristic regions due to historical factors: namely, as we now know, the evolution of the species and also the evolution of the planet.

As did Humboldt, De Candolle chiefly considered two phenomena: 1) endemic genera, with many species confined to one region (Buffon's Law); but also 2) related species showing broad geographical disjunctions.

> *"This fact leads us to the idea ... that stations are determined uniquely by physical causes actually in operation, and that habitations are probably determined by geological causes which no longer exist today. According to this hypothesis one may easily conceive why plant species that are never found native in a certain area will nevertheless live there if they are introduced. But this theory is touched by the uncertainty, one must admit, of all of the ideas relating to the ancient state of our globe and to the primitive origin of living things"*
>
> De Candolle, 1820, pages 415-416

Floristic regions are supposed to be characterised by particular regional floras, and differ significantly from other such regions in their species composition. Takhtajan (1986) describes it explicitly as an 'area of endemism' (see also Nelson, 1978), defined statistically by numbers of taxa only found within that region: a majority of species should be confined to a particular floristic region. For plants, schemes of floristic regions of the world have been proposed by Engler (e.g. Engler & Diels, 1936), by Good (1974) and by Takhtajan (1986). These various schemes differ, but only in detail, not in their aims or methods. However, knowledge of all species' distributions is still imperfect, and the statistical criteria cited by Takhtajan, for example, are not applied uniformly for discriminating all his regions. Nor is it clear how to distinguish appropriately between differently-sized areas, nested inside each other, which may each contain a majority of endemic species; because floristic regions are intended to be mutually exclusive, one floristic region cannot contain another inside of it.

The floristic region is therefore only one level in a hierarchy of similarity analogous to the Linnean taxonomic hierarchy. The aim is to show relationships between areas. Realms are the largest areas, with the most general similarity in floristic composition, progressing then through regions, provinces and districts (Takhtajan, 1986). However, as with taxonomic classes, these entities are hard to define in a non-arbitrary way. This leads to the differences between the various authors of such schemes, as different areas may be circumscribed based on different degrees of similarity. For example, Takhtajan (1986), following White (1983), considered coastal regions of east Africa to be distinct enough to merit recognition as the Zanzibar-Inhambane floristic region; Good (1974) included this area within his much

larger East African Steppe region, but the same species are still found there. An objective test of the concept of an area hierarchy was provided by McLaughlin (1989, 1992) for the western U.S.A.; the idea did seem to model the reality of plant distributions in this study. However, the prospect of such a detailed numerical ordination being carried out at species level on a world-wide basis seems a distant one.

A rigid area hierarchy means that floristic areas are delimited as contiguous and mutually exclusive. In reality, plant distributions are seldom so regular; there is rarely a hard-and-fast boundary between different floristic regions. Although Takhtajan (1986) in his introduction discusses the principles of delimiting floristic areas, and the broad and complex transition regions often to be found between them, these thoughts do not seem to have affected his practice, which has been to create a set of mutually exclusive contiguous areas. Takhtajan's work can be seen as a compromise between simplifying patterns to a clear-cut but rigid conceptual model and expressing the complexity of the real situation with a less structured and therefore less coherent classification. A more flexible approach is that of White (1983, 1993), for example, who defined different types of floristic region in his chorological classification of Africa. These were:

1. Regional centres of endemism (traditional 'floristic regions') e.g. Guineo-Congolian region.
2. Archipelago-like regional centres of endemism, composed of several small areas isolated geographically. e.g. Afromontane region
3. Archipelago-like regions of extreme floristic impoverishment (differing from the above only in terms of species diversity) e.g. Afroalpine region
4. Regional transition zones e.g. Kalahari-Highveld region
5. Regional mosaics e.g. the vegetation around Lake Victoria

However much this approach better reflects the underlying patterns of distribution, it nevertheless does not seem to have been adopted for other areas of the world (c.f. Takhtajan, 1986). Assessing the existing hierarchy of floristic areas and the strength of floristic relationships between different areas of the world is taken further in Chapter 5.

1.6.3 Floristic elements

Under the definition used in this thesis, a floristic element is a component of the flora of a particular region which is defined by reference to another region. It is both a part of the flora of one area but also found within another area, and so demonstrates floristic links between different areas. Thus, the distributions of floristic elements themselves do not necessarily conform to a particular type – floristic elements may have distributions which can be either widespread and contiguous (or at least, more widespread than the limits of that area) or localised within that area but then often disjunct between areas, but a floristic element within a flora implies the opposite of an endemic taxon – it is used here explicitly to

mean something which is also found somewhere else. Also implicit in the concept of a floristic element is that it is made up of several-to-many taxa; unless they should be completely different from any other distribution pattern, individual, idiosyncratic distributions tend to be discounted. Floristic elements are thus repeating distribution patterns, and the larger the floristic element, the more important a component it will be of that flora. However, a given floristic element should always be only a small component of the complete flora of a region. Should a floristic element encompass the majority of the constituent taxa within a given area this often becomes the justification for recognising that area as a separate floristic region (cf. Takhtajan, 1986); this component then ceases to be a floristic element of that area (something also found outside of it) but becomes instead the endemic element confined to it. The identification of floristic elements forms the basis of Chapter 6 of this thesis.

1.6.4 Widespread taxa

The difference between widespread and endemic taxa is not well defined; range-sizes of taxa can range from those found almost throughout the world (e.g. *Pteridium aquilinum* (L.) Kuhn, bracken) through to genera known only from a single tree on a small island (*Ramosmania rodriguesii* Tirveng.). To a large extent this is simply a question of scale, as a taxon widespread at a local scale may well be localised at a broader scale. However, certain taxa are undeniably widespread regardless of scale. The genus *Acacia*, for example, has a very broad distribution through tropical and subtropical regions; genera such as *Senecio* are found in most parts of the world (subcosmopolitan), while *Carex* L. and *Cyperus* L. (both Cyperaceae), and *Plantago* L. (Plantaginaceae) are the most widespread genera of angiosperms, found in all regions used in this thesis except for Antarctica. To De Candolle (1820) widespread taxa, which broke Buffon's Law, were products of wide ecological tolerances and active and efficient dispersal mechanisms; although shared floristic elements between areas indicated their biogeographical similarity, very widespread taxa were found in so many areas as to obscure biogeographical relationships. Modern vicariance biogeographers also view widespread taxa as obstacles to identifying and understanding more generalised and meaningful patterns of taxon distributions (Humphries & Parenti, 1999).

1.6.5 Disjunct taxa

Disjunct taxa best emphasize the debate in biogeography between dispersalist and vicariant explanations. Darwin and Hooker had both travelled extensively, particularly in the southern hemisphere (see Section 1.5). Darwin was struck by Buffon's Law on his *Beagle* voyage:

> "*In the southern hemisphere, if we compare large tracts of land in Australia, South Africa and western South America, between latitudes 25° and 35°, we shall find these parts extremely*

similar in all their conditions, yet it would not be possible to point out three faunas and floras more utterly dissimilar"

Darwin, 1859, page 347

Hooker, however, on the *Erebus* and *Terror* voyage around the Antarctic, was more impressed by the floristic similarities between widely separate land masses and made a study of disjunct taxa from which he argued that the widely disjunct areas visited by the Antarctic voyage must have, in the past, formed a single area – a land mass occupying a continent larger than that of the Antarctic Ocean. Whereas the dispersalist tradition of Darwin and Wallace focused on the history and dispersal capability of specific organisms in isolation from each other, Hooker's approach explicitly considered similar distributions of many unrelated taxa in conjunction, set within a vicariant framework of changing continental geography. In this way, he was drawing on an older, Continental tradition of biogeographic thought that itself goes back to Humboldt and De Candolle (Humphries & Parenti, 1999). However, as Hooker himself realised, either the dispersal theory or the land-bridge theory could adequately explain the situation; the choice of one over the other was more a matter of temperament than of evidence (Grehan, 1990).

Disjunct distributions, especially those of the southern hemisphere, for example *Nothofagus* Blume (Humphries, 1981; Linder and Crisp, 1995) and the Proteaceae (Weston & Crisp, 1987), have long been a favourite subject of study by biogeographers (Humphries & Parenti, 1999). The advent of plate tectonic theory was eagerly seized upon by biogeographers as a mechanism to explain previously-puzzling disjunct distributions (e.g. Raven & Axelrod, 1974; Humphries, 1981; Linder & Crisp, 1995), and as a counter-argument to 'old-fashioned' dispersalist explanations (c.f. Darlington, 1965; Brundin 1966). However, there is more to disjunct distributions than 'Gondwanic' taxa; apart from eastern Asia/eastern North American disjunctions (e.g. Wen, 1999; Donoghue *et al*, 2001), most other patterns have not been studied on detail. Thorne (1972), however, in a rare comprehensive study of angiosperm distributions, offered a classification of disjunct distribution patterns:

I. Eurasian-North American
   1. Arctic
      a. Circum-Arctic
      b. Beringian-Arctic
      c. Amphi-Atlantic-Arctic
   2. Boreal
      a. Circum-Boreal
      b. Beringian-Boreal
      c. Amphi-Atlantic-Boreal

3. Temperate
    a. Circum-North Temperate
    b. North and South Temperate
    c. Fragmentary North Temperate

II. Amphi-Pacific Tropical

III. Pantropical

IV. African-Eurasian(-Pacific)
1. African-Mediterranean
2. African-Eurasian
3. African-Eurasian-Malesian
4. African-Eurasian-Pacific
5. African-Eurasian-Australasian
6. Indian Ocean-Eurasian

V. Amphi-Indian Ocean

VI. Asian-Pacific
1. Asian-Papuan
2. Asian-Papuan-Melanesian
3. Asian-Papuan-Pacific Basin
4. Asian-Papuan-Australasian

VII. Pacific Ocean

VIII. Pacific-Indian-Atlantic Oceans

IX. American-African

X. North America-South American

XI. South America-Australasian
1. South American-Australasian
2. South American-Australasian-Asian
3. South American-Australasian-Madagascan

XII. Temperate South America-Asian

XIII. Circum-South Temperate

XIV. Circum-Antarctic

This classification of disjunct distribution patterns, and indeed the usefulness of the three different categories of plant distribution, will be explored again in the context of the global classification of genus distribution patterns for angiosperms undertaken in Chapter 6.

1.7 Discussion

As a discipline, biogeography has a long but diverse history; despite over two centuries of biogeographic research its aims and methods and the fundamental concepts underpinning the discipline are still debated (Nelson, 1978; Humphries, 2000). Even the most appropriate way of characterising and representing plant distribution patterns is still not agreed upon. Probably there is no single best system; different methods are needed to show different aspects of the same phenomenon. Currently within biogeography there are several: traditional delimitation of floristic regions within a rigid hierarchy (e.g. Takhtajan, 1986), or a more pluralistic one (White, 1983); or representing distribution patterns as generalised tracks (Croizat, 1964; Craw, 1989; Craw *et al.*, 1999) or more formally as minimal-spanning trees (Page, 1987); or representing area relationships as area cladograms (Nelson and Platnick, 1981). Within systematics, on the other hand, there is a generally-agreed-upon conceptual model: the Linnean hierarchy of taxonomic ranks. The Linnean hierarchy is an inclusive nested hierarchy, where taxa at lower levels are included within taxa at higher levels but taxa within one level are mutually exclusive; that is, several species can be included within one genus, for example (or genera within a family), but a taxonomically-recognised species cannot be included within another species (nor a genus within another genus). This contrasts with a phylogenetic tree, where clades do not show any taxonomic ranks and do nest one inside the other (Brummitt, 1997). The production of an exclusively-monophyletic classification (e.g. APG, 1998, 2003) therefore principally involves deciding which rank to assign to which clade, but under the strict criterion that ranks within any one clade must always decrease towards the terminal tips of the phylogenetic tree (and all taxa thus contain, in the taxonomic sense, all known descendent taxa).

Constraining an unranked hierarchy to be exclusively monophyletic, however, with ranks only decreasing within clades, can lead to two disadvantages: situations in which either well-supported monophyletic groups must be subsumed into a taxon of higher rank because they nest within a larger clade, and would thus cause it to be paraphyletic if accorded the same rank; or situations in which poorly-supported monophyletic groups must be created from such paraphyletic groups in order to maintain at an equivalent taxonomic rank to these the well-supported, monophyletic taxa nested within them (Albach *et al.*, 2004). The practical problem with this situation is that the same taxon can thus be either accepted taxonomically or not without there being any doubts over its monophyletic status, and so in these cases taxonomic stability is not actually improved. This has led to the Linnean hierarchy recently being challenged as the most appropriate conceptual model for systematics. Phylogenetic systematics (Hennig, 1966) can represent patterns of relationship in a more detailed and explicit way than can the Linnean hierarchy, which must inevitably show less information for those higher taxa which contain more than two subsidiary taxa – there are simply not enough ranks to accommodate each subsequent node of a

phylogenetic tree as a taxon of subsidiary rank (Forey, 1992). Some taxonomists are therefore rejecting the Linnean hierarchical model of nested ranks in favour of purely rank-less (phylogenetic) tree-based methods of nomenclature (de Queiroz and Gaulthier, 1990, 1994; Cantino, 2000), although this is currently a very controversial issue in systematics.

Within biogeography, cladistic biogeography is similarly founded on the assumption that a tree-based method of representing area relationships is a more appropriate conceptual model than the traditional nested hierarchy of floristic areas (Humphries and Parenti, 1999), although this is an idea that has itself been criticised (Cracraft, 1988; Sober, 1988; Hovenkamp, 1997). Areas of endemism, notwithstanding problems over their definition (Harold & Mooi, 1994; Morrone, 1994), are not discrete entities (historical individuals; Ghiselin, 1974; Hull, 1980) as taxa are; the patterns of historical relationships may be too complex to represent in a simple dichotomous diagram. The difficulties of identifying non-overlapping areas of endemism as the units of study for cladistic biogeography also means that incongruence between similar patterns shown at different geographical scales, which may not have a common historical basis, may be largely due to this difference in scale. Without adequate criteria for comparing distribution patterns prior to a cladistic biogeographic analysis, it is unclear when two or more distributions are actually 'the same' and should be studied together using component analysis, for example. There has therefore not been a single overall methodology or rigid philosophical viewpoint, such as vicariance biogeography, followed in this thesis; rather, different approaches and methods have been used as appropriate for each of the aspects of plant biogeography studied. In turn, this has hopefully given a more balanced picture of the patterns of plant diversity and distribution. However, the general view of cladistic biogeography that a rigid hierarchy of areas may not be the most appropriate way of representing biogeographic patterns (in this case historical relationships between areas) is one that has been taken up in this thesis for studying patterns of floristic relationship and patterns of distribution (see Chapters 5 and 6).

As not all genera of flowering plants have yet been analysed phylogenetically, and the degree of sampling for large genera is still small, it is possible and even likely that not all the genera contained in the Vascular Plant Families and Genera database will prove to be monophyletic. Therefore, in analysing patterns of diversity and distribution at each taxonomic scale (i.e. at family, genus and species levels), it is possible that not all units of the analysis are exactly equivalent, and thus like is not being compared with like. However, since there are no objective criteria for ranking clades (APG, 1998; Scotland & Sanderson, 2004), and suggested methodologies for comparing clades (e.g. Barraclough *et al.*, 1999) are still sensitive to taxon size and sampling effort, it is not clear how to ensure that comparisons of different clades do actually compare like with like. That is, is a clade assigned the rank of genus really equivalent to another clade given the rank of genus, or is it perhaps equivalent to another clade assigned a different rank? Furthermore, despite the sizes and distributions of many individual taxa changing under an exclusively

monophyletic classification, changes in one direction are counteracted by changes in the opposite direction, and the overall frequencies of taxon size and range size are insensitive to this (i.e. many small families are sunk, but other small families are newly-recognised in turn) (see Chapter 3). The results of the analyses presented in this thesis, therefore, will hopefully remain robust to future changes in the status and the distribution of individual families and genera.

1.8 Aims and objectives of this thesis

The data on which this thesis is based – global distributions at regional level for all angiosperm genera – is currently the most comprehensive (though not the most detailed by far) such data yet compiled. The overall aim of this thesis is to use this comprehensive data set to analyse biogeographic and ecological ideas in a global context which were originally proposed from analyses of much smaller data sets or at finer geographical or taxonomic scales. Do they still hold at broader scales or with increased sampling of taxa? The different chapters of this thesis analyse different aspects of plant distribution patterns, each using a separate method appropriate to that question, and so the different chapters stand on their own to some extent. Each represents a separate piece of a larger puzzle, and the picture will hopefully be clear with the last piece (the last chapter of the thesis) in place. Previous analyses of plant distribution patterns have tended to each focus only on particular taxa or on particular types of distribution. There is nothing intrinsically wrong with this – it is often the only possible approach – but this narrow focus does not represent the whole picture of plant distributions and has therefore not been followed here.

It is hoped that by analysing a more comprehensive global data set, the patterns observed and the conclusions drawn from the analyses will thus prove to be more universal and more robust, and the biogeographic picture more rounded, than with more local analyses. The primary unit of study for this thesis is therefore the genus; genera are taken as representing real evolutionary units, although it is possible that not all genera are strictly equivalent to each other (see above). Patterns in the diversity and distributions of families are also investigated (see Chapters 3, 4, 5 and 6), although as the units are more detailed, more attention is given to patterns at genus level. Also, patterns in the diversity of species have been estimated (see Chapter 2). Although it was not possible to study comprehensively patterns in the distribution of species, patterns in the diversity and distribution of angiosperm genera are nevertheless assumed to be similar to patterns in the diversity and distribution of angiosperm species (see Chapters 3 and 7). Collectively, distributions of these genera are very heterogeneous (see Chapters 3 and 6), from endemic to widespread to widely disjunct taxa. However, just how best to characterise distribution patterns is still not agreed upon, and thus no single analysis of global plant distribution patterns is really equivalent to the one presented here.

The specific objectives of this thesis are therefore as follows:

- To establish what are the broad patterns in the diversity of families, genera and species in different regions around the world (Chapter 3).
- To investigate how the size of a region, its latitudinal position and the distances between regions influence the number of taxa found within that region (Chapters 3 and 4).
- To investigate the roles of available area and continental shape in determining the range sizes of taxa and the richness of tropical regions (Chapter 4).
- To describe the floristic relationships between different regions, and to compare these with existing global schemes of floristic classification (Chapter 5).
- To study patterns of plant distribution around the world and produce global classifications of family and genus distribution patterns (Chapter 6).
- To study the relationship between the number of genera within a region and the number of distribution patterns within a region (Chapter 6).
- To investigate how these various factors together interact to form the patterns of plant diversity around the world (Chapter 7).

CHAPTER 2

MATERIALS & METHODS

This chapter presents the details of the data used in this study and reviews the analytical methods chosen. Although the database upon which the majority of this research is based was discussed in the previous chapter, additional data sources are discussed here. Methods for particular analyses are outlined in each chapter as relevant, but here a more discursive review of methods of multivariate statistics is presented which would otherwise appear interjected into the flow of the thesis.

2.1 Taxonomic surrogates for biodiversity

2.1.1 What *is* biodiversity?

The precise meaning of the much-bandied term 'biodiversity' is as confused and confusing as the question of exactly how this biodiversity should be measured. Most commonly, 'biodiversity' is used as a synonym for 'the variety of life' (Gaston, 1996a), though the concept may be extended to include process-based as well as pattern-based approaches (Noss, 1990). Of many similar definitions, perhaps the simplest and most comprehensive is the definition given by Bibby *et al.* (1992):

> "*Biodiversity is the total variety of life on earth. It includes all genes, species and ecosystems and the ecological processes of which they are a part.*"

Though the exact words may differ, three conventionally recognised dimensions of biodiversity are: genetic diversity, the underlying diversity upon which natural selection operates; species or taxonomic diversity, the products of that natural diversity; and ecosystem diversity, the context in which natural selection occurs (Gaston, 1996a). Implicit in such a broad definition is the impossibility of finding a single measurement to encapsulate all of biodiversity: the complexity of the system cannot just be reduced to a simple statistic. The different aspects of biodiversity have been studied using different techniques, and choice of a particular means of measuring biodiversity will obviously depend on the questions being asked. That there is a preponderance of studies dealing with species richness only at the local scale perhaps reflects the impossibility of large-scale, comprehensive sampling at the species level, doubts as to how ecosystem diversity can best be measured, and what such measures really represent (Gaston, 1996a).

Species richness remains the optimum measure of biodiversity (Gaston, 1996a). However, the possibility of attaining detailed, comprehensive information on global patterns of species richness still seems remote. Given this impasse, much attention has been given to the question of whether there are indirect methods for measuring species richness. Three broad approaches have been developed (Gaston & Williams, 1993), based on the correlation between total species richness and: i) the species richness of one or more 'indicator' taxa (e.g. Mittermeier, 1998; Bibby *et al.*, 1992); ii) values of one or more environmental or habitat variables (e.g. Miller *et al.*, 1987; Braithwaite *et al.*, 1989); and iii) the richness of higher taxonomic groups to which the taxa of interest belong (e.g. Gaston & Williams, 1993; Williams & Gaston, 1994). In the context of this thesis, the study of patterns in the diversity and distribution of families and genera is of greater interest if higher taxa can be shown to accurately reflect patterns of diversity at species level; patterns in the distribution of families and genera may therefore be indicative of those at species level.

2.1.2 Higher taxa as biodiversity surrogates

The use of taxonomic surrogates for studying species diversity patterns is appealing because the numbers of higher taxa are obviously far fewer, their identification is often easier, yet they give an approximation of wholesale biodiversity which neither indicator taxa nor environmental variables can do (Williams & Gaston, 1994). There is inevitably a trade-off between the exactitude of analysing complete species-level data and the savings of time and resources with the higher-taxon approach. The 'higher' the taxonomic surrogate (in the sense of the Linnean hierarchy), the more detail is lost and the less reliable the results will be (Williams & Humphries, 1996; La Ferla *et al.*, 2002). The use of higher taxa as surrogates for species diversity has been widely used in palaeontology, where the probability of a particular family, representing many species, being fossilised is far greater than that of a single species. Despite concern over the inherent assumption that the number of species per family has been constant over time, patterns of higher-taxon richness are thought to be a good approximation of patterns of species richness (Sepkoski, 1992).

Williams & Gaston (1994) found a positive correlation ($p < 0.001$) between family richness and species richness for each of: ferns and fern-allies in Britain and Ireland; butterflies, also in Britain and Ireland; passerine birds in Australia; and North and Central American bats. La Ferla *et al.* (2002) found that a high proportion of the variation in species richness was explained by each of genus richness ($R^2 = 0.93$), family richness ($R^2 = 0.86$) and ordinal richness ($R^2 = 0.84$) for flowering plants in 1 degree x 1 degree grid squares throughout continental Africa. The uniformity of results in this diversity of taxa, scales and latitudes suggests that this is a widely-applicable phenomenon. An additional advantage of higher-taxon surrogates over indicator taxa and environmental variables is that this method retains information on the identity of taxa within each area and therefore there is some knowledge of the spatial turnover of taxa between areas (β diversity) (Williams *et al.*, 1994), giving

relative estimates of endemism within areas and from this a measure of complementarity (Vane-Wright *et al.*, 1991) important for assigning conservation priorities.

Potential problems with the use of higher-taxon surrogates for species richness come principally from the inherent difficulty of establishing non-arbitrary higher taxa, and from genuine differences in taxonomic richness relationships between areas. For example, some island groups such as the Hawai'ian Islands contain many higher taxa that have undergone evolutionary radiations in isolation, resulting in unusually low generic diversity with high species diversity (Wagner & Funk, 1995). Conversely, St Helena is an island with a very low species : genus ratio; most genera are represented by only one species (Cronk, 2000). A situation comparable to that of the Hawai'ian Islands, with large radiations in few genera, can be found in continental areas, for example the Cape region of South Africa (e.g. Richardson *et al.*, 2001b). A difference in the richness relationship is also likely between the terrestrial and marine realms for much higher taxonomic ranks, since the marine fauna contains many more phyla of animals, yet fewer species than does the terrestrial realm (May, 1988). As yet these variations in taxon-richness relationships between areas remain understudied. For plant taxa, however, the effectiveness of families and genera as higher-level surrogates of species across the globe is assessed in Chapter 3.

2.2 Estimating regional species richness

2.2.1 Extrapolating from the World Checklists and Bibliographies database

In addition to the Vascular Plant Families and Genera database used throughout this thesis, RBG Kew compiles and maintains a number of other taxonomic databases. Among these is the mammoth World Checklists and Bibliographies series, which details complete taxonomic checklists at the species and sub-specific level, with places of publication, distributions and habit notes for each species, and accepted species names for all synonyms. Working on a family-by-family basis, principally from existing publications, this series has to date published the families Betulaceae, Corylaceae, Euphorbiaceae, Fagaceae, Magnoliaceae, Sapotaceae and all conifers, and the further families Araceae, Gramineae, Myrtaceae, Orchidaceae, Palmae and Rubiaceae have already been completed and are undergoing editing prior to publication. Many more family-level checklists are also in preparation. The total number of species names so far entered into this database is about half of the total number of species names listed in Index Kewensis (an index to the place of publication of all validly-published generic and species names for spermatophytes – that is, flowering plants and gymnosperms), and the number of accepted genera listed is also about half the number of genera in the Vascular Plant Families and Genera database, so this database is estimated to represent a combined total of some 50% of spermatophyte species (R. Govaerts, pers. comm.).

It should be possible to use data from this independently-derived dataset to test some general patterns of genus distribution from the Vascular Plant Families and Genera database, and compare genus-level patterns of diversity with those at the species level (see Chapter 3). The World Checklists and Bibliographies database was therefore queried for the number of species found in each TDWG Level 2 region. Working on the basis that the total in this database was 50% of the total number of accepted species names in Index Kewensis, these totals were simply multiplied by 2 to give total species richness estimates for each Level 2 TDWG region (cf. Govaerts, 2001, 2003); these totals are listed in Table 2.1, along with numbers of families and genera from the Vascular Plant Families and Genera database. The accuracy of these extrapolated figures rests on two assumptions: that rates of synonymy are equivalent between the families contained in the World Checklists and Bibliographies database and those families not yet tackled; and that there has been no geographical bias in the selection of families already treated, i.e. that the database contains an equivalent proportion of predominantly-tropical or predominantly-temperate families as really exist. Until such time as there is a complete species checklist for the world, however, it is difficult to know how realistic these assumptions are.

2.2.2 Published literature estimates

In an attempt to further verify the accuracy of the estimated species richness totals, published estimates of total species numbers were sought for as many comparable areas as possible. Insofar as it was possible to gather figures considered reliable, these published estimates, along with cited references and the percentage difference between these totals and the totals given in Table 2.1, are given in Table 2.2. Gathering such figures proved to be a very difficult task. In many cases, estimates for exactly equivalent areas could not be found, and totals had be estimated from those given for comparable areas (e.g. Region 70; Subarctic America). In some cases, published estimates were either for just flowering plants or for all vascular plants, rather than specifically for spermatophytes (e.g. Region 43, Papuasia). In other cases, the most recent and reliable published estimate for an area was itself still very unreliable, a mere 'guesstimate' (e.g. that for Region 42, Malesia), and this may or may not be the total number of known native species for that area or the total predicted number of native species, including the number not discovered as yet (e.g. that for Region 84, Brazil).

Not having a reliable-enough bench-mark, however, it is difficult to say which of the estimated and published totals is over-estimated or under-estimated relative to the other. In this respect, it is notable that amongst the published estimates, several stand out as being from recently-published complete floras or checklists of areas exactly equivalent to TDWG Level 2 Regions (those of Regions 21, Macaronesia; 22, West Tropical Africa; 27, Southern Africa; 28 Middle Atlantic Ocean; 32, Central Asia; 36, China; 37, Mongolia; 50, Australia; 62, Northwestern Pacific; 63, North-Central Pacific). It is these species totals where the greatest agreement should be expected. However,

inspecting Table 2.2, these appear no more nor less likely to agree with species totals estimated from the World Checklists and Bibliographies database than do any others. It is difficult to assert with any degree of confidence the significance of such differences in patterns of species richness: either set of figures may be equally wrong. Notwithstanding the sometimes large discrepancies between database estimates and published estimates of species numbers, however, for those regions where there are both values there is a high overall correlation between them (Spearman's r_s: 0.96; $n = 38$, $p < 0.01$).

Despite this, the broader geographical variation in species richness is robust to these differences within individual regions: patterns of generic and familial richness are well correlated with both the database estimates used here and also the available published literature estimates of plant species richness across the Earth (Spearman's r_s for literature estimates of species numbers: between species richness and genus richness, 0.94; between species richness and family richness, 0.88; $n = 38$, $p < 0.01$ in each case), although correlation coefficients are slightly lower for published literature estimates than for database estimates (between species richness and genus richness, 0.94 *vs.* 0.97, respectively; between species richness and family richness, 0.88 *vs.* 0.93, respectively).

2.2.3 Three taxonomic levels of diversity

Given the comprehensive geographical coverage of the World Checklists and Bibliographies database, the one-to-one correspondence to the TDWG Level 2 Regions used in the Distributions of Vascular Plant Families and Genera database, and the overall correlation between these database estimates and previously-published literature estimates, these extrapolated species counts for each region were used in this study. Numbers of families and genera of flowering plants for each region are taken from the Distributions of Vascular Plant Families and Genera database as discussed in Chapter 1. Family delimitation generally follows APG I (1998), which has been implemented as an optional family-level classification within the Distributions of Vascular Plant Families and Genera database; however, the research which forms this thesis had progressed too far to accommodate the subsequent changes presented in APG II when that appeared (APG, 2003). This study therefore utilises known numbers of families and genera, and estimates of numbers of species, for each TDWG Level 2 Region around the world (see Table 2.1). These are all used to estimate the general patterns of diversity shown in the following chapter. In the absence of comprehensive species-level distribution data, which is not yet available, subsequent chapters then explore in greater detail how the underlying distribution patterns of families and (chiefly) genera contribute to the global patterns of higher taxon diversity, and the final chapter then argues that, if there is a strong correlation between diversity at different taxonomic levels, insights from the distributions of higher taxa may also provide insights into how species diversity patterns may be structured by the underlying patterns of species distribution.

Table 2.1 Numbers of families, genera and species for TDWG Level 2 Regions.

TDWG Region		Number of families	Number of genera	Estimate of species number
10	Northern Europe	115	577	1678
11	Middle Europe	123	709	3358
12	Southwestern Europe	142	1031	6514
13	Southeastern Europe	148	1093	7778
14	East Europe	124	803	3546
20	Northern Africa	135	971	5102
21	Macaronesia	111	491	1694
22	West Tropical Africa	203	1574	7808
23	West-Central Tropical Africa	214	1856	15898
24	Northeast Tropical Africa	199	1578	7950
25	East Tropical Africa	218	1861	12614
26	South Tropical Africa	214	1806	12422
27	Southern Africa	209	1864	19706
28	Middle Atlantic Ocean	21	44	60
29	Western Indian Ocean	205	1398	12868
30	Siberia	106	606	3170
31	Russian Far East	122	573	2504
32	Central Asia	116	917	9726
33	Caucasus	134	873	5388
34	Western Asia	160	1476	20662
35	Arabian Peninsula	137	885	2386
36	China	264	2420	27664
37	Mongolia	96	478	1758
38	Eastern Asia	214	1334	7482
40	Indian Subcontinent	259	2572	18786
41	Indo-China	249	2085	19724
42	Malesia	243	2162	35628
43	Papuasia	223	1535	17202
50	Australia	227	1874	22122
51	New Zealand	111	351	2084
60	Southwestern Pacific	173	909	6392
61	South-Central Pacific	91	264	816
62	Northwestern Pacific	103	339	1086
63	North-Central Pacific	86	242	1084
70	Subarctic America	80	335	1142
71	Western Canada	121	548	1682
72	Eastern Canada	137	519	1652
73	Northwestern U.S.A.	126	707	3600
74	North-Central U.S.A.	151	694	2672
75	Northeastern U.S.A.	150	637	2284
76	Southwestern U.S.A.	159	1067	6862
77	South-Central U.S.A.	174	1020	4524
78	Southeastern U.S.A.	195	1018	4164
79	Mexico	234	2212	22308
80	Central America	226	1933	17800
81	Caribbean	198	1538	12954
82	Northern South America	226	2053	18308
83	Western South America	237	2701	46012
84	Brazil	222	2489	37662
85	Southern South America	216	1706	12400
90	Subantarctic Islands	46	91	228
91	Antarctic Continent	2	2	2

Table 2.2 Comparison of database *versus* published literature estimates of species numbers for TDWG regions.

TDWG Region		Database estimate	Literature estimate	% difference
10	Northern Europe	1678		
11	Middle Europe	3358		
12	Southwestern Europe	6513		
13	Southeastern Europe	7778		
14	East Europe	3545	4100-4300[1]	-18.48
20	Northern Africa	5102	10000[2]	-96.00
21	Macaronesia	1694	3106[3]-3200[4]	-86.13
22	West Tropical Africa	7808	7343[5]	+5.96
23	West-Central Tropical Africa	15898		
24	Northeast Tropical Africa	7950	7000+[6]	+11.95
25	East Tropical Africa	12614	12290[7]	+2.57
26	South Tropical Africa	12422	10000[8]	+19.50
27	Southern Africa	19706	21087[9]	-7.01
28	Middle Atlantic Ocean	61	61[10]	0.00
29	Western Indian Ocean	12868	12000+[6]	+6.75
30	Siberia	3170	4200[11]	-32.49
31	Russian Far East	2504	4000[12]	-59.74
32	Central Asia	9726	8096[13]	+16.76
33	Caucasus	5388	6200+[6]	-15.07
34	Western Asia	20662	c.22000[14]	-6.48
35	Arabian Peninsula	2386	3060[15]	-28.25
36	China	27664	26766[16]-27283[17]	+2.31
37	Mongolia	1758	2239[18]	-27.36
38	Eastern Asia	7482		
40	Indian Subcontinent	18786	c.20000[6]	-6.46
41	Indo-China	19724	20000-25000[6]	-14.07
42	Malesia	35628	26000-31,500[19]	+19.31
43	Papuasia	17202	17000[6]	+1.17
50	Australia	22122	15638[20]-18095[21]	+23.75
51	New Zealand	2084	2000[22]	+4.03
60	Southwestern Pacific	6392	4156+[6]	+34.98
61	South-Central Pacific	816	675+[23]	+17.28
62	Northwestern Pacific	1086	1228[24]	-13.08
63	North-Central Pacific	1084	956[25]	+11.81
70	Subarctic America	1142	1280+[26]	-12.08
71	Western Canada	1682		
72	Eastern Canada	1652		
73	Northwestern U.S.A.	3600		
74	North-Central U.S.A.	2672		
75	Northeastern U.S.A.	2284	c.4418[6,27]	-93.43
76	Southwestern U.S.A.	6862		
77	South-Central U.S.A.	4524		
78	Southeastern U.S.A.	4164		
79	Mexico	22308	21600[28]	+3.17
80	Central America	17800	17000[29]	+4.49
81	Caribbean	12954	13000[30]	-0.36
82	Northern South America	18308	c.22000[31]	-20.17
83	Western South America	46012	c.65000[31]	-41.27
84	Brazil	37662	c.55000[32]	-46.04
85	Southern South America	12400	c.14000[31]	-12.90
90	Subantarctic Islands	228		
91	Antarctic Continent	3	2[33]	+33.33

Sources for literature estimates of regional species diversity given in Table 2.2; literature estimates could not always be found for exactly comparable regions. Where two literature estimates are provided for a single region, percentage difference from the database estimate was calculated from the mean value.

1. Webb, D.A. (1978). *Flora Europaea* – a retrospect. *Taxon* **27**: 3-14.
2. Quézel, P. (1985). Definition of the Mediterranean region and the origin of its flora. In Gomez-Campo (ed.). *Plant Conservation in the Mediterranean Area.* Dr. W. Junk Publishers, Dordrecht, Netherlands.
3. Hansen, A. & Sunding, P. (1993). Flora of Macaronesia: checklist of vascular plants, 4th edition. *Sommerfeltia* **17**.
4. Humphries, C.J. (1979). Endemism and evolution in Macaronesia. Pp. 171-199 in Bramwell, D. (ed.) *Plants and Islands.* Academic Press, London.
5. Hepper, F.N. (1989). West Africa. Pp. 189-197 in Campbell, D.G. & Hammond, H.D. (eds.) *Floristic Inventory of Tropical Countries.* New York Botanical Garden, New York.
6. Frodin, D.G. (2001). *Guide to Standard Floras of the World*, 2nd edition. Cambridge University Press, Cambridge, U.K.
7. H.J. Beentje, pers. comm. 2001.
8. G. Pope, pers. comm. 2001.
9. Arnold, T.H. & de Wet, B.C. (1993). *Plants of southern Africa: names and distribution.* Memoirs of the Botanical Survey of South Africa, no. 62. National Botanical Institute, Pretoria, South Africa.
10. Cronk, Q.C.B. (2000). *The Endemic Flora of St Helena.* Anthony Nelson, Oswestry, England.
11. Krasnoborov, I.M. *et al.* (1987-1999). *Flora Sibiri*, Vols 1-14. 'Nauka', Novosibirsk.
12. Charkevicz, S.S. (ed.) *Plantae Vasculares Orientis Extremi Sovietici.* Vol 1- . 'Nauka', St. Petersburg.
13. Vvedensky, A.I. *et al.* (1968-1993). *Conspectus Florae Asiae Mediae*, Vols 1-10. 'FAN', Tashkent.
14. Boulos, L., Miller, A.G. & Mill, R.R. South West Asia and the Middle East: regional overview. In Davis, S.D. *et al.* (1994) (eds.) *Centres of Plant Diversity: a guide and strategy for their conservation, volume 1.* WWF & IUCN Publication Unit, Cambridge, U.K.
15. A.G. Miller, pers. comm. 2001.
16. Keng, Hsuan, Hong Der-Yuan & Chen Chia-Jui. (1993). *Orders and Families of Seed Plants of China.* World Scientific, Singapore.
17. Editorial Board of Flora of Zhejiang (1989-1993). *Flora of Zhejiang.* Zhejiang Science & Technology Press.
18. Grubov, V.I. (2000). *Key to the Vascular Plants of Mongolia (with an atlas), Vols. 1 and 2.* Science Publishers, Inc. Enfield, New Hampshire, U.S.A. & Plymouth, U.K.
19. Davis, S.D. *et al.* (1995) (eds.) *Centres of Plant Diversity: a guide and strategy for their conservation, volume 2, Asia, Australasia and the Pacific.* WWF & IUCN Publication Unit, Cambridge, U.K.
20. Hnatiuk, R.J. (1990). *Census of Australian Vascular Plants.* Australian Fauna and Flora Series, no. 11. Bureau of Flora and Fauna, Canberra, Australia.
21. Crisp, M.D., West, J.G. & Linder, H.P. (1999). Biogeography of the terrestrial flora. Pp. 321-367 in Orchard, A.E. (ed.) *Flora of Australia, Volume 1: Introduction*, 2nd edition. ABRS/CSIRO, Canberra, Australia.
22. New Zealand: regional overview. In Davis, S.D. *et al.* (1995) (eds.) *Centres of Plant Diversity: a guide and strategy for their conservation, volume 2, Asia, Australasia and the Pacific.* WWF & IUCN Publication Unit, Cambridge, U.K.
23. Fosberg, F.R., Sachet, M.-H. & Oliver, R.L. (1979). A geographical checklist of Micronesian Dicotyledonae. *Micronesica* **15**: 41-295. *Idem*, (1982). A geographical checklist of Micronesian Pteridophyta and Gymnospermeae. *Micronesica* **18**: 23-82. *Idem*, (1987). A geographical checklist of Micronesian Monocotyledonae. *Micronesica* **20**: 19-129.
24. Florence, J. (1997). *Flore de la Polynésie Française*, Vol. 1- . Éditions ORSTOM, Paris.
25. Wagner, W.L., Herbst, D.R. & Sohmer, S.H. (1999). *Manual of the Flowering Plants of Hawai'i*, revised edition. Bishop Museum Special Publication, no. 97. Bishop Museum Press, Honolulu, Hawai'i.
26. Hultén, E. (1941-1950). *Flora of Alaska & Yukon.* 10 parts. *Acta Universitatis Lundensis* **37-46**.
27. Fernald, M.L. (1950). *Gray's Manual of Botany*, 8th edition. American Book Co., New York.
28. Rzedowski, J. (1988). Diversidad y orígenes de la flora fanerogámica de México. *Simposio Diversidad Biológica de México.* Universidad Nacional Autónoma de México, Oaxtepec, México.
29. Middle America: regional overview. In Davis, S.D. *et al.* (1997) (eds.) *Centres of Plant Diversity: a guide and strategy for their conservation, volume 3, the Americas.* WWF & IUCN Publication Unit, Cambridge, U.K.
30. Caribbean Islands: regional overview. In Davis, S.D. *et al.* (1997) (eds.) *Centres of Plant Diversity: a guide and strategy for their conservation, volume 3, the Americas.* WWF & IUCN Publication Unit, Cambridge, U.K.
31. Introduction to Davis, S.D. *et al.* (1997) (eds.) *Centres of Plant Diversity: a guide and strategy for their conservation, volume 3, the Americas.* WWF & IUCN Publication Unit, Cambridge, U.K.
32. Prance, G.T. (1979). History of exploration: South America. Pp. 55-70 in Hedberg, I. (ed.) *Systematic Botany, Plant Utilization & Biosphere Conservation.* Almqvist & Wiksell International, Stockholm, Sweden.
33. Skottsberg, C.J.F. (1954). Antarctic flowering plants. *Botanisk Tidskrift* **51**: 330-338.

2.3 General analytical methods

Each analysis used in this thesis is described in a separate methodology section within the relevant chapter. However, it is useful to present here some overview of the many analytical techniques used. The full range of all analytical methods used in every chapter is not presented here, as most of the analyses in both Chapter 2 and Chapter 3 use standard univariate statistics such as correlation and regression analysis as outlined within each of these chapters. What is useful to emphasize here, however, are the differences in the types of data used in different chapters. Although all the data, apart from counts of species richness for each TDWG Level 2 Region (see Section 2.2), is drawn from the Distributions of Vascular Plants Families and Genera database as described in Chapter 1, this is extracted, manipulated and analysed in different ways within each chapter. Chapter 3 which follows uses counts of taxa per region, assessing diversity on a region-by-region basis and studying the factors of area and distance between regions by their effect on these counts of taxa. Chapter 4 studies a single gradient, latitude, and so uses only latitudinal ranges of all taxa across the world in these analyses, and not distributions of taxa by particular regions. Chapter 4 also uses a Monte Carlo simulation technique to predict numbers of taxa at different latitudes, although these results are then assessed against empirical taxon richness with standard regression statistics and a 2-dimensional Kolmogorov-Smirnov test. However, Chapters 5 and 6 use the complete matrix of distributions, for both families and genera, analysing relationships between regions and between taxa, respectively, with multivariate statistics. Due to the reliance on multivariate statistics in this thesis, and the complexity and range of many different available methods, these are discussed more fully below, along with a short statement as to why a particular technique was chosen (or not chosen) for a particular stage in the analysis.

2.4 Review of multivariate techniques

Within the taxon by area data matrix, the same (or different) techniques may group the areas into floristic regions by the similarity of their taxon-composition (see Chapter 5); alternatively, taxa may be grouped by similarities of their distribution into common floristic elements (see Chapter 6). The 52 TDWG Level 2 regions in the data matrix used in this thesis represent the objects, while the taxa found in those regions represent the attributes of those objects. Analysing the relationships amongst objects, or amongst the attributes of those objects, are known respectively as Q-mode and R-mode analysis (Kent & Coker, 1992; McCune & Grace, 2002), respectively. Chapters 5 and 6 therefore undertake Q-mode and R-mode analysis of the same data matrix, respectively. As a Q-mode analysis, therefore, what is of interest in Chapter 5 is the common floristic relationships between different geographical regions. In the R-mode analysis presented in Chapter 6 what is of interest is the presence of taxa in common floristic elements, i.e.

shared distribution patterns. Notwithstanding the interest of floristic classification, for the purposes of this thesis what are of more interest are the common floristic elements which actually generate the floristic relationships between regions. Since both Chapter 5 and Chapter 6 involve large-scale multivariate analyses, a brief overview and discussion of several multivariate techniques used is given below, together with a short statement justifying the selection of that technique. Two further techniques, Two-Way INdicator SPecies ANalysis (TWINSPAN) and Detrended Correspondence Analysis (DCA), are discussed; each was investigated for this analysis but discarded in favour of superior methods. The discussion below indicates the reasons for not pursuing either of these two techniques, despite them still being widely used by ecologists.

2.5 Data transformation – Beals' smoothing

Beals' smoothing (Beals 1984) is a data transformation which reduces the heterogeneity of data in data sets with a large proportion of zeros (McCune, 1994; McCune & Grace, 2002); effectively it transforms qualitative (presence-absence) data into quantitative proportion data by evaluating the 'favourability' of a particular site for a particular taxon. This is done through calculating, for a given taxon in a region, the proportion of distributions which are shared with each other taxon also found in that region but outside of the region in question, i.e. if a taxon within a region has many shared distribution records outside of that region with the other taxa which are found in that region, then that region is very 'favourable' for that taxon. So, a given genus would have a high favourability score for a given region if in that region were also found many other genera with similar distributions. However, should a genus have a distribution which is very dissimilar to the other genera in that region, then that region would be regarded as being very 'unfavourable' for that genus. Favourability scores for taxa within regions are recorded on a scale from 0 (completely unfavourable) to 1 (extremely favourable). Given that this transformation was designed to enhance analysis of data sets with a high proportion of zeros, and that most genera, and many families, show small range-sizes (modal value for generic range-size = 1 region while mean range-size for genera = 5 regions – i.e. the data does have a large proportion of zeros), it was decided to apply Beals' smoothing before the ordination analyses undertaken here. In order to evaluate the importance of Beals' smoothing to the analysis, however, each ordination was run with the data both transformed and untransformed.

2.6 Distance measures

With most multivariate techniques, the first stage of the analysis is to calculate a matrix of similarity (or dissimilarity, distance) between the objects in the original data matrix whose relationships are the focus of study. There are many different types of distance measure, some of which differ markedly in their mathematical properties compared with other distance measures, and therefore the choice of an appropriate (or inappropriate) distance measure at the outset can have an overwhelming effect on the results of the subsequent analysis. Broadly speaking, distance measures can be divided into two main classes: Euclidean and city-block (Manhattan). Euclidean distances measure along the shortest possible path between two points, i.e. in a straight line; city block or Manhattan distances, however, are constrained to only pass along two dimensions of a space, one dimension at a time, in a way analogous to trying to take a diagonal path across a city such as Manhattan through a grid of rectangular blocks – one can only travel along streets which are at right-angles to one another. Since the number of turns taken by a particular path through city block space will not alter the length of the journey, many separate paths can consequently all be of equal distance. Not all distance measures fit into either the Euclidean or city block types, however: other examples are correlation distance, defined as the angle between two points relative to the centroid of the whole group of all points, and Mahalanobis distance, measured between the centroids of two groups of points.

Ecological data often does not conform to the ideal of a normal distribution with linear interactions; instead, many ecological phenomena show 'hollow curve' (Willis, 1922) or 'dust bunny' (McCune & Grace, 2002) distributions, with the modal value found at one extreme of the data range (usually the left), and species' distributions each governed by different environmental factors which interact with each other in a non-linear fashion. City-block distances, which measure along the edges of species space, generally perform better with ecological data than do Euclidean distances because the actual distribution of that data more closely follows a city block than a Euclidean path. The shortest distance between two points in species space, as measured by Euclidean distance, may well be a portion of species space devoid of species, since in reality species are often confined to the edges of species space defined by particular environmental variables (McCune & Grace 2002).

Much of classical statistics is based on the assumption of Euclidean distance measures, which is therefore incompatible with the use of proportional city block coefficients. This means many standard multivariate techniques such as Principal Components Analysis are usually inappropriate for ecological data. The utility of city block coefficients with ecological data has therefore led to a multitude of multivariate techniques developed specifically for ecology. However, even established ecological techniques may be prone to distortion. For example, Minchin (1987a) argued that chi-squared distance, a

Euclidean measure, exaggerated the distinctiveness of samples containing rare species, since it accords high weight to species with low abundance and low weight to species with high abundance. Despite being found to perform poorly, however (Faith *et al.*, 1987; Minchin, 1987a), chi-squared remains the underlying distance measure in several widely-used multivariate methods such as Detrended Correspondence Analysis (DCA, DECORANA) and Canonical Correspondence Analysis (CCA, CANOCO). Although city block distances are foreign to much of classical statistics, mathematical justification for their use was provided by Roberts (1986), who derived proportion coefficients from fuzzy set theory.

Sensitivity of distance measures to increasing data heterogeneity has been investigated by Beals (1984) and Faith *et al.* (1987), using synthetic data with a known underlying structure. Results showed that all distance measures declined in sensitivity to some degree as the distance along the known environmental gradient increased; however, proportional city block coefficients such as Sørensen's or Jaccard's, where distances are expressed as a proportion of the maximum possible distance, were more robust than were Euclidean distance measures. With more heterogeneous datasets (those with a greater average value of distance coefficient), the distortion becomes greater for all distance measures, but again proportional city-block distances perform better than do Euclidean distances such as Euclidean, squared Euclidean or chi-squared distance (McCune & Grace, 2002), while correlation distance consistently performed more poorly than did either Euclidean or proportional city block distances.

All the analyses presented in both Chapters 5 and 6 are based on proportional city block distance coefficients. Sørensen's (Czekanowski; Bray-Curtis) coefficient is effectively twice the abundance shared between two samples divided by the total abundance summed across all species for those two samples, whereas Jaccard's coefficient is the proportion of combined abundance not shared between two samples (McCune & Grace, 2002). Sørensen's coefficient and the relativised Sørensen (Kulczynski) coefficient, which standardises areas of unequal richness by their number of taxa, have been the distance measure of choice throughout these multivariate analyses, as Sørensen's coefficient is a proportional city block coefficient which is intuitively simple to understand and express, and has been shown by several simulation studies to be among the most effective in preserving ecological distances and most robust to increasing data heterogeneity (Faith *et al.*, 1987; Minchin, 1987a,b; McCune & Grace, 2002).

2.7 Classification methods

Classification methods group data units into higher ranking units based on a measure of similarity (distance). A set of objects each of which contain many attributes can be partitioned into groups either by

beginning with the whole set of objects and successively dividing them into more-tightly defined, more-similar groups (divisive clustering), or beginning with a unique point, either a data object or an objectively-defined point such as the centroid, and successively adding together less-and-less similar points into groups (agglomerative clustering). Given the intuitive understanding of the word 'clustering' as being an agglomerative process – combining objects into clusters – 'clustering' for the purposes of this project refers to agglomerative rather than divisive techniques, notwithstanding the development of such divisive methods (see the discussion of TWINSPAN below). The majority of classification techniques are hierarchical, displaying results in the form of a tree (dendrogram). From this, a ranked hierarchical classification can be produced by assigning arbitrary cut-off points for the different ranks and their constituent elements. However, for situations like the analysis of floristic relationships presented here, where the units of study show multiple floristic relationships each with different areas, non-hierarchical methods have also been developed. Non-hierarchical clustering methods assign units to a fixed number of groups, but do not rank these groups into a formal hierarchical model. This is particularly important where the units of study show a great deal of overlap in their attributes and there is therefore not an obvious dichotomous hierarchy of inter-nested elements, as is the case with genus distribution patterns (see Chapter 6). One non-hierarchical technique, *k*-means partitioning (or clustering), is the basis of the analysis of distribution patterns in Chapter 6, and in order to set it in the context of the other multivariate methods used in this thesis it is discussed below.

2.7.1 Hierarchical Cluster Analysis

The general algorithm for hierarchical clustering was described by Lance and Williams (1967, 1968) and Wishart (1969). It is outlined below:

1. For a data matrix of $n \times p$ elements, a distance (dissimilarity) matrix of $n \times n$ elements in p-dimensional space is calculated, and each value is then squared.
2. The smallest value in the distance matrix, which defines the closest pair of elements, is found.
3. Distances between this pair of elements and each other element are recalculated.
4. The smallest distance value is again sought, and these two elements (one of which may or may not be the original pair of elements) joined.
5. The process is re-iterated, re-calculating distance measures between each element (or cluster) and each other element (or cluster), finding the smallest distance each time, until all elements are joined in clusters and the process is complete.

There are numerous variants on the general hierarchical clustering algorithm, principally in the choice of linkage method. Using the linkage method in the algorithm above, 'single linkage' can result in concatenation of successive single elements in the clustering process, a phenomenon known as 'chaining',

which distort the underlying data structure. Its opposite, 'farthest neighbour' or 'complete linkage', when the least-closest pair of elements is joined at each stage, also distorts the underlying data structure and can impose the appearance of distinct groups even when none exist. Both methods, since distances are calculated relative to only single pairs of elements and without taking into account the structure of the group of elements as a whole, may produce spurious results. More reliable is 'group average' or 'unweighted pair-group method' (UPGMA; Sokal & Michener, 1958), which finds average distances for each pair of elements between two groups (i.e. one from each group), and 'median' (Gower, 1969) or 'centroid clustering'.

Ward's method (Ward, 1963; Orlóci, 1967) minimises the sum of squared distances from each element to the centroid of its group (the error sum of squares) – each stage of clustering yields the smallest increase in the error sum of squares, so minimising the variance in inter-group distances. Ward's method has been shown to perform robustly with ecological data (McCune & Grace, 2002), but is incompatible with proportional distance coefficients such as Sørensen's. This is because the inter-group distances calculated at each iteration of the clustering are calculated assuming Euclidean distance, whereas the initial distance matrix calculated with Sørensen's coefficient assumed city-block distances. However, a combinatorial procedure called 'flexible beta' is said to produce results comparable to Ward's method when the combinatorial coefficient $\beta = -0.25$ (Lance & Williams, 1967). Combinatorial methods calculate successive linkages from the original distance matrix, whereas in non-combinatorial methods the whole dissimilarity matrix is re-calculated anew with each iteration; the choice of initial distance measure with combinatorial methods is thus unimportant. The parameter β is a coefficient of the basic combinatorial equation, the value of which determines the way in which distances from two groups are fused into a set of new distances for this new group (McCune & Grace, 2002). However, the user must specify the value of the parameter β with flexible beta linkage, and values of β strongly affect the degree of chaining in the resulting dendrogram (McCune & Grace, 2002).

Cluster analysis is effective with large, heterogeneous data sets which contain several underlying dimensions, and it could be considered an advantage of hierarchical clustering that, unlike non-hierarchical clustering, the user need not decide on the number of groups prior to the analysis. An initial clustering into groups can be used as a basis for more detailed analyses of the relationships within each group separately. There is no hard-and-fast rule as to how or when to 'prune' the dendrogram into groups, however; it involves a compromise between the homogeneity of the resulting groups and the number of groups obtained. The more groups, the less heterogeneous they become yet the less clustered are the data, defeating the object of data reduction; the larger the clusters, the more heterogeneous they are, and the more difficult it becomes to interpret the clusters in a biologically meaningful way.

A number of objective procedures have been developed, however, the most effective being Indicator Species Analysis – not to be confused with Two-Way INdicator SPecies ANalysis (TWINSPAN) (see below) – which produces 'indicator values' for each attribute for each cluster, and groups can then be defined on the basis of maximising indicator values across the taxa in the group (Dufrêne & Legendre, 1997; McCune & Grace, 2002). However, the principal drawback with hierarchical clustering as a whole remains its sensitivity to the choice of both distance measure and linkage method. Despite cluster analysis first being developed as a method of taxonomic analysis (Sneath & Sokal, 1973), this sensitivity to distance measure and linkage method, and the concomitant effects on the results, has led to its almost-complete abandoning within taxonomy (Scotland, 1992); however, the techniques still prove useful where the aim is to study patterns of similarity rather than patterns of evolutionary descent. Hierarchical cluster analysis was used in this thesis in order to represent floristic relationships as a dichotomous hierarchy that could then be explicitly compared with global schemes of floristic classification (De Candolle, 1820; Engler, 1934; Good, 1964; Takhtajan, 1986), although the units used in this analysis are not assumed to be biogeographically equivalent to floristic regions. Because of the sensitivity of hierarchical cluster analysis to the distance measure and the linkage method, several different variations were undertaken and compared with each other (see Chapter 5).

2.7.2 Two-Way INdicator SPecies ANalysis (TWINSPAN)

The most widely used classification method for analysing floristic data is Two-Way INdicator SPecies ANalysis, or TWINSPAN (Hill, 1979b; Kent & Coker, 1992), a polythetic divisive method of numerical classification. TWINSPAN is based on a prior ordination analysis by Correspondence Analysis, and can be thought of as trying to group data points in multidimensional space by placing divisions through the areas of least-density of those points. Though TWINSPAN was explicitly devised to cope with quantitative abundance data, by converting it into qualitative presence/absence by the derivation of 'pseudospecies', the qualitative data type for the analysis here means that this initial stage will not be necessary. A product of TWINSPAN analysis is a two-way ordered table summarising variation in species composition, with taxa and sample units simultaneously ranked along adjacent sides. This produces a two-way table very similar to the those of the Braun-Blanquet European tradition of phytosociology (Braun-Blanquet, 1965; Mueller-Dombois & Ellenberg, 1974). A dendrogram of area relationships can then be produced from the two-way table.

For each dichotomy in the dendrogram, three successive ordinations are required. A primary ordination is carried out on the first axis ordination from a prior Correspondence Analysis ordination, and the data points grouped either side (positive or negative) of the centroid of their combined values. An indicator value, I, is assigned to each taxon, $_j$, from their distribution in areas either side of the centroid:

$$I_j = \frac{n_j^+}{n_+} - \frac{n_j^-}{n_-}$$

where n_j^+ is the number of areas on the positive side which have the taxon $_j$, and n_j^- is the number of areas on the negative side which have the taxon $_j$. Thus, when a taxon occurs in all and only all the areas on the positive side (a perfect positive indicator), $I_j = 1$; when a taxon occurs in all and only all the areas on the negative side (a perfect negative indicator), $I_j = -1$; a taxon with a cosmopolitan distribution, found in all areas, will have $I_j = 0$. For a refined ordination, each area is then allocated an indicator score by adding +1 for each positive indicator and –1 for each negative indicator. Where there are multiple indicator taxa for an area and none of them is a perfect indicator, each indicator is given a value of either +1 or –1 depending on whether they are a positive or a negative indicator.

An indicator score for each area is calculated by summing the number of positive indicators less the number of negative indicators; these indicator scores are used to construct the classification. TWINSPAN limits the number of indicator taxa per area to five (by default), thus the indicator score for each area may range from +5 to –5. The aim is to minimise the differences between the refined TWINSPAN ordination and the original Correspondence Analysis ordination. This is done through establishing an indicator threshold which will assign an area to either one group or another based on its indicator score; the indicator threshold is established for each division in the classification based on a measurement of closeness-of-fit between the refined ordination and the original ordination. Areas with indicator scores above the threshold go into one branch of the division; those with indicator scores below the threshold go into the other. Indicator thresholds are calculated afresh for each division through an iterative relocation algorithm. To run an analysis to completion would leave each area in a single terminal branch, with similarity between areas shown by the number of nodes from the start of the tree – most dissimilar areas would have few nodes, being quickly separated from the other areas; most similar areas, those with least distinctive floras, would require many iterations, hence many nodes, to separate them out. In practice, most TWINSPAN analyses are limited to four or five divisions.

Although it remains a widely-used technique in ecology and biogeography, McCune & Grace (2002) argue that TWINSPAN has inherited several faults from Correspondence Analysis. When data are structured by several underlying gradients, then TWINSPAN performs poorly; it is an assumption of the method that data can be best ordered by a single dominant gradient (Belbin & McDonald, 1993). When data sets are large, cover large spatial scales and are inherently heterogeneous, therefore, more explicit clustering methods perform better. The underlying Correspondence Analysis ordination is based on an implicit chi-squared distance, whereas proportional city-block coefficients such as Sørensen's, Relative Sørensen's or Jaccard's have been shown to perform better (Faith *et al.*, 1987; Minchin, 1987a,b).

Furthermore, the creation of 'pseudospecies' introduces an extra step necessary only for the production of the two-way indicator species table, but the cut-off levels imposed for creating the pseudospecies will influence the two-way ordered table and therefore the resulting dendrogram: more explicit clustering methods have no need of pseudospecies (McCune & Grace, 2002). Although this last point will not apply to the presence-absence data used here, based on the other criticisms given above TWINSPAN was not considered further as a technique to be used in this thesis.

2.7.3 Non-hierarchical clustering by *k*-means

Although the dendrogram resulting from a hierarchical cluster analysis is intuitively simple and easy to interpret, the imposition of hierarchical structure on the data may not necessarily be justified: the relationships in the data may not be best represented by a hierarchical diagram. With non-hierarchical clustering methods the number of groups is specified *a priori* and items are organised into groups by optimizing a particular statistical attribute of those groups. In *k*-means clustering, each data object is assigned to a group based on its proximity in multi-dimensional space to a number of points, defined by the user, along an artificial dimension following the line of maximum diversity through that space. The general procedure is outlined below (adapted from McCune & Grace, 2002).

1. The initial multi-dimensional space is divided by a single dimension from the origin to a hypothetical point of maximum diversity (a point where all species would be present).
2. This artificial dimension is then divided into *k* equal-length segments (the number is defined by the user).
3. Each data object is assigned to the nearest *k*-segment (with the shortest distance to the midpoint of the segment); objects are now arranged along this single artificial dimension.
4. The centroid of each group of data objects is calculated and each object is assigned to the group with the closest centroid.
5. Centroids for each group are iteratively re-calculated and each data object reassigned to the group with the closest centroid until no further re-arrangement of objects into groups occurs.

As distributions of taxa show great overlap with many other distribution patterns, the relationships in the data are too complex to represent as a simple hierarchical diagram. Non-hierarchical clustering performs better than does hierarchical cluster analysis in such situations with units that have highly overlapping attributes, a situation epitomised by the overlapping distribution patterns of different taxa that are studied in Chapter 6. With respect to the data analysed in this thesis, disjunct distributions in particular were not amenable to hierarchical cluster analysis since they were found to group arbitrarily with taxa from either one side or the other of that distribution. For example, a genus widespread in North America and also in China would group with North American taxa, while a genus widespread in eastern Asia but

also with a locality in North America would group with Asian taxa, rather than these two examples forming an 'eastern Asia / North America disjunct' group as one might intuitively expect. Non-hierarchical clustering by *k*-means was therefore used to assign taxa to a pre-set number of groups based on the similarity of their distribution patterns (see Chapter 6). Since the optimum number of groups being sought is often unknown beforehand, group membership is often determined iteratively (i.e. all units are classified sequentially into partitions, each of which has a different possible number of groups). Group membership for each size partition (ie. 5 groups, 6 groups, 10 groups etc.) can then be assessed with various statistics, which usually calculate the ratio between within-group homogeneity and between-group heterogeneity, with the optimum number of groups having the highest value of that statistic. In this study the Calinski-Harabasz pseudo-*F*-statistic (Calinski & Harabasz, 1974) was used to assess the structure of groups, and the partition with the greatest value of this statistic was chosen as being the optimal number of groups. The Calinski-Harabasz pseudo-*F*-statistic has been found to be the most effective at returning the optimal number of groups in simulation studies (Milligan & Cooper, 1985). This technique proved to be an effective method of classifying objects with overlapping attributes.

2.8 Ordination methods

The term ordination refers to the arrangement of data units along an axis, as in the 'ordinate' of a standard two-dimensional graph. The purpose of this arranging is usually to reduce the complexity of the relationships amongst the data units being analysed by graphically summarising this information along one or a few dominant axes. Ordination methods order the data units in terms of their similarity to each other; whilst this is generally done by taxon composition of area units, producing an area-ordination (Q-mode analysis), the inverse or transposed analysis using the same methods produces a taxon-ordination (R-mode analysis). An important feature of ordination methods is that they do not impose a hierarchical structure on the data; instead they represent the data as a scatter (ordination) diagram. Whereas individual taxa are assumed to have a unique evolutionary history, areas are often known to show biogeographic relationships with more than one other area. Thus, multiple relationships between areas are more easily visualised by ordination techniques than by classification techniques, which generally constrain the results into a tree structure.

The most widely-used ordination methods, such as Principal Components Analysis (PCA), are now acknowledged to be very limited in their suitability for analysing floristic data. PCA assumes that correlations between variables will be linear, which they rarely are (Kent & Coker, 1992), and is constrained to having ordination axes orthogonal to each other, and thus uncorrelated, although axes are often interpreted as representing one of several auto-correlated environmental gradients. If the underlying

distribution of the data is not normal or not linear, as with most biogeographical data, then PCA is not an appropriate technique and will severely distort relationships within the data (Kent & Coker, 1992; McCune & Grace, 2002). This is because the first step in PCA is usually to compute a correlation matrix for the data objects being studied, and in the correlation matrix shared absences between objects are interpreted as evidence of a positive correlation. Given that most presence/absence data sets over large geographical areas are inherently heterogeneous, with most taxa found in only a few areas, then areas with only a few taxa in common will be treated as closely related. For the analysis of qualitative presence/absence data, other ordination techniques are more suitable. Given the multitude of ordination methods, however, it is worth examining some in more detail.

2.8.1 Bray-Curtis (polar) ordination

One of the first ordination techniques to be developed specifically for the analysis of ecological data is Bray-Curtis ordination (Bray & Curtis, 1957). Though it is one of the earliest-developed and simplest ordination techniques, simulation studies have shown that Bray-Curtis ordination remains a reliable and effective technique (Beals, 1984); Roberts (1986) achieved almost identical results from mathematical fuzzy set theory. It has the advantage of not being dependent on a particular distance measure, so long as it is metric, and can equally-well cope with either Euclidean or proportional city block coefficients (McCune & Grace, 2002). Originally, the technique was designed to be computable with a pair of compasses, as outlined below; elements in the calculation are illustrated in Figure 4.1:

1. A dissimilarity matrix is calculated.
2. A primary axis is defined between two reference points (Q_1, Q_2), usually the pair of entities with the greatest single dissimilarity value.
3. Each subsequent entity (Q_i) is placed with reference to the two initial reference points: two arcs, each of a distance proportional to the dissimilarity value between the entity being placed and one of the reference points (d_{1i}, d_{2i}), are drawn with compasses; a line perpendicular to the axis is drawn from the intersection of those arcs (h_i), and the position of that entity on the primary axis is given by the point on the axis perpendicular to the intersection of those arcs (P_i).
4. A secondary axis is defined by two dissimilar entities positioned near the middle of the first axis, and each entity positioned on the secondary axis as explained above.
5. The two axes of the ordination represent axes of a bivariate scatter plot, with the positions of the entities on the axes representing the co-ordinates in the scatter plot.
6. A third axis for a 3-dimensional solution may be defined by a pair of reference values near the middle of the bivariate scatter plot which nonetheless show a high dissimilarity value.

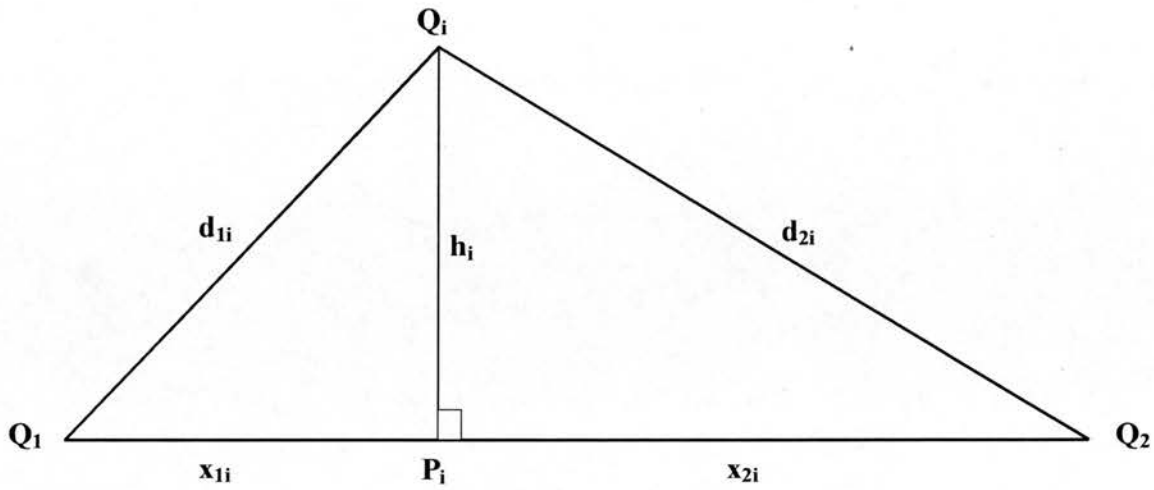


Figure 2.1 Elements in the calculation of Bray-Curtis ordination; after Causton, 1988.

In positioning each entity along an axis, each arc is proportional to the dissimilarity distance between that entity and either one of the reference entities which define that axis (see step 3). The point for a particular entity formed by the intersection of the two arcs creates two right-angled triangles with the line dropped perpendicularly between that point and the axis, and the two reference entities (as illustrated in Figure 2.1). The dissimilarity distance (d_{1i}, d_{2i}), which is to say the radius of each arc, thus forms the hypotenuse of each triangle; the line perpendicular to the axis (h_i) forms one side of each triangle; the third side of each triangle is formed by the distance from each reference point to the position of that entity along the axis (x_{1i}, x_{2i}). Subsequently to its initial formulation (Bray & Curtis, 1957), computer routines using Pythagoras' theorem have obviated the need for compass construction to position each point, and the position of each point on an axis can be calculated straight from the (dis)similarity matrix (Causton, 1988).

The single biggest step influencing the outcome of Bray-Curtis ordination is the selection of endpoints which define the primary axis. Although Bray & Curtis (1957) advised taking the two most-dissimilar entities, this in effect over-emphasises any outliers within the data, so that the ordination results tend to show a cluster of points around one endpoint, with the second endpoint, if it is a true outlier, isolated to one extreme of the ordination. Instead, more modern packages such as PC-ORD offer alternative endpoint selection methods (McCune & Mefford, 1999). The preferred method for endpoint selection is the variance-regression method (Beals, 1984; McCune & Grace 2002):

1. The first endpoint has the greatest variance in its inter-point distances (i.e. has the largest variation in its dissimilarity with each other entity) – this is biased against outliers, since outliers are consistently more dissimilar to other points, which gives a lower variance; instead, an entity from the edge of the main cluster of points tends to be selected.

2. The distances from this initial endpoint to all other points are then calculated
3. Distances from each other point to all remaining points are also calculated.
4. Each set of distances, from each point to all other points, is regressed against distances from the first endpoint to all other points, using simple linear regression.
5. The point whose distance-regression shows the largest negative regression coefficient against the initial endpoint is chosen as the second endpoint; in practice, this usually results in a point being chosen from the opposite side of the main cluster of points from the initial endpoint
6. Remaining points are then positioned on this axis as described above.

As it is a conceptually simple yet ecologically robust method (McCune & Grace, 2002), Bray-Curtis ordination was the first technique investigated here, with genus-level data, both with and without the Beals' smoothing transformation (to investigate the effect of Beals' smoothing).

2.8.2 Detrended Correspondence Analysis (DCA)

Detrended Correspondence Analysis (DCA or DECORANA), is a technique which gained tremendous popularity through the 1980's for the ordination of ecological data. DCA (Hill, 1979a; Hill & Gauch, 1980) is a development of correspondence analysis or reciprocal averaging designed to solve inherent problems of data distortion in the earlier technique. From an initial presence/absence matrix, row and column totals are calculated, representing respectively the number of areas in which the taxon occurs, and the number of taxa occurring in each area. In the absence of any environmental factors (an indirect gradient analysis), weights are assigned to each of the taxa evenly across the range of 0-100. Area scores are derived by multiplying taxon values for that area by the weighting score for that taxon, and then summing and averaging these values by dividing by the number of taxa found in each area. Taxon scores are then calculated by multiplying the taxon value by the *new* area scores, summing these results and then averaging them over the number of areas in which the taxon occurs. After each iteration, taxon scores are rescaled over the range 0-100 (i.e. the lowest value is re-set at 0, the highest at 100). New area scores are calculated from the rescaled taxon scores, and new taxon scores then calculated and rescaled, and so on.

This reciprocal averaging process is repeated until the amount of change in the newly calculated area and taxon scores is minimal. Ultimately the decision as to when to halt the iterations is arbitrary, depending on the precision needed. The rescaled scores in the final taxon column represent the first axis of the taxon ordination, while the rescaled scores in the final area column represent the first axis of the area ordination. The eigenvalues of the first axes, which are a measure of the proportion of the total variation explained by the axis, are calculated from the total range of variation in the final *unscaled* scores as a proportion of the range of variation of the *rescaled* scores of the previous column (i.e. 100). Second axes

are derived with a new set of initial scores (scores for both taxa and areas are calculated simultaneously) taken from 'near the end' of the number of iterations used for the first axis. After each iteration for calculating the second axes a multiple of the first axis is subtracted from the new scores, or the iterations will tend towards to first axis (Hill, 1973).

Correspondence analysis or reciprocal averaging suffers from two major artefacts: the ends of the first axis become compressed with respect to the middle; and the second axis may be a quadratic distortion of the first – this produces ordination plots with a strong 'arch effect' (Kent & Coker, 1992). DCA removes the 'arch effect' by 'detrending': the first axis is divided into a number of segments, and within each segment the second axis scores are recalculated so that they have an average of zero. Done for all segments, this means that second axis scores are expressed as deviations from a mean of zero. Detrending is done for the second axis scores at the end of each iteration. The axis compression effect is overcome by expanding terminal segments and contracting middle segments of the taxon ordination. Taxon ordinations are thus adjusted so that the area scores are the weighted mean values of the scores of the taxa that occur within them (Kent & Coker, 1992). A matrix algebra solution for DCA has also been provided (Hill, 1979a; Legendre & Legendre, 1998).

DCA has recently been the subject of some criticism within ecology, for several reasons (McCune & Grace, 2002). Firstly, the technique implicitly relies on a chi-squared distance measure (McCune & Grace, 2002), which cause it to perform less-well than other techniques for ecological ordination such as non-metric multidimensional scaling (Minchin, 1987a). Also, the detrending process can seek to impose homogeneous axis scores when this is not justified by the data (ter Braak, 1986). There is no way of distinguishing between "horse-shoe" or "arch" distortions and non-linear relationships between species distributions and environmental conditions (Minchin, 1987a); real ecological patterns, caused for example by bi-modal species distributions, may therefore be lost (Beals, 1984). Furthermore, Tausch *et al.* (1995) discovered that the order of the sample units within the data set affected the scores produced by the DCA analysis, although this bug has since been fixed in recent software packages. In view of all the above criticisms, and the availability of superior methods (see below), after initial testing DCA was not pursued as an ordination technique in this thesis.

2.8.3 Non-metric Multidimensional Scaling (NMDS)

Non-metric multidimensional scaling is an ordination technique, developed by Kruskal (1964a,b) from an idea by Shepherd (1962a,b), which is particularly suited to data that are non-normally distributed, or are on arbitrary or discontinuous scales, such as much ecological data; Clarke (1993) and McCune & Grace (2002) describe it as the 'method of choice' for analysing patterns of relationship in multivariate

ecological data. It contains fewer assumptions and implicit methodological constraints than do other frequently-used ordination methods, such as DCA, and the user is free to control the initial stages of the analysis. These are: the standardisation and transformation of data, if appropriate to the data-set, and the construction of a similarity matrix of objects, with an appropriate similarity coefficient chosen by the user. Unlike PCA and DCA, NMDS is not an eigenanalysis method; instead NMDS is grouping objects so that the *order* of relative distances in the ordination matches, insofar as is possible, their relative distances in the underlying similarity matrix. A parameter termed the 'stress' of the ordination measures the discrepancy between objects' position in the ordination and in the original similarity matrix; stress tends to zero if the two sets of relative ordering of the objects are in perfect agreement (Clarke, 1993).

A clever analogy (taken from Clarke, 1993) is: given a matrix of (great-circle) distances between a number of large cities across the world, a NMDS ordination (in two dimensions) seeks to cluster those cities such that the ordination plot actually represents those positions in real space, i.e. the ordination plot would represent a map of the world, with the different cities in their correct positions. With ecological data, however, distances are only relative, defined by a particular similarity coefficient, so in effect the analysis seeks to re-order data objects to minimise the difference between the *rank order* of distances between the objects in the ordination, and the rank order between the objects in the original similarity (distance) matrix, and then re-plot the objects (in two-or multi-dimensional space) with the corresponding ordination distances between them. In practice, the configuration of the objects is found by an iterative steepest-descent algorithm, with objects initially placed at random within the multi-dimensional space, and the positions of different objects relative to each other then continually refined with each iteration until the 'stress' parameter ceases to fall. With each iteration, a new arrangement is retained if the rank order of the objects is closer to that of the distance matrix, and so-on and so-on until there is no further change.

The general procedure is as follows:

1. Calculate a dissimilarity matrix between each pair of sample units.
2. Assign sample units to a starting configuration (e.g. using a random number generator) in multi-dimensional space, for a given number of dimensions.
3. Normalize the starting configuration (by subtracting the axis means for each axis and dividing by the overall standard deviation of scores).
4. Calculate a matrix of Euclidean interpoint distances between sample units in multi-dimensional space.
5. Rank the elements of the original dissimilarity matrix in ascending order, and order the elements of the Euclidean interpoint distance matrix in the same order.
6. Calculate 'stress', defined as the departure from monotonicity between the two sets of rank orders of sample units.

7. Normalize stress values over a scale of 1-100.
8. Minimize stress by changing the configuration of sample units in multi-dimensional space, using a steepest descent algorithm. This completes one iteration.
9. Re-iterate the procedure from step 3 until either a set number of iterations is completed, or stress values stabilize.

Since the NMDS algorithm is iterative, it is possible that a particular analysis will find a local, sub-optimal solution of low stress which is nevertheless still higher than the global solution. Clark (1993) therefore recommends repeating the whole analysis eight or nine times, or until a solution with equally low stress is found in several of the repeat runs. Clarke (1993) also proposed the following rules of thumb for assessing stress values:

- < 5 excellent representation with no prospect of misinterpretation
- < 10 a sound ordination with accurate representation insensitive to the number of dimensions
- < 20 not too much reliance should be placed on the details of the ordination plot; another plot with more dimensions could reveal a different picture
- > 20 plots become dangerous to interpret sensibly at any scale
- > 40 objects are effectively randomly placed, and the ordination is meaningless

Non-metric multidimensional scaling was the ordination method of choice for the analyses presented both in Chapter 5 and also in Chapter 6. As an ordination method it contains fewer assumptions and implicit methodological constraints than do other frequently-used ordination methods, such as DCA, and the user is free to control the standardisation and transformation of data and select the appropriate similarity coefficient. It has also shown a superior performance in simulation studies (Minchin, 1987a; McCune, 1994).

2.9 Summary

- Numbers of families and genera for each TDWG Level 2 Region have been taken from the Distributions of Vascular Plant Families and Genera database.
- These have been supplemented by extrapolated numbers of species for the same regions taken from the World Checklists and Bibliographies database.
- Hierarchical clustering techniques were selected to study floristic relationships between regions.
- Ordination by Bray-Curtis ordination and by non-metric multidimensional scaling were also selected to study floristic relationships between regions.

- Non-metric multidimensional scaling together with non-hierarchical clustering by *k*-means was selected to study biogeographical relationships between taxa.

CHAPTER 3

ESTIMATING GENERAL PATTERNS OF DIVERSITY

This chapter deals with several separate, yet inter-related, issues concerning the broad-scale global distribution of plant biodiversity. These issues are: the correlation between patterns of diversity at different taxonomic scales and the use of higher taxa as biodiversity surrogates; patterns of taxon range-size and their similarity to patterns of taxon size; the relationship between area and diversity for different regions of the world; the geographical variation in taxonomic richness at different scales, and the relative degrees of endemism of different regions; and the degree of compositional change and the decay in floristic similarity with increasing distance between regions. Methods are outlined in the text for each section. The relationship between diversity and latitude is granted the following chapter to itself. Having established these general patterns of diversity, questions concerning floristic relationships between individual regions, or the distributions of individual families or genera, are addressed in more detail in subsequent chapters. The final chapter then revisits these general patterns in the light of the results from the more detailed analyses.

3.1 Taxonomic surrogates for biodiversity

Species richness (simply the total number of species) remains the most common measurement, the 'common currency', of biodiversity (Gaston, 1996a, 1996b, 1998; see also Chapter 2) – there are no complicated indices needed to calculate species richness, while it is a widely understood and (relatively) easily measurable parameter which seems to capture much of the 'essence' of biodiversity, and for which much data already exists (Gaston 1996b). Species richness is itself often broken down into three separate components (Whittaker, 1972), which reflect the variation in patterns of diversity at different spatial scales: alpha (α) or point diversity is simply number of species, measured within a uniform area or habitat at (very) local scales, often in a standard-sized plot; gamma (γ) diversity is also number of species, but measured across large ('landscape' or 'regional') scales – often taken as the total cumulative species diversity of many alpha diversity plots; beta (β) diversity, the degree of change in species composition along an environmental gradient or between sites, connects the two. Patterns of gamma diversity at all taxonomic levels are tackled in Section 3.5, and related to previously-published studies of species-level alpha diversity in Section 3.7; patterns of beta diversity at genus level are studied in Section 3.6.

To complement the numbers of families and genera per TDWG Level 2 Region, numbers of species for each region were extrapolated from the World Checklists and Bibliographies database as described in Chapter 2 (page 37), multiplying each value by 2 to give total species richness estimates for each Level 2 TDWG region, on the basis that the total in this database was 50% of the total number of accepted species names in Index Kewensis (cf. Govaerts, 2001, 2003). As well as being used to analyse the distribution of species diversity itself, these extrapolated species numbers can also be correlated with numbers of families and genera, hopefully to provide some additional justification for studying patterns in the diversity and distribution of higher taxa (see also Chapter 7) if they do indeed stand as reliable surrogates of species-level diversity. As described in Chapter 2 also, the accuracy of these extrapolated figures rests on two assumptions, both of which cannot yet be validated: that rates of synonymy are equivalent between the families contained in the World Checklists and Bibliographies database and those families not yet tackled; and that there has been no geographic bias in the selection of families already treated, i.e. that the database contains an equivalent proportion of predominantly-tropical or predominantly-temperate families as really exist.

Bearing these assumptions in mind, there is a reassuringly strong correlation between numbers of species and numbers of genera and families for each Region, although a power-law or double-logarithmic relationship provides a better fit for the data than does an untransformed one (Figure 3.1; see also Enquist *et al.*, 2002). Comparing the three taxonomic scales, patterns of relative diversity within a region are highly correlated within regions, irrespective of the size of the region (Spearman's r_s: between species richness and genus richness, 0.97; between genus richness and family richness, 0.96; between species richness and family richness, 0.93; $n = 52$, $p < 0.01$ in each case), confirming that higher taxonomic levels may indeed be a reasonable surrogate for species richness (Gaston, 1996b; La Ferla *et al.* 2002; but see also Prance, 1994). Although a non-linear relationship between diversity at different taxonomic ranks is shown in Figure 3.1, the non-parametric Spearman's rank correlation, being based on relative ranks rather than directly on numbers of taxa, is insensitive to the form of the diversity relationship (i.e. whether or not the relationship between different taxonomic ranks is linear or non-linear).

The relationship between numbers of species and numbers of higher taxa in a region is only partly dependent on the size of that region. For example, Regions 12, Southeastern Europe and 13, Southwestern Europe are similar in size (see Table 1.1) and also have similar generic (1031 and 1093 genera, respectively) and specific (6513 and 7778 species, respectively) diversities. However, Regions 25, East Tropical Africa and 26, South Tropical Africa) have similar generic and specific diversities but are very different in size (1773068 km^2 and 3298998 km^2), while Regions 32, Central Asia and 34, Western Asia are similar in size but have quite different numbers of genera (917 and 1476, respectively) and species (9726 and 20662, respectively). This relationship is further complicated by latitude: with two regions of equivalent size but of different latitudes (for example, Regions 78, Southeastern U.S.A. and 82, Northern South America) the region of lower latitude has many more

species (4164 and 18308 species, respectively) than would be predicted simply by the number of genera (1018 and 2053 genera, respectively). The number of species in two regions of the same size is therefore not always accurately predicted simply by the number of genera in that region.

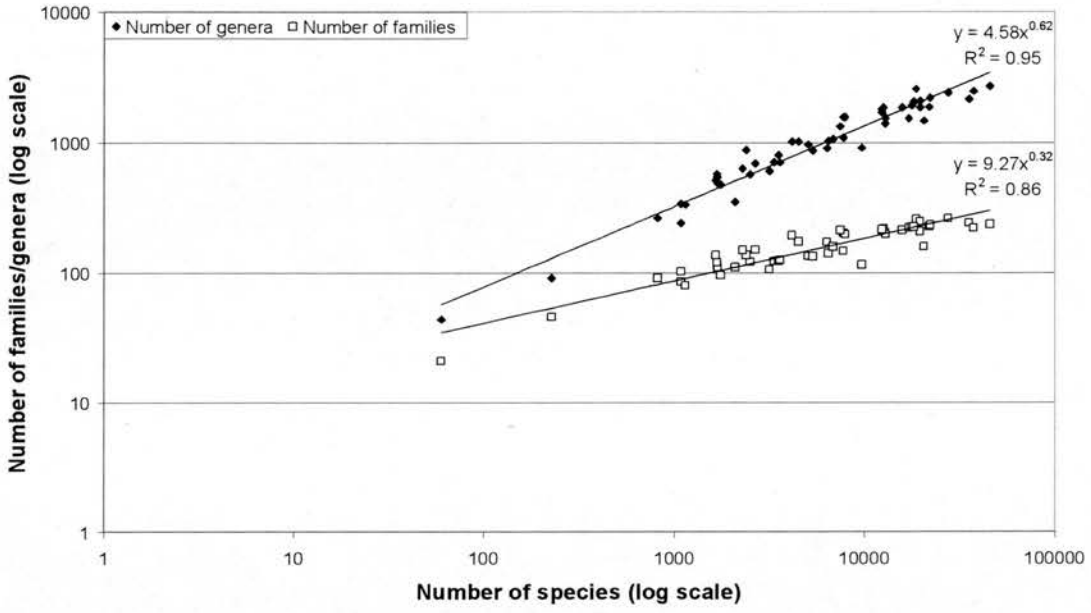


Figure 3.1 Relationship of numbers of higher taxa (families and genera) to estimated numbers of species for TDWG Regions (excluding Region 91, Antarctic Continent). A power-law relationship, presented on double-logarithmic axes, provides the best fit of the data.

3.2 Patterns in genus-level frequency distributions

3.2.1 'Hollow curve' frequency distributions

The frequency distribution of taxonomic categories is strongly skewed: there are many monotypic taxa, and very few, very large taxa (Clayton, 1972, 1974; Cronk, 1989; see also Chapter 1). This is observed at successive taxonomic levels (Clayton, 1972, 1974; Cronk, 1989) – species within genera, genera within families, etc. – although only the level of genera within families may be investigated here with the data from this database. Figure 3.2 presents a frequency distribution of family size, measured as numbers of genera per family, for those 395 families with fewer than fifty genera; family size has a clear modal value of only one, and there are more than three times as many families with only one genus as there are with only two genera. Conversely, however, the 61 families which have more than fifty genera cumulatively account for some 11143 (82%) of genera. This pattern is independent of questions of monophyly: Figure 3.3 shows numbers of both 'traditionally' recognised families (Brummitt, 1992), some of which may be paraphyletic, and exclusively-monophyletic families (APG, 1998) against sizes of families (again measured by numbers of genera),

on log-transformed axes. Both 'traditional' and exclusively-monophyletic classifications occupy the same space on the graph, with few large families and many small families. Differences between the two classifications are usually in the position of the numerous small families, since large families are generally distinct crown clades with well-established monophyly and are therefore recognised in either classification.

The shape of the frequency distribution of family size (Figure 3.2) is the well-known 'hollow-curve' (Willis, 1922; Williams, 1943), where the modal (most frequent) value of the graph is also the lowest value. The hollow curve is shown not just by taxon size but also by spatial taxon distribution, for both families and genera (many narrowly distributed taxa and few very widely distributed taxa; Colwell & Lees, 2000; Gaston, 2003; see Figures 3.4 and 3.5), and also temporal taxon distribution (many short-lived taxa and few very long-lived taxa; Rosenzweig, 1995), further suggesting that it is not just an artefact. Figure 3.4 shows the range size frequency distribution of genera, where range size is measured simply as a count of TDWG Regions per genus; the inset graph reproduces this information as a scatter plot on log-transformed axes: the relationship is negative and approximately linear. Figure 3.5 shows the range size frequency distribution for families: though the modal value is again 1, there are greater numbers of intermediate- and large-ranged families than expected by a simple hollow-curve distribution or shown by the genus range size frequency distribution. Though species-level data are not included here, species range-sizes would be expected to be an even steeper hollow-curve than for genera (even more localised species and proportionally fewer widespread ones); however, this would probably be more a reflection that the geographical scale used here is too coarse to record species distributions than an accurate representation of species range-sizes. Figure 3.6 presents a frequency distribution of distribution patterns: the number of genera found with each unique combination of TDWG regions. As with the other frequency distributions, Figure 3.6 is highly skewed towards many genera in a few common distribution patterns and most distribution patterns only being shown by single genera; the inset graph shows only the 25 most common distribution patterns, with that combination of TDWG Regions shown for each: 15 of these 25 (60%) are distribution patterns which are endemic to single regions.

3.2.2 The potential number of possible distribution patterns

A distribution pattern is defined for the purposes of this study as any unique combination of native occurrence(s) in any of the 52 TDWG Level 2 Regions. With 52 regions, and any individual genus able to occur or not occur in each of those regions, there will be a total of 2^{52} possible distribution patterns, or unique combinations of those regions, making up the complete set of distribution patterns. The maximum distribution range is obviously a genus present in every region (although this is not actually shown by any genus); the minimum distribution is not those genera endemic to a single region, but a null distribution absent from every region, since in theory this is also a potential distribution. However, since recently-extinct genera are here treated as native in their

former range, and intergeneric hybrids have been excluded from this analysis, in practice no genus actually shows this null distribution. Assuming therefore that we are not interested in the single empty (null) distribution pattern containing no genera, then this still leaves $2^{52}-1$ potential generic distribution patterns – or more than 4.5×10^{15} possible distributions!

However, for angiosperm genera only 2817 separate combinations of regions are actually found. There obviously cannot be more than 14,304 actual generic distributions, since there are only 14,304 genera of angiosperms, and since some 5062 (38% of genera) of these are endemic to single regions (46 regions out of 52 have endemic genera), 62% (8868 genera) must be found in the 2771 distribution patterns occurring in more than one region. The number of possible unique generic distributions is therefore much less than the potential number mathematically. The frequency distribution of distribution patterns (as distinct from range-sizes) shows, as might be expected, a very pronounced hollow curve (see Figure 3.6). That is, as well as the majority of distribution patterns occurring in more than one area, the majority of distribution patterns are also unicate, only shown by one genus, and only a small number of actual unique distribution patterns are each shown by many genera. Of the 2817 distributions, c. 2200 are unicate (shown by only one genus); only c. 600 distributions are shown by more than one genus, but collectively these 600 distributions account for about 11000 genera. Respectively, only just over 20% of distribution patterns account for just over 80% of genera, while just under 80% of distribution patterns account for only about 20% of angiosperm genera.

Therefore, the majority of angiosperm generic diversity is shown by a few repeating distribution patterns over several regions (see Chapter 6). Some 38% of genera are endemic to a single region, but there is no genus found in every region, and only three genera (*Carex* L. and *Cyperus* L., both Cyperaceae; and *Plantago* L., Plantaginaceae) in 51 regions, found everywhere except Antarctica. That the frequency distribution of distribution patterns is so right-skewed is not necessarily evidence that genera of angiosperms are distributed in a non-random pattern; if all the angiosperm genera in the world were randomly distributed the frequency distribution of distribution patterns would still be right-skewed. However, the frequency of the actual individual distribution patterns might well than be different – that is, those 25 most common distribution patterns in Figure 3.6 would be unlikely to occur if genera were distributed completely at random. The dependence of numbers of genera on factors such as area (see Section 3.3), latitude (see Chapter 4) and distance from neighbouring regions (see Chapter 6) and the correspondence of distribution patterns to floristic regions (see Chapter 5) imply that the particular frequencies of genera within individual distribution patterns are not random, even though a random distribution of genera may well produce a similarly-shaped frequency distribution of distribution patterns.

3.2.3 The relationship between range size and diversity

Figure 3.5 gives the range size frequency distribution for families of angiosperms; there are proportionally more widespread families than there are genera (see Figure 3.4). Though not all widespread families have many genera, all families with many genera are themselves widespread – and conversely all families which have small distributions have few genera. However, the overall correlation between family size and family range size is poor (Spearman's r_s: 0.40; n = 456, $p <$ 0.001) because there are many families which have few genera but are nonetheless very widespread (e.g. Plantaginaceae; 3 genera with a combined distribution of 51 regions). Most of these small-but-widespread families are aquatic, perhaps indicating greater dispersal capacity in these families between areas of aquatic habitats which themselves have an almost global distribution.

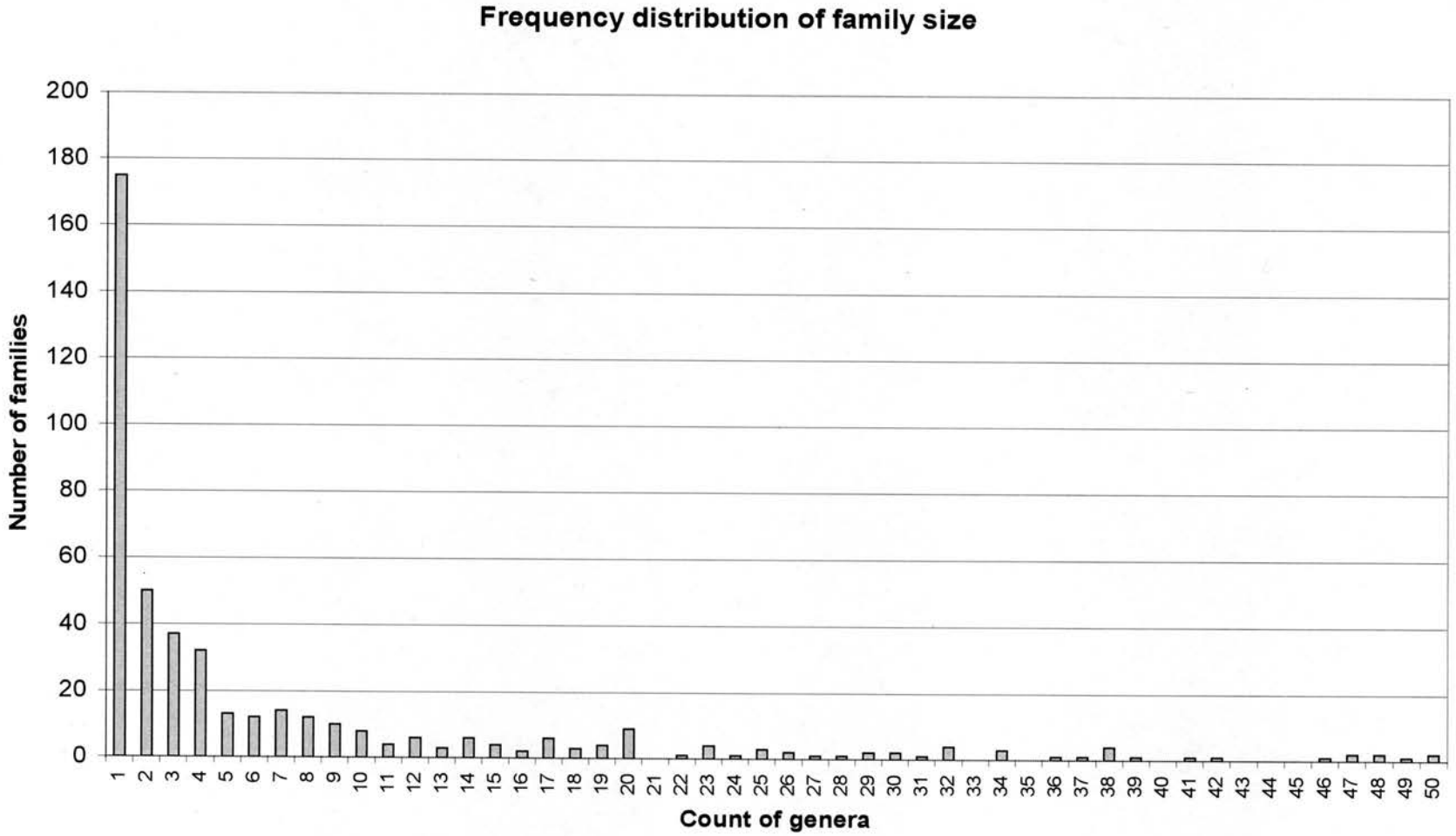


Figure 3.2 Frequency distribution of family size, measured as number of genera: there is a clear modal value of only 1 genus per family. Due to space constraints only families with fewer than fifty genera are shown.

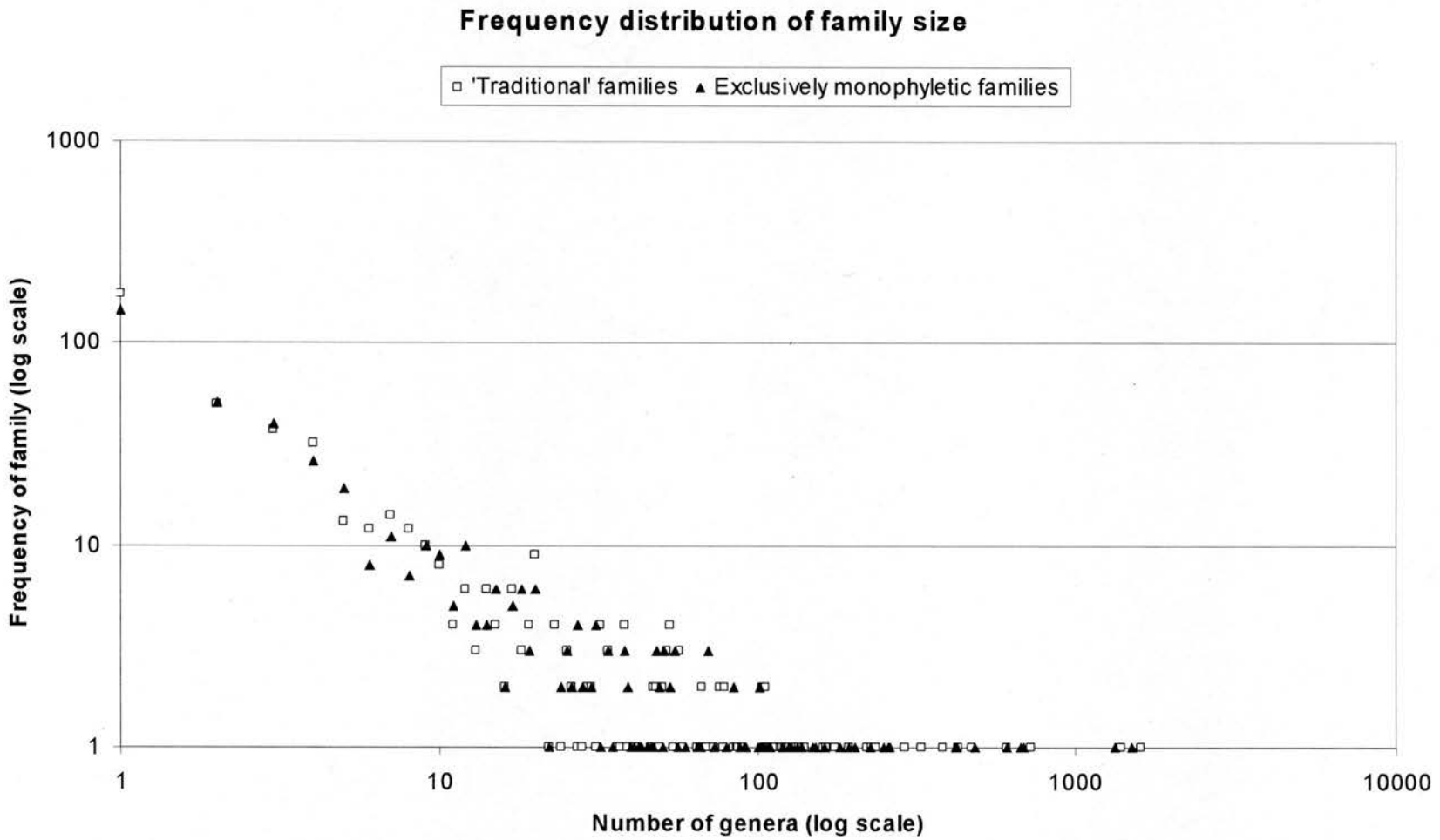


Figure 3.3 Double-logarithmic frequency distribution of family sizes for both 'traditional' (including non-monophyletic) families (Brummitt, 1992) and exclusively-monophyletic families (APG, 1998). Though the status and circumscriptions of individual families may differ between the two classifications, the overall size-distribution of all families does not.

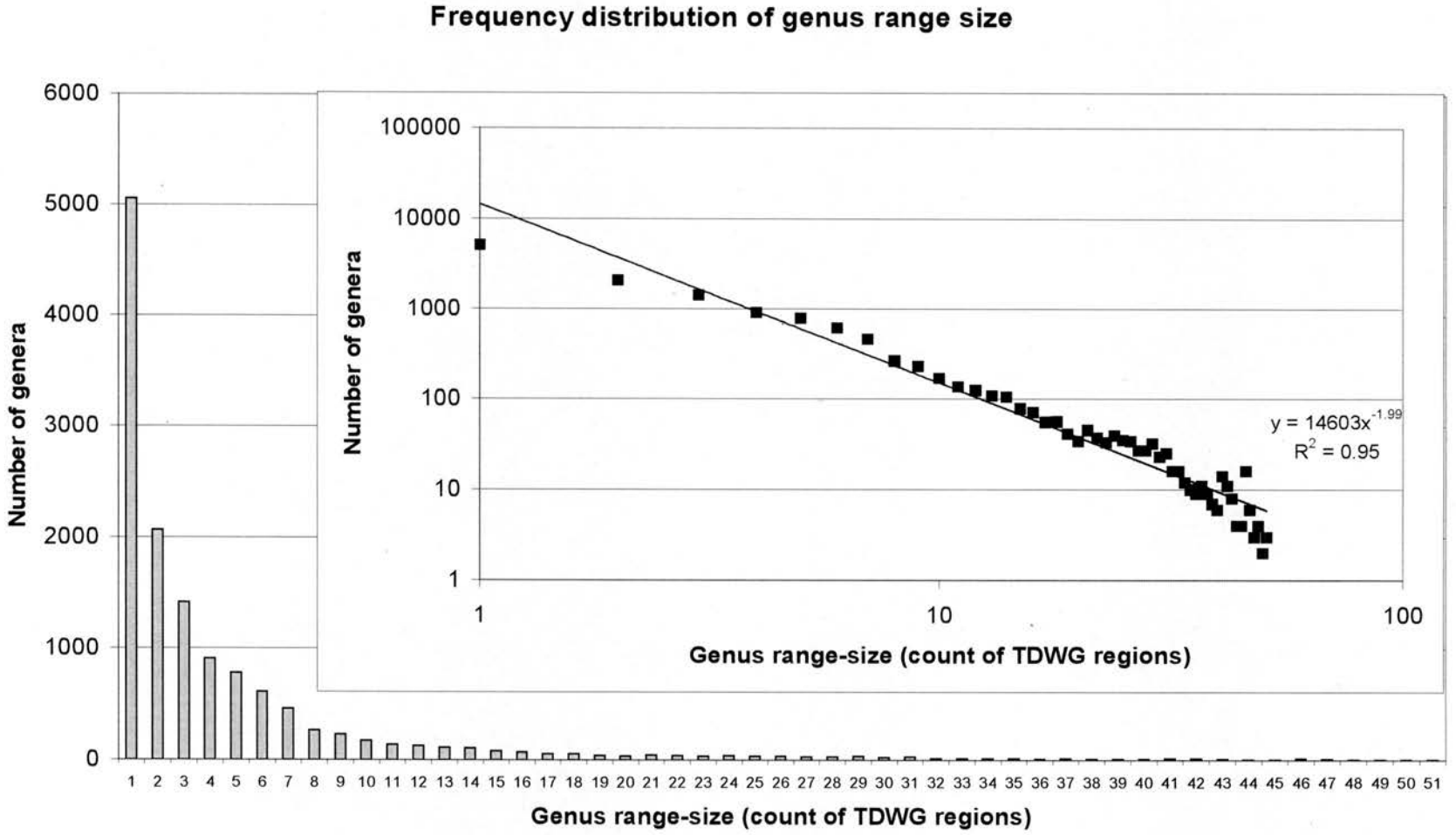


Figure 3.4 Frequency distribution of genus range size, measured as the number of TDWG regions per genus: the modal value is clearly 1; the inset graph displays this same data as a scatter diagram on log-transformed axes.

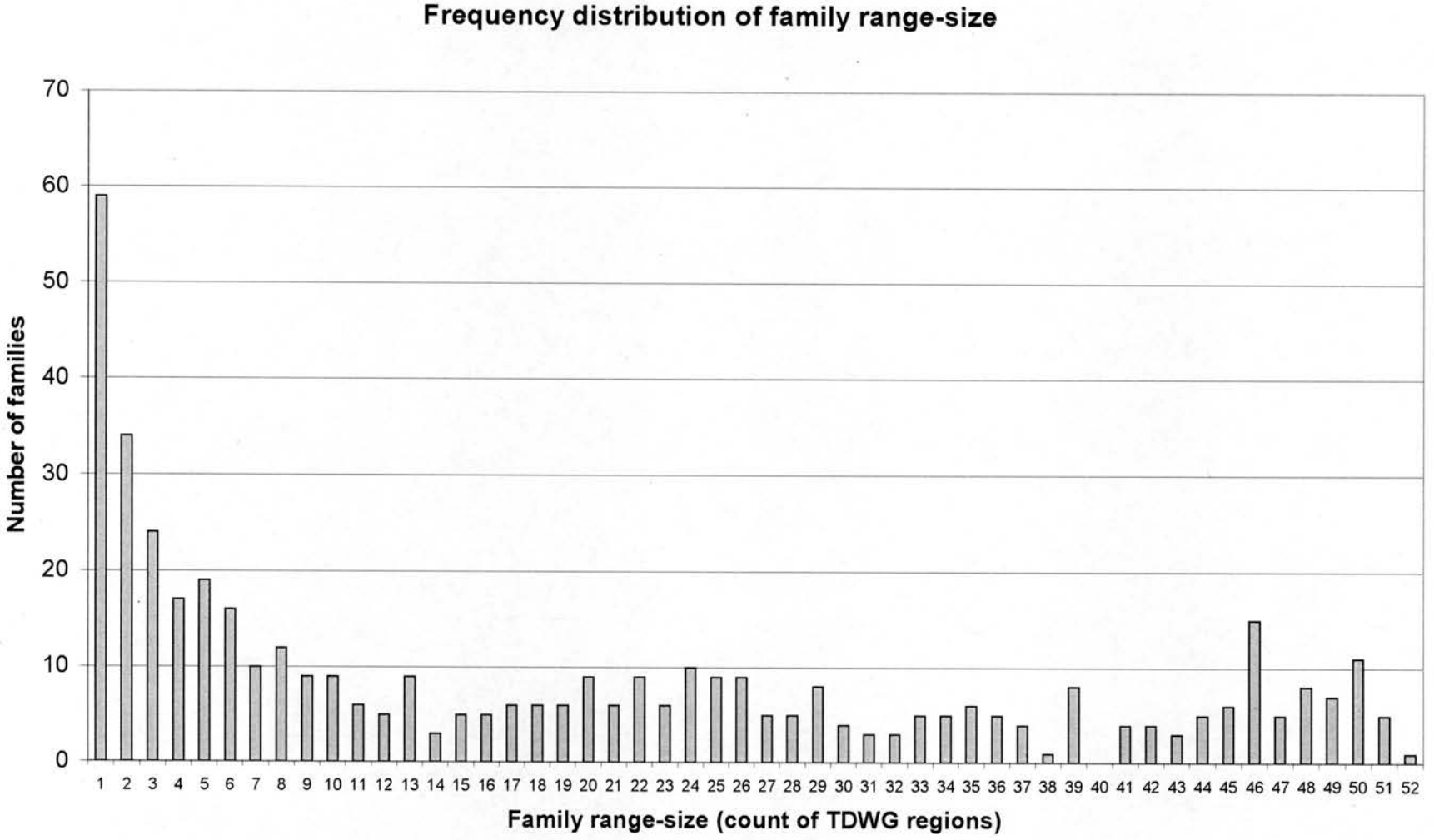


Figure 3.5 Frequency distribution of family range size, measured as the number of TDWG regions per family.

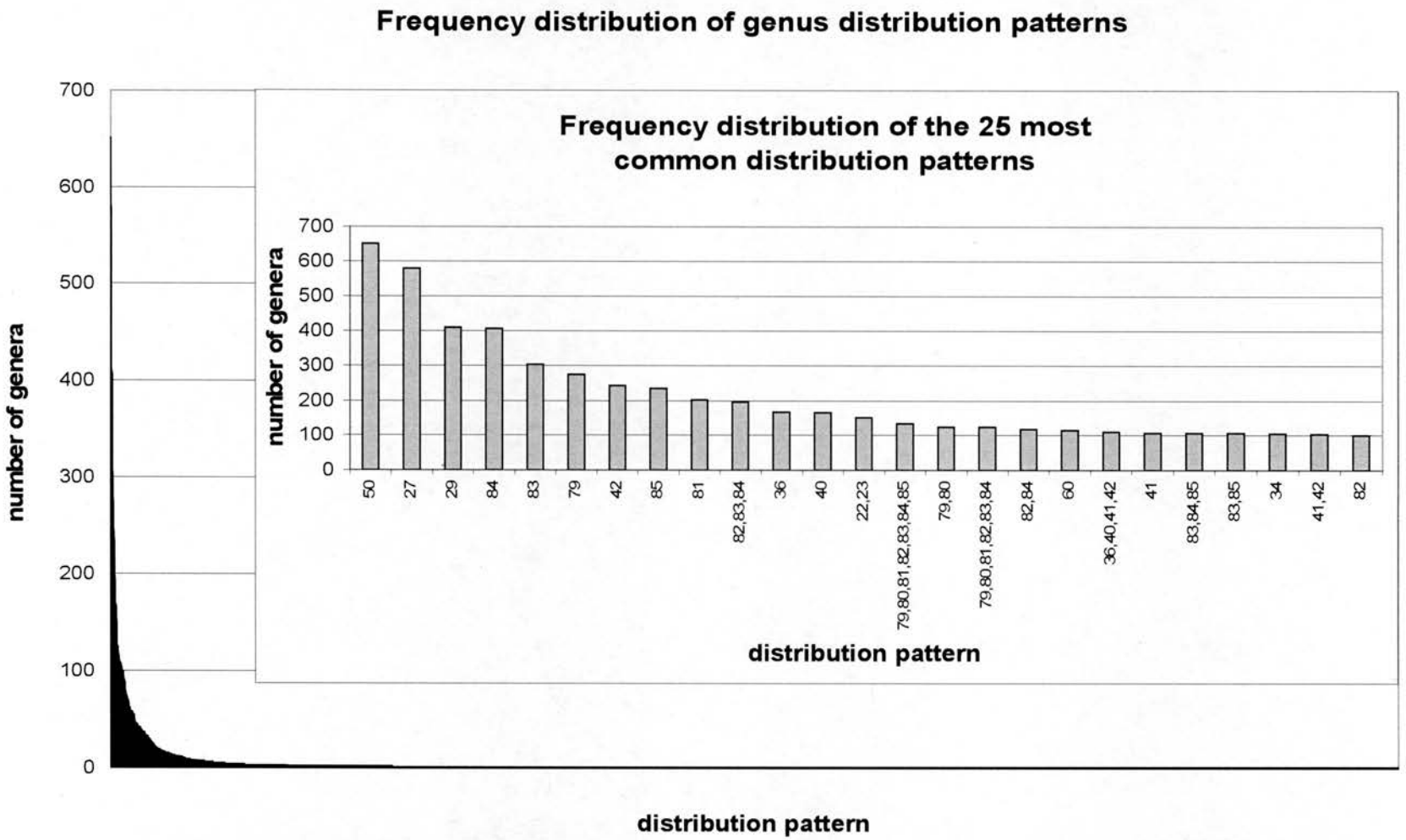


Figure 3.6 Frequency distribution of genus distribution patterns; the inset graph shows only the 25 most common distributions, with the TDWG regions indicated below each.

3.3 The relationship of diversity to area

3.3.1 Species-area relationships

It seems self-evident that larger areas should contain more species than smaller areas, yet the relationship between area and the number of species present in that area has been described as 'one of the most important patterns in biogeography' (Lomolino, 1989) and 'one of ecology's few universal regularities' (Schoener, 1986). Although the exact form of the relationship is still not agreed upon (Connor & McCoy, 1979), it is conventionally expressed by the model

$$S = cA^z$$

where S = number of species, A = area and c and z are fitted constants. This is a power-law or double-logarithmic relationship, which is also known as the Arrhenius equation, and may be expressed as

$$\log S = z \log A + \log c$$

The classic species-area relationship with a nested set of sub-plots forms a straight line plot in logarithmic space, and the two fitted constants, which can be estimated through simple linear regression, are interpreted as the slope (z) and the intercept (c) of the regression line. In reality, however, the slope must intercept at the origin (an area of zero size must have zero species in it), so the coefficient c is more correctly interpreted as the slope of a graph of S (y-axis) and A^z (x-axis) (Rosenzweig, 1995). A log-log graph of regional family, genus and species richness against area is presented in Figure 3.7. [Note: graphs and regressions do not include values for Antarctica, a region of more than 12 million square kilometres but with only two native angiosperm genera; including Antarctica considerably reduces the slope without being informative of global diversity patterns]

At each taxonomic level the relationship shown in Figure 3.7 is broadly linear in log-log space, with larger regions showing greater numbers of taxa than do smaller regions. The spread of points obviously increases from family- to genus- to species-level, with the exponent of the species area relationship likewise increasing from 0.12 to 0.26 to 0.35, respectively. The increasing exponent values simply reflect the structure of the taxonomic hierarchy: there can be several species within one genus or genera within one family, but not vice versa; therefore the numbers of taxa will inevitably increase with decreasing taxonomic rank. However, the increasing spread of data points with decreasing taxonomic rank reflects the increase in the proportion of tropical taxa at lower ranks: the ratio of tropical species : genera is greater than the ratio of temperate species : genera. That is to say, the strength of the latitudinal gradient of diversity increases with decreasing taxonomic rank (see also Section 3.5).

In Figure 3.8 only genus-level data, the primary interest of this thesis, is presented, with individual regions identified by their two-digit code (see Table 1.1). Inspecting Figure 3.8, there is an obvious cluster of points towards the upper right hand corner. This reflects the efforts made in the TDWG Geographical Scheme (Brummitt, 2001) to divide the world up into equal-sized political

subdivisions, as far as is possible; most areas are within one order of magnitude, from 1 to 10 million square kilometres. A choropleth map of absolute genus richness for each region of the world is given as Figure 3.9. Examining the regression residuals, the spread of points around the line is obviously due to latitude. Tropical areas are mostly found above the trend line, in a gently arcing line, from small tropical island systems (Region 28, Middle Atlantic Ocean [St. Helena and Ascension I.]) to large continental areas (Region 84, Brazil); the arcing is due to the geographically scattered (non-nested) regions. Temperate areas are mostly found below the trend line, steadily increasing in latitude down to the areas farthest from the equator (Region 30, Siberia, and Region 70, Subarctic America), which are the least diverse continental areas. This can also be seen among the smaller island regions: the Subantarctic Islands (Region 90) are much less diverse than comparably-sized tropical islands (North-Central Pacific [Hawaiian Islands], Region 63); the same is true for New Zealand (Region 51) *versus* the Caribbean (Region 81). Latitudinal patterns in genus diversity are studied further in Chapter 4.

It is possible to detect geographic factors within the distribution of tropical regions. Neotropical areas are most diverse, with tropical Asian areas slightly below and African regions slightly below them (see also Figure 3.9). Regions both directly north (Regions 34, 36 and 38) and south (Regions 27, 50 and 85) of the equator have greater diversity than would be expected purely from their latitude; this is presumably due to 'spill-over' of genera found in geographically adjacent tropical regions. Below that, there is a band of 'warm-temperate' regions which is clustered in two distinct groups: Regions 12 & 13 (collectively southern Europe) with Regions 76, 77 & 78 (collectively southern U.S.A.), and Regions 20 (Northern Africa), 32 (Middle Asia) & 35 (Arabian Peninsula) with Region 14 (Eastern Europe). In the case of Regions 20 and 35, this is presumably due to great aridity over most of these areas (Archibold, 1995) suppressing the diversity expected at that latitude (much lower than Regions 14 and 32); however, it may be argued that it is the aridity itself which is expected at that latitude, not high diversity. Below these clusters is a diffuse group of north temperate regions, both Old and New World (Regions 10, 11, 30, 31, 37, 70, 71, 72, 73, 74 & 75).

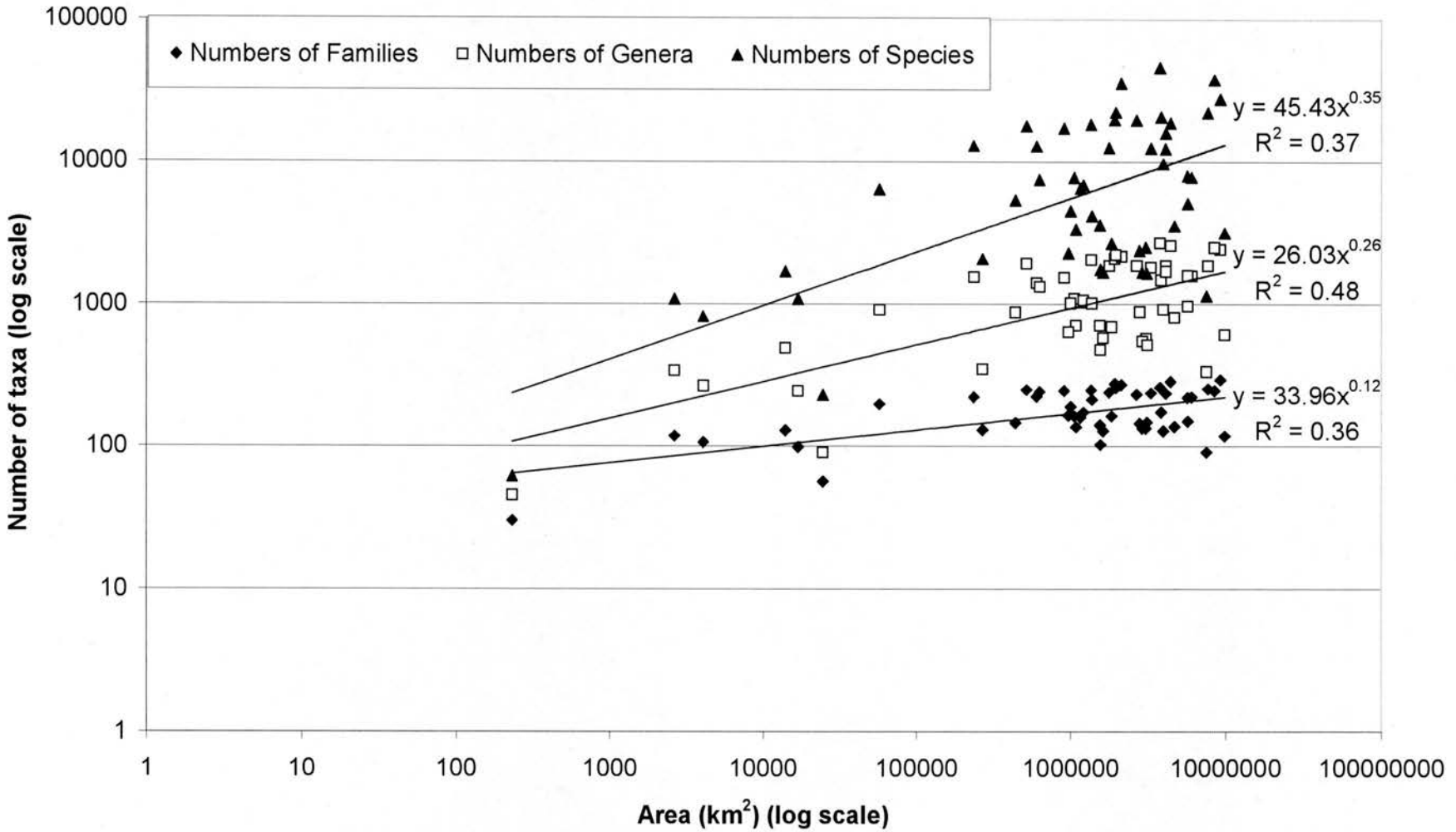


Figure 3.7 The relationship between diversity and area for three taxonomic levels in 51 TDWG Level-2 regions across the world (excluding Antarctica).

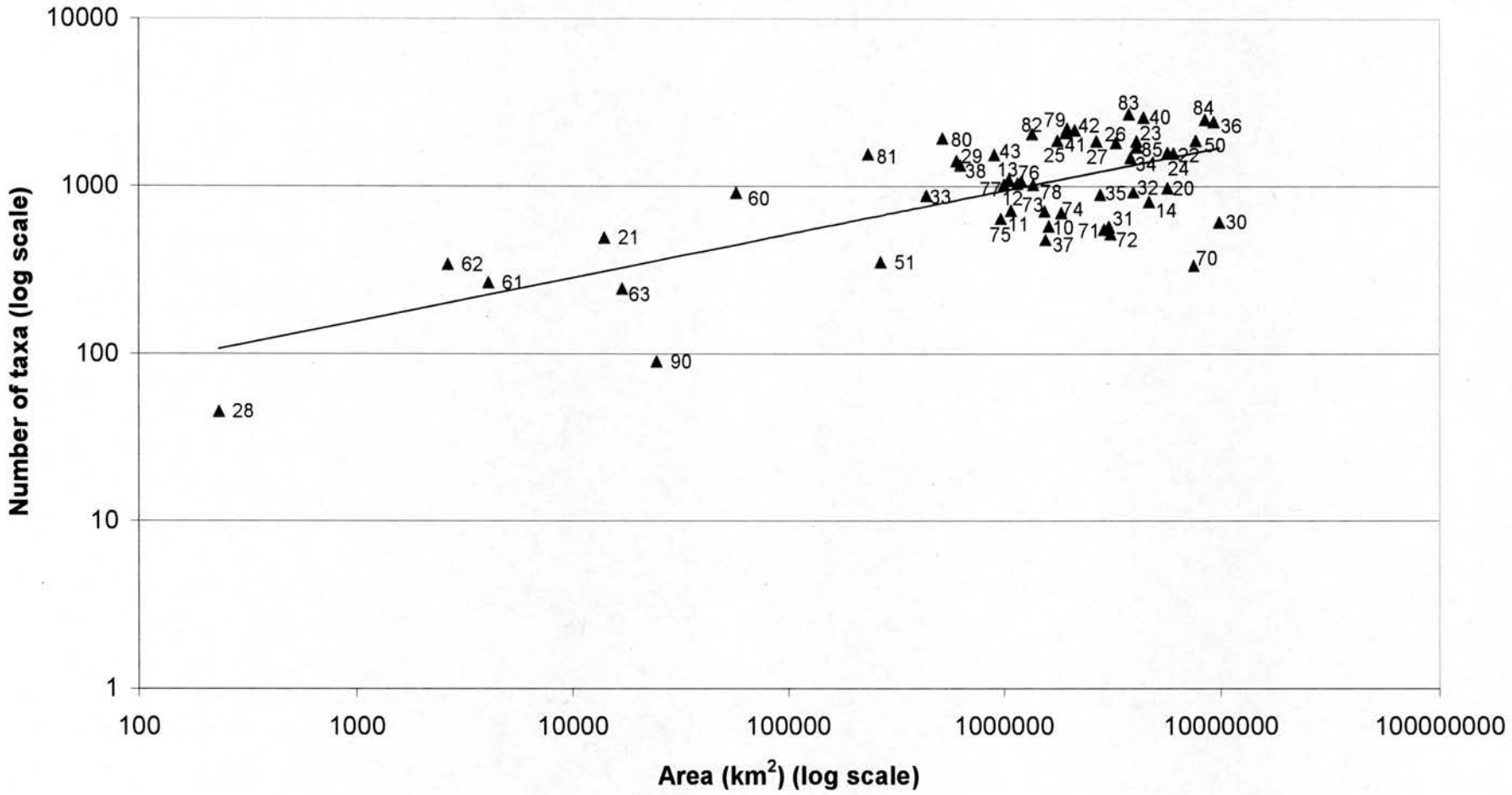


Figure 3.8 Genus-area plot for angiosperms in 51 TDWG Level-2 Regions across the world (excluding Antarctica); regions are identified by their two-digit code (see Table 1.1).

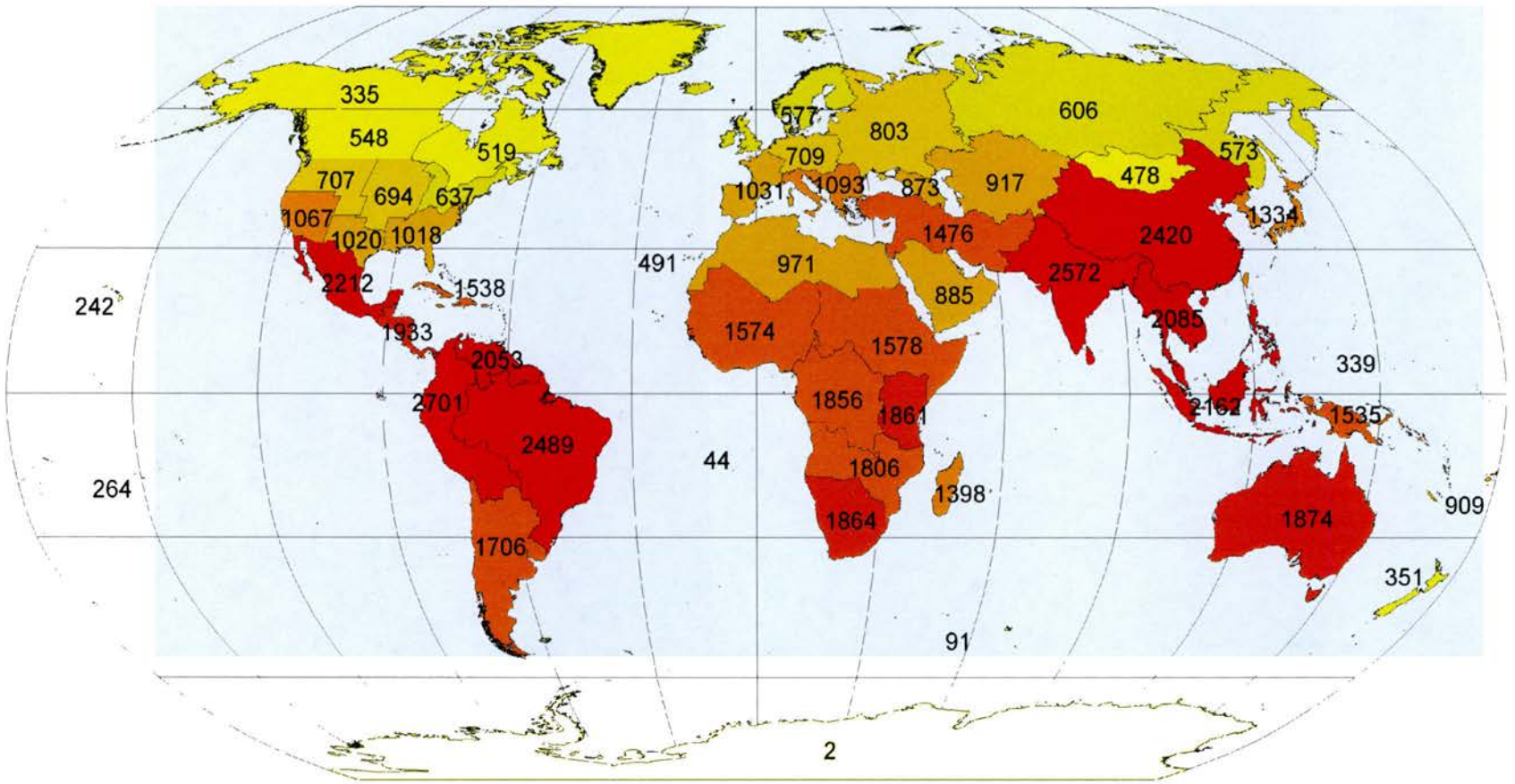


Figure 3.9 Choropleth map of absolute numbers of genera for TDWG Level 2 regions across the world; graduated colours range from yellow (low diversity) to red (high diversity), and actual numbers of genera are given in boxes for each region (see also Table 2.1)

3.3.2 Nested genus-area curves

Although the 52 TDWG Regions used here are contiguous geo-political entities (see Figure 1.1), by querying successively for numbers of genera in ever-larger groups of regions it is possible to approximate the nested design of classical species-area plots (MacArthur & Wilson, 1967; Rosenzweig, 1995). This was done several times, beginning each time in a different TDWG Region and finishing with the whole world. Individual nested genus-area plots are shown together in Figure 3.10; regression statistics, including intercept and slope values, are given in Table 3.1 below. Note that, because most genera in any one region are also found in adjacent regions (since total degree of generic endemism is 38%; see Section 3.2), totals used in nested species-area plots must be queried each time and not simply summed from total numbers of genera between regions, in order to avoid including twice genera which occur in adjacent regions; cumulative area, however, which must be mutually exclusive, can simply be summed between regions.

Classical, nested species-area plots have high r^2 values approaching unity (i.e. the relationship is almost a straight line in double-logarithmic space; Rosenzweig, 1995), and this is also true for patterns of genera in the world (see Table 3.1). Since each analysis begins in a region with a different initial diversity, but all of them converge on the whole world, intercept and slope parameters (c and z values) differ between different initial starting positions. However, although these values depend to a small extent on the sequence of addition of regions (and hence cumulative area) of the nested design, in each case there is a remarkably high r^2 value (with the exception of that for Region 74, North-Central U.S.A., it is always greater than 0.9) given that collectively the relationship between generic diversity and sizes of regions had an r^2 value of only 0.48. The plot of diversity against area for different regions presented in Figure 3.7 does not show as strong a relationship because distributions of genera overlap between regions and diversity of individual regions is not spatially independent.

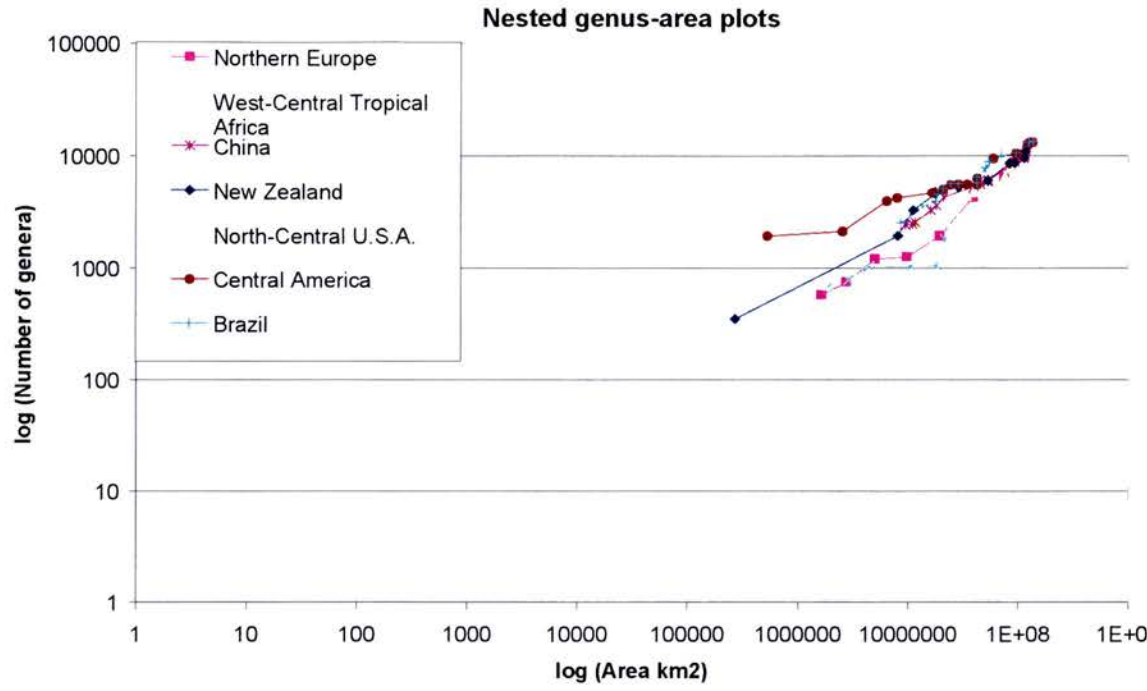


Figure 3.10 Genus-area plots on double-logarithmic axes, with a successively-nested design, beginning in different areas of the world. Regression statistics for these plots are given in Table 3.1.

TDWG Region	no. of genera	r^2	Intercept	Slope
Northern Europe	577	0.98	-1.73	0.71
West-Central Tropical Africa	1856	0.94	-0.69	0.58
China	2420	0.98	-0.69	0.58
New Zealand	351	0.94	0.71	0.40
North-Central U.S.A.	694	0.89	-2.19	0.77
Central America	1933	0.92	1.13	0.36
Brazil	2489	0.97	-0.62	0.58

Table 3.1 Regression statistics for nested genus-area plots for the world, each beginning in a different TDWG Region (see also Figure 3.9).

3.4 Patterns of generic endemism

The level of Regional endemism for families is 10%; family distributions in general are therefore not spatially-restricted enough for patterns of family endemism to be of much interest (see the family range-size frequency distribution in Figure 3.5); and, although an overall level of species endemism was calculated (see Section 3.6), it has not been possible to calculate species endemism values for each TDWG Level 2 Region. Table 3.2, however, gives numbers of endemic genera and relative degrees of endemism for each TDWG Region. Patterns of generic diversity and patterns of generic endemism are less well correlated than is diversity at different taxonomic ranks (Spearman's r_s 0.81, n = 52, $p < 0.01$); the areas with the greatest numbers of genera are not those with the greatest degree of endemism, and areas of high endemism are not necessarily particularly rich in genera. The relationship between number of endemic genera and area of TDWG regions is given in Figure 3.11. A choropleth map of absolute numbers of endemic genera is given in Figure 3.12, while Figure 3.13 gives a choropleth map of percentage generic endemism.

Though all the regions with highest genus richness have moderate degrees of endemism, those regions with the highest degree of endemism (Region 27, Southern Africa; Region 29, Western Indian Ocean [≈ Madagascar]; and Region 50, Australia) have themselves moderate genus richness (see Figure 3.12). It is notable that amongst tropical regions, no tropical African region has generic endemism greater than 10%; indeed, Region 22, West Tropical Africa; Region 25, East Tropical Africa; and Region 26, South Tropical Africa, all have levels of generic endemism (2.2%, 2.8% and 1.9%, respectively) lower than that for Region 12, Southwestern Europe (3.0%). In SE. Asia and the Neotropics, on the other hand, Region 42 Malesia [10.8%] and all of Region 79, Mexico [11.8%]; Region 81, Caribbean [12.6%]; Region 83, Western South America [10.8%]; Region 84, Brazil [15.7%]; and Region 85, Southern South America [13.6%] have values for generic endemism greater than 10%. Assuming still that patterns in the distribution of genera truly reflect underlying patterns in species distribution, this implies that levels of speciation have been far greater in extra-African tropical regions, and in the Neotropics in particular.

Amongst temperate regions, some (Region 32, Central Asia [6.0%]; Region 34, Western Asia [6.6%]; Region 76, Southwestern U.S.A. [8.4%]) have levels of generic endemism greater than many tropical regions (all of tropical Africa; Region 40, Indian Subcontinent [6.1%]; Region 43, Papuasia [5.7%]; Region 80, Central America [3.7%]; and Region 82, Northern South America [4.7%]). Other, cold-temperate regions, however, have absolutely no endemic genera (Region 10, Northern Europe; Region 11, Middle Europe; Region 71, Western Canada; Region 72, Eastern Canada; Region 74, North-Central U.S.A.; and, not surprisingly, Region 91, the Antarctic Continent) – and Region 14, Eastern Europe and Region 70, Subarctic America, have only a single one apiece. Several isolated island regions (Region 28, Middle Atlantic Ocean [St Helena and Ascension I.]; Region 51, New Zealand; Region 63, North-Central Pacific [Hawaiian Islands]) have moderate degrees of endemism but relatively low genus richness. In fact, Region 28, Middle Atlantic Ocean, (by far the smallest

TDWG Level 2 Region at only 232 km^2, at least one order of magnitude smaller than the next smallest region) has a remarkable 11 genera endemic out of only 44 native angiosperm genera.

The relationship between generic endemism and the sizes of regions is shown in Figure 3.11 below; there is almost no relationship ($r^2 = 0.06$) – although small regions never have large numbers of endemic genera, large regions do not necessarily always have large numbers of endemic genera. The first- and third-highest scoring regions for endemic genera, 50 (Australia) and 84 (Brazil) are both amongst the largest five regions (5th and 4th, respectively); other regions with many endemic genera (27, Southern Africa [2nd highest] and 29, Western Indian Ocean [4th highest]) are not particularly large (positions 22 and 41, respectively), however, and some very large regions (30, Siberia [2nd largest]; 70, Subarctic America [6th largest]) have almost no endemic genera (7 genera and 1 genus endemic, respectively). The regions which have many endemic genera are all at least partly tropical, while large regions with few endemic genera are all at high latitudes; the relationship between numbers of endemic genera and sizes of regions is obviously influenced greatly by the latitudinal position of those regions.

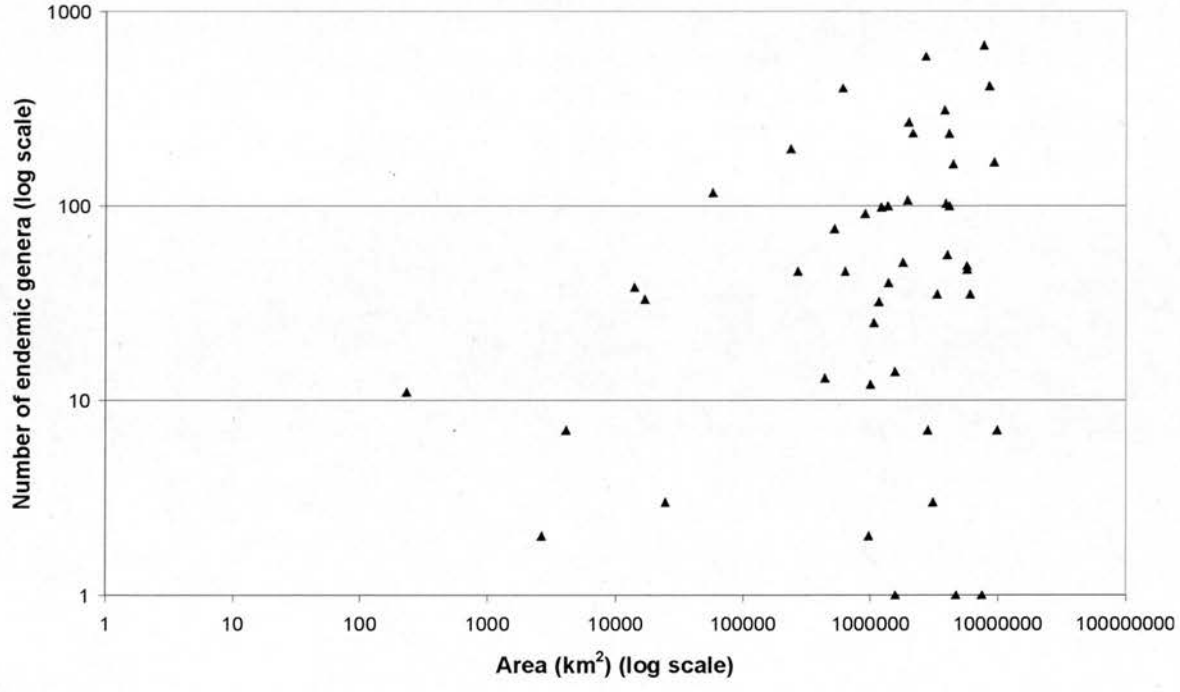


Figure 3.11 The relationship between numbers of endemic genera and area of TDWG regions.

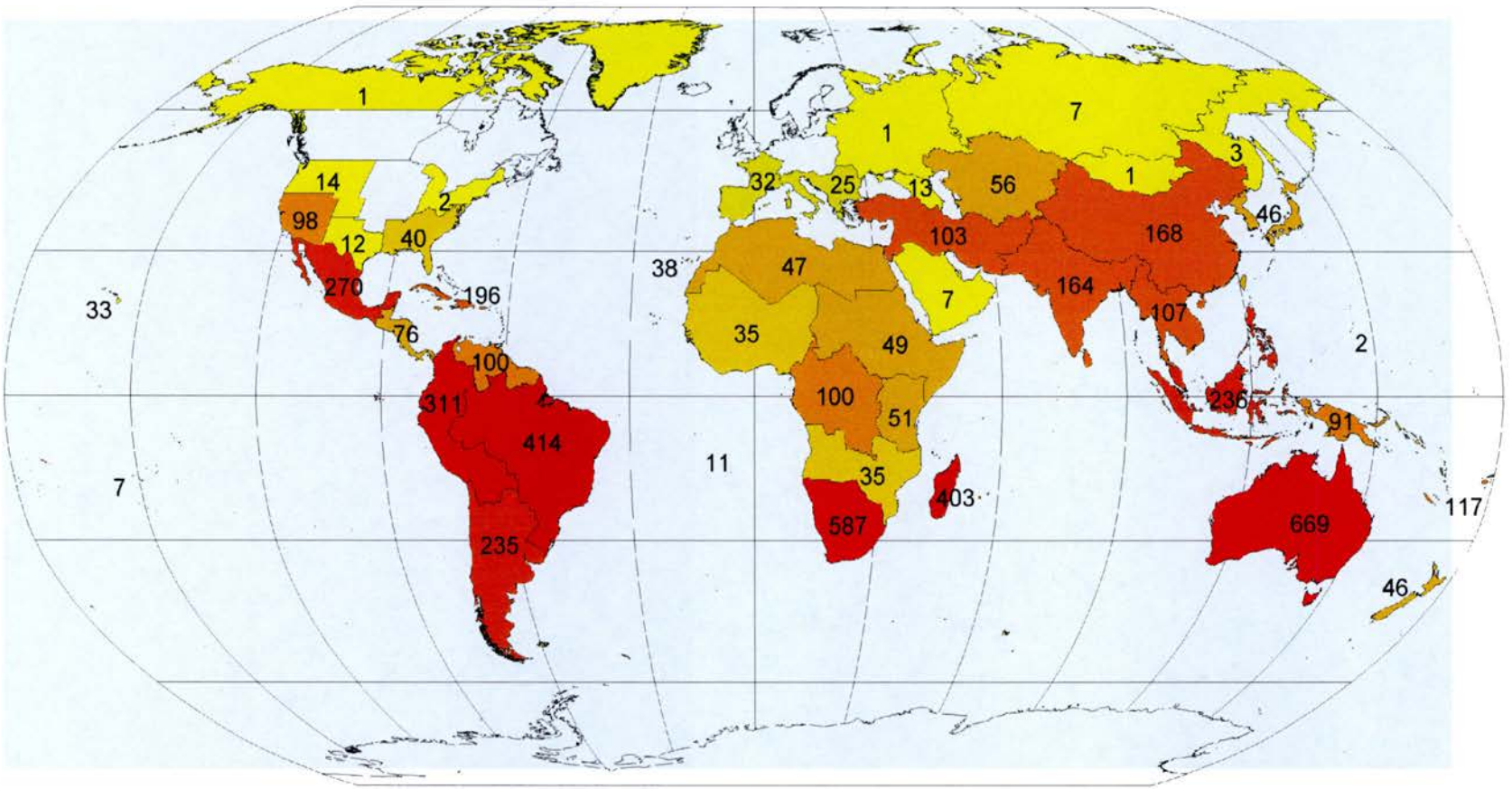


Figure 3.12 Choropleth map of absolute numbers of endemic genera for TDWG Level 2 regions across the world; regions with no endemic genera are without colour (white), graduated colours range from yellow (low diversity) to red (high diversity), and actual numbers of endemic genera are given in boxes for each region (see also Table 3.1).

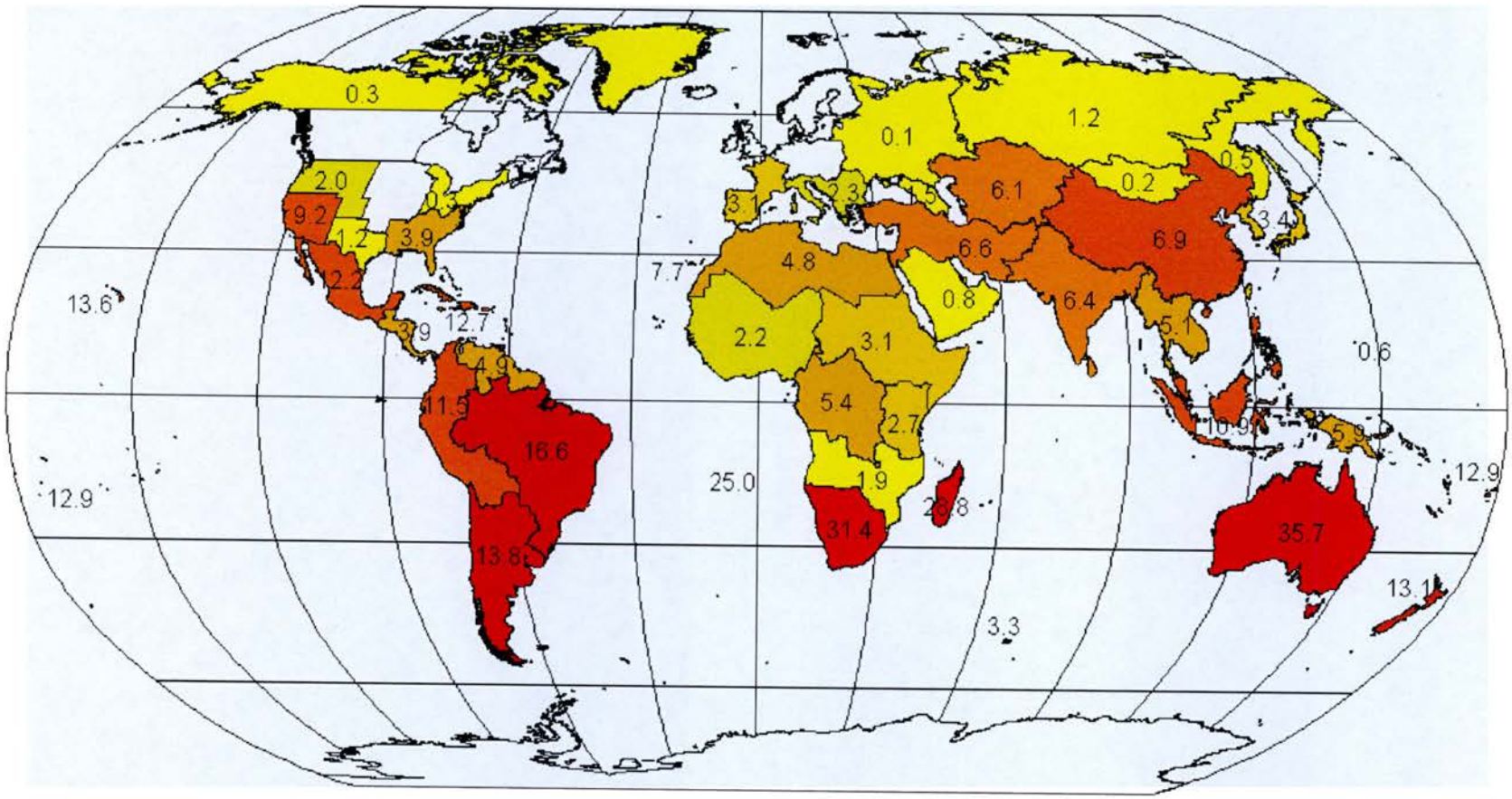


Figure 3.13 Choropleth map of percentages of generic endemism for TDWG Level 2 regions across the world; regions with no endemic genera are without colour (white), graduated colours range from yellow (low diversity) to red (high diversity), and values given for each region are percentage endemism (see also Table 3.1).

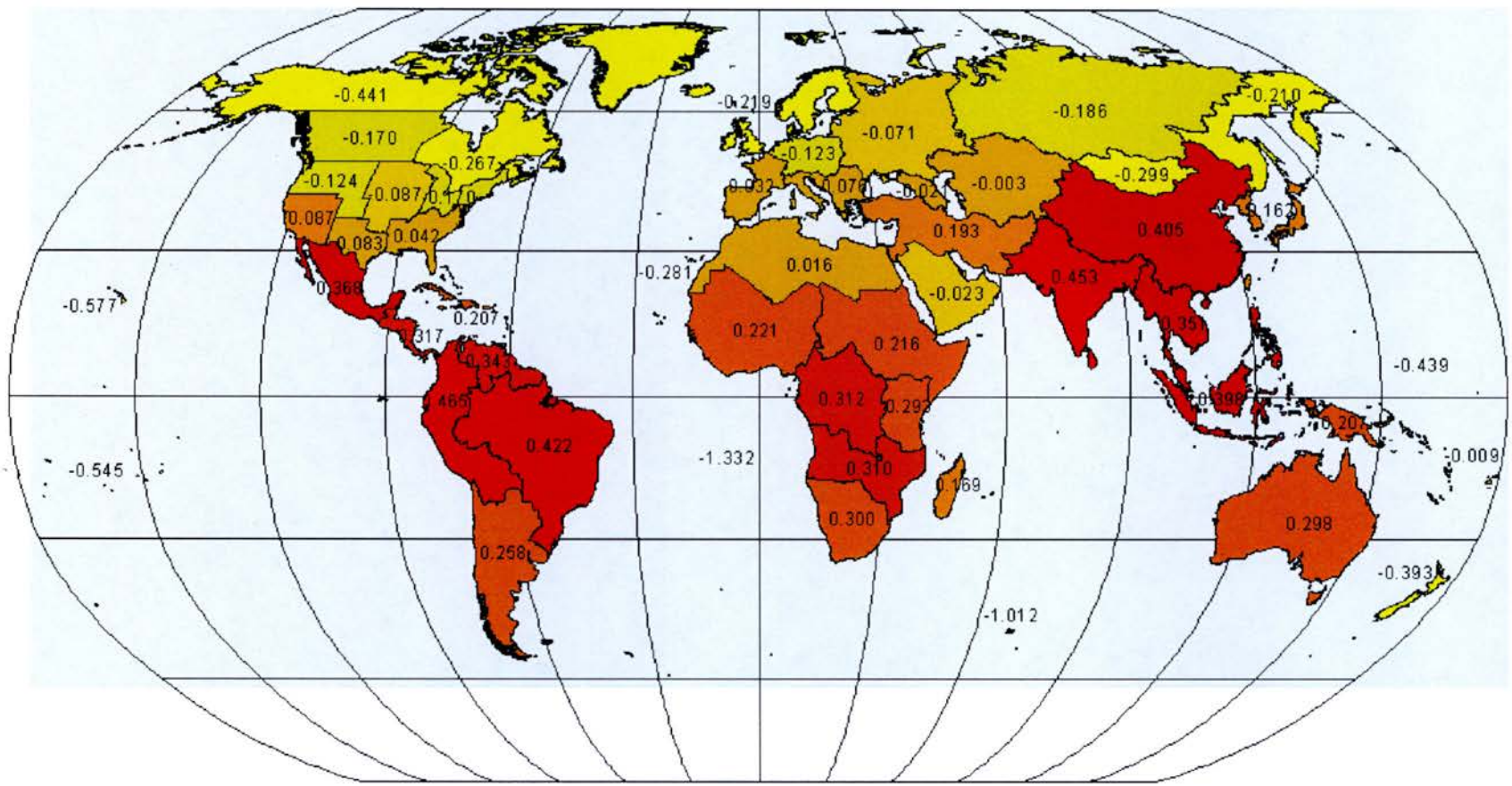
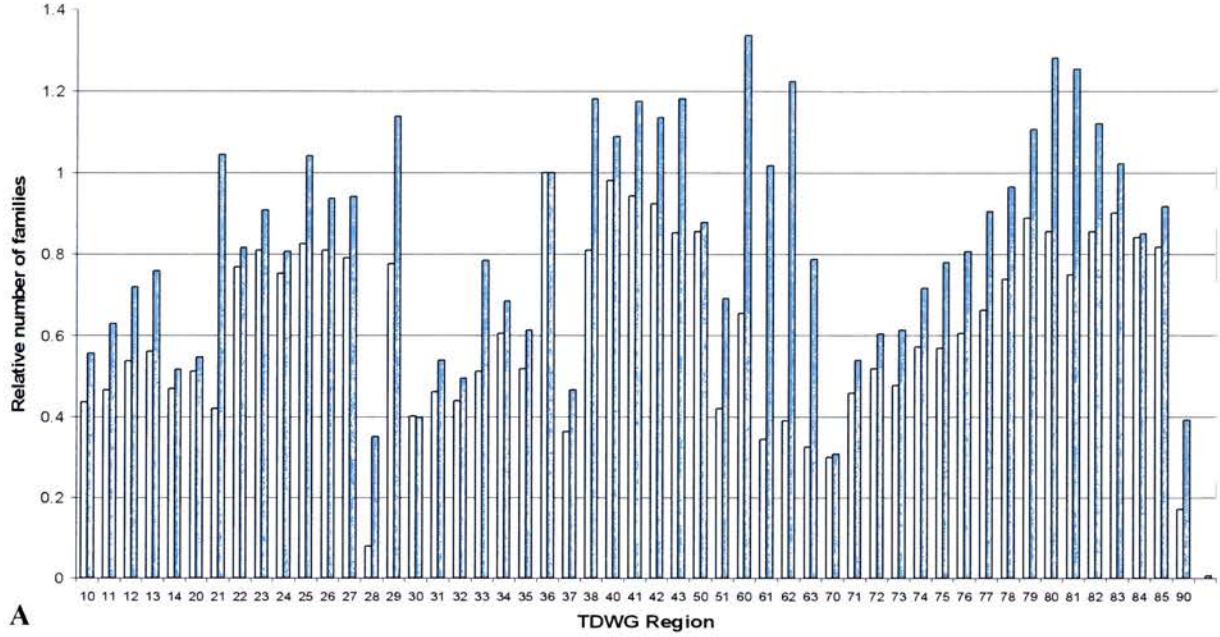
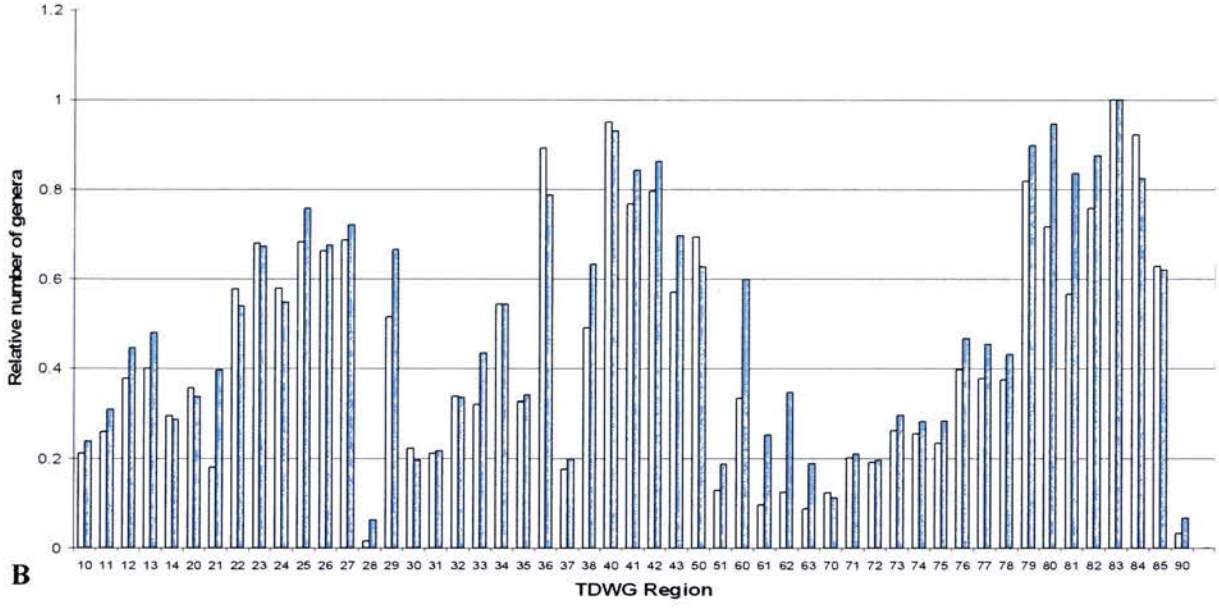


Figure 3.14 Distribution map of the residuals from the regression of log-transformed numbers of genera against area for 51 TDWG regions (excluding Antarctica).

Table 3.2 Degree of generic endemism for TDWG regions.

TDWG Region		Number of genera	Number of endemic genera	% endemism
10	Northern Europe	577	0	0
11	Middle Europe	709	0	0
12	Southwestern Europe	1031	32	3.1
13	Southeastern Europe	1093	25	2.3
14	East Europe	803	1	0.1
20	Northern Africa	971	47	4.8
21	Macaronesia	491	38	7.7
22	West Tropical Africa	1574	35	2.2
23	West-Central Tropical Africa	1856	100	5.4
24	Northeast Tropical Africa	1578	49	3.1
25	East Tropical Africa	1861	51	2.7
26	South Tropical Africa	1806	35	1.9
27	Southern Africa	1864	587	31.4
28	Middle Atlantic Ocean	44	11	25.0
29	Western Indian Ocean	1398	403	28.8
30	Siberia	606	7	1.2
31	Russian Far East	573	3	0.5
32	Central Asia	917	56	6.1
33	Caucasus	873	13	1.5
34	Western Asia	1476	103	6.6
35	Arabian Peninsula	885	7	0.8
36	China	2420	168	6.9
37	Mongolia	478	1	0.2
38	Eastern Asia	1334	46	3.4
40	Indian Subcontinent	2572	164	6.4
41	Indo-China	2085	107	5.1
42	Malesia	2162	236	10.9
43	Papuasia	1535	91	5.9
50	Australia	1874	669	35.7
51	New Zealand	351	46	13.1
60	Southwestern Pacific	909	117	12.9
61	South-Central Pacific	264	7	2.7
62	Northwestern Pacific	339	2	0.6
63	North-Central Pacific	242	33	13.6
70	Subarctic America	335	1	0.3
71	Western Canada	548	0	0
72	Eastern Canada	519	0	0
73	Northwestern U.S.A.	707	14	2.0
74	North-Central U.S.A.	694	0	0
75	Northeastern U.S.A.	637	2	0.3
76	Southwestern U.S.A.	1067	98	9.2
77	South-Central U.S.A.	1020	12	1.2
78	Southeastern U.S.A.	1018	40	3.9
79	Mexico	2212	270	12.2
80	Central America	1933	76	3.9
81	Caribbean	1538	196	12.7
82	Northern South America	2053	100	4.9
83	Western South America	2701	311	11.5
84	Brazil	2489	414	16.6
85	Southern South America	1706	235	13.8
90	Subantarctic Islands	91	3	3.3
91	Antarctic Continent	2	0	0





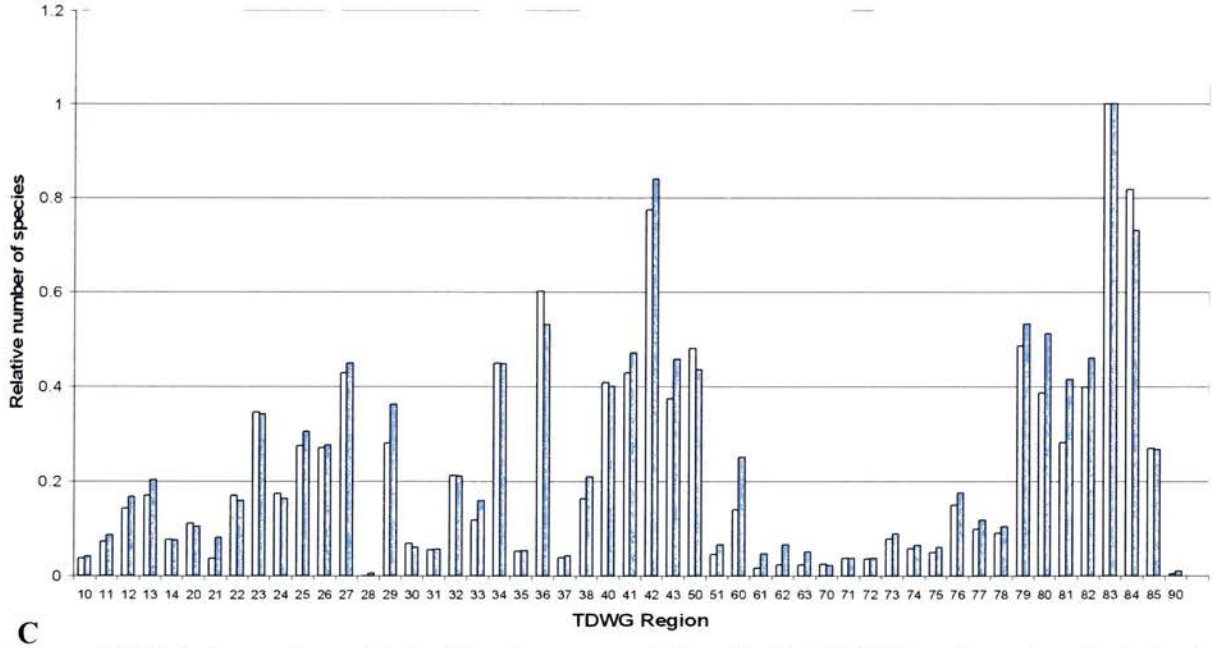


Figure 3.15 Relative numbers of **A** families; **B** genera; and **C** species for TDWG Level 2 regions. Both absolute (light blue) and area-rescaled values (dark blue) are given for each region. Within each taxonomic rank, values are given on a 0 – 1 scale relative to that region with the highest absolute number of taxa. See text for details.

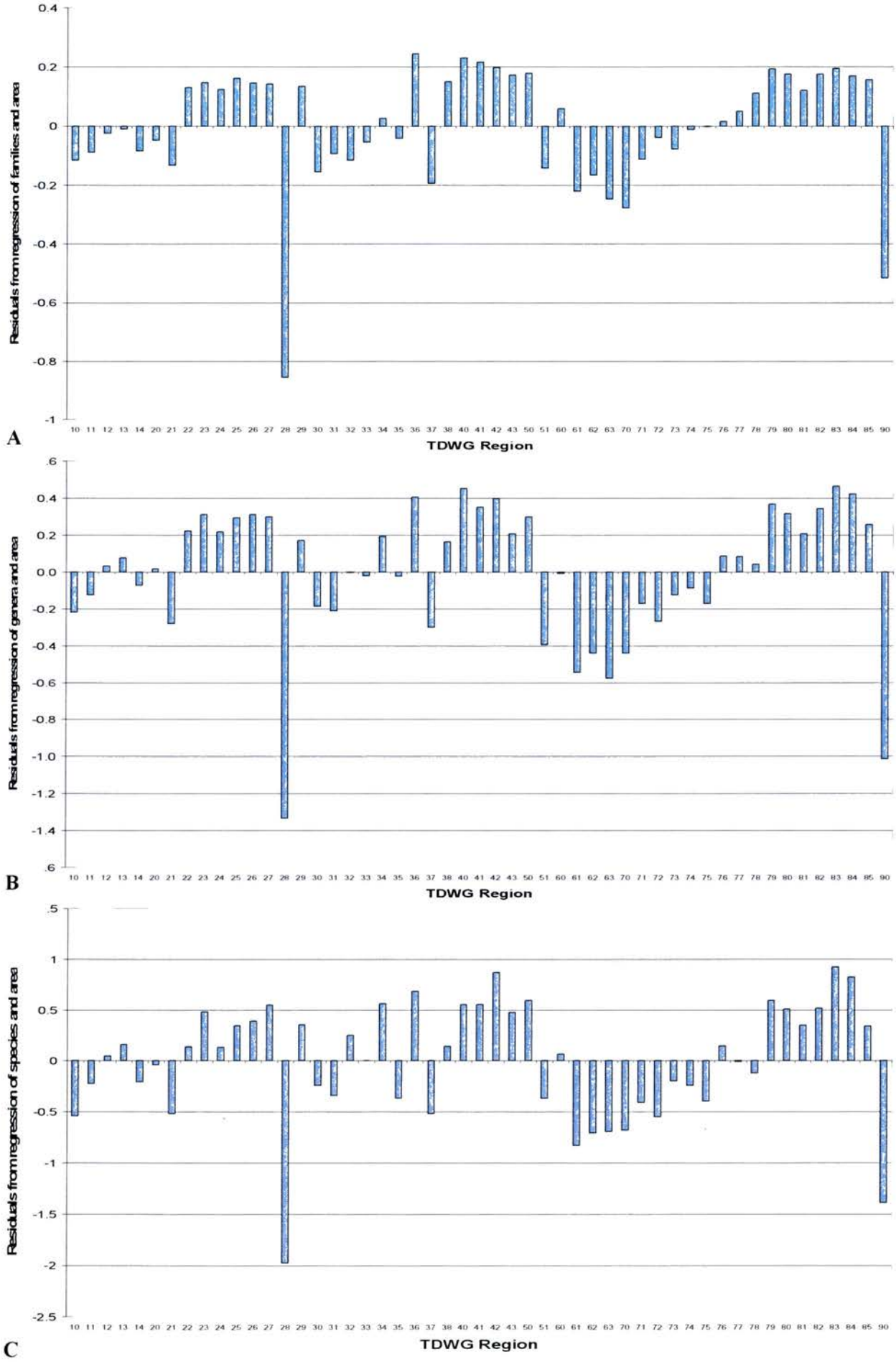


Figure 3.16 Residuals from the regression of log-transformed diversity of **A** families; **B** genera; and **C** species for TDWG Level 2 regions against area of regions.

3.5 Global patterns of angiosperm richness – gamma diversity

Gamma (landscape or regional) diversity is a concept which has been applied to areas of many different sizes; it is not clear whether or not the term should be restricted to analyses of any one particular scale. In particular, it is not clear how to delimit a landscape or region in a non-arbitrary way, except as that region across which a certain study was carried out. However, since it was explicitly coined for large areas encompassing a variety of habitats (Whittaker, 1960, 1972), counts of taxon richness for arbitrary, geo-political TDWG Regions will here be referred to as the gamma diversity of that region. Scores of gamma diversity can thus be compared across different regions of the world at successive taxonomic levels.

3.5.1 Rescaling absolute richness by the species-area relationship

In order to compare gamma diversity values for regions which have different sizes, however, it is first necessary to produce relative diversity scores that factor out the confounding effects of area. Simply dividing species number by area to give unscaled species-area ratios leads to highly spurious results (Connor & McCoy, 1979) – they have the effect of over-estimating the diversity of small areas. Implicit in this approach is an assumption of a simple linear relationship between area and diversity, whereas ecologists have long known that such a linear relationship does not exist (Arrhenius, 1921; Williams, 1943; MacArthur & Wilson, 1967). The true relationship is a power function, as shown by the equation $S = cA^z$. If:

$S = cA^z$, then $c = S/A^z$

and thus the constant c is the ratio of diversity (S) to A^z (Rosenzweig,1995).

So in order to obtain realistic scores of relative diversity, area needs to be scaled by a suitable exponent value (z), and relative values for c are then calculated with S/A^z, and this can be used to give relative c values for areas of different size. A crucial question is therefore how to choose an appropriate value of the exponent z. It is not appropriate here to simply use the z value shown by the regression of each taxonomic level against area shown in Figure 3.7, since the slope of the regression varies with geographical scale (species-area curves across different biogeographical provinces are much steeper than are those across smaller areas within a particular biogeographic province; MacArthur & Wilson, 1967; Rosenzweig, 1995). Figure 3.7 shows the relationship of diversity to area between different regions, rather than within each individual region. The slopes of the regressions shown in Figure 3.7 are therefore steeper than for the species-area relationship within each of the regions. What is ideally needed is a set of detailed species-area studies, at least one within each TDWG region. The absolute diversity of each TDWG region could then be re-scaled by an exponent value known to be appropriate for that region. However, such detailed studies have simply not been performed across each region of the world, so this was not a feasible solution.

Choice of an appropriate z value was therefore determined from the work of MacArthur and Wilson (1967), who list typical values of z from species-area studies in different regions. The majority of the 52 TDWG Level 2 regions conform to what MacArthur and Wilson (1967) call 'non-isolated mainland continental regions',, which have typical z values given as between 0.12 – 0.17. An intermediate value of 0.14 from this range was therefore chosen. In practice, however, although rescaling by a suitable exponent value or not does greatly affect relative diversity scores, these are not greatly affected by the precise value of that exponent (Brummitt & Nic Lughadha, 2003). For the data presented in Table 2.1, relative diversity at each rank calculated with either $z = 0.14$ or $z = 0.25$ (a value in middle of the range of species-area studies over island archipelagos; MacArthur & Wilson, 1967) are highly correlated (Spearman's r_s between relative diversities: for families, 0.89; for genera, 0.93; for species, 0.98; $n = 52$, $p < 0.0001$ in each case). Another way to compare relative diversities without the effects of area would be to examine the relative sizes of the residual variation for each region: regions with large positive residuals are particularly diverse; regions with large negative residuals are particularly depauperate. To corroborate the re-scaling method used here, residuals from the linear regression of log-transformed diversity scores (families, genera and species) against log-transformed area were also calculated and compared with results from re-scaling diversity scores by $z = 0.14$. A map of the distribution of the residuals from the linear regression of the log-transformed variables is given in Figure 3.14.

Both absolute and area-rescaled diversity at each taxonomic rank for $c = S/A^z$ where $z = 0.14$ are given in Figure 3.15. In Figure 3.16 the residuals from the regression of the log-transformed variables are plotted against absolute diversity at each taxonomic rank. In order to make absolute and re-scaled data comparable within each taxonomic rank, both sets of data at each rank have been normalised by dividing values by the largest value for that rank (i.e. absolute and rescaled numbers of families for each region were both divided by the respective absolute and rescaled values for the region with greatest absolute number of families which was 36, China; for both genera and species, where the region with the greatest absolute and rescaled numbers of taxa was the same both times, values for each region were divided by respective values for 83, Western South America), in both sets of graphs. Thus the absolute and relative scores are on comparable scales within taxonomic ranks, but the scales are not equivalent between taxonomic ranks (although relative positions of regions will be). Genus- and species-level values are therefore given on a scale of 0 – 1 relative to that of Western South America, which has a maximum relative value of 1.0 for both absolute and area-rescaled data; for family-level values, because the highest absolute value and the highest area-rescaled value are from different regions, after normalising both absolute and rescaled values by those for China (the region with the greatest absolute number of families) rescaled values of some regions are greater than 1.0 in Figure 3.15A.

Several salient points can be inferred from Figure 3.15. Firstly, for each taxonomic rank, three areas of tropical diversity can be clearly made out: Africa (+ Madagascar), SE. Asia and the

Neotropics, increasing in that order. However, the proportion of taxa in the tropics also increases with decreasing taxonomic ranks, as was also evident from the respective slope values for each taxonomic rank from Figure 3.7. The region with the highest absolute number of families, Region 36, China, is not especially rich when diversity is rescaled by area; in fact only 17th out of 52. The high absolute number of families in China is therefore a product of its great size; amongst area-rescaled values it is actually the SW. Pacific which truly has the greatest number of families, as also shown by Williams *et al.* (1994). In second and third place for rescaled numbers of families are two regions of the Neotropics; 80, Central America and 81, Caribbean. The region with the greatest number of both genera and species, both for absolute and for area-rescaled values, is also Neotropical: Region 83, Western South America. With the area-rescaled values, this is followed at genus level by Region 80, Central America, and at species level by Region 42, Malesia, followed in turn by Region 84, Brazil.

Region 40, Indian Subcontinent, has the third highest relative genus-level diversity, a surprisingly high position when its relative species-level diversity is not high. At genus level, regions from tropical Asia are almost comparable with Neotropical regions in their area-rescaled values, but at species level only Region 42, Malesia, can really approach the diversity of Western South America. In this analysis, three regions of the world really stand out as having exceptionally high species-level diversity: Region 42, Malesia; Region 83, Western South America; and Region 84, Brazil. When absolute species diversity estimates are rescaled by relative area, most regions in both tropical Asia and the Neotropics show an increase relative to absolute values. Three regions with very large absolute species numbers show a decrease after rescaling relative to that of Region 83, Western South America, however. These are the huge countries of China (Region 36), Australia (Region 50) and Brazil (Region 84), where the large areas appear to inflate the absolute numbers of species but rescaling diversity by the species-area relationship reveals their true relative diversities.

At each taxonomic rank, Africa is consistently less diverse than is either tropical Asia or the Neotropics. Within Africa furthermore, Regions 22 (West Tropical Africa) & 24 (Northeast Tropical Africa) have considerably lower diversity than do other tropical regions when rescaled by area; in fact, for area-rescaled relative species diversity, even below that of southern Europe (Regions 12, Southwestern Europe, & 13, Southeastern Europe) and SW. Asia (Regions 32, Middle Asia and 33, Caucasus, and much lower than Region 34, Western Asia) and comparable with SW. U.S.A. (Regions 76, Southwestern U.S.A.). This surprising result may be partly explained by the almost-barren Sahara Desert covering large expanses of Regions 22 and 24, giving lower diversities than would be expected for regions of that size. Also, however, differences in relative diversities between regions are exaggerated somewhat by being normalised by Region 83, Western South America, which has an absolute number of species more than twice as large as other tropical regions except for Region 42, Malesia and Region 84, Brazil.

Plotting the residuals from the regression of the log-transformed variables (see Figure 3.16) gives a similar picture at lower taxonomic levels as does re-scaling absolute diversity scores by a *z* value of 0.14. For both methods, Region 83, Western South America, still emerges as most diverse for both numbers of genera and of species, after area has been accounted for. At family level, however, Region 36 (China) appears as the region with the greatest family diversity (largest positive residual value), in contrast to the area-rescaled graph (Figure 3.15A) and to Williams *et al.* (1994) where the SW. Pacific (Region 60) was the richest. Note that with the residual method, however, the largest residuals are not necessarily for those regions which appear to be furthest from the regression line in Figure 3.7, since the axis is logarithmically scaled. For example, in Figure 3.8 it appears to be Region 80, Central America, which has the largest positive residual, whereas in fact the residual for Region 83, Western South America, is larger but this is obscured by this region being higher on the y axis (i.e. Region 83 has greater absolute diversity than does Region 80).

3.6 Beta diversity across regions

3.6.1 Measuring beta diversity

Beta diversity (Whittaker, 1960, 1972) is a measure of the change in taxon composition across spatial gradients. A number of slightly different measures have been proposed by different authors (Wilson & Shmida, 1984), none of which seems to hold any logical priority over the others (Blackburn & Gaston, 1996; Koleff & Gaston, 2001). These measures were first developed explicitly to study change along directional gradients, such as increasing altitude or decreasing rainfall, for example. Also implicit in the original concept of beta diversity was having to sample several local patches of vegetation within a larger area along that particular gradient, so that beta diversity was a measure of the change in species composition between these sampled patches, originally defined (Whittaker, 1960) as total regional (gamma) diversity divided by average local (alpha) diversity (Harrison, 1997). Beta diversity has also been measured between only two adjacent or contiguous areas, for example either within (Blackburn & Gaston, 1996) or between (Koleff & Gaston, 2001) latitudinal bands. Under the more specific term 'beta turnover' (Wilson & Shmida, 1984), change in taxon composition is restricted to the direct study of gradients; however, simply as the ratio between the total number of taxa and the average number of taxa it can also be used as a measure of overall heterogeneity in a data matrix (McCune & Grace, 2002). More recently, other non-directional means of measuring beta-diversity have been proposed based purely on distance, which measure the decay in overall similarity (measured with a similarity coefficient) between areas with increasing distance between them (Nekola & White, 1999; Condit *et al.*, 2002).

A choice therefore needs to be made of which way and with which measure to assess levels of beta diversity. A useful first approximation is to simply use Whittaker's (1960) original formulation

(β_w), or Williams' (1996) alternative formulation (β_j) of this, as a general, non-directional measure of overall heterogeneity within the data (McCune & Grace, 2002); this was done at family, genus and species levels (see Table 3.3). Although this does not account for differences in sizes between regions, the fact that values at each taxonomic level represent exactly the same regions means that values of beta diversity will nevertheless be comparable between the different taxonomic levels. Whittaker's original (1960) formula for beta diversity was given as

$$\beta_w = (S/\alpha) - 1$$

where S is the overall number of species recorded in the whole region and α is the average number of species found throughout that region, whereas Williams' (1996) alternative was

$$\beta_j = 1 - (\alpha_{max}/S)$$

where α_{max} is the maximum number of species in any unique area within the whole region. Values for both these formulations, for all the 454 families, 14304 genera, and 370,000 species are presented in Table 3.3 below. Figures for the numbers of families and genera are from the Vascular Plant Families and Genera database, while the estimate of the total number of species was kindly provided by R. Govaerts:

Taxonomic level	$\beta_w = (S/\alpha) - 1$	$\beta_j = 1 - (\alpha_{max}/S)$
Family	1.81	0.42
Genus	11.23	0.81
Species	35.93	0.88

Table 3.3 Beta diversity values for numbers of families, genera and species in 52 TDWG Level-2 regions around the world calculated with two different beta diversity indices, Whittaker's (β_w; Whittaker, 1960) and Williams' (β_j; Williams, 1996).

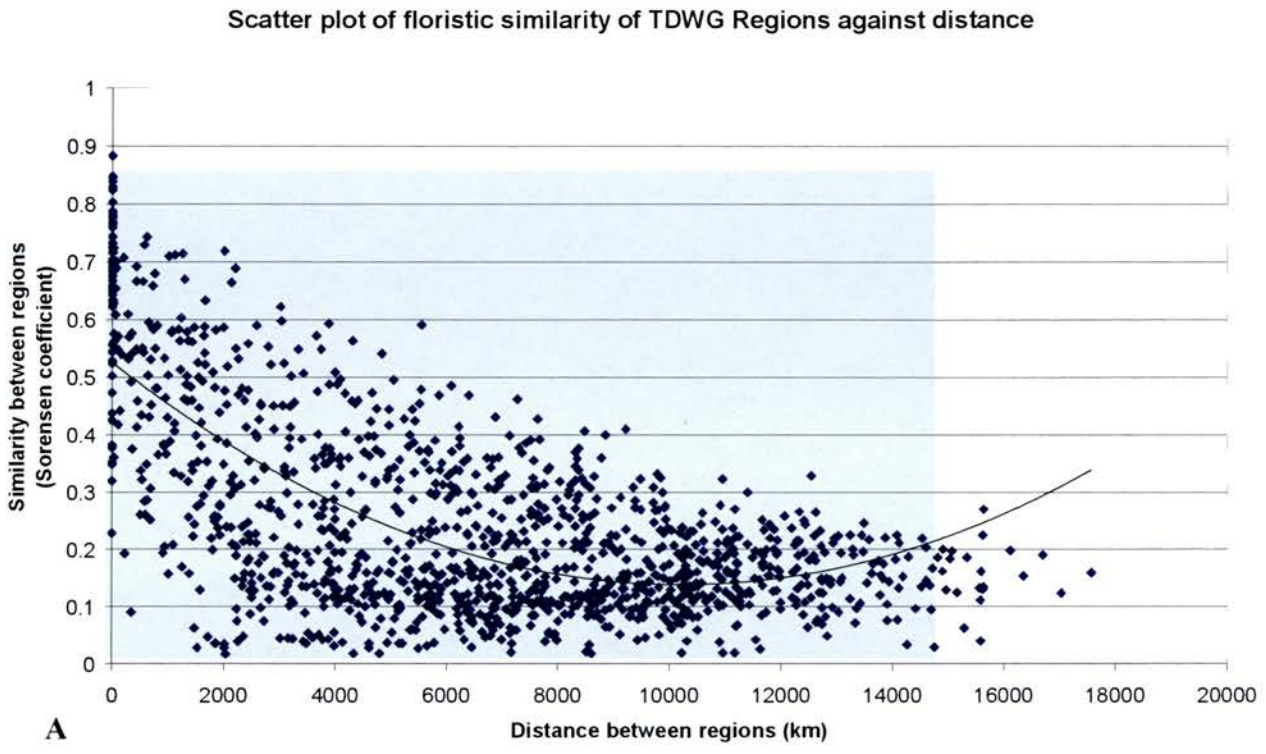
McCune & Grace (2002) warn that the larger the value of β_w, then the more difficult multivariate analyses, and particularly ordination, become. They list 'rule of thumb' values for β_w of below 1 as 'low' and above 5 as 'high', which would imply problems ahead with multivariate analysis (see Chapters 5 and 6). However, they also point out that the maximum value of β_w is achieved when there is no taxon in common between any of the areas, and so will be only one less than the total number of taxa in the study. In the context of such large datasets which each have so many taxa, therefore, values of β_w for genus- and species-level data are actually rather low. Partly to avoid this problem of β_w not having any uniform maximum value and thus not being comparable between different datasets, β_j was suggested by Williams (1996) as an index with clear minimum and maximum values, from 0 to 1 (where 0 implies no turnover, and 1 implies no taxa in common). With these data, family distributions therefore show a moderate level of beta diversity, whereas distributions of genera and species both show very high levels of beta diversity. Since beta diversity measures compositional change in taxa between areas, high levels of beta diversity imply high levels of endemism, and vice versa. However, this might be partly due to the large size of most of the regions used in this study, which means that the degree of endemism of both genera and species, and

hence beta diversity, will be high. The large variation in the sizes of the total set of regions has opposite effects on the two measures of beta diversity used here: the bigger the average value (the greater the proportion of large areas), the smaller the value of β_w; in contrast, high maximum values relative to the total number of taxa (a product of large areas) lead to low beta diversity values for β_j.

3.6.2 Distance decay of floristic similarity between regions

The regions used in this study are all defined by irregular geo-political boundaries, or consist of isolated islands (see Figure 1.1), and a single irregularly-shaped area can have contiguous boundaries with several others. It is therefore not obvious in which direction beta diversity should be measured, unless there is some pre-determined direction which is of particular biological interest. Such a direction might be towards the equator, and latitudinal gradients in diversity are explored in more detail in the next chapter. It would be possible instead to measure beta diversity across all contiguous boundaries, but without having a unifying direction of interest, such as latitude, this would not be very satisfactory. Also, there is a large number of possible pairs of contiguous boundaries and calculating beta diversity for each of these would approximate a distance matrix that is more conventionally studied with similarity coefficients. Instead, therefore, the more general representation of beta diversity as the distance decay in floristic similarity between all possible pairs of regions (Nekola & White, 1999; Condit *et al.*, 2002) was studied here with genus-level data for the world.

Floristic similarity was assessed using both Jaccard's and Sørensen's similarity indices, and as well as using minimum distance between pairs of regions, the difference in their respective minimum latitudinal distance from the equator was also calculated (i.e. two regions the same distance from the equator have a small latitudinal difference, whereas if one region is close to the equator while the other is far from the equator the two regions show a large latitudinal difference). Difference in respective latitudinal distance between regions was used in preference to purely latitudinal distance as this latter measure itself approximates true distances (and if measured between two regions in opposite northern and southern hemispheres it is identical to true distances) and thus gives almost identical results. Graphs of these relationships using Sørensen's similarity index are presented in Figure 3.17; because of the overwhelming similarity between the two sets of graphs, those for Jaccard's coefficient are not shown. Trendlines for both graphs are 2nd order polynomial regressions. For all combinations of floristic similarity (Jaccard's or Sørensen's index) and distance (minimum distance and difference in latitudinal distance from the equator), similarity in generic composition declined with increasing distance between regions. All possible pairs of regions seem to fall within a constraint envelope whose upper limit is bounded by a maximum rate of decline in similarity of 0.1 per 2300km, or per 7° difference in latitudinal distance between regions. A linear relationship was found between floristic similarity and latitudinal difference (Figure 3.17A), whereas with simple distance, similarity seemingly increases slightly at very long distances (Figure 3.17B). The upturn visible in Figure 3.17B may appear more pronounced than is perhaps warranted by the data, due to the regression used here.



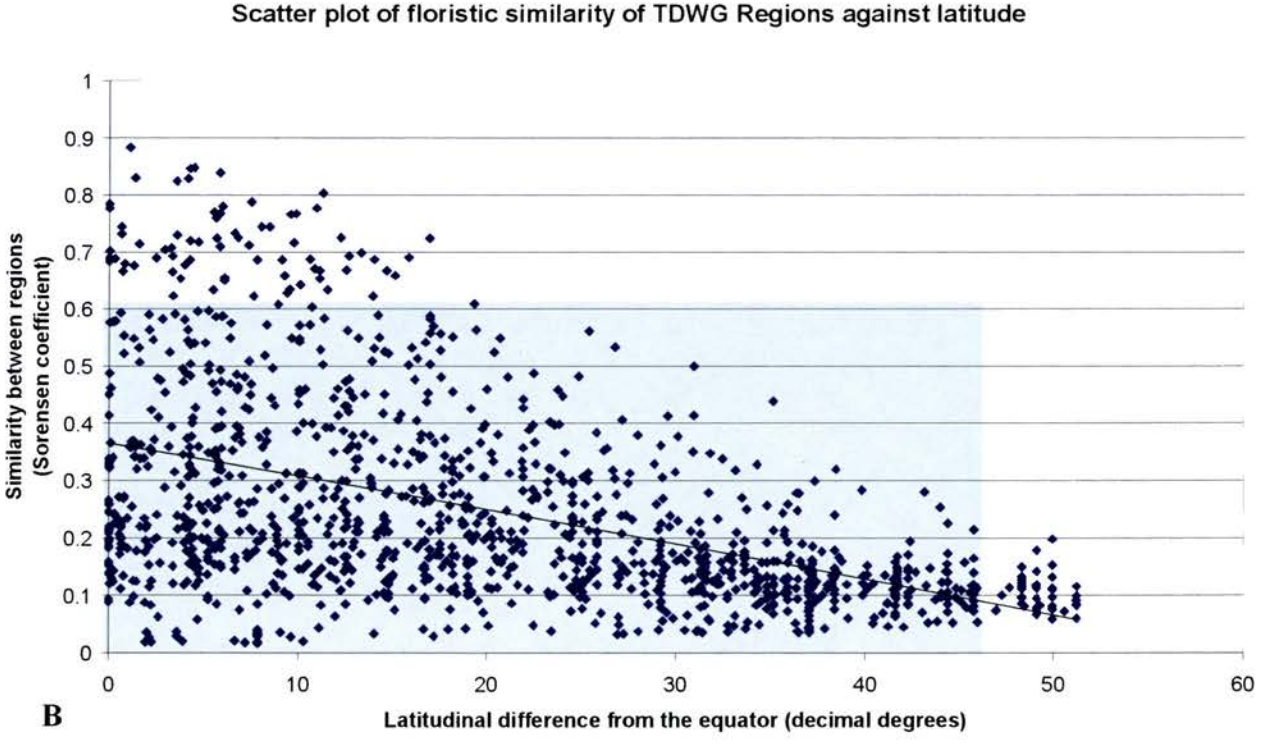


Figure 3.17 Generic turnover between regions can be expressed by a measure of floristic similarity. Graphs show the relationship between floristic similarity (Sorensen coefficient), for each pair of TDWG regions, and: **A** physical distance between regions (km); and **B** the difference between regions in the respective latitudinal distance to the equator (decimal degrees). The trendline in both cases is a 2nd order polynomial regression. Two regions which are a given distance apart at the same latitude have greater floristic similarity than do two regions the same distance apart but at different latitudes

3.7 Hotspots

Norman Myers and colleagues at Conservation International have recently published a huge compendium of information on the distribution of biodiversity, in collaboration with over 100 scientists world-wide, which identified 25 regional 'biodiversity hotspots' for urgent conservation action (Mittermeier *et al.*, 1999; Myers *et al.*, 2000). Their criteria for inclusion as a hotspot are twofold: an area must contain at least 1,500 endemic vascular plant species (0.5% of an estimated global total of 300,000); and remaining primary vegetation should be no more than 30% of its original extent. Data on vascular plant and (non-fish) vertebrate diversity are clearly presented and subjected to a series of simple analyses, ranking hotspots first by both absolute numbers of species and of endemic species for each taxonomic group (birds, mammals, reptiles, amphibians, vascular plants) regardless of the size of hotspots, and also by relative species numbers assessed in relation to both the original extent of natural vegetation and to the proportion of natural vegetation remaining intact. Perhaps recognising that a single measure of biodiversity, capturing all its different facets, remains elusive, this information is presented in some 46 separate tables. Myers *et al.* (2000), however, took this a stage further, and combined several of these aspects into a single list of 'hottest hotspots' (Table 6, p. 857, Myers *et al.*, 2000).

The chief result of all these analysis is how irregularly distributed the world's biological diversity is: Myers *et al.* (2000) estimate that 44% of vascular plant diversity is contained within only 1.4 % of the world's land surface (the cumulative total for the remaining intact vegetation of 25 hotspots). Despite the intuitive appeal of the concept, however, the selection of hotspots in general has been criticised on several grounds. (i) Reliable quantitative data are generally only available for the most conspicuous and popular groups of organisms (vascular plants, vertebrates), which are by no means the most speciose (Margules *et al.*, 1994) – and it is generally assumed rather than proven that areas of diversity for one group will be concordant with areas of diversity of unsampled groups (Prendergast *et al.*, 1993). (ii) Without a measure of complementarity between hotspots there is no way of knowing how many species are 'conserved twice' in adjacent hotspots (Margules & Pressey, 2000). (iii) Simply conserving maximum species numbers is not the same as conserving maximum species diversity, since distantly related taxa are 'worth more' in terms of phylogenetic diversity than are numerous closely related species (Vane-Wright *et al.*, 1991; Williams & Humphries, 1994). (iv) The huge size of some hotspots makes effective conservation action impractical, since it must involve co-ordination through many national governments – designation of such areas as the Mediterranean Basin (2,362,000 km^2) or Indo-Burma (2,060,000 km^2) as biodiversity hotspots can hardly be said to represent 'tight targeting of conservation efforts.' Although no-one is taking issue with the assertion that small areas of the world are exceptionally rich biologically, all of the above criticisms may be levelled at the work of Mittermeier *et al.* (1999) and Myers *et al.* (2000) (Mace *et al.*, 2000; Humphries, 2001).

Another weakness of the study is that this is essentially a single-scale analysis with no study of how biodiversity values change with the available area of habitat. For example, the diversity contained within a hotspot may be just as irregularly distributed, with small areas of land accounting for a large proportion of diversity. The actual delimitation of these areas is not discussed and they do not actually list their data sources for numbers of species or areas of remaining habitat. In contrast to the TDWG geo-political regions, their analysis has been focused on whole biogeographic regions (although they do not come out and actually say as much) but their hotspot areas are comparable in scale to the TDWG regions and make an interesting comparison with the current study. Unfortunately, however, in addition to the above criticisms, the hotspot prioritisation analysis in both Mittermeier *et al.* (1999) and in Myers *et al.* (2000) contains a commonplace but fundamental error. To compare regions of different sizes it is essential to produce relative diversity scores which factor out the effects of area. To do this, Mittermeier *et al.* (1999) and Myers *et al.* (2000) simply divided species richness and endemism values by area to produce values of species per unit area (100 km^2) (see Tables 21 – 44, Mittermeier *et al.*, 1999, and Table 4, Myers *et al*, 2000).

Although it is true that, assuming the whole hotspot can be conserved, an unscaled species-area ratio can provide a comparative measure of the cost per species of conserving each hotspot, unscaled species-area ratios do not accurately reflect true relative diversities (Connor & McCoy, 1979) – they have the effect of over-estimating the relative diversity of small areas (see Section 3.5). Tables 45 a-d and 46 a-d in Mittermeier *et al.* (1999, pp. 54-55) present the richest hotspots ranked by species diversity and species endemism, both separately and combined, per unit area (100 km^2) of both original extent of natural vegetation and remaining intact primary vegetation. The top five hotspots in each of these eight tables are all amongst the ten smallest when ranked either by original extent of native vegetation or by remaining intact natural vegetation (Mittermeier *et al.*, 1999, Table 2, p. 33). The Eastern Arc Mountains and Coastal Forests of Kenya and Tanzania appears as the most diverse hotspot in all of these eight tables except Table 45d, where it lies third. But it is also smallest of all the hotspots ranked by remaining intact natural vegetation and second smallest by original extent of native vegetation (Mittermeier *et al.*, 1999, Table 2, p. 33). Species-area ratios per unit area for both plant diversity and endemism and total vertebrate diversity and endemism are all significantly negatively correlated (Spearman's r_s -0.62 – -0.82; $p < 0.01$) with both original extent of natural vegetation and remaining intact natural vegetation; i.e. smaller areas always appear relatively more diverse.

The raw data presented by Mittermeier et al. (1999) was re-analysed, re-scaling species-per-unit-area values by the power law species-area relationship (Section 3.5), to investigate how their rank-based prioritisation analysis of hotspots might be affected (Brummitt & Nic Lughadha, 2003). When the data is re-analysed, areas of the Neotropics then emerge as most diverse in terms of total diversity and/or total endemism, in separate, combined, and combined and weighted analyses, whether based on original or current extent of natural vegetation (see Table 3.4 in this thesis). The 'hottest

hotspot' in each analysis is either the Tropical Andes or Mesoamerica. The Tropical Andes are placed first (or joint first) in six of the eight analyses specified by Mittermeier *et al.*, and in second place in the remaining two analyses. Mesoamerica is placed first (or joint first) in three of the eight analyses and is second to the Tropical Andes in a further three analyses. The Neotropics consistently emerge as the most bio-diverse areas of the world, while the enormous apparent diversity of the Eastern Arc Mountains and Coastal Forests is greatly over-exaggerated. In contrast, the Tropical Andes never ranks above eighth highest in the analysis of Mittermeier *et al.* and does not even feature in their table of fifteen hotspots ranked by combined equally-weighted species diversity and endemism per unit area of remaining intact natural vegetation, where in fact it ranks highest in our re-analysis.

In addition to this large-scale study, Davis *et al.* (1994-1997) have compiled an extensive global analysis of 203 smaller-scale 'Centres of Plant Diversity' ranging in size from 53 km^2 to 1100000 km^2 (although 89% of these areas are smaller than 100000 km^2). Recalculations accounting for the species-area relationship give results comparable to those obtained in other comparative studies. Calculating relative richness in the same way for over 200 centres of plant diversity around the world, from data in Davis *et al.* (1994-1997), also shows three neotropical areas – La Amistad (Costa Rica/Panama), the Upper Rio Negro Region (Brazil/Colombia/Venezuela) and Braulio Carillo-La Selva (Costa Rica) – to be considerably more diverse than any others. Furthermore, Barthlott *et al.* (1996) listed six areas where vascular plant diversity is estimated to exceed 5,000 species per 10000 sq. km. and designated them global diversity maxima on the assumption that plant maxima correspond to maxima of total biodiversity. Although this assumption is questionable, it is notable that of these six areas the top three are all neotropical: Chocó – Costa Rica Centre; Tropical Eastern Andes Centre; Atlantic Brazil Centre.

Also, Clarke *et al.* (2000) find that the Eastern Arc and the Coastal rain forests of East Africa, treated separately, are each considerably less diverse than is the Guineo-Congolian rain forest, when scaled using a power function relationship. By using simple species-area ratios, the enormous apparent diversity of the Eastern Arc Mountains and Coastal Forests has been greatly exaggerated. Finally, in a global analysis of species richness data from wet tropical forests, based on standard-sized 0.1 ha plots, Gentry (1988) reported five plots with at least 240 tree species exceeding 2.5 cm diameter at breast height. The richest two plots were in Colombia, the third and fifth in Peru; only one of the top five plots was from outside the Neotropics – Semengoh Forest in Sarawak (Borneo). At all geographical scales of analysis, therefore, and at both genus-level and species-level, using both data presented here, published data re-analysed here, and comparisons with previously-published studies, a consistent picture emerges: various parts of the Neotropics, in particular southern Central America and especially the eastern slopes of the Andes, are the most biodiverse areas of the planet.

Table 3.4 Unscaled and area-rescaled ranking totals for 25 biodiversity hotspots.

Hotspot	Original Extent of Natural Vegetation								Remaining Intact Natural Vegetation							
	Total Diversity[1]		Total Endemism[2]		Combined Diversity & Endemism[3]		Equally-Weighted Combined Diversity & Endemism[4]		Total Diversity[1]		Total Endemism[2]		Combined Diversity & Endemism[3]		Equally-Weighted Combined Diversity & Endemism[4]	
Tropical Andes	**46**	10	**36**	12	**82**	22	**40**	7	**44**	3	**32**	5	**76**	8	**40**	
Mesoamerica	**46**	5	**35**	8	**81**	13	**27**		**43**		**34**	1	**77**	1	**25**	
Sundaland	**27**		**29**		**56**		**28**		**29**		**28**	5	**57**	5	**28**	1
Caribbean	**24**	29	**25**	29	**49**	58	**25**	23	**25**	34	**26**	34	**51**	68	**26**	27
Madagascar and Indian Ocean Islands	**13**	3	**31**	16	**44**	19	**18**	5	**12**	8	**31**	22	**43**	30	**17**	8
Atlantic Forest Region	**15**	1	**18**	1	**33**	2	**12**		**16**	6	**19**	11	**35**	17	**13**	3
Indo-Burma	**23**		**11**		**34**		**14**		**31**	4	**15**	2	**46**	6	**17**	1
Philippines	**3**	11	**20**	27	**23**	38	**8**	12	**3**	31	**22**	33	**25**	64	**10**	23
Chocó-Darién-Western Ecuador	**19**	28	**12**	17	**31**	45	**9**	13	**17**	11	**10**	8	**27**	19	**4**	5
Eastern Arc Mountains and Coastal Forests	**14**	48		44	**14**	92	**4**	33	**14**	50		49	**14**	99	**6**	40
Western Ghats and Sri Lanka	7	26	7	21	**14**	47	**3**	12	7	41	**8**	32	**15**	73	**4**	27
Wallacea	**5**	7	**18**	18	**23**	25	7	6	**4**	9	**17**	16	**21**	25	**5**	5
Mediterranean Basin	7		**8**		**15**		**15**		**9**	3	**8**		**17**	3	**17**	3
Polynesia/Micronesia		24	**9**	28	**9**	52	**1**	34		22	**8**	22	**8**	44		29
Cape Floristic Province	**1**	34	**5**	12	**6**	46	**6**	24		24	**4**	7	**4**	31	**4**	18
Mountains of South-Central China	**11**		**4**		**15**		**3**		**10**	5	**4**	3	**14**	8	**2**	

Hotspot	Original Extent of Natural Vegetation				Remaining Intact Natural Vegetation			
	Total Diversity[1]	Total Endemism[2]	Combined Diversity & Endemism[3]	Equally-Weighted Combined Diversity & Endemism[4]	Total Diversity[1]	Total Endemism[2]	Combined Diversity & Endemism[3]	Equally-Weighted Combined Diversity & Endemism[4]
Guinean Forests of West Africa	**10**	**1**	**11**		**10** 4	**1**	**11** 4	
Southwest Australia	1	**6** 1	**6** 2	1	3	**7** 2	**7** 5	**1** 2
Brazilian Cerrado	**4**		**4**		**1**		**1**	
New Caledonia	30	36	66	38	20	23	43	29
Succulent Karoo	18	5	23	12				
California Floristic Province								
Caucasus								
Central Chile								
New Zealand								

Table 3.4 (contd.) Figures in **bold**: re-analysis of original data from Mittermeier *et al.* (1999), assuming $z = 0.14$; figures in non-bold: original hotspot ranking totals taken from Tables 45 a-d and 46 a-d of Mittermeier *et al.* (1999), pp. 54-55. All calculations presented here are based on data for species richness and endemism and for original extent of and remaining intact natural vegetation presented by Mittermeier *et al.* (1999), which differ slightly in some cases from that given in Myers *et al.* (2000).

[1] *Combined total for the 10 most diverse hotspots ranked by species diversity for each of birds, mammals, reptiles, amphibians and vascular plants [max. score = 50]*

[2] *Combined total for the 10 most diverse hotspots ranked by species endemism for each of birds, mammals, reptiles, amphibians and vascular plants [max. score = 50]*

[3] *Summed total of both species diversity and species endemism rankings [max. score = 100]*

[4] *Summed total of both species diversity and species endemism rankings, with vascular plants and combined non-fish vertebrates weighted equally [max score = 40]*

3.8 Discussion

3.8.1 Higher taxa as surrogates of species-level diversity

Estimates of species richness used here are well-correlated with numbers of families and genera per region (see Section 3.1). This, however, does not necessarily mean that these estimates are themselves reliable, and this cannot really be gauged until a complete species-level checklist for the world has been completed (Govaerts, 2003). Unfortunately, there are few reliable published literature estimates, and many are themselves little more than guesses. Nevertheless, a strong correlation between numbers of species and numbers of higher taxa increases the relevance of the work presented in this thesis by suggesting that patterns in the diversity and distribution of families and genera, which are the focus of study here, are indeed representative of patterns in the diversity and distribution of species.

3.8.2 Patterns in frequency distributions

Of the attributes of plant taxa studied here (taxon size, range size and frequency of distribution patterns), all show exceptionally skewed frequencies, with the most common value being the smallest. Though the degree of skew is greater at lower taxonomic ranks (see Figures 3.4 and 3.5), frequency distributions for higher ranks are themselves still skewed. This means that any biological mechanism generating such skewed frequency distributions must likewise be applicable at all taxonomic scales. Given that a skewed range size frequency distribution is still evident at such a coarse geographical resolution (c.f. Gaston, 2003), any explanation for this must equally be applicable over many different geographical scales. Understanding what might be the cause of such patterns is still lacking. The shape of the genus range size frequency distribution shows that the majority of genera are only found in very few regions (see Figure 3.4), a pattern which is echoed by the frequency distribution of distribution patterns (see Figure 3.6). The majority of genera are found in only a small number of unique distribution patterns, the majority of which are only found in one or a few TDWG Regions. This should mean that adjacent regions share many more genera than do regions more distant from each other.

3.8.3 The relationship of diversity to distance

The floristic similarity between two TDWG regions did indeed decline as the distance between them increased; however, measuring 'distance' by the difference in respective latitudinal positions of regions showed a more linear relationship than did simply measuring absolute distance. Nekola and White (1999) studied distance decay of floristic similarity in temperate communities of distances up to 6000km and found that the model with the most linear relationship was a semi-logarithmic one, with the similarity values log-transformed and the distance untransformed, which

implies an exponential rate of distance decay. However, none of the systems studied by Nekola & White (1999) included a significant latitudinal gradient. It is possible that Nekola and White (1999) did not study the distance-decay of similarity over large enough distances to begin to see an increase in similarity. In a study of lowland tropical forest plots in Panama, Ecuador and Peru, Condit *et al.* (2002) found that similarity initially fell rapidly at distances of 3 – 5 km before continuing to fall more slowly, and they did not find a linear relationship under any transformation. Again, this may be due to differences in scale between studies: they acknowledge that Nekola and White (1999) did not have such fine-scale data as they did.

In the analysis presented here (see Section 3.6), log-transforming the floristic similarity as did Nekola & White (1999) helped to linearize its relationship with distance, but only until distances of approximately 8,000 km, after which similarity between regions appears to increase; the relationship between similarity and latitudinal difference between regions was already linear without transforming either variable. Although linear trendlines could be fitted to either Figure 17A or Figure 17B, it is noticeable nonetheless that there are fewer data points at both very small distances with very low floristic similarity and also at very large distances in Figure 17A compared with Figure 17B, and a greater concentration of data points at very small distances with very high floristic similarity and at intermediate distances with very low floristic similarity in Figure 17A compared with Figure 17B. Assuming therefore that this increase in similarity at very large distances is not simply an artefact of the polynomial regression used in Figure 17, it may be due to very widespread genera. Very large distances imply pairs of regions either side of the tropics, rather than a pair of one temperate and one tropical region: because more genera are found in the tropics (see Figure 3.9 and Chapter 4), they have a smaller proportion of very widespread or cosmopolitan genera also common to temperate regions (and pantropical genera are not usually found in temperate regions); therefore, as they are missing the exclusively tropical taxa and also have smaller overall diversities, the proportion of genera in common is greater in two temperate regions from opposite hemispheres even though the absolute number of cosmopolitan genera will be more-or-less uniform in different regions (see also Chapter 6).

3.8.4 The relationship of diversity to area

The present study differs in two major ways from most published species-area studies: the rank of taxonomic focus (genus-level rather than species-level) and the geographical scale (global rather than regional). Most species-area studies concentrate on a particular island system with individual islands of differing sizes (e.g. Wright, 1981); those studies which have compared geographically separate areas have done so at comparable latitudes (e.g. Cody, 1975), rather than across latitudes as here. In addition, the biological meaning of the coefficient (c) and the exponent (z) still remain poorly understood. The 'classical' values for z approach 0.25 (MacArthur & Wilson, 1967; Gould, 1979; Rosenzweig, 1995), with higher values usually attributed to smaller island areas and smaller values to less diverse continental areas. MacArthur & Wilson (1967) grouped z values

into two classes: island archipelagos, typically having z values of 0.25 – 0.45; and continental areas within biogeographic provinces (≈ floristic regions), typically having z values of 0.14 – 0.15; a third class of relationships, between biogeographic provinces, is discussed by Rosenzweig (1995), with z values typically of 0.9, approaching or even exceeding unity (1). Rosenzweig (1995) lists several examples of such inter-provincial z values. Gould (1979) has suggested that comparison of c values for slopes of constant z (and calculated by the same methods) will reveal additional factors to explain variance beyond just area, for example latitude and distance from source area for islands.

Slope values were also calculated for the data from Davis *et al.* (1994 – 1997) and Mittermeier *et al.* (2000), which list species numbers for areas of the world particularly rich in vascular plant species. In the former case, this is for over 200 areas ranging in size from a few tens of km^2 to complete biomes and biogeographic provinces of several thousand km^2; in the latter case for the 25 most species-rich regions, all of the scale of biogeographic provinces, which cumulatively account for almost 12% of the total land area of the Earth. Data from Davis *et al.* (1994 – 1997) yield a z value of 1.13, while those of Mittermeier *et al.* (2000) yield a z value of 1.26. The z values from the data in Davis *et al.* (1994 – 1997) and Mittermeier *et al.*(2000) may be even higher than those listed by Rosenzweig (1995) because those latter data only included sites with similar ecological conditions and from within the same latitudinal band, whereas the former data analysed here include areas of widely different latitudes. In this context the relatively low z value for species-level data from Figure 3.7 is puzzling; as arbitrary geo-political units which cross biogeographic provinces, diversity across TDWG Regions should surely fall under the 'inter-provincial' class of species-area relationships, yet a z value of 0.35 rather than approaching unity is comparable to those for island archipelagos and not inter-provincial patterns.

This low z value is perhaps partly explained by data points which cover not just different biogeographic provinces but also both island and continental systems, and not just within similar latitudes but also between dissimilar latitudes. Within the set of TDWG regions, the majority are all large regions which are continental or single land masses, which show small z values, while all small regions are island archipelagos with higher z values. Despite Rosenzweig (1995) listing several examples of inter-provincial species-area curves with very high z values, none of these is truly global in coverage, unlike the scope of this thesis. Given that both island archipelago systems and non-isolated continental systems have lower z values than do inter-provincial systems on their own, a global coverage of species-area patterns might be expected to show a z value in between the high values of inter-provincial systems and the low values of within-province continental systems; the value of 0.35 for the species-level data in this thesis might be said to do this.

With a lack of similar studies to this at higher taxonomic levels it is also difficult to make a meaningful comparison of z values for families and genera. The exponent for family-level data is only 0.12, and also the residuals of the regression are smallest for family richness data, indicating that there

is less of a decrease in family diversity with increasing latitude (see also Figure 3.15), i.e. that families are generally more widespread than are genera (see also Figures 3.4 and 3.5). It is still possible to distinguish between two broad bands of regions, one tropical and the other temperate. This must mean that, despite families in general being more widespread than are genera, there are still many more exclusively tropical families than there are widespread or exclusively temperate families (see Figure 3.15). For genera, these patterns are exaggerated further (proportionally fewer widespread genera and proportionally even more exclusively tropical genera), while for species they are exaggerated further still. The ratio of species to both genera and families is also lower in the small, island archipelago regions than in the large, continental regions, further reducing the slope of the regression at higher taxonomic levels.

3.8.5 Global patterns of angiosperm diversity

The Neotropics in general, and in particular western South America and southern Central America, consistently emerge as the most diverse areas of the world for plants, at all spatial scales. Results from the analysis of the data presented in Table 2.1 (see Section 3.5) are supported by the re-analysis of data presented by both Mittermeier *et al.* (1999) and by Davis *et al.* (1994 – 1997) and by the analyses of Barthlott *et al.* (1996) and Gentry (1988), all of which list areas of southern Central America (Costa Rica/Panama) and northwest South America (northwest Colombia and eastern Andean Colombia and Peru) as being the most diverse of all the areas studied (see Section 3.7). In the study of Gentry (1988), the only 0.1 ha plot in the five richest (for numbers of tree species) from outside of the Neotropics was from Sarawak in Borneo; similarly, in Figure 3.15 and in Figure 3.16, Region 42 (Malesia), which includes Borneo, has the second highest relative species diversity from this analysis behind only Western South America, in which the other four out of five richest sites in Gentry's analysis are found. The studies listed above all explicitly addressed species diversity, whereas this study has confirmed the prominence of Western South America for generic diversity also (see Table 2.1 and Figure 3.15).

That Central America is among the most diverse was not expected, the small area of this region masking its apparent diversity (see Figure 3.15). However, this may be due to Central America being both the southern limit of distribution for many northern hemisphere taxa, and the northern limit of distribution for many tropical and southern hemisphere taxa, given that degree of generic endemism is not especially high for Central America (3.7%). That is to say, Central America straddles different biogeographic regions, in addition to there also being many pantropical taxa or taxa from throughout the Neotropics there. This same argument applies to Mexico just to the north, although the flora of Mexico shows a more pronounced northern bias (many North American genera do not extend south beyond Mexico, while many South American genera do not extend north beyond Central America). Geologically, the mountains today forming the backbone of the Central American isthmus have their origin in the Andean orogeny caused by the meeting of the North American and South American

plates, with a recent period of uplift under the northwestern Andes from approximately 5 million years ago (mya), and a growing series of volcanic islands between the two continents then coalescing to form the Isthmus of Panama in the Pliocene, approximately 3.5 my (Marshall *et al.*, 1980; Coates & Obando, 1996).

3.8.6 Biogeographical hypotheses

Gentry (1982) proposed that the exceptionally high floristic diversity of the northwestern Andean and southern Central American region was principally due to rapid, sympatric *in situ* speciation caused by the uplift of the Andes mountains, on top of an already-rich tropical flora, and he estimated that this explosive speciation might account for approximately half of the total number of Neotropical species. Gentry (1982) distinguished three broad floristic elements within the Neotropics as a whole (notwithstanding the tremendous floristic complexity of such a large area at more detailed scales): Laurasian taxa, which were originally absent from the isolated South American continent but which migrated south from North America as the Isthmus of Panama closed; Gondwanan, Amazonian-centred trees and lianas with almost half of their species complement in Amazonia; and Gondwanan, Andean-centred taxa with only a minority of species present in Amazonia. Plant groups with a northern-Andean-centred distribution tend to be the abundant guilds of epiphytes, under-storey shrubs and palmettos that are less prominent in the Palaeotropics. These groups are characterised by numerous local endemic species in small habitat patches subject to frequent disturbance, with rapid generation times and specific pollinator relationships, all factors promoting rapid speciation (Gentry, 1982). This hypothesis remained untested until borne out in a recent study of the species-rich genus *Inga* (Leguminosae), where molecular estimates of the mean divergence time of all species from the most-recent common ancestor of the genus from two separate gene regions were between 3.5 and 5.9 mya, respectively (after application of non-parametric rate smoothing; Richardson *et al.*, 2001a).

Gentry (1982) further hypothesized considerable floristic interchange between Northern and Southern America following the closing of the Panamanian isthmus, though he suggests that in this case the interchange occurred without significant amounts of speciation (at least for woody taxa) in either the northern or southern elements. Also, there seems to be some ecological differentiation between the two elements, with many Laurasian (northern) taxa today confined to the montane Andes and absent from lowland South America, and Gondwanan, Amazonian-centred (southern) taxa dominant in lowland forest types (Gentry, 1982). Furthermore, the woody Laurasian taxa show a higher proportion of wind-pollination, longer generation times and low numbers of species. Gondwanan, Amazonian-centred trees and lianas, however, show restricted, allopatric distributions within a widespread lowland rain forest habitat, with species of unrelated groups found in the same small areas which have been proposed as Pleistocene forest refuges within a (then) more widespread savannah (Prance, 1973; Gentry, 1982) or seasonally-deciduous forest habitat (Pennington *et al.*, 2000).

This scenario of both northern and southern migrations into Mexico, Central America and western South America may also apply to Region 81, the Caribbean. The proto-Antilles was created at the junction of the Northern and Southern American plates, and was then isolated by the pinching of the Caribbean plate from the easternmost section of the East Pacific plate when North and South America eventually joined, forming the Greater Antilles we know today (Gentry, 1982; Hedges, 2001). Gentry (1982) emphasised the dual northern and southern affinities of the Caribbean flora, which Rosen (1975) had suggested was due to the original vicariant stocking of a proto-Antillean island arc which was located between separate North and South American continents. The age of the Greater Antilles (Cuba, Hispaniola, Puerto Rico and Jamaica) is estimated at about 60 – 45 my and for the Lesser Antilles at about 10 my (Rosen 1975, 1985; MacPhee & Grimaldi, 1996; Hedges, 2001). Rosen (1975) identified two major terrestrial elements of the Caribbean biota, between each of North America and South America, and the Caribbean, respectively, which 'represent extensions of the original biotas into the Caribbean region', as well as two marine elements between both the eastern Atlantic and the eastern Pacific, and the Caribbean. In this vicariance model, the biotic interchange occurred primarily through the proto-Antilles archipelago, not the modern-day Greater and Lesser Antilles, as it later became. However, a series of recent studies on different animal taxa (reviewed in Hedges, 2001) calculate divergence times for Caribbean lineages which are too young to have been the product of vicariance and which imply later (Cenozoic) over-water dispersal from modern-day continents to various Caribbean islands.

The Caribbean also shows proportionally higher generic endemism (12.6%) than do either western South America (10.8%) or Central America (3.7%), and in this case one might think the fragmentary nature of the region is contributing to increased generic diversity. In fact, however, this is not so; some two-thirds of Caribbean genera (134 of 201) are confined to either or both of only two islands: Cuba and Hispaniola. It is the isolation of this region from the continental Americas that is important, not the fragmented nature of the region itself, and the predominance of endemic genera on only the oldest islands of the Caribbean archipelago furthermore might perhaps give more weight to a vicariant rather than a dispersalist origin for these endemics, in contrast to zoological taxa, where the distribution of modern-day higher-level taxa is often widespread, but diversity of both modern-day and fossil higher-level taxa is low (Hedges, 2001).

A single tectonic episode, the fusion of the two American continents, would therefore have allowed floristic interchange between two isolated floras across what became the most-diverse, Neotropical regions, 79 (Mexico), 80 (Central America), 81 (Caribbean), and 83 (Western South America). However, this interchange would have occurred at different times: in the late Mesozoic / early Cenozoic for the Caribbean (Rosen, 1975, 1985; Hedges, 2001); during the Pliocene for Central America and western South America (Gentry, 1982; Richardson *et al.*, 2001a). In the former case, the scenario is one of initial floristic interchange followed by isolation and then additional later dispersal

to the Caribbean, resulting in high endemism. In the latter case, this initial diversity from floristic interchange is then thought to have been coupled with high speciation rates following isolation of populations by the successive phases of the Andean Orogeny in the Central American / northwest South American region.

3.9 Summary

- Biodiversity is difficult to define, but the most-widely used measurement of it is species richness.
- 'Hollow-curve' frequency distributions are found with taxonomic, spatial and temporal phenomena, although their true significance is debated.
- The majority of angiosperm diversity is explained by relatively few common distribution patterns.
- Patterns in the distribution of genera mirror those seen at species level, and genera can be used as a reliable surrogate for species-level diversity.
- Data at different taxonomic levels fit the species-area model, though there is considerable scatter in the spread of points, mostly due to latitude.
- There is a poor correlation between the diversity of an area and the degree of endemism there.
- True relative diversity scores for regions of different size are given by re-scaling with the species-area relationship.
- Tropical regions contain many more genera than temperate regions, with the Neotropics more diverse than SE. Asia, which in turn is more diverse than Africa.
- Areas of the Neotropics consistently emerge as the most bio-diverse areas of the world at all geographical scales of analysis, and at both genus-level and species-level.
- The most diverse regions of all, in this analysis, are those which straddle biogeographic provinces.

CHAPTER 4

THE LATITUDINAL GRADIENT OF DIVERSITY

The previous chapter focused on patterns of diversity around the world and the roles of area and distance in shaping these patterns. The principal focus of this chapter is the so-called latitudinal gradient of diversity, or the fact that the tropical regions are so much more bio-diverse than temperate regions. It is a phenomenon first brought to the attention of biologists over two hundred years ago, and yet, despite a multitude of hypotheses, it still lacks a comprehensive, widely-agreed explanation. Though not all of the many hypotheses to explain the latitudinal gradient have been (or could have been) studied here, at a deeper level, the latitudinal gradient must be manifested in differences between the relative ranges of taxa, a subject which can very much be studied here. Whatever the ultimate explanation for the latitudinal gradient of diversity, it involves the inter-relationships of the three principal aspects of biogeographical macroecology: range-size, richness and turnover. Fundamentally, differences in the relative range-sizes of taxa (the range-size frequency distribution), cause both the biological richness of different areas relative to each other, and also the turnover of taxa in areas relative to other areas. The study of relative range sizes of taxa in determining the latitudinal gradient of diversity is the approach which has been taken in this chapter. As concerns over the state of the natural world have increased, so efforts to summarize the existing knowledge of the world's biodiversity have grown also. This increasing knowledge of the distributions of many taxonomic groups has led to a resurgence of interest in long-standing ecological questions, including that of the latitudinal gradient of diversity (Stevens, 1989; Gaston, 1996b; Chown & Gaston, 2000).

4.1 Introduction

From the earliest botanical exploration of the tropics, one of the most notable features of the region has been its enormous wealth of plant diversity. Arriving in Venezuela in 1799, Alexander von Humboldt famously wrote home to his brother:

"We rush around like the demented; in the first three days we were unable to classify anything; we pick up one object to throw it away for the next. Bonpland keeps telling me he will go mad if the wonders do not cease." A. von Humboldt, letter to Wilhelm von Humboldt, 1799.

while in a different context, reminiscing on his journeys in the forests of Malesia from 1854 - 1862, Alfred Russell Wallace wistfully commented:

"If the traveller notices a particular species and wishes to find more like it, he may often turn his eyes in vain in every direction. Trees of varied forms, dimensions and colours are all around him, but he rarely sees any one of them repeated." A.R. Wallace, 1878, p. 65.

The latitudinal gradient of diversity has been described as "one of the most prominent features of the natural world" (Rohde, 1992; Taylor & Gaines, 1999), but also as "the major, unexplained pattern in natural history" (R. E. Ricklefs, quoted in Lewin, 1989). For despite these early descriptions, and decades of subsequent ecological research, an adequate understanding of the cause(s) of the latitudinal gradient of diversity remains lacking. Greater numbers of species are found in the tropics for almost all groups of organisms and in almost every type of habitat, both terrestrial and aquatic, and so, if we are searching for a comprehensive understanding of the latitudinal gradient, the explanation must then apply equally well to nearly all organisms and environments – something which has not always been the case (Rohde, 1992).

In a recent review, as many as 28 possible mechanisms that were claimed to be in current use were discussed (Rohde, 1992). These were divided into three broad categories: external 'ecological' factors such as area, seasonality, productivity, tropical environmental stability, environmental instability (for example, glaciation), environmental heterogeneity; circular, self-reinforcing factors such as competition, predation, niche-breadth and population growth rate where increased diversity is in essence the product of high existing diversity usually of other taxonomic groups; and extended temporal factors of increased evolutionary speed due to shorter generation times, higher mutation rates and accelerated selection processes in the tropics. These last factors must themselves be the outcome of both external factors such as greater energy availability in the tropics and self-reinforcing factors such as competition and niche-breadth (Rohde, 1992), and so represent a pluralistic hypothesis which acknowledges that a single causal process might not be found.

The great diversity of the tropics is evident not only at species level, but also at genus and family level, as the three peaks for Africa, SE. Asia and the Neotropics in Figure 3.12 (Chapter 3) showed. Figure 4.1 plots standardised genus diversity of each region against minimum distance from the equator of that region (≈ minimum latitude of that region). Since any genus record will be included for that region if it barely crosses over the border, and diversity is known to decline away from the equator, this should have the effect of under-estimating the latitudinal gradient of diversity, if anything. From Figure 4.1 it is clear that there is a strong relationship between the latitude of a region and its genus richness. The values plotted in Figure 4.1 are re-scaled genus richness for regions against latitude; both northern and southern hemispheres are shown. Re-scaled genus richness values, calculated after correction by the species-are relationship (see Chapter 3), standardise for the effects of different sizes of regions. Thus the latitudinal gradient in diversity evident from Figure 4.1 is independent of any considerations of land area.

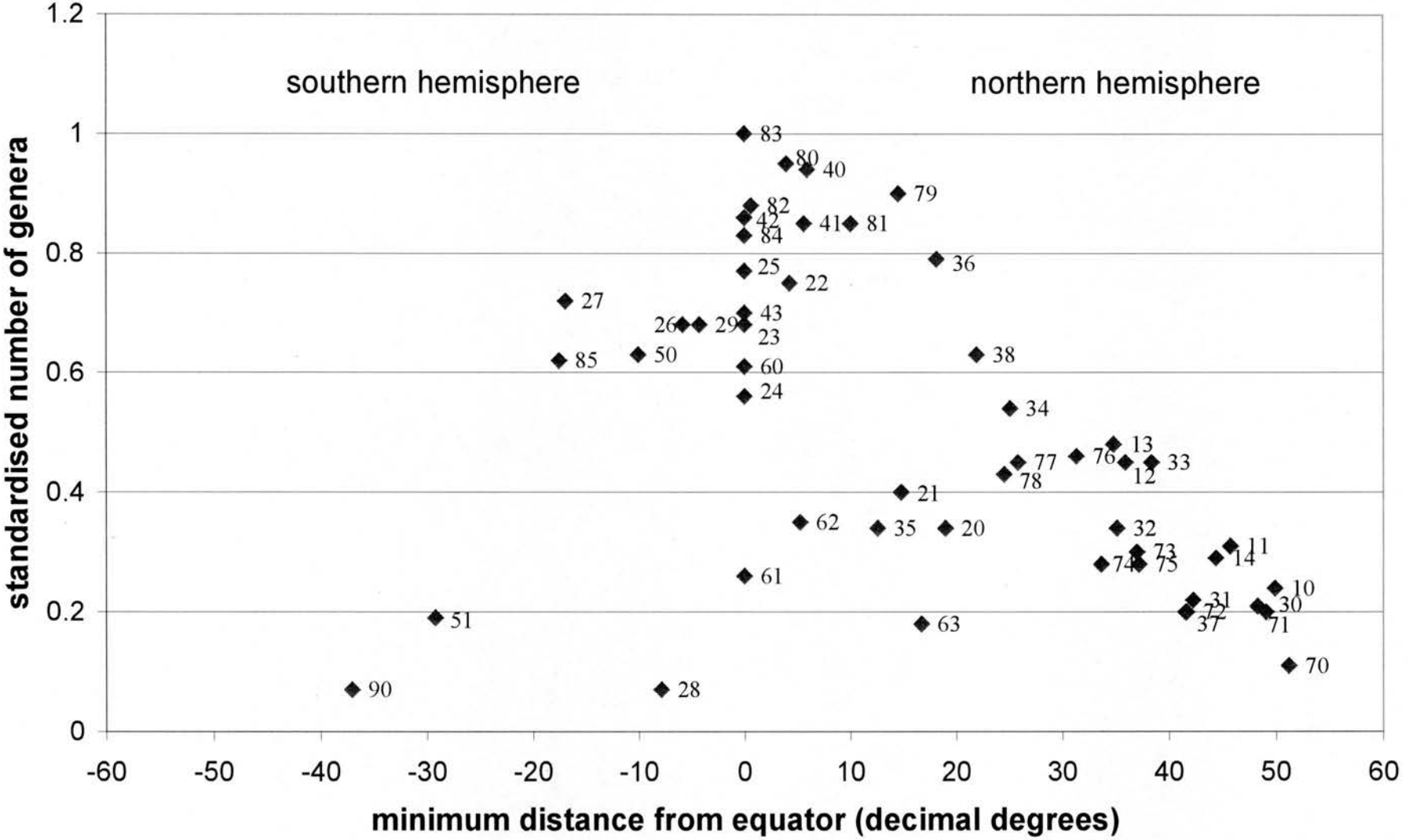


Figure 4.1 The diversity of angiosperm genera is much greater in tropical regions. Regions are indicated by their 2-digit codes, and generic diversity is scaled relative to Region 83, Western South America and standardised to be independent of sizes of regions (see Chapter 3); negative latitudes arbitrarily denote the Southern Hemisphere.

The shape of the graph is roughly parabolic: diversity increases at the same rate in either hemisphere towards the equator. Also, it is again possible to see that, in general, the Neotropics is more diverse than is SE. Asia, which is again more diverse than is Africa. Regions falling below the apex of the distribution (that is, not as diverse as would be expected) can mostly be explained on a case-by-case basis. Regions 20 (Northern Africa) and 35 (Arabian Peninsula) are both covered by extensive deserts, the Sahara and Arabian Deserts, respectively; their low diversity is surely a product of their excessive aridity. Regions 28 (Middle Atlantic Ocean [St Helena and Ascension I.]), 61 (South-Central Pacific) and 63 (North-Central Pacific [Hawaiian Islands]) are all isolated island systems. However, the low diversity of Region 24 (Northeast Tropical Africa) remain difficult to explain, excepting that the Sahara Desert also occupies large expanses of the northern portions of this region.

A better measure of the latitudinal gradient of diversity, however, is by the minimum and maximum latitudinal extent of the distribution of each genus. This was measured by fitting a Minimum Enclosing Rectangle around the entire distribution; the maximum and minimum latitudinal extent are then simply the most northerly and southerly point of all regional polygons for that genus, respectively. Figure 4.2 shows this data for all angiosperm genera across the globe. The tropics are represented by a roughly constant plateau in the graph of between 7,000 and 8,000 genera. Individual peaks and troughs within this plateau represent latitudinal bands which happen to intersect a greater or lesser number of regions at that latitude, respectively. The step-like nature of the increase in diversity across the southern hemisphere is a consequence of the relatively coarser dividing of the world into large polygons in an area of much less land area, while the tail towards the South Pole at the left hand side of the graph is caused by the two genera to reach the Antarctic Continent, *Deschampsia* (Gramineae) and *Colobanthus* (Caryophyllaceae).

This method is of course likely to overestimate the latitudinal extent of each genus, simply because the TDWG Level 2 polygons are so coarse. The actual distribution of any genus will, most likely, not extend to fully the northernmost or southernmost point of those polygons. Two possible methods might be used to try to overcome this: reduce both the maximum and minimum latitude by some amount and re-calculate the range-size between them, or use some other measure of most northerly and southerly points, such as the centroid or median latitude of polygons. Both these were tried (truncating by 10%; measuring polygon centroid); in neither case, however, is the latitudinal diversity gradient greatly affected (see Figure 4.3).

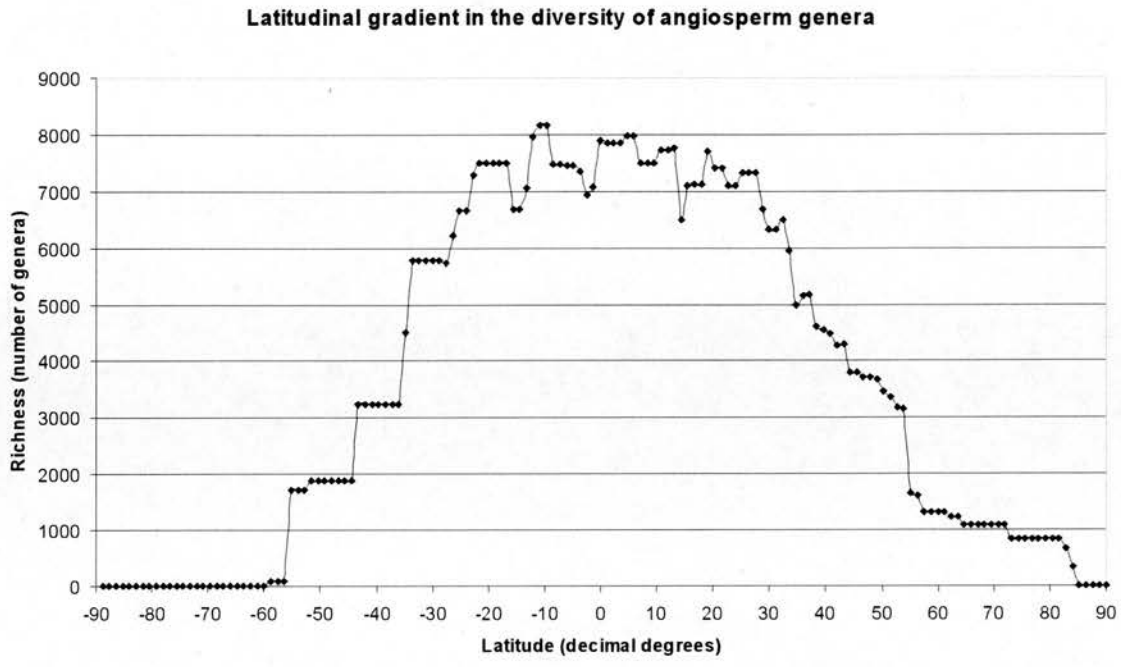


Figure 4.2 Frequency distribution of genus diversity across the latitudinal gradient of the World, showing that many more genera are found in the tropics.

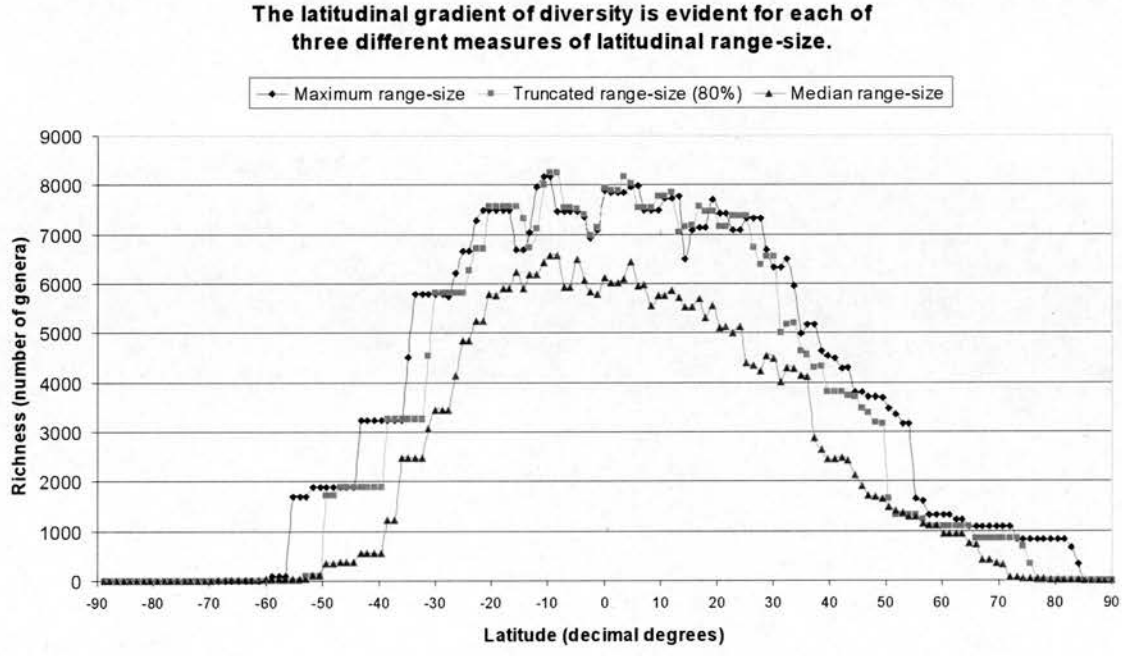


Figure 4.3 The latitudinal gradient of diversity is evident for each of three different measures of latitudinal range-size. Negative latitudes denote the Southern Hemisphere.

4.2 The area effect

4.2.1 Introduction

Rosenzweig (Rosenzweig, 1992) has vigorously championed Terborgh's (1973) argument that the latitudinal gradient in diversity is mainly due to the greater extent of climatically–equivalent land available in the tropics. This had not been noticed before, he suggested, because areas had usually been calculated from a Mercator-like map projection, where lines of longitude appear in parallel. This reduces the apparent size of tropical latitudes. In fact, lines of longitude converge on the poles and are at their most distant on the equator. The tropics, therefore, are actually of much greater extent than usually recognised and, under the species-area model, greater areas would be expected to show logarithmically greater numbers of species. Rather than use a Mercator projection, areal extent should therefore be calculated on an equal-area projection, such as Peters' projection used for Figure 4.4 below, which shows the actual amount of available land at different latitudes. Clearly there is far greater land available in the northern hemisphere than in the tropics or the southern hemisphere.

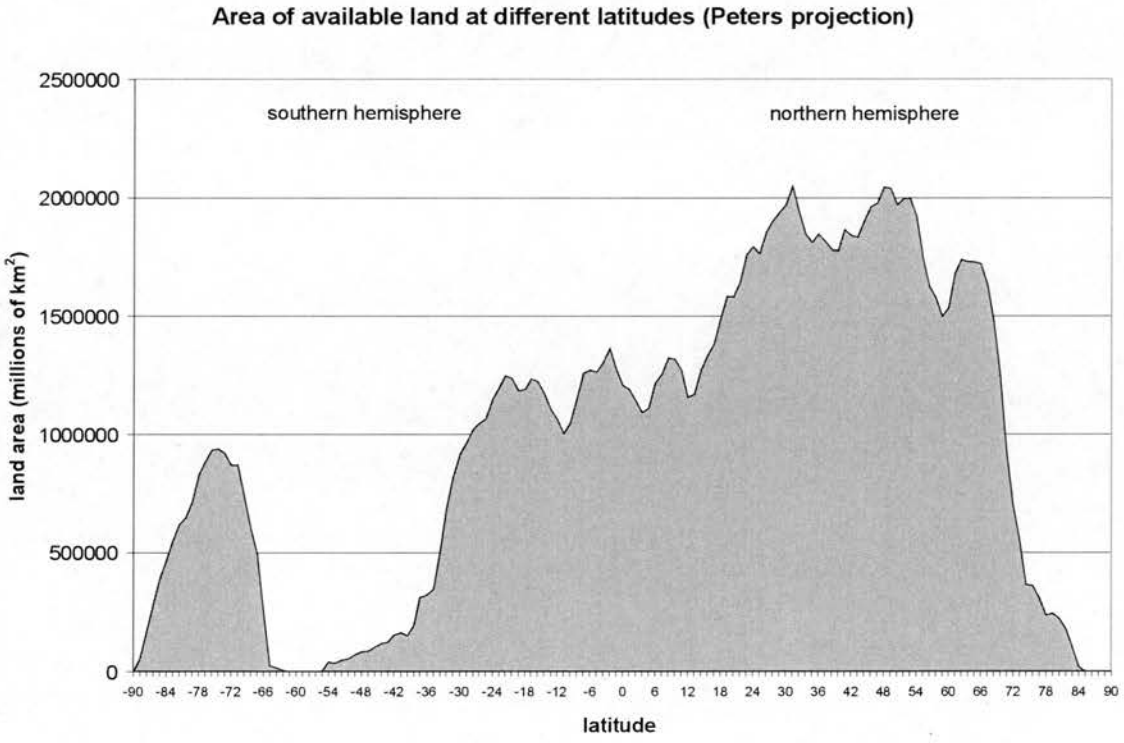


Figure 4.4 Amount of available land area differs with latitude across the world.

Rohde (1997) also claimed that the temperate regions held the greatest land area and disputed Rosenzweig's claim for greater available land in the tropics. However, in his paper (Rohde, 1997) all land between 23.5° and 66.5° was grouped into a single temperate biome. In fact, climatic variables vary more rapidly over shorter latitudinal distances outside of the tropics than they do within the

tropics (Terborgh, 1973; Rosenzweig, 1992); therefore, extra-tropical biomes are defined more tightly than is the tropical biome. Also, tropical areas abut on each side of the equator, so the tropical latitudinal band, which stretches from 23.5°N to 23.5°S, is wider than any single subtropical or temperate band. Furthermore, since there is a relatively small amount of land in the southern hemisphere, this means that even after summing the corresponding northern and southern temperate bands, each of these will be smaller than the contiguous tropics. So on these three counts, the amount of available area is considerably greater in the tropics.

Rosenzweig (1992) divided the world into five biomes: the tropics (± 26°); the subtropics (26° – 36°); temperate (36° – 46°); boreal (46° – 56°); and tundra (>56°). This treatment has been re-assessed here. Figure 4.5 below shows relative proportions of land at different latitudinal bands, with the data from Figure 4.4 summed into 10 degree latitudinal bands outside of the tropics. Clearly, Rosenzweig's tropical regions contain more than twice as much land area as does any other of his biomes, which reflects the greater latitudinal width of his tropical zone. Since he admits this division of the world is largely arbitrary (Rosenzweig, 1992; Rosenzweig & Sandlin, 1997), here three different latitudinal-band schemes were tried, with different numbers of bands and different latitudinal widths of bands. Table 4.1 below shows land area in millions of kilometres squared for the earth under three different latitudinal-band schemes, dividing the world both more finely and more coarsely than did Rosenzweig (1992), and each time the land area of the tropics is significantly greater.

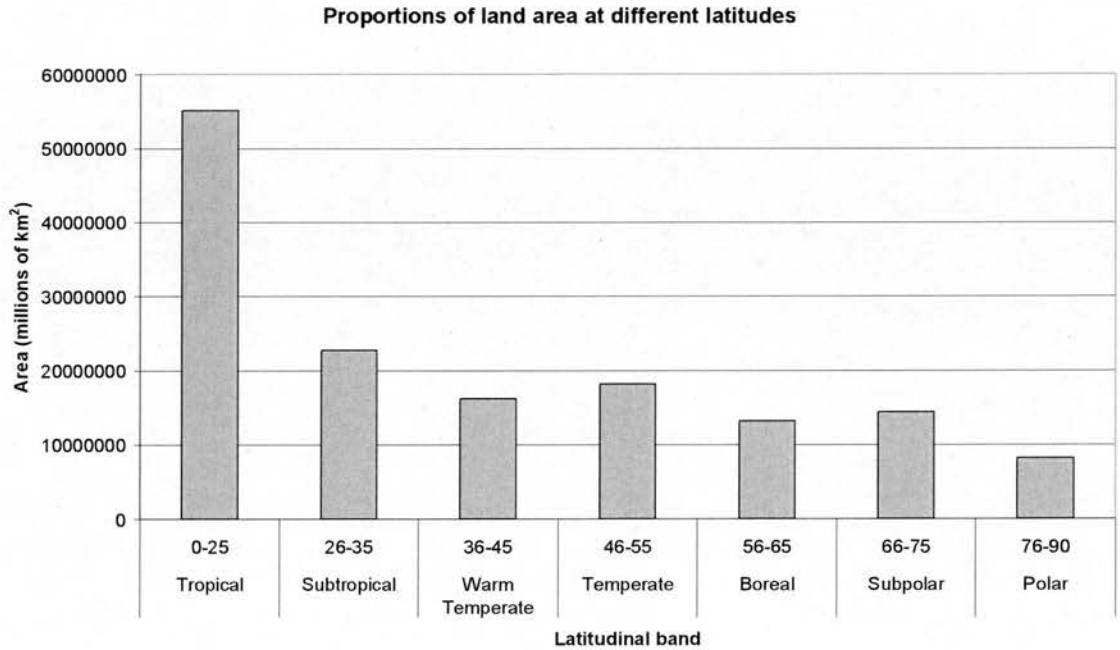


Figure 4.5 Amount of land at different latitudes; tropical regions contain more land area than do any other climatic zone.

Rosenzweig (1992, 1995; Rosenzweig & Sandlin, 1997) proposed that the greater area of the tropics would lead to increased diversity by increasing the speciation rate and dampening the extinction rate. If the tropics has a greater extent of climatically-equivalent land area, then tropical taxa should have a potentially greater geographical extent. Greater geographic ranges lead to increased rates of allopatric speciation, since the larger the geographic range, the greater the chance of crossing a geographic barrier and isolating a peripheral population which will result in the formation of a new species. In addition, larger geographic ranges reduce the probability of extinction, since larger ranges mean both that population sizes for species will be larger (and larger populations have a lower probability of extinction) and that each species has a greater probability of finding favourable niche refuges to provide sanctuary in the event of fluctuating climatic conditions.

Latitude	Region	Land area (millions of km^2)
0°-25°	Tropical	55084484
26°-35°	Subtropical	22669737
36°-45°	Warm temperate	16196218
46°-55°	Temperate	18177394
56°-65°	Boreal	13176101
66°-75°	Subpolar	14340722
76°-90°	Polar	8229627
0°-25°	Tropical	55084484
26°-45°	Warm temperate	38865955
46°-65°	Boreal	31353495
66°-90°	Polar	22570349
0°-25°	Tropical	55084484
26°-35°	Warm temperate	22669737
36°-55°	Temperate	34373612
56°-90°	Boreal & polar	35746450

Table 4.1 Division of the world into three arbitrarily-defined latitudinal-band schemes to test the idea that the land area of the tropics is always greater than is any other latitudinal band.

The greater land area of the tropics is therefore thought to increase the speciation rate and reduce the extinction rate through allowing greater geographical range sizes, which results in a higher steady-state diversity than in temperate regions (Rosenzweig, 1992). Rosenzweig's area model makes several assumptions or predictions which might allow the hypothesis to be verified: that available geographic area and geographic range size of taxa are positively correlated; that larger range sizes result in increased speciation rates; and that larger range sizes result in reduced extinction rates. A recent review of the subject found only equivocal or circumstantial evidence in support of these assumptions underlying the model (Chown & Gaston, 2000), even while admitting to the fundamental

importance of area availability in determining diversity gradients. However, given the importance of available area in determining diversity (see Chapter 3), the first of these testable predictions was studied in this chapter.

Rosenzweig (1992, 1995; Rosenzweig & Sandlin, 1997) proposed that the greater area of the tropics would lead to increased diversity by increasing the speciation rate and dampening the extinction rate. If the tropics has a greater extent of climatically-equivalent land area, then tropical taxa should have a potentially greater geographical extent. Greater geographic ranges lead to increased rates of allopatric speciation, since the larger the geographic range, the greater the chance of crossing a geographic barrier and isolating a peripheral population which will result in the formation of a new species. In addition, larger geographic ranges reduce the probability of extinction, since larger ranges mean both that population sizes for species will be larger (and larger populations have a lower probability of extinction) and that each species has a greater probability of finding favourable niche refuges to provide sanctuary in the event of fluctuating climatic conditions.

The greater land area of the tropics is therefore thought to increase the speciation rate and reduce the extinction rate through allowing greater geographical range-sizes, which results in a higher steady-state diversity than in temperate regions (Rosenzweig, 1992). Rosenzweig's area model makes several assumptions or predictions which might allow the hypothesis to be verified: that available geographic area and geographic range-size of taxa are positively correlated; that larger range-sizes result in increased speciation rates; and that larger range-sizes result in reduced extinction rates. However, a recent review of the subject found only equivocal or circumstantial evidence in support of these assumptions underlying the model (Chown & Gaston, 2000), even while admitting to the fundamental importance of area availability in determining diversity gradients.

4.2.2 'Zonal bleeding' might hide the area effect

Rosenzweig himself proposed two ameliorating factors which might obscure the strength of the tropical area - tropical diversity relationship: firstly, diversity is reduced in the geographically-extensive northernmost latitudes (defined as >56°; Rosenzweig, 1992) by the very low productivity of these regions; secondly, that the latitudinal gradient of diversity declines over the northern hemisphere more gradually (see Figure 4.2) than expected by the amount of corresponding available area (see Figure 4.5) might be due to the ranges of essentially tropical taxa spilling over into the temperate zones and smoothing the gradient by inflating the diversity of subtropical and warm temperate regions; Rosenzweig (1992) termed this phenomenon 'zonal bleeding' – distributions of taxa bleed into adjacent zones. If area were the single most important factor in determining diversity then the latitudinal gradient of diversity might be expected to show a step-like decline: high diversity in the tropics, low (but constant) diversity outside of it. If 'zonal bleeding' is obscuring a steeper latitudinal diversity gradient caused by greater area within tropical regions, Rosenzweig (1992) predicted that the

smooth latitudinal gradient should disappear if taxa with partly-tropical ranges were omitted from patterns of extra-tropical diversity (Rosenzweig, 1992). 'Zonal bleeding' might thus provide indirect support for the geographic area hypothesis to explain the latitudinal diversity gradient. Though this does only provide an indirect test, the subsequent discovery of zonal bleeding in New World birds (Blackburn & Gaston, 1997) and South American mammals (Ruggiero, 1999) provides some support for the area model.

4.2.3 Methodology

Zonal bleeding was tested for in this dataset using the regression methodology outlined by Blackburn and Gaston (1997). Taxon-ranges were measured by their maximum latitudinal extent (i.e. the northernmost and southernmost points of any regions in which a genus occurs) and defined as either tropical (latitudinal range endpoints <25°); strictly extra-tropical (latitudinal range endpoints >25°); partly extra-tropical (latitudinal range midpoints <25°; latitudinal range endpoints >25°); or principally extra-tropical (latitudinal range midpoints >25°; latitudinal range endpoints <25°). Thus the partly extra-tropical taxa have a midpoint within the tropics but extend outside of it and artificially inflate the diversity of biomes outside of the tropics, and their exclusion should eliminate the latitudinal gradient of diversity and reveal a strong relationship between the diversity of the remaining taxa and available area. However, as discussed by Blackburn and Gaston (1997), omitting all strictly extra-tropical taxa introduces an artefact into the analysis, since by setting a hard boundary at 25° a greater proportion of taxa whose midpoints lie close to this latitude will be excluded than taxa whose midpoints lie at higher latitudes (Colwell & Hurtt, 1994), and so diversity might appear to decline towards the tropics when this may not be the case. Blackburn and Gaston (1997) recommend studying zonal bleeding using only principally extra-tropical taxa, which have a midpoint outside of the tropics yet extend into the tropics, and that recommendation is followed here. With the scale of the data used in this thesis it is impossible to know the exact latitudinal extent of a genus within any one region: it may occupy all of that region or it may only just extend over the border into that region. Setting the range-limits for taxa as their maximum latitudinal extent will therefore under-estimate the strength of the latitudinal gradient of diversity and give extra weight to any significant results.

Extracting only those principally extra-tropical genera created a dataset containing 3613 genera (approximately 28% of the global total), strictly enforcing a cut-off point for latitudinal range midpoints of 25°. 1988 of these genera had northern hemisphere midpoints and 1625 had southern hemisphere midpoints. Because of the differences in the total latitudinal ranges between different hemispheres, richness values were calculated for different numbers of bins, giving a common latitudinal bin-size of 1° ±0.1°: northern hemisphere, 90 bins; southern hemisphere 60 bins; both hemispheres, 150 bins. Richness values are simply the summed number of genera that have range midpoints falling within particular latitudinal bins. By using range midpoints, adjacent bins are statistically independent of each other (i.e. each genus is only counted once, in the bin in which its

range-midpoint falls. Regression analyses were performed between genus richness values and both available area and latitude, both separately and as a multiple regression with two explanatory variables, both for the northern and southern hemispheres separately, and for the two hemispheres combined (see Table 4.2). Because the size of regions is so much greater than the size of bins used (1° ±0.1°), many adjacent bins were of equal richness values; for these land area was summed and latitudinal midpoints taken for that area, while richness was held constant (i.e. not summed). For the northern hemisphere, the resulting number of unique richness values returned was 18; for the southern hemisphere, 5. Due to the power-law relationship between diversity and area, richness values, amount of available land area and latitudinal midpoint of biomes were all log-transformed before regression, and latitudes treated as absolute values (i.e. northern and southern hemispheres were not distinguished). Results are presented in Table 4.2 below.

The effect of area was also studied by dividing the world into equal-area rather than equal-latitude bands. This was done on an equal-area projection of the world (Peters' projection) with a grid-cell size of 1km x 1km. By counting the number of cells for all the world's land surface and dividing this by the number of bands desired (an arbitrary 150), the size of each equal area band was obtained (≈ number of grid-cells). Then, the latitude reached after counting each equal number of grid-cells was returned, resulting in latitudinal limits of 150 equal-area bands. Various sizes of equal-area bands could be obtained by taking the latitude of every 5th, 6th or 10th band, depending on which factor of 150 was required. A straight regression of richness values against latitudinal midpoint of equal-area bands should therefore exclude area effects.

Regression covariate	Northern hemisphere	Southern hemisphere	Both hemispheres
Land area	$r^2 = 0.53$	$r^2 = 0.92$	$r^2 = 0.78$
Latitude	$r^2 = 0.82$	$r^2 = 0.61$	$r^2 = 0.29$
Land area + latitude	$r^2 = 0.85$	$r^2 = 0.96$	$r^2 = 0.80$

Table 4.2 Regression statistics for the relationship between principally extra-tropical flowering plant genus diversity and either land area, latitude or both combined for the northern (n = 18) and the southern (n = 5) hemispheres individually and for both combined.

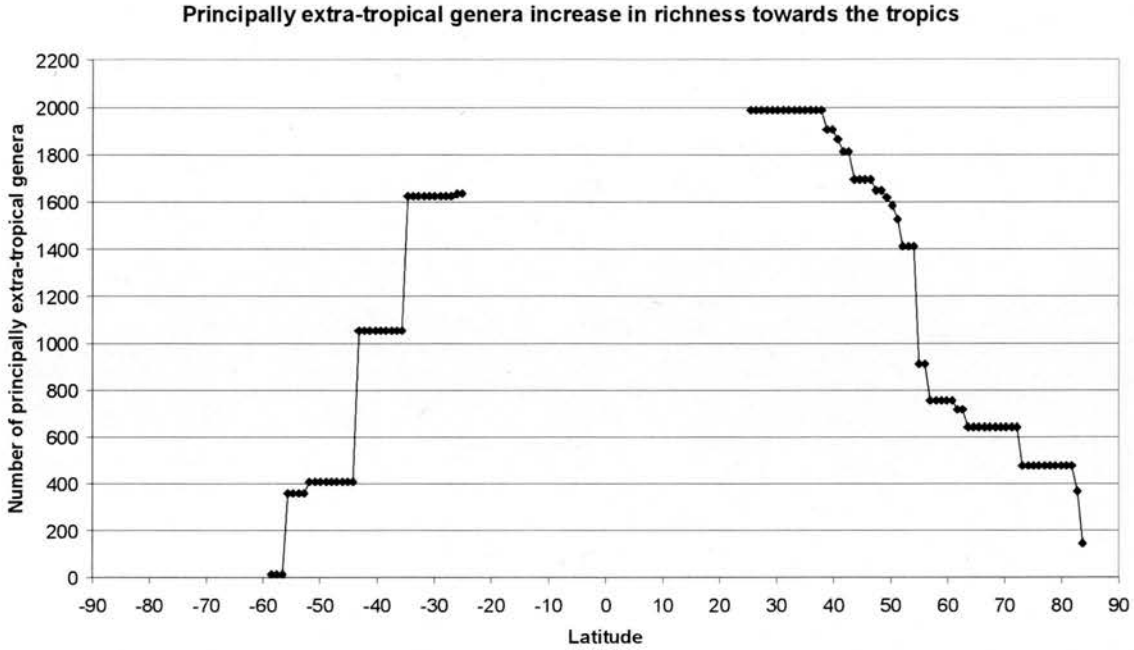


Figure 4.6. Principally extra-tropical genera show a latitudinal diversity gradient; since the definition of principally extra-tropical genera is those with range midpoint found at <25° latitude, no genera are shown beyond this point. Negative latitudes arbitrarily denote the southern hemisphere.

4.2.4 Results

A plot of numbers of principally extra-tropical genera against latitude for both the northern and the southern hemispheres is given in Figure 4.6. For the northern hemisphere, the relationship between richness and latitude was stronger than that between richness and land area; however, for the southern hemisphere this pattern was reversed, with the relationship between richness and land area very strong but that between richness and latitude not significant (see Table 4.2). This may well be due simply to the disparity in sample sizes between the two hemispheres, with 18 samples for the northern hemisphere but only 5 for the southern. This in turn is a reflection of the fewer underlying TDWG polygons in the southern hemisphere than in the north, so that the southern hemisphere portion of Figure 4.6 has a step-like appearance which is more an artefact of the presence/absence records in fewer polygons than the actual shape of the gradient. However, there is still a strong relationship between principally extra tropical diversity and land area for both hemispheres combined that is only marginally improved (r^2 increasing by 0.02) in the multiple regression on both available area and latitude. Indeed, the results for both variables in combination only marginally improve on the strongest relationship in either hemisphere, either separately or in combination.

As shown in Figures 4.7 and 4.8, richness plots for both equal-latitudinal bands and equal-area bands show little difference, both for the whole world and for genera confined to the New World, except perhaps over the far-southern hemisphere where land area is considerably reduced (see Figure 4.8). This is a surprising result, especially for the New World which is so longitudinally-constricted

through Central America and the Isthmus of Panama that an equal-area band across this region has a very broad latitudinal amplitude. It is presumably again due simply to the coarse size of the regions underlying the latitudinal ranges; for differences in the latitudinal limits between the equal-latitude and equal-area bins to be seen as differences in their respective generic richness, latitudinal positions of the bin ranges would have to cross the latitudinal limits of the regions. For such large underlying regions this is unlikely to occur, even with only 15 bins as in Figures 4.7 and 4.8, so that the limits of either equal-latitude or equal-area bins will fall within the same TDWG region and therefore return the same values of generic richness.

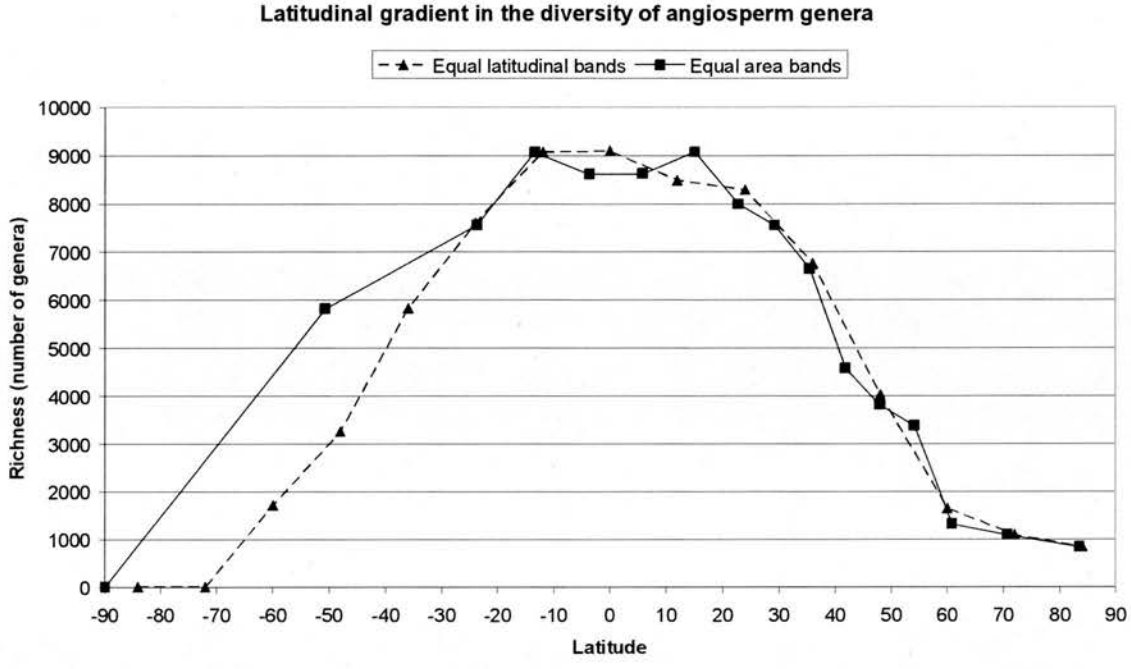


Figure 4.7 Using either equal-area or equal-latitude bands has little effect on patterns of genus richness. Negative latitudes denote the southern hemisphere.

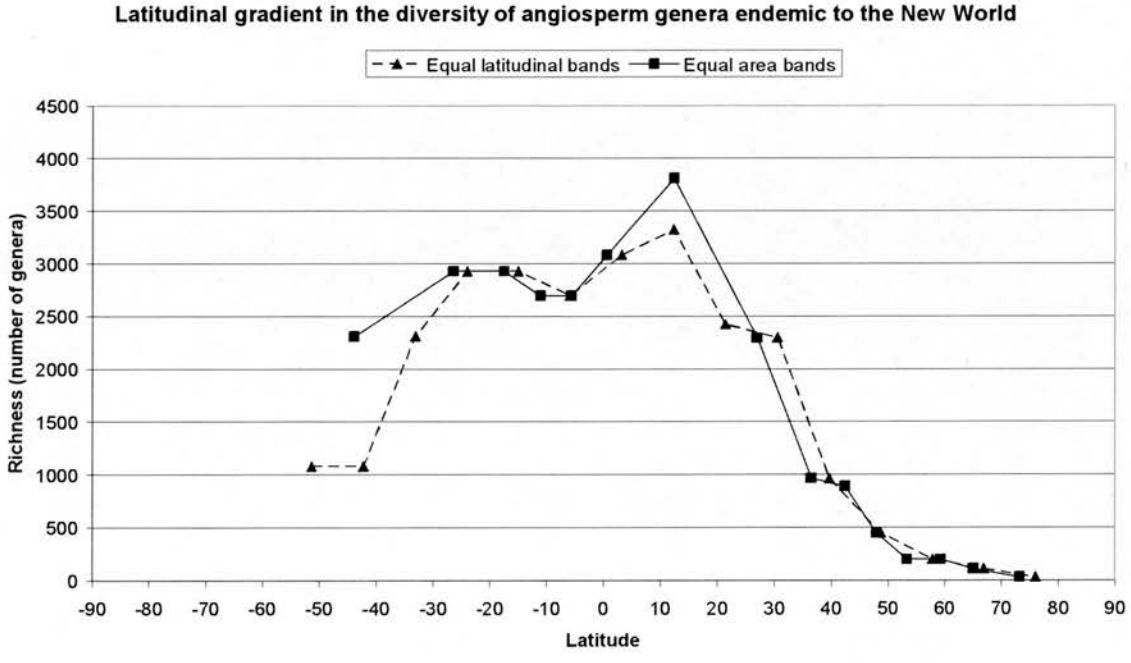


Figure 4.8 Using either equal-area or equal-latitude bands has little effect on patterns of genus richness even for the hourglass-shaped New World. Negative latitudes denote the southern hemisphere.

4.2.5 Discussion

The results presented in Table 4.2 seem to give moderate or perhaps equivocal support for the geographic area hypothesis: a strong relationship between principally extra tropical diversity and land area is only marginally improved by the addition of latitude in a multiple regression. Unfortunately, however, the analysis as outlined by Blackburn and Gaston (1997) does not quite conform to the thought experiment imagined by Rosenzweig (1992). In his original paper, Rosenzweig (1992) defined as being 'tropical' any taxon which crossed into the tropical region for any part of its range (i.e. latitudinal range endpoints <26° [his cut-off point]). Under this definition only strictly extra-tropical taxa (those with latitudinal range endpoints >26°) remain and it is these with which Rosenzweig was predicting a strong relationship with available area. However, strictly extra-tropical taxa may cause an artefactual decline in diversity towards the tropics as discussed by Blackburn and Gaston (1997) and above, since taxa with midpoints closer to the cut-off (26°) are more likely to be excluded from the analysis and so bias the analysis against finding support for the geographic area hypothesis, and therefore the analysis presented here only uses those principally extra-tropical taxa. An element of circularity is thus unavoidably introduced since Rosenzweig's original test of the area model cannot be analysed confidently without the possibility of introducing statistical artefacts, yet the more rigorous statistical test does not conform to Rosenzweig's original idea.

Assuming, however, that principally extra-tropical genera can be used as a test of the area model, then the results presented here give moderate support for the idea. Ruggiero (1999), in an analysis of the geographic distributions of South American mammal species, also found moderate support for the geographic area hypothesis, but focusing, however, on the geographic area of biomes rather than just geographic area *per se*. The species density pattern (values of species richness estimated from equal-area samples) in mammal species as a whole was found to support the geographic area hypothesis after excluding tropical species from the analysis in the same manner as did Blackburn & Gaston (1997). The area of the biome was found to be the best predictor of the number of principally extra-tropical mammal species per unit area, even after the effect of latitude and differences in mean net primary productivity of the different macrohabitats had been controlled for. Ruggiero (1999) concludes, however, that the ability to capture the impact of the spillover effect of tropical species into extra-tropical latitudes might depend on the scale of habitat analysis. It is certainly true from this study that the scale of the analysis has affected the results; such large regions as used here conflate generic richness values into areas of great size and latitudinal amplitude, and this makes discriminating between the effects of area and the effects of latitude very difficult. However, a more explicit study of the variation in range size with latitude is provided in the next section.

4.3 Rapoport's Rule

4.3.1 Introduction

Stevens, (1989) drew attention to a phenomenon first encapsulated by E.H. Rapoport (1982), that species' range sizes increase with latitude, i.e. tropical species have smaller ranges than do temperate species. On closer study of Rapoport's *Areography*, this contention is seen to rest on analyses of: the latitude of the subspecies with the largest range relative to the latitude of the type subspecies for 197 mammal species from Northern America and 33 bird species from Eurasia; and the relative numbers of 'micro-areal' species (i.e. those with restricted range, defined by Rapoport as less than the first quartile of the frequency distribution of range sizes) at different latitudes for the same Northern American mammals and for 87 species of passerine birds from Africa and 211 bird species from South America. In all cases, total range size decreased with latitude.

Stevens (1989), however, narrowed the idea of range size specifically to latitudinal range size (maximum latitudinal extent of a species):

> *"When the latitudinal extent of the geographic range of organisms occurring at a given latitude is plotted against latitude, a simple positive correlation is found"*
>
> G.C. Stevens, 1989.

and he termed this correlation 'Rapoport's Rule', making an explicit comparison between this newly-discovered ecological pattern and more venerable evolutionary counterparts such as Cope's Rule (that body-size increases over evolutionary time within a lineage) or Bergmann's Rule (that body-size increases with decreasing temperature within widespread warm-blooded species) (Gaston *et al.*, 1998). Later, this idea was extended to any geographical gradient, such as altitude (Stevens, 1992) or depth (Stevens, 1996):

> *"There exists a correlation between the mean geographic breadth of taxa occurring at a particular point along a biogeographic gradient and the relative position of the point along the gradient"*
>
> G.C. Stevens, 1996.

Stevens also went further, however, claiming that Rapoport's Rule also helped to explain the latitudinal gradient in diversity. The argument is as follows: temperate taxa, with larger latitudinal ranges, are subject to a greater range of environmental and climatic conditions than are tropical taxa, and thus the narrower geographic ranges shown by tropical species implies that they also tolerate a much narrower range of climatic conditions (Stevens, 1989). Stevens described the tropics as "a finer

mosaic of distinctive microclimates", with the consequence that species' dispersal may often extend beyond the favourable microclimate into unfavourable areas. This would increase the richness of a given area through continual input of propagules from areas of favourable microclimate into areas where the species can survive but cannot maintain their populations, a process identified by Shmida & Wilson and termed the 'mass effect' (Shmida & Wilson, 1985). As in island biogeography theory, the persistence of enhanced species richness therefore depends not on the favourableness of local conditions for each population but on the distance of these populations from favourable source areas; Stevens termed this phenomenon the 'rescue effect' (Stevens, 1989).

In effect, the greater species richness of the tropics is explained as the result of prolonging the coexistence of species by reducing the effects of competition: species which would otherwise drive other species to competitive exclusion are prevented from doing so because populations of species may be maintained through continual dispersal from stable populations in favourable areas (Stevens, 1989). Toward the equator, where ranges are smaller, this process is more prevalent because the ratio of 'rescue effect' area to geographic range would increase as geographic ranges decrease. The latitudinal gradient of diversity is then seen to be a by-product of Rapoport's Rule and the rescue effect (Stevens, 1989; Gaston *et al.*, 1998). Stevens' initial publication attracted a great deal of interest among ecologists, and Rapoport's Rule has since been the subject of a large number of studies evaluating its presence in a variety of taxonomic groups. However, nearly all of these have been on animals, and almost all of the studies to date have focused at the species level, rather than also addressing the phenomenon at genus or family level (Gaston *et al.*, 1998). Also, given that both Rapoport's Rule and the rescue effect are eminently testable phenomena, it is very evident that it has been the former phenomenon, rather than the latter, which has received all the attention (Taylor & Gaines, 1999), presumably due to the ready availability of existing taxonomic datasets and the difficulties involved in the large amounts of tropical fieldwork necessary to substantiate the 'rescue effect'. Given the lack of studies of Rapoport's Rule right across the world, and also on plants and at higher taxonomic levels, it was felt desirable to test for Rapoport's Rule with this comprehensive dataset of distributions of plant genera.

Rohde (Rohde *et al.*, 1993; Rohde & Heap, 1996) pointed out two fundamental objections to Stevens' (and, by implication, Rapoport's) original formulation of the phenomenon. Firstly and most importantly, the ranges used to calculate mean latitudinal range size at any particular latitudinal band will also overlap in many other latitudinal bands, which means that the mean latitudinal range size of every species within any particular latitudinal band will include a majority of ranges which also occur in other latitudinal bands, and so data points will not be spatially independent of each other (Rohde *et al.*, 1993). Rohde *et al.* (1993) instead proposed their 'midpoint' method, where mean latitudinal ranges are calculated only for those taxa whose midpoint falls within that latitudinal band (see also Blackburn & Gaston, 1996). Each taxon is thus counted only once and data points become spatially independent. The second objection with Rapoport's Rule is that longitudinal extent of available land

area at different latitudes may be causing range size to co-vary with latitude; this was a particular concern since both Rapoport and Stevens first enumerated the pattern partly over the latitudinal gradient across North America, where the shape of the continent also becomes much smaller towards the tropics. Roy *et al.* (1994) also criticised the use of mean latitudinal ranges, since range size distributions are often strongly skewed towards small ranges, and so also presented median and modal range sizes.

4.3.2 Methodology

Rapoport's Rule was studied with both the Stevens' method and the 'midpoint' method. To explore this discrepancy between the two methods proposed for studying Rapoport's Rule further, the same analyses were carried out using taxa endemic to the New World, to the Old World, to Europe & Africa and to Asia & Australasia. The composition of the geographical subdivisions is as follows: New World: Regions 70 – 85; Old World: Regions 10 – 51; Europe & Africa: Regions 10 – 29; Asia & Australasia: Regions 30 – 51. Each subdivision thus constitutes a separate latitudinal gradient. Taxa need to be endemic to each region studied to be sure of measuring true latitudinal ranges, since non-endemic taxa might otherwise exceed the minimum or maximum latitude outwith the region; for this reason the larger area of the subdivided regions is advantageous, since larger areas have a greater percentage of generic endemism. Results of Stevens' plots of mean latitudinal range of all taxa within a latitudinal band were systematically compared with those of 'midpoint' plots by using a Spearman non-parametric rank correlation of range size against absolute latitude (i.e. northern and southern hemispheres were not distinguished between). Though statistical independence of data points is achieved using the 'midpoint' method, so that linear regression might be a more appropriate technique (Fleishman *et al.*, 1998), using Spearman's correlation coefficients means results can be compared directly with those from the Stevens' method, where regression would not be statistically valid. These results are presented in Table 4.3 below. Note that, although Table 4.3 includes the total number of genera and the percentage generic endemism for each region, correlations were only performed between latitude of bins and mean latitudinal ranges for each bin (Fleishman *et al.*, 1998), with a bin-range of 30 bins. Mean, median and modal range sizes were also calculated for Rohde's 'midpoint method' for each of the geographical subdivisions, and these results are presented in Table 4.4 below.

4.3.3 Results

Figure 4.9 shows the mean latitudinal range size for all genera within each latitudinal band – a so-called Stevens' plot (Rohde *et al.*, 1993; Colwell & Lees, 2000). Mean latitudinal range sizes show a general decline through the temperate regions and are conspicuously smaller in the tropical regions – apparently supporting Rapoport's Rule. Using the original Stevens' method, the smooth decline in mean latitudinal range sizes is very apparent across the northern hemisphere (see Figure 4.9; c.f. Figures 1 – 5 in Stevens, 1989). Figure 4.10 shows the same data as in Figure 4.9 but plotted

using Rohde's 'midpoint method'. In contrast to Figure 4.9, using this method mean latitudinal range size actually peaks to the north of the equator and does not show any general decline towards the tropics. This contradicts the claim that, if Rapoport's Rule is supported by one method, then it should be shown by the other (Gaston *et al.*, 1998).

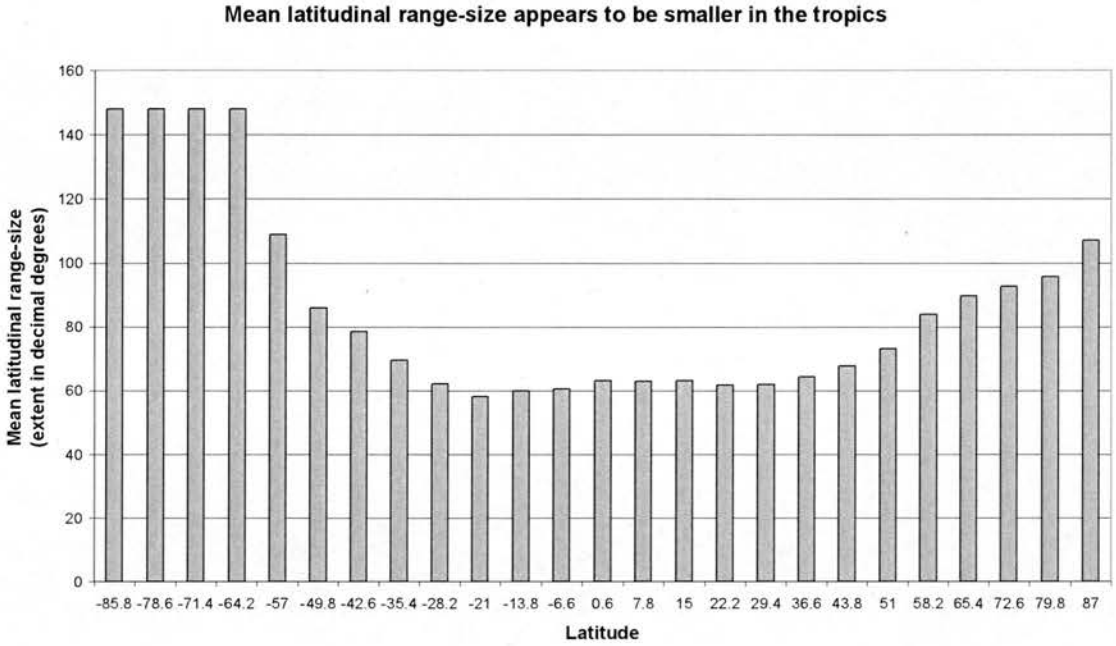


Figure 4.9 Stevens' plot of mean latitudinal range-size against latitude. Negative latitudes denote the Southern Hemisphere.

Using the Stevens' method of the mean of total range sizes there is a positive correlation between latitude and range size for every geographical subdivision, although one of these is not significant and only two might be argued to be 'strong' (see Table 4.3). Using the midpoint method, however, there is a negative correlation in every case, although only three are significant, and coefficients are at best moderate (see Table 4.3). Despite the generally poor correlations found using the midpoint method, clearly this method provides no 'simple positive correlation' between range size and latitude, and so no support for 'Rapoport's Rule'. When using the midpoint method, however, a decline in mean (and median and modal) range sizes is apparent from 50°N to 35°N for the New World, the general latitudinal range studied by Stevens (1989), although the general tendency is for range sizes to be greater in the tropics (see Figure 4.11).

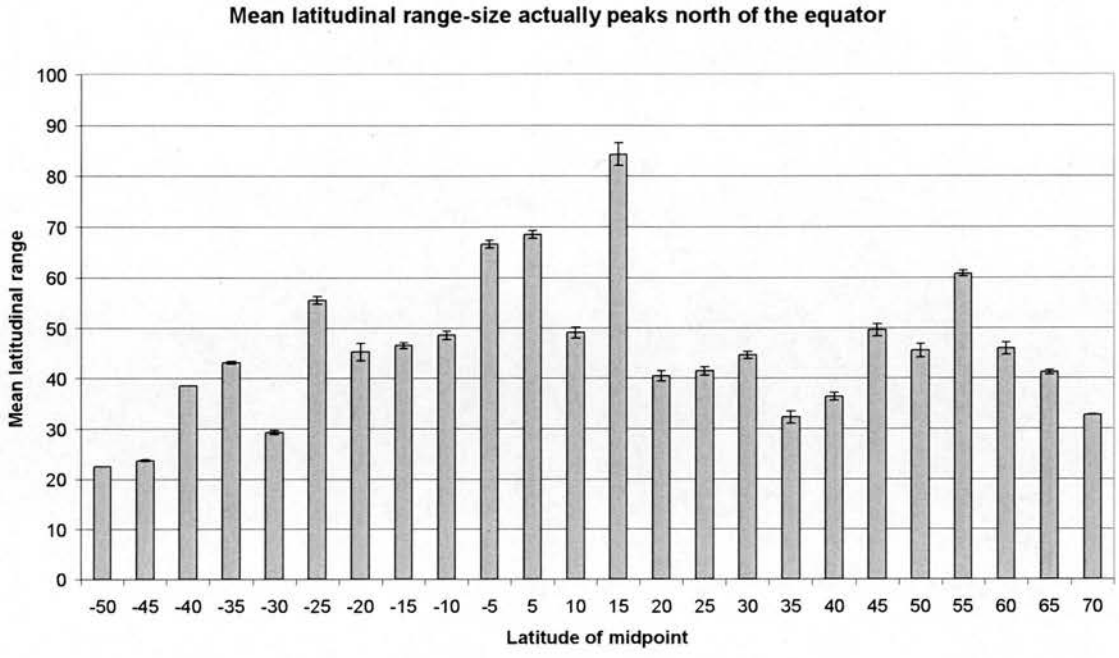


Figure 4.10 'Midpoint method' plot of mean latitudinal range-size against latitude. Negative latitudes denote the Southern Hemisphere; error bars ± 1 standard error of the mean.

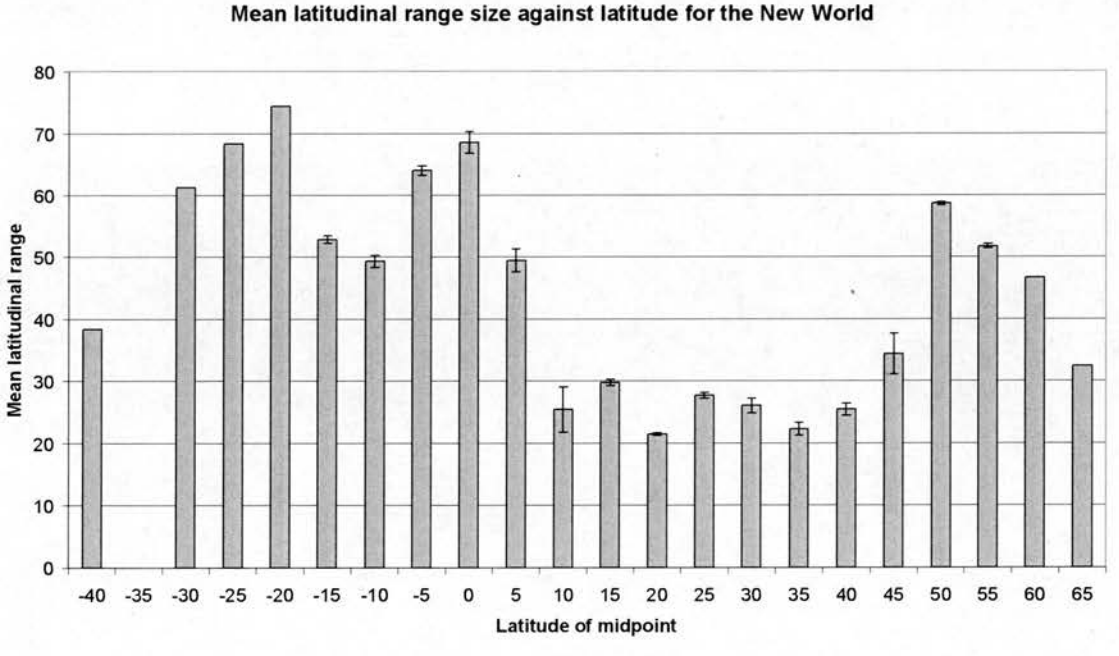


Figure 4.11 'Midpoint method' plot of mean latitudinal range-size against latitude for taxa endemic to the New World. Negative latitudes denote the southern hemisphere; error bars ± 1 standard error of the mean. As shown by Stevens (1989), mean latitudinal range-sizes decrease from 50°N to 35°N in the New World.

Spearman's rank correlation (rs)	No. of genera (% endemism)	Stevens' method	Midpoint method
World	13121 (100%)	0.702; $p > 0.005$	-0.499; $p > 0.005$
Old World	6845 (80%)	0.664; $p > 0.005$	-0.323; $p > 0.1$
New World	4260 (76%)	0.549; $p > 0.005$	-0.229; n.s.
Europe & Africa	2540 (54%)	0.247; n.s.	-0.113; n.s.
Asia & Australasia	3193 (54%)	0.457; $p > 0.01$	-0.368; $p > 0.05$

Table 4.3 Spearman's non-parametric correlation coefficients for mean range-size against latitudinal bins (mean bin size 4.82°) using various subdivisions of the world, comparing results of the Stevens' method of mean total range-sizes for all taxa within a latitudinal band against Rohde's 'midpoint method' of mean range-sizes only for taxa with midpoints falling within a latitudinal band.

Maximum range	Mean	Median	Mode
World	10°N – 15°N	10°N – 15°N	10°N – 15°N
Old World	20°N – 25°N	20°N – 25°N	40°N – 45°N
New World	25°S – 20°S	25°S – 20°S	25°S – 20°S
Europe & Africa	45°N – 50°N	45°N – 50°N	45°N – 50°N
Asia & Australasia	5°S – 0°	5°S – 0°	5°S – 0°

Table 4.4 Positions of greatest mean, median and modal latitudinal range-sizes for the whole world, for taxa endemic to the New World, to the Old World, to Europe & Africa and to Asia & Australasia.

The greatest latitudinal ranges using three separate measures of central tendency for the World all fell within the same latitudinal band: 10 – 15 degrees north (see Table 4.4 above) – just south of the peak in latitudinal range sizes found in New World birds at 17°N (Blackburn & Gaston, 1996). If the data in Figure 4.10 is plotted in single-degree latitudinal bins, the exact peak in range size falls on 14°N (figure not shown). None of Figures 4.9 – 4.11, however, show a uniform trend; all have principal and secondary peaks (see Figure 4.9, for example). The peak in modal range size for the Old World at 40°N – 45°N represents the secondary peak in Old World mean and median range sizes. Figure 4.12 re-plots the raw data of range size versus position of midpoint as a single scatter-plot, with northern and southern hemispheres combined. The peak in latitudinal range size at approximately 14° becomes more apparent, but what is more obvious is the huge degree of scatter within the data: there are many genera with small ranges at all latitudes, but the mean latitudinal range size gradually decreases towards the poles.

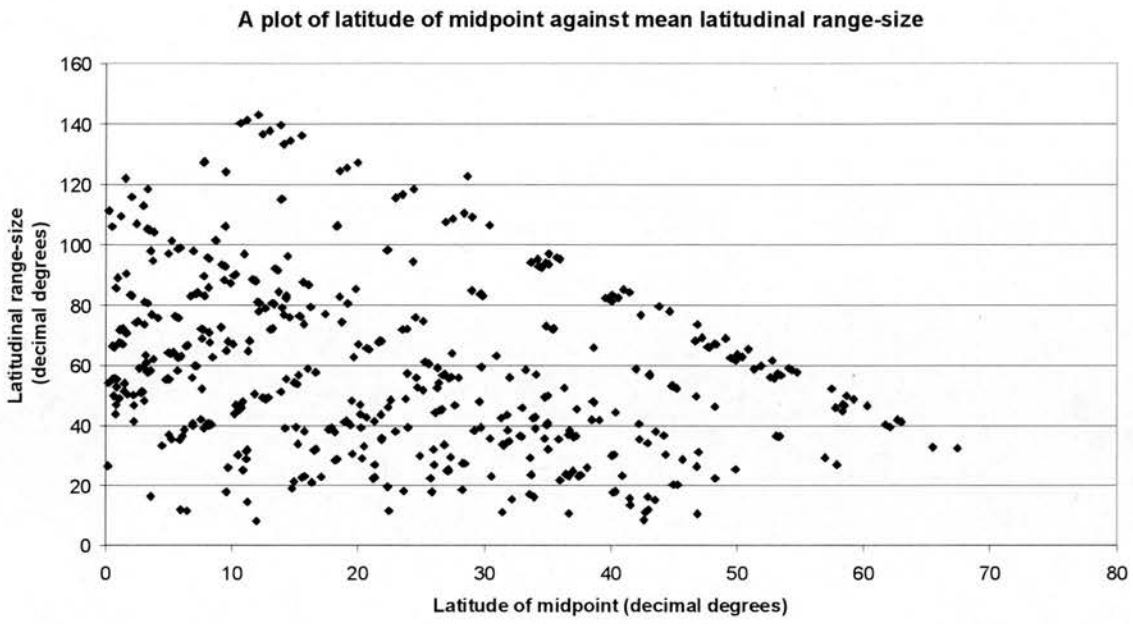


Figure 4.12 A plot of latitude of midpoint against mean latitudinal range-size for genera shows considerable scatter. Note that northern and southern hemispheres are not differentiated here.

The midpoint of the Earth's latitude is obviously the equator, so that cannot be causing the peak in latitudinal range-size at 14°N – the midpoint in the largest mean ranges does not coincide with the midpoint of the greatest available latitude. Neither can it be due to the latitude of the maximum longitudinal extent, since from Figure 4.4 it can be seen that there are two almost-equal-sized peaks, one at 31°N and one at 48°N. However, it is possible to produce a measure incorporating both latitudinal extent and longitudinal extent – the weighted centroid, which may be thought of as the point exactly in the middle of the available land area: if a map of the world's land were hung on its weighted centroid, the map would be perfectly balanced about this point. The weighted centroid for the Earth was calculated on a Peters' equal area cylindrical projection of the world using ArcView GIS's "convert polygons to centroids" function, and the latitude of the Earth's weighted centroid turned out to be at exactly 14°N also. This was then done for each of the other geographical subdivisions listed in Table 4.4, as well as comparing each with the median latitude and the latitude of the greatest longitudinal extent for that region (see Table 4.5).

Midpoint plots for all five geographical subdivisions of the world are given in Figure 4.13. In no other case did the latitude of the centroid coincide with the latitude of the maximum mean latitudinal range (see Table 4.5). However, the peak in range-size for the Old World does coincide with the latitude of the maximum longitudinal extent for the Old World at approximately 24°N, while the median latitude coincides with the second peak in mean latitudinal range-size for the whole world and the latitude of maximum longitudinal extent coincides with the second peak in mean latitudinal range-size for the New World (see Figure 4.13). Each continent was also treated separately to find its maximum mean latitudinal range and compare this with its centroid, but even at this scale of

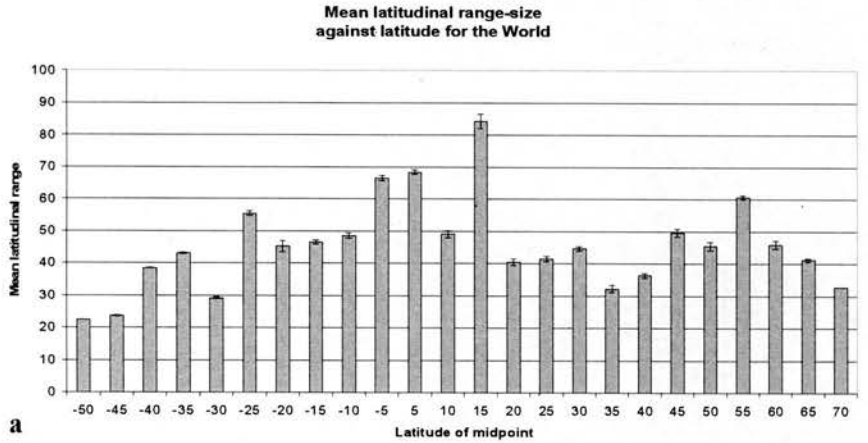
geographical resolution it was found that the low number of polygons within each continent caused only artefacts to appear in the data. For example, South America only contains four regions: 82, Northern South America; 83, Western South America; 84, Brazil; and 85, Southern South America. The histogram of mean latitudinal range-size against latitude of midpoint for South America is dominated by a peak at 23°S, but this is merely the latitudinal midpoint of the continent, and so the midpoint of all the genera found in regions 82, 83, 84 & 85.

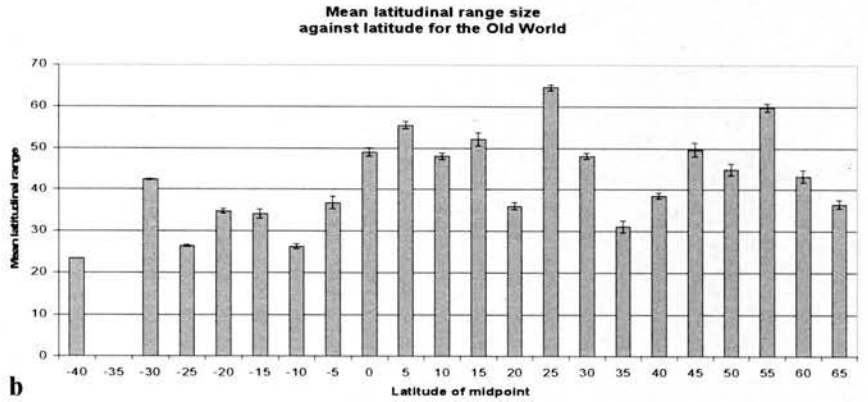
Geographical extent	Maximum mean latitudinal range	Latitude of Centroid	Median latitude	Latitude of max. longitude
World	10°N – 15°N	14°N	0°	65°N
Old World	20°N – 25°N	43°N	15°N	24°N
New World	25°S – 20°S	35°N	14°N	5°S
Europe & Africa	45°N – 50°N	24°N	18°N	10°N
Asia & Australasia	5°S – 0°	36°N	15°N	40°N

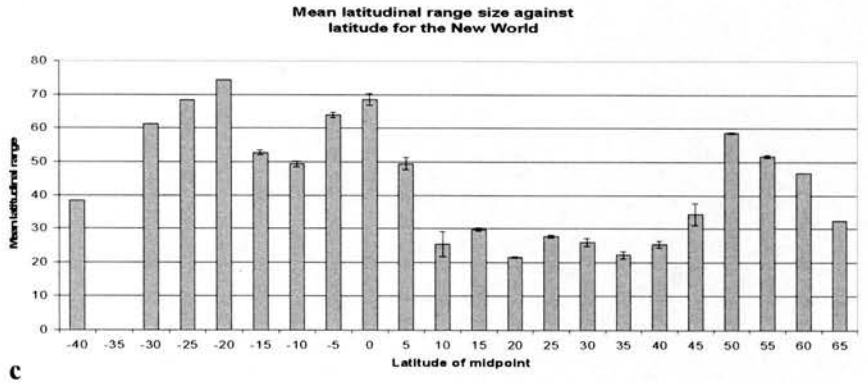
Table 4.5 Latitude of maximum mean latitudinal range-size (within 5-degree latitudinal bands) compared against latitudes of three separate measures of geographic meridians (to the nearest degree), for various geographical subdivisions of the world (see also Figure 4.13).

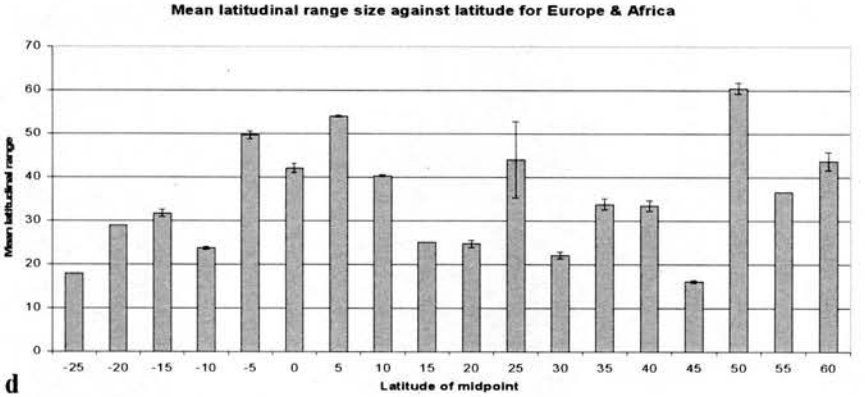
4.3.4 Discussion

It would be fair to say that in the twelve or so years since Stevens' first publication identfying the phenomenon (Stevens, 1989), Rapoport's Rule has lived fast, but died young: despite the considerable initial interest, it is now generally regarded as an artefact of the methods used to estimate it. The original formulation of Rapoport's Rule (Stevens, 1989) has been criticised for its lack of geographical breadth, despite claiming to be a solution to the latitudinal gradient of diversity (Rohde & Heap, 1996). In none of the examples originally cited is the lower latitudinal limit less than 25°N; that is, latitudinal range sizes are not shown even entering the tropics, let alone across it (Stevens, 1989). With marine molluscs, which were found not to support Rapoport's Rule despite being cited as such by Stevens (Stevens, 1989), species' ranges in the Pacific were limited by the boundaries of biogeographic regions, rather than explicitly by latitude (Roy *et al.*, 1994). Substantiating this idea, the northern limit of the Neotropical regions has also, but not unexpectedly, been found to be a region of very high faunal turnover (Blackburn & Gaston, 1996), i.e. not many species cross it. As Stevens' examples did not actually extend to tropical latitudes it is not possible to see this pattern in his initial paper (Stevens, 1989), but the cause of the second peak at 0° for the New World (see Figure 4.11) may well be that tropical genera reach their northern distribution limits at about 15°N, while extending far south into South America.









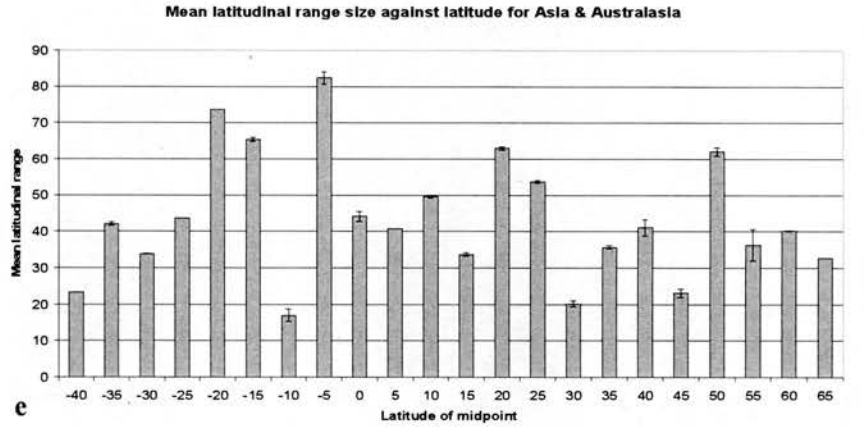


Figure 4.13 Mean latitudinal range plotted against latitude of midpoint for each of: **a)** genera of the whole world; **b)** genera endemic to the Old World; **c)** genera endemic to the New World; **d)** genera endemic to Europe & Africa **e)** genera endemic to Asia & Australasia. See text for fuller discussion. Negative latitudes denote the Southern Hemisphere; error bars ± 1 standard error of the mean.

Considering all the above, it thus seems as if the generality of Rapoport's Rule cannot be substantiated, and it is merely a 'local phenomenon' (Rohde & Heap, 1996) which cannot then be used to explain the latitudinal gradient of diversity. In fact, results of the midpoint method (see Figure 4.10) show that in general range sizes actually increased towards the equator and were smaller towards the poles. The contention that 'in most cases if a relationship between range size and latitude is documented using Stevens' method it is also documented using the midpoint method' (Gaston *et al.*, 1998) is not supported here, although the claim that the spatial non-independence of the Stevens' method gives a stronger relationship (Gaston *et al.*, 1998) is. The most plausible explanation for this is that ranges are indeed co-varying in size with available area, and the narrowing of North America towards the Isthmus of Panama is causing the decline in latitudinal range sizes over these latitudes. Are the secondary peaks within Figure 4.10 then geographically significant, or are they just artefacts of the data? Could the shape of the continents be influencing the maximum range sizes?

The midpoint of the Earth's latitude is obviously the equator, so that cannot be causing the peak in latitudinal range size at 14°N – the midpoint in the largest mean ranges does not coincide with the midpoint of the greatest available latitude. Neither can it be due to the latitude of the maximum longitudinal extent, since from Figure 4.4 it can be seen that there are two almost-equal-sized peaks, one at 31°N and one at 48°N. However, it is possible to produce a measure incorporating both latitudinal extent and longitudinal extent – the weighted centroid, which may be thought of as the point exactly in the middle of the available land area: if a map of the world's land were hung on its weighted centroid, the map would be perfectly balanced about this point. The weighted centroid for the Earth was calculated on a Peters' equal area cylindrical projection of the world using ArcView GIS's "convert polygons to centroids" function, and the latitude of the Earth's weighted centroid turned out to be at exactly 14°N also. This was then done for each of the other geographical subdivisions listed in Table 4.4, as well as comparing each with the median latitude and the latitude of the greatest longitudinal extent for that region (see Table 4.5).

In no other case did the latitude of the centroid coincide with the latitude of the maximum mean latitudinal range (see Table 4.5). However, the peak in range size for the Old World does conincide with the latitude of the maximum longitudinal extent for the Old World at approximately 24°N, while the median latitude coincides with the second peak in mean latitudinal range size for the whole world and the latitude of maximum longitudinal extent coincides with the second peak in mean latitudinal range size for the New World (see Figure 4.12). Each continent was also treated separately to find its maximum mean latitudinal range and compare this with its centroid, but even at this scale of geographical resolution it was found that the low number of polygons within each continent caused only artefacts to appear in the data. For example, South America only contains four regions: 82, Northern South America; 83, Western South

America; 84, Brazil; and 85, Southern South America. The histogram of mean latitudinal range size against latitude of midpoint for South America is dominated by a peak at 23°S, but this is merely the latitudinal midpoint of the continent, and so the midpoint of all the genera found in regions 82, 83, 84 & 85. However, results from the analysis of Rapoport's Rule do suggest that continental geography is influencing the range sizes and hence the diversity of taxa within continents. This idea is explored further in the following section.

The 'mid-domain' effect

4.4.1 Introduction

The 'mid-domain' effect is a product of a geometric null model which predicts a peak in diversity in the middle of a one-dimensional domain, purely through geographic boundary constraints (Colwell & Hurtt, 1994; Colwell & Lees, 2000). It is defined more precisely as 'the increasing overlap of species ranges towards the centre of a shared geographic domain due to geometric boundary constraints in relation to the distribution of species' range sizes and midpoints' (Colwell & Lees, 2000). Imagine a random set of distributions each occupying a space along any line, such that the extent of the distribution and the position of its midpoint can be measured relative to the length of the line. If a member of this set of distributions is found at only one point along the line, then that point may be positioned anywhere from either end to the exact middle; however, if a member of this set of distributions has an extent equal to the length of the line, then the midpoint of that distribution must be in the middle of the line. Since no distribution can overlap the ends of the line, but may overlap with other distributions, by chance there will be a greater number of overlapping distributions near the middle of the line than at either end. This is the essence of the 'mid-domain' effect – since species' distributions are bounded by geographical barriers, there will inevitably be a peak of diversity at points roughly midway between those barriers.

Initially, geometric null models were proposed as a test of Rapoport's Rule for the explanation of the latitudinal gradient of diversity (Colwell & Hurtt, 1994), which was thought to be a spurious phenomenon. In an effort to model possible patterns of distribution which could account for Rapoport's Rule, it was discovered that randomly placed geographic ranges would always produce a peak of diversity towards the middle of the domain being modelled (Colwell *et al.*, 2004). The 'domain' for the latitudinal gradient of diversity is then the World, since no species can be found north of the North Pole or south of the South Pole, and the middle of this domain is the tropical regions. The 'mid-domain' effect thus predicts greater diversity within tropical regions, simply because latitudinal ranges overlap more in this region. It is in this context that the idea is applied here. Interestingly, in his discussion of the latitudinal increase in range-size, later to be called Rapoport's Rule (Stevens, 1989), Rapoport (1982) first dismisses and then hints at the idea of geometric constraints on range-sizes himself:

> *"The 'latitudinal effect'* [of increasing range-size], *however, can be a deceiving concept because species are not influenced by parallels or meridians but by temperature, solar radiation, and other climatic parameters"* E.H. Rapoport, 1982, page 160

" ... the fact is that once again we get the idea that the smallness or 'micro-arealism' of species can somehow be determined by the size of the continent."

E.H. Rapoport, 1982, page 162

" ... although the width of the continent does not seem to influence the extension of the species and subspecies, maybe this is not valid for the case of micro-areal species."

E.H. Rapoport, 1982, page 164

He did not, however, hypothesise any correlation between increased diversity and the geometric constraints on range-size, the central thesis of the mid-domain effect. The mid-domain effect has since been discovered in: elevational gradients of birds in South America (Rahbek, 1997); and bathymetric gradients of gastropods and polycheates in the northwest Atlantic (Pineda & Caswell, 1998); latitudinal gradients of bats and marsupials of the New World (Willig & Lyons, 1998); latitudinal and longitudinal gradients of animals in Madagascar (Lees *et al.*, 1999); latitudinal and longitudinal gradients of birds in Africa (Jetz & Rahbek, 2001, 2002); and elevational gradients of plants in Nepal (Grytnes & Vetaas, 2002). The size of the mid-domain effect depends on the frequency distribution of range sizes in any sample: since widespread taxa must by definition occur over much of the domain, the greater is the number of widespread species in a sample, the greater the extent of the mid-domain effect will be; conversely, the greater is the number of local species, the smaller the mid-domain effect will be. Originally the mid-domain effect was proposed along a single 1-dimensional gradient, although in principal it is equally applicable to areas (two dimensions) and volumes (three dimensions). However, it is a major criticism of the concept that, as yet, most of the several proposed models have only applied to a single dimension (Hawkins & Diniz-Filho, 2002; Zapata *et al.*, 2003; Colwell *et al.*, 2004).

The simplicity and power of this idea may mean that biogeographers and ecologists have had the wrong null model in their minds by assuming that, were there no climatic, physical or biological gradients, species richness would be the same at all latitudes, elevations and depths. Purely geometric constraints suggest that this cannot be so: there will inevitably be some peak in diversity at midpoints along any dimension. As such, the mid-domain effect not only serves as a partial explanation of diversity patterns but, importantly, as a null model against which to measure deviations from the expected results. Departure from the expected richness peak, under an appropriate null model, but not the peak itself, requires biological or historical explanation at geographic scales (Colwell & Lees, 2000).

4.4.2 Methodology

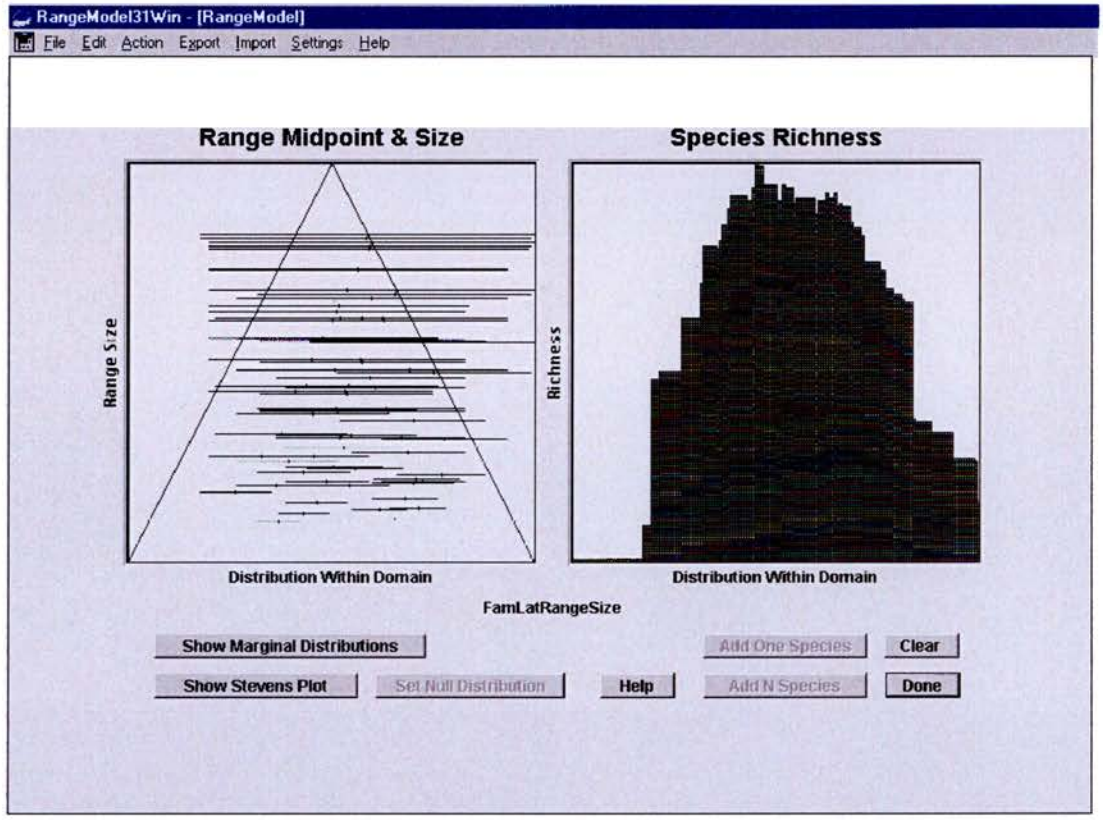


Figure 4.14 Screen-shot of RangeModel software showing the mid-domain effect, in the right-hand window, produced by the overlapping ranges of the taxa shown in the left-hand window. Data are entered as paired range-sizes and midpoints for each taxon and the relationship between these parameters is evident in the left-hand window, where for a given range-size (range-size increases towards the top of the window), the associated midpoint must fall within the triangle.

The mid-domain effect was explored as a possible explanation of the latitudinal gradient of diversity with the RangeModel simulation software (Colwell, 2000) (see Figure 4.14). This package implements up to 5 different geometric null models, based on the publications of Colwell and Hurtt (1994) and Colwell and Lees (2000). In the left-hand window of Figure 4.14 an array of taxa is represented each by their respective ranges, from the widest range at the top of the window to the smallest range at the bottom; the midpoint of the range must fall within the triangle shown in the left-hand window. For large ranges, therefore, the midpoint of the range must fall approximately within the middle of the domain, since the whole range is constrained to lie within the domain limits, whereas taxa with small

ranges can be found almost anywhere within the domain. Since any taxon with a range more than half as large as the limits of the domain must by definition cross the midpoint of the domain, for collections of taxa showing a high proportion of large ranges there will inevitably be a peak in diversity towards the middle of the domain. This is shown in the right-hand window of Figure 4.14, where the ranges of individual taxa overlap to give a count of diversity which can be measured at several points to show the changes in richness across the domain.

Data is imported simply as paired columns of midpoints and range-sizes representing each distribution. Models 1 – 3 of RangeModel were not assessed, as these employ a theoretical range size frequency distribution (RSFD) which may introduce unrealistic biological assumptions into the analysis (Colwell *et al.*, 2004), for example that the maximum range of the RSFD must equal the maximum dimension of the domain (Colwell & Lees, 2000). Comparing results from randomly resampling a theoretical RSFD with the observed pattern of taxonomic richness across the domain will therefore introduce unnecessary artefacts into the analysis (Koleff & Gaston, 2000; Colwell *et al.*, 2004). Instead, if an empirical RSFD is available for a group of taxa, this should be resampled in preference to an empirical model (Colwell *et al.*, 2004). Two models were assessed: the first (Model 4 of RangeModel) takes the empirical ranges from the inputted data and randomly assigns midpoints to each range, under the constraint that the range with the randomly-assigned midpoint must still fall entirely within the domain; the second model (Model 5 of RangeModel) takes the empirical midpoint from the inputted data, and randomly assigns range sizes to these, again under the constraint that the entire range must still fall within the domain. Area across the longitudinal dimension of the domain was accounted for using the method described in Chapter 3, rescaling the empirical richness values by the area of that latitudinal band using the power law species-area relationship and assuming $z = 0.14$. Model 5 has been criticised by Koleff & Gaston (2000) for being too closely constrained by the empirical data, a criticism accepted by Colwell *et al.* (2004). If an empirical peak in richness lies towards the middle of the domain, then empirical midpoints will also be clustered towards the middle of the domain, and so by keeping empirical midpoints a mid-domain peak in simulated richness is bound to occur, no matter how wide or narrow the randomisation process makes the ranges of taxa around those empirical midpoints.

The data used in this thesis is highly structured, since ranges can only be estimated to the boundaries of regions, rather than their actual latitudinal extent. Numerous genera which in reality have range limits at different latitudes will therefore appear to be of the same extent for the purposes of this analysis (i.e. the northernmost or southernmost region in which they are found is the same). When these ranges are randomised, however, range endpoints may then fall anywhere within the domain and are no longer constrained to falling within hard boundaries, other than within the limits of the whole domain. The

empirical data thus has a much more pronounced step-like increase in diversity towards the equator than does the randomised data, which will reduce the possible strength of the relationship between them. It might perhaps be more appropriate to use bins in the randomisations which equal the size and positions of actual TDWG regions; however, the option to vary bin sizes (rather than simply number of bins) within a single randomisation was not available with the RangeModel software. A further complication of doing this would be that larger bins, which would correspond in position to larger TDWG regions, would also be more diverse simply because they are larger. Latitudinal extent of bins would therefore need to be taken into account in a multiple regression model in addition to longitudinal extent of bins.

Simulations of 1000 runs were performed on several different formulations of both genus and family ranges: firstly with all of the data, with minimum and maximum latitudes determined by the most-southerly and most-northerly point of the combined perimeter of all the polygons for that taxon, as described before (see Section 4.1); then, since the mid-domain effect is disproportionately caused by widespread taxa, sensitivity analyses were performed by truncating the data set by removing the largest (most widespread) 5% (following Willig & Lyons, 1998) and 25% of both families and genera, and then, as a comparison, removing only the smallest (least widespread) 25%. The fit of observed data to that predicted by the null model was assessed with simple linear regression (see Table 4.6). However, since data points in the regression are not statistically independent (they are spatially autocorrelated), tests of significance cannot be applied to the results of the regression given here.

Simple linear regression		Empirical range, random midpoint		Empirical midpoint, random range	
		slope	r^2	slope	r^2
All Flowering Plant families	MDE alone	1.29	0.89	1.12	0.89
	MDE + area	0.47	0.68	0.42	0.62
All Flowering Plant genera	MDE alone	1.11	0.69	0.76	0.87
	MDE + area	1.11	0.83	1.05	0.91
95% of families	MDE alone	0.97	0.73	0.84	0.85
	MDE + area	0.23	0.14	0.21	0.22
95% of genera	MDE alone	1.21	0.78	1.04	0.94
	MDE + area	0.32	0.33	0.40	0.55
Largest 75% of families	MDE alone	1.09	0.75	1.15	0.83
	MDE + area	0.76	0.73	0.64	0.62
Largest 75% of genera	MDE alone	1.50	0.82	1.11	0.94
	MDE + area	1.05	0.86	1.03	0.93
Smallest 75% of families	MDE alone	1.01	0.71	1.19	0.91
	MDE + area	0.25	0.23	0.30	0.38
Smallest 75% of genera	MDE alone	1.44	0.78	0.96	0.94
	MDE + area	0.39	0.41	0.46	0.61

Table 4.6 Results of linear regression of observed against expected patterns of genus richness from two separate null models, both with and without accounting for the variation in land area with latitude [MDE = mid-domain effect].

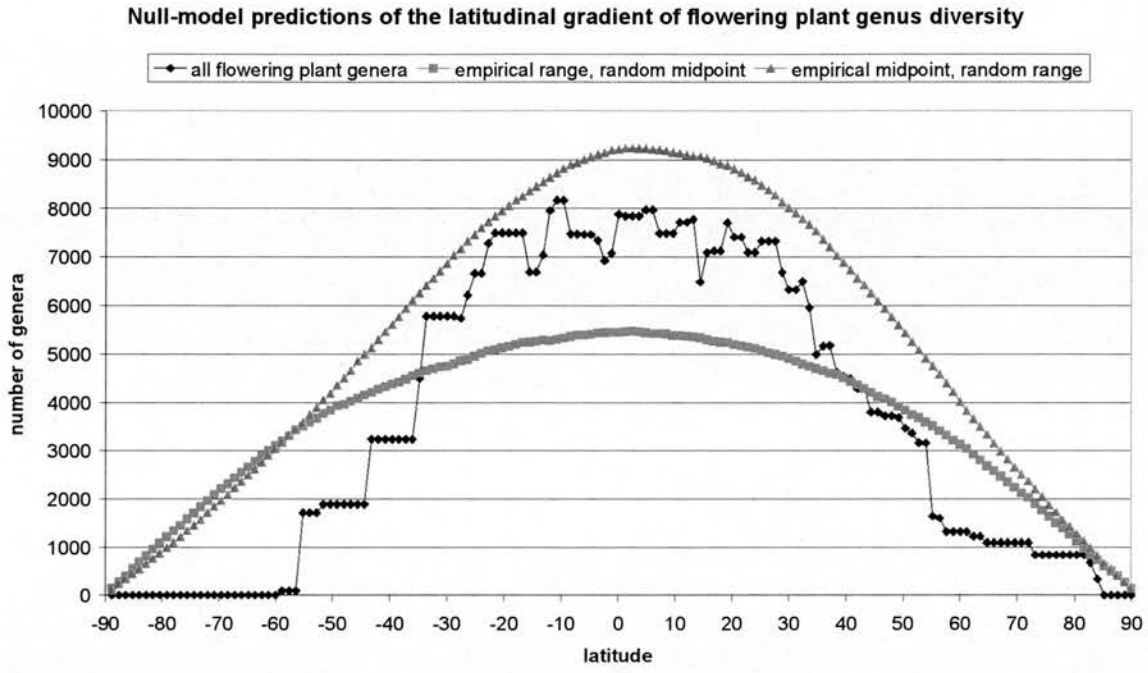


Figure 4.15 Null-model predictions of the latitudinal gradient of diversity for flowering plant genera.

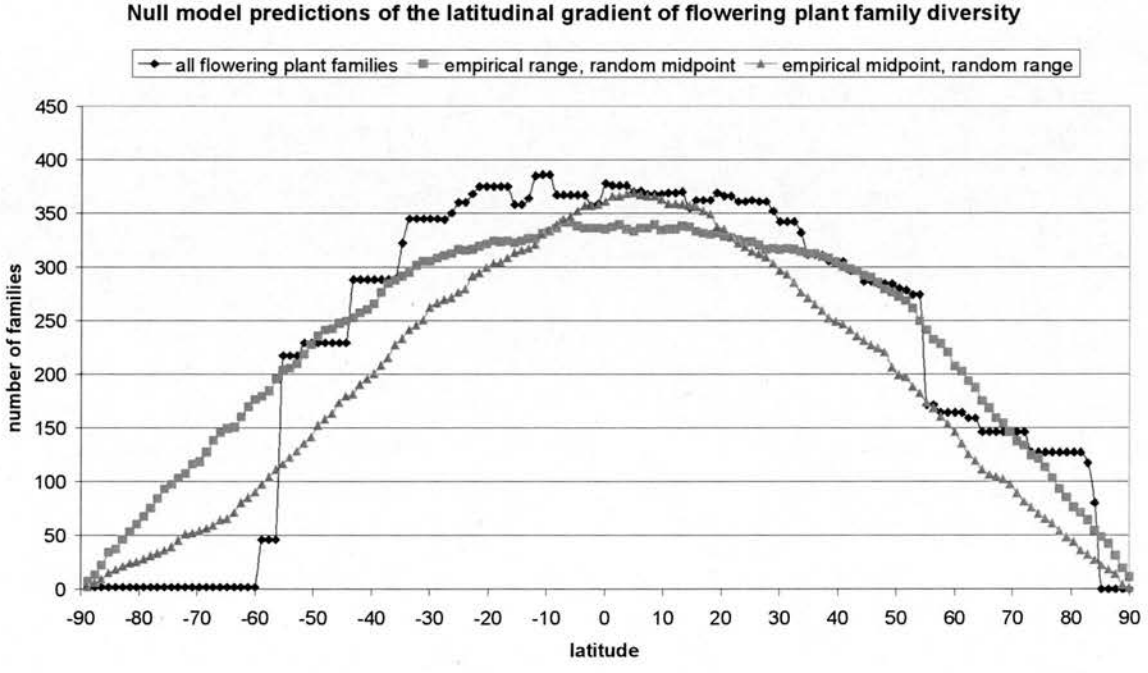


Figure 4.16 Null-model predictions of the latitudinal gradient of diversity for flowering plant families.

4.4.3 Results

In Figure 4.15 above, empirical ranges for all flowering plant genera are shown with predicted richness patterns from both null models 4 (empirical range, random midpoint) and 5 (empirical midpoint, random range) of RangeModel; results for all flowering plant families are shown in Figure 4.16. Results of the null model simulations show a strong mid-domain peak at all geographical and taxonomic scales studied. For families, the large mean latitudinal range-size may constrain the position of midpoints in the first null model; for genera, with smaller ranges, empirical midpoint positions and a greater than observed number of random small ranges cause a pronounced Rapoport effect. Production of a Rapoport effect seems to be an artefact of the data, rather than of the second null model; since regions are very large, latitudinal range-size tends to be overestimated, and also mean latitudinal range-size is greater for genera than for species. The size of the mid-domain peak in richness predicted by the two null models in Figure 4.15 differs greatly between them: model 5 (empirical midpoint, random range) predicts a tropical peak of more than 9000 genera, whereas model 4 (random midpoint, empirical range) predicts only about 5500 genera. This is despite the boundaries of the domain being the same for both null models. In fact, model 4 underestimates diversity relative to the empirical data, while model 5 overestimates diversity relative to the empirical data.

This difference might be caused by the way in which RangeModel counts overlapping ranges to give the richness scores: rather than actually summing ranges of all taxa within the entire bin, the program only counts ranges intersected at a particular point within that bin (R. Colwell, pers, comm.; Bachman *et al.*, 2004). Since numbers of endemic genera are high (38% of the total; see Chapter 3), this means that the range size frequency distribution of genera is very steep, i.e. that genera mostly have small ranges on a global scale. If empirical ranges are used with random midpoints (model 4), some of these ranges might therefore be falling between the points of intersection at which richness is summed, and so richness will be underestimated. This might explain why such a large difference in null model predictions is not evident in Figure 4.16, for family level data, since families have a much shallower range-size frequency distribution (i.e. are generally more widespread) than do genera (see Chapter 3). Since the positions of empirical midpoints are constrained by the finite number of regions in the TDWG classification (i.e. many taxa have exactly the same range midpoint), and these inevitably tend to be clustered towards the middle of the domain where richness is greater, a greater proportion of random range sizes under model 5 are less likely to fall between these points of intersection at which RangeModel sums the number of taxa to give the richness score.

Results of a linear regression between empirical observed richness across the gradient and predicted patterns of richness from the null model analysis are shown in Table 4.6. If observed and predicted richness values were identical, this would result in a slope of 1.0 with a coefficient of determination (r^2) also of 1.0. Results in Table 4.6 thus indicate the degree of deviation of predicted results from empirical results: slope values less than 1.0 imply that values of predicted richness are over-estimated relative to observed richness values, while values greater than 1.0 imply that predicted results are under-estimating observed richness; the lower the coefficient of determination, the greater the variability between the observed and the predicted richness values. Comparing purely the observed richness values with the predicted richness values from the null models, there is always a strong relationship between the two sets of values (slope values = 0.76 – 1.44; r^2 = 0.69 – 0.94); except in the case of all flowering plant families, the relationship is consistently stronger with the second null model (empirical midpoint, random range; r^2 = 0.83 – 0.94) than with the first (empirical range, random midpoint; r^2 = 0.69 – 0.89). However, given the criticism of Koleff & Gaston (2000), that the model is too closely constrained by the empirical data and a mid-domain peak in richness is inevitable, not too much confidence can be placed in the results from model 5 of RangeModel: the relationship between observed and expected diversity would be expected to be stronger under this model than under the model of empirical range and random midpoint.

However, variation in observed patterns of taxon richness across the latitudinal gradient may be at least partly due to the variation in land area with latitude, since latitudinal bands with larger land area would be expected to contain greater numbers of taxa. After accounting for area, using the methodology developed in Chapter 3 and scaling both the area-rescaled empirical richness and the predicted richness from the null model analysis onto comparable scales from 0 to 1 by dividing each set of values through by the maximum value of that set (so that r^2 values would be directly comparable with the previous results that do not account for area), a different picture emerges. In Table 4.7, the relationship between area-rescaled observed richness values and predicted richness values increased for ranges of all genera (both models) and for the largest 75% of genus ranges – i.e. omitting the smallest 25 % of ranges (model 1; empirical range, random midpoint), although for this latter case with model 2 (empirical midpoint, random range) and for the largest 75% of family ranges with model 1 the r^2 values are comparable before and after accounting for area; and the area-rescaled results are still fairly strong for all family ranges (r^2 = 0.68, model 1; r^2 = 0.62, model 2) and for the largest 75% of family ranges with model 2 (r^2 = 0.62). However, after accounting for area, the r^2 values declined considerably in all cases where large ranges were trimmed from the analyses (leaving the largest 95% or 75% of ranges); this was true for both family ranges or genus ranges, although the effect is more pronounced for family ranges.

4.4.4 Discussion

The mid-domain effect null model challenges the implicit biological assumption of previous biogeographers that taxon richness would be uniform within an area in the absence of biological, environmental and historical factors, by assuming instead that a peak in richness towards the middle of the domain will always be produced as the result of stochastic processes that constrain geometrically the possible distribution of range sizes (Colwell *et al.*, 2004). The predicted richness values from mid-domain effect null models are very sensitive to the shape of the underlying range size frequency distribution (RSFD) (Colwell & Lees, 2000; Colwell *et al.*, 2004): RSFD's with a high proportion of large ranges show a stronger mid-domain effect, since large ranges must inevitably cross the middle of the domain and therefore increase the mid-domain peak in species richness; RSFD's with few large ranges, however, show a small mid-domain effect since small ranges can be found more uniformly across the domain and therefore do not necessarily cross the middle of the domain. This helps to explain the area-rescaled results from Table 4.7, where after accounting for area, removing the largest distributions from the analysis considerably weakened the relationship between observed and predicted richness values, whereas keeping the RSFD intact (i.e. not removing any taxa) or only removing the smallest ranges did not greatly affect the strength of the relationship. Furthermore this effect was greater for family ranges than for genus ranges since the family-level RSFD is less skewed than is the genus RSFD – i.e. it has a greater proportion of large ranges (see Chapter 3).

In mid-domain effect analyses, it is the empirical RSFD that should be resampled to create predicted richness patterns, rather than measuring observed richness patterns against any of the proposed theoretical RSFD's (Models 1 – 3 of RangeModel). Using the empirical RSFD implicitly incorporates into the model taxon-specific biological characteristics (such as vagility, size, or population density) that are logically independent of spatial patterns of richness within the domain, but which all interact to determine the empirical RSFD (Colwell *et al.*, 2004). Using a purely theoretical RSFD might seem less subject to biological assumptions than using an empirical RSFD and therefore a more truly "null" model (Koleff & Gaston, 2001). In fact, however, these theoretical models make the implicit biological assumptions that the range size frequency distribution for all species, will be precisely the same for different taxa, and that the maximum species range will cover the entire domain. If an MDE model is based on an RSFD that differs substantially from the empirical RSFD, the observed species richness gradient cannot be reproduced by any spatial arrangement of the empirical ranges, random or otherwise. Therefore, comparing an empirical richness pattern with an expected richness pattern based on an unrealistic theoretical RSFD tends to underestimate the role of the mid-domain effect because the predicted pattern is unobtainable (Colwell *et al.*, 2004).

Another important point is that in this study there is an inherent assumption that taxa occur in all latitudinal bands between the minimum and maximum observed values. Such interpolation is typical of analyses involving estimates of taxon richness from taxon range sizes (Whittaker *et al.*, 2001, Grytnes & Vetaas, 2002). It has been suggested that richness estimates based on interpolation may overestimate richness towards the centre of the gradient, because taxa are only strictly observed in bands at the extreme ends of each species range and may not in fact be present throughout that range, and also richness may be underestimated at the periphery because it cannot be interpolated beyond the limits of taxon ranges (Grytnes & Vetaas, 2002). However, interpolating range sizes is a pragmatic solution to an otherwise-intractable analytical problem, which may not noticeably alter the underlying trends in richness, and besides there is no evidence that taxa are not found where the ranges have been interpolated (Lees *et al.*, 1999; Bachman *et al.*, 2004).

Results of the simulation analyses nevertheless suggest that geometric constraints account for a large proportion of the variation in plant diversity across the latitudinal gradient. This is a surprising result: whilst the great diversity of the tropics has been known for over two hundred years, the role of geometric constraints was not explicitly formulated until 1994 (Colwell & Hurtt, 1994). However, the mid-domain effect has not been without its critics. The most obvious criticism is that, as yet, geometric null-models have mostly been applied across single-dimensional gradients. Although variation in area within the orthogonal dimension (longitude) can be controlled for in each latitudinal band, the challenge of applying these concepts in a single, explicit null model across two and even three dimensions still remains (Colwell & Lees, 2000; Koleff & Gaston, 2001; Colwell *et al.*, 2004; but see Jetz & Rahbek, 2001, 2002, for a 2-dimensional application). Despite some of the stronger attacks on the mid-domain effect as a concept (Hawkins & Diniz-Filho, 2002; Zapata *et al.*, 2003), it was not proposed as the sole explanation for patterns of taxon richness; it is simply an intrinsic property of the distribution of diversity within geometrically-bounded areas which must be accounted for before invoking other candidate explanations such as energy, temperature, potential evapo-transpiration, topography, precipitation, or isolation from source areas (Colwell *et al.*, 2004). Null model analysis therefore provides a powerful tool for eliminating these non-biological factors, and mid-domain effect predictions really need to be evaluated statistically on an equal footing with additional historical and environmental factors that accord equally well with empirical patterns of taxonomic diversity. However, these latter effects will only supplement the underlying pattern of overlapping range sizes.

4.5 General discussion on the latitudinal gradient of diversity

The approach taken in this chapter for studying the latitudinal gradient of diversity, essentially analysing the relationship between range size and latitude, is perhaps not the expected one. There has been no attempt to correlate taxon diversity at different latitudes with climatic or ecological variables such as solar energy or potential evapotranspiration, for example. The regions used in initially compiling this data are so large, and each therefore has such great internal variability, that such a simple correlation would be subject to a great many biases. Even if a clear relationship between diversity and climatic factors was found, it would be difficult to completely rule out artefacts in either the data or the analysis as the cause of this relationship. This is not to deny that strong relationships between diversity and climatic variables might be found at more detailed scales of analysis, as has often been demonstrated (Pianka, 1966). However, as pointed out by Currie (1991), there is no logical reason why increased energy in the tropics should lead to increased numbers of species, rather than just increased numbers of individuals within existing species (Willig *et al.*, 2003). There is also a lack of consensus among published empirical studies of the productivity-diversity relationship, with unimodal and linear relationships (both positive and negative) being found (Waide *et al.*, 1999; Willig *et al.*, 2003). The amount of available energy alone cannot account for the latitudinal gradient of diversity, therefore.

Instead of directly analysing the relationships between taxon diversity and ecological or climatic factors, the analyses in this chapter have taken a purely geographical approach to the latitudinal gradient in diversity, which focuses on the geographical variation in range sizes of taxa. It is possible that a latitudinal diversity gradient could be produced if all taxa have equivalently-sized ranges; there would then just need to be more of them in the tropics than in the temperate regions. However, range sizes of taxa are known not to be equivalent in size (see Chapter 3) and this variation in range size is also known to vary with latitude (this chapter). The latitudinal gradient of diversity is, therefore, underpinned by a latitudinal variation in range sizes of taxa. Individual taxon range sizes are known to be limited by ecological and/or climatic factors (Gaston, 2003), as evidenced by the recent latitudinal expansion of species ranges as a consequence of global warming (Parmesan & Yohe, 2003; Root *et al.*, 2003). Therefore, the latitudinal variation in range sizes can be interpreted as a surrogate measure of the ecological and/or climatic factors influencing ranges of individual taxa, and thus results based purely on the analysis of range sizes (as presented here) may well prove to have a climatic as well as a purely geographical explanation. Given the problems with the scale of the data outlined above, however, only those purely geographic analyses have been attempted here.

As taxa show distributions of different spatial extents and land area varies with latitude (see Section 4.2), each analysis in this chapter involves both latitudinal range size and land area. The predictions of Rapoport's Rule (that range sizes are smaller in the tropics) and the mid-domain effect (that tropical diversity is a consequence of the greater likelihood of overlap of large-ranged tropical taxa) contradict each other. Indeed, the formulation of the mid-domain effect was a by-product of the critical analysis of Rapoport's Rule (Colwell & Hurtt, 1994; Colwell *et al.*, 2004). Initially, a Stevens' plot of the latitudinal variation in geographical range sizes seems to support Rapoport's Rule, with mean latitudinal range-sizes decreasing towards the equator, but this is an artefact; as is shown with the midpoint method (Rohde, 1992), mean range-size is actually greatest rather than smallest for those genera centred on the tropics (see Section 4.3). Rapoport's Rule (Stevens, 1989) is therefore not the explanation of the latitudinal gradient of diversity but is better explained as a local phenomenon occurring over short latitudinal ranges in the northern hemisphere in which the amount of land area also decreases (Rapoport, 1982; Gaston *et al.*, 1998), i.e. range sizes co-vary with available land area.

As well as range sizes being greatest for genera centred on the tropics, the tropics also contain far more land area than do other climatic zones (Terborgh, 1973; Rosenzweig, 1992); if range sizes co-vary with available land area, therefore, genera centred on the tropics should be expected to have larger ranges. Since the degree of overlap between distributions is greater for large distributions than for small distributions (see Section 4.4), genera with large distributions, which are centred on the tropics, disproportionately increase the latitudinal gradient of diversity. This was shown by the distributions of tropical taxa 'bleeding' into temperate zones; and their exclusion revealed a strong relationship between the diversity of the remaining principally extra-tropical taxa and available area (see Section 4.2). Confirming the importance of area for range sizes, the principal peak in latitudinal range sizes coincides with the weighted centroid for the world (the latitude with the greatest land area) at 14°N. The greatest latitudinal ranges for the World (either mean, median or mode) all fell within the latitudinal band of 10 – 15 degrees north – just south of the peak in latitudinal range sizes found in New World birds at 17°N (Blackburn & Gaston, 1996). Although other competing explanations have not been explored in this thesis, overall, the geometry of the land area of the Earth seems to play a large part in explaining the presence of the latitudinal gradient of diversity.

The larger the latitudinal range of a genus, the more likely it is that the latitudinal midpoint of that range will be found in the tropics; put another way, genera with midpoints in the tropical regions are more likely to have large ranges simply because there is more land available to them either side of the tropics. The range-size frequency distribution of these latitudinal ranges around the equator is explained well by a 1-D geometrically-constrained null model known as the mid-domain effect, although the predicted richness values are very sensitive to the shape of the underlying range size frequency

distribution (RSFD) (Colwell & Lees, 2000; Colwell *et al.*, 2004): the mid-domain effect is more pronounced in RSFD's with a high proportion of large ranges, since these large ranges must by definition cross the middle of the domain and so increase the mid-domain peak in taxon richness. After accounting for the amount of area within each latitudinal band, the r^2 values for the relationship between observed and predicted richness values declined considerably in all cases where large ranges were trimmed from the analyses (leaving the smallest 95% or 75% of ranges); this was true for both family ranges or genus ranges, although the effect is more pronounced for family ranges. Despite growing evidence in support of the mid-domain effect (reviewed in Colwell *et al.*, 2004), however, and despite admissions by its critics that it is "a property of all biologically realistic null models based on range overlap counts" (Laurie & Silander, 2002), as a possible explanation of the latitudinal gradient of diversity, the mid-domain effect remains controversial (Hawkins & Diniz-Filho, 2002; Zapata *et al.*, 2003; Colwell *et al.*, 2004; Pimm & Brown, 2004). Further studies demonstrating the prevalence of the mid-domain effect are needed if the idea is to be more widely recognised.

The field of macroecology has been defined, in one sense, as the search for, and explanation of, emergent statistical properties in the patterns of ecological data (Brown, 1995). In the search for these emergent properties, it is possible that there is a variety of phenomena, and different explanations are valid at different scales. For example, the application of the mid-domain effect to the latitudinal gradient of diversity must necessarily be applied over very large ($\approx$ global) spatial scales, whilst the most successful applications of modelling energy-diversity relationships have been applied over continental scales (Currie & Paquin, 1987; Currie, 1991; O'Brien, 1998; O'Brien *et al.*, 1998). Studies of geometric constraints on patterns of diversity, however, which frequently show a pronounced mid-domain effect, are predominantly over regional scales (Colwell *et al.*, 2004, and references therein); but if the mid-domain effect is prevalent over both global and regional scales, it seems unlikely that it does not also come into play over continental scales. What is currently lacking in the assessment of the mid-domain effect are comprehensive studies with the mid-domain effect results analysed along with other candidate factors for explaining patterns of taxon richness. The mid-domain effect could be only one of a number of explanatory factors, therefore. Where comprehensive multivariate analyses have been done, the mid-domain effect still explains a large proportion (though not all) of the variance in taxon richness (e.g. Lees *et al.*, 1999); however, such studies remain few. Incorporating additional explanatory factors and estimating the influence of the mid-domain effect at finer scales remains as work for the future.

4.6 Summary

- There is a latitudinal gradient of diversity which is roughly symmetrical about the equator.
- The tropics contain far more land area than do other climatic zones, and this partly explains the greater biological diversity of tropical regions.
- Latitudinal range-sizes appear to decrease towards the equator, but this is an artefact; Rapoport's Rule is better explained as a local phenomenon over a short range of latitudes in the Northern Hemisphere.
- Latitudinal range-sizes in fact show a general increase in tropical regions, although the frequency distribution has many secondary peaks.
- The principal peak in latitudinal range-size coincides with the weighted centroid of land area for the world at 14°N.
- The range-size frequency distribution of latitudinal ranges around the equator is explained well by a geometrically-constrained null model known as the mid-domain effect.
- Overall, the geometry of the land area of the Earth seems to play a large part in explaining the presence of the latitudinal gradient of diversity.

CHAPTER 5

MULTIVARIATE ANALYSIS OF FLORISTIC RELATIONSHIPS

5.1 Introduction

This chapter and the following chapter present opposing multivariate analyses of regions and taxa, respectively. These questions have been tackled using a variety of multivariate statistical techniques; as such, they are partly an investigation into the methods as well as into the data themselves. This chapter asks the question: how are the floras of different areas of the world related to each other, i.e. what are their floristic relationships? This analysis of floristic relationships between regions is an extension of the beta diversity analysis of floristic similarity carried out in Chapter 3. However, whereas that was simply measuring overall similarity with regard to other measurable parameters such as latitude and distance between regions, in this chapter the analysis is finer-grained, attempting to tease-out the strengths of the various distribution patterns which separately contribute to the overall relationships. The actual distribution patterns and the number of taxa in each are then investigated in greater detail in the following chapter. The multivariate techniques used in both this and the next chapter were reviewed in Chapter 2. The general format followed in the analysis of floristic relationships in this chapter is set out below.

- All records from the Antarctic Continent were excluded, since there are only 2 native genera and this outlying region will greatly affect positions of regions in the subsequent analysis whilst being of little biological interest (both genera are very widespread).
- Genera occurring in only one region were excluded from the analysis, since they cannot be informative of relationships between regions, but will increase dissimilarity between regions.
- Floristic relationships between regions were investigated using hierarchical cluster analysis by UPGMA (Sørensen, Jaccard and Kulczynski similarity coefficients) and flexible beta (Sørensen similarity coefficient; $\beta = -0.25$ and $\beta = 0$).
- Beals' smoothing was then applied to the data to bring out the structure within the matrix and reduce the associated noise for ordination analyses.
- Bray-Curtis ordination, with the variance-regression endpoint selection method, was used to group regions in a non-hierarchical framework.
- Non-metric multidimensional scaling was then used to evaluate the fine structure of inter- and intra-group relationships.
- Results were compared with previous global analyses of floristic relationships.

5.2 Methodology

The outline of the methodology followed in the analysis of floristic relationships between regions is given above; greater methodological details for each of the different analyses are given below. Initial ordinations firstly revealed Region 91, Antarctic Continent, to be a very distant outlier, with this region at one end of the ordination diagram and all other regions clustered towards the other end of the diagram. Therefore for all analyses in this chapter, records from the Antarctic Continent were excluded, since there are only 2 native genera (*Colobanthus* (Caryophyllaceae) and *Deschampsia* (Gramineae)) which are of little biological interest (both genera are very widespread). Genera occurring in only one region were also excluded from the analysis, since they cannot be informative of relationships between regions, but will increase dissimilarity between regions. Whether or not to exclude endemic genera depends on exactly what is meant by 'floristic relationships'. For regions with a high proportion of endemic genera, such as Region 50, Australia, excluding endemic genera will reveal the floristic links with neighbouring regions which are otherwise overlooked because the Australian flora is so distinctive. Including the endemic genera would thus reduce the similarity of region with high endemism relative to the similarities between regions with low endemism. However, notwithstanding the interest in the whole floras of each region, what is really of interest in this chapter is precisely to reveal those underlying floristic relationships between regions, rather than the distinctiveness of each region (for this, see Chapter 3). All the analyses below were run in PC-ORD version 4.0 (McCune & Mefford, 1999).

5.2.1 Hierarchical cluster analysis

Cluster analysis of the genus-level data matrix was undertaken with the UPGMA linking algorithm, using both Sørensen's coefficient and Jaccard's coefficient, and also, since these two similarity coefficients are very similar to each other, with Kulczynski's coefficient. Clustering was also undertaken with the flexible beta linking algorithm, using $\beta = -0.25$ (which emulates Ward's linking method for Euclidean distance measures) and $\beta = 0$.

5.2.2 Bray-Curtis ordination

Bray-Curtis ordination of genus-level data was conducted using the Sørensen similarity coefficient, with the variance-regression method of endpoint selection, both with and without running the Beals' smoothing transformation.

5.2.3 Non-metric Multidimensional Scaling

Ordination by non-metric multidimensional scaling, both with and without transformation of the data with the Beals' smoothing function, was carried out under the 'Autopilot' mode in PC-ORD Version 4, which sequentially calculates each successive stage of the analysis. The analysis was run separately using

two different distance measures, Søresen and Kulczynski (relativised Sørensen, which standardises regions by their number of genera), each time stepping down in dimensionality from six axes to one, with the following Autopilot settings:

1. Distance measure = RELATIVISED SORENSEN [or SORENSEN]
2. Number of axes (max. = 6) = 6
3. Maximum number of iterations = 400
4. Starting coordinates (random or from file) = RANDOM
5. Reduction in dimensionality at each cycle = 1
6. Step length (rate of movement toward minimum stress) = 0.20
7. Random number seeds (use time vs. user-supplied) = USE TIME
8. Number of runs with real data = 40
9. Number of runs with randomized data = 50
10. Autopilot = YES
11. Stability criterion, standard deviations in stress over last 15 iterations = 0.00001
12. Speed vs. thoroughness = THOROUGH

This initial run found the optimal number of axes (between 1 and 6) for each ordination. The optimal number of axes is determined as the step beyond which there is little further reduction in stress. Again for each ordination, a final analysis was then run with no step-down of dimensionality, using as a starting point the configuration of objects from the optimal (that with the least stress) 3-dimensional ordination from the 40 preliminary runs with real data, with the following Autopilot settings:

13. Distance measure = RELATIVISED SORENSEN [or SORENSEN]
14. Number of axes (max. = 6) = 3
15. Maximum number of iterations = 400
16. Starting coordinates (random or from file) = FROM FILE
17. Reduction in dimensionality at each cycle = 3
18. Step length (rate of movement toward minimum stress) = 0.20
19. Random number seeds (use time vs. user-supplied) = USE TIME
20. Number of runs with real data = 1
21. Number of runs with randomized data = 0
22. Autopilot = YES
23. Stability criterion, standard deviations in stress over last 15 iterations. = 0.000010
24. Speed vs. thoroughness = THOROUGH

This ordination technique was used for both family-level and genus-level data. For the family-level data, several analyses were run, all with the Autopilot settings used in the genus-level ordinations,

except that, based on the results of the genus-level analysis, the Kulczynski similarity coefficient was chosen for all analyses. With the family-level data, it is possible to either summarise the distribution of each family simply as present or absent in each region, or instead to total the number of genera in that family found within that region (analogous to the 'abundance' of different species in an area). Both of these were tried, each with or without applying Beals' smoothing (since the number of genera of a particular family within a region is itself some measure of the 'favourability' of that area for that family, it was of interest to see if applying the Beals' smoothing function would create much difference). For the genus-level data, ordinations were run both with and without Beals' smoothing for both the Sørensen and Kulczynski (relativised Sørensen) similarity coefficients. Lastly, to investigate the floristic relationships within each region in more detail, this procedure was also repeated for each region, using the worldwide distributions of all the genera known from that individual region.

5.3 Results

5.3.1 Hierarchical cluster analysis

Results from the cluster analysis are given in Figures 5.1 – 5.5. Dendrograms for the UPGMA linking algorithm, using both Sørensen's coefficient and Jaccard's coefficient, are given in Figures 5.1 and 5.2, respectively. Dendrograms for the flexible beta linking algorithm, using $\beta = -0.25$ and $\beta = 0$, are given in Figures 5.3 and 5.4, respectively. A dendrogram for UPGMA clustering using Kulczynski's coefficient is given in Figure 5.5. There is strong agreement in the topologies of all five dendrograms; the differences between them are relatively minor. In both UPGMA dendrograms (Figures 5.1 and 5.2), and with flexible beta when $\beta = 0$ (Figure 5.4), the separation of Region 28 from all other regions is the first division of all, followed by a group of other island regions separate from the rest of the bulk of the regions. This second group comprises a group of three of the four regions of Pacific Islands (Region 61, South-Central Pacific, Region 62, Northwestern Pacific; and Region 63, North-Central Pacific), and the two regions 51 (New Zealand) and 90 (Subantarctic Islands), which always appear grouped together except for in UPGMA clustering with Kulczynski's coefficient (Figure 5.5), where Regions 28 (Middle Atlantic Ocean).and 90 (Subantarctic Islands) appear together outside all other groupings.

The next division in Figure 5.5 is of all the remaining island regions, with Region 51, New Zealand, adjacent to a group of three Pacific island regions (Region 61, South-Central Pacific, Region 62, Northwestern Pacific; and Region 63, North-Central Pacific). In Figure 5.3, with flexible beta when $\beta = -0.25$, Region 28 groups with Regions 51 and 90 in a group adjacent to the group of three Pacific regions (Regions 61, 62 and 63) and next to a large group of exclusively- or predominantly-tropical regions. All

of these island regions are small in size and have relatively low generic diversities; they will therefore show weaker floristic relationships with other, more diverse, regions than those more-diverse regions will show with each other. To some extent, therefore, Sorensen's coefficient is sensitive to the richness of the regions. However, the differences in generic diversity between regions is an inherent property of those regions and the weaker floristic relationships thus shown are a reflection of that real difference.

The next major division, in all five dendrograms, is the separation of a large group of predominantly tropical regions from a large group of predominantly temperate regions. There is almost no overlap in the composition of these two major divisions by any of the regions; in Figures 5.1 – 5.4, each region falls into the same group, either the 'tropical' group or the 'temperate' group. However, in Figure 5.5, Region 35, Arabian Peninsula, lies adjacent to the tropical group of Africa and Madagascar and not adjacent to the broad temperate group as in Figures 5.1 & 5.2, or adjacent to a 'Mediterranean and Middle East' group as in Figures 5.3 & 5.4. Taking the 'tropical' group first, there is again almost perfect separation within this group in all five dendrograms into three major elements, the position of Region 35, Arabian Peninsula in Figure 5.5 notwithstanding: an 'Africa and Madagascar' group (Region 22, West Tropical Africa; Region 23, West-Central Tropical Africa; Region 24, Northeast Tropical Africa; Region 25, East Tropical Africa; Region 26, South Tropical Africa; Region 27, Southern Africa; Region 29, Western Indian Ocean [Madagascar and nearby islands]); then a predominantly 'Asian' group, a broad geographical area of eastern Asia, tropical Asia, Australia and the islands of the SW. Pacific (Region 36, China; Region 38, Eastern Asia; Region 40, Indian Subcontinent; Region 41, Indo-China; Region 42, Malesia; Region 43, Papuasia; Region 50, Australia; Region 60, Southwestern Pacific); then a clear 'Neotropical' group (Region 79, Mexico; Region 80, Central America; Region 81, Caribbean; Region 82, Northern South America; Region 83, Western South America; Region 84, Brazil; Region 85, Southern South America).

Within each of these three tropical groups, there is almost perfect agreement between all five dendrograms in the grouping of all the regions within both the 'Asian' group and the 'Neotropical' group, and only slight disagreements in the positions of regions within the 'Africa and Madagascar' group and of Region 81, Caribbean within the Neotropical group in Figure 5.5. In the 'Asian' group, two subgroups are evident, one consisting of Regions 41 (Indo-China) and 42 (Malesia) and Regions 36 (China) and 40 (Indian Subcontinent) paired together and these two pairs themselves paired, with both pairs then grouped with Region 38, Eastern Asia; the other subgroup within the 'Asian' group consists of Region 43 (Papuasia [New Guinea, Bismarck Archipelago and Solomon Islands]) paired with Region 50 (Australia), both of which together are then paired with Region 60 (Southwestern Pacific). The 'Neotropical' group consists of seven regions from Central and South America, with a tight group of Region 82 (Northern South America) paired with Region 84 (Brazil), and this next to Region 83 (Western South America), next

to a group of Region 81 (Caribbean) adjacent to the pair of Region 79 (Mexico) and Region 80 (Central America), with Region 85 (Southern South America) below all other 'Neotropical' Regions. In Figure 5.5, however, Region 81, Caribbean, groups outside of the main group of truly 'Neotropical' regions (Regions 79, Mexico; 80, Central America; 82, Northern South America; 83, Western South America; and 84, Brazil) but inside Region 85, Southern South America.

With Africa, however, two regions (Region 24, Northeast Tropical Africa, and Region 27, Southern Africa) show different positions in the dendrogram of flexible beta clustering when $\beta = -0.25$ (Figure 5.4) from the dendrograms resulting from other clustering operations. Region 22 (West Tropical Africa) is always paired with Region 23 (West-Central Tropical Africa), and Region 25 (East Tropical Africa) is always paired with Region 26 (South Tropical Africa); these two pairs are themselves grouped together with UPGMA clustering using either Sørensen's (Figure 5.1), Jaccard's (Figure 5.2) or Kulczynski's (Figure 5.5) coefficient or with flexible beta clustering when $\beta = 0$ (Figure 5.4), but Region 24 (Northeast Tropical Africa) is grouped between these two pairs, outside Regions 25 and 26, in Figure 5.3. Region 27, which in Figures 5.1, 5.2 and 5.4 is placed outside a group consisting of Regions 22, 23, 24, 25 and 26, is instead grouped with Region 29 (Western Indian Ocean [Madagascar and nearby islands]) in Figure 5.3, whereas in Figures 5.1, 5.2, 5.4 and 5.5, Region 29 is grouped outside all of the African regions (Regions 22, 23, 24,2 5, 26 and 27).

Within the large group of 'temperate' regions, there is greater disparity between the dendrograms. Surprisingly, two geographically-distant regions, Region 21 (Macaronesia) and Region 35 (Arabian Peninsula) group together in four out of five dendrograms (Figures 5.1 – 5.4) but not in Figure 5.5. In UPGMA analyses with either Sørensen's (Figure 5.1) or Jaccard's (Figure 5.2) coefficient, Regions 21 and 35 form the one half of the first division, next to all the other 'temperate' regions (Regions 10, 11, 12, 13, 14, 20, 30, 31, 32, 33, 34, 37, 70, 71, 72, 73, 74, 75, 76, 77, and 78). However, with flexible beta clustering when $\beta = -0.25$, Regions 21 and 35 group next to a group comprising European, Mediterranean and Middle Eastern regions (see Figure 5.3) – Region 10, Northern Europe; Region 11, Middle Europe; Region 12, Southwestern Europe; Region 13, Southeastern Europe; Region 14, Eastern Europe; Region 20, Northern Africa; Region 32, Middle Asia; Region 33, Caucasus; and Region 34, Western Asia. In this group, Regions 12 (Southwestern Europe) and 13 (Southeastern Europe) form a 'Mediterranean' group with Region 20 (Northern Africa), next to a 'Middle Eastern' group of Regions 32 (Middle Asia), 33 (Caucasus) and 34 (Western Asia), which collectively group with Regions 10 (Northern Europe), 11 (Middle Europe) and 14 (Eastern Europe) (see Figure 5.3). In Figure 5.5, Region 21, Macaronesia, is still left outside a group of all other temperate regions, while Region 35, Arabian Peninsula, groups with Africa and Madagascar in the 'tropical' part of the dendrogram.

In UPGMA analyses with either Sørensen's or Jaccard's coefficient, however, Regions 12 and 13 group with Regions 10, 11 and 14 to make a 'European' group, while Region 20 is grouped outside all of the other European and Middle Eastern regions (Regions 10, 11, 12, 13, 14, 32, 33 and 34) (see Figures 5.1 and 5.2). In UPGMA clustering with Kulczynski's coefficient (Figure 5.5) there is a more diverse group of European, Mediterranean and Middle Eastern regions (Regions 10, 11, 12, 13, 14, 20, 32, 33 and 34) next to a group of more temperate Asian regions (Regions 30, 31 and 37) which group with another region which seems to lie out of its geographical position – Region 70, Subarctic America. With flexible beta clustering when $\beta = 0$ (Figure 5.4), the Mediterranean and Middle Eastern regions (Regions 12, 13, 20, 32, 33 and 34) still group together, along with Region 21 (Macaronesia) and Region 35 (Arabian Peninsula), but the more northerly European regions (Regions 10, 11, 14) instead group with the more northern Asian 'boreal' regions (Region 30, Siberia; Region 31, Russian Far East; and Region 37, Mongolia), and this group as a whole is positioned next to Region 70, Subarctic America. A group comprising Regions 30, 31, 37 and 70 appears in all other analyses: with UPGMA (Sørensen's [Figure 5.1], Jaccard's [Figure 5.2] and Kulczynski's coefficient [Figure 5.5]) this 'northern Asia and subarctic America' group is next to the large group of European, Mediterranean and Middle Eastern regions (Regions 10, 11, 12, 13, 14, 20, 32, 33 and 34), but with flexible beta clustering when $\beta = -0.25$ (Figure 5.3) this group (Regions 30, 31, 37 and 70) is placed next to a group consisting entirely of North American regions – Region 71, Western Canada; Region 72 Eastern Canada; Region 73, Northwestern U.S.A.; Region 74, North-Central U.S.A.; Region 75, Northeastern U.S.A.; Region 76, Southwestern U.S.A.; Region 77, South-Central U.S.A.; and Region 78, Southeastern U.S.A.

This exclusively North American group (but excluding Region 70, Subarctic America) is another group found throughout the five dendrograms, either positioned next to a group of cold-temperate regions (Regions 30, 31, 37 and 70) as in Figure 5.3; or next to this group with the inclusion of temperate European Regions 10 (Northern Europe), 11 (Middle Europe) and 14 (Eastern Europe) as in Figure 5.4; or next a whole group of temperate Eurasian Regions (10, 11, 12, 13, 14, 20, 30, 31, 32, 33, 34 and 37) with the addition of Region 70 (Subarctic America) as in the UPGMA clustering of Figures 5.1, 5.2 and 5.5. Whatever the position of this North American group, however, it always maintains the same composition in four of the analyses (Figures 5.1 – 5.4), and with the same relationships. In these dendrograms there are two subgroups within the group, one of Western North America, with Region 76 (Southwestern U.S.A.) next to a pair formed by Region 71 (Western Canada) and Region 73 (Northwestern U.S.A.), and another subgroup of Central and Eastern North America, with a pair formed by Region 77 (South-Central U.S.A.) and Region 78 (Southeastern U.S.A.) next to a group of three regions, with Region 74 (North-Central U.S.A.) next to a pair formed by Region 72 (Eastern Canada) and Region 75 (Northeastern U.S.A). In Figure 5.5, however, Region 76 (Southwestern U.S.A.) groups with Regions 77 (South-Central U.S.A.) and 78 (Southeastern U.S.A.) to form a 'Southern U.S.A.' group which is adjacent to the rest of the North

American regions (except Region 70, Subarctic America) that are themselves split into a 'Northwestern North America' group (Regions 71 and 73) and a 'Central-Northeastern' group (Regions 72, 74 and 75).

To sum up the above results, the differences between the dendrograms are:

- The position of Region 28 (Middle Atlantic Ocean) is either basal to the whole group of other regions, or is grouped with other island regions (Regions 50, 61, 62, 73 and 90).
- Region 24 (Northeast Tropical Africa) either groups below other tropical African groups (Regions 22, 23, 25 and 26), or between Region 22 and 23 and Regions 25 and 26, below Regions 25 and 26.
- Region 27 (Southern Africa) either groups below all tropical African regions (Regions 22, 23, 24, 25 and 26), or as a pair with Region 29, themselves next to the tropical African regions.
- Regions 21 (Macaronesia) and 35 (Arabian Peninsula) together group either next to all other 'temperate' regions (Regions 10, 11, 12, 13, 14, 20, 30, 31, 32, 33, 34, 37, 70, 71, 72, 73, 74, 75, 76, 77, and 78); or next to just the European, Mediterranean and Middle Eastern regions (Regions 10, 11, 12, 13, 14, 20, 32, 33 and 34); or next to only the Mediterranean or Middle Eastern part of this group (Regions 12, 13, 20, 32, 33, 34); or Region 35 (Arabian Peninsula) groups next to the group of African and Madagascan regions.
- The 'boreal' northern Asian regions (Regions 30, 31 and 37) group either with the North America group (Regions 71, 72, 73, 74, 75, 76, 77 and 78); or with the European, Mediterranean and Middle Eastern group (Regions 10, 11, 12, 13, 14, 20, 32, 33 and 34); or with the more-northern European regions (Regions 10, 11 and 14).
- Regions 10, 11, and 14 are positioned either next to Regions 12 and 13 in a 'European' group, or with Regions 12, 13, 20, 32, 33 and 34 in a 'European, Mediterranean and Middle Eastern' group; or with Regions 30, 31 and 37 in a 'Northern Eurasian' group.

Therefore, the following groups seem to be robustly defined:

- A group of three Pacific island regions, Regions 61, 62 and 63.
- A 'tropical' group consisting of Regions 22, 23, 24, 25, 26, 27, 29, 36, 38, 40, 41, 42, 43, 50, 60, 79, 80, 81, 82, 83, 84 and 85.
- An 'Africa and Madagascar' group consisting of Regions 22, 23, 24, 25, 26, 27 and 29.
- A predominantly 'Asian' group consisting of Regions 36, 38, 40, 41, 42, 43, 50 and 60.
    - Within this an 'eastern and southern Asian' group consisting of Regions 36, 38, 40, 41 and 42.

    - Also within the 'Asian' group, a second 'Papuasian, Australian and SW. Pacific' group consisting of Regions 43, 50 and 60.
- A 'Neotropical' group consisting of Regions 79, 80, 81, 82, 83, 84 and 85.
    - Within this, a 'Central American and Caribbean' group of Regions 79, 80 and 81.
    - Also within this, a 'tropical South America' group of Regions 82, 83 and 84.
- A 'temperate' group consisting of Regions 10, 11, 12, 13, 14, 20, 21, 30, 31, 32, 33, 34, 35, 37, 70, 71, 72, 73, 74, 75, 76, 77 and 78.
- A 'temperate Europe' group consisting of Regions 10, 11 and 14.
- A 'Northern Asia' group consisting of Regions 30, 31 and 37.
- A 'North America' group consisting of Regions 71, 72, 73, 74, 75. 76, 77 and 78.
    - Within this, a 'Western North America' group consisting of Regions 71, 73 and 76.
    - Also within this, a 'Central and Eastern North America' group consisting of Regions 72, 74, 75, 77 and 78.

And the following groups seem to be poorly defined:

- The group of Region 51 (New Zealand) and Region 90 (Subantarctic Islands) may or may not include Region 28.
- The group of 'European, Mediterranean and Middle Eastern' Regions 10, 11, 12, 13, 14, 20, 32, 33 and 34 may or may not include either or both of Regions 21 and 35.
- The 'Northern Temperate' group of Regions 10, 11, 14, 30, 31, 37, 70, 71, 72, 73, 74, 75, 76, 77 and 78 may or may not include the 'Mediterranean and Middle Eastern' group of Regions 12, 13, 20, 32, 33 and 34, plus either or both of Regions 21 and 35.
- The 'North America' group of Regions 71, 72, 73, 74, 75, 76, 77 and 78 may or may not be next to a group of Regions 30, 31, 37 and 70

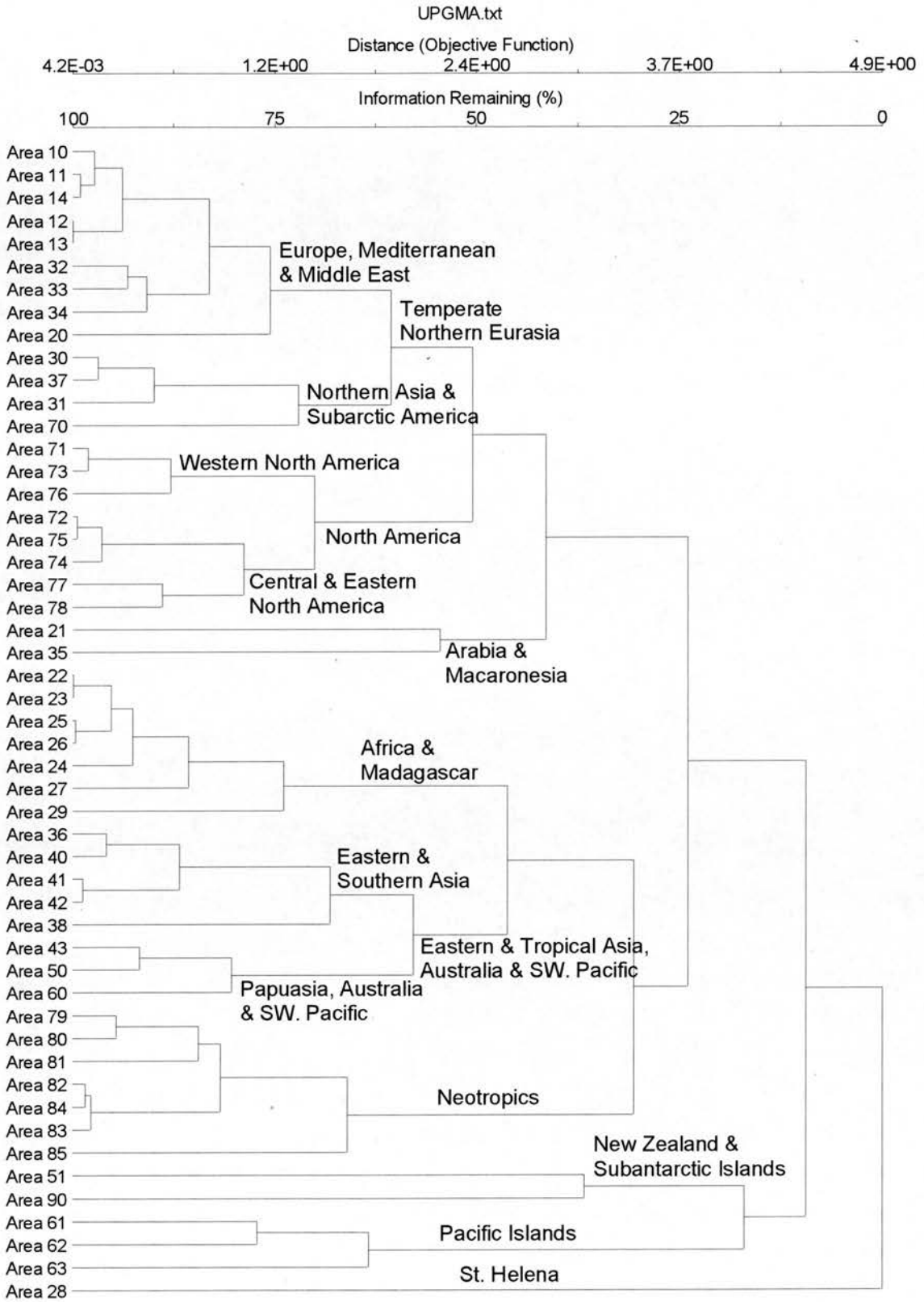


Figure 5.1 Dendrogram from UPGMA cluster analysis of TDWG regions by genera, Sørensen similarity coefficient, scaled by Wishart's objective function.

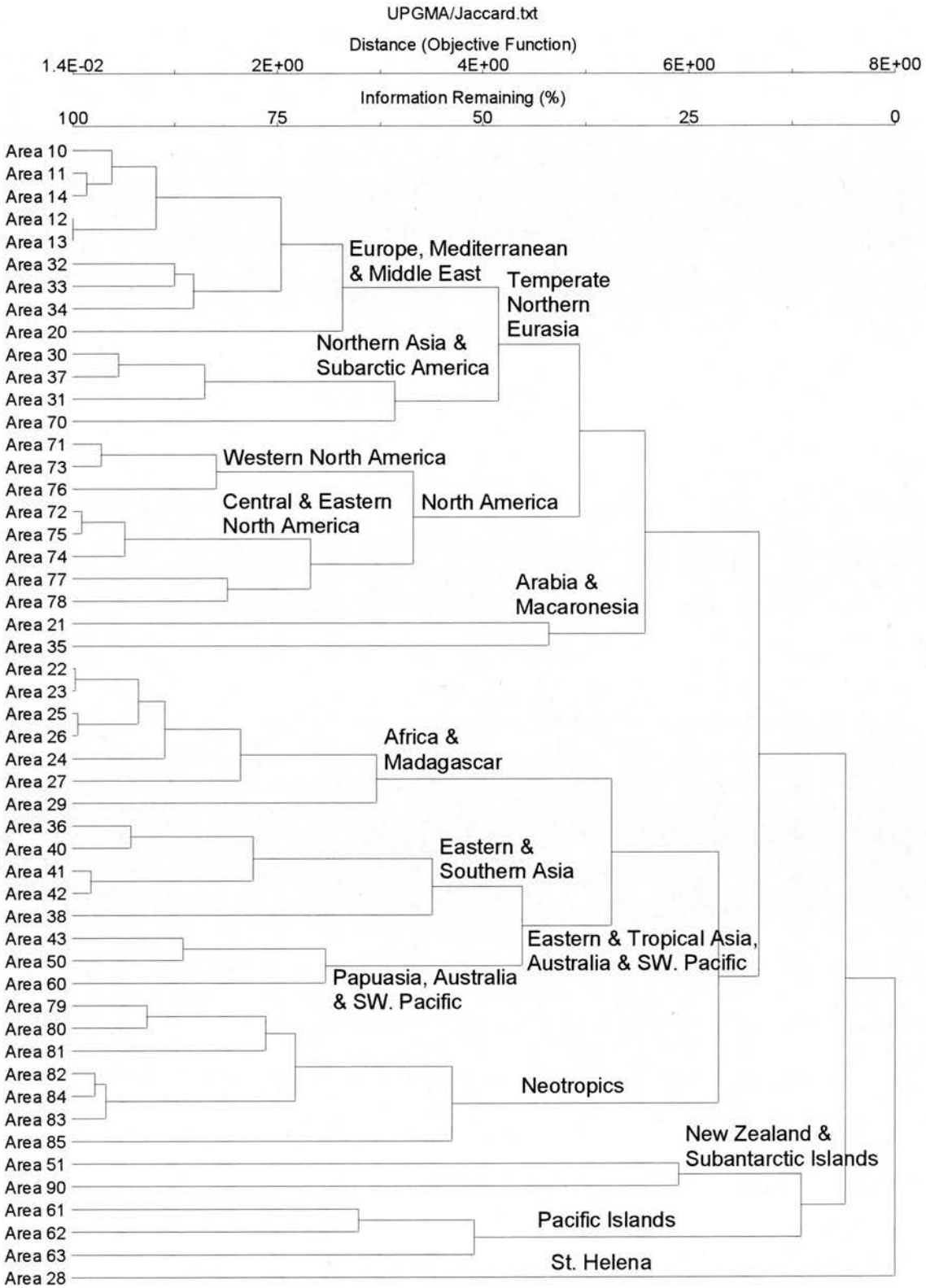


Figure 5.2 Dendrogram from UPGMA cluster analysis of TDWG regions by genera, Jaccard similarity coefficient, scaled by Wishart's objective function.

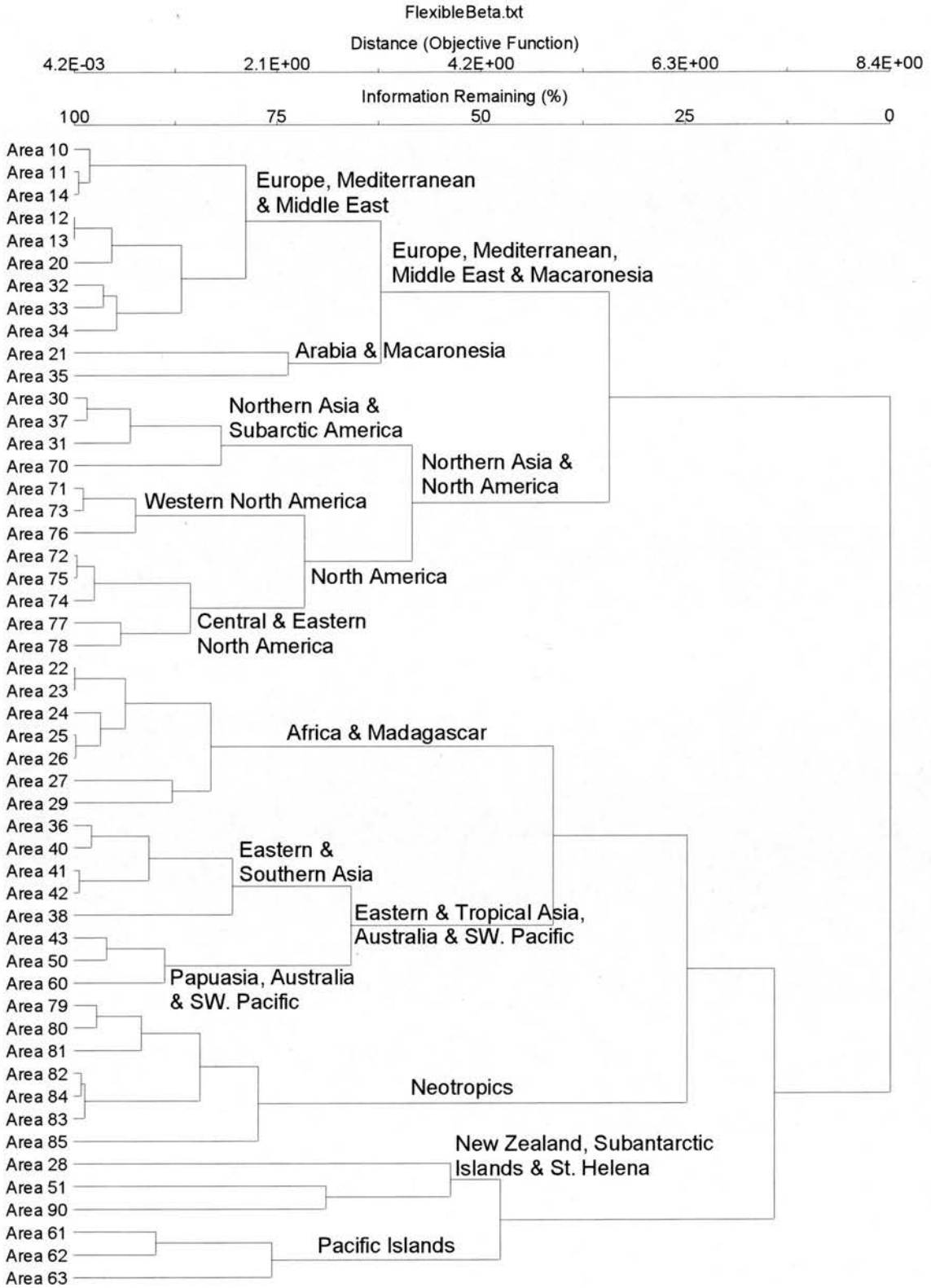


Figure 5.3 Dendrogram from flexible-beta cluster analysis of TDWG regions by genera, Sørensen similarity coefficient, beta = -0.25, scaled by Wishart's objective function.

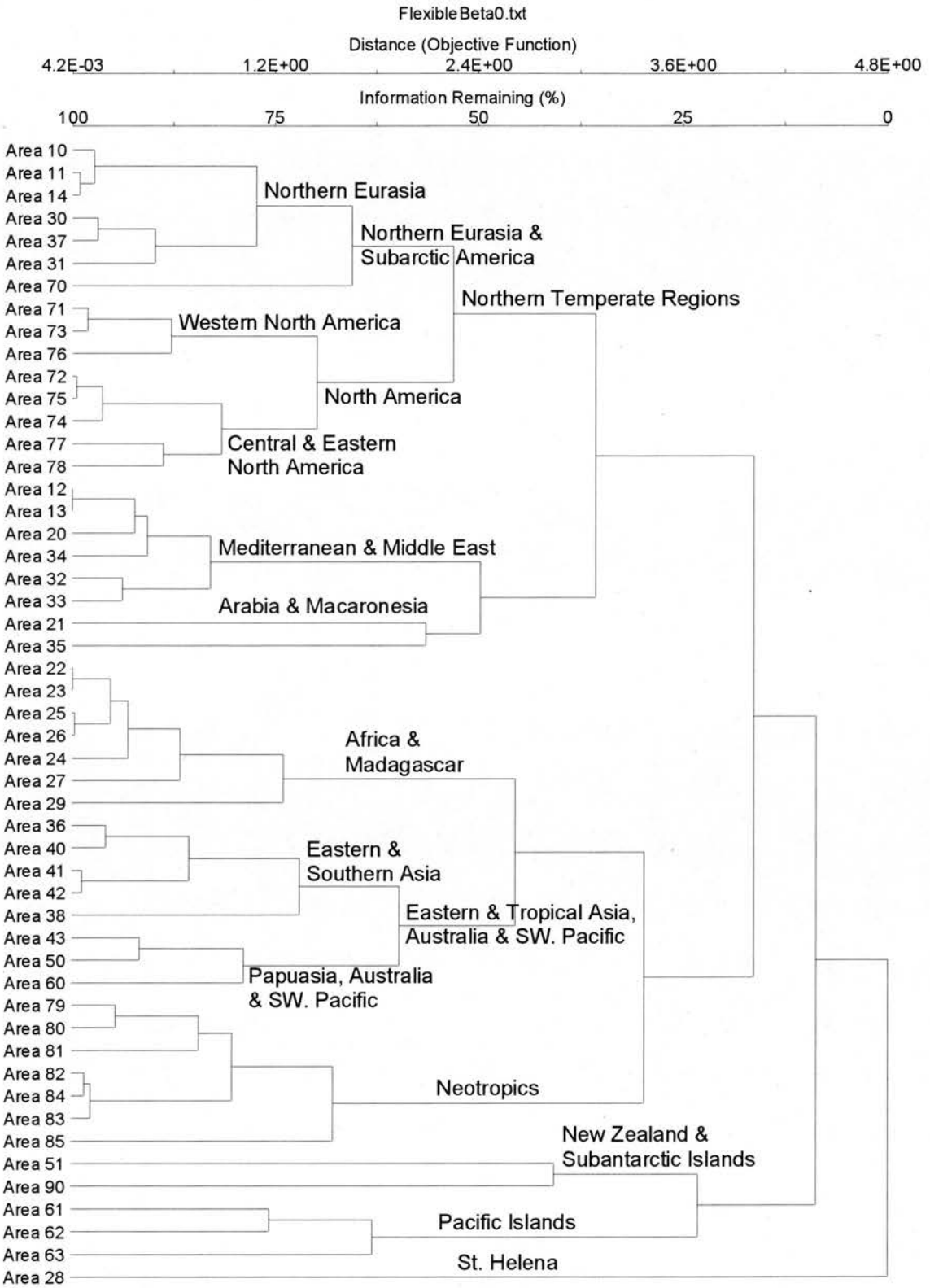


Figure 5.4 Dendrogram from flexible-beta cluster analysis of TDWG regions by genera, Sørensen similarity coefficient, beta = 0, scaled by Wishart's objective function.

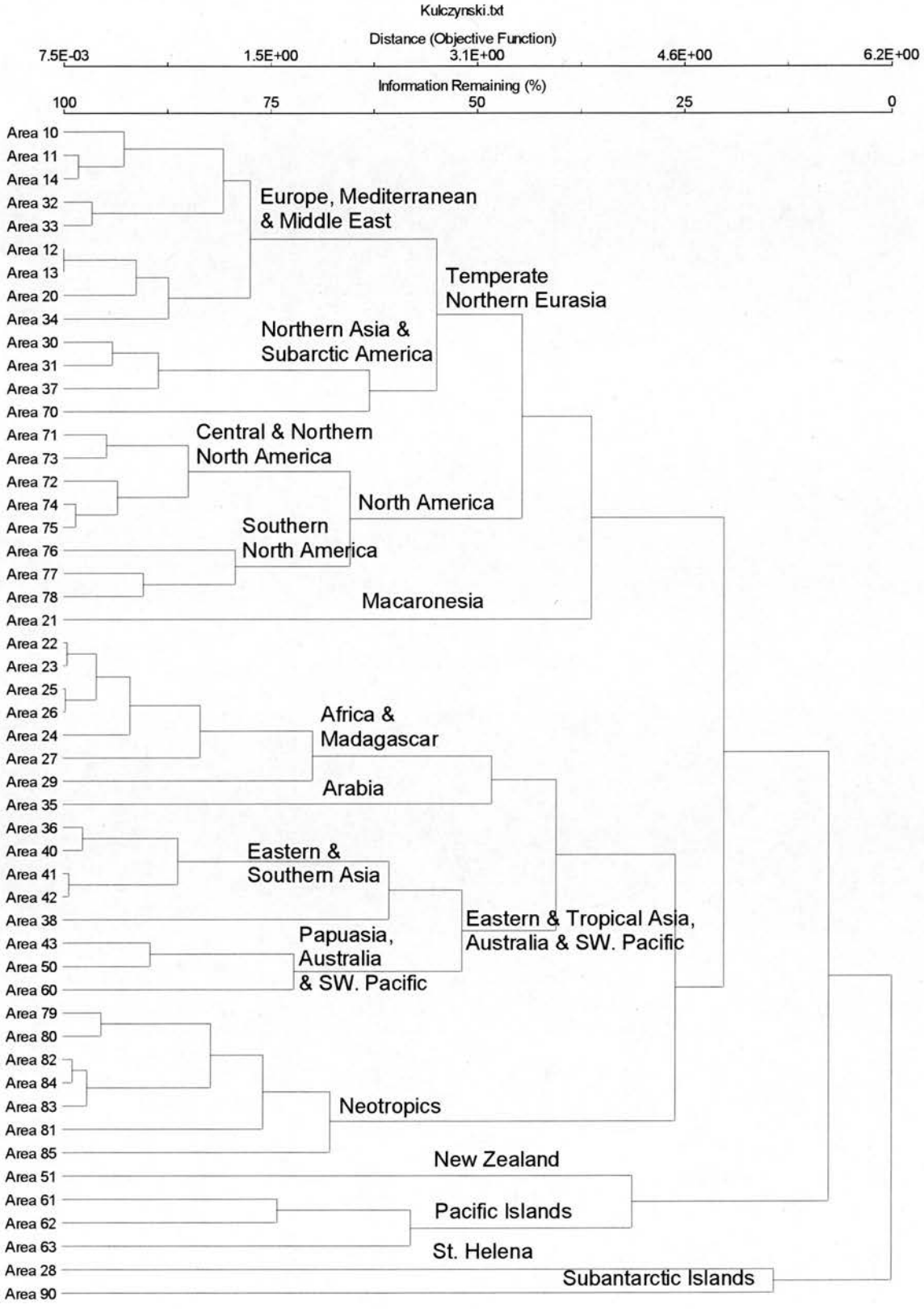


Figure 5.5 Dendrogram from UPGMA cluster analysis of TDWG regions by genera, Kulczynski similarity coefficient, scaled by Wishart's objective function.

	TDWG Region	Before Beals' smoothing		After Beals' smoothing	
		Skewness	Kurtosis	Skewness	Kurtosis
10	Northern Europe	3.36	9.28	2.63	6.82
11	Middle Europe	2.94	6.65	2.55	6.21
12	Southwestern Europe	2.31	3.36	2.40	5.18
13	Southeastern Europe	2.19	2.81	2.38	5.04
14	Eastern Europe	2.71	5.32	2.51	6.06
20	Northern Africa	2.45	3.99	2.25	4.73
21	Macaronesia	3.89	13.11	2.38	5.90
22	West Tropical Africa	1.60	0.57	1.67	1.63
23	West-Central Tropical Africa	1.40	-0.05	1.63	1.43
24	Northeast Tropical Africa	1.61	0.59	1.76	2.21
25	East Tropical Africa	1.35	-0.18	1.66	1.58
26	South Tropical Africa	1.38	-0.09	1.65	1.52
27	Southern Africa	1.39	1.58	1.77	2.14
28	Middle Atlantic Ocean	15.43	236.00	2.37	6.10
29	Western Indian Ocean	1.81	3.32	1.97	3.31
30	Siberia	3.28	8.76	2.62	7.02
31	Russian Far East	3.38	9.45	2.57	7.02
32	Middle Asia	2.57	4.62	2.27	4.86
33	Caucasus	2.58	4.63	2.39	5.29
34	Western Asia	1.78	1.16	2.16	4.19
35	Arabian Peninsula	2.54	4.44	2.01	3.96
36	China	1.01	-0.99	1.40	1.02
37	Mongolia	3.77	12.24	2.54	6.66
38	Eastern Asia	1.88	1.55	1.77	2.62
40	Indian Subcontinent	0.90	-1.18	1.41	1.15
41	Indo-China	1.21	-0.54	1.38	0.79
42	Malesia	1.25	-0.44	1.49	1.13
43	Papuasia	1.70	0.88	1.73	2.13
50	Australia	1.37	1.90	1.99	3.42
51	New Zealand	4.60	21.73	2.47	6.63
60	Southwestern Pacific	2.57	5.44	1.97	3.35
61	South-Central Pacific	5.37	26.79	2.26	5.14
62	Northwestern Pacific	4.62	19.36	2.06	3.91
63	North-Central Pacific	6.02	34.23	2.49	6.67
70	Subarctic America	4.64	19.57	2.98	9.56
71	Western Canada	3.47	10.02	2.92	9.11
72	Eastern Canada	3.59	10.86	2.88	8.84
73	Northwestern U.S.A.	2.98	6.90	2.77	7.94
74	North-Central U.S.A.	2.98	6.90	2.74	7.87
75	Northeastern U.S.A.	3.16	7.99	2.80	8.25
76	Southwestern U.S.A.	2.35	3.52	2.36	5.69
77	South-Central U.S.A.	2.29	3.26	2.30	5.43
78	Southeastern U.S.A.	2.35	3.51	2.42	6.19
79	Mexico	1.23	-0.48	1.53	1.51
80	Central America	1.30	-0.30	1.38	0.73
81	Caribbean	1.82	1.30	1.56	1.48
82	Northern South America	1.23	-0.50	1.27	0.24
83	Western South America	0.91	-1.18	1.23	0.19
84	Brazil	1.12	-0.74	1.26	0.24
85	Southern South America	1.67	0.78	1.49	1.37
90	Subantarctic Islands	9.49	88.12	2.73	9.02
Average		2.88	11.76	2.10	4.32
Coefficient of Variation		**58%**		**19%**	

Table 5.1 Effect of Beals' smoothing on skewness, kurtosis and coefficient of variation between regions.

5.3.2 Beals' smoothing

Beals' smoothing has the effect of reducing the skewness, kurtosis and coefficient of variation within the data (McCune & Grace, 2002). The effects of applying Beals' smoothing to the data used in the Bray-Curtis and non-metric multidimensional scaling ordinations is shown in Table 5.1; although neither skewness nor kurtosis is necessarily reduced for each region, average skewness declines from 2.88 to 2.10, while average kurtosis declines from 11.76 to 4.32. The coefficient of variation, furthermore, shows a significant reduction from 58% to 19%. Given that the data are qualitative (presence or absence of genera in regions) the skewness and kurtosis of the original data should be an simple function of the numbers of taxa in each region (Table 5.2), thus diverse tropical regions show less skewness and kurtosis than do temperate regions (Table 5.1). Although in general they are reduced across the board after Beals' smoothing, both skewness and kurtosis still remain strongly correlated with diversity (Table 5.2).

Spearman rank correlation	Skewness	Kurtosis
Before Beals' smoothing	$r_s = 1.00$	$r_s = -0.99$
After Beals' smoothing	$r_s = -0.85$	$r_s = -0.88$

Table 5.2 Skewness and kurtosis (as calculated by PC-Ord version 4.0) are both strongly negatively correlated with diversity of regions (Spearman non-parametric rank correlation; $n = 51$).

5.3.3 Bray-Curtis Ordination

Bray-Curtis ordination of genus-level data was conducted using the Sørensen similarity coefficient, with the variance-regression method of endpoint selection, both with and without running the Beals' smoothing transformation. Results from Bray-Curtis ordinations are given in Table 5.3. With Beals' smoothing, the majority of variation (71.96%) was accounted for after only two axes had been calculated, while after three axes had been calculated 85.73% cumulative variance was explained. Without Beals' smoothing, however, both the % variance explained and also the regression coefficients for each axis are much lower, with a total of only 36.61% cumulative variance explained after three axes had been calculated. The Bray-Curtis ordination diagram is given in Figure 5.6 for the first two axes. Results from ordinations both with (Figure 5.6A) and without Beals' smoothing (Figure 5.6B) are shown; as well as increasing the amount of variation explained by the first two axes, the effect of Beals' smoothing on the ordination is to tighten the clusters of regions – but the group into which each region falls is not affected. The end points of the axes in Figure 5.6 are as follows, with lowest-scoring regions given first for each axis: before Beals' smoothing: axis 1 (Region 14, Region 82); axis 2 (Region 74, Region 24); after Beals' smoothing: axis 1 (Region 32, Region 82); axis 2 (Region 76, Region 23).

Four main clusters are apparent in Figure 5.6, which correspond to large geographical regions. There is a small, tight cluster of 'temperate Eurasian' regions on the left-hand side of the ordination (individual regions are too close together to distinguish); a very large, poorly defined group of principally 'Old World' tropical regions, which also includes the Pacific regions (Regions 60, 61, 62 and 63) but also Region 51 (New Zealand) and Region 90 (Subantarctic islands); subsidiary groups of African and Asian regions, respectively, are evident within this groups, but the separation between them is poor. There is a well-defined 'Neotropical' region to the right-hand side of the ordination; and a group of 'North American' regions at the bottom of the ordination. The first cluster of temperate Eurasian regions is so tight because one of the endpoint elements (both before and after Beals' smoothing) is from this cluster (i.e. one of these regions has the greatest variance in inter-point distances), but then other regions show great floristic similarity to it. In each case the ordination clusters after Beals' smoothing are more tightly defined and closer to each other than are ordination clusters without Beals' smoothing.

Bray-Curtis Ordination	**Axis 1**		**Axis 2**		**Axis 3**	
	% variance	r^2	% variance	r^2	% variance	r^2
Before Beals' smoothing	22.59	0.28	12.17	0.17	1.55	0.17
After Beals' smoothing	42.82	0.58	29.14	0.31	13.77	0.07

Table 5.3 The effect of Beals' smoothing on Bray-Curtis ordination: percentage variance explained and regression coefficients are both greater for the first two axes following Beals' smoothing.

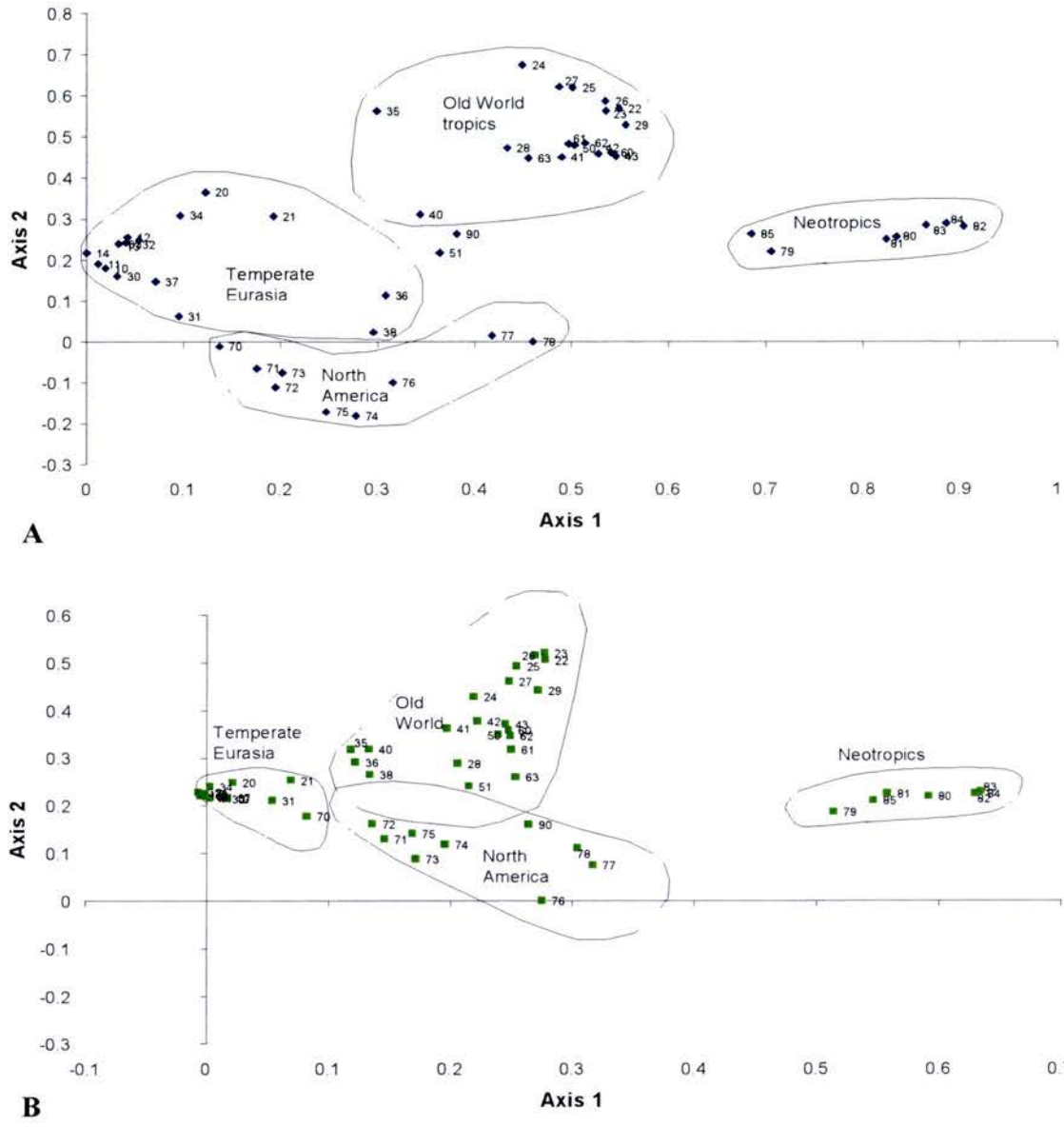


Figure 5.6 Bray-Curtis ordination diagram of genus-level data on 2 axes, Sørensen similarity coefficient, with the variance-regression method of endpoint selection, **A** before Beals' smoothing and **B** after Beals' smoothing. As the end points of the axes are different regions in each graph, both sets of data cannot be plotted together. Four broad groups are recovered with both ordinations, but groups are more tightly defined after Beals' smoothing; positions of groups overlap slightly between ordinations.

5.7.4 Non-metric Multidimensional Scaling of genus-level data

For each combination of matrix parameters for the non-metric multidimensional scaling ordination of genus-level data, Table 5.4 gives the amount of stress measured for each number of axes for both real and randomised data, and results of a Monte Carlo test of significance of difference between them. The aim of non-metric multidimensional scaling is to reduce stress between the configuration of points in the ordination space and the configuration of points in the original ecological space. In every case, there is a greater stress for randomised data than for real data, showing the degree of biogeographical structure within the data. The effect of Beals' smoothing on the ordination is to further reduce the degree of stress at higher numbers of dimensions, but for only one dimension the ordinations without Beals' smoothing actually have lower stress than do those with Beals' smoothing. However, not much is gained by reducing a large number of regions with complex floristic relationships down to a single dimension, effectively just placing them all along a line of similarity.

Given the reduction of stress at higher dimensions with Beals' smoothing, this transformation was retained for running the subsequent final ordinations. Degree of stress differed slightly between the different similarity coefficients – for ordination with the Sørensen coefficient stress was slightly greater than with the Kulczynski coefficient, although not by a great deal, and at low numbers of dimensions the ordination with Beals' smoothing and the Sørensen coefficient showed slightly lower stress values than did the ordination with Beals' smoothing and the Kulczynski coefficient (see Table 5.4). Table 5.4 shows that stress no longer falls significantly after three dimensions have been calculated using real data; this reduction in stress is also shown graphically in Figure 5.7 (an NMDS 'scree-plot'). Therefore a three-dimensional representation is the most appropriate for this data set. In Figure 5.7 the stress of the real data is also considerably lower than the stress in the random data, revealing that there is a significant degree of structure within the data matrix and that the ordination results would therefore not be expected by chance. This is also reflected in the final column of p values in Table 5.4; all p values are significant, being lower than 0.02, meaning there is a less than 2% chance that the stress in the ordination would be that low (i.e. degree of structure would be that high) by chance.

For each combination of ordination parameters (with or without Beals' smoothing, Sørensen or Kulczynski similarity coefficient) a 3-dimensional final solution was found. Table 5.5 gives the stress value, instability value, number of iterations and the proportion of variation explained by each axis (r^2 value). The preferred final solution found by NMDS (Beals' smoothing, Kulczynski similarity coefficient) reached a low stress value of 1.83, with an instability value of 0.00001. These values are both remarkably low, and so indicate a very strong ordination of highly structured data. Although the amount of variation explained by the first axis is small (only 27%), 98% of variation has been accounted for after 3 axes have

been constructed. The plot of iteration *vs.* stress is given in Figure 5.8 below and shows that as the number of iterations increases beyond 20, stress suddenly drops very sharply until it levels off at beyond about 40 iterations. By about 65 iterations there is no further decline in stress, and the analysis terminated at 80 iterations. Clark (1993) recommended an ideal stress value of about 5, although McCune & Grace (2002) claim this is 'rarely achieved' with ecological data. The value of 1.83 obtained here, however, is below that given as an indication of an 'excellent' non-metric multidimensional scaling ordination proposed by Kruskal (1964a) of 2.5. Similarly, the instability value obtained here is below that recommended by McCune & Grace of 0.001. The ordination plot for this solution is given as Figure 5.9.

	Stress in real data (40 runs)			**Stress in random data (50 runs)**			
No. axes	Minimum	Mean	Maximum	Minimum	Mean	Maximum	*p*
No Beals' smoothing, Sørensen coefficient							
1	26.706	35.305	56.561	51.448	54.646	56.556	< 0.02
2	13.646	14.908	17.599	33.435	35.122	36.117	< 0.02
3	7.545	7.768	11.745	25.551	26.235	26.969	< 0.02
4	4.035	4.036	4.036	20.653	21.197	21.914	< 0.02
5	3.160	3.204	3.567	17.325	17.797	18.370	< 0.02
6	2.487	2.533	2.696	15.007	15.311	15.798	< 0.02
Beals' smoothing, Sørensen coefficient							
1	28.35	44.26	56.55	47.04	50.50	56.59	< 0.02
2	8.32	11.62	40.23	30.91	33.23	40.30	< 0.02
3	2.47	2.48	2.50	23.51	24.54	25.75	< 0.02
4	1.55	1.63	2.09	19.15	20.03	21.06	< 0.02
5	0.92	0.94	1.06	16.14	16.98	17.94	< 0.02
6	0.68	0.72	0.77	14.19	14.69	15.48	< 0.02
No Beals' smoothing, Kulczynski coefficient							
1	24.525	34.845	56.590	49.188	52.468	56.744	< 0.02
2	12.333	13.419	39.759	32.451	34.506	37.692	< 0.02
3	6.678	6.679	6.681	24.819	26.245	27.523	< 0.02
4	3.768	3.880	3.905	20.396	21.554	25.092	< 0.02
5	3.040	3.277	3.498	17.333	18.395	21.532	< 0.02
6	2.707	2.862	2.948	14.688	16.273	22.397	< 0.02
Beals' smoothing, Kulczynski coefficient							
1	29.811	41.962	56.464	45.860	51.243	56.592	< 0.02
2	8.880	11.909	40.102	31.016	32.664	34.833	< 0.02
3	1.834	1.840	1.849	23.199	24.388	25.151	< 0.02
4	1.205	1.297	1.575	19.158	19.926	20.534	< 0.02
5	0.786	0.837	1.091	15.941	16.872	17.486	< 0.02
6	0.663	0.697	0.787	13.765	14.588	15.156	< 0.02

Table 5.4 Results of ordination by non-metric multidimensional scaling for genus-level data, with an initial starting dimension of 6 axes, stepping down sequentially to 1 axis; column *p* represents the 95% Monte Carlo probability that ordinations with each decrease in dimensionality are significantly different from random [*p* = proportion of randomized runs with stress < or = observed stress i.e., *p* = (1 + no. permutations <= observed)/(1 + no. permutations)], which in this case they are ($p < 0.05$).

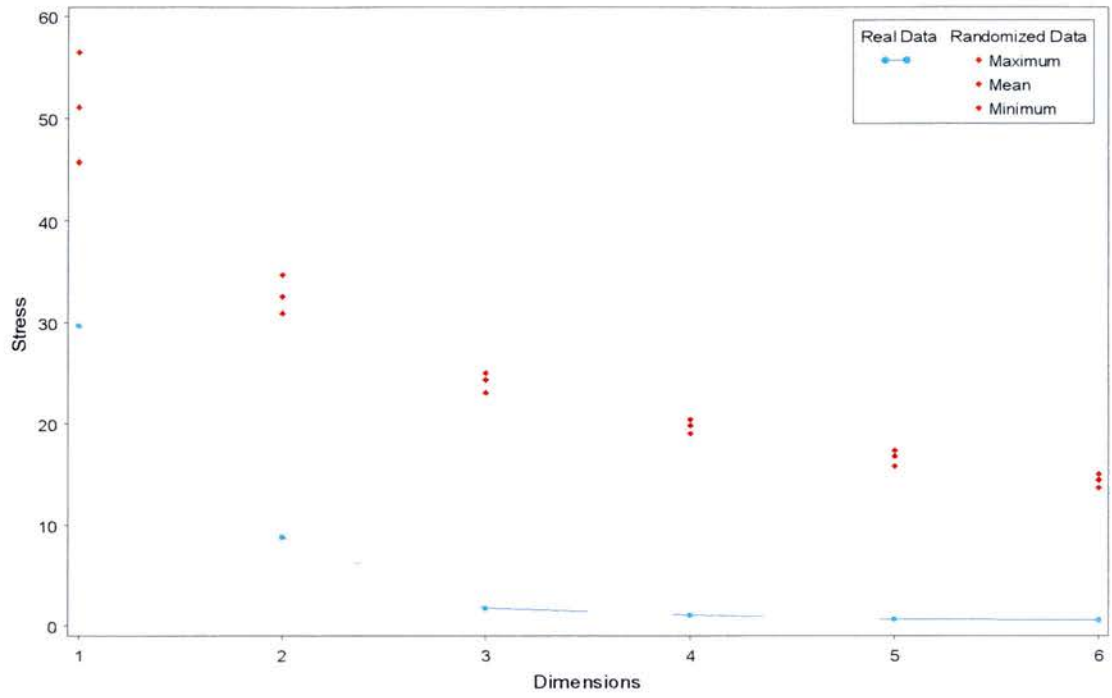


Figure 5.7 Non-metric multidimensional scaling 'scree-plot' showing reduction in stress with increasing dimensionality of the ordination (Beals' smoothing, Kulczynski coefficient). Reduction in stress is maximised after 3 dimensions, and results using real data show much lower stress than do results from random data, implying a high degree of structure in the original data matrix.

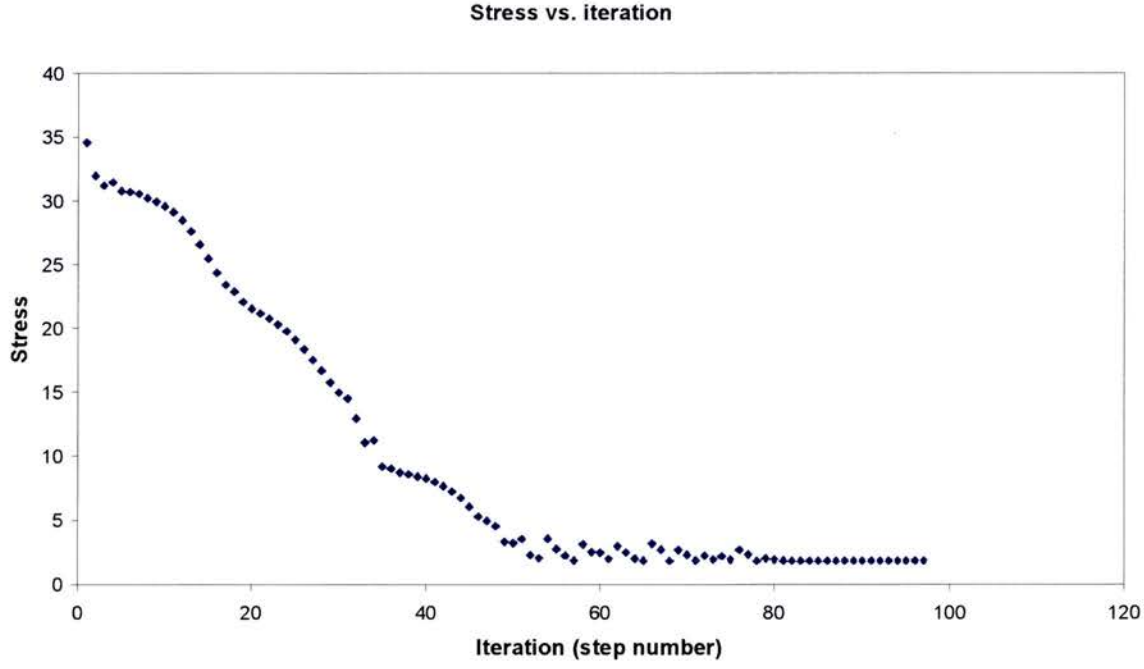


Figure 5.8 Stress in the final NMDS ordination (Beals' smoothing, Kulczynski coefficient) declines with successive iterations; the iterations cease once stress ceases to fall.

5.7.5 Non-metric Multidimensional Scaling of family-level data

Table 5.6 gives stress value, instability value, number of iterations and proportion of variation explained by each axis for the final ordination of all TDWG Regions by non-metric multidimensional scaling with the Kulczynski similarity coefficient for family-level data; ordinations were run using either genus totals per region, or scoring each family only as present or absent in a region, both with or without applying Beals' smoothing. In each case a 2-dimensional ordination was found. Stress values for family distributions scored simply as presence-absence data, for this data transformed with Beals' smoothing and for Beals' smoothing applied to numbers of genera per family per region are all very similar (6.62, 6.79 and 6.56, respectively), while that for the numbers of genera per family per region without applying Beals' smoothing is higher (9.39). Though these are all above the 'ideal' stress value of 5, it is still regarded as 'a good ordination with no real risk of drawing false inferences' (Clarke, 1993), and below Kruskal's (1964a) criterion for a 'fair' result. The preferred final solution (Beals' smoothing of genus numbers per family per region), however, was chosen for the lowest stress value and the smallest number of iterations, together with the (joint) highest percentage of explained variance at two dimensions (97%). Although the amount of variation accounted for in the first axis of this ordination is again small (10%), a further 87% is accounted for by the second axis. The ordination plot for this solution is given as Figure 5.10.

The ordination of families scored simply as presence-absence data was terminated due to exceeding the maximum number of iteration (400), rather than because of reaching a stable solution. This ordination has the same matrix parameters as the 'no Beals' smoothing, Kulczynski coefficient' ordination of genus-level data given in Table 5.5. It is noticeable that the stress value and percentage variance explained are similar for the final solutions for both these ordinations, implying that the strength of overall floristic relationships between regions are similar at both genus and family level. For the other ordinations, however, Beals' smoothing improved the genus-level data far better than it did the family-level data; indeed, for family data scored simply as presence-absence there was no appreciable decline in stress between the ordination applying Beals' smoothing and that which did not, whereas for genus-level data the stress value declined from 6.67 before Beals' smoothing to 1.83 after Beals' smoothing. This difference is presumably because the genus-level dataset has a much greater proportion of zeros than does the family-level data.

Genus-level data – matrix parameters	Stress value	Instability value	No. of iterations	Axis 1 (r^2)	Axis 2 (r^2)		Axis 3 (r^2)	
				Increment	Increment	Cumulative	Increment	Cumulative
No Beals' smoothing, Sørensen coefficient	7.54	0.0000	90	0.38	0.18	0.56	0.23	0.79
Beals' smoothing, Sørensen coefficient	2.47	0.0001	80	0.51	0.36	0.87	0.11	0.98
No Beals' smoothing, Kulczynski coefficient	6.67	0.0001	67	0.36	0.19	0.55	0.16	0.71
Beals' smoothing, Kulczynski coefficient	1.83	0.0000	97	0.27	0.53	0.80	0.18	0.98

Table 5.5 Stress value, instability value, number of iterations and proportion of variation explained by each axis for the final ordination of all TDWG Regions by non-metric multidimensional scaling of genus-level data; ordinations using either the Sørensen or the Kulczynski coefficient were run either with or without applying Beals' smoothing – in each case a 3-dimensional solution was found.

Family-level data – matrix parameters	Stress value	Instability value	No. of iterations	Axis 1 (r^2)	Axis 2 (r^2)		Axis 3 (r^2)	
				Increment	Increment	Cumulative	Increment	Cumulative
Genus totals (without further modification)	9.39	0.00001	129	0.73	0.21	0.94	—	—
Presence-absence transformation	6.62	0.00018	400	0.67	0.12	0.79	—	—
Beals' smoothing of un-modified data	6.56	0.00001	87	0.10	0.87	0.97	—	—
Beals' smoothing of presence-absence data	6.79	0.00001	162	0.92	0.05	0.97	—	—

Table 5.6 Stress value, instability value, number of iterations and proportion of variation explained by each axis for the final ordination of all TDWG Regions by non-metric multidimensional scaling with the Kulczynski similarity coefficient for family-level data; ordinations were run using either genus totals per region, or scoring each family only as present or absent in a region, both with or without applying Beals' smoothing – in each case a 2-dimensional solution was found.

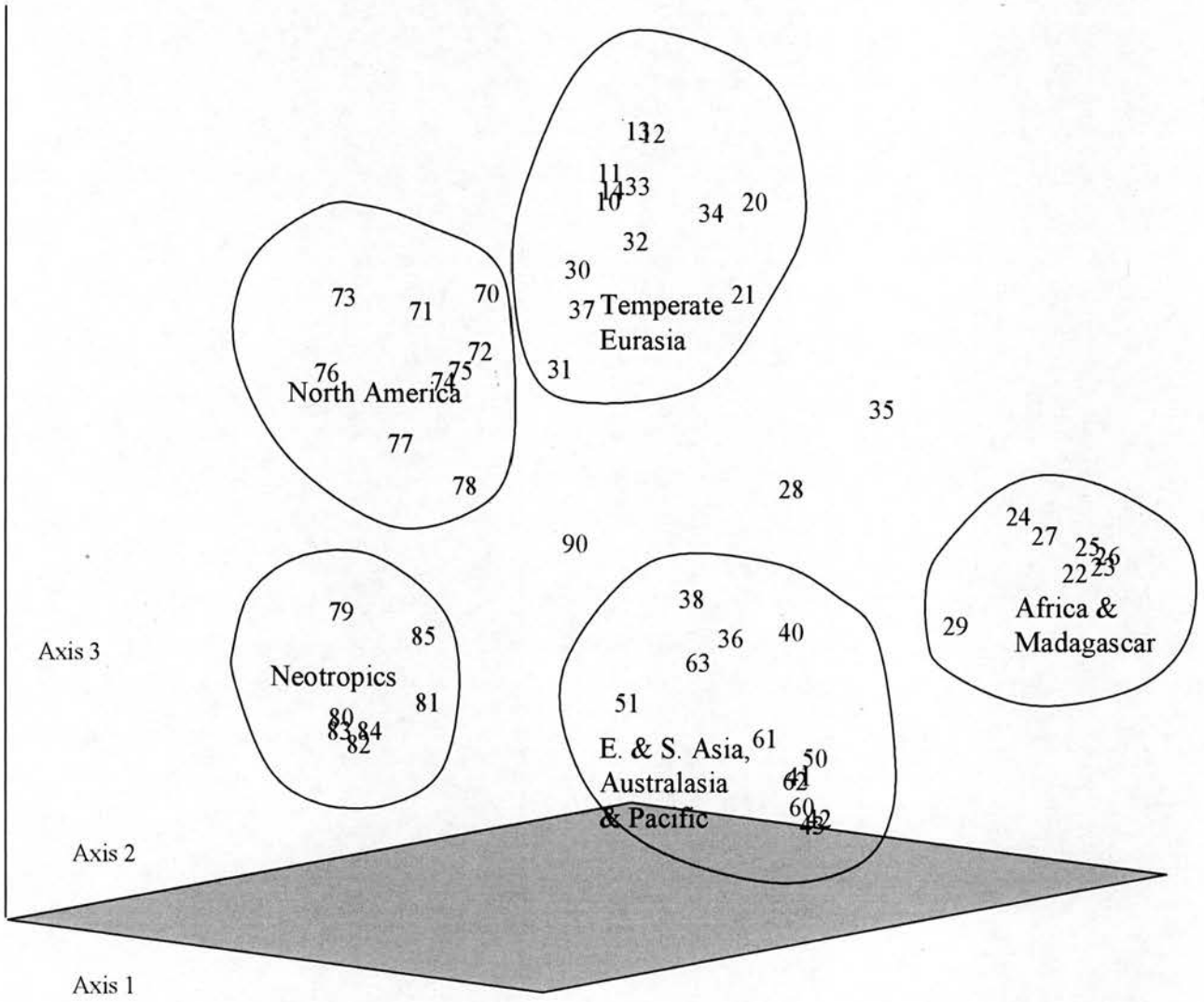


Figure 5.9 3-dimensional ordination plot from non-metric multidimensional scaling of genus-level data (Beals' smoothing, Kulczynski coefficient). Five groups are readily apparent, but not all regions fit into them.

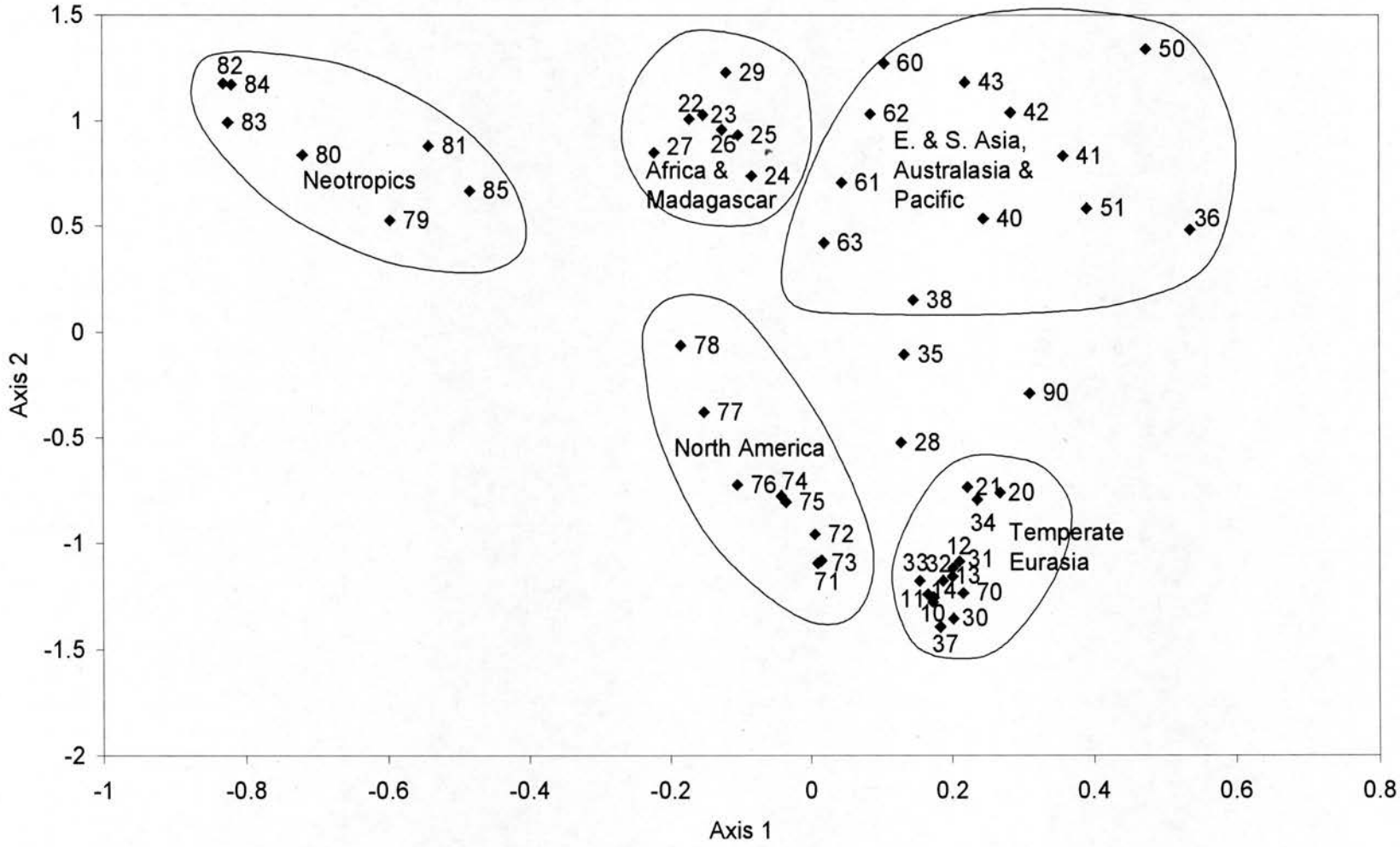


Figure 5.10 Ordination plot for family-level data from non-metric multidimensional scaling (Beals' smoothing of genus numbers per family per region; Kulczynski coefficient). The same groups are apparent as in Figure 5.9, but still the same regions do not fit into them.

5.7.6 Interpretation of the ordination plots

From the NDMS ordination diagrams given in both Figure 5.9 and Figure 5.10, five well-defined groups are apparent (although the 2-dimensional representation of 3-dimensional relationships reduces the interpretability of the diagram). The five groups are: an 'Africa and Madagascar' group, comprising Region 22, West Tropical Africa; Region 23, West-Central Tropical Africa; Region 24, Northeast Tropical Africa; Region 25, East Tropical Africa; Region 26. South Tropical Africa; Region 27, Southern Africa and Region 29, Western Atlantic Ocean; a distinct 'Neotropical' group, comprising Region 79, Mexico; Region 80, Central America; Region 81, Caribbean; Region 82, Northern South America; Region 83, Western South America; Region 84, Brazil and Region 85, Southern South America; a diverse group of eastern and southern Asian regions (Region 36, China; Region 38, Eastern Asia; Region 40, Indian Subcontinent; Region 41, Indo-China; Region 42, Malesia and Region 43, Papuasia), plus Region 50, Australia and the Pacific (Region 60, Southwestern Pacific; Region 61, South-Central Pacific [Polynesia]; Region 62, Northwestern Pacific [Micronesia] and Region 63, North-Central Pacific [Hawaiian Islands]; then a 'North American' group (Region 70, Subarctic America; Region 71, Western Canada; Region 72, Eastern Canada; Region 73, Northwestern U.S.A.; Region 74, North-Central U.S.A.; Region 75, Northeastern U.S.A.; Region 76, Southwestern U.S.A.; Region 77, South-Central U.S.A. and Region 78, Southeastern U.S.A.); and lastly there is a group of 'Temperate Eurasian' regions (Region 10, Northern Europe; Region 11, Middle Europe; Region 12, Southwestern Europe; Region 13, Southeastern Europe; Region 14, Eastern Europe; Region 20, Northern Africa; Region 21, Macaronesia; Region 30, Siberia; Region 31, Russian Far East; Region 32, Middle Asia; Region 33, Caucasus; Region 34, Western Asia and Region 37, Mongolia).

Several regions, however, do not fall neatly into any of these groups: Region 28, Middle Atlantic Ocean [St. Helena and Ascension Island] appears in between the 'Africa and Madagascar' group and the 'Asia, Australia and Pacific' group; it lies not far from Region 35, Arabian Peninsula, which is almost equidistant between the 'Africa and Madagascar' group and the furthest outlier of the 'Temperate Eurasia' group – Region 21, Macaronesia. Region 51, New Zealand, lies peripherally to the Pacific end (Regions 60, 61, 62 and 63) of the large 'Asia, Australia and Pacific' group in Figure 5.9, but no closer to this group than to Region 90, Subantarctic Islands. Region 90 itself, however, lies between the 'Asia, Australia and Pacific' group, Region 51 and the 'North America' group, and is itself not far from the isolated Region 28. Furthermore, within the very distinct 'Neotropical' group, Region 79 (Mexico) is seen approaching Regions 77 (South-Central U.S.A.) and 78 (Southeastern U.S.A.); in the 'Temperate Eurasia' group in Figure 5.9, Region 21 (Macaronesia) approaches Region 35 (Arabian Peninsula) and the 'Africa and Madagascar' group beyond that; and Region 70 (Subarctic America) is very close at one end of the 'North America' group to Regions 30 (Siberia) and 31 (Russian Far East) at one end of the 'Temperate Eurasia' group.

TDWG Region	No. axes	Stress in real data			Stress in random data			p	Stress value	Instability value	No. of iterations	Axis 1 (r^2)	Axis 2 (r^2)	
		Min.	Mean	Max.	Min.	Mean	Max.					Increment	Increment	Cumulative
10	2	1.33	2.37	5.20	33.67	35.61	36.87	0.0196	1.33	0	152	0.301	0.698	0.999
11	2	1.18	3.26	40.23	33.80	35.42	36.70	0.0196	5.03	0.00001	93	0.661	0.329	0.989
12	2	1.16	2.14	40.36	33.34	35.01	36.20	0.0196	1.16	0.00001	81	0.957	0.042	0.999
13	2	1.16	1.43	6.34	33.24	35.10	36.48	0.0196	1.16	0	95	0.969	0.030	0.999
14	2	1.40	1.94	5.95	34.10	35.51	36.83	0.0196	1.40	0	140	0.942	0.056	0.999
20	2	1.17	1.17	1.17	33.62	34.84	36.74	0.0196	1.17	0.00001	62	0.852	0.147	0.999
21	2	1.11	4.33	40.36	33.79	34.95	36.35	0.0196	1.11	0.00001	72	0.093	0.907	0.999
22	1	5.05	40.69	56.58	46.48	50.38	56.58	0.0196	5.05	0	55	0.990	—	—
23	1	4.29	40.19	56.56	43.35	48.26	56.58	0.0196	4.29	0	44	0.993	—	—
24	2	1.59	1.81	6.23	32.18	33.94	40.36	0.0196	1.59	0.00001	102	0.819	0.179	0.998
25	1	5.21	39.25	56.59	45.20	50.79	56.59	0.0196	5.21	0	88	0.989	—	—
26	1	4.86	36.42	56.59	42.37	48.70	56.59	0.0196	4.86	0.00001	81	0.991	—	—
27	1	5.98	37.24	56.57	44.47	50.26	56.57	0.0196	5.98	0.00001	45	0.987	—	—
28	2	4.09	6.31	40.36	26.96	29.90	32.46	0.0196	4.09	0.00001	106	0.881	0.111	0.993
29	1	4.50	37.06	56.55	49.99	54.77	56.59	0.0196	4.50	0	70	0.991	—	—
30	2	1.14	2.04	5.14	33.94	35.81	40.36	0.0196	1.14	0	101	0.853	0.146	0.999
31	1	6.34	35.88	56.58	53.34	55.60	56.54	0.0196	6.34	0.00001	71	0.982	—	—
32	2	1.10	1.77	4.58	32.73	34.61	36.13	0.0196	1.10	0	100	0.052	0.948	0.999
33	1	5.82	42.99	56.58	51.74	54.99	56.59	0.0196	5.82	0.00001	53	0.988	—	—
34	2	2.27	3.41	40.29	32.67	34.57	36.02	0.0196	2.27	0	86	0.482	0.515	0.997
35	2	1.82	2.88	16.29	34.12	35.20	36.52	0.0196	1.82	0	51	0.605	0.392	0.998
36	2	2.66	2.66	2.66	33.00	34.47	36.45	0.0196	2.66	0	121	0.589	0.407	0.996
37	1	5.01	35.54	56.58	52.18	55.58	56.56	0.0196	5.06	0.00001	61	0.991	—	—
38	2	2.07	4.20	10.39	34.41	35.71	37.20	0.0196	3.84	0.00001	70	0.775	0.217	0.992
40	2	3.98	5.37	40.30	33.16	34.78	36.47	0.0196	3.98	0.00001	128	0.808	0.183	0.991

41	1	5.21	37.27	56.59	48.07	52.78	56.59	0.0196	5.21	0.00001	157	0.987	—	—
42	1	4.54	40.63	56.59	47.49	51.94	56.59	0.0196	4.54	0.00001	76	0.991	—	—
43	1	5.05	39.21	56.59	47.50	53.02	56.59	0.0196	5.05	0.00001	66	0.990	—	—
50	2	1.93	2.56	3.23	33.47	34.88	36.19	0.0196	1.93	0.00001	95	0.960	0.038	0.998
51	1	7.45	28.44	56.55	45.31	50.54	56.59	0.0196	7.45	0.00001	62	0.975	—	—
60	1	4.80	42.21	56.56	50.52	54.48	56.56	0.0196	4.80	0.00001	68	0.991	—	—
61	1	2.92	40.61	56.56	53.37	55.89	56.59	0.0196	2.92	0	133	0.997	—	—
62	1	3.08	43.04	56.56	52.22	55.56	56.59	0.0196	3.08	0.00001	70	0.996	—	—
63	1	5.05	41.62	56.57	51.04	55.62	56.58	0.0196	5.05	0	45	0.990	—	—
70	1	3.39	40.13	56.59	53.66	55.87	56.56	0.0196	3.39	0	48	0.995	—	—
71	2	1.62	40.87	56.57	51.73	54.99	56.58	0.0196	1.62	0	147	0.832	0.166	0.998
72	2	1.14	1.29	1.63	33.96	35.38	36.84	0.0196	1.14	0.00001	64	0.872	0.127	0.999
73	2	1.80	4.67	9.45	31.56	33.39	34.89	0.0196	1.80	0.00001	131	0.797	0.201	0.998
74	2	1.69	1.69	1.69	33.29	34.82	36.30	0.0196	1.69	0	58	0.144	0.854	0.998
75	2	1.34	1.55	9.58	33.07	35.27	37.10	0.0196	1.34	0.00001	74	0.514	0.484	0.999
76	2	1.47	1.74	12.10	30.78	32.98	34.93	0.0196	1.47	0.00001	136	0.318	0.681	0.999
77	2	1.56	3.03	20.96	32.25	33.82	35.53	0.0196	1.56	0	64	0.376	0.623	0.998
78	2	1.38	8.26	15.95	33.89	35.11	36.72	0.0196	1.38	0.00001	108	0.733	0.265	0.999
79	2	1.01	2.11	8.35	30.52	31.72	33.83	0.0196	1.01	0	100	0.848	0.151	0.999
80	1	4.97	36.55	56.59	42.27	48.61	56.59	0.0196	4.97	0.00001	53	0.989	—	—
81	1	4.92	32.98	56.56	43.84	48.54	56.59	0.0196	4.92	0	69	0.990	—	—
82	1	3.44	35.88	56.59	43.53	46.37	56.57	0.0196	3.44	0.00001	54	0.995	—	—
83	1	4.73	34.81	56.59	40.47	45.37	56.53	0.0196	4.73	0.00001	47	0.990	—	—
84	1	3.83	33.32	56.59	41.43	47.10	56.59	0.0196	3.83	0	47	0.994	—	—
85	2	2.68	5.31	23.53	29.70	31.15	32.78	0.0196	2.68	0.00001	136	0.876	0.120	0.997
90	2	3.58	5.78	15.34	28.57	30.84	33.29	0.0196	3.58	0.00451	400	0.585	0.410	0.994

Table 5.7 Non-metric multidimensional scaling results only for genera found in each TDWG Region; Beals' smoothing, Kulczynski similarity coefficient.

5.8 Discussion

5.8.1 Comparison of multivariate techniques

The ultimate aim of multivariate analyses of this type is to reduce the complexity within the data by representing the principal relationships within a smaller number of dimensions. With all these techniques this data reduction necessarily comes at the expense of some distortion of the information within the data; in general, the greater the reduction in data complexity, the greater the distortion of the true patterns in the data (McCune & Grace, 2002). However, there is no real 'right' or 'wrong' method of multivariate analysis – there are merely more- or less-appropriate techniques. Since techniques differ in their statistical details of how best to measure floristic similarity there will not be a definitive answer to the question of which regions are floristically most similar to which other regions. As no region shows 100% endemism (see Chapter 3) all regions contain many genera found elsewhere and so show relationships with many other regions. Therefore it is worthwhile exploring several different techniques, and it is not unexpected that different techniques will give slightly different results. Indeed, it would have been surprising had they not.

The topologies of all five dendrograms produced by cluster analysis (see Figures 5.1 – 5.5) are very similar; the differences between them are only minor, reflecting different positions of individual regions. There is actually no difference at all in the topologies of the dendrograms between UPGMA clustering with Sørensen's coefficient (Figure 5.1) and UPGMA clustering with Jaccard's coefficient (Figure 5.2); the only difference is in the percentage information remaining, with Jaccard's coefficient consistently accounting for more information with each stage of the clustering process. This is because, whereas Sørensen's similarity coefficient is twice the shared abundance divided by total abundance, Jaccard's similarity coefficient is the shared abundance divided by the total non-shared abundance, so creating greater distances between groups (McCune & Grace, 2002). The Kulczynski coefficient relativises similarities between regions by the diversities of those regions, an important property when regions themselves differ greatly in size and diversities as they do here. This should prevent the analysis from becoming too heavily weighted by strong relationships caused simply by greater numbers of shared genera between diverse regions. Bearing this in mind, there is a perhaps surprising degree of congruence in the topologies of the four non-relativised dendrograms (Figures 5.1 – 5.4) and the dendrogram from UPGMA clustering of the Kulczynski coefficient, and between the dendrograms from the cluster analysis and the ordination diagrams (produced using the Kulczynski coefficient).

In all five dendrograms, there is an initial division which can broadly be characterised as 'island regions' *vs.* 'continental regions' (or, perhaps, less-diverse regions *vs.* more-diverse regions); then within

'continental regions' there is a major division between (again, broadly speaking) tropical *vs.* temperate regions. Within the tropical regions there is a clear separation of the three tropical areas: Africa and Madagascar, the Neotropics and a diverse region including tropical Asia but also China and Eastern Asia, Australia, and the SW. Pacific. Within temperate regions, the major division is between temperate Eurasia (but always excluding China and Eastern Asia) and North America; within each of these two divisions there is further separation usually between westerly regions (Europe and the Mediterranean region and western North America, respectively) and easterly regions (north temperate Asia and central and eastern North America, respectively).

From the genus-level NDMS ordination five large, well-defined groups are apparent (see Figure 5.9), which are: an 'Africa and Madagascar' group, a distinct 'Neotropical' group, a diverse group of eastern and southern Asian regions, a 'North American' group, and lastly a group of 'Temperate Northern Eurasian' regions. The following regions, however, lie outside of these groups: Middle Atlantic Ocean (28), Arabian Peninsula (35) and Subantarctic Islands (90). There is also remarkable agreement between the genus-level ordination plot shown in Figure 5.9 and that for family level shown in Figure 5.10. The same large geographical groups are recovered and the same unplaced regions occur in both analyses. The five groups recovered by non-metric multidimensional scaling correspond largely to the continental groups found by UPGMA cluster analysis at the level of 50% similarity or less. The 'Africa and Madagascar' group and the 'Neotropical' group are both identical in the two ordination analyses and in the UPGMA clustering with both Sørensen's coefficient and Jaccard's coefficient. In the ordination, Subarctic America (70) shows greater similarity to other North American regions than it does in the cluster analysis, but otherwise this North American group is identical in composition also. Macaronesia (21), which grouped with northern temperate regions with cluster analysis, and Arabian Peninsula (35) group more strongly with the 'Temperate Northern Eurasia' group under ordination. The major difference between the clustering and the ordination analyses is thus in the placing of the Pacific regions 61 (South-Central Pacific), 62 (Northwestern Pacific) and 63 (North-Central Pacific). Although they consistently group with other island regions in cluster analysis, there is no equivalent 'island' group in the ordination analysis; instead they lie peripherally to the large group of eastern, tropical and Australasian regions.

Ordination statistics from separate NMDS ordinations for all the genera found within each TDWG Region, all run using the Autopilot function of PC-ORD version 4.0 (McCune & Mefford, 1999) and with Beals' smoothing and the Kulczynski similarity coefficient, are given in Table 5.7. Stress values range from 7.45 (51, New Zealand) to only 1.10 (32, Middle Asia). Unlike the ordination of the whole genus-level dataset, for no individual region was an optimal 3-dimensional ordination ever recovered: 28 regions (55%) returned a 2-dimensional solution, while 23 regions (45%) returned only a 1-dimensional solution. However, the lowest percentage variation explained by any of these ordinations (either 1 or 2

axes) was as much as 97.5 % (for region 51, New Zealand). Those regions with only 1 axis of variation, accounting for the great majority of their floristic diversity, are predominantly tropical (19 out of 23 regions), but, irrespective of whether they are tropical or temperate, they are mostly areas which show very strong local floristic relationships – as can be seen in the cluster analysis dendrograms where they show great floristic similarity with neighbouring regions. Interestingly, however, the other regions with very simple floristic relationships are those Pacific regions (60, 61 and 63) which show up as the main difference between the clustering and the ordination analyses.

In general, the different results presented here are remarkably congruent with each other: the same large continental clusters of regions (Africa and Madagascar; Asia, Australia and at least the SW. Pacific; Neotropics; North America and Temperate Eurasia) are picked out in all analyses; in addition to the placing of the Pacific regions discussed above, the differences lie in the treatment of those regions which do not show clear-cut floristic relationships: Region 21, Macaronesia; Region 28, Middle Atlantic Ocean [St. Helena and Ascension Island]; Region 35, Arabian Peninsula; Region 51, New Zealand and Region 90, Subantarctic Islands. It is no coincidence that these are all amongst the least diverse regions, in terms of number of genera (ranked 43rd, 51st, 31st, 45th and 50th out of 51 regions [excluding Antarctica], respectively). These strong continental clusters of regions, which are in broad agreement in both the clustering and ordination analyses, are a product of the highly-skewed frequency distribution of distribution patterns: the majority of generic distributions are within individual continents. However, some differences are to be expected, since the 3-dimensional, non-hierarchical framework of the ordination reveals subsidiary floristic links between regions which are lost when the results are constrained onto a 2-dimensional dendrogram, which can only show the relationships of maximum similarity. The ordination results therefore more truly reflect the complexity of the underlying distribution patterns, but are correspondingly less easy to interpret.

Non-metric multidimensional scaling performed remarkably well in this analysis. The chief advantage of NMDS over other multivariate ordination techniques is that, being based on ranked distances, it has a greater ability to extract information from non-linear relationships (McCune & Grace, 2002). Much of classical multivariate statistics, for example PCA, is inappropriate for ecological data as it implicitly assumes linear relationships between species and either environmental factors or other species, which do not exist in reality. Secondly, the flexibility of NMDS means it can be based on a superior distance measure, such as the Sørensen (Czekanowski; Bray-Curtis) and Kulczynski (relativised Sørensen) coefficients used here, whereas DCA is implicitly built around a chi-squared distance measure inherent in the original Correspondence Analysis algorithm (Faith *et al.*, 1987; Minchin, 1987a). Manhattan (city-block) distance measures such as the Sørensen (Czekanowski; Bray-Curtis) and Jaccard coefficients have repeatedly been found to perform better with ecological data than do distance measures such as chi-

squared distance, which measure distance through Euclidean (i.e. straight-line) space (Minchin, 1987a). Euclidean space causes a greater sensitivity to outliers, as compared with Manhattan (city-block) distances; this means that single large differences in species composition between areas are weighted more heavily than are several small differences, so that using similarity coefficients which measure across Euclidean space can result in artificial groupings (McCune & Grace, 2002).

Floristic Regions of the World, according to Good (1974)

BOREAL KINGDOM

1. Arctic and Subarctic Region
2. Euro-Siberian Region
3. Sino-Japanese Region
4. Western and Central Asiatic Region
5. Mediterranean Region
6. Macaronesian Region
7. Atlantic North American Atlantic Region
8. Pacific North American Atlantic Region

PALAEOTROPICAL KINGDOM

AFRICAN SUBKINGDOM

9. North African – Indian Desert Region
10. Sudanese Park Steppe Region
11 Northeast African Highland Region
12. West African Rain-forest Region
13. East African Steppe Region
14. South African Region
15. Madagascan Region
16. St. Helena and Ascension Region

INDO-MALAYSIAN SUBKINGDOM

17. Indian Region
18. Continental South-east Asiatic Region
19. Malaysian Region

POLYNESIAN SUBKINGDOM

20. Hawaiian Region
21. Region of New Caledonian
22. Region of Melanesia and Micronesia
23. Region of Polynesia

NEOTROPICAL KINGDOM

24. Caribbean Region
25. Region of Venezuela and Guiana
26. Amazon Region
27. South Brazilian Region
28. Andean Region
29. Pampas Region
30. Region of Juan Fernandez

SOUTH AFRICAN KINGDOM

31. Cape Region

AUSTRALIAN KINGDOM

32. North and east Australian Region
33. South-west Australian Region
34. Central Australian Region

ANTARCTIC KINGDOM

35. New Zealand Region
36. Patagonian Region
37. Region of the South Temperate Oceanic Islands

Floristic Regions of the World, according to Takhtajan (1986)

HOLARCTIC KINGDOM

BOREAL SUBKINGDOM

1. Circumboreal Region
2. Eastern Asiatic Region
3. North American Atlantic Region
4. Rocky Mountain Region

TETHYAN SUBKINGDOM

5. Macaronesian Region
6. Mediterranean Region
7. Saharo-Arabian Region
8. Irano-Turanian Region

MADREAN SUBKINGDOM

9. Madrean Region

PALAEOTROPICAL KINGDOM

AFRICAN SUBKINGDOM

10. Guineo-Congolian Region
11. Uzambara-Zululand Region
12. Sudano-Zambesian Region
13. Karoo-Namib Region
14. St. Helena and Ascension Region

MADAGASCAN SUBKINGDOM

15. Madagascan Region

INDOMALESIAN SUBKINGDOM

16. Indian Region
17. Indochinese Region
18. Malesian Region
19. Fijian Region

POLYNESIAN SUBKINGDOM

20. Polynesian Region
21. Hawaiian Region

NEOCALEDONIAN SUBKINGDOM

22. Neocaledonian Region

NEOTROPICAL KINGDOM

23. Caribbean Region
24. Region of the Guayana Highlands
25. Amazonian Region
26. Brazilian Region
27. Andean Region

CAPE KINGDOM

28. Cape Region

AUSTRALIAN KINGDOM

29. Northeast Australian Region
30. Southwest Australian Region
31. Central Australian or Eremaean Region

HOLANTARCTIC KINGDOM

32. Fernandezian Region
33. Chile-Patagonian Region
34. Region of the South Subantarctic Islands
35. Neozeylandic Region

5.8.2 Comparison with global classifications of floristic regions

This chapter has presented a global analysis of floristic relationships between the 52 large geo-political TDWG regions of the world. To interpret the findings from this analysis in a broader context, comparisons can be drawn with the global hierarchical classifications of floristic regions that have been proposed several times in the past (e.g. Schouw, 1823; Engler & Diels, 1936; Good, 1974; Takhtajan, 1986). The basis for constructing such a floristic classification is the observation that plant distributions are discontinuous (not all plants are found in all places) and that because of this discontinuity between plant distributions some areas of the world are more similar floristically to each other than to other areas. As discussed before (see Chapter 1), an inter-nesting hierarchy of floristic regions is modelled on the Linnean taxonomic hierarchy. The justification for creating a hierarchical classification is therefore to express the floristic relationships between different areas of the world by placing closely-related areas in the same higher-level floristic unit, much as closely-related species are placed in the same genus. Also, the classification aims to reflect different degrees of endemism, if regions are placed within their own higher floristic unit (e.g. Madagascan Floristic Region alone within the Madagascan Subkingdom; Neocaledonian Floristic Region alone within the Neocaledonian Subkingdom).

The analysis presented in this chapter obviously differs fundamentally from previous floristic classifications in the delimitation of the areas being studied – large geo-political areas rather than biogeographically-defined floristic regions – and furthermore, this study is based entirely on genus-level data, whereas degree of species endemism is an important criterion in determining the floristic distinctness of an area. Both Good (1974) and Takhtajan (1986) explicitly use a criterion of endemism to define floristic regions: "each ... supporting a flora of its own ... which ... has largely developed within the region" (Good, 1974, page 27); "regions ... are established on the basis of high amounts of species and generic endemism" (Takhtajan, 1986, page 2). In practice, however, neither Good nor Takhtajan actually had enough detailed data to perform a robust statistical analysis, and their classifications were therefore rather subjective. Given that the data on which the analysis of floristic relationships in this chapter is based is so much more comprehensive than that used for any of these previous studies, it is useful to evaluate the relationships proposed in these previous systems of floristic regions against the floristic relationships evident from this analysis. To simplify the discussion below between the previous floristic hierarchies and the current results, the TDWG Level 2 Regions used in this analysis are always prefixed with their two-digit code, whereas any of Good's (1974) or Takhtajan's (1986) floristic regions are named with the suffix 'Floristic Region'.

The general pattern apparent from all the multivariate analyses presented in this chapter is one of separate continental clusters of regions. There are five broad groups from these analyses: the temperate

Eurasia group; the North America group; the Africa and Madagascar group; the tropical Asia group; and the Neotropical group (see Figures 5.1 – 5.5, Figure 5.9 and Figure 5.10). These continental groups relate to the floristic kingdoms or subkingdoms of the traditional floristic hierarchies such as Good's (1974) or Takhtajan's (1986) (see above). However, hierarchical cluster analysis consistently distinguished between continental regions and regions made up only of oceanic islands (see Figures 5.1 – 5.4), whereas neither Good (1974) or Takhtajan (1986) did. Instead, both of these authors included island regions alongside neighbouring continental regions (e.g. St. Helena and Ascension Floristic Region following Karoo-Namib Floristic Region; Takhtajan, 1986). Oceanic islands contain a large proportion of widespread and cosmopolitan taxa (Fosberg *et al.*, 1979, 1982, 1987; see also Chapter 6), so reducing their floristic similarity with any particular region; there may have been a tendency on the part of Good and Takhtajan to subjectively downweight the cosmopolitan element within these regions in order to reveal floristic relationships with nearby continental areas that they felt were more representative of these floras.

Both Good (1974) and Takhtajan (1986) had broadly-defined Boreal Kingdoms which contained both temperate Eurasia and North America. These two groups appear separately in the analyses within this chapter (see Figures 5.1 – 5.5) at a comparable level of similarity (approximately 50% information remaining) to that of the three large tropical regions (Africa and Madagascar; a broad tropical Asian group; and the Neotropics). It could be argued, however, that at only a slightly lower level of similarity there is a single group of all-temperate regions, with the main distinction being between 'temperate' continental regions and 'tropical' continental regions (see Figures 5.1 – 5.5), and this single group of 'temperate' regions largely corresponds to the broad Boreal Kingdoms of Good (1974) and Takhtajan (1986). Within the temperate Eurasia group, there is some separation within three of the clustering dendrograms between the European regions (Region 10, Northern Europe; Region 11, Middle Europe; Region 12, Southwestern Europe; Region 13, Southeastern Europe; Region 14, Eastern Europe) and the Asian regions (Region 30, Siberia; Region 31, Russian Far East; Region 32, Middle Asia; Region 33, Caucasus; Region 34, Western Asia; Region 35, Arabian Peninsula; Region 37, Mongolia) (see Figures 5.1 – 5.5), reflecting the floristic hierarchy of Good (1974), who separated a Euro-Siberian Floristic Region into separate European and Asiatic Sub-regions. However, this separation is not so apparent from either the Bray-Curtis ordination (Figure 5.6) or the NMDS ordinations of either family-level data (Figure 5.10) or genus-level data (Figure 5.9). Takhtajan (1986), on the other hand, recognised a much broader Circumboreal Floristic Region, which even included subarctic areas of North America. This latter delimitation is closer to the results of the analyses in this chapter, where, despite the separation of European and Asian regions within Figures 5.1 – 5.5, Region 70 (Subarctic America) consistently groups with other northern Asian regions (30, 31, 37) rather than other North American regions, except perhaps in the NMDS ordination of genus-level data (Figure 5.9), where it appears closer to other North American regions.

The large North American group of regions (Region 71, Western Canada; Region 72, Eastern Canada; Region 73, Northwestern U.S.A.; Region 74, North-Central U.S.A.; Region 75, Northeastern U.S.A.; Region 76, Southwestern U.S.A.; Region 77, South-Central U.S.A.; Region 78, Southeastern U.S.A. – but not Region 70, Subarctic America, see above) consistently emerges from all these analyses. In all clustering analyses there is a separation between Central and Eastern North America, on the one hand, and Western North America on the other (see Figures 5.1 – 5.5). This mirrors the division of temperate North America into the North American Atlantic Floristic Region and the Rocky Mountain and Madrean Floristic Regions, respectively (Takhtajan, 1986), even though the degree of endemism within the Madrean Floristic Region was sufficiently great for Takhtajan (1986) to place this within its own Subkingdom rather than in the Boreal Subkingdom with the Rocky Mountain Floristic Region. Good (1974) similarly imposes a longitudinal division on North America, recognising a Pacific North American Floristic Region and an Atlantic North American Floristic Region (which nevertheless includes the Great Lakes, Prairies and Mississippi Basin). Qian (2001), however, found that floristic similarity between latitudinal bands was greater than between longitudinal bands, although the use of a Euclidean distance metric may have complicated the results from this analysis. Although there are clearly latitudinal differences in the numbers of families and genera within North American regions (see Chapters 3 and 4), floristic similarity was found to be greater between regions of similar longitude than between regions of similar latitude.

Again, as with the Boreal Kingdom, both Good (1974) and Takhtajan (1986) have a broadly-defined Palaeotropical Kingdom including both Africa and tropical Asia. All the analyses within this chapter show these two large tropical areas to be more similar to each other than either is to the neotropical regions. Within Africa, Takhtajan's (1986) system of floristic regions largely follows that of White (1983), though without his more-pluralistic approach to delimiting alternative regional concepts such as archipelago-like centres of endemism (see Chapter 1). Quantitative support for White's broad floristic regions has been provided by Linder (1996), although a later study also found more restricted areas of species endemism within these broad floristic regions (Linder, 2001). In the present study, Regions 22 (West Tropical Africa) and 23 (West-Central Tropical Africa) always group together; collectively they correspond to the Guineo-Congolian Floristic Region (White, 1983; Takhtajan, 1986). Northern elements within these two regions, however, fall into the Sudano-Zambesian Floristic Region (White, 1983; Takhtajan, 1986), which also includes most of Regions 25 (East Tropical Africa) and 26 (South Tropical Africa). These latter two TDWG regions are also always grouped together, either with Region 24 (Northeast Tropical Africa) or with this outside of the group of all four preceding regions (22, 23, 25, 26). Region 27 (Southern South America) and Region 29 (Western Indian Ocean) then consistently group as less similar to this group of five tropical African regions than they do to each other.

The tropical Asian group found consistently in all the analyses in this chapter is the area of the analysis which differs most from the traditional view of this part of the world. The composition of the area is very broad, encompassing not just the strictly tropical Asian regions 40 (Indian Subcontinent), 41 (Indo-China), 42 (Malesia) and 43 (Papausia) but also the neighbouring regions both to the north (Region 36, China; Region 38, Eastern Asia) and to the southwest (Region 50, Australia; Region 60, Southwestern Pacific). In traditional hierarchies of floristic regions, China has always been grouped with other areas of temperate Asia (Sino-Japanese Floristic Region, Good, 1974; Eastern Asia Floristic Region, Takhtajan, 1986). The southern portion of China, however, is within Takhtajan's (1986) tropical Indochinese Floristic Region rather than the rest of China in his Eastern Asiatic Floristic Region. However, the position of Region 38, Eastern Asia, in this analysis, suggests the affinities of this broader Sino-Japanese or Eastern Asiatic Floristic Region itself are more tropical than they are temperate. The strong tropical elements within the East Asian flora have been recently studied by Qian *et al.* (2003), although without an explicit consideration of floristic relationships. The positions of China and Eastern Asia in this analysis imply that the strength of tropical floristic elements, at generic level, is greater than has previously been thought. Although Region 60, Southwestern Pacific, has long been considered closely related floristically to tropical Asia (e.g. Takhtajan, 1986), Australia has always been accorded special status. However, this is due to its great endemism (Crisp, *et al.*, 2001) and despite its strong floristic links with both New Guinea and the Southwest Pacific (see also van Welzen *et al.*, 2003). Indeed, in the hierarchical cluster analysis, Region 43, Papuasia [New Guinea and the Solomon Islands] consistently appears to be more closely related floristically to Region 50, Australia and Region 60, Southwestern Pacific, than it is to other areas of tropical Asia (see Figures 5.1 – 5.5). Neither Good (1974) nor Takhtajan (1986) suggested this, although it does agree with results of more recent analyses (Turner *et al.*, 2001; van Welzen *et al.*, 2003). However, the position of Region 50 in the results of the analyses in this chapter are influenced by the decision to exclude all endemic genera from the analyses; including the endemic genera would have reduced the similarity between this region and the neighbouring regions (Regions 43 and 60) – but the aim was explicitly to reveal those floristic similarities with neighbouring regions.

The group of seven neotropical regions (79, Mexico; 80, Central America; 81, Caribbean; 82, Northern South America; 83, Western South America; 84, Brazil; 85, Southern South America) is the most distinct and consistently-found in all the analyses in this chapter. It corresponds to the Neotropical Kingdom of Good (1974), although Takhtajan (1986) sets the boundary of his Neotropical Kingdom at a more northerly limit, excluding the subtropical areas of South America and placing these in his Holantarctic Kingdom. However, Cox (2001), following Conran (1995), advocates abandoning the Holantarctic Kingdom and placing all of southern South America within the Neotropical Kingdom. This neotropical group therefore seems to be robust, implying very high levels of generic endemism within the

neotropical region as a whole. Within this neotropical group, the three tropical South American regions (82, Northern South America; 83, Western South America; 84, Brazil) cluster particularly tightly (see Figures 5.1 – 5.5, and Figures 5.9 and 5.10). The high endemism within this area, and the strength of floristic relationships between these TDWG Regions, is also shown by the analysis of distribution patterns given in Chapter 6.

5.9 Summary

- Floristic relationships between regions have been investigated with a variety of multivariate statistical techniques.
- Ordination is regarded as a more effective technique since it does not impose any hierarchical structure onto the data, and more complex floristic relationships between regions are revealed.
- In general, strong continental clusters of regions emerge from these analyses.
- These distinct continental clusters correspond to the floristic kingdoms and subkingdoms of traditional floristic hierarchies.
- There is broad agreement between the analyses presented here and traditional hierarchical schemes of floristic regions.
- However, some regions consistently show stronger relationships with regions other than those shown in these floristic hierarchies.

CHAPTER 6

MULTIVARIATE ANALYSIS OF PLANT DISTRIBUTION PATTERNS

6.1 Introduction

This chapter presents a multivariate analysis of distribution patterns. Whereas Chapter 5 analysed biogeographic similarities between regions, this chapter analyses biogeographic similarities between taxa. Fundamentally, the outcome of the analysis presented in this chapter is a classification of distributions for all flowering plant families of the Angiosperm Phylogeny Group (APG, 1998), and for all genera held in the Vascular Plant Families and Genera database at the time of performing the analysis. The motivation for undertaking such a large analysis was the knowledge, from when compiling the data, that certain distribution patterns accounted for a very large number of genera, while other distribution patterns seemed to be essentially similar, but were just a small range-extension from other patterns. It therefore seemed that it should be possible to combine together the more 'similar' distributions, which did not necessarily correspond to other previously-published schemes of geographical classification such as Good's (1974) or Takhtajan's (1986) Floristic Regions of the World, and so be left with a less complex data set with a smaller number of distribution patterns than the 2817 unique combinations of TDWG regions (see Chapter 3).

These repeating distribution patterns occur again and again in unrelated genera of flowering plants, presumably because these same plants are each responding independently to the same underlying factors constraining their distribution, such as patterns of rainfall and temperature. Furthermore, since the majority of genera in each region are not endemic to that region (see Chapter 3), it is through an analysis of distribution patterns that the floristic links between regions (see Chapter 5) are revealed. The generic composition of any one region is a complex geographical confluence of a multitude of overlapping, separate floristic elements, and these together comprise the diversity within that region. Given that patterns of generic richness for TDWG regions were strongly correlated with both area (see Chapter 3) and latitude (see Chapter 4) while this richness is itself a product of these multiple floristic elements, was there any correlation between the number of floristic elements and the diversity within any one region? Is the generic diversity of a region a product of its floristic complexity?

While a few distribution patterns are very common, however, certain other genera show distribution patterns not found in any other genus. Many of these are apparently 'palaeo-endemics', while others simply defy easy explanation. The classification of distribution patterns throws these genera into sharper relief. Since I have often been asked the question: 'What is the most unusual distribution pattern of all?', this serves as a perfect opportunity to provide the answer – it is the genus *Oxygyne* in the Burmanniaceae, which has 2 species in Cameroon and 2 further species known only from Japan.

6.1.1 True R-mode analysis of distribution patterns

Identification of floristic elements has one major drawback – although shared floristic elements are identified, individual taxa are not assigned to them. Fundamentally, it is still a Q–mode rather than an R–mode analysis (see Chapter 5). However, a complete R–mode analysis of 14,304 distributions is a daunting task, both because of the complexity as well as the volume of the data. We would wish to have a less complex and more tractable data set than 14,304 genera in 2817 unique distribution patterns in 52 dimensions, especially if we consider some of these unique patterns to be essentially 'the same' as others (perhaps differing by the presence or absence in single regions). But how should we best measure geographical similarity? The general problem is illustrated in Figure 6.1: is the distribution shown in **A** 'the same' as that shown in **B**, which only differs by presence within one additional region? If so, is the distribution in **C** then 'the same' as that shown in **B**, since again they only differ by a single region? If the answer is again yes, then is the distribution in **A** the same as in **C**? If so, then where can the line be drawn between 'different' distributions when, for the totality of the data, there is so much overlap? If the answer is no, then are we left with having to deal with all 2817 unique distributions?

In general, one would need to run a transpose analysis of the data analysed in Chapter 5, perhaps using the same techniques. Cluster analysis was therefore tried, but not found to be particularly effective. There were three main problems with using cluster analysis, the additional inherent problems with using different similarity coefficients or different clustering strategies notwithstanding.

- Firstly, and most importantly, disjunct distributions were found to group arbitrarily with taxa from either one side or the other of that distribution: for example, a genus widespread in North America and also in China would group with North American taxa, while a genus widespread in eastern Asia but also with a locality in North America would group with Asian taxa, rather than these two examples forming an 'eastern Asia / North America disjunct' group as one might expect.
- Secondly, interpretation of the results became very difficult and time-consuming since the dendrogram resulting from the cluster analysis was over 200 pages in length (these had to be taped together and placed on a roll, like a scroll).

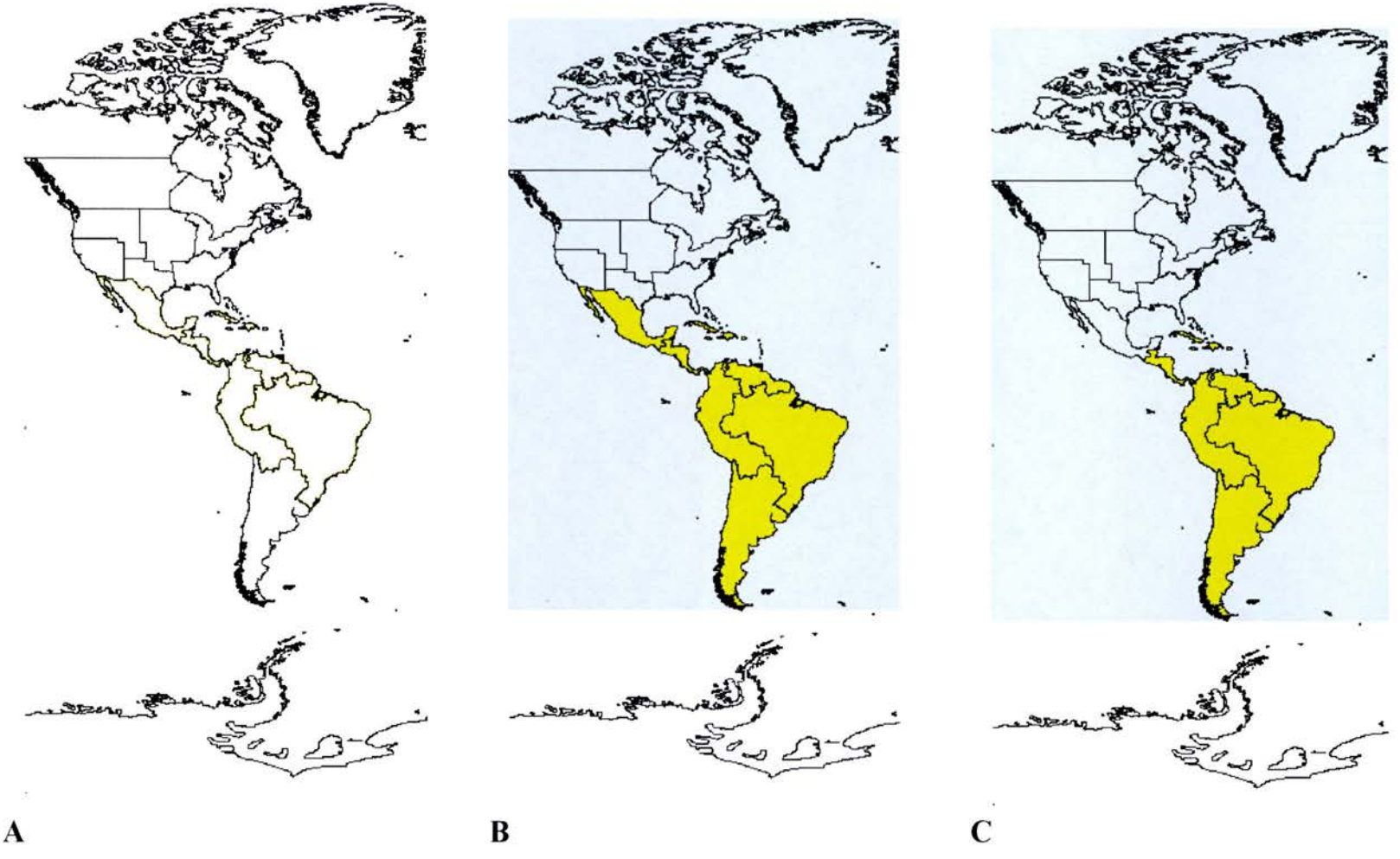


Figure 6.1 Three exemplar distributions. **A**. S. Mexico and Caribbean south to Bolivia and Brazil. **B**. S. Mexico and Caribbean south to N. Argentina. **C**. Guatemala and Caribbean south to N. Argentina. All three distributions are widespread through the Neotropics, but are they 'the same' or are they 'different'?

- Thirdly, since the dendrogram itself is the main output of cluster analysis, these results proved to be too intractable – taxa had to be designated by codes to fit into the constraints of the available software, and having done this it was then a separate stage to list the group membership for each cluster and then query the database to reveal what the distributions of those taxa actually were. Only then could the performance of the analysis be assessed. The results from this analysis were thus not amenable to further manipulation and analysis and so it proved difficult to take this approach further than production of the dendrogram.

Instead, building on the review of multivariate techniques outlined in Chapter 2, a non-hierarchical clustering approach was implemented using *k*-means partitioning (Lance & Williams, 1967; Bailey & Gatrell, 1995; Legendre & Gallagher, 2002), the technique again being described in Chapter 2. Since the number of partitions being sought was unknown beforehand, partition number was determined iteratively and the partition with the greatest value of the Calinski-Harabasz pseudo-*F*-statistic was chosen (Calinski & Harabasz, 1974; Milligan & Cooper, 1985). The main advantage of using this technique was in the lack of structure imposed on the data: forcing overlapping distributions into a discrete hierarchical framework caused the arbitrary resolution of disjunct distributions in one or another group discussed above, while this artefact was avoided within a non-hierarchical framework. However, since *k*-means partitioning can only be applied within a Euclidean geometric space, distributions first needed to be transformed from a non-Euclidean geometric space in 52 dimensions (the raw data matrix) to a lesser number of dimensions in Euclidean space. Ordination by non-metric multidimensional scaling, as used also in Chapter 5, was employed for this purpose. Group membership was outputted from the *k*-means partitioning analysis as simple ASCII text files, which could then be imported back into the Vascular Plant Families and Genera database to inspect the geographical distributions of each cluster. A non-parametric test of statistical significance was then applied between pairs of groups formed (Biondini *et al.*, 1988).

This approach is conceptually similar to an unsupervised classification of remotely sensed images; except that, unfortunately, there is no off-the-shelf software with which to carry out *k*-means partitioning of ecological data, and since the data format was not an even raster grid as with remotely sensed images it did not prove possible to use commercially available image processing software such as Erdas Imagine® with these data. Therefore the methodology was developed step-by-step, utilising several different software packages and transferring data between packages for separate stages of the analysis.

The methodology followed in the analysis of distribution patterns is outlined more fully below and presented in the accompanying flow diagram (Figure 6.2). The priority at each stage was to reduce the size and complexity of the data set but without losing or distorting the geographical inter-relationships

between the taxa too much. The different steps of this analysis therefore generally involve reduction of either the number of taxa to be analysed, or the number of dimensions in which the analysis operated. The analysis of family distributions followed the same methodology as that for genera, except that for families the initial step (excluding taxa endemic to a single region) was not necessary since the entire data set, being so much smaller, could be analysed *in toto*, and was.

6.2 Methodology

See Figure 6.2 for the complete flow diagram of each stage of this analysis.

1. Genera endemic to a single region were excluded from the analysis, on the assumption that each endemic distribution would remain a distinct pattern from the others, and from the wider distributions as well, while excluding these (almost 38% of the total) would reduce the size from c. 14,304 genera to c. 8,200 genera, thus speeding computation time considerably. Furthermore, single-region endemics would not greatly affect the results of an analysis of distributions based ultimately on a similarity matrix of overlapping distributions, as was this one. This step was unnecessary with family-level data since the number of families is so much smaller than the number of genera.

2. This presence-absence matrix was imported into PC-Ord version 4.0 and Beals' smoothing (see Chapter 2) was carried out to transform the data from presence-absence data into continuous data. The analysis becomes far more tractable and the degree of signal in the data far stronger after this transformation.

3. A non-metric multidimensional scaling ordination of genera, with the Sørensen similarity coefficient, was carried out using the 'medium' thoroughness setting of the PC-Ord Autopilot function, with 15 runs using real data and Monte Carlo tests based on 30 randomised runs, stepping down in dimensionality from 4 dimensions to 1 dimension for each run and with a maximum of 200 iterations per run, each from different random starting configurations. The 'slow and thorough' setting was abandoned after initial tests revealed it would take many weeks to run this for c. 9,200 genera. For the smaller family-level data set, the 'slow and thorough' setting was implemented, with 30 real runs and 45 randomised runs, stepping down in dimensionality from 6 dimensions to 1 for each run, with a maximum of 400 iterations per run, each from different random starting configurations. The result of these ordinations was a reduction from 52 dimensions to only 3 dimensions.

4. In each case a 3-dimensional solution was found, and ordination scores for these analyses were imported into Excel and sorted numerically. On inspection of these results, it was found that identical distribution patterns (that is, identical combinations of TDWG regions) always resulted in identical ordination scores along each three axes, so the set of complete ordination scores for all taxa was held in an Access database while only that subset of unique distributions (unique combinations of ordination scores) was re-exported to the next stage of the analysis. This resulted in a reduction in numbers of distribution from c. 8,200 genera to only 2817 unique distribution patterns.

5. Ordination scores were imported as ASCII text files into a *k*-means partitioning program downloaded from http://www.fas.umontreal.ca/biol/legendre/indexEnglish.html to objectively identify the clusters (= floristic elements) within the ordination space. The *k*-means partitioning algorithm works sequentially from a large number of groups down to a smaller number of groups, at each step combining the two groups whose inter-centroid distance is smallest. Since the initial *k*-means partition was into only a small number of geographically-diverse groups, this stage was implemented recursively, saving the partition (number of groups) with the maximum value of the Calinski-Harabasz pseudo-F-statistic from each pass and subsequently running separate partitions on each of these groups. The initial assignment of objects to groups was at random, with 100 replicates of each random assignment for each partition; since the input data were the ordination scores, these were treated as unweighted and unstandardized, the ordination itself already having standardised the raw data.

6. However, because *k*-means partitioning operates within a Euclidean space, the groups so defined tend to demonstrate a maximally-spherical shape – the drawback of so-called 'hyper-sphericity' (Legendre & Legendre, 1998) – and so a proportion of elements within the 2817 unique distribution patterns were apparently 'misclassified' when the group composition was inspected: they should have been in an adjacent group (based on their distribution pattern) despite being closer to the centroid of a different cluster (based on their position in the ordination space). The VPFG database was queried for presence-absence distributions for each group defined by *k*-means clustering and those misclassified elements were re-assigned manually after visual inspection of all 2817 distribution patterns (for genera), after each pass of *k*-means partitioning.

7. Groups defined by *k*-means partitioning were tested for statistical significance using pairwise non-parametric multi-response permutation procedures (MRPP), a test of the significance of between-group differences. The advantage of this technique is that it can be done on the raw presence-absence distribution data for all taxa within that distribution, rather than carrying out, for example, parametric canonical variates (= discriminant functions) analysis on the ordination scores for each unique

distribution pattern. Not only does canonical variates analysis require the usual assumptions of normal distributions and equality of variances of parametric statistics, it would also treat each unique distribution as statistically equivalent; using MRPP on all individual taxa within that distribution pattern effectively gives greater weight to the distinctness of more frequent distributions (those shown by more taxa). Though weighting the results in this way may seem to be an undesirable procedure on purely statistical grounds, in practice this is what has always been intuitively done by practising biogeographers: less frequent distribution patterns are grouped into larger classes (unless they are so distinct as to not fall into any other pattern), while more frequent distributions stand on their own. This is because unique distribution patterns do not allow for broader generalisations which will also explain the distribution patterns of other taxa.

8. Groups not significantly different were recombined manually and the analysis repeated from step 5 onwards. However, in practice it was discovered that after only five passes of *k*-means partitioning were large numbers of groups proving not to be significantly different. These groups were then recombined and the *k*-means partitioning halted. Because the recursive *k*-means partitioning in effect imposed a divisive hierarchy on the data, dividing it into successively smaller groups, a single pass was also made requesting the optimal number of groups, after five passes, plus or minus 20 groups and this result compared with that from five successive partitions – even though the partition from the single pass would not have been the optimal group number (i.e. this partition did not produce the maximum value of the Calinski-Harabasz pseudo-*F*-statistic).

9. Lastly, the groups of distribution patterns produced by the *k*-means partition were combined with those single-region endemic distributions whose genera had initially been excluded from the analysis, resulting in a final global classification of distribution patterns, at both family and generic levels.

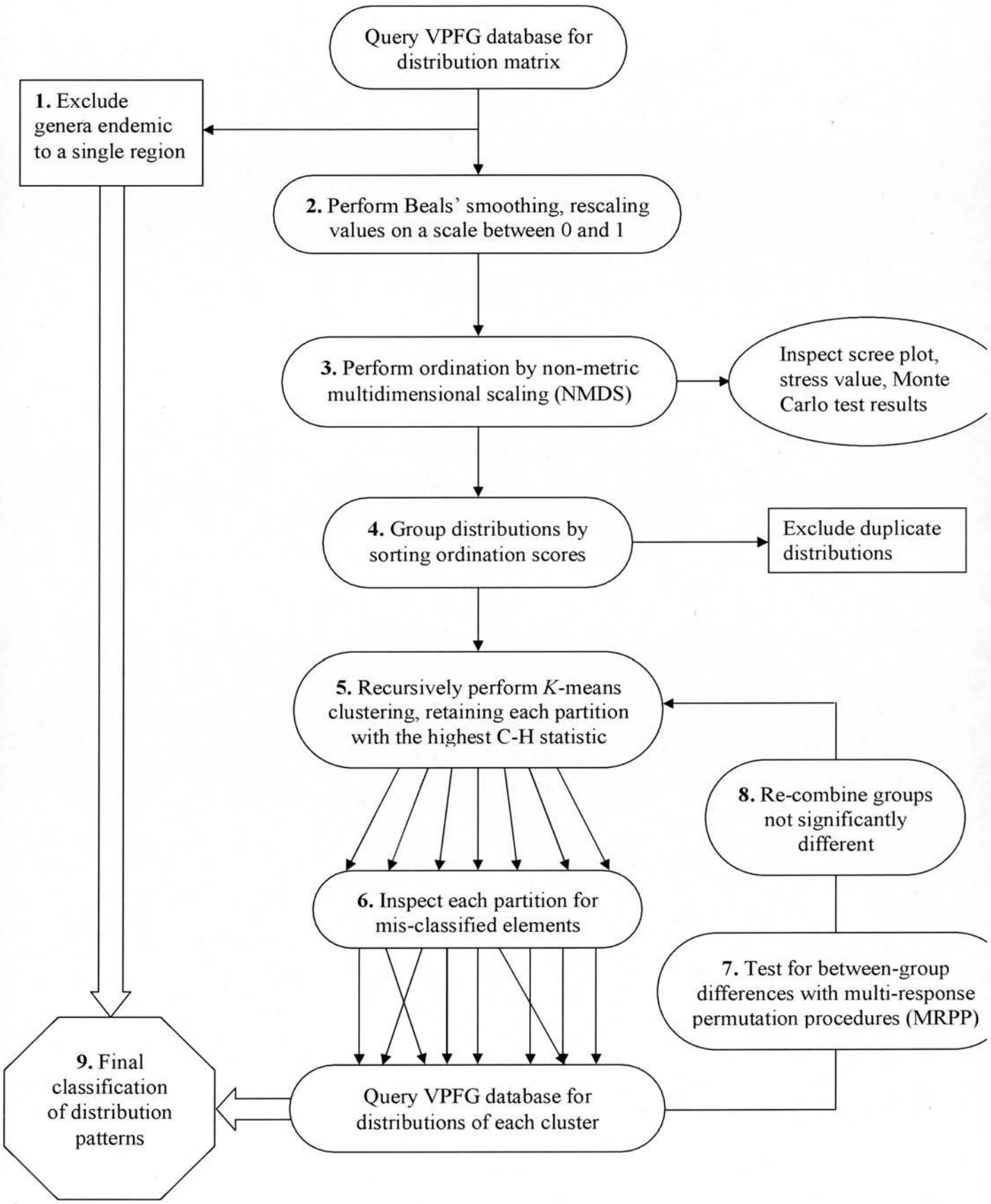


Figure 6.2 Schematic flow diagram of methodology followed in the analysis of distribution patterns.

6.3 Results

6.3.1 Ordination of families

A scree-plot for the final solution found by NMDS is shown in Figure 6.3. A 3-dimensional solution provided the optimal reduction in stress; with additional dimensions there was little subsequent reduction. The scree-plot also shows the values of the randomised data; results from a Monte Carlo test are given in Table 6.1, with stress values for each dimension significantly different from those for randomised data. Figure 6.4 shows the decline in stress with number of iterations. Although it had declined rapidly up to 100 iterations, the degree of stress was still fluctuating markedly until the maximum number (400) of iterations had been reached, despite not falling below a value of about 8. This is in contrast to the ordination of regions in Chapter 5, where the number of iterations terminated at 80 since beyond 65 iterations there was no further change in the value of stress. The stress value for the final solution is 8.19, and instability is 0.033, much higher than that for the ordination of regions (stress = 1.83), although Clark (1993) claims that with stress values below 10 there is 'no real risk of drawing false inferences', while McCune and Grace (2002) quote typical stress values of between 10 and 20 for ecological community data. Each axis represents roughly equal proportions of the variation, with a total cumulative variance of 94% explained. The ordination diagram is given in Figure 6.5.

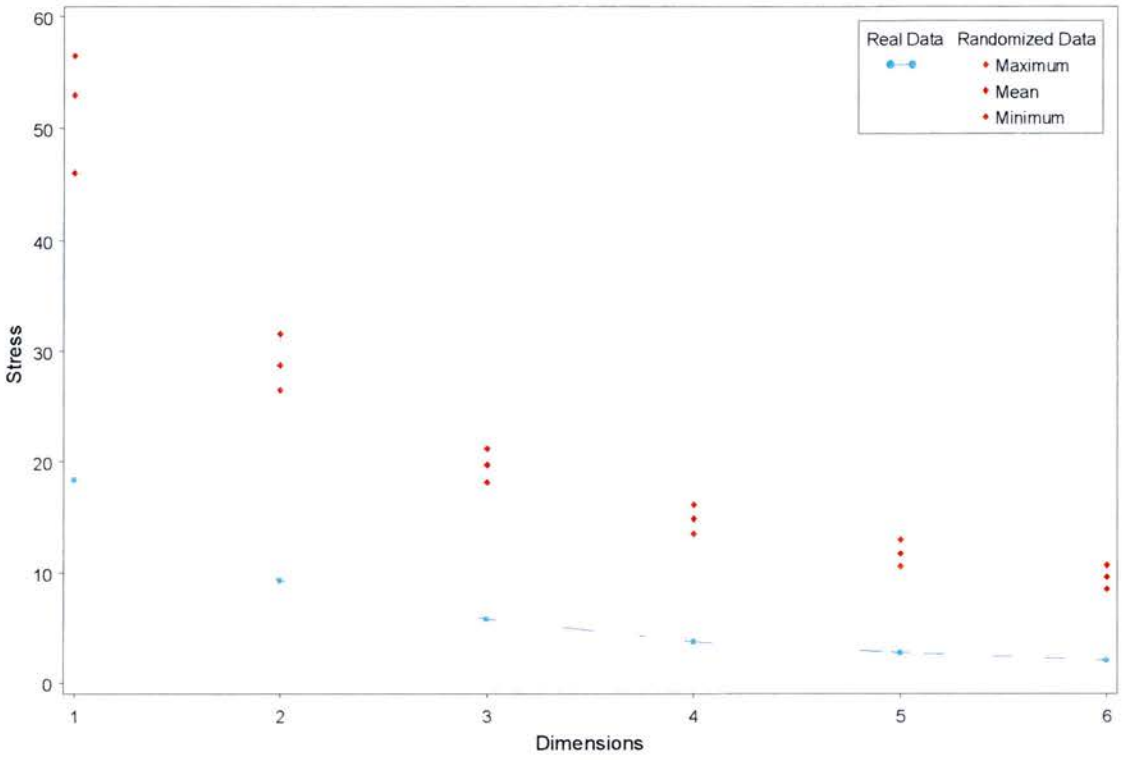


Figure 6.3 Scree-plot of reduction in stress with dimensionality from non-metric multidimensional scaling ordination of 414 APG families. The optimal reduction of stress was found with a 3-dimensional solution

	Stress in real data (40 runs)			Stress in random data (50 runs)			
No. axes	**Minimum**	**Mean**	**Maximum**	**Minimum**	**Mean**	**Maximum**	***p***
1	42.67	48.95	57.97	54.74	56.02	57.59	< 0.02
2	14.62	18.21	22.26	37.94	38.55	41.91	< 0.02
3	8.19	11.81	29.46	29.77	30.08	30.36	< 0.02
4	6.93	13.99	47.69	24.75	27.07	27.54	< 0.02
5	10.79	17.79	49.98	21.32	23.36	24.05	< 0.02
6	9.76	15.32	43.27	18.82	21.02	21.59	< 0.02

Table 6.1 Monte Carlo test results from non-metric multidimensional scaling ordination of family distributions. p = proportion of randomized runs with stress < or = observed stress i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations).

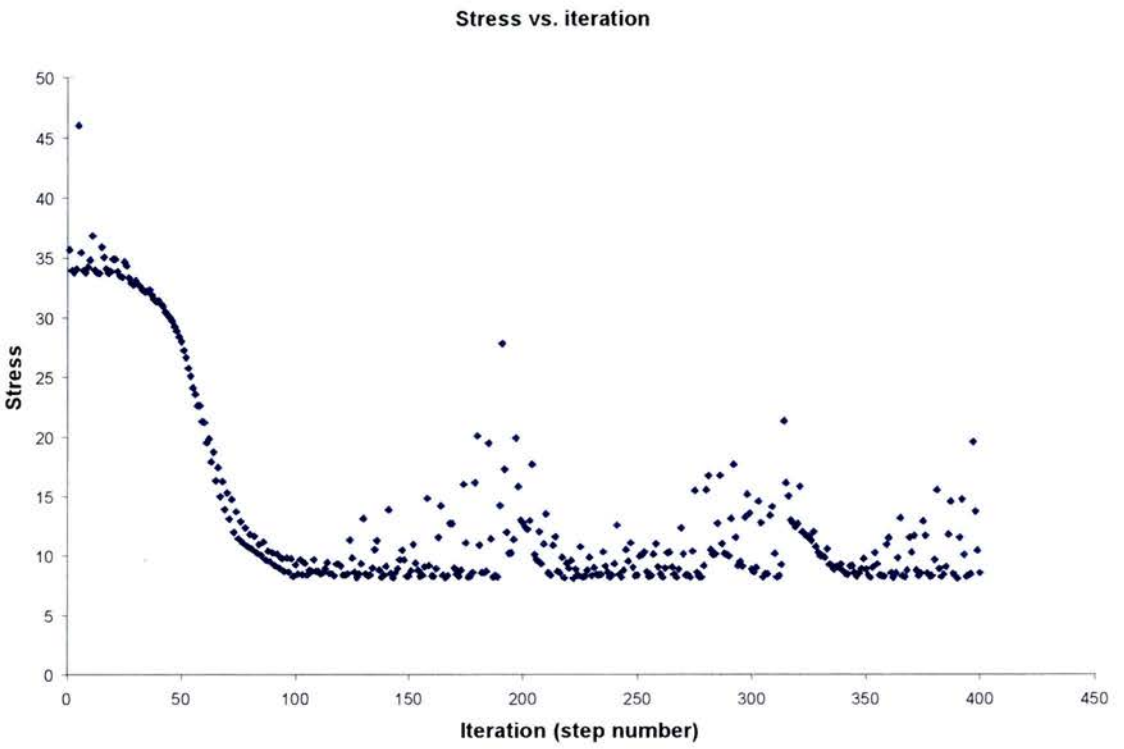


Figure 6.4 Plot of reduction in stress with iteration from non-metric multidimensional scaling ordination of 414 APG families. Stress was still fluctuating despite the maximum number of iterations having been reached.

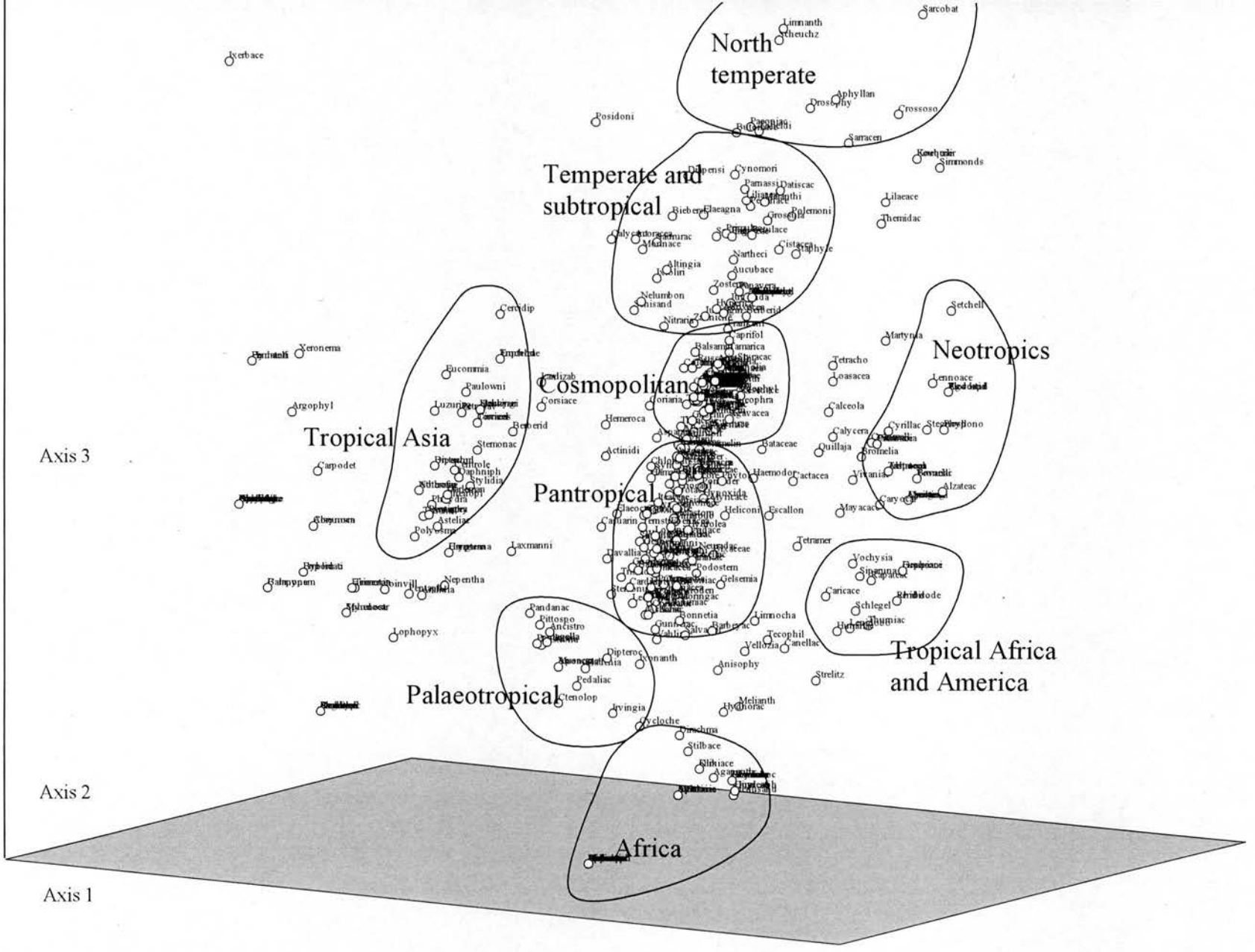


Figure 6.5 Non-metric multidimensional ordination plot of APG families. Groups are identified for illustration, but do not correspond to the classification of family distributions.

6.3.2 Ordination of genera

A scree-plot for the final solution found by NMDS is shown in Figure 6.6. A 3-dimensional solution again provided the optimal reduction in stress; with additional dimensions there was little subsequent reduction. The scree-plot also shows the values of the randomised data; results from a Monte Carlo test are given in Table 6.2, with stress values for each dimension still significantly different from those for randomised data. Figure 6.7 shows the decline in stress with number of iterations for generic data. Although there was a decline up to 30 iterations, the degree of stress shows no sign of stabilising before the maximum number (200) of iterations is reached. This is in contrast to the ordination of families, where stress seems to be converging on a stable value even though this has not been reached by 400 iterations. However, for generic data the solution does seem to be stable despite the fluctuations in stress, as shown by Figure 6.8. Ordination statistics for the final solution are given in Table 6.3, with those for family data. The stress value for the final solution is 9.76, and instability is 0.028; the stress is still below 10 and so with 'no real risk of drawing false inferences' (Clark, 1993), and still below those stress values of between 10 and 20 quoted by McCune and Grace (2002) as typical for ecological community data. Each axis represents roughly equal proportions of the variation, with a total cumulative variance of 91% explained. It seems remarkable that this technique can give such a robust and stable solution with such a large and complex data set – since the number of units (14,304 genera) being analysed is so much greater than is the number of attributes (52 regions) there is little discriminatory power with which to distinguish different distribution patterns. In contrast, with the floristic relationships studied in the previous chapter, each region contained so many genera that differences in composition between regions were marked. The ordination diagram for the genus-level analysis of distribution patterns is given in Figure 6.9.

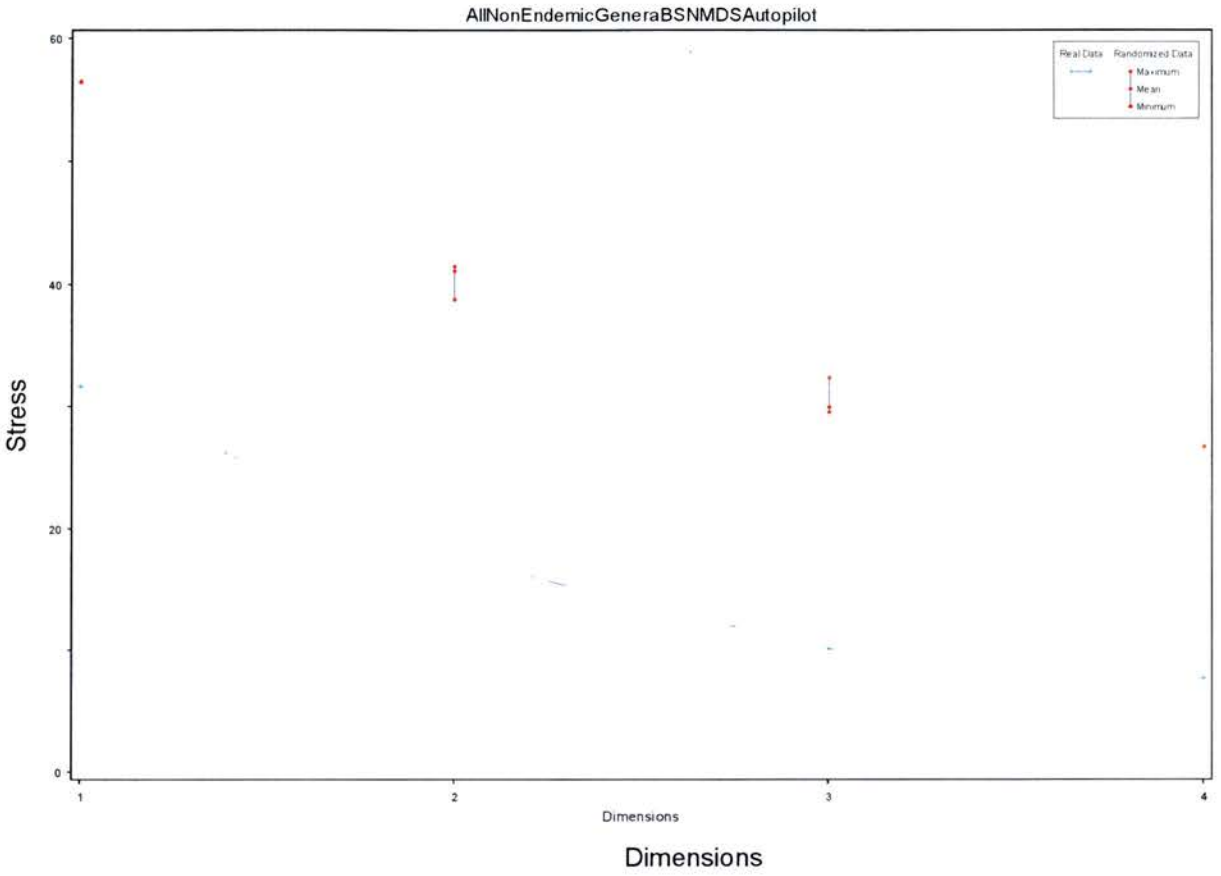


Figure 6.6 Scree-plot of reduction in stress with dimensionality from non-metric multidimensional scaling ordination of 8188 non-endemic genera. A 3-dimensional solution provides the optimal reduction of stress.

	Stress in real data (15 runs)			Stress in random data (30 runs)			
No. axes	**Minimum**	**Mean**	**Maximum**	**Minimum**	**Mean**	**Maximum**	***p***
1	31.70	41.40	46.92	56.46	56.52	56.56	≤0.0323
2	17.62	19.49	23.17	38.77	41.13	41.50	≤0.0323
3	9.76	13.82	22.32	29.59	30.01	32.40	≤0.0323
4	7.76	12.46	22.98	26.72	26.74	26.75	≤0.0323

Table 6.2 Monte Carlo test results from non-metric multidimensional scaling ordination of genus distributions; fewer runs were performed than with family distributions due to the extreme slowness of processing such a large matrix. p = proportion of randomized runs with stress < or = observed stress i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations).

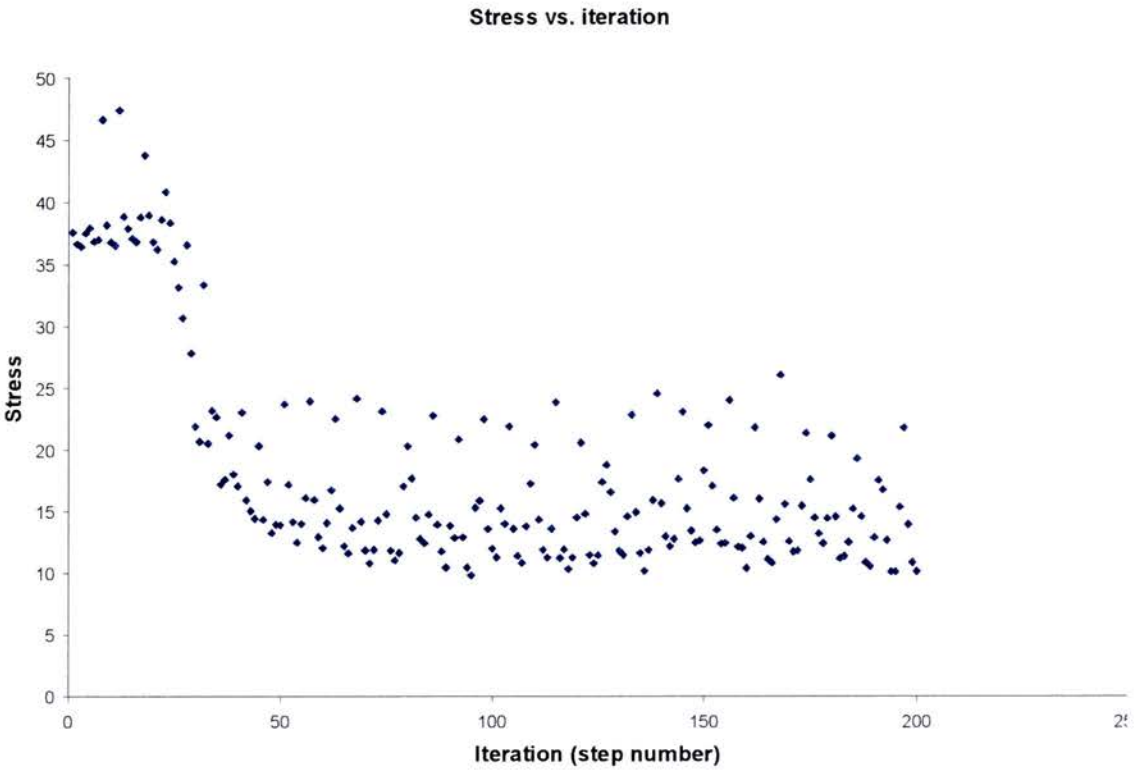


Figure 6.7 Plot of reduction in stress with iteration from non-metric multidimensional scaling ordination of 8188 non-endemic genera. Stress had not yet stabilised despite reaching the maximum number of iterations.

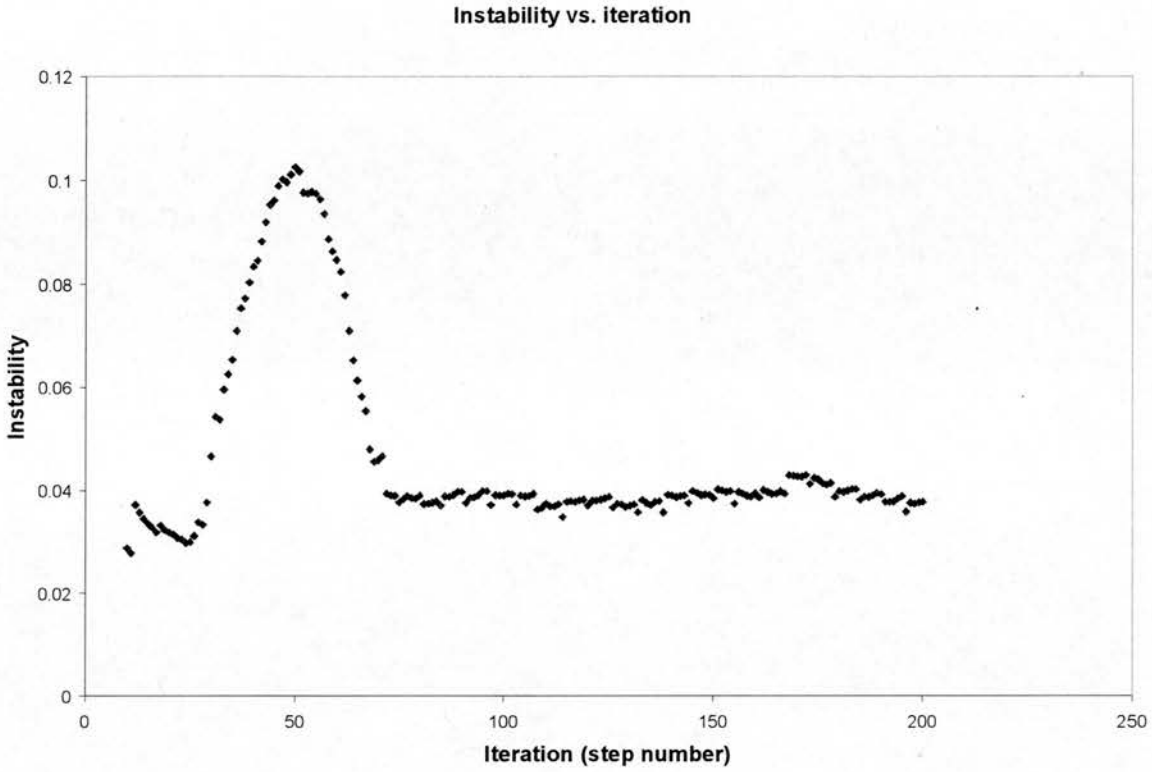


Figure 6.8 Plot of change in instability with iteration from non-metric multidimensional scaling ordination of 8188 non-endemic genera. Instability settles around 0.04 after 70 iterations, despite stress continuing to fluctuate (see Figure 6.7).

Matrix	Stress value	Instability value	No. of iterations	Axis 1 (r^2)	Axis 2 (r^2)		Axis 3 (r^2)	
				Increment	Increment	Cumulative	Increment	Cumulative
Families	8.19	0.033	400	0.31	0.35	0.66	0.28	0.94
Genera	9.76	0.028	200	0.27	0.42	0.69	0.22	0.91

Table 6.3 Ordination statistics from final best runs of non-metric multidimensional scaling for matrices of distributions of both families and genera.

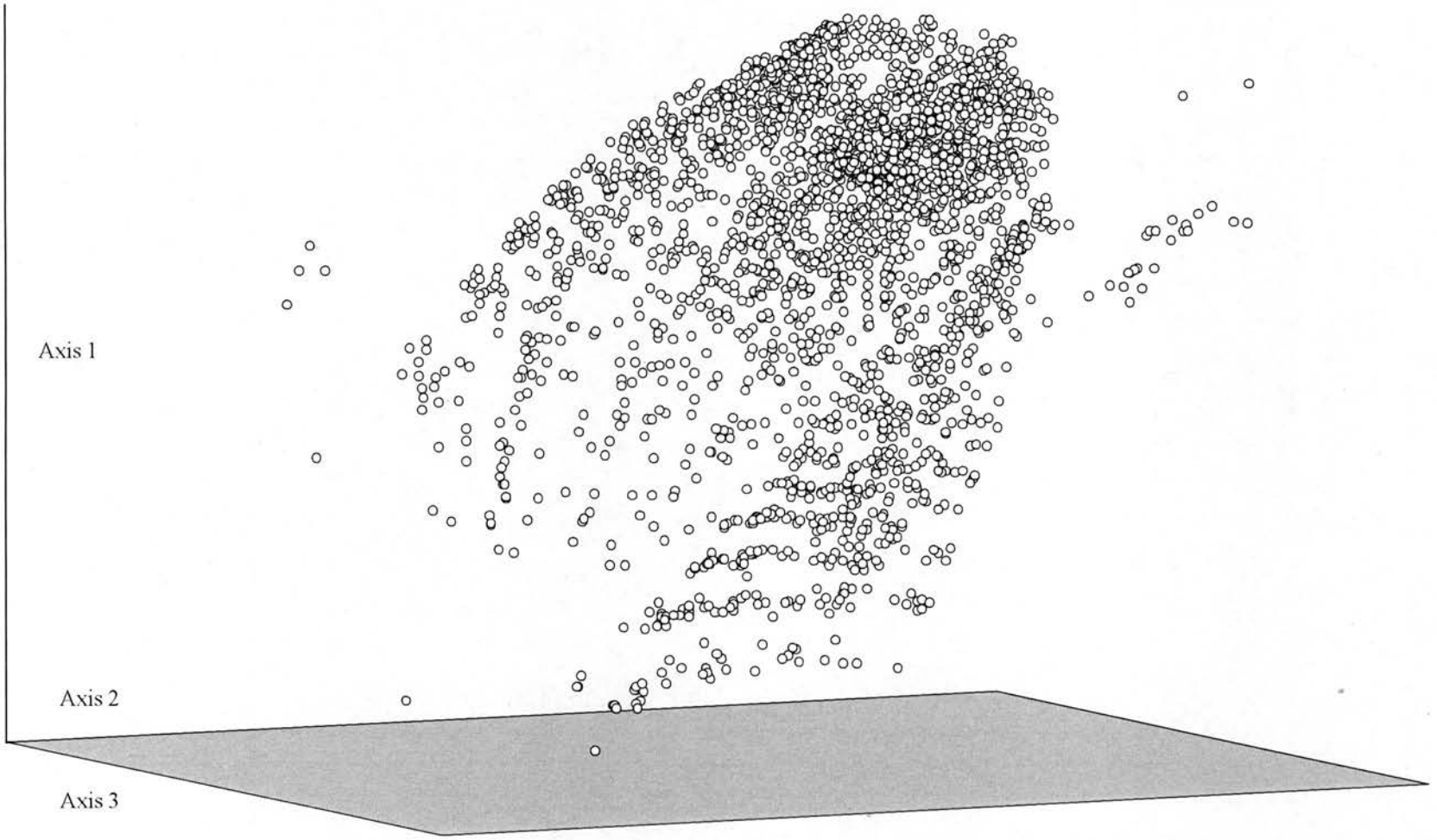


Figure 6.9 Non-metric multidimensional scaling ordination plot of 8188 non-endemic genera; individual points represent unique distribution patterns. Obvious groups are not apparent, indicating how much different distribution patterns overlap, but small clusters can be seen at finer scales, corresponding to groups found by *k*-means partitioning.

6.3.3 Interpretation of the ordination plots

The ordination of families shown in Figure 6.5 is a 2-dimensional representation of a 3-dimensional ordination plot; therefore the precise relationships between the positions of the families may be obscured by the lack of 'depth' inherent in a flat piece of paper. Geographical groups have been identified in Figure 6.5, although these are purely for illustration and do not necessarily correspond to the groups found in subsequent *k*-means partitioning. What they illustrate, however, is the basic principle behind the ordination – that families close together in the ordination plot are close together geographically. A dense group of cosmopolitan families occupies the centre of the ordination space, while the three tropical areas of Asia, Africa and the Neotropics occupy different 'faces' of the 3-dimensional space. Pantropical families are positioned 'between' all three of these tropical areas in the ordination space, while intermediate tropical distributions are found 'between' the respective tropical areas. Palaeotropical families, for example, are found between tropical Asia and tropical Africa, while families from both tropical Africa and tropical America are found between these two groups. Many more detailed groups can be identified, and in addition some families remain isolated between clusters and do not fit easily into any one group.

In Table 6.4 the 5 highest- and lowest-scoring families, and their distribution, for each axis of the ordination are given. What is most notable in this table is that some families occur at one of the ends of each of the three axes: Aphyllanthaceae from the Mediterranean region as a whole and Drosophyllaceae from SW. Europe and Morocco are two of the highest scoring taxa on both axes 1 and 3 and two of the lowest scoring taxa on axis 2. Since the highest-scoring families on Axis 2 are all from the Neotropics, those other families which score highly on Axes 1 and 3 but are found in North America (Crossosomataceae, Limnanthaceae, Sarcobataceae and Scheuchzeriaceae) do not have similarly low scores on Axis 2 as do Aphyllanthaceae and Drosophyllaceae. The endpoints of each axis are all parts of the Earth geographically distant from each other. The only exception to this is the position of Posidoniaceae (which is found disjunct between the Mediterranean region and Australia) on Axis 2; this would be because the Mediterranean region is represented as four regions (Region 12, Southwestern Europe; Region 13, Southeastern Europe; Region 20, Northern Africa; and Region 34, Western Asia), whereas Australia is only one region, so the geographical similarity of Posidoniaceae is greatest with other Mediterranean families and Posidoniaceae have been 'pulled away' from Australian families.

In turn, therefore, Axis 1 runs from, roughly speaking, north temperate regions to Australasia; Axis 2 runs from the Neotropics to the Mediterranean; and Axis 3 again runs from north temperate regions but this time to Madagascar. Also, notice that highest- and lowest-scoring families are geographically localised, with the exception of Scheuchzeriaceae (widespread in north temperate regions) and

Butomaceae (throughout temperate Eurasia); this is because the more widespread a family distribution is, the greater the number of other families with which that distribution will overlap. Therefore more widespread distributions are found more towards the middle of the ordination space (for example, the tight cluster of cosmopolitan families) since they show similarities with the largest number of other families.

The ordination for generic data is shown in Figure 6.9, except that here no groups have been picked out. This was because there does not initially seem to be any structure in the ordination plot at all, just a continuous smear of points, although on closer inspection many small clusters can be seen at a much finer scale. Genera from the 5 highest- and lowest-scoring distribution patterns are indicated in Table 6.5. The obvious difference between the genus ordination and the family ordination is that whereas cosmopolitan families formed a small cluster in the centre of the ordination space, cosmopolitan genera are here found at the end of one of the axes. However, the difference is not so simple as that; the number of cosmopolitan or widespread genera is greater than cosmopolitan families, and cosmopolitan genera do not just represent the endpoint of one of the axes, but instead form a large diffuse cluster along one side of the ordination, since the short-hand 'cosmopolitan' could apply to many different combinations of a large number of regions. Cosmopolitan genera aside, however, the other axis endpoints are represented by taxa from small, geographically-localised areas – remembering that endemic taxa have been omitted from this analysis to improve computational speed.

Since the objects in this analysis are unique distribution patterns, each object may represent many different genera, and therefore the genera listed in Table 6.5 are merely exemplars of that distribution pattern (although the textual distributions quoted in the table are the actual distributions for those genera). As with the family-level ordination, endpoints of the genus-level ordination are geographically distant areas. Axis 1 runs from Neotropical regions to S. Europe/Mediterranean and the Middle East; Axis 2 runs from Subantarctic Islands / New Zealand / Pacific Islands to Cosmopolitan; Axis 3 runs from Tropical Africa to W. North America. It is interesting that Axis 2 runs not just from geographically-localised to geographically-most-widespread, but from the most physically isolated distribution patterns to the least physically isolated. This may be because for the other geographically-localised, but less geographically isolated, patterns at the endpoints of the other axes there is a greater degree of overlap with other patterns.

Family	Score on axis	Distribution
Axis 1 – highest		
Sarcobataceae	1.2647	W. U.S.A.
Aphyllanthaceae	1.2257	Mediterranean region
Drosophyllaceae	1.1639	SW. Europe, Morocco
Scheuchzeriaceae	0.8822	Widespread in north temperate regions
Crossosomataceae	0.8511	W. U.S.A. and into Mexico
Axis 1 – lowest		
Ixerbaceae	-1.5069	New Zealand
Hydatellaceae	-1.4593	S. Australia and Tasmania, New Zealand
Pennantiaceae	-1.4593	E. Australia and New Zealand
Xeronemataceae	-1.4566	New Caledonia, New Zealand
Akaniaceae	-1.4176	Australia
Axis 2 – highest		
Peridiscaceae	1.0961	Venezuela, Guyana and Brazil
Rhabdodendraceae	1.0961	Brazil and the Guianas
Euphroniaceae	1.0806	Colombia, Venezuela and Brazil
Goupiaceae	1.0806	Colombia, Peru, Venezuela and the Guianas, Amazonian Brazil
Alzateaceae	1.0519	Costa Rica to Bolivia
Axis 2 – lowest		
Drosophyllaceae	-1.0552	SW. Europe, Morocco
Aphyllanthaceae	-1.0115	Mediterranean region
Cynomoriaceae	-0.9399	Canary Is. and Mediterranean region east to Mongolia
Posidoniaceae	-0.8622	Mediterranean region; Australia
Butomaceae	-0.8532	Throughout temperate Eurasia
Axis 3 – highest		
Sarcobataceae	1.4701	W. U.S.A.
Scheuchzeriaceae	1.3867	Widespread in north temperate regions
Limnanthaceae	1.3775	British Columbia to California
Aphyllanthaceae	1.2801	Mediterranean region
Drosophyllaceae	1.2589	SW. Europe, Morocco
Axis 3 – lowest		
Asteropeiaceae	-1.0897	Madagascar
Barbeuiaceae	-1.0897	Madagascar
Didiereaceae	-1.0897	Madagascar
Melanophyllaceae	-1.0897	Madagascar
Physenaceae	-1.0897	Madagascar

Table 6.4 The five highest- and lowest-scoring families for each axis of the ordination. Each family was a separate object in the ordination; where the distribution is the same, the score on each axis will be identical for those families.

Genus	Score on axis	Distribution
Axis 1 – highest		
Lophopterys	0.8259	Venezuela and the Guianas, Brazil
Acioa	0.8259	Brazil, Venezuela and French Guiana
Caryomene	0.7775	Peru and Brazil
Mona	0.7746	Colombia and Venezuela
Diogenesia	0.7746	Colombia and Venezuela, Ecuador, Peru
Axis 1 – lowest		
Ceratocarpus	-0.8483	S. Europe and Middle East to Siberia
Ventenata	-0.8464	S. Europe and Middle East
Sternbergia	-0.8460	S. Europe and Middle East
Myrrhoides	-0.8459	Mediterranean and Middle East
Xeranthemum	-0.8447	Mediterranean and Middle East
Axis 2 – highest		
Pleurophyllum	2.2059	Antipodean Is. and Macquarie I.
Stilbocarpa	1.9013	Stewart I., Antipodean Is. and Macquarie I.
Morelotia	1.7926	New Zealand; Hawaiian Is.
Kadua	1.7356	Hawaiian Is., Society Is. and Tubuai Is.
Nesoluma	1.6492	Hawaiian Is., Society Is., Rapa, Pitcairn I.
Axis 2 – lowest		
Cyperus	-1.8728	Cosmopolitan
Rubus	-1.7974	Cosmopolitan
Poa	-1.7884	Cosmopolitan
Oxalis	-1.7365	Cosmopolitan
Polygala	-1.7363	Cosmopolitan
Axis 3 – highest		
Poggea	0.7544	Cameroon to Zaire and Angola
Viridivia	0.7367	SW. Tanzania and Zambia
Cleistochlamys	0.7367	S. Tanzania to Zimbabwe and Mozambique
Diogea	0.7337	Nigeria to Zaire
Octolobus	0.7268	Sierra Leone to Angola
Axis 3 – lowest		
Canbya	-0.7389	Oregon and California
Nothochelone	-0.7298	W. North America
Redfieldia	-0.7149	Colorado and Utah to Oklahoma and Kansas
Jamesia	-0.7147	California to Wyoming and New Mexico
Podistera	-0.7060	Alaska to California

Table 6.5 The five highest- and lowest-scoring generic distribution patterns for each axis of the ordination. Since the units in the genus-level analysis were the 2817 unique distribution patterns, each of these may represent many different genera; the generic names quoted are therefore merely illustrative, although the distributions quoted are correct for those genera.

6.3.4 *k*- means partitioning

The complete breakdown of *k*-means partitioning is presented here for the more complex genus-level analysis. This gives the flavour of the recursive approach, where each partition with the maximum Calinski-Harabasz pseudo-*F*-statistic was saved after each pass through the data, and this partition then partitioned further until the optimum number of groups was reached. An example table of Calinski-Harabasz pseudo-*F*-statistic scores from one of the partitions is given in Table 6.6; the values peak at an intermediate number of clusters (13). To determine the optimum number of groups, each pair of groups was subjected to a non-parametric multi-response permutation procedure (MRPP) in PC-ORD version 4.0 (McCune & Mefford, 1999), which tests for statistically significant differences in the geographical distributions between groups. Three examples are given in Table 6.7; there is not enough room to reproduce the entire half-matrix (it contains 16,200 cells). With initial passes of the data, groups were, not surprisingly, highly significantly different. However, the test was applied to all pairs of groups after each pass of *k*-means partitioning, since similar distribution patterns could be partitioned into different groups in an earlier pass through the data (i.e. a group of genera from the fourth pass of *k*-means partitioning might be geographically similar to another group of genera from the fourth pass but which had already been partitioned into a different group at the third pass – therefore the MRPP test was applied to all pairs of groups from all partitions after each pass, and not just to the groups within that partition). As partitioning progressed through successive passes, groups became progressively less statistically distinct until, on the 5th pass, a large proportion of groups no longer satisfied the test of statistical distinctness. These groups were then re-combined manually and *k*-means partitioning halted.

Complete breakdown of *k*-means partitioning of genus distribution patterns

For each partition the output lists: number of objects (genera) in that partition; the number of groups found within that partition; the value of the error sum of squares (SSE) of that result; the value of the Calinski-Harabasz (C-H) pseudo-*F*-statistic of that result; group membership indicates the number of objects (genera) in each group found within that partition.

1st Pass

Run #1 2679 distributions 2 clusters
K = 2 groups: SSE = 956.37032 C-H = 1921.92178 (5 iterations) (Random start 1)
Group membership: 1554 1215

2nd Pass

Run #1 1554 distributions 4 clusters
K = 4 groups: SSE = 102.28752 C-H = 1241.04595 (5 iterations) (Random start 1)
Group membership: 527 372 397 258

3rd Pass

Run #1 527 distributions 9 clusters
K = 9 groups: SSE = 5.32083 C-H = 342.94145 (9 iterations) (Random start 3)
Group membership: 137 64 81 31 27 46 2 98 41

Run#2 372 distributions 2 clusters
K = 2 groups: SSE = 15.31855 C-H = 253.36977 (4 iterations) (Random start 1)
Group membership: 166 206

4thPass

3rd pass Run#2 Cluster#1 171 distributions 2 clusters
K = 2 groups: SSE = 6.06635 C-H = 140.32957 (3 iterations) (Random start 24)
Group membership: 100 71

3rd pass Run#2 Cluster#2 209 distributions 20 clusters
K =20 groups: SSE = 0.50627 C-H = 203.70104 (3 iterations) (Random start 6)
Group membership: 9 11 6 10 8 23 24 13 15 12 17 6 1 14 10 12 8 1 1 8

3rd Pass

Run#3 397 distributions 2 clusters
K = 2 groups: SSE = 10.47192 C-H = 401.82636 (2 iterations) (Random start 1)
Group membership: 215 182

4thPass

3rd pass Run#3 Cluster#1 196 distributions 3 clusters
K = 3 groups: SSE = 1.83966 C-H = 180.87549 (5 iterations) (Random start 2)
Group membership: 74 75 47

3rd pass Run#3 Cluster#2 160 distributions 5 clusters
K = 5 groups: SSE = 0.49505 C-H = 202.04950 (2 iterations) (Random start 5)
Group membership: 48 40 31 24 17

3rd Pass

Run#4 258 distributions 10 clusters
K =10 groups: SSE = 3.27940 C-H = 156.00186 (9 iterations) (Random start 36)
Group membership: 53 14 16 1 24 32 31 47 29 11

2nd Pass

Run #2 1215 distributions 5 clusters
K = 5 groups: SSE = 172.55158 C-H = 764.02269 (5 iterations) (Random start 6)
Group membership: 320 343 289 130 133

3rd Pass

Run#5 320 distributions 13 clusters
K =13 groups: SSE = 3.44836 C-H = 207.53329 (13 iterations) (Random start 21)
Group membership: 14 22 38 28 25 11 22 33 13 36 21 36 21

Run#6 343 distributions 3 clusters
K = 3 groups: SSE = 22.50246 C-H = 225.71121 (10 iterations) (Random start 1)
Group membership: 126 108 109

4thPass

3rd pass Run#6 Cluster#1 92 distributions 18 clusters

K =18 groups: SSE = 0.12751 C-H = 109.89093 (2 iterations) (Random start 72)
Group membership: 9 2 7 2 5 5 4 6 4 4 8 6 2 4 1 9 9 5

3rd pass Run#6 Cluster#2 93 distributions 17 clusters
K =17 groups: SSE = 0.24893 C-H = 119.55667 (3 iterations) (Random start 28)
Group membership: 12 10 5 3 12 3 5 5 4 6 3 4 4 1 2 12 2

3rd pass Run#6 Cluster#3 102 distributions 16 clusters
K =16 groups: SSE = 0.24530 C-H = 120.15724 (2 iterations) (Random start 86)
Group membership: 5 4 6 7 6 7 9 1 4 7 8 8 9 10 4 7

3rd Pass

Run#7 289 distributions 19 clusters
K =19 groups: SSE = 1.96224 C-H = 227.28931 (5 iterations) (Random start 56)
Group membership: 24 20 21 6 18 11 17 11 14 13 18 21 5 13 13 24 8 19 13

Run#8 130 distributions 16 clusters
K =16 groups: SSE = 1.81672 C-H = 117.56827 (2 iterations) (Random start 35)
Group membership: 9 18 13 5 7 8 1 2 14 1 18 12 6 3 8 5

Run#9 133 distributions 19 clusters
K =19 groups: SSE = 1.39559 C-H = 116.79102 (2 iterations) (Random start 56)
Group membership: 15 16 1 9 2 22 6 7 9 5 1 7 9 3 1 10 2 2 6

No. of groups (k)	C-H pseudo-F-statistic
2	80.46
3	153.46
4	203.69
5	188.20
6	173.31
7	195.56
8	179.08
9	192.55
10	203.75
11	205.14
12	204.17
13	207.53
14	205.54
15	203.41
16	202.43
17	200.24
18	197.66
19	194.59
20	195.87
21	193.19
22	193.55
23	192.97
24	192.42
25	188.91
26	187.64
27	182.34
28	176.71
29	175.71
30	172.07

Table 6.6 Calinski-Harabasz psuedo-F-statistic scores from the partition of the 3rd Pass Run#5 cluster. The C-H score peaks at 13 groups, indicating optimal within-group clustering and between-group separation.

Distribution pair	Test statistic	Within-group agreement	p value
Cosmopolitan / north temperate	-113.9855	0.3847	0.0000
Warm temperate / pantropical	-7.1793	0.0387	0.0000
Mediterranean to W. Himalaya/ Mediterranean to C. Asia + Him.	-0.7872	0.0267	0.1799

Table 6.7 Examples of pair-wise MRPP test scores of between-group distinctness. The first pair are highly significantly different; the last pair are not significantly different from each other, and were combined.

65 groups were found for 414 angiosperm families; 181 floristic elements were found for c. 14,034 angiosperm genera. For families, the cosmopolitan distribution pattern was by far the most common, whereas for genera, distributions throughout one tropical region or endemic to a single region remain the most common. This reflects the differences seen in the respective range-size frequency distributions for families and genera (see Chapter 3). Given that family distributions are the sum of the distributions for all genera in that family, it is not surprising that more families are more widespread than are genera. Table 6.8 repeats the 12 most frequent distribution patterns for angiosperm genera found by *k*-means partitioning. Not only is tropical South America the most common distribution pattern found by this analysis, but two other widespread Neotropical patterns are within the top twelve. Distributions endemic to either Region 83, Western South America, or Region 84, Brazil, are also in Table 6.8; since each of these regions also present in the three more widespread Neotropical patterns, both Regions 83 and 84 are therefore represented by 4 distribution patterns out of the 12 most common.

Distribution pattern	Number of genera
Tropical South America	682
Endemic to Australia	669
Endemic to Southern Africa	587
Throughout tropical Africa	500
Endemic to Brazil	414
Endemic to Madagascar or Mascarene Is.	403
Tropical Africa	343
Central America and tropical South America	317
Endemic to Western South America	311
Mexico and Caribbean south to Southern South America	306
Endemic to Mexico	270
S. China and throughout tropical Asia	264

Table 6.8 Twelve most-frequent distribution patterns found by *k*-means partitioning of genus distributions.

Complete schemes of distribution patterns found by *k*-means partitioning for both all angiosperm families and all angiosperm genera are presented as Appendix 2 and as Appendix 3, respectively. Family delimitation here follows the Angiosperm Phylogeny Group (APG, 1998), which in some cases has resulted in widely different circumscriptions from the traditional concepts of those families. The schemes are deliberately non-hierarchical, although geographically similar distribution patterns are grouped together for comparative purposes. It is important to note that what seem to be geographically nested

distributions are here treated independently. For example, there is a distribution of 'Eastern Asia and North America', as well as a distribution pattern of 'China, Japan and E. U.S.A.'. Also, although some of the distribution patterns are indeed very similar to each other, the limitations of briefly summarising a complex pattern with a simple textual description mean that some distributions appear more similar than the underlying data actually are. Choropleth maps showing numbers of genera in each region for each floristic element for all angiosperm genera are given in Appendix 4.

6.4 Discussion

6.4.1 The effectiveness of the analysis

Trying to summarise simply the overall results of such a complex analysis is extremely difficult. Figure 6.10 shows, however, that even after *k*-means partitioning has reduced the size of the data from 2817 unique distribution patterns to fewer than 200 distinct geographical clusters (= floristic elements), the frequency distribution is still extremely skewed; the majority of diversity is still accounted for by a few distribution patterns. Only 38 distribution patterns out of 181 (21%) produced by the *k*-means partitioning are shared by 100 genera or more; but together these 38 distribution patterns account for 8785 genera, or 61% of the total number of all angiosperm genera. With no *k*-means partitioning, 20% of unique distribution patterns accounted for 80% of generic diversity (see Chapter 3). The *k*-means partitioning analysis has therefore grouped the bulk of the unique distribution patterns, while maintaining the overall geographical structure within the dataset. Inspecting the distribution patterns of individual floristic elements, the majority of unique distribution patterns are clearly within the broader, more widespread floristic elements such as 'cosmopolitan', 'north temperate' or 'pantropical'. However, the number of possible unique combinations of regions will rise from 52 possible single-region endemic distributions to peak at an intermediate number of regions, (i.e. c. 26 out of 52), before declining again to a single possible distribution pattern for all 52 regions, so the greatest heterogeneity in distribution patterns should be for those taxa of intermediate range-size. This is not actually shown by these data: distributions of intermediate range size are not the most variable geographically. That the greatest heterogeneity in this dataset is found for widespread genera is because the majority of distribution patterns for genera of intermediate range-size would be widely disjunct distributions which are not actually found in nature.

The results of the *k*-means partitioning analysis, however, reveal several unfortunate drawbacks with the technique. The main drawback was the so-called problem of 'hyper-sphericity' (McCune & Grace, 2002): the tendency for the groups found within the 3-dimensional ordination space to be as

spherical as possible, when the actual geographical similarity between those entities was not represented by a spherical cluster. For example, in the ordination of genus distribution patterns shown in Figure 6.9, several small clusters can be made out at fine scales in which objects apparently chain together from left to right across the ordination plot. In these cases, *k*-means partitioning tended to assign some of these objects to an adjacent cluster, while sometimes including other objects from another adjacent cluster which were actually more dissimilar geographically but closer to the centroid of this group than to the centroid of that adjacent group. This meant that after each pass of *k*-means partitioning the assignment of all of the 2817 unique distribution patterns being analysed (for the genus-level analysis) had to be inspected visually and a small proportion of apparently 'mis-classified' objects re-assigned to adjacent clusters by hand. This was done by inspecting the axis scores for each axis of the NMDS ordination. The proportion of mis-classified elements in each pass ranged from just below 10% to almost 16% of all objects. In almost all cases the misclassified elements were those objects with ordination scores at the outer limit of the group composition, such that in the output result file, these were either the first or the last distribution patterns within that group, and needed to be moved to or exchanged with the first or the last elements within the preceding or succeeding group.

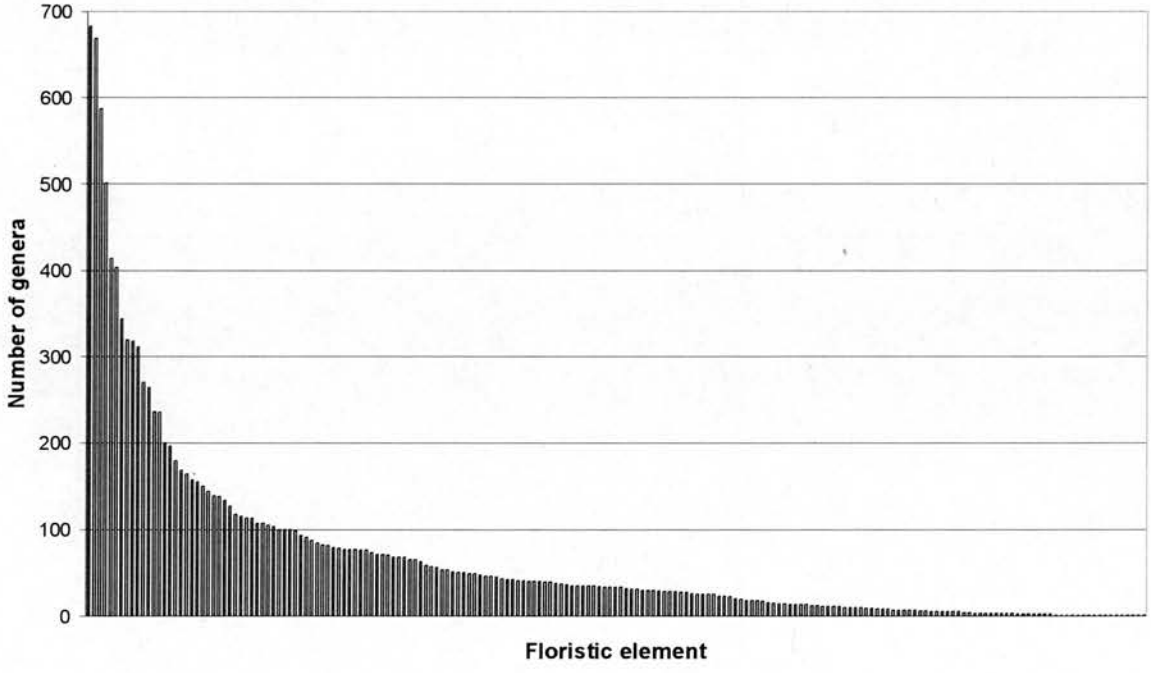


Figure 6.10 Taxon-size frequency distribution for genus distributions found by *k*-means partitioning; the frequency distribution remains extremely skewed, with 20% of distribution patterns still accounting for as much as two-thirds of the number of genera.

When looked at in their entirety, there is no single distribution which is distinct from every other distribution – they all overlap to some degree with at least some others (see Appendix 4). If there were any completely non-overlapping distributions, then that would mean that those regions would not show any floristic relationships with any other region; the analysis of floristic relationships (Chapter 5) shows that this cannot be the case – except for the Antarctic Continent (which was omitted from this analysis), all regions show some degree of floristic similarity with every other region. Some floristic elements are geographically extremely similar to others (see, for example, elements 12, 13 and 14 in Appendix 4); although the *k*-means partitioning analysis has distinguished between these as being distinct, they might not appear so to the observer. This might be because the *k*-means partitioning was implemented recursively, taking each initial partition and partitioning that separately from the others, thereby effectively imposing a divisive hierarchy on the classification which the data may not merit. From a pragmatic point of view, however, since the initial partitions created only very few groups, there was little alternative than to adopt this approach. The consequence may have been forcing of similar distributions into separate groups; however, the MRPP test should have led to very similar groups being recombined and leaving only those which remain significantly different from each other in their distribution.

For distributions of small range-size, however, the *k*-means partitioning analysis cannot distinguish between several unique but overlapping combinations of only two or three regions; the degree of overlap with only few regions is proportionally greater than for most distributions found in larger numbers of regions. This means that most small distributions are grouped with other small distributions (e.g. element 174, Appendix 4), when at finer geographical scales they may well appear distinct from each other. This has therefore created a disjunction in the range-sizes of floristic elements, since the class of regional endemics was initially excluded from the analysis, and later re-combined with the *k*-means partitioning analysis of all non-endemic genera. In the *k*-means partitioning analysis there are few floristic elements which are shown from fewer than three regions, whereas there are then an additional 46 endemic distributions each from one single region. Given the shape of the range-size frequency distribution (Chapter 3), this cannot be because of a lack of genera known from only two regions – it is just that these 2-region distributions are submerged within other floristic elements in the *k*-means partitioning.

It might therefore be more appropriate to keep the endemic genera separate and describe the results of the *k*-means partitioning as a classification of non-endemic distribution patterns. As defined in Chapter 1, however, floristic elements are explicitly regarded as non-endemic distribution patterns. The upshot of this is that very small distributions, which account for many genera, will be clustered together, so perhaps artificially increasing the size (frequency) of those distribution patterns. Intuitively one might be inclined to break up these large clusters of small distributions, but doing this was not supported either by the *k*-means partitioning procedure nor by the MRPP test. This is perhaps why the repeating

distribution patterns found do not correspond easily with schemes of floristic regions (phytochoria), although of course the resolution of the analysis (both geographically and taxonomically) will also mitigate against this. Floristic regions (e.g. Takhtajan, 1986) are comparable in size but not circumscription to the 52 TDWG Level 2 regions used in this analysis – most floristic regions are large but cross international boundaries – so even though small-range distributions are the most common, it is difficult to explicitly match these distributions to floristic regions and say whether or not those genera are confined to those regions. For example, a genus found only in Regions 83 (Western South America) and 84 (Brazil) could be endemic to either the Amazonian Region or to the Brazilian Region (Takhtajan, 1986), but there is not enough detail in the distribution data to be able to say which.

The final drawback with the analysis was the discovery of nested floristic elements which occur in the same areas but differ in range-size. For example, genera only known from part of tropical Africa and part of SE. Asia (see element 87, Appendix 4) have been grouped as distinct from genera which were strictly confined to the Palaeotropical region but were widespread through it (see element 84, Appendix 4), which have been distinct from genera widespread through Palaeotropical and warm temperate regions (see element 83, Appendix 4). Although it cannot be denied that there are true differences in range-size between taxa with geographically-similar distributions, a less-stringent classification may have been content to leave all these as simply Palaeotropical. These nested distributions occur because there was no relativization of taxa by distribution size prior to the ordination; it was felt that standardizing the distribution size by relativizing the row totals in the ordination was introducing an additional assumption about what constituted 'similar' geographical distributions, whereas it would be better to let the data 'speak for themselves'.

There are two further drawbacks which are a consequence of the data used rather than the analytical techniques: firstly, a floristic element revealed by this analysis may well consist of genera which still vary in their geographical distribution (since the aim was to group similar-but-different distribution patterns, that should be expected), but also genera with identical distributions as TDWG regions may in fact differ ecologically within those regions – genera within a floristic element may therefore not actually be ecologically consistent with each other; secondly, genera which simply do not match anything else are returned as belonging to no other group (see, for example, elements 62 – 66, Appendix 4). The first of these problems is an artefact of this analysis which can only be resolved with more-detailed data; the second is an inherent problem of noise in broad-scale distributional data.

6.4.2 Comparison of family-level *vs.* genus-level analysis

The main difference between the family-level and genus-level analyses – the high proportion of cosmopolitan families compared to genera – is due to the differences in their respective range-size frequency distributions (see Chapter 3). The range-size frequency distribution of genera is much more strongly skewed than it is for families. Most genera are (relatively) localised, found in only a few regions (see Table 6.8), whereas at family level 'cosmopolitan' distributions are the most common (i.e. the family range-size frequency distribution is not so skewed; see Chapter 3 and Appendix 2); for genus-level data, on the other hand, 'cosmopolitan' and other widespread distribution patterns account for only a small proportion of the total number of genera. With a greater proportion of small-range taxa, 'nesting' of distribution patterns is therefore more apparent with genus-level data than with family-level data. This inherent difference in range sizes is evident from the respective NMDS ordinations of families and genera, the axes of which represent different geographical extremes. Although no groups are indicated in the ordination plot of the genus-level data in Figure 6.9, the distributions of genera at the end-points of each axis given in Table 6.5 contrasts strongly with those for the family-level ordination given in Table 6.4.

The endpoints of each axis in the family-level ordination are all parts of the Earth which are geographically distant from each other. Axis 1 runs from, roughly speaking, north temperate regions to Australasia; Axis 2 runs from the Neotropics to the Mediterranean; and Axis 3 again runs from north temperate regions but this time to Madagascar. The highest- and lowest-scoring families are geographically localised, with the exception of Scheuchzeriaceae (widespread in north temperate regions) and Butomaceae (throughout temperate Eurasia); this is because the more widespread a family distribution is, the greater the number of other families there will be with which that distribution overlaps. Therefore more widespread distributions are found more towards the middle of the ordination space (for example, the tight cluster of cosmopolitan families) since they show great similarity with the largest number of other families. Conversely, since not many families are confined to small areas, those which are will have localised distributions distinct from the majority of other families and form the outliers in this ordination which are selected as axis endpoints.

The obvious difference between the genus ordination and the family ordination is that whereas cosmopolitan families form a small cluster in the centre of the family-level ordination space, cosmopolitan genera are here found at the end of one of the axes. With a smaller proportion of genera having large distributions than do families, cosmopolitan genera are therefore less similar to most other genera, which have smaller distributions. Cosmopolitan genera aside, however, the other axis endpoints are represented by genera from small, geographically-localised areas – remembering that endemic taxa have been omitted from this analysis to improve computational speed. As with the family-level ordination, endpoints of the

genus-level ordination are geographically distant areas. Axis 1 runs from Neotropical regions to S. Europe/Mediterranean and the Middle East; Axis 2 runs from Subantarctic Islands / New Zealand / Pacific Islands to Cosmopolitan; Axis 3 runs from Tropical Africa to W. North America. It is interesting that Axis 2 runs not just from geographically-localised to geographically most widespread, but more specifically from the most physically isolated distribution patterns to the least physically isolated. This may be because, for the other geographically-localised but less geographically-isolated patterns at the endpoints of the other axes, there is a greater degree of overlap of these latter patterns with other patterns.

In the ordination of genus-level data there does not seem to be any structure in the ordination plot at all (see Figure 6.9), just a continuous smear of points – although on closer inspection many small clusters can be seen at a much finer scale. The reason for this lack of structure is again the greater degree of geographical overlap at generic level, and especially with so many objects in the ordination (8188), so that any distribution will overlap with another to some extent. This is the problem which was illustrated at the beginning of this chapter in Figure 6.1, and is the reason that classifying large numbers of distribution patterns becomes such a complicated and daunting task. This is more apparent with genera, since the smaller the range-size (number of regions) the more similar two distributions will appear to be – the presences and absences in different regions are more likely to be identical. To some extent this degree of overlapping will therefore be an artefact of only scoring distributions in 52 large regions. However, this is only a coarse representation of the underlying distribution, and the actual similarities of the spatial extent of genera are sufficient for this problem to still occur if distributions are recorded at finer scales – distributions will be less likely to be identical, but will still be likely to show a large degree of similarity.

6.4.3 The relationship between generic diversity and floristic complexity

One way of further simplifying the complexity of all of the overlapping floristic elements identified is to summarise all those elements shown within each region individually; in effect this gives some measure of the relative floristic complexity for each region. That is, many different floristic elements overlap within an individual region, and the greater the number of floristic elements which overlap, the more 'complex' the flora of that region can be described as. Furthermore, having summarised the floristic elements for each region, we might try to return to the question posed at the beginning of this chapter and ask 'Is there a relationship between floristic richness and floristic complexity?' The answer seems to be, 'yes'. Figure 6.11 below shows a strong correlation between the number of genera in a region, and the number of floristic elements (as produced by *k*-means partitioning) shown by those genera – what I refer to as the 'complexity' of the floristic composition. Not surprisingly, there is considerable scatter in this graph. However, indicating the identity of these regions and superimposing a crude geographical classification on top, as has been done for Figure 6.11, reveals a striking uniformity within the evident

clusters of the graph. Regions within each cluster are not necessarily geographically adjacent, or even geographically close, but do share important geographical characteristics. This mirrors the results from the beta diversity analysis of Chapter 3, where the floristic similarity between a pair of regions showed a stronger relationship to the latitudinal difference in the positions of each of those regions than it did simply to the distance between those regions. Some of these groups in Figure 6.11, furthermore, are essentially the same as those found in the analysis of floristic relationships presented in Chapter 5.

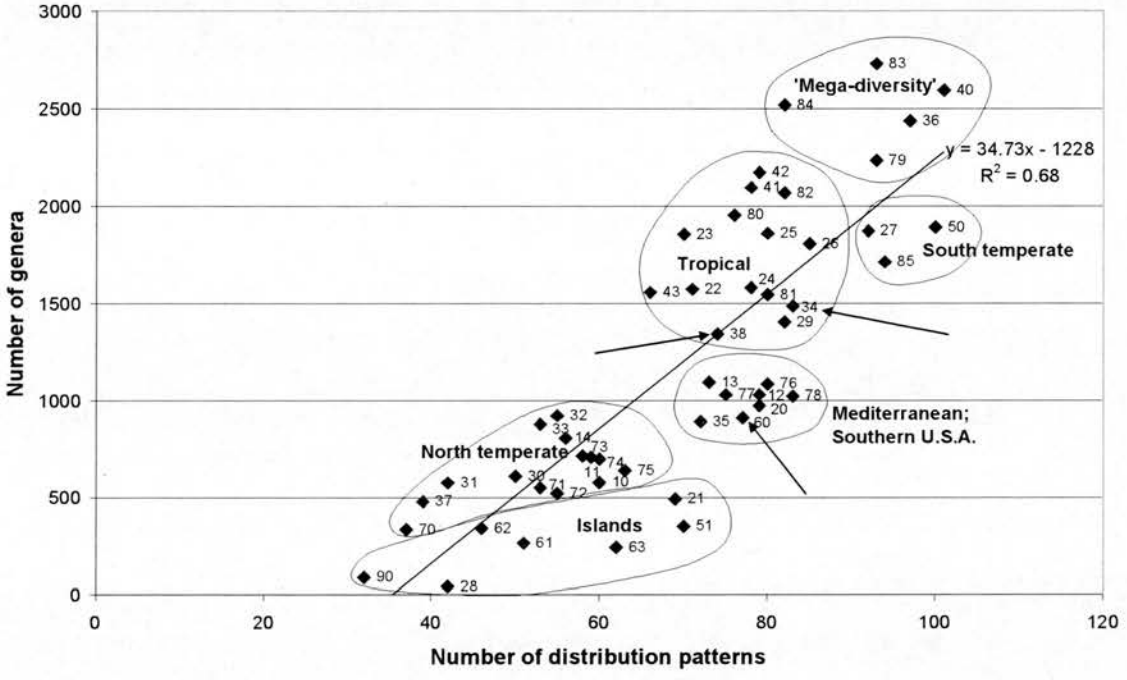


Figure 6.11 Relationship between number of distribution patterns and number of genera for TDWG regions. Although there is considerable scatter, clusters of geographically similar regions are evident; arrows indicate apparently 'misplaced' regions.

Moving upwards through Figure 6.11, at the bottom is a diffuse cluster of regions consisting only of islands: Region 21, Macaronesia; Region 28, Middle Atlantic Ocean [St. Helena and Ascension Island]; Region 51, New Zealand; Region 61, South-Central Pacific; Region 62, Northwestern Pacific; Region 63, North-Central Pacific [Hawaiian Islands]; Region 90, Subantarctic Islands. Island floras are not large (since most islands are small) and generally consist of a mixture of endemic elements and cosmopolitan elements; hence both the diversity and the complexity of these floras are low. This group of island regions, either including Region 28, Middle Atlantic Ocean or with this region as distinct from all other regions, consistently appears separate from other regions in the dendrograms given in the analysis of floristic relationships in Chapter 5. Immediately above this is a large group of northern temperate regions; to the

left of the group the most floristically depauperate of these (Region 31, Russian Far East; Region 37, Mongolia; Region 70, Subarctic America) show considerably fewer distribution patterns than do the rest of this group (Region 10, Northern Europe; Region 11, Middle Europe; Region 14, Eastern Europe; Region 30, Siberia; Region 32, Middle Asia; Region 33, Caucasus; Region 71, Western Canada; Region 72, Eastern Canada; Region 73, Northwestern U.S.A.; Region 74, North-Central U.S.A.; Region 75, Northeastern U.S.A.).

Next come a group of more southerly but still predominantly temperate regions which I have labelled 'Mediterranean; Southern U.S.A.'. The North American component of this group is truly the southern U.S.A. (Region 76, Southwestern U.S.A.; Region 77, South-Central U.S.A.; Region 78, Southeastern U.S.A.) while the 'Mediterranean' component (Region 12, Southwestern U.S.A.; Region 13, Southeastern U.S.A; Region 20, Northern Africa) is missing one Mediterranean region (Region 34, Western Asia) but has gained another (Region 35, Arabian Peninsula). Though this group consists of two geographically separate entities, these two entities are very similar in their latitude (see Table 1.1). Collectively, they show many more floristic elements (have a more 'complex' flora) than the regression line predicts for regions of such generic richness. These two geographical groups (the northern temperate and the Mediterranean/southern U.S.A.) together form the large division of temperate continental regions which groups together in all of the cluster analyses of floristic relationships in Chapter 5, although in these dendrograms there is no such apparent separation by latitude of the 'Mediterranean/southern U.S.A.' group – the Mediterranean regions group with adjacent European and Middle Eastern regions, while the regions of the southern U.S.A. group with adjacent regions from North America (see Chapter 5).

A large and diffuse group of tropical regions follows, representing all three of the tropical areas of Africa (Region 22, West Tropical Africa; Region 23, West-Central Tropical Africa; Region 24, Northeast Tropical Africa; Region 25, East Tropical Africa; Region 26, South Tropical Africa); SE. Asia (Region 41, Indo-China; Region 42, Malesia; Region 43, Papuasia); and the Neotropics (Region 80, Central America; Region 81, Caribbean; Region 82, Northern South America), with many more genera and with correspondingly more distribution patterns. To the right of these, with similar numbers of genera but with many more distribution patterns, is a group of three geographically disparate but biogeographically inter-related regions: Region 27, Southern Africa; Region 50, Australia and Region 85, Southern South America. The existence of pan-Southern Hemisphere taxa has long been known (Hooker, 1853), but in this thesis only now with the analysis of distribution patterns have these shared floristic elements been revealed. In the analysis of floristic relationships (Chapter 5), the greatest similarity of each of these regions was with adjacent tropical areas. Each of these regions is so large, however, that at their northernmost limit they each extend into the tropics, so it is perhaps a moot point whether or not they should really be labelled as 'temperate'.

Three regions, indicated by arrows in Figure 6.11, do not 'fit' into this simple geographical classification: they have both more genera and more distribution patterns than they 'ought to' from their geographic location. Region 60, Southwestern Pacific, falls within the 'Mediterranean; Southern U.S.A.' cluster, not with the group of 'Island' regions at the base of the graph; however, it did not group with these 'Island' regions in the analysis of floristic relationships (Chapter 5) either. The possible reasons for the anomalous position of the Southwestern Pacific region in Figure 6.11 include: catching the tail end of the distributions of many SE. Asian genera which reach no further into the Pacific; New Caledonia within this region is a continental fragment derived from Gondwanaland, and shows not only an exceptional degree of generic endemism but also many floristic relationships with other regions (New Guinea, Australia, New Zealand) which are not shown by any other islands in the vicinity. Then within the tropical group are two predominantly temperate Asian areas: Region 34, Western Asia and Region 38, Eastern Asia. However, although this is not the case for Region 34, Western Asia, it was clear from the analysis of floristic relationships in Chapter 5 (both clustering and ordination analyses) that for Region 38, Eastern Asia, the floristic relationships with the tropical Asian flora were much stronger than they were with the temperate Asian and Eurasian flora.

Finally, at the top of the graph is a group of what has been labelled 'mega-diversity' regions (Region 36, China; Region 40, Indian Subcontinent; Region 79, Mexico; Region 83, Western South America; Region 84, Brazil), which have much greater numbers of genera than do the other regions. These same regions were revealed as especially diverse both with respect to area (Chapter 3) and to latitude (Chapter 4), although in the analysis of floristic relationships (Chapter 5), they each grouped with geographically adjacent regions. One of them (Region 84, Brazil) may be regarded as exclusively tropical; three of the others (Region 36, China; Region 40, Indian Subcontinent; Region 79, Mexico) cross the important temperate-tropical boundary, so accumulating both exclusively temperate and exclusively tropical taxa; the final region (Region 83, Western South America) is the richest region at genus-level both in absolute and in relative terms (see Chapter 3), but, as might Region 40, Indian Subcontinent, it might be thought of as crossing the temperate/tropical divide in elevation rather than strictly by latitude, as these two regions are also by far the most mountainous, containing the bulk of, respectively, the great mountain chains of the Andes and the Himalaya. For these regions, therefore, more than for any other region, it is the combination of temperate and tropical elements which make such a contribution to their great generic diversity.

The relationship between number of genera and floristic 'complexity' is partly dependent on region size. The 'megadiversity' regions, with many genera and many distribution patterns, are all large. However, several large regions from temperate areas (30, Siberia; 31, Russian Far East; 70, Subarctic

America), have noticeably few distribution patterns compared with other regions (see Figure 6.11), while smaller regions (e.g. 12, Southwestern Europe; 76, Southwestern U.S.A.; 78, Southeastern U.S.A) show many more distribution patterns than do larger temperate regions. Island regions, on the other hand, which may be large in geographical extent (e.g. regions 61, South-Central Pacific; 62, Northwestern Pacific; 90, Subantarctic Islands) although they are all small in land area, all show characteristically few distribution patterns (and have fewer genera). Some small continental regions such as 11, Middle Europe, or 77, South-Central U.S.A., border a greater number of other regions than do some larger regions such as 27, Southern Africa or 84, Brazil. These regions might therefore be expected to show a greater degree of floristic complexity simply because they border a greater number of other regions, genera found in adjacent regions being more likely to be found in that region. However, it should be apparent from Figure 6.11 that this factor does not seem to be affecting the floristic complexity of the regions, Regions 11 and 77 having low floristic complexity while Regions 27 and 84 show high floristic complexity.

The overall relationship between numbers of genera and numbers of distribution patterns for the 52 TDWG regions of the world therefore seems to be predominantly dependent on three factors: size of region, latitude of region, and distance between regions. Regions isolated from other regions (i.e. island regions), have both fewer genera and fewer distribution patterns; for contiguous (continental) regions, those closer to the equator have greater numbers of genera than those with a more temperate distribution, although they do not always show greater numbers of distribution patterns; within broad latitudinal bands, however, larger regions contain greater numbers of genera, and also usually show greater numbers of distribution patterns, than do smaller regions.

6.5 Summary

- Global floristic classifications have been presented for both families and genera of angiosperms.
- Non-metric multidimensional scaling has proved a robust and powerful technique for grouping taxa by their geographical similarity.
- *k*-means partitioning has proved to be a robust technique for delimiting common floristic elements within this ordination space.
- The result of this classification was to reduce the noise in the data, representing the diversity of genera within a region by fewer floristic elements, while preserving the overall biogeographic structure within the original distribution data.
- There is a relationship between the generic diversity of a region and the floristic complexity of a region, which can be measured as the number of floristic elements.

- Floristically, regions can be grouped into collections of regions showing common geographical characteristics rather than simple geographical proximity.
- The least diverse regions are islands which consist of a combination of endemic and widespread taxa, and thus show little floristic similarity with other regions.
- The most diverse regions of all are those which contain strong floristic links with both temperate and tropical floras.

CHAPTER 7

DISCUSSION AND CONCLUSIONS

7.1 Justification of this thesis

This thesis has presented several separate, yet inter-connected, analyses of patterns of diversity for different taxonomic scales, and also analyses of patterns of distribution, focusing at the level of genus, on a global scale for all angiosperms. Angiosperms represent a large and highly diverse crown group of established monophyly (Soltis *et al.*, 1999; Savolainen *et al.*, 2000) and increasing species number (Niklas, 1988; Niklas & Tiffney, 1994); both pteridophytes and gymnosperms, on the other hand, represent formerly-more-diverse groups now restricted in their size (Kenrick & Crane, 1997). Furthermore, notwithstanding the interest in the biogeography of pteridophytes and gymnosperms, many pteridophyte genera show highly scattered and irregular distributions when compared with angiosperm genera, presumably due to their spores being easily wind-dispersed (Tryon & Lugardon, 1990), which makes the interpretation of the kind of analyses presented here, and comparison of these with those of angiosperm genera, considerably more difficult.

By 'patterns of diversity', I have meant taxonomic richness, or counts of all taxa (of the same rank) naturally occurring within an area, compared between separate areas. 'Patterns of distribution' therefore refers to the comparison of all native occurrences for taxa or collections of taxa, throughout different areas. The two aspects are intimately inter-connected – they are like two sides of the same coin; indeed, the former patterns are themselves a product of the latter. For many genera, there are other genera showing exactly the same distribution pattern (at this scale); conversely, however, for each individual genus that distribution pattern is different from the majority of other genera. Therefore, since only a minority of genera within any particular region are endemic to it (see Chapter 3), the majority of genera in each region used in this study display a variety of distribution patterns extending outside of it.

The set of taxa occurring within a particular geographical region can thus be thought of as the intersection of many different patterns of distribution. It is here argued that the diversity within that region depends not only on features unique to that region itself, but also on the degree of overlap for the set of non-endemic genera of the wider distributions outwith that region: richer regions have more complex patterns of floristic relationship. This diversity in the distribution patterns of non-endemic genera I here term the 'floristic complexity' of a region. Given the similarity between the results of the analyses of diversity patterns set out in this thesis with previous findings at different spatial and taxonomic scales (Gentry, 1988; Barthlott *et al.*, 1996; Brummitt & Nic Lughadha, 2003), it is predicted that these previous findings would also be explained well by the floristic complexity of their non-endemic

taxa, should there be further plant data-sets of comparable size and scope to that used in this thesis with which to extend these other studies.

7.2 The use of higher taxa

That different genera can be treated as somehow 'equivalent' in evolutionary status to each other is the key assumption underpinning this thesis. The analyses presented in the previous chapters have all treated the individual genera as unique and independent, and therefore comparable to each other. However, in practice it is almost impossible to really say whether or not all the genera used in these analyses may actually be 'equivalent' to each other. This therefore remains an inherent uncertainty in the work presented here, and it remains to be seen whether or not these results will change in the light of ongoing taxonomic research. Genera consist of species which are each assumed to have their own unique niche and distribution, and the distribution of a genus is no more than the combined distributions of all of its component species. In its definition a genus is therefore less of a biological entity and more of a purely taxonomic one, since it is the component species of a genus that are maintained as a homogeneous entity through interbreeding and gene flow, and their individuals that are subject to natural selection, rather than the genus itself. A single genus may therefore be subject to a range of different selection pressures which precludes a single explanation for its distribution However, the literature on what constitutes a species and whether or not species are really equivalent to each other (e.g. Ereshefsky, 1992) is voluminous and still lacking in real consensus. Nevertheless, this does not prevent biologists from holding a strong feeling that the concept of a species does have a biological reality, or evolutionary ecologists from constructing models and theories to explain patterns of species' distribution, even if the precise definition of a species and its role in evolution is not yet widely agreed upon.

The argument used to justify the value of genus-level analyses in this thesis is twofold. Firstly, it is that the 'hollow curve' distribution, which is repeated at different taxonomic levels, is a representation of a fractal-like underlying phylogenetic structure (Minelli *et al.* 1990), a slice at a particular level through the phylogenetic tree. Each genus is therefore assumed to represent a portion of a larger phylogenetic tree, with an independent evolutionary and biogeographic history. It still remains a somewhat arbitrary decision as to which node represents a taxon worthy of formal recognition (although of course that node may well mark the position of a readily-observable morphological synapomorphy which would serve as a character for generic recognition). At any particular level, therefore, although the status of individual taxa may change, the overall frequency distribution of sizes (and range-sizes, and distribution patterns) of taxa should not. This was borne out in a comparison of the sizes (numbers of genera) in traditional *versus* exclusively-monophyletic families (see Chapter 3). That is, some genera of uncertain status may be sunk, but this will be compensated by other genera becoming newly-recognised. For the analyses presented here, therefore, this uncertainty in the status of genera may not affect the results (see Chapter 3). The second justification of using genera as the primary units of analysis in this thesis is their utility as taxonomic surrogates for species-level patterns of diversity and distribution, as has

been demonstrated in a range of recent studies (Sepkoski, 1992; Williams & Gaston, 1994; Balmford *et al.*, 1996; Gaston, 1996; Williams & Humphries, 1996; La Ferla *et al.*, 2002), as well as in this thesis. If patterns of diversity are well correlated at different taxonomic levels (see Chapter 3), it is predicted that patterns of genus distribution also mirror patterns of distribution at species level.

7.3 General Discussion

7.3.1 Diversity within a region is well correlated at all taxonomic scales

Perhaps the most obvious question is first to ask how many taxa are there in each region (Gaston, 1996). Comparing the three taxonomic scales of family, genus and species, patterns of relative diversity within a region are highly correlated, irrespective of the sizes or geographical positions of the regions (Spearman's r_s: between species richness and genus richness, 0.97; between genus richness and family richness, 0.96; between species richness and family richness, 0.93; $n = 52$, $p < 0.01$ in each case; see Chapter 3), confirming that higher taxonomic levels may indeed be a reasonable surrogate for species richness (Balmford *et al.*, 1996; Gaston, 1996; La Ferla *et al.*, 2002; but see also Prance, 1994). The spread of data points in the log-log taxon-area graph increases from family- to genus- to species-level, with the exponent of the taxon-area relationship likewise increasing from 0.12 to 0.26 to 0.35, respectively. The increasing exponent values simply reflect the structure of the taxonomic hierarchy: there can be several species within one genus or genera within one family, but not vice versa; therefore the numbers of taxa will inevitably increase with decreasing taxonomic rank. However, in addition to this, the increasing spread of data points with decreasing taxonomic rank reflects the increase in the proportion of tropical taxa at lower ranks: the ratio of tropical species : genera is greater than the ratio of temperate species : genera. In other words, the strength of the latitudinal gradient of diversity increases with decreasing taxonomic rank (Willig *et al.*, 2003).

7.3.2 But differences in size between regions mask their true relative diversities

Comparing diversities of different regions is confounded by differences in the relative sizes of regions (Rosenzweig, 1995). Given the positive relationship between diversity and area (Arrhenius, 1921; Williams, 1943), larger areas would be expected to be richer simply because they are larger. True relative diversities between areas of different size can, however, be established through re-arranging the power-law species-area relationship and calculating area-rescaled diversity figures (Rosenzweig, 1995; Brummitt & Nic Lughadha, 2003). For relative generic richness, three areas of tropical diversity can be clearly made out: the Neotropics, SE. Asia and Africa (+ Madagascar), decreasing in that order (see Chapter 3). Overall, it is clear that Western South America (83) is the most diverse region of all for both absolute and relative numbers of genera. Comparison between absolute and re-scaled rankings for the Neotropical regions shows that in absolute terms the large regions Mexico (79) and Brazil (84) appear

very diverse in numbers of genera; in fact, however, this is an artefact of their large size. The two smallest Neotropical regions, Central America (80) and Caribbean (81), which do not appear to be particularly diverse in terms of absolute numbers of genera, are revealed as surprisingly diverse when numbers of genera are scaled by area; indeed, for relative generic richness, Central America (80) is second only to Western South America (83). Collectively, four of the five most diverse regions are in the Neotropics, (Western South America, 83; Central America, 80; Mexico, 79; Northern South America, 82, in descending order of relative diversity), the exception being the Indian Subcontinent (40). However, for both China (36) and the Indian Subcontinent (40) very high absolute genus richness scores are inflated by their huge size. The western Pacific is also markedly diverse, though this is only apparent from the rescaled data: the small area of this region masks its true diversity. However, West Tropical Africa (22), West-Central Tropical Africa (23) and Northeast Tropical Africa (24) are all of considerably lower relative diversity than are other tropical regions; in fact comparable to southern Europe (Southwestern Europe, 12; Southeastern Europe, 13), to SW Asia (Caucasus, 33; Western Asia, 34) and to southern U.S.A. (Southwestern U.S.A., 76; South-Central U.S.A., 77; Southeastern U.S.A., 78). Again, this is masked in the absolute genus richness scores for these regions. This surprising result may be partly explained by the almost-barren Sahara Desert covering large expanses of West Tropical Africa (22) and Northeast Tropical Africa (24), giving lower generic diversities than would be expected for regions of that size (Archibold, 1995).

7.3.3 Patterns of relative taxonomic richness echo those found in previous studies

If genus-level patterns really are acceptable surrogates for species-level diversity (Balmford *et al.*, 1996; Gaston, 1996; La Ferla *et al.*, 2002), perhaps inferences drawn from these analyses may also be applicable at species-level. For example, the Neotropics in general, and in particular western South America and Central America, consistently emerge as the most diverse areas of the world for plants, at all spatial scales. At small scales, this was previously shown by Gentry (1988), in a global analysis of species-richness data from standard-sized 0.1-ha wet forest plots, where only one out of the five richest sites was not found in the Neotropics (Semengoh Forest in Sarawak [Borneo]); of the Neotropical sites, two were in Colombia and two in Peru. It has also been shown by a similar re-analysis of the raw data from Davis *et al.* (1994-1997) and from Mittermeier *et al.* (1999), again in each case re-scaling absolute species numbers by the species-area relationship where $z = 0.14$ (Brummitt & Nic Lughadha, 2003). For the data from Davis *et al.* (1994-1997), the richest areas were found to be La Amistad (Costa Rica / Panama), the region of the upper Rio Negro (Brazil / Colombia / Venezuela) and Braulio Carillo – La Selva (Costa Rica). Similarly, re-analysing the 'hotspots' data of Mittermeier *et al.* (1999) reveals the Tropical Andes and Mesoamerica hotspots to be by far the most diverse for both total vascular plant species and endemic vascular plant species (Brummitt & Nic Lughadha, 2003). Furthermore, in the pioneering study by Barthlott *et al.* (1996), three of the six areas where vascular plant diversity was estimated to exceed 5,000 species/10,000 km^2 were found in the Neotropics: the Chocó-Costa Rica centre;

the Tropical Eastern Andes centre; and the Atlantic Brazil centre. In this thesis it is again western South America which is the richest region of the world, this time for genera of flowering plants.

7.3.4 But areas richest in genera are not necessarily also richest in endemic genera

Patterns of generic diversity and patterns of generic endemism are less well correlated than is diversity at different taxonomic ranks (Spearman's r_s 0.81, n = 52, $p < 0.01$; see Chapter 3). Though all the regions with highest genus richness have moderate degrees of generic endemism, those regions with the highest degree of endemism (Southern Africa, 27 [31%]; Western Indian Ocean, 29 [29%] and Australia, 50 [36%]) have themselves moderate genus richness. It is notable that amongst tropical regions, no tropical African region has generic endemism greater than 10%; indeed West Tropical Africa (22); East Tropical Africa (25); and South Tropical Africa (26), all have levels of generic endemism (2.2%, 2.8% and 1.9%, respectively) lower than that for Southwestern Europe (12; 3.0%). In SE Asia and the Neotropics, on the other hand, Malesia (42; 10.8%) and all of Mexico (79; 11.8%); Caribbean (81; 12.6%); Western South America (83; 10.8%); Brazil (84; 15.7%); and Southern South America (85; 13.6%) have values for generic endemism greater than 10%. Assuming still that patterns in the distribution of genera truly reflect underlying patterns in species distribution, this implies that levels of speciation have been far greater in extra-African tropical regions, and in the Neotropics in particular (Gentry, 1982; Richardson *et al.*, 2001a; Dick *et al.*, 2003; see also Bramley *et al.*, 2004), unless wholesale extinction of genera has been greater in Africa than in both SE Asia and the Neotropics (Whitmore, 1990).

Amongst temperate regions, some (Central Asia [32; 6.0%]; Western Asia [34; 6.6%]; Southwestern U.S.A. [76; 8.4%]) have levels of generic endemism greater than many tropical regions (all of tropical Africa; Indian Subcontinent [40; 6.1%]; Papuasia [43; 5.7%]; Central America [80; 3.7%]; and Northern South America [82; 4.7%]). Other, cold-temperate regions, however, have absolutely no endemic genera (Northern Europe, 10; Middle Europe, 11; Western Canada, 71; Eastern Canada, 72; North-Central U.S.A., 74; and, not surprisingly, the Antarctic Continent, 91) – and Eastern Europe (14) and Subarctic America (70), have only a single one apiece. Several isolated island regions (Middle Atlantic Ocean, 28; New Zealand, 51; North-Central Pacific, 63) have moderate degrees of endemism but relatively low genus richness. In fact, Middle Atlantic Ocean (by far the smallest TDWG Region at only 232 km^2, and at least one order of magnitude smaller than the next-smallest region), has a remarkable 11 endemic genera out of only 44 native angiosperm genera.

There is almost no relationship between generic endemism and the sizes of regions ($r^2 = 0.06$) – although small regions never have large numbers of endemic genera, large regions do not necessarily always have large numbers of endemic genera (see Chapter 3). This is the corollary of patterns of endemism not being that well correlated with patterns of diversity. The regions which have many endemic genera are all at least partly tropical, while large regions with few endemic genera are all at high

latitudes; the relationship between numbers of endemic genera and sizes of regions is obviously influenced greatly by the latitudinal position of those regions. However, conclusions from patterns of generic endemism remain tentative, since these genera are likely to be subject to future taxonomic change. Many endemic genera, particularly from isolated oceanic islands (e.g. St. Helena; Richardson *et al.*, 2001b) or ecological islands (e.g. mountain tops; Comes & Kadereit, 2003) in temperate regions, contain only one or a few species marked by strong morphological adaptation to isolated, localised, often extreme environments. The phylogenetic rationale for recognising many of these genera, the small temperate endemic genera in particular, has not yet been established; many may be shown to nest within other genera and may lose their taxonomic status in the future.

7.3.5 The larger land area of the tropics helps to explain the latitudinal diversity gradient

Tropical regions emerge as clearly more diverse than are temperate regions, and there is a strong latitudinal gradient of diversity of angiosperm genera which is roughly symmetrical about the equator (see Chapter 4). However, an adequate theoretical understanding of the cause(s) of the latitudinal gradient of diversity remains lacking (Rosenzweig, 1995; Hawkins *et al.*, 2003; Willig *et al.*, 2003; Hawkins & Diniz-Filho, 2004). Greater numbers of species are found in the tropics for almost all groups of organisms and in almost every type of habitat, both terrestrial and aquatic, and so, if we are searching for a comprehensive understanding of the latitudinal gradient, the explanation must then apply equally well to nearly all organisms and environments. The tropics contain far more land area than do other climatic zones (Terborgh, 1973; Rosenzweig, 1992), and this partly explains the greater biological diversity of tropical regions (Blackburn & Gaston, 1997; see Chapter 4). Mean latitudinal range-sizes appear to decrease towards the equator, but this is an artefact: by excluding overlapping genera in adjacent latitudes from the analysis and only counting each genus once (Rohde, 1992), range-size is shown to be greatest for those genera centred on the tropics (see Chapter 4). Rapoport's Rule (Stevens, 1989) is therefore not the explanation of the latitudinal gradient of diversity but is better explained as a local phenomenon occurring over short latitudinal ranges in the northern hemisphere in which the amount of land area also decreases (Rapoport, 1982; Gaston *et al.*, 1998). Mean latitudinal range-sizes in fact show a general increase in tropical regions, but the histogram of range size with latitude has many secondary peaks. The principal peak in latitudinal range-sizes does, however, coincide with the weighted centroid for the world at 14°N. The range-size frequency distribution of latitudinal ranges around the equator is explained well by a 1-D geometrically-constrained null model known as the mid-domain effect, although as an explanation of the latitudinal gradient of diversity this idea remains controversial (Hawkins & Diniz-Filho, 2002; Zapata *et al.*, 2003; Colwell *et al.*, 2004; Pimm & Brown, 2004). Although other competing explanations have not been explored in this thesis, overall, the geometry of the land area of the Earth seems to play a large part in explaining the presence of the latitudinal gradient of diversity.

7.3.6 Strong floristic clusters are evidence of localised genera

In the analysis of floristic similarity between regions (see Chapter 5), five large, well-defined groups are apparent from the ordination by non-metric multidimensional scaling, which are: an 'Africa and Madagascar' group, a distinct 'Neotropical' group, a diverse group of eastern and southern Asian regions, then a 'North American' group, and lastly there is a group of 'Temperate Northern Eurasian' regions. These groups agree well with the kingdoms and subkingdoms of traditional floristic hierarchies (Good, 1974; Takhtajan, 1986) and also the relationships confirmed by later studies (e.g. Conran, 1995; Linder, 1996). Several regions, however, do not fall neatly into any of these groups: Middle Atlantic Ocean (28), Arabian Peninsula (35) and Subantarctic Islands (90). The five groups recovered by non-metric multidimensional scaling also largely correspond to the continental groups found by UPGMA cluster analysis at the level of 50% similarity or less. The 'Africa and Madagascar' group and the 'Neotropical' group are both identical in the two analyses. In the ordination analysis, Subarctic America (70) shows greater similarity to other North American regions than it does in the cluster analysis, but otherwise this North American group is identical in composition also. Macaronesia (21), which grouped with northern temperate regions with cluster analysis, and Arabian Peninsula (35) group more strongly with the 'Temperate Northern Eurasia' group in the ordination. The biggest single difference between the two analyses is with the position of the Pacific regions (61, 62, 63), which lie peripherally to the large group of eastern, tropical and Australasian regions by ordination analysis but as a distinct group with other 'island' regions by cluster analysis. These strong continental clusters of regions, which are in broad agreement in both the clustering and ordination analyses, are a product of the highly-skewed frequency distribution of distribution patterns: the majority of generic distributions are within individual continents. However, some differences between the ordination and the clustering analyses are to be expected, since the 3-dimensional, non-hierarchical framework of the ordination reveals subsidiary floristic links between regions that are lost when the results are constrained onto a 2-dimensional dendrogram, which can only show the relationships of maximum similarity (McCune & Grace, 2002). The ordination results therefore more truly reflect the complexity of the underlying distribution patterns, but are correspondingly less easy to interpret.

7.3.7 Only a small number of possible distributions are found

If localised genera produce strong floristic relationships between adjacent regions, by investigating further it is possible to identify just which distribution patterns link which regions. Given the diversity of taxa within any one region and the possible number of potential distributions for those taxa, the floristic relationships are expressed in a remarkably small number of distribution patterns that account for a remarkably high number of genera. With 52 regions, and any individual genus observed either to occur or not occur in each of those regions, there will be a total of 2^{52} possible unique combinations of those regions making up the complete set of distribution patterns (see Chapter 6). The maximum distribution range is obviously a genus present in every region (although this is not actually

shown by any genus); the minimum distribution is not actually those genera endemic to a single region, but a null distribution absent from every region, since in theory this is also a potential distribution. However, since recently-extinct genera are here treated as native in their former range, and intergeneric hybrids have been excluded from this analysis as they have often arisen in horticulture and so have no natural distribution, in practice no genus actually shows this null distribution. Assuming therefore that we are not interested in the single empty (null) distribution pattern containing no genera, then this still leaves $2^{52}-1$ potential generic distribution patterns – or more than 4.5×10^{15} possible distributions! There cannot be more actual generic distributions than there are actual numbers of angiosperm genera (14,304), however, so the vast majority of mathematically possible distribution patterns are never shown by these data. For angiosperm genera, only 2817 separate combinations of regions are actually found. Since some 38% of genera are endemic to a single region (see Chapter 3), 62% (8868 genera) must be found in the 2771 distribution patterns occurring in more than one region, and the majority of these are those small range distributions which give rise to the strong floristic relationships between adjacent TDWG regions (see Chapter 5).

7.3.8 Small-range distributions are the most common

Of the 2817 actual distributions, c. 2200 are unicate (shown by only one genus); only c. 600 distributions are shown by more than one genus, but collectively these 600 distributions account for about 11000 genera. Respectively, therefore, only just over 20% of distribution patterns account for just over 80% of genera, while just under 80% of distribution patterns account for only about 20% of angiosperm genera. That is to say, the frequency distribution of genus distribution patterns is extremely right-skewed, with a modal (most common) value of only one genus per distribution pattern. This 'hollow-curve' frequency distribution (Willis, 1922; Williams, 1964) of distribution patterns is also shown by the frequency distribution of range sizes giving rise to the mid-domain effect, a pattern found in the majority of groups (Colwell & Lees, 2000; Gaston, 2003). As well as the majority of distribution patterns being unicate, only shown by one genus, the majority of distribution patterns also occur in more than one region (i.e. are non-endemic). Although the frequency distribution of genus range-sizes is also extremely right-skewed, the modal value of genus range-sizes, 1, only accounts for 38% of genera (i.e. only 38% of genera are endemic to a single region). The highest rate of endemism (36% of taxa for any single region [Region 50, Australia]) does not exceed 5% of the global total number of genera; furthermore, there is no genus found in every region, and only three genera (*Carex* L. and *Cyperus* L., both Cyperaceae; and *Plantago* L., Plantaginaceae) found in every region except Antarctica. Since the total set of endemic distributions only numbers 46 separate regions (endemic genera are not found in all regions), the remainder of the 554 repeating distribution patterns must therefore be shared between different regions, as is revealed by the analysis of floristic relationships (see Chapter 5). Therefore, the majority of angiosperm generic diversity is shown by a few repeating distribution patterns over several regions (see Chapter 6). What are these repeating distribution patterns which account for such a high proportion of the total diversity of angiosperm genera?

7.3.9 Genera can be further grouped into clusters of repeating distribution

The strength of floristic relationships between regions of the same continent (see Chapter 5) suggests that these common, repeating distribution patterns are themselves confined to areas within single continents, an idea reinforced by the shape of the range-size frequency distribution that shows a great number of small-range distributions (see Chapter 3). Returning therefore to the bigger question suggested by the frequency distribution of genus distribution patterns, we can ask: what actually are these different distribution patterns, and how frequently are they found. However, defining just what is meant by 'different distribution patterns' proves to be no easy task. For a heterogeneous collection of distribution patterns, where every single pattern is overlapped at least partially by another, where should the discontinuities between them be drawn? For example, can two widespread distributions which only differ in one of the genera also being present in a single additional region be considered to be essentially 'the same' distribution and grouped together? If so, then what about a distribution pattern which differs from this group by only one region, and then where can the line be drawn between 'different' distributions when, for the totality of the data, there is so much overlap? If not, then are we left with having to deal with all 2817 unique distributions? If this is the case, then we have made little progress with trying to analyse distribution patterns themselves, the raw data actually underlying all of the previous analyses, and questions of plant distribution will therefore remain essentially intractable. If, however, classifying distribution patterns is indeed possible, perhaps the great mass of 2,200 unicate distributions can be reduced to a manageable quantity without losing biogeographic detail. Furthermore, can the different distribution patterns be classified into an optimal number for all angiosperm genera? An 'optimal' number of distribution patterns, as in any type of classification, implies a balance between having small enough groups so that each group is internally homogeneous, but not so many groups that the total number is no longer conceptually manageable. There would be considerably fewer than 2817, in other words, but not so few that any one group contained a great diversity in the distribution patterns. Lastly, would this classification of distribution patterns reveal any geographical patterns in the constituent floras of the different regions?

7.3.10 There is a strong relationship between regional richness and floristic complexity

A *k*-means partitioning analysis (Bailey & Gatrell, 1995; Legendre & Legendre, 1998) found an optimal number of fewer than 200 geographically-distinct clusters out of 2817 unique distribution patterns (see Chapter 6). However, trying to summarise the results of such a large and complex analysis in simple terms proves difficult. The shape of the frequency distribution is still extremely skewed – the majority of diversity is still accounted for by only a few distribution patterns. Only 38 distribution patterns out of 181 are shared by 100 genera or more, but together these account for 8785 genera, or roughly two-thirds of the total number for all angiosperms. These distribution patterns are not all mutually exclusive; they might themselves overlap. Furthermore, we can study the total set of distribution

patterns given by all of the taxa of any one region and we can then pose the question 'Is there a relationship between the number of genera within a region, and the number of distribution patterns (as produced by *k*-means partitioning) shown by the non-endemic genera outwith that region?' – what we might refer to as the 'complexity' of the floristic composition of a region. The answer seems to be: 'yes', there is a strong correlation between floristic richness and floristic complexity (see Chapter 6). Not surprisingly, there is considerable scatter in the relationship. However, indicating the identity of each region and superimposing a crude geographical classification on top reveals a striking uniformity within the evident clusters of the graph. Regions within each cluster are not necessarily geographically adjacent, or even geographically close, but do share important geographical characteristics. Some of these groups, furthermore, are essentially the same as those found in the analysis of floristic relationships presented earlier (see Chapter 5).

With the lowest diversities and least complex floras is a diffuse cluster of regions consisting only of islands. Island floras are not large (since most islands are small) and generally consist of a mixture of endemic elements and cosmopolitan elements, hence both the diversity and the complexity of these floras are low. This group of island regions was found repeatedly in the analysis of floristic relationships. Immediately above this is a large group of northern temperate regions; next comes a group of more southerly but still predominantly temperate regions comprising the Mediterranean area and the southern U.S.A. Though this latter group consists of two geographically separate entities, both these entities are very similar in their latitude. Collectively, they show many more distribution patterns (have a geographically more 'complex' flora) than the regression line would predict for regions of such generic richness. A large and diffuse group of tropical regions follows, with many more genera but with correspondingly more distribution patterns. To the right of these, with similar numbers of genera but with many more distribution patterns, is a group of three geographically disparate but biogeographically inter-related regions: Southern Africa (27); Australia (50); Southern South America (85). The existence of pan-Southern Hemisphere taxa has long been known (e.g. Hooker, 1853), but in this study not until now, with the analysis of distribution patterns, have these shared floristic elements been revealed. In the analysis of floristic relationships, the influence of greater tropical diversity meant that the greatest similarity of each of these regions was with adjacent tropical areas, whereas now their more complex biogeographic relationships with distant areas are highlighted. Finally, there is a group of what I have labelled 'mega-diversity' regions (c.f. Mittermeier, 1988) which have much greater numbers of genera than do the other regions. These same regions were revealed as especially diverse with respect to both area and latitude. One of them (Brazil, 84) may be regarded as exclusively tropical; three of the others (China, 36; Indian Subcontinent, 40; Mexico, 79) cross the important temperate-tropical latitudinal boundary, so accumulating both exclusively-temperate and exclusively-tropical taxa; the final region (Western South America, 83) is the richest region at genus-level both in absolute and in relative terms, but, as might the Indian Subcontinent (40), it might be thought of as also crossing the temperate/tropical divide but in elevation rather than strictly by latitude, as these two regions are also by far the most mountainous, containing the bulk of, respectively, the great mountain chains of the Andes and the Himalaya. Both

therefore also contain the mix of tropical and temperate genera that characterise these 'mega-diversity' regions.

7.4 Integrating biogeographic patterns

An overall aim of this thesis has been to try to demonstrate that different analyses of the same data will give different results that will lead to different interpretations of general patterns of plant distribution. However, since these analyses have largely been based on a single, comprehensive data set, the different analyses merely represent alternative viewpoints of the same picture; the results are therefore inter-related. The inter-relationships between the main findings of this study can perhaps be best illustrated with reference to those parts of the world which are particularly diverse in both numbers of taxa and numbers of distribution patterns. The richest part of the world at both genus- and species-level is Western South America, and the eastern slopes of the Andes in particular; by comparison with other studies, this result seems to be consistent for plants, at all spatial scales. Gentry (1982) proposed that the exceptionally high floristic diversity of the northwestern Andean and southern Central American region was principally due to rapid, sympatric *in situ* speciation caused by the uplift of the Andes mountains, on top of an already-rich tropical flora, and he estimated that this explosive speciation might account for approximately half of the total number of Neotropical species. In this study, the regions of particularly high diversity are generally regions which show a high degree of range overlap, rather than just being those areas with very high proportions of endemic taxa. The three regions with the highest percentage endemism (27, Southern Africa; 29, Western Indian Ocean; 50, Australia) do not have conspicuously rich floras when compared with regions of the neotropics. Regions of the neotropics, however, have both conspicuously rich floras and also particularly complex floras. Central America, for example, the second-richest region for genera, is both the southern limit of distribution for many northern hemisphere taxa, and the northern limit of distribution for many tropical and southern hemisphere taxa, but the degree of generic endemism for Central America is low (3.7%). At a regional scale, Western South America shows great floristic complexity compared with other regions, containing as it does both tropical and also many temperate floristic elements, yet at the same time predominantly consisting of a characteristic neotropical flora with moderate levels of generic endemism at the regional scale and very strong floristic relationships with other adjacent neotropical regions.

Regions with many local taxa and a high degree of endemism create the impression of a very rich flora, since all of these taxa are 'new' to the botanist; however, endemic taxa are only a small component of the total flora of an area. The evolutionary and conservation interest of local endemic taxa and geographical areas of endemism notwithstanding, these alone do not make a region particularly rich in taxa; in this study at least, it is the combination of both local endemic taxa and also widespread taxa that generates high values of gamma diversity within a region. At the very broad geographical scales used in this thesis, more to the point, it is the overlapping of many different distribution patterns within the one

region which helps account for the great diversity shown by some regions. Local patterns of alpha diversity connect to regional patterns of gamma diversity through beta diversity, the change in taxonomic composition with distance; therefore the spatial turnover of taxa between regions goes a long way to determining their relative diversities. This spatial turnover is expressed in the range size frequency distribution for those taxa found within each region, and the distance decay of similarity between regions which in turn is a measure of the pairwise similarity shown by the floristic relationships between regions. This implies relatedness in a floristic rather than a historical sense, however, and one obvious omission from this study is any analysis of distribution patterns from a phylogenetic viewpoint. Although it must be true that all taxa possess a unique evolutionary history, and that shared patterns of distribution can have a common biogeographic origin, not all biogeographic studies seek to understand the historical relationships between areas or the biogeographic histories of particular taxa. While phylogenetic considerations may help understand which taxa are found in a particular area, each taxon is only one component of a larger flora, and the overall number of taxa is the outcome of many independent historical events. The strong relationships between diversity and broad geographical factors such as area and latitude imply that ecological factors rather than phylogenetic factors have a greater role in determining relative numbers of and ranges of taxa. Furthermore, in studies of the variation in range sizes and distribution patterns between phylogenetically-independent taxon pairs, little phylogenetic correlation between distributions was found (Gaston, 2003 and references therein; Davies *et al.*, 2004) – sister taxa tend to have different distributions from each other, and range sizes appear to vary randomly with respect to phylogeny.

The results from the various aspects of this study demonstrate the importance of the range size frequency distribution in understanding patterns of diversity and distribution. The overall range size frequency distribution for genera shows a smooth decline in frequency from the modal class of a single TDWG region to the least-frequent class of widespread taxa found in 51 regions. Large ranges underpin the very broad-scale patterns of the mid-domain effect and the latitudinal gradient of diversity, even though these large ranges are uncommon: the cause of the latitudinal diversity gradient was thought to be small latitudinal ranges in the tropics (Rapoport's Rule), although mean latitudinal range size actually increases in the tropics, and the removal of large-ranged taxa from the analyses of the geographic area hypothesis and the mid-domain effect reduced the effectiveness of either of these analyses. Small ranges, on the other hand, are more frequent than are large ranges and are what underpin the strong floristic relationships between adjacent regions; they remain robust to grouping with other similar patterns in the classification of floristic elements. Diversity within a region is the intersection of many different overlapping distribution patterns, the majority of which extend beyond that region, and each region displays a varied range size frequency distribution of both few large and many small ranges. However, range size frequency distributions for each region show important differences between regions. Since the modal class for different regions is most often 1, but endemic genera for each region are only ever a small component of the total generic flora (maximum generic endemism is only 35.7%), the majority of diversity within that region is contained in the genera with small ranges found in that region and the

regions immediately beyond it. The greater frequency of small ranges is revealed in the strong floristic relationships between adjacent regions; however, at a broader scale, the more widespread taxa are an important component of the total diversity within any one region. This is evident from the relationship between diversity and floristic complexity: the greater the number of the floristic elements within that region, the greater is the diversity of that region. The contrasting influences of taxa of different range sizes is only apparent from the analysis of a comprehensive dataset; it would not be shown by studying either just the taxa within one region, or the distributions of just one taxon.

The strong relationships found between broad-scale patterns of diversity in higher taxa and geographical factors such as latitude and area and the correlation of these with established patterns of diversity at finer scales suggest that these former patterns are a manifestation of ecological processes applying to species and individuals (Hubbell, 2001; Enquist *et al.*, 2002). It has not been possible to also analyse fine-scale patterns of species distribution in this study; however, this has been the focus of a very large number of previous case studies. The fine-scale processes determining species' distributions are also expressed in range size frequency distributions for those species, which generally show the same pattern at higher taxonomic ranks (Gaston, 2003). The range size frequency distributions of individual regions will therefore express the patterns of distribution of the constituent genera, as is shown for three exemplar regions (from the analysis in Chapter 6) in Figure 7.1 below. In the context of this study, particularly-diverse regions (e.g. 40, Indian Subcontinent; see Figure 7.1A) have shallower, more evenly-declining range size frequency distributions, containing both greater numbers of locally-distributed taxa (high alpha diversity and strong floristic links between adjacent areas) and also large numbers of widespread taxa (high gamma diversity and greater floristic complexity). In contrast, island regions which are both taxonomically and floristically depauperate (e.g. 28, Middle Atlantic Ocean; See Figure 7.1B) have much steeper range size frequency distributions as these floras tend to consist of a simpler mixture of endemic taxa (which do not reveal any floristic relationships) and cosmopolitan taxa (which only indicate the most general floristic relationships). Regions with both low endemism and low diversity (e.g. 31, Russian Far East; see Figure 7.1C) show a flatter range size frequency distribution with less variation in frequency between different range size classes. As yet, however, a widely-accepted mechanistic explanation for any highly-skewed range size frequency distribution, which would perhaps be based on high degrees of habitat and topographic heterogeneity, is not agreed upon.

Most previous studies seeking to explain patterns of high diversity have focused solely on factors intrinsic to high-diversity regions rather than looking also at the extra-regional diversity, the floristic diversity in the global distribution patterns of the non-endemic genera, that lies extrinsic to those regions. A generally narrow geographical focus is understandable given the lack of comprehensive, detailed information on plant distributions that still exists. However, this extra-regional diversity forms an important component of the intra-regional diversity, and this study has shown a strong relationship between numbers of genera within that region and numbers of distribution patterns outside of that region for the non-endemic genera. Is it then possible to say whether the diversity within the region is *caused by*

the confluence of these genera from outwith the region (plus the endemic genera of that region)? I.e. are the patterns of richness due to factors external to those regions of high diversity, rather than factors internal to it? If they are, then relative range sizes of taxa, or, looked at another way, whichever factors are limiting the range-sizes of taxa, would become a key parameter for explaining patterns of gamma diversity. A region might therefore be able to support a particular genus by providing the conditions in which that genus is also found outside of that region, and so on for each genus found with the region. Following this chain of argument, high habitat diversity within a region ≈ high taxon diversity for that region. This simple relationship would become further complicated by additional factors such as dispersal ability of those taxa (assuming that plants within a region had the ability [by whatever means] to disperse to suitable habitat within the region from suitable habitat outwith the region), and amount of available area of each habitat (since different habitats have different intrinsic carrying capacities; regions with more rain forest will have more species for that reason). In this context, Gaston (2003) has suggested that the underlying cause of the range size frequency distribution of taxa is the patchy distribution of suitable environmental conditions for each taxon: most sets of suitable environmental conditions are themselves narrowly distributed and only a few are widely distributed. However, this hypothesis would be very difficult to test with the coarse scale of the data used for this study. This study has therefore suggested many further avenues of research to be pursued in the future.

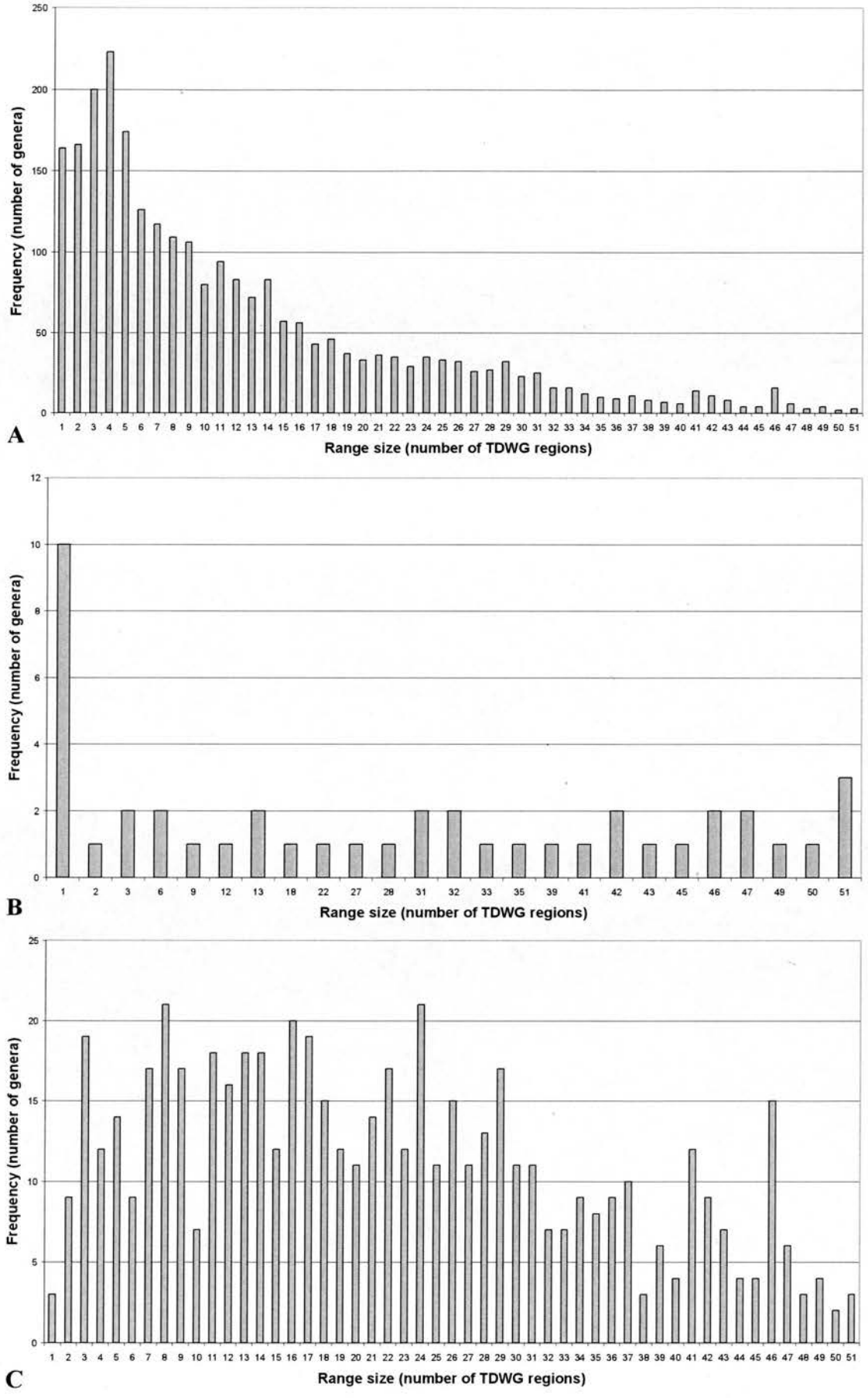


Figure 7.1 Range size frequency distributions for genera found in regions **A** 40, Indian Subcontinent (taxonomically rich and floristically complex); **B** 28, Middle Atlantic Ocean (taxonomically and floristically poor but with high endemism); and **C** 31, Russian Far East (taxonomically and floristically poor but with low endemism).

7.5 Limitations of this thesis

7.5.1 Practical and pragmatic limitations

Inevitably, perhaps, the work presented in this thesis has suffered from several constraints. From a purely pragmatic point of view, the most limiting of these has simply been the availability of suitable software packages capable of a) running appropriate analyses that ask relevant and biologically interesting questions – many new ideas are presented in research papers but without supplying software for the use of other researchers – and b) simply handling a data set of this size within a reasonable period of time, or even at all – many otherwise appropriate packages for ecological analysis are limited to handling only a few hundred taxa. Several of the analyses presented here each required data sets of hundreds of thousands of cells to be calculated and formatted, while a lack of programming and advanced computing skills unfortunately precluded the development of new applications.

7.5.2 Scale – both geographical and taxonomic

The most obvious limitation of this thesis has been the scale at which the original data has been compiled and then analysed. The regions used are simply too large to reveal the finer-scale variation in distribution patterns and the corollary of this, the detailed topographic, climatic and habitat differences within each region. Furthermore, as discussed in Chapter 1 and earlier in this chapter, the real biological interest of analyses such as those undertaken here is at the levels of species and also individuals – the levels at which biological interactions are actually occurring. The assumption underpinning all of this work is that genera are valid evolutionary units which act as surrogates for the more detailed patterns at species level, but for which there is still inadequate data; however, inevitably this must remain an assumption. It was nevertheless felt that the work presented in this thesis would be a valuable undertaking, given the comprehensive scope of the data, covering all families and genera across the world, and it simply was not feasible within the time constraints to compile a greater volume of data at a finer scale. Confirmation of some of the analyses presented here must therefore wait until such time as this data is available.

Scoring genera as present or absent in geo-political regions can also create problems; plants occurring in different habitats which have the same regional-level distributions can actually appear to have the 'same' distribution, while plants just extending over the border into another region will appear to have 'different' distributions. The *k*-means clustering analysis should avoid the latter problem, but cannot solve the former one. For example, within Africa, the Guineo-Congolian Regional Centre of Endemism of White (1983) and its surrounding Regional Transition Zone, has a distribution through most of TDWG Level-2 Regions 22 and 23 and also touches on Regions 24, 25, and into 26; the Sudano-Zambesian Regional Centre of Endemism also touches on all of these TDWG regions. Many genera which at a finer

scale may well be endemic to either the Guineo-Congolian or the Sudano-Zambesian zones, and thus each part of distinct floristic elements, would have to be recorded with the same TDWG Level-2 distribution. Recording distributions across habitats rather than geo-political regions initially would have been the ideal solution for cases such as this, except that this information is extremely hard to determine for widespread taxa (which tend by definition to be present in many different habitat types); it was often extremely difficult just to determine in which regions some taxa really were native. Given that it would have taken much longer to score distributions by habitat, this was unfortunately not feasible in the timescale of this project.

The coarse scale of the data has meant that the detailed investigation of causal mechanisms such as water (e.g. O'Brien *et al.*, 1998b; O'Brien *et al.*, 2000; Hawkins *et al.*, 2003), available energy (e.g. Currie, 1991; Barraclough & Savolainen, 2001; Bromham & Cardillo, 2003; Wright *et al.*, 2003) and climate more generally (Francis & Currie, 2003; Currie & Francis, 2004) in determining taxonomic richness within regions has had to be omitted, since generalising over such large areas with great internal heterogeneity will lose much of the detail upon which a reliable verification of these relationships should rest. That is, relationships evident at fine scales may not be evident at coarser scales, but this would not necessarily prove that those relationships did not exist; it could just be that that scale was inappropriate for demonstrating them (c.f. Dungan *et al.*, 2002). These criticisms notwithstanding, however, the results presented here are thought to be sufficiently robust that they will still be valid if they could be repeated at finer geographic scales. Moreover, the very strong correlations between species level and higher taxonomic levels (see Chapter 3) imply that results from analyses of patterns of distribution at generic level are also going to apply at species level. However, the causal mechanisms for many of these generic-level patterns should be sought at species or sub-specific levels; but beyond demonstrating a strong correlation, exact causation is difficult to determine.

7.5.3 Possible biases in the data

The question of bias in the data revolves around its representativeness – do the data accurately represent actual distributions of genera found in nature, and do these generic distributions accurately represent underlying species' distributions? Although no rigorous comparisons between large herbaria have ever been done, the Herbarium at the Royal Botanic Gardens, Kew is widely regarded as the most comprehensive in the world, with not only one of the largest collections of specimens but also the greatest taxonomic and geographical spread within that collection. The collections cover all families, over 97% of genera and an estimated 70% of species known in the world (R. Govaerts, pers. comm.). An extensive, though not exhaustive, literature search when compiling the data used in this thesis augmented the records from the herbarium specimens by only another 4% (see Chapter 1). Furthermore, the presence-absence structure of the data means that a single reliably-identified specimen (or literature reference) is sufficient to stand as a distribution record; the vast majority of specimens at RBG Kew are thus superfluous to the genus distribution.

From the point of view of scoring genus distributions by the large, geo-political TDWG regions, therefore, there is great redundancy in the herbarium collections, and they are judged to be comprehensive enough to give accurate large-scale distributions of genera. Taxonomic research at RBG Kew has long focused on large, economically important families with great tropical diversity, such as Compositae, Gramineae, Leguminosae, Palmae; however, this does not mean that other families or geographical regions are under-represented as general collecting has been carried out all over the world and active exchange programmes are also underway with many institutions all over the world. It would be impossible for RBG Kew staff to have collected all of the specimens in its own collections; most have been given as gifts by other institutions. Therefore the herbarium collections at RBG Kew do not solely or even primarily reflect the research and collecting efforts of its own staff but of generations of botanists from around the world, producing a huge collection of unmatched comprehensiveness.

At species level, however, the RBG Kew collections are less comprehensive, and it is possible that in some cases species on the extremity of a genus' distribution are not represented at RBG Kew and the distribution of that genus will therefore be more widespread than the collections show. Consulting authoritative regional floras has hopefully picked up many of these, although as this was not done systematically for each genus this may not be the case. In general, the tropical regions are represented by a greater number of more-recently collected specimens than are the temperate regions in the RBG Kew collections, although of course the tropical regions are themselves inherently much more species-rich than the temperate regions and the rate of discovery of new species is greater for the tropics than for temperate regions (R. Davies, pers. comm.). The key factors in assessing the representativeness of the RBG Kew collections are that the TDWG regions are so large that few distribution records for any genus are not represented by at least some specimens from that region, and also that even the areas of the world which are poorly represented by the standards of the RBG Kew Herbarium are nevertheless well represented for the purposes of this thesis.

If there are biases, they are likely to be towards areas of historical and continuing research activity by RBG Kew (the tropics, particularly the Old World tropics; and large, economically-important families), although it is more the case that these areas are over-represented (with more duplication of specimens of each taxon) than that other areas are under-represented (with a higher proportion of taxa completely unrepresented). However, judging the potential biases in the data remains very difficult because there is simply no reliable data against which to compare the RBG Kew collections: we cannot be sure if they under-represent certain areas because we don't know what the diversities of those areas should actually be. Any remaining biases in the representativeness of the collections will obviously influence the results of the analyses, however. For example, under-representation of temperate taxa will inflate the strength of floristic relationships between tropical regions (see Chapter 5); and the number of unique, widespread taxa is likely to be inflated if genera from temperate regions are poorly and patchily represented (see Chapter 6). The latitudinal gradient of diversity would likely be overestimated (see

Chapter 4), and so the strength of the results from the mid-domain effect simulations would therefore be inflated.

A more serious potential bias is the reliability or otherwise of the estimated counts of species richness for each TDWG region (see Chapter 2); simple extrapolation by two from a half-completed taxonomic edit of the Index Kewensis carries with it several implicit assumptions (i.e. that the families tackled in the first-completed half are representative of the rates of synonymy, the distribution and the size of those families in the uncompleted half). If the extrapolation of species numbers proves to be wrong, then obviously the correlation between patterns of diversity at different taxonomic levels and thus the whole issue of the use of higher taxa as surrogates for species level patterns is brought into question. Patterns of diversity of genera around the world would therefore be overestimating the richness of the tropical regions relevant to the temperate regions (see Chapter 3). In this respect the corroboration between the results presented in this thesis and those of independent studies is reassuring.

7.5.4 Why classify distribution patterns?

A further question which also needs to be addressed concerns the utility and objectivity of floristic classifications – what is actually the point of trying to produce a classification of plant distribution patterns? An uncompromising cladistic position would be that only systems in which there is known to be *a priori* a recoverable, discoverable hierarchical pattern are amenable to objective classification (in this case, a pattern of sister relationships produced by the process of evolution) – which can be represented as a branching diagram estimated by phylogenetic techniques (Hennig, 1966; Scotland, 1992; Nelson *et al.*, 2003). It can be argued that biogeographic patterns classifications do not seek to reconstruct a recoverable hierarchical pattern which has resulted from a single historical process, and that therefore such classifications cannot hope to be objective and repeatable (but see also McLaughlin, 1989, 1992). Undoubtedly important as this philosophical stance has been and continues to be within systematics, however, it should be clear that it has not been followed here. To follow this line of argument would be to overlook many successful applications of classification procedures in other fields, including biogeography, which lie outside the limits of systematics in the strict sense. For example, the *k*-means partitioning methodology followed in this thesis is directly analogous to the image-processing methodologies used in remote sensing studies of vegetation and in habitat classifications (Lillesand & Keifer, 1994).

These techniques were first adopted by this field because there was an obvious need to classify vegetation – after all, a forest is clearly different from a prairie – but upon a more stable, objective and mathematically-rigorous basis, in order to better study questions of how vegetation patterns vary over space and time: for example, studying rates of deforestation and habitat loss. Furthermore, this approach is conceptually similar to pattern recognition and 'data-mining' procedures developed principally within the burgeoning bioinformatics field, which are often based upon variations on the *k*-means clustering

techniques (e.g. Wu *et al.*, 2002; Likas *et al.*, 2003), and which also lie behind many internet-based search engines. Indeed, the mathematical techniques are widely used in other areas of science wherever there is a need to analyse complex patterns of variation *not* caused purely by evolutionary descent – and the question of how to conceptualise and delimit spatial variation lies at the heart of the whole discipline of geography (Couclelis, 1999). So why classify plant distribution patterns? Because there is discontinuous variation within plant distribution patterns which suggests that different groups of plant taxa are responding in different ways either to the same evolutionary and ecological processes or to different evolutionary and ecological processes. In order to study what these processes might be and how they affect plant distribution patterns for different groups, it is first necessary to analyse these differences in distribution patterns between different groups of plants.

7.5.5 Lack of comparable studies

Having produced a global classification of generic plant distribution patterns, however, and having demonstrated the correlation between diversity and floristic complexity within regions, it has been difficult to then corroborate this correlation at other taxonomic and spatial scales. Studies similar to this one, analysing both patterns of diversity and patterns of distribution, are almost none. In general, the focus is either geographical, comparing relative diversities between different regions, or taxonomic, comparing patterns of distribution within a particular taxonomic group. For most parts of the world, especially the diverse tropical regions, comprehensive distribution information about the constituent taxa remains lacking. It is predicted, however, that areas confirmed by other studies as being especially diverse would also show a similar relationship between high diversity and high floristic complexity. Further work is then needed to try to establish causality: whether or not a region has a large number of taxa because they happen to overlap in distribution there, or whether there are intrinsic qualities within that region, for example heterogeneous habitats supporting a wide range of taxa of differing ecological tolerances, which allow such overlap in distribution patterns to occur in the first place. However, what is not clear from the analyses in this thesis is just how distribution patterns which differ at the global scale themselves differ within any one region, or just how subtle the ecological differences within a region need to be to support a diversity of taxa having such a range of distribution patterns on a global scale. That is, finer spatial differentiation between taxa would be evident at larger spatial scales within a region, and this spatial differentiation between taxa of different distribution patterns might account for the 'floristic complexity' of a region.

7.5.6 Time

Also missing from this thesis is an explicit consideration of temporal scales. Over long time scales, plant distribution patterns are compressed or expanded subject to climatically-driven fluctuations in available habitat, undergo partial or wholesale extinction from former ranges, and/or radiation into new niches and habitats. The lack of detailed data on how different plant distributions vary over time,

however, makes accounting for these changes in distribution patterns problematic. There is evidence that broad-scale patterns of plant distribution such as the latitudinal gradient of diversity extend well back into the geological past (Lidgard & Crane, 1988), but whether this has been a constant or a periodic feature is less certain: for example, it is also well-established from fossil evidence that plant groups now confined to tropical regions were more widespread at various times in the past. There is little doubt that global diversity of angiosperms has increased greatly since the Cretaceous, more than offsetting the concomitant decline of pteridophyte and gymnosperm taxa, and it therefore remains uncertain that areas (and taxa) which show great diversity today have always been so; indeed for taxa this cannot be true, as they must evolve from single common ancestors. Furthermore, recent developments in molecular phylogenetics have demonstrated that large radiations of new species can happen over extremely short time scales (Richardson *et al.*, 2001a; Richardson *et al.*, 2001b). It therefore remains difficult to set some of the results presented in this thesis into an appropriate temporal context – the patterns may only be evident at this period in time.

Within biology as a whole, the body of theory underlying certain fields can shape the entire outlook, or paradigm, under which further work in that field is carried out and interpreted; if two fields operating under different paradigms are investigating closely-related questions, these divergent paradigms can create an impasse in the theoretical understanding of those issues. In particular, a fundamental difference sometimes emerges in the respective outlooks of ecologists and evolutionists over the generation and maintenance of species diversity. For much of ecological theory, a central assumption is the existence of steady-state equilibria in total diversity (Rosenzweig, 1995); for much of evolutionary theory, however, the focus is on particular clades of species under the assumption of individual evolutionary histories for each, and the historical dimension is paramount. In addition to existing controversies within both ecological (e.g. Willig *et al.*, 2003) and evolutionary theory, this difference in outlook has often meant that advances in one field are not easily understood in the context of the other, and *vice versa*. Despite recent attempts to bridge this gap, this problem remains. Davies *et al.* (2004) demonstrated that results are phylogenetically random with respect to distribution size and available energy input. Phylogenetic randomness does not necessarily mean ecological randomness, however. Large-scale ecological patterns such as the relationship between area and diversity, the correlation in diversity at different taxonomic levels and the latitudinal gradient of diversity remain so strong and so universal that it is possible that phylogenetic considerations play only a small role in determining these. Something constrains the levels of diversity at a particular point on the globe, and these constraints seem more likely to be ecological than phylogenetic.

7.6 Possible directions for future work

To take the work presented here further forward would first require improvements in those areas highlighted above as currently being its limitations. Above all, there should be an increase in the scale of

the analyses. This could be done in one of two ways, or with both combined. Distributions of genera could be scored more finely – at either TDWG Level-3 units or at country-by-country level, or in terms of habitat classifications. Finer geographical resolution (scoring at TDWG Level-3 or country-by-country) would be more appropriate for widespread taxa, when, as indicated above, habitat information is either difficult to come by or the taxon is found in a large variety of habitats. Finer ecological resolution (scoring by habitat classifications), on the other hand, would be more appropriate for localised taxa, whose restricted ecological distribution would be more informative of the ecological and evolutionary relationships of that habitat. Gathering sufficient information on either of these two aspects would necessitate the investment of several person-years, however. There would then be the problem of circularity if different habitats were analysed for their relative diversities and floristic relationships, since these features depend on just how the habitats have been classified in the first place – change the habitat classification and you change the results. A possible solution to this would be to use a more objectively defined remotely-sensed classification, although several exist. Also, in genera where different species have evolved to occupy different habitats then scoring these as present in several habitats would obscure the finer ecological differentiation visible at species level.

The other approach to improve the scale of these analyses would therefore mean working at the level of species and below. Given that a feature of this thesis has explicitly been that the analysis of a comprehensive global dataset reveals a more balanced picture of large-scale patterns, however, the obstacles to undertaking similar studies at the species level – a more-or-less complete global species checklist, for example – would probably remain insurmountable. Furthermore, correlation between patterns of generic diversity and existing studies done at species level (see Chapters 3 and 4) suggests that these are robust patterns which are unlikely to change with the addition of more data. Species-level data to test this assertion might be chosen from a random sample of genera, but further work at species-level might perhaps be better directed into areas of particular interest such as finer distributions for those genera already identified as localised – for example genera endemic to a particular region – or instead maybe a random sample of floristic elements identified in Chapter 6 could be selected and species-level distributions analysed for each to better assess the levels of spatial heterogeneity within clusters – perhaps also taking this analysis down to the level of individual georeferenced herbarium specimens. A further issue which needs to be addressed, as indicated above, is to compare the spatial and ecological heterogeneity within a particular region for taxa whose distributions differ outwith that region: if the distributions of the different floristic elements differ within the region, this implies that it is the heterogeneity intrinsic to the region which accounts for its diversity; if the distributions appear not to differ within the region, this implies that it is the wider ecological tolerances intrinsic to the taxa, rather than to the region, which differentiate them outside of that region but allow them to overlap there.

All of these above efforts, however, would necessarily result in a larger and even more complex data set. Concomitant with any of these approaches, therefore, would have to be parallel developments in technical ability (methods used in this thesis would be too slow and cumbersome to process such large

quantities of information) and also, and perhaps most importantly, further theoretical advances. Current debates about causes of the latitudinal gradient of diversity, for example, (e.g. Hawkins *et al.*, 2003; Willig *et al.*, 2003), a pattern known for over two hundred years (see Chapter 4), suggest that it is lack of insight rather than lack of good data which is preventing its understanding, the plethora of causal explanations which have been put forward for it notwithstanding. The issue of scale is central to future progress, with some patterns seemingly only apparent at certain scales while others seem to be evident at repeating scales. Ascertaining the correct scale at which to analyse the question of interest is therefore critical. Perhaps traditional, static approaches comparing different patterns at the same scale, or studying single patterns separately at different scales, can be replaced by more innovative approaches to problems of scale utilising fractal mathematics (Enquist & Niklas, 2001; Hubbell, 2001; Enquist *et al.*, 2002).

7.7 Conclusions

What can be concluded from the work presented in this thesis? The dataset analysed in this study is comprehensive but extremely coarse; this has meant that many factors responsible for plant distribution patterns that show finer patterns of geographical variation, such as temperature or rainfall or combinations of these such as potential evapo-transpiration, have not been studied. Nevertheless, some strong results have emerged from this study that, generally, corroborate patterns found in previous studies at finer geographical and/or taxonomic scales and imply that the results from the analysis of generic distribution patterns may also hold true at species level. Returning therefore to the specific objectives of this thesis set out in Chapter 1, and given in italics below, the following conclusions can be drawn.

To establish what are the broad patterns in the diversity of families, genera and species in different regions around the world (Chapter 3).

Patterns in the diversity of genera (and families) around the world mirror patterns in the diversity of species, and so at a regional scale genera can be used as a reliable surrogate for species-level diversity. The frequency distributions of range sizes and distribution patterns are both highly skewed, so that, at genus level, the majority of angiosperm diversity is explained by relatively few common distribution patterns. The different sizes of regions complicates comparison of their diversities; data at different taxonomic levels loosely fit the species-area model, though there is considerable scatter in the spread of points, mostly due to latitude. True relative diversity scores for regions of different size can be estimated by re-scaling with the species-area relationship. Areas of the Neotropics consistently emerge as the most bio-diverse areas of the world at all geographical scales of analysis, and at both genus-level and species-level. The most diverse regions of all, in this analysis, are those which straddle biogeographic regions.

To investigate how the size of a region, its latitudinal position and the distances between regions influence the number of taxa found within that region (Chapters 3 and 4).

There is a moderate relationship between the diversity of a region and the area of that region. If this relationship is analysed with successively-nested regions, however, there is a very strong relationship between diversity and area. However, the correlation between the diversity of an area and the number of endemic genera found there is poorer; some regions (27, Southern Africa; 29, Western Indian Ocean; 50, Australia) have far greater numbers of endemic genera than either their size or their latitudinal position would suggest. There is a strong latitudinal gradient in diversity which is roughly symmetrical around the equator. Tropical regions contain many more genera than do temperate regions, with the Neotropics more diverse than SE. Asia, which in turn is more diverse than Africa. Regions closer together have more genera in common than do more distant regions, although this relationship is more a product of the latitudinal difference between regions than it is simply of geographical distance between them, implying that genera are constrained in latitudinal rather than longitudinal bands.

To investigate the roles of available area and continental shape in determining the range sizes of taxa and the richness of tropical regions (Chapter 4).

The tropics contain far more land area than do other climatic zones, and this partly explains the greater biological diversity of tropical regions. Latitudinal range-sizes appear to decrease towards the equator (Rapoport's Rule), but this is an artefact and in fact latitudinal range sizes show a general increase in tropical regions; Rapoport's Rule is better explained as a local phenomenon over a short range of latitudes in the Northern Hemisphere. The principal peak in latitudinal range-size coincides with the weighted centroid for the world at 14°N (the latitude of the world with greatest land area), although the frequency distribution of latitudinal range size has many secondary peaks. The range-size frequency distribution of latitudinal ranges around the equator is explained well by a geometrically-constrained null model known as the mid-domain effect. However, expression of the mid-domain effect is highly dependent on large-ranged genera. Overall, the geometry of the land area of the Earth seems to play a large part in explaining the presence of the latitudinal gradient of diversity.

To describe the floristic relationships between different regions, and to compare these with existing global schemes of floristic classification (Chapter 5).

In general, strong continental clusters of regions emerge from the analysis of floristic relationships between regions. This pattern is a product of genera predominantly having small ranges and adjacent regions therefore being more similar floristically than are distant regions. These distinct continental clusters correspond to the floristic kingdoms and subkingdoms of traditional floristic hierarchies. There is broad agreement between the analyses presented here and traditional hierarchical schemes of floristic regions. However, some regions (for example 36, China, and 38, Eastern Asia) consistently show stronger relationships with regions other than those shown in these floristic hierarchies, and neither are the degrees of endemism of different regions taken into account in this study.

To study patterns of plant distribution around the world and produce global classifications of family and genus distribution patterns (Chapter 6).

Distribution patterns can be grouped into larger floristic elements based on their biogeographic similarity. The result of such a classification was to reduce the noise in the data, representing the diversity of genera within a region by fewer floristic elements, while preserving the overall biogeographic structure within the original distribution data. After grouping genera into repeating distribution patterns, the majority of diversity was still explained by few distributions; however, the number of unicate distributions was reduced considerably. When analysing this classification region by region there is a fairly strong relationship between the generic diversity of a region and the floristic complexity of a region, which can be measured as the number of floristic elements. Floristically, regions can be grouped into collections of regions showing common geographical characteristics rather than simple geographical proximity. The least diverse regions are islands which consist of a combination of endemic and widespread taxa, and thus show little floristic similarity with other regions. The most diverse regions of all are those which straddle major biogeographic regions and thus contain strong floristic links with both temperate and tropical floras.

REFERENCES

Angiosperm Phylogeny Group (1998). An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* **85(4)**: 531-553.

Angiosperm Phylogeny Group (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. *Botanical Journal of the Linnean Society* **141**: 399-436.

Albach, D.C., M.M. Ortega-Martínez, M.A. Fischer & M.W. Chase (2004). A new classification of the tribe Veroniceae – problems and a possible solution. *Taxon* **53(2)**: 429–452.

Archibold, O.W. (1995). *Ecology of World Vegetation*. Chapman & Hall, London, U.K.

Arnold, T.H. & B.C. de Wet (1993). *Plants of southern Africa: names and distribution*. Memoirs of the Botanical Survey of South Africa, no. 62. National Botanical Institute, Pretoria, South Africa.

Arrhenius, O. (1921). Species and area. *Journal of Ecology* **9**: 95–99.

Bachmann, S., W. J. Baker, N.A. Brummitt, J. Dransfield & J. Moat (2004). Elevational gradients, area and tropical island diversity: an example from the palms of New Guinea. *Ecography* **27(3)**: 299-310.

Backlund, A. & K. Bremer (1998). To be or not to be — principles of classification and monotypic plant families. *Taxon* **47(2)**: 391–400.

Bailey, T.C. & A.C. Gatrell (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical, Harlow, U.K.

Balmford A., M.J.B. Green & M.G. Murray (1996). Using higher-taxon richness as a surrogate for species richness.1. Regional tests. *Proceedings of the Royal Society, Series B* **263** (1375): 1267–1274.

Barraclough, T.G., A.P. Vogler & P.H. Harvey (1999). Revealing the factors that promote speciation. Pp. 202–219 in A.E. Magurran & R.M. May (eds.) *Evolution of Biological Diversity*. Oxford University Press, Oxford, U.K.

Barraclough, T.G. & V. Savolainen (2001). Evolutionary rates and species diversity in flowering plants. *Evolution* **55(4)**: 677–683.

Barthlott, W., W. Lauer, & A. Placke (1996). Global distribution of species diversity in vascular plants. *Erdkunde* **50**: 317–327.

Bateman, R.M. & W.A. DiMichele (1994). Saltational evolution of form in vascular plants: a neoGoldschmidtian synthesis. Pp. 61–100 in D.S. Ingram & A. Hudson, (eds.) *Shape and Form in Plants and Fungi*. Linnean Society Symposium Series 16. Academic Press, London, U.K.

Beals, E.W. (1984). Bray-Curtis ordination: and effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research* **14**: 1–55.

Belbin, L. & C. McDonald (1993). Comparing three classification strategies for use in ecology. *Journal of Vegetation Science* **4**: 341–348.

Bibby, C.J., N.J. Collar, M.J. Crosby, M.F. Heath, C. Imboden, T.H. Johnson, A.J. Long, A.J. Stattersfield & S.J. Thirgood (1992). *Putting biodiversity on the map: priority areas for global conservation*. ICBP (Birdlife International), Cambridge, U.K.

Biondini M.E., P.W. Mielke, Jr. & K.J. Berry (1988). Data-dependent permutation techniques for the analysis of ecological data. *Vegetatio* **75**: 161–168.

Blackburn T.M. & K.J. Gaston (1996). The distribution of bird species in the New World: patterns in species turnover. *Oikos* **77(1)**: 146–152.

Blackburn T.M. & K.J. Gaston (1997). The relationship between geographic area and the latitudinal gradient in species richness in New World birds. *Evolutionary Ecology* **11(2)**: 195–204

Braithwaite, L.W., M.P. Austin, M. Clayton, J. Turner & A.O. Nicholls (1989). On predicting the presence of birds in *Eucalyptus* forests. *Biological Conservation* **50**: 33–50.

Bramley, G.L.C., R.T. Pennington, R. Zakaria, S.S. Tjitrosoedirdjo & Q.C.B. Cronk (2004). Assembly of tropical plant diversity on a local scale: *Cyrtandra* (Gesneriaceae) on Mount Kerinci, Sumatra. *Biological Journal of the Linnean Society* **81(1)**: 49–62.

Braun-Blanquet, J. (1965). *Plant Sociology: the study of plant communities*. Hafner, London, U.K.

Bray, J.R. & T.J. Curtis (1957). An ordination of the upland forest communities in southern Wisconsin. *Ecological Monographs* **27**: 325–349.

Bromham, L & M. Cardillo (2003). Testing the link between the latitudinal gradient in species richness and rates of molecular evolution. *Journal of Evolutionary Biology* **16(2)**: 200–207.

Brown, J.H. (1995). *Macroecology*. University of Chicago Press, Chicago, Illinois, U.S.A.

Brummitt, N.A. & E. Nic Lughadha (2003). Biodiversity – where's hot and where's not. *Conservation Biology* **17(5)**: 1442–1448.

Brummitt, R.K. (comp.) (1992). *Vascular Plant Families and Genera*. Royal Botanic Gardens, Kew, U.K.

Brummitt, R.K. (1997). Taxonomy versus cladonomy: a fundamental controversy in biological systematics. *Taxon* **46(4)**: 723–734.

Brummitt, R.K. (2001). *World Geographical Scheme for Recording Plant Distributions*, 2nd edition. Hunt Institute for Botanical Documentation, Pittsburgh, U.S.A.

Brundin, L. (1966). Transantarctic relationships and their significance as evidenced by midges. *Kungliga Svenska Vetenskapsakademiens Handlinger* (Series 4) **11**: 1–472.

Buffon, G.L.L. Comte de (1776) . *Histoire naturelle générale et particulière, supplement III: Servant de suite à l'histoire des animaux quadrupèdes.* A Paris de l'Imprimerie Royale, France.

Burlando, B. (1990). The fractal dimension of taxonomic systems. *Journal of Theoretical Biology* **146**: 99–146.

Calinski, T. & J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics* **3**: 1–27.

Cantino, P.D. (2000). Phylogenetic nomenclature: addressing some concerns. *Taxon* **49**: 85-93.

Cardillo M., J.S. Huxtable & L. Bromham (2003). Geographic range size, life history and rates of diversification in Australian mammals. *Journal of Evolutionary Biology* **16(2)**: 282-288.

Causton, D.R. (1988). *An Introduction to Vegetation Analysis: principles, practice and interpretation*. Unwin Hyman, London, U.K.

Charkevicz, S.S. (ed.) *Plantae Vasculares Orientis Extremi Sovietici*. Vol. 1–. 'Nauka', St. Petersburg, Russia.

Chown, S.L., K.J. Gaston (2000). Areas, cradles and museums: the latitudinal gradient in species richness. Trends in Ecology & Evolution **15(8)**: 311–315.

Clarke, G.P., K. Vollesen & L.B. Mwasumbi (2000). Vascular plants. Pp. 129–147 in N.D. Burgess & G.P Clarke, eds. *Coastal Forests of Eastern Africa*. World Conservation Union, Gland, Switzerland & Cambridge, U.K.

Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18**: 117–143.

Clayton, W.D. (1972). Some aspects of the genus concept. *Kew Bulletin* **27 (2)**: 281–287.

Clayton, W.D. (1974). The logarithmic distribution of Angiosperm families. *Kew Bulletin* **29**: 271–279.

Clayton, W.D. (1983). The genus concept in practice. *Kew Bulletin* **38 (2)**: 149–153.

Coates, A.G. & J.G. Obando (1996). The geologic evolution of the Central American isthmus. Pp. 21–56 in J.B.C. Jackson, A.F. Budd & A.G. Coates, eds. *Evolution and Environment in Tropical America*. University of Chicago Press, Chicago, U.S.A.

Cody, M.L. (1975). Towards a theory of continental diversities: bird distribution over mediterranean habitat gradients. Pp. 214–257 in M.L. Cody & J.M. Diamond (eds.) *Ecology and Evolution of Communities*. Belknap Press, Cambridge, Massachusetts, U.S.A.

Colwell, R. K. (2000). RangeModel: a Monte Carlo simulation tool for assessing geometric constraints on species richness. Version 3. User's guide and application. http://viceroy.eeb.uconn.edu/asn.

Colwell, R.K. & G.C. Hurtt (1994). Nonbiological gradients in species richness and a spurious Rapoport effect. *American Naturalist* **144(4)**: 570–595.

Colwell, R.K. & D.C. Lees (2000). The mid-domain effect: geometric constraints on the geography of species richness. *Trends in Ecology and Evolution* **15**: 70–76.

Colwell R.K., C. Rahbek & N.J. Gotelli (2004). The mid-domain effect and species richness patterns: What have we learned so far? *American Naturalist* **163(3)**: 1–23.

Comes, H.P. & J.W. Kadereit (2003). Spatial and temporal patterns in the evolution of the flora of the European Alpine System. *Taxon* **52(3)**: 451–462.

Condit, R., N. Pitman, E.G. Leigh, Jr., J. Chave, J. Terborgh, R.B. Foster, P. Núñez V., S. Aguilar, R. Valencia, G. Villa, H.C. Muller-Landau, E. Losos & S.P. Hubbell (2002). Beta-diversity in tropical forest trees. *Science* **295**: 666-669.

Conran, J.G. (1995). Family distributions in the Liliiflorae and their biogeographical implications. *Journal of Biogeography* **22**: 1023–1034.

Connor, E.F. & E.D. McCoy (1979). The biology and statistics of the species-area relationship. *The American Naturalist* **113**: 791–833.

Cox, C.B. (2001). The biogeographic regions reconsidered. *Journal of Biogeography* **28(4)**: 511–523.

Cox, C.B. & P.D. Moore (1993). *Biogeography: an ecological and evolutionary approach*. Blackwell's, Oxford, U.K.

Couclelis, H. (1999). Space, time, geography. Pp. 29–38 in P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (eds.) *Geographical Information Systems. Volume 1: principles and technical issues*, 2nd edition. John Wiley & Sons, New York, U.S.A.

Cracraft, J. (1987). Species concepts and the ontology of evolution. *Biology and Philosophy* **2**: 329–346.

Cracraft, J. (1988). Deep-history biogeography: retrieving the historical pattern of evolving continental biotas. *Systematic Zoology* **37(3)**: 231–236.

Craw, R. (1989). Quantitative panbiogeography: introduction to methods. *New Zealand Journal of Zoology* **16(4)**: 485–494.

Craw, R.C., J.R. Grehan & M.J. Heads (1999). *Panbiogeography : tracking the history of life*. Oxford University Press, Oxford, U.K.

Crisp, M.D., J.G. West & H.P. Linder (1999). Biogeography of the terrestrial flora. Pp. 321–367 in A.E. Orchard (ed.) *Flora of Australia, Volume 1: Introduction*, 2nd edition. ABRS/CSIRO, Canberra, Australia.

Crisp, M.D., S. Laffan, H.P. Linder & A. Monro (2001). Endemism in the Australian flora. *Journal of Biogeography* **28(2)**: 183–198.

Croizat, L. (1964). *Space, time, form: the biological synthesis*. Published by the author, Caracas, Venezuela

Cronk, Q.C.B. (1989). Measurement of biological and historical influences in plant classifications. *Taxon* **38**: 357–370.

Cronk, Q.C.B. (2000). *The Endemic Flora of St Helena*. Anthony Nelson, Oswestry, U.K.

Cronk, Q.C.B. & J.L. Fuller. (2001). *Plant invaders : the threat to natural ecosystems*. Earthscan, London, U.K.

Currie D.J. & V. Paquin (1987). Large-scale biogeographical patterns of species richness of trees. *Nature* **329** (6137): 326–327.

Currie, D.J. (1991). Energy and large-scale patterns of animal-species and plant-species richness. *American Naturalist* **137(1)**: 27–49.

Currie D.J. & A.P. Francis (2004). Taxon richness and climate in angiosperms: Is there a globally consistent relationship that precludes region effects? Reply. *American Naturalist* **163(5)**: 780–785.

Darlington, P.J., Jr. (1965). *Biogeography of the Southern End of the World*. Harvard University Press, Cambridge, Massachusetts, U.K.

Darwin, C. (1859). *On the Origin of Species by the Means of Natural Selection, or the preservation of favoured races in the struggle for life*. John Murray, London, U.K.

Davies T.J., T.G. Barraclough, M.W. Chase, P.S. Soltis, D.E. Soltis & V. Savolainen (2004). Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences, U.S.A.* **101(7)**: 1904–1909.

Davis, P.H. & V.H. Heywood (1963). *Principles of Angiosperm Taxonomy*. University of Edinburgh Press, Edinburgh, U.K.

Davis, S.D., V.H. Heywood, O. Herrera-MacBryde, J. Villa-Lobos & A.C. Hamilton (1994-1997). *Centres of Plant Diversity: a guide and strategy for their conservation*. 3 volumes. World Wide Fund for Nature and World Conservation Union, Cambridge, U. K.

de Candolle, A.P. (1820). Essai élémentaire de géographie botanique. In *Dictionaire des Sciences Naturelles*, Volume 18. Flevrault, Strasbourg, France.

de Queiroz, K. & Gauthier, J. (1990). Phylogeny as a central principle in taxonomy: phylogenetic definitions of taxon names. *Systematic Zoology* **39**: 307-322.

de Queiroz, K. & J. Gauthier (1994). Toward a phylogenetic system of biological nomenclature. *Trends in Ecology and Evolution* **9**: 27–31.

Dial, K.P. & J.M. Marzluff (1989). Nonrandom diversification within taxonomic assemblages. *Systematic Zoology* **38**: 26–37.

Dick, C.W., K. Abdul-Salim & E. Bermingham (2003). Molecular systematic analysis reveals cryptic tertiary diversification of a widespread tropical rain forest tree. *American Naturalist* **162(6)**: 691–703.

Doebley, J. & R.-L. Wang (1997). Genetics and the evolution of plant form: an example from maize. *Cold Spring Harbour Symposia on Quantitative Biology* **22**: 361–367.

Donoghue, M.J., C.D. Bell & J.H. Li (2001). Phylogenetic patterns in Northern Hemisphere plant geography. *International Journal of Plant Sciences* **162**, Suppl. 6: S41–S52.

Dufrêne, M. & P. Legendre (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* **67**: 345–366

Dungan, J.L, J.N. Perry, M.R.T. Dale, P. Legendre, S. Citron-Pousty, M.-J. Fortin, A. Jakomulska, M. Miriti & M.S. Rosenberg. (2002). A balanced view of scale in spatial statistical analysis. *Ecology* **25**: 626–640.

Editorial Board of Flora of Zhejiang (1989–1993). *Flora of Zhejiang*. Zhejiang Science & Technology Press, China.

Engler, A. & L. Diels (1936). *Syllabus der Pflanzenfamilien*, auflage 11. Verlag von Gebrüder Borntrager, Berlin, Germany.

Enquist, B.J. & K.J. Niklas (2001). Invariant scaling relations across tree-dominated communities. *Nature* **410**: 655–660.

Enquist, B.J., J.P. Haskell & B.H. Tiffney (2002). General patterns of taxonomic and biomass partitioning in extant and fossil plant communities. *Nature* **419**: 610–613.

Ereshefsky, M. (1991). Species, higher taxa, and the units of evolution. *Philosophy of Science* **58**: 84–101.

Ereshefsky, M. (1992). *The Units of Evolution: essays on the nature of species*. MIT Press, Cambridge, U.S.A.

Erwin, D.H. (2000). Macroevolution is more than repeated rounds of microevolution. *Evolution and Development* **2**: 78–84.

Faith, D.P., P.R. Minchin & L. Belbin (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* **69**: 57–68.

Fernald, M.L. (1950). *Gray's Manual of Botany*, 8th edition. American Book Co., New York, U.S.A.

Fleishman, E., G.T. Austin & A.D. Weiss (1998). An empirical test of Rapoport's rule: elevational gradients in montane butterfly communities. *Ecology* **79(7)**: 2482–2493.

Florence, J. (1997). *Flore de la Polynésie Française*, Vol. 1–. Éditions ORSTOM, Paris, France.

Forey, P.L. (1992). Formal classification. Pp. 160–169 in P.L. Forey, C.J. Humphries, I.L Kitching, R.W. Scotland, D.J. Siebert & D.M. Williams, *Cladistics: a practical course in systematics*. Systematics Association Publication No. 10, Oxford University Press, Oxford, U.K.

Fosberg, F.R., M.-H. Sachet & R.L. Oliver (1979). A geographical checklist of Micronesian Dicotyledonae. *Micronesica* **15**: 41–295.

Fosberg, F.R., M.-H. Sachet & R.L. Oliver (1982). A geographical checklist of Micronesian Pteridophyta and Gymnospermeae. *Micronesica* **18**: 23–82.

Fosberg, F.R., M.-H. Sachet & R.L. Oliver (1987). A geographical checklist of Micronesian Monocotyledonae. *Micronesica* **20**: 19–129.

Francis, A.P. & D.J. Currie (2003). A globally consistent richness-climate relationship for angiosperms. *American Naturalist* **161(4)**: 523–536.

Frodin, D.G. (2001). *Guide to Standard Floras of the World*, 2nd edition. Cambridge University Press, Cambridge, U.K.

Furley, P.A., J. Proctor & J.A. Ratter (1992). *Nature and Dynamics of Forest-Savanna Boundaries*. Chapman & Hall, London, U.K.

Gaston, K.J. (1996a). What is biodiversity? Pp. 1–9 in K.J. Gaston (ed.) *Biodiversity: a biology of numbers and difference*. Blackwell Science, Oxford, U.K.

Gaston, K.J. (1996b). Species richness: measure and measurement. Pp. 77–113 in K.J. Gaston (ed.) *Biodiversity: a biology of numbers and difference*. Blackwell Science, Oxford, U.K.

Gaston, K.J. (1998a). Biodiversity – the road to an atlas. *Progress in Physical Geography* **22(2)**: 269–281.

Gaston, K.J. (1998b). Species-range size distributions: products of speciation, extinction and transformation. *Philosophical Transactions of the Royal Society of London, Series B* **353** (1366): 219–230.

Gaston, K.J. (2003). *The Structure and Dynamics of Geographic Ranges*. Oxford University Press, Oxford, U.K.

Gaston, K.J., T.M. Blackburn & J.I. Spicer (1998). Rapoport's Rule: time for an epitaph? *Trends in Ecology & Evolution* **13(2)**: 70–74.

Gaston, K.J. & P.H. Williams (1993). Mapping the world's species – the higher taxon approach. *Biodiversity Letters* **1**: 2–8.

Gentry, A.H. (1982). Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleistocene climatic fluctuations, or an accident of the Andean orogeny? *Annals of the Missouri Botanical Garden* **69**: 557–593.

Gentry, A.H. (1988). Changes in plant community diversity and floristic composition on environmental and geographical gradients. *Annals of the Missouri Botanical Garden* **75**: 1–34.

Ghiselin, M.T. (1974). A radical solution to the species problem. *Systematic Zoology* **23**: 536–544.

Good, R. (1974). *The Geography of the Flowering Plants*, 4th edition. Longmans, London, U.K.

Gould, S. J. (1979). An allometric interpretation of species-area curves: the meaning of the coefficient. *American Naturalist* **114(3)**: 335–343.

Govaerts, R. 2001. How many species of seed plants are there? *Taxon* **50**: 1085–1090.

Govaerts, R. 2003. How many species of seed plants are there? – a response. *Taxon* **52**: 583–584.

Gower, J.C. (1967). A comparison of some methods of cluster analysis. Biometrics 23: 623–637.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult taxonomic problem? *Systematic Biology* **47(1)**: 9–17.

Grehan, J.R. (1990). Panbiogeography: past, present and future. *New Zealand Journal of Zoology* **16**: 513–525.

Grubov, V.I. (2000). *Key to the Vascular Plants of Mongolia (with an atlas), Vols. 1 and 2.* Science Publishers, Inc. Enfield, New Hampshire, U.S.A. & Plymouth, U.K.

Grytnes, J.-A. & O.R. Vetaas (2002). Species richness and altitude: a comparison between null models and interpolated plant species richness along the Himalayan altitudinal gradient, Nepal. *American Naturalist* **519**: 294–304.

Hansen, A. & P. Sunding (1993). Flora of Macaronesia: checklist of vascular plants, 4th edition. *Sommerfeltia* **17**.

Harrison, S. (1997). How natural habitat patchiness affects the distribution of diversity in Californian serpentine chaparral. *Ecology* **78(6)**: 1898-1906.

Harold, A.S. & R.D. Mooi (1994). Areas of endemism: definition and recognition criteria. *Systematic Biology* **43(2)**: 261–266.

Hawkins, B.A. & J.A. Diniz-Filho (2002). The mid-domain effect cannot explain the diversity gradient of Nearctic birds. *Global Ecology and Biogeography* **11(5)**: 419–426.

Hawkins B.A. & J.A. Diniz-Filho (2004). 'Latitude' and geographic patterns in species richness. *Ecography* **27(2)**: 268–272.

Hawkins, B.A., R. Field, H.V. Cornell, D.J. Currie, J.F. Guegan, D.M. Kaufman, J..T Kerr, G.G. Mittelbach, T. Oberdorff, E.M. O'Brien, E.E. Porter & J.R.G. Turner (2003). Energy, water, and broad-scale geographic patterns of species richness. *Ecology* **84(12)**: 3105–3117.

Hedges, S.B. (2001). Biogeography of the West Indies: an overview. Pp. 15–33 in C.A. Woods & F.E. Sergile, eds. *Biogeography of the West Indies: patterns and perspectives, 2nd edition*. CRC Press, Boca Raton, U.S.A.

Hennig, W. (1966). *Phylogenetic Systematics*. University of Illinois Press, Urbana, Illinois, U.S.A.

Hepper, F.N. (1989). West Africa. Pp. 189–197 in D.G. Campbell & H.D. Hammond (eds.) *Floristic Inventory of Tropical Countries*. New York Botanical Garden, New York, U.S.A.

Hill, M.O. (1973). Reciprocal averaging: and eigenvector method of ordination. *Journal of Ecology* **61**: 237–249.

Hill, M.O. (1979a). DECORANA – a Fortran program for detrended correspondence analysis and reciprocal averaging. Ecology and Systematics, Cornell University, Ithaca, New York, U.S.A.

Hill, M.O. (1979b). TWINSPAN – a Fortran program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Ecology and Systematics, Cornell University, Ithaca, New York, U.S.A.

Hill, M.O. & Gauch, H.G. (1980) Detrended correspondence analysis: an improved ordination technique. *Vegetatio* **42**: 47–58.

Hillis, D.M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* **47(1)**: 3–8.

Hnatiuk, R.J. (1990). *Census of Australian Vascular Plants*. Australian Fauna and Flora Series, no. 11. Bureau of Flora and Fauna, Canberra, Australia.

Hooker, J.D. (1853). *The botany of the Antarctic voyage of HM discovery ships* Erebus *and* Terror *in the years 1839–1843. Volume II: Flora Novae-Zelandiae, Part 1; Flowering plants.* Lovell Reeve, London, U.K.

Hovenkamp, P. (1997). Vicariance events, not areas, should be used in biogeographical analysis. *Cladistics* **13**: 67–79.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton University Press, Princeton, New Jersey, U.S.A.

Hull, D.L. (1965). The effect of essentialism on taxonomy — two thousand years of stasis. *British Journal for the Philosophy of Science* **15**: 314–326.

Hull, D.L. (1978). A matter of individuality. *Philosophy of Science* **45**: 335–360.

Hull, D.L. (1980). Individuality and selection. *Annual Review of Ecology and Systematics* **11**: 311–332.

Hultén, E. (1941–1950). *Flora of Alaska & Yukon*. 10 parts. *Acta Universitatis Lundensis* **37–46**.

Humboldt F.A. von & A.J.A. Bonpland (1805). *Essai sur la géographie des plantes.* Levrault, Schoell & Co., Paris, France.

Humphries, C.J. (1979). Endemism and evolution in Macaronesia. Pp. 171–199 in D. Bramwell (ed.) *Plants and Islands.* Academic Press, London, U.K.

Humphries, C.J. (1981). Biogeographical methods and the southern beeches (Fagaceae: *Nothofagus*). Pp. 177–207 in Funk, V.A. & Brooks, D.R. (eds.) *Advances in Cladistics: proceedings of the first meeting of the Willi Hennig Society.* The New York Botanical Garden, New York, U.S.A.

Humphries, C.J. (2000). Form, space and time; which comes first? *Journal of Biogeography* **27(1)**: 11–15.

Humphries, C.J. (2001). Hotspots: going off the boil? *Diversity and Distributions* **7**: 104–105.

Humphries, C.J. & L.R. Parenti (1999). *Cladistic Biogeography: interpreting patterns of plant and animal distributions*, 2nd edition. Oxford University Press, Oxford, U.K.

Jetz, W. & C. Rahbek, (2001). Geometric constraints explain much of the species richness pattern of African birds. *Proceedings of the National Academy of Sciences, U.S.A.* **98**: 5661–5666.

Jetz, W. & C. Rahbek, (2002). Geographic range size and determinants of avian species richness. *Science* **297**: 1548–1551.

Keng, Hsuan, Hong Der-Yuan & Chen Chia-Jui. (1993). *Orders and Families of Seed Plants of China*. World Scientific, Singapore.

Kenrick, P. & P.R. Crane (1997). *Origin and Early Diversification of Land Plants: a cladistic study.* Smithsonian Institution Press, Washington, U.S.A.

Kent, M. & P. Coker (1992). *Vegetation Description and Analysis: a practical approach.* Belhaven Press, London, U.K.

Knox, E.B. (1998). The use of hierarchies as organizational models in systematics. *Biological Journal of the Linnean Society* **63(1)**: 1–49.

Koleff, P. & K. J. Gaston (2001). Latitudinal gradients in diversity: real patterns and random models. *Ecography* **24**: 341–351.

Krasnoborov, I.M. *et al.* (1987–1999). *Flora Sibiri*, Vols 1–14. 'Nauka', Novosibirsk, Russia.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika* **29**: 1–27.

Kruskal, J.B. (1964b). Non-metric multidimensional scaling: a numerical method. *Psychometrika* **29**: 115–129.

La Ferla, B., J. Taplin, D. Ockwell & J.C. Lovett (2002). Continental scale patterns of biodiversity: can higher taxa accurately predict African plant distributions? *Botanical Journal of the Linnean Society* **138 (2)**: 225–235.

Lance, G.N. & W.T. Williams (1967). A general theory of classification sorting strategies. I: Hierarchical systems. *Computer Journal* **9**: 373–380.

Lance, G.N. & W.T. Williams (1968). A general theory of classification sorting strategies. II: Clustering systems. *Computer Journal* **10**: 271–277.

Laurie, H. & J.A. Silander, Jr. (2002).Geometric constraints and spatial patterns of species richness: a critique of range-based models. *Diversity and Distributions* **8**: 351–364.

Lees, D. C., C. Kremen & L. Andriamampianina (1999). A null model for species richness gradients: bounded range overlap of butterflies and other rainforest endemics in Madagascar. *Biological Journal of the Linnean Society* **67**: 529–584.

Legendre, P. & E. Gallagher (2002). Ecologically meaningful transformations for ordination of species data. *Oecologia* **129(2)**: 271–280.

Legendre, P. & L. Legendre (1998). *Numerical Ecology*, 2nd English edition. Elsevier Science, Amsterdam, The Netherlands.

Lewin, R. (1989). Biologists disagree over bold signature of nature. *Science* **244** (4904): 527–528

Lidgard S. & P.R. Crane (1988). Quantitative analyses of the early angiosperm radiation. *Nature* **331** (6154): 344–346.

Lidgard S. & P.R. Crane (1990). Angiosperm diversification and Cretaceous floristic trends: a comparison of palynofloras and leaf macrofloras. *Paleobiology* **16**: 77–93.

Likas, A., N. Vlassis & J.J. Verbeek (2003). The global *k*-means clustering algorithm. *Pattern Recognition* **36**: 451–461.

Lillesand, T.M. & R.W. Keifer (1994). *Remote Sensing and Image Interpretation*, 3rd edition. John Wiley & Sons, New York, U.S.A.

Linder, H.P (1996). Numerical analyses of African plant distribution patterns. Pp. 67-86 in C.R. Huxley, J.M. Lock & D.F. Cutler (eds.) *Chorology, Taxonomy and Ecology of the Floras of Africa and Madagascar*. Royal Botanic Gardens, Kew, U.K.

Linder, H.P. (2001). Plant diversity and endemism in sub-Saharan tropical Africa. *Journal of Biogeography* **28(2)**: 169–182.

Linder, H.P. & M.D. Crisp (1995). *Nothofagus* and Pacific biogeography. *Cladistics* **11**: 5–32.

Linnaeus, C. (1737a). *Genera Plantarum*, 1st edition. Leyden, The Netherlands.

Linnaeus, C. (1737b). *Critica Botanica*. Leyden, The Netherlands.

Linnaeus, C. (1753). *Species Plantarum*, 1st edition. Stockholm, Sweden.

Lomolino, M. V. (1989). Interpretations and comparisons of constants in the species-area relationship: an additional caution. *American Naturalist* **133(2)**: 277–280.

MacArthur, R.H. & E.O. Wilson (1967). *The Theory of Island Biogeography*. Princeton University Press, Princeton, New Jersey, U.S.A.

Mace, G.M., A. Balmford, L. Boitani, G. Cowlishaw, A.P. Dobson, D.P. Faith, K.J. Gaston, C.J. Humphries, R.I. Vane-Wright, P.H. Williams, J.H. Lawton, C.R. Margules, R.M. May, A.O. Nicholls, H.P Possingham , C. Rahbek & A.S. van Jaarsveld (2000). From hotspots towards conservation consensus. *Nature* **405**: 393.

MacPhee, R.D.E & D.A. Grimaldi (1996). Mammal bones in Dominican amber. *Nature* **380**: 489–490.

Margules, C.R., I.D. Cresswell & A.O. Nicholls (1994). A scientific basis for establishing networks of protected areas. Pp. 327–350 in P.L. Forey, C.J. Humphries & R.I. Vane-Wright, eds. *Systematics and Conservation Evaluation*. Systematics Association special volume no. 50. Clarendon Press, Oxford, U.K.

Margules, C.R. & R.L. Pressey (2000). Systematic conservation planning. *Nature* **405**: 243–253.

Marshall, L.G., S.D. Webb, J.J. Sepkoski & D.M. Raup (1980). Mammalian evolution and the great American interchange. *Science* **215**: 1351–1357.

May, R.M. (1988). How many species are there on Earth? *Science* **241**: 1441–1449.

Mayr, E. (1976). Is the species a class or an individual? *Systematic Zoology* **25**: 192.

Mayr, E. (1988). The ontology of the species taxon. Pp. 335–358 in E. Mayr, *Toward a New Philosophy of Biology*. Belknap Press, Cambridge, Massachusetts, U.S.A.

McCune, B. (1994). Improving community analysis with the Beals smoothing function. *Ecoscience* **1**: 82–86.

McCune, B. & M.J. Mefford (1999). *PC-ORD: Multivariate Analysis of Ecological Data, version 4.0*. MjM software design, Gleneden Beach, Oregon, U.S.A.

McCune, B. & Grace, J.B. (2002). *Analysis of Ecological Communities*. MjM software design, Gleneden Beach, Oregon, U.S.A.

McLaughlin, S.P. (1989). Natural floristic areas of the western United States. *Journal Of Biogeography* **16(3)**: 239–248.

McLaughlin, S.P. (1992). Are floristic regions hierarchically arranged? *Journal of Biogeography* **19**: 21–32.

Miller, R.I., S.P. Bratton & P.S. White (1987). A regional strategy for reserve design and placement based on an analysis of rare and endangered species' distribution patterns. *Biological Conservation* **39**: 255–268.

Milligan, G.W. & M.C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**: 159–179.

Minchin. P.R. (1987a). An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* **69**: 89–107.

Minchin, P.R. (1987b). Simulation of multidimensional community patterns - towards a comprehensive model. *Vegetatio* **71**: 145–156.

Minelli, A. *et al.* (1991). Self-similarity in biological classifications. *BioSystems* **26**: 89–97.

Mittermeier, R.A. (1988). Primate diversity and the tropical forest: case studies from Brazil and Madagascar and the importance of megadiversity countries. Pp. 145–154 in E.O. Wilson, ed. *Biodiversity*. National Academy Press, Washington DC, U.S.A.

Mittermeier, R.A., N. Myers, P.R. Gil & C.G. Mittermeier (1999). *Hotspots: Earth's biologically richest and most endangered terrestrial ecoregions*. CEMEX, Conservation International and Agrupacion Sierra Madre, Mexico.

Morrone, J.J. (1994). On the identification of areas of endemism. *Systematic Biology* **43(3)**: 438–441.

Mueller-Dombois, D. & H. Ellenberg (1974). *Aims and Methods of Vegetation Ecology*. John Wiley & Sons, New York, U.S.A.

Myers, A.A. & P.S. Giller (eds.) (1988). *Analytical Biogeography: an integrated approach to the study of animal and plant distributions*. Chapman & Hall, London, U.K.

Myers, N., R.A. Mittermeier, C.G. Mittermeier, G.A.B. da Fonseca & J. Kent (2000). Biodiversity

hotspots for conservation priorities. *Nature* **403**: 853–858.

Nekola, J.C. & P.S. White (1999). The distance decay of similarity in ecology and biogeography. *Journal of Biogeography* **26**: 867-878.

Nelson, G. (1978). From Candolle to Croizat: comments on the history of biogeography. *Journal of the History of Biology* **11**: 269–305.

Nelson, G. & Platnick, N. (1981). *Systematics and Biogeography: cladistics and vicariance*. Columbia University Press, New York, U.S.A.

Nelson, G., D.J. Murphy & P.Y. Ladiges (2003). Brummitt on paraphyly: a response. Taxon **52(2)**: 295–298.

Niklas, K.J. (1988). Patterns of vascular plant diversification in the fossil record – proof and conjecture. *Annals of the Missouri Botanical Garden* **75(1)**: 35–54.

Niklas, K.J. & B.H. Tiffney (1994). The quantification of plant biodiversity through time. *Philosophical Transactions of the Royal Society, Series B* **345** (1311): 35–44.

Noss, R.F. (1990). Indicators for monitoring biodiversity: a hierarchical approach. *Conservation Biology* **4**: 355–364.

O'Brien, E.M. (1998). Water-energy dynamics, climate, and prediction of woody plant species richness: an interim general model. *Journal of Biogeography* **25(2)**: 379–398.

O'Brien, E.M., Whittaker R.J. & R. Field (1998). Climate and woody plant diversity in southern Africa: relationships at species, genus and family levels. *Ecography* **21(5)**: 495–509.

O'Brien, E.M., R. Field & R.J. Whittaker (2000). Climatic gradients in woody plant (tree and shrub) diversity: water-energy dynamics, residual variation, and topography. *Oikos* **89(3)**: 588–600.

Oliveira-Filho, A.T. & J.A. Ratter (1995). A study of the origin of central Brazilian forests by the analysis of plant species distribution patterns. *Edinburgh Journal of Botany* **52(2)**: 141–194.

Orlóci, L. (1967). An agglomerative method for the classification of plant communities. *Journal of Ecology* **55**: 193–206.

Page, R.D.M. (1987). Graphs and generalised tracks: quantifying Croizat's panbiogeography. *Systematic Zoology* **36(1)**: 1–17.

Parmesan, C. & G. Yohe (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**: 37–42.

Pennington R.T., D.E. Prado & C.A. Pendry (2000). Neotropical seasonally dry forests and Quaternary vegetation changes. *Journal of Biogeography* **27(2)**: 261–273.

Pianka, E.R. (1966). Latitudinal gradients in species diversity: a review of concepts. *American Naturalist* **100**: 33–46.

Pimm, S.L. & J.H. Brown (2004). Domains of diversity. *Science* **304** (5672): 831–833.

Pineda, J. & H. Caswell (1998). Bathymetric species-diversity patterns and boundary constraints on vertical range distributions. *Deep-Sea Research Part II – Topical Studies in Oceanography* **45**: 83–101.

Poe, S. (1998). Sensitivity of phylogenetic estimation to taxonomic sampling. *Systematic Biology* **47(1)**: 18–31.

Prance, G.T. 1973. Phytogeographic support for the theory of Pleistocene forest refuges in the Amazon Basin based on evidence from distribution patterns in Caryocaryaceae, Chrysobalanaceae, Dichapetalaceae and Lecythidaceae. Acta Amazonica 3: 5–28.

Prance, G.T. (1979). History of exploration: South America. Pp. 55–70 in I. Hedberg (ed.) *Systematic Botany, Plant Utilization & Biosphere Conservation*. Almqvist & Wiksell International, Stockholm, Sweden.

Prance, G.T. (1994). A comparison of the efficacy of higher taxa and species numbers in the assessment of biodiversity in the neotropics. *Proceedings of the Royal Society, Series B* **345** (1311): 89–99.

Prendergast, J.R., R.M. Quinn, J.H. Lawton, B.C. Eversham & D.W. Gibbons (1993). Rare species, the coincidence of diversity hotspots, and conservation strategies. *Nature* **365**: 335–337.

Quézel, P. (1985). Definition of the Mediterranean region and the origin of its flora. In Gomez-Campo (ed.). *Plant Conservation in the Mediterranean Area*. Dr. W. Junk Publishers, Dordrecht, The Netherlands.

Qian, H. (2001). Floristic analysis of vascular plant genera of North America north of Mexico: spatial patterning of phytogeography. *Journal of Biogeography* **28**: 525–534.

Qian, H., J.-S. Song, P. Krestov, Q. Guo, Z. Wu, X. Shen & X. Guo (2003). Large-scale phytogeographic patterns in East Asia in relation to latitudinal and climatic gradients. *Journal of Biogeography* **30(1)**: 129–141.

Rahbek, C. (1997). The relationship among area, elevation, and regional species richness in Neotropical birds. *American Naturalist* **149**: 875–902.

Rahbek, C. & G.R. Graves (2001). Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences, U.S.A.* **98**: 4534–4539.

Rapoport, E.H. (1982). *Areography: geographical strategies of species*. Fundación Bariloche Series, Volume 1. Pergamon Press, Oxford, U.K.

Raven, P.H. & D.I. Axelrod (1974). Angiosperm biogeography and past continental movements. *Annals of the Missouri Botanical Garden* **61**: 539–673.

Richardson, I.B.K. (1978). Endemic taxa and the taxonomist. Pp 245–262 in H.E. Street (ed.) *Essays in Plant Taxonomy*. Academic Press, London, U.K. & New York, U.S.A.

Richardson, J.E., R.T. Pennington, T.D. Pennington & P.M. Hollingsworth (2001a). Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science* **293**: 2242–2245.

Richardson, J.E., F.M. Weitz, M.F. Fay, Q.C.B. Cronk, H.P. Linder, G. Reeves & M.W. Chase (2001b). Phylogenetic analysis of *Phylica* L. (Rhamnaceae) with an emphasis on island species: evidence from plastid *trn*L-F and nuclear internal transcribed spacer (ribosomal) DNA sequences. *Taxon* **50(2)**: 405–427.

Rico Arce, M. de L., M. Sousa S. & S. Fuentes S. (1999). *Guinetia*: a new genus in the tribe Ingeae (Leguminosae: Mimosoideae) from Mexico. *Kew Bulletin* **54(4)**: 975–981.

Roberts, D.W. (1986). Ordination on the basis of fuzzy set theory. *Vegetatio* **66**: 123–131.

Rohde, K. (1992). Latitudinal gradients in species diversity – the search for the primary cause. *Oikos* **65 (3)**: 514–527.

Rohde, K., M. Heap & D. Heap (1993). Rapoport's rule does not apply to marine teleosts and cannot explain latitudinal gradients in species richness. *American Naturalist* **142(1)**: 1–16.

Rohde, K. & M. Heap (1996). Latitudinal ranges of teleost fish in the Atlantic and Indo-Pacific Oceans.
American Naturalist **147(4)**: 659–665.

Rohde, K. (1997). The larger area of the tropics does not explain latitudinal gradients in species diversity. *Oikos* **79(1)**: 169–172.

Root, T.L., J.T. Price, K.R. Hall, S.H. Schneider, C. Rosenzweig & J.A. Pounds (2003). Fingerprints of global warming on wild animals and plants. *Nature* **421**: 57–60.

Rosen, D.E. (1975). A vicariance model of Caribbean biogeography. *Systematic Zoology* **24**: 431–461.

Rosen, D.E. (1985). Geological hierarchies and biogeographic congruence in the Caribbean. *Annals of the Missouri Botanical Garden* **72**: 636–659.

Rosenzweig, M.L. (1992). Species-diversity gradients – we know more and less than we thought. *Journal of Mammalogy* **73(4)**: 715–730.

Rosenzweig, M.L. (1995). *Species Diversity in Space and Time*. Cambridge University Press, Cambridge, U.K.

Rosenzweig, M.L. & E.A. Sandlin (1997). Special diversity and latitudes: listening to area's signal. *Oikos* **80(1)**: 172–176.

Roy K., D. Jablonski & J.W. Valentine. (1994). Eastern pacific molluscan provinces and latitudinal diversity gradient – no evidence for Rapoport's rule. *Proceedings of the National Academy of Sciences, U.S.A.* **91(19)**: 8871–8874.

Ruggiero, A. (1999). Spatial patterns in the diversity of mammal species: a test of the geographic area hypothesis in South America. *Ecoscience* **6**: 338–354.

Rzedowski, J. (1988). Diversidad y orígenes de la flora fanerogámica de México. *Simposio Diversidad Biológica de México*. Universidad Nacional Autónoma de México, Oaxtepec, México.

Savolainen, V., M.W. Chase, S.B. Hoot, C.M. Morton, D.E. Soltis, C. Bayer, M.F. Fay, A.Y. de Bruijn, S. Sullivan & Y.-L. Qiu (2000). Phylogenetics of flowering plants based on combined analyses of plastid *atp*B and *rbc*L gene sequences. *Systematic Biology* **49(2)**: 306–362.

Schimper, A.F.W. (1903). *Plant Geography upon a Physiological Basis*. Clarendon Press, Oxford, U.K.

Schoener, T.W. (1986). Patterns in terrestrial vertebrates *vs.* arthropod communities: do systematic differences in regularity exist? Pp. 556–586 in J. Diamond & T.J. Case (eds.) *Community Ecology*. Harper & Row, New York.

Schouw, J.F. (1823). *Grundzüge einer alldemeinen Pflanzengeographie*. Reimer, Berlin, Germany.

Sclater, P.L. (1858). On the general geographic distribution of the members of the class Aves. *Journal of the Linnean Society of London, Zoology* **2**: 130–145.

Scotland, R.W. (1992). Cladistic theory. Pp. 3–13 in P.L. Forey, C.J. Humphries, I.L Kitching, R.W. Scotland, D.J. Siebert & D.M. Williams, *Cladistics: a practical course in systematics*. Systematic Association Publication No. 10, Oxford University Press, Oxford, U.K.

Scotland, R.W. & M.J. Sanderson (2004). The significance of few versus many in the tree of life. *Science* **303**: 643.

Sepkoski, J.J, Jr. (1992). Phylogenetic and ecologic patterns in the Phanerozoic history of marine biodiversity. Pp. 77–100 in N. Eldredge (ed.) *Systematics, ecology and the biodiversity crisis*. Columbia University Press, New York, U.S.A.

Shepherd, R.N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* **27**: 125–139.

Shepherd, R.N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika* **27**: 219–246.

Shmida, A. & M.V. Wilson (1985). Biological determinants of species-diversity. *Journal of Biogeography* **12(1)**: 1–20.

Shubin, N.H. & C.R. Marshall (2000). Fossils, genes, and the origin of novelty. *Palaeobiology* **26(4)**: 324–340.

Skottsberg, C.J.F. (1954). Antarctic flowering plants. *Botanisk Tidskrift* **51**: 330–338.

Sneath, P.H.A & R.R. Sokal (1973). *Numerical Taxonomy: the principles andpractice of numerical classification*. W.H. Freeman, San Francisco, U.S.A.

Sober, E. (1988). The conceptual relationship of cladistic phylogenetics and vicariance biogeography. *Systematic Zoology* **37(3)**: 245–253.

Sokal, R.R. & C.D. Michener (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**: 1409–1438.

Soltis, P.S., D.E. Soltis and M.W. Chase (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**: 402–404.

Stearn, W.T. (1957). An introduction to the *Species Plantarum* and cognate botanical works of Carl Linnaeus. Introduction to Linnaeus, C. *Species Plantarum*, a facsimile of the first edition. Ray Society, London, U.K.

Stern, D. (1998). A role of *ultrabithorax* in morphological differences between species. *Nature* **396**: 463–466.

Stevens, G.C. (1989). The latitudinal gradient in geographical range - how so many species coexist in the tropics. *American Naturalist* **133(2)**: 240–256.

Stevens, G.C. (1992). The elevational gradient in altitudinal range - an extension of Rapoport's latitudinal rule to altitude. *American Naturalist* **140**: 893–911.

Stevens, G.C. (1996). Extending Rapoport's rule to Pacific marine fishes. *Journal of Biogeography* **23(2)**: 149–154.

Stott, P.A. (1981). *Historical Plant Geography*. George Allen & Unwin, London, U.K.

Takhtajan, A.L. (1986). *Floristic Regions of the World*. University of California Press, Berkeley & Los Angeles, U.S.A.

Tausch, R.J., D.A. Charlet, D.A. Weixelman & D.C Zamudio (1995). Patterns of ordination and classification instability resulting from changes in input order. *Journal of Vegetation Science* **6**: 897–902.

Taylor, P.H. & S.D. Gaines (1999). Can Rapoport's rule be rescued? Modelling causes of the latitudinal gradient in species richness. *Ecology* **80(8)**: 2474–2482.

ter Braak, C.J.F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167–1179.

Terborgh, J. (1973). On the notion of favourableness in plant ecology. *American Naturalist* **107**: 481–501.

Thorne, R.F. (1972). Major disjunctions in the geographic ranges of seed plants. *Quarterly Review of Biology* **47(4)**: 365–411.

Tryon, A.F. & Lugardon, B. (1990). *Spores of the Pteridophyta*. Springer-Verlag, New York, U.S.A.

Turner, H., P. Hovenkamp & P.C. van Welzen (2001). Biogeography of Southeast Asia and the West Pacific. *Journal of Biogeography* **28(2)**: 217–230.

van Welzen, P.C., H. Turner & P. Hovenkamp (2003). Historical biogeography of Southeast Asia and the West Pacific, or the generality of unrooted area networks as historical biogeographic hypotheses. *Journal of Biogeography* **30(2)**: 181–192.

Vane-Wright, R.I., C.J. Humphries & P.H. Williams (1991). What to protect? – systematics and the agony of choice. *Biological Conservation* **55**: 235–254.

Veech, J.A. (2000). A null model for detecting nonrandom patterns of species richness along spatial gradients. *Ecology* **81(4)**: 1143–1149.

Vvedensky, A.I. (ed.) (1968–1993). *Conspectus Florae Asiae Mediae*, Vols 1–10. 'FAN', Tashkent, Uzbekistan.

Wagner, W.L. & V.A. Funk (eds.) (1995). *Hawaiian Biogeography*. Smithsonian Institution Press, Washington DC., U.S.A.

Wagner, W.L., D.R. Herbst & S.H. Sohmer (1999). *Manual of the Flowering Plants of Hawai'i*, revised edition. Bishop Museum Special Publication, no. 97. Bishop Museum Press, Honolulu, Hawai'I, U.S.A.

Waide, R.B., M.R. Willig, C.F. Steiner, G.G. Mittelbach & L. Gough (1999). The relationship between productivity and species richness. *Annual Review of Ecology and Systematics* **30**: 247–300.

Wallace, A.R. (1876). *The Geographical Distribution of Animals* (2 Volumes). Macmillan, London, U.K.

Wallace, A.R. (1878). *Tropical Nature and other essays*. Macmillan, London, U.K.

Wallace, A.R. (1880). *Island Life: or the phenomena and causes of insular faunas and floras including a revision and attempted solution of the problem of geological climates*. Macmillan, London, U.K.

Ward, J.H. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association* **58**: 236–244.

Webb, D.A. (1978). *Flora Europaea* – a retrospect. *Taxon* **27**: 3–14.

Wen, J. (1999). Evolution of Eastern Asian and Eastern North American disjunct distributions in flowering plants. *Annual Review of Ecology and Systematics* **30**: 421–455.

Weston, P.H. & M.D. Crisp (1987). Evolution and biogeography of the waratahs. Pp. 17–34 in J.A. Armstrong (ed.) *Waratahs: their biology, cultivation and conservation*. Occasional Publication No. 9. Australian National Botanic Gardens, Canberra, Australia.

White, F. (1983). The vegetation of Africa: a descriptive memoir to accompany the Unesco/AETFAT/UNSO vegetation map of Africa. *Natural Resources Research* **20**. Unesco, Paris.

White, F. (1993). The AETFAT chorological classification of Africa: history, methods and applications. *Bulletin des Jardins Botanique National de Belgique* **62**: 225–281.

Whitmore, T.C. (1990). *An Introduction to Tropical Rain Forests*. Oxford University Press, Oxford, U.K.

Whittaker, R. J., K.J. Willis & R. Field 2001 Scale and species richness: toward a general hierarchical theory of species diversity. *Journal of Biogeography* **28**: 453–470.

Whittaker, R.H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* **30**: 279–338.

Whittaker, R.H. (1972). Evolution and measurement of species diversity. *Taxon* **21**: 213–251.

Williams, C.B. (1943). Area and the number of species. *Nature* **152**: 264–267.

Williams, C.B. (1964). *Patterns in the Balance of Nature*. Academic Press, London, U.K. and New York, U.S.A.

Williams, P.H. (1994). *WORLDMAP: priority areas for biodiversity.* Version 3.18. Published by the author, London, U.K.

Williams, P.H. (1996). Mapping variations in the strength and breadth of biogeographic transition zones using species turnover. *Proceedings of the Royal Society of London, series B* **263**: 579–588.

Williams, P.H. & K.J. Gaston (1994). Measuring more of biodiversity: can higher-taxon richness predict wholesale species richness? *Biological Conservation* **67**: 211–217.

Williams, P.H. & C.J. Humphries (1994). Biodiversity, taxonomic relatedness and endemism in conservation. Pp. 269–287 in P.L. Forey, C.J. Humphries & R.I. Vane-Wright, eds. *Systematics and Conservation Evaluation*. Systematics Association special volume no. 50. Clarendon Press, Oxford, U.K.

Williams, P.H. & C.J. Humphries (1996). Comparing character diversity among biotas. Pp. 54–76 in K.J. Gaston (ed.). *Biodiversity: a biology of numbers and difference*. Blackwell Science, Oxford, U.K.

Williams, P.H., C.J. Humphries & K.J. Gaston (1994). Centres of seed-plant diversity: the family way. *Proceedings of the Royal Society of London, series B* **256**: 67–70.

Willig, M.R., & S.K. Lyons (1998). An analytical model of latitudinal gradients of species richness with an empirical test for marsupials and bats in the New World. *Oikos* **81**: 93–98.

Willig, M.R., D.M. Kaufman & R.D. Stevens (2003). Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis. *Annual Review of Ecology, Evolution and Systematics* **34**: 273–309.

Willis, J.C. (1922). *Age and Area: a study in the geographical distribution and spread of species*. Cambridge University Press, Cambridge, U.K.

Wilson, M.V. & A. Shmida (1984). Measuring beta diversity with presence-absence data. *Journal of Ecology* **72**: 1055–1064.

Wishart, D. (1969). An algorithm for hierarchical classifications. *Biometrics* **25**: 165–170.

Wright, S.J. (1981). Intra-archipelago vertebrate distributions: the slope of the species-area relation. *American Naturalist* **118**: 726–748.

Wright, S.D., R.D. Gray & R.C. Gardner (2003). Energy and the rate of evolution: Inferences from plant rDNA substitution rates in the western Pacific. *Evolution* **57(12)**: 2893–2898.

Wu, Y.Q., K. Ianakiev & V. Govindaraju (2002). Improved *k*-nearest neighbour classification. *Pattern Recognition* **35**: 2311–2318.

Zapata, F.A., K.J. Gaston & S.L. Chown (2003). Mid-domain models of species richness gradients: assumptions, methods and evidence. *Journal of Animal Ecology* **72(4)**: 677–690.