New Phytologist

WILEY

Determination of the Number of Zones in a Biostratigraphical Sequence

Author(s): K. D. Bennett

Source: *The New Phytologist*, Jan., 1996, Vol. 132, No. 1 (Jan., 1996), pp. 155-170

Published by: Wiley on behalf of the New Phytologist Trust

Stable URL: https://www.jstor.org/stable/2558863

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

New Phytologist Trust and *Wiley* are collaborating with JSTOR to digitize, preserve and extend access to *The New Phytologist*

JSTOR

This content downloaded from 154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

New Phytol. (1996), **132**, 155–170

Determination of the number of zones in a biostratigraphical sequence

By K. D. BENNETT

Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

(*Received 10 November 1994; accepted 15 September 1995*)

SUMMARY

Current methods for numerical zonation of biostratigraphic sequences neither assess the reliability of zones identified nor provide any means of determining the number of zones that are 'significant' (based on structure in the data set, rather than the stochastic element). These problems can be overcome by using the broken-stick model to assess the significance of zones against a model of random distribution of zones within a sequence. The broken-stick model is described and its application demonstrated on actual data sets. Additionally, simulation modelling is used to assess the uncertainty of the location of individual zone markers, given the errors of the original data. These widely-applicable methods make it possible to identify zones rigorously and consistently. The potential in biostratigraphy and correlation includes the recognition, correlation and subdivision of chronostratigraphic units in long Quaternary sequences.

Key words: Zonation, biostratigraphy, correlation, pollen analysis.

INTRODUCTION

Sequences of stratigraphical data are cumbersome to describe and interpret without some reduction of the data set to manageable units. This reduction, normally termed 'zonation', might be subjective or objective, might be derived solely from information within the sequence or might incorporate information from other sequences. In the case of pollen data, the aims of zonation, in addition to ease of description, are typically to identify zones of uniform pollen and spore content (pollen assemblage zones), which can then be compared with other such zones from other sites by means of a time-scale (Hedberg, 1976). In Quaternary palaeoecology, this approach was pioneered by Cushing (1967), advocated by West (1970), and formed the background to the development of numerical zonation schemes by Gordon & Birks (1972). A key difficulty, however, is the determination of the number of zones that can be reasonably recognized in a sequence. Birks (1986) suggested that binary divisive analyses should proceed 'until little further reduction in variance occurs', leaving the judgement entirely to the analyst.

Most numerical zonation has been concerned with late glacial and Holocene sequences in which some objective criterion for determining the number of zones is desirable but probably not critical. However, when a Quaternary sediment sequence is much longer, the assemblage zones identified may have some relation to formal chronostratigraphic units of the Quaternary (Eemian, Holsteinian, etc.), and then some objective criterion for determining the number of pollen zones becomes necessary for description of individual sequences, for correlation between sequences and for comparison with established chronostratigraphical schemes. Gordon & Birks (1972) were mainly concerned with establishing the viability of using numerical methods and with contrasting results directly with existing zonation schemes derived by visual inspection. They suggested that 'the recognition and delimitation of hierarchical units within a biostratigraphic sequence must rest, however, with the investigator' (Gordon & Birks, 1972). Grimm (1987) notes that 'the hierarchical nature of the cluster analysis leads to qualitative decisions concerning criteria for zone definition'.

The aims of this paper are to address two aspects of numerical zonation; firstly, to show how randomized datasets and the broken-stick model (MacArthur, 1957) can be used to determine the number of zones in a sequence more objectively and secondly to use simulation modelling to investigate the uncertainty associated with the location of zone markers, in order to establish some measure of confidence in these. The effects of varying aspects of input data sets (number of samples, number of taxa) and of zonation method and data transformation on

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

the number of zones determined numerically are also examined. The methods and approaches are generally applicable and are not confined to the data sets or numerical techniques chosen for analysis. Four different data sets and five zonation methods are used to explore the factors that control output from numerical zonation analyses and the reliance that can be placed on this output.

METHODS

Choice of datasets and statistical approaches

Four Quaternary percentage pollen and spore datasets were analysed (Table 1), chosen to cover a variety of different patterns of vegetation change within the Holocene (two sites) and over multiple glacial–interglacial oscillations (two sites). Details of the zonation schemes developed by the original authors can be found in the references cited. Initially, for each site, all terrestrial pollen and spore types of vascular plants exceeding a threshold of 5 % at any one level were included in a reduced data set, following the recommendation of Birks & Berglund (1979) and Birks (1986), on the grounds that types with lower proportions have little influence on numerical zonations. Proportions were then recalculated to the sum of the included types. The data sets were not transformed before analysis. The effects of varying values for the threshold and of using transformations are discussed below. In this section, methods used for zonation are described, then the methods used to explore zonation output and assess, firstly, the number of zones that should be recognized, secondly, the uncertainty associated with zone marker locations and, thirdly, the effect of the number of samples in the datasets.

Zonation

Zonation of the data sets was carried out divisively using binary and optimal techniques, developed by Gordon & Birks (1972), Birks (1974), and Birks & Gordon (1985), and by agglomeration using constrained cluster analysis (CONISS), developed by Grimm (1987).

The binary approach splits the data set into successively smaller groups by placing divisions within existing zones. The technique first determines the best location for a marker to divide the data set into two. That marker is fixed and the technique then looks for the best location for a marker to divide one of the existing zones itself into two and so on. Results for any given number of zones are thus an extension of results for all results with fewer zones and can be viewed hierarchically.

The optimal approach starts afresh for each successive number of zones, determining the best place for $n-1$ markers for a division into n zones. There will not necessarily be any correspondence between results for division into different numbers of zones, and there is not necessarily any hierarchy of zones. Optimal splitting can be expensive in computer time but, in principal, it is more satisfactory than binary splitting.

For both binary and optimal splitting, the 'best' location for a marker is the location that results in the greatest reduction in variance over the data set as a whole. Reduction in variance may be determined by minimizing either sum-of-squares or information content.

CONISS is based on cluster analysis, with the constraint that clusters are formed by hierarchical agglomeration of stratigraphically-adjacent samples. Grimm (1987) argues that such hierarchical agglomeration is better suited to zonation than divisive methods because the clusters are built up locally, whereas divisive techniques depend on the entire sequence, so the positions of splits might change if the sequence is truncated. However, the results of hierarchical agglomerative techniques can be difficult to interpret during transitional phases (Gordon & Birks, 1972; Birks & Gordon, 1985).

Table 1. *Datasets used*

Site	Dallican Water	Hockham Mere	Ioannina	Valle di Castiglione
Region	Shetland	Norfolk, UK	Epirus, Greece	Roma, Italy
Latitude	60° 23′ 17″ N	52° 30′ 23″ N	39° 40′ N	41° 53′ 30″ N
Longitude	1° 05′ 47″ W	0° 50′ 55″ E	20° 51′ E	12° 45′ 35″ E
Age range*	0–10000	1600–12600	0–> 423000	0–270000
Samples	80	163	217	326
Mean pollen sum	570	996	411	405
Published zones	6	8	44	33
Pollen taxa†	17	14	20	28
Pollen diagram‡	Fig. 1	Fig. 2	Fig. 3	Fig. 4

* Radiocarbon years for Dallican Water and Hockham Mere; calendar years, by correlation with marine cores, for Ioannina and Valle di Castiglione (all years B.P.).

† Number reaching a threshold of 5 % in one sample within the sequence.

‡ In this paper.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

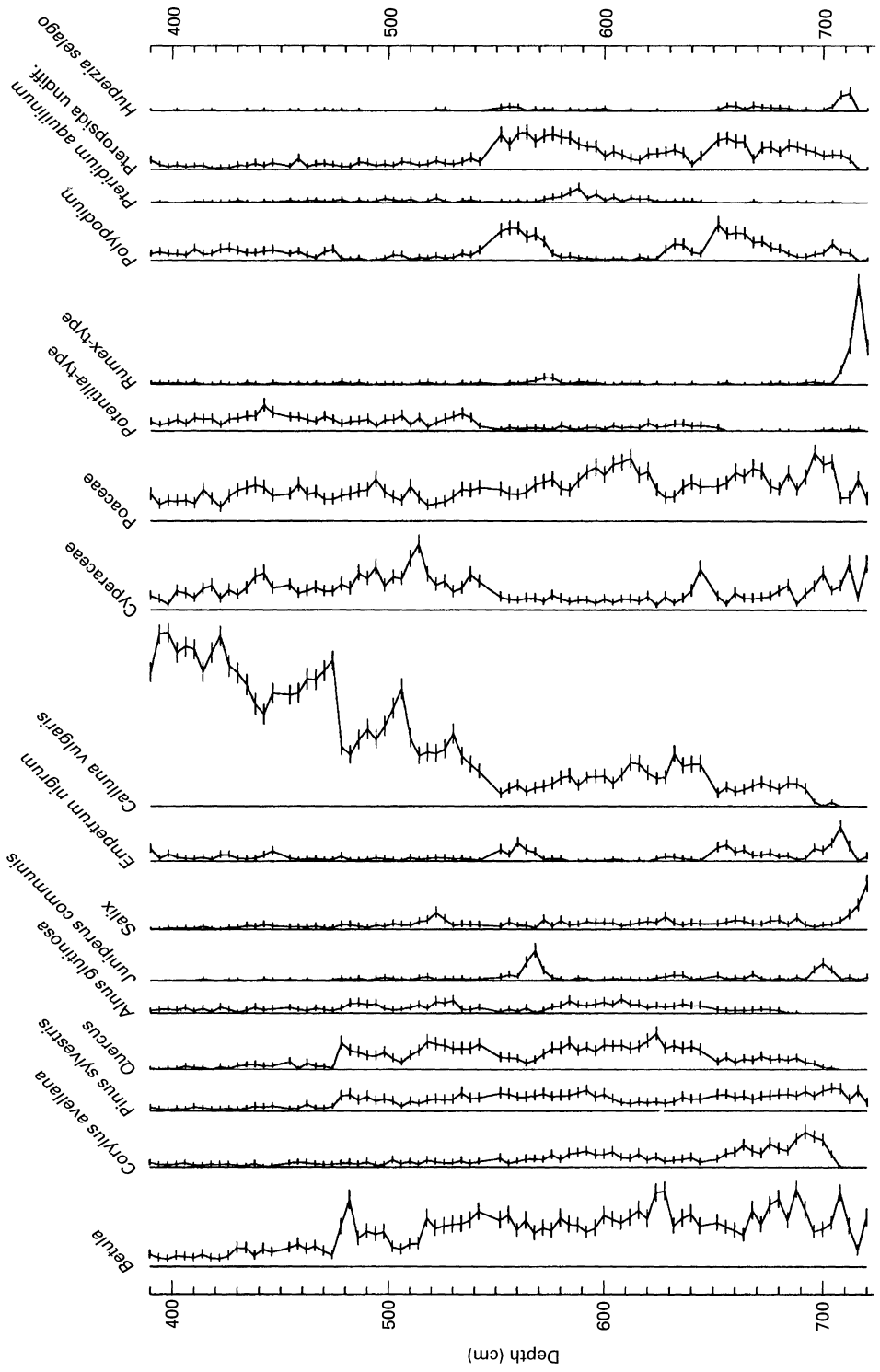

Figure 1. Diagram of selected pollen and spores from the Dallican Water dataset (Bennett *et al.*, 1992), plotted against depth, with 95 % confidence intervals (Maher, 1972). The types selected are those used in this paper, presented as a proportion of total terrestrial pollen and spores.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

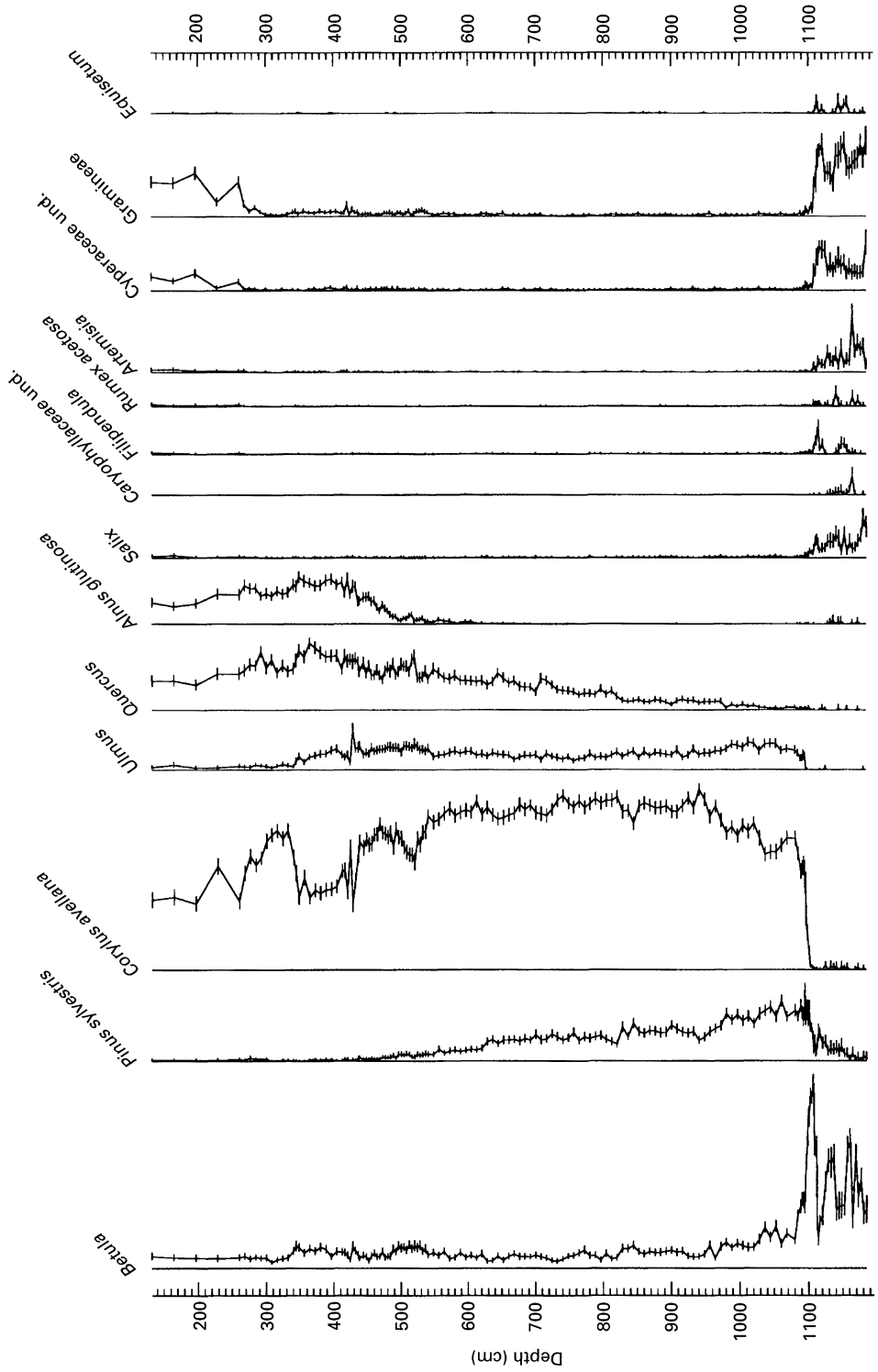

Figure 2. Diagram of selected pollen and spores from the Hockham Mere dataset (Bennett, 1983), plotted against depth, with 95% confidence intervals (Maher, 1972). The types selected are those used in this paper, presented as a proportion of total terrestrial pollen and spores.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

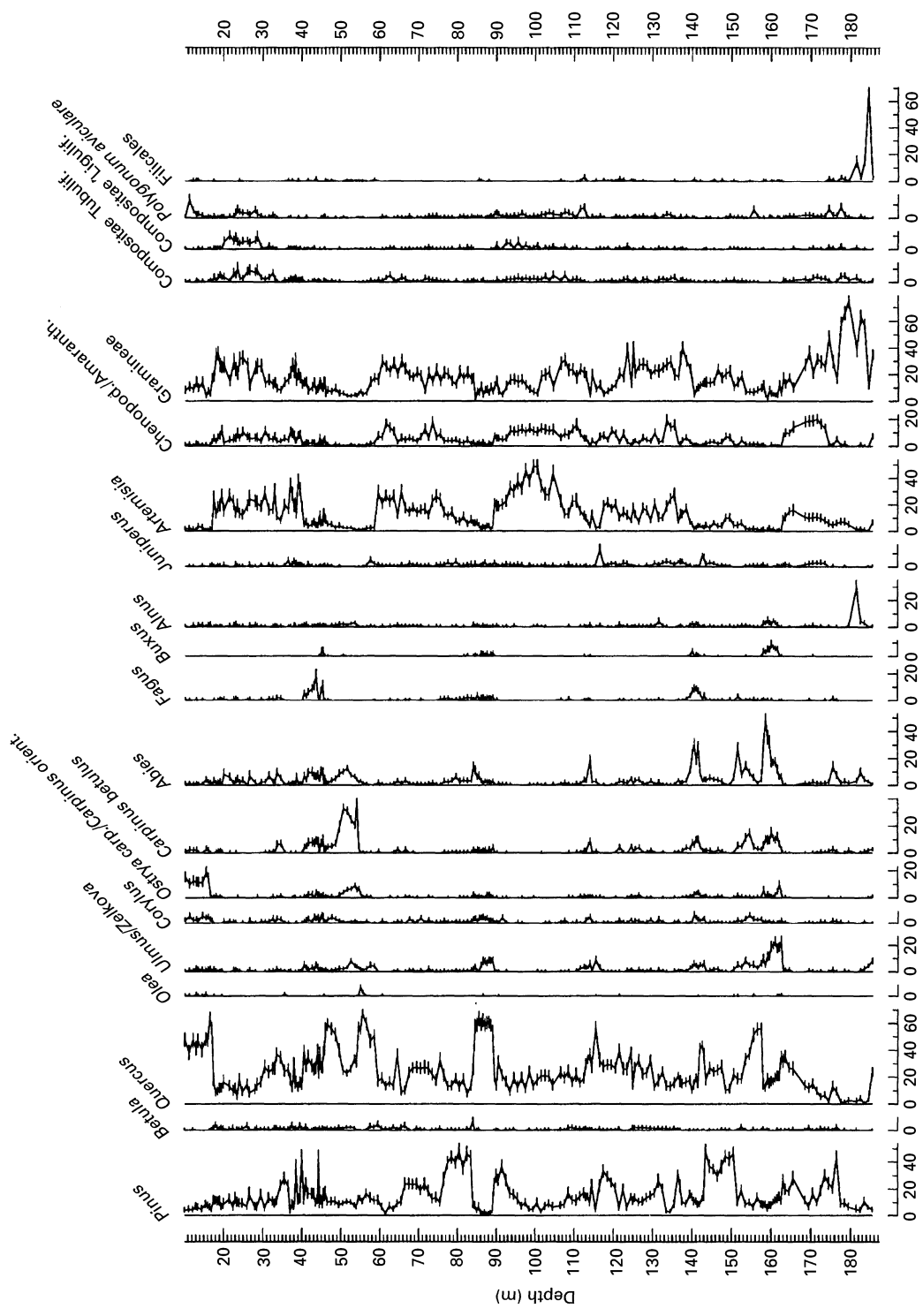

Figure 3. Diagram of selected pollen and spores from the Ioannina dataset (Tzedakis, 1991, 1993, 1994*b*), plotted against depth, with 95% confidence intervals (Maher, 1972). The types selected are those used in this paper, presented as a proportion of total terrestrial pollen and spores.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

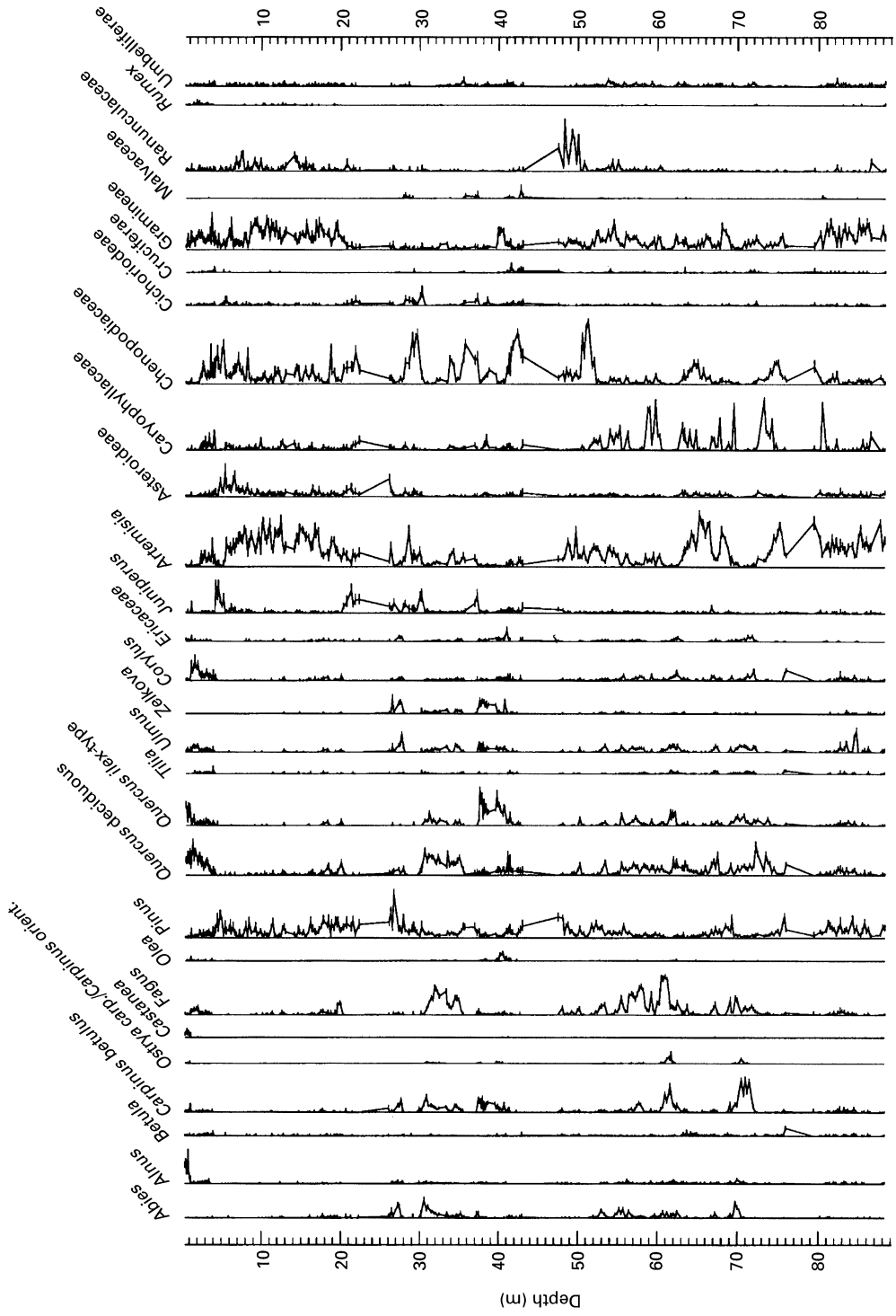

Figure 4. Diagram of selected pollen and spores from the Valle di Castiglione dataset (Follieri, Magri & Sadori, 1988, 1989), plotted against depth, with 95% confidence intervals (Maher, 1972). The types selected are those used in this paper, presented as a proportion of total terrestrial pollen and spores.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

Determining the number of zones

Two methods, randomized data sets and the broken-stick model, are used to help determine how many zones can be recognized numerically in a sequence.

Randomized data sets. If data sets consisted of samples that were drawn from the same, perfectly uniform, body of sediment, their differences would be due to stochastic (random) variation. One way of assessing the success of a zonation analysis is to examine the change in residual variance as new zones are established and to compare this with the way that the residual variance would change if samples in the data set were distributed along the sequence at random. A successful zone is then one which produces a greater fall in total variance than would be obtained from the randomized data set. Randomized data sets are generated by shuffling the samples of the existing data set. Two random integers (i,j) are obtained, and the ith and jth samples in the data set swapped. This was repeated ($10 \times$ number of samples) to achieve a thorough shuffling.

This approach is similar to that of Legendre, Dallot & Legendre (1985), who clustered samples along a chronological sequence until the samples within each cluster behaved as if they resulted from random sampling. The aim in both cases is to separate variance that results from structure in the data set from variance that results from stochastic processes.

Broken-stick model. Zonation analyses, in their various forms, work by partitioning the total variance into components, each related to the allocation of samples into zones. The total variance can be considered as a stick of unit length, with $n-1$ markers positioned on it at random. The lengths of the n resulting segments are the proportion of total variance that would be owing to each level of zonation if the sequence consisted of samples with no stratigraphic structure. Thus, if the reduction in variance for a particular zone exceeds the proportion expected from this model, the zone concerned accounts for more variance than would have been expected if the data set consisted of samples arranged at random and consequently may be considered 'significant'. The idea is that each successive 'split' accounts for part of the total variance in the data set and this part is tested against the portion expected from the model. The formula for calculating values is:

$$Pr = \frac{1}{n} \sum_{i=k}^{n} \frac{1}{i},$$

where Pr is the expected proportion for the kth component out of n. It can be readily obtained on a calculator (for small n) or a spreadsheet, and results compared with the appropriate measure of variance from the output of a numerical zonation method. A similar application of this model (termed 'broken-stick' by MacArthur (1957)) has been adopted to interpret the eigenvalues of principal components analysis (Legendre & Legendre, 1983; Jackson, 1993).

This procedure is readily applied to binary divisive analyses. Each successive split accounts for a portion of residual variation, until the data set is split completely, and residual variance is nil. However, there is no hierarchy of successive splits with optimal divisive procedures, since each level of zonation is begun anew. If an optimal splitting procedure into n zones ($n-1$ divisions) happened to be the same as a binary split into the same number of zones, then there would be the same residual variance. An optimal splitting will always result in a drop in variance that is at least as great as would be obtained from binary splitting. Therefore, if the variance reduction after an optimal split into n zones is compared with the broken-stick model for $n-1$ divisions, it is less likely to exceed the broken-stick proportions than the binary split into the same number of zones. Comparison of an optimal split result with a broken-stick model can therefore help recognize levels of splitting that are significant (those that have variance reductions greater than those expected from the broken-stick model) but cannot show that a level of splitting is not significant.

CONISS is an agglomerative procedure. Samples are combined into clusters, increasing a measure of variance termed 'dispersion' (Grimm, 1987). Each increase corresponds to the variance associated with each cluster (potentially a zone). The analysis thus generates a total dispersion (after all clusters are combined), and values associated with each stage which sum to give the total dispersion. This is all we need to apply the broken-stick model: the increase of dispersion associated with the final stage as a proportion of the total dispersion is equivalent to the first break of the stick, the increase due to the penultimate stage is equivalent to the second break, and so on.

Uncertainty of zone marker locations

It is not possible to obtain directly confidence statistics for results of numerical zonation methods. This is partly because of the complexity of the calculations and partly because of the complexity of the output format. However, it is possible to get some measure of the repeatability of the output variables (zone markers) by repeated sampling from the distributions of input variables (proportional pollen data), and repeated calculation of the zonation until there is enough information to describe the reliability of the zonation with a desired precision. This is an application of the 'Monte Carlo' method, known as distribution sampling (Kleijnen, 1974).

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

The proportions of individual pollen types in a sample follow a binomial distribution (Mosimann, 1965; Maher, 1972). When an analyst sees a grain, it is identified as, for example, either *Betula* or not *Betula*. A binomial distribution is completely characterized by knowing the number of observations made (the pollen sum) and the proportion of the pollen type of immediate interest. On completion of a pollen count, one set of observations has been obtained but repeating the count would not produce identical results. Mosimann (1965) and Maher (1972) show how to obtain confidence intervals for a particular pollen count. Here, however, the purpose is slightly different because the complexity of the numerical methods involved in zonation does not allow propagation of confidence intervals through the analysis. The solution is to simulate the data set by using random numbers to draw from the observed distributions of the pollen and spores in each sample, and thereby obtain simulated 'data sets' each as likely to have been observed as the one that actually was observed but different from it. The full procedure is:

(1) For each sample, obtain the pollen sum, and the proportion of each pollen and spore type of interest within that sum (i.e. all taxa with values of $> 5\%$ somewhere among the samples in the whole data set). The original count can, of course, be obtained from the proportion and the sum.
(2) For each sample and for each taxon, draw a simulated 'count' from a binomial distribution with the same proportion and same total number of observations, using a random number generator.
(3) For each sample, sum the simulated 'counts' and derive a new set of proportions. The simulated sum will not normally be the same as the observed sum.
(4) Run the zonation analysis with the simulated data set and save the results.
(5) Repeat as many times as desired, accumulating results.

Such an approach has certain advantages and disadvantages relative to analytical calculation of confidence intervals. The solution is not exact but bears a probabilistic relationship to the (unknown) exact solution in relation to the number of times the sampling was repeaed. Random numbers are needed to draw from the distribution of the input variables, so the solution will usually be slightly different with each run on the same data. The need for repeated sampling makes the technique slow relative to an analytical solution. On the other hand, a solution will be obtainable with any data, through any zonation method. Given the crudeness of the basic data (e.g. single estimates of the proportion of a particular pollen type in a particular sample), it might be unwise to rely on spuriously exact statistical solutions, in any case. A comparable approach is used by Bennett (1994*b*) to develop confidence intervals for pollen and for rates of accumulation of sediment.

Selection of samples

The effect of varying the number of samples on output was achieved by either omitting every nth sample (to remove fewer than half the samples) or leaving in only every nth sample (to remove more than half the samples). The value of n was varied to give a set of data sets with a reasonable spread of sample numbers. For a given value of n (where $n \neq 1$), there is more than one possible set of subsamples but only one was used here (starting the count of n from the first sample in the data set) and this accounts for some of the variation in the results. As more samples are included within a data set, there is a greater chance of minor taxa exceeding 5% (the threshold value for inclusion) and thus being included in the analysis. Thus, removing samples might reduce the number of taxa with values exceeding 5%, depending on which samples were removed. Such taxa could have been left in but have been omitted as a better model of how a data set with fewer samples would appear to an analyst who was gradually increasing the number of samples.

Implementation

All data-handling and analysis were carried out within the ANSI C program 'psimpoll' (Bennett, 1994*a*). The functions that carry out divisive zonation were written from the algorithms in Birks & Gordon (1985), using the dynamic programming algorithm for optimal splitting. The function that carries out CONISS was written from the original FORTRAN code in Grimm (1987). The functions were tested using the example data sets in Birks & Gordon (1985) and Grimm (1987). The program was run on IBM-compatible personal computers where possible and UNIX mainframe computers for the more computer-intensive aspects of this work.

Simulation results were obtained from the functions 'bnldev' (to obtain random deviates drawn from a binomial distribution), and 'ran1' (the source of uniform random deviates) (Press *et al.*, 1992), both implemented within 'psimpoll'.

Random integers for shuffling datasets are also generated via 'ran1'.

RESULTS

Results to be presented

The full range of possible results from the combinations of methods and data used is too large to present here and what follows is the presentation of a selection to show the nature of the results obtained.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

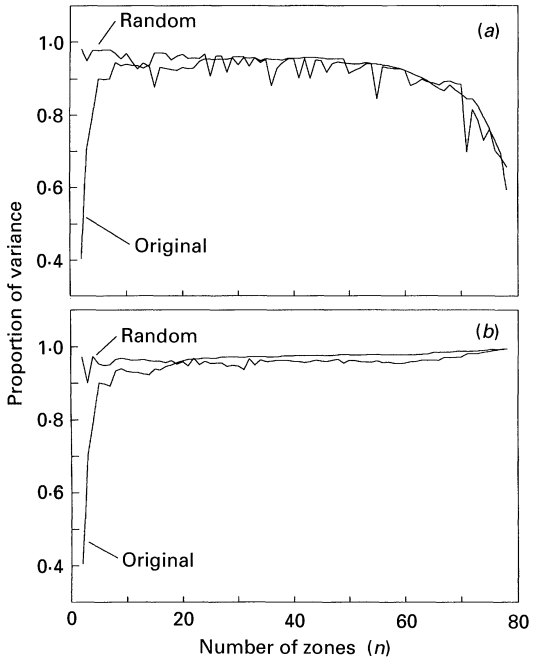

Figure 5. Variance accounted for the nth zone as a proportion of the residual variance after $n-1$ zones, comparing the original data-set with the same data set after randomization of the samples. Zonation method: (*a*) binary divisive and (*b*) optimal divisive, both using the sum-of-squares statistic. Data set: Dallican Water.

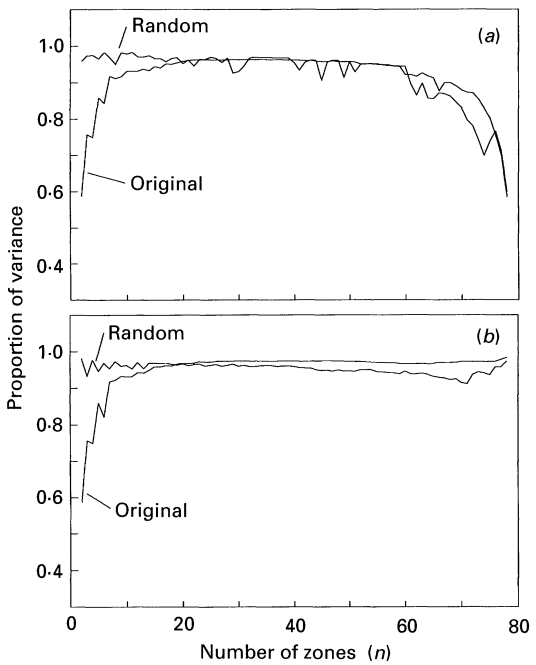

Figure 6. Variance accounted for the nth zone as a proportion of the residual variance after $n-1$ zones, comparing the original data set with the same data set after randomization of the samples. Zonation method: (*a*) binary divisive and (*b*) optimal divisive, both using the information content statistic. Data set: Dallican Water.

Information from randomized data sets

Zonation of randomized data sets and comparison with results from actual data-sets helps determine the number of zones that can be reliably distinguished. The most useful means of comparison are plots of the number of zones (n) against the residual variation after n zones as a proportion of the residual variance after $n-1$ zones (Figs 5–7). The value of n where the two lines converge gives an indication of the maximum number of reliable zones. With divisive procedures (Figs 5, 6), the variance accounted for by the nth zone of an original data set is initially a low proportion of the variance accounted for by the $(n-1)$th zone but increases rapidly. It is then approximately constant for most splits but becomes low again with binary division only as the number of zones approaches the number of samples. When the data set is randomized, the variance accounted for by the nth zone is a high proportion of the variance accounted for by the $(n-1)$th zone, declining as n becomes large during binary division. The decreasing proportion of variance for high numbers of zones during binary division is probably a result of greater efficiency of that zonation as the sequence becomes completely split, catching up with a slight inefficiency relative to optimal division for low numbers of zones.

A corresponding pattern is found with CONISS (Fig. 7), an agglomerative procedure. As clustering

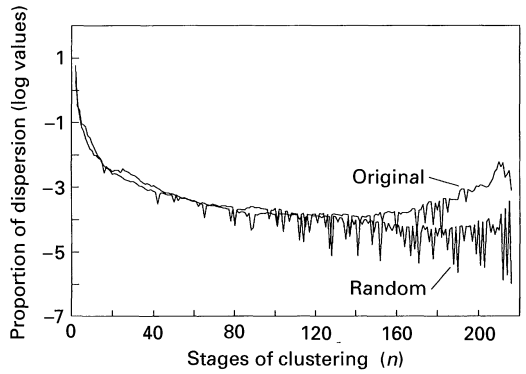

Figure 7. Dispersion increase as a proportion of the running total dispersion as clustering proceeds towards the total number of samples, comparing the original data set with the same data set after randomization of the samples. Zonation methods: CONISS. Data set: Ioannina.

proceeds, the pattern of increase in dispersion of the original data set is similar to that for the randomized data set. As the clustering becomes complete, the patterns diverge as the rate of increase of dispersion for the original data set itself increases, reflecting the combination of clusters with differing contents because of the structure in the data set. The randomized data set, lacking structure, does not show such an increase (Fig. 7).

This suggests that the zonation procedures account for the structure in these data sets within a

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

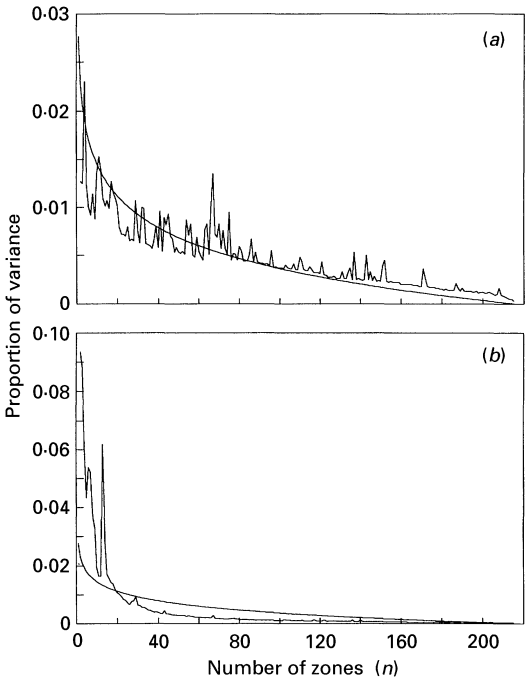

Figure 8. Variance accounted for by the *n*th zone as a proportion of the total variance (fluctuating curve) compared with values from a broken-stick model (smooth curve): (*a*) randomized data set, (*b*) original data set. Zonation method: binary divisive using the information content statistic. Data set: Ioannina.

number of zones that is low relative to the number of samples and the variance accounted for by subsequent divisions is essentially noise. The point at which the curves for original and randomized data sets become similar (divisive procedures) or diverge (agglomerative procedures) gives an indication of the maximum number of zones worth considering. For Dallican Water, this would be about four zones (Fig. 5) or six zones (Fig. 6), depending on the splitting criterion and for Ionnina certainly no more than 50–60 zones (Fig. 7).

This approach demonstrates that much of the variance in these data sets is essentially stochastic and that there is a point beyond which further subdivision of the variance is worthless. It also gives some information on the number of useful zones but greater precision is needed.

Information from the broken-stick model

Zonation of data sets and comparison of the resulting division of total variance with that expected from the broken-stick model provides a more precise recommendation of the number of zones that may be reliably distinguished. The most useful means of comparison are plots of the number of zones (n) against the residual variation after n zones as a proportion of the total variance (Fig. 8). When n is small, if there is structure in the data set, the proportion of variance accounted for in the real dataset should be greater than the proportion predicted by the broken-stick model. The value of n where the two lines cross gives an indication of the maximum number of reliable zones.

For example, using the Ioannina data set (Fig. 8), comparison of the expected distribution of total variance using the broken-stick model with the distribution of total variance from binary division of the randomized dataset shows similar patterns (Fig. 8*a*). Comparison of the original dataset (Fig. 8*b*) with the same broken-stick curve shows that splits for the first 18 zones have decreases in residual variance that exceed those expected from the broken-stick model but splits from all subsequent zones have decreases that are less than the broken-stick model. This implies that the first 18 zones only are significant. They contain series of samples that differ from other series in some way. When these zones are further subdivided, the differences between the new, smaller, zones are the result of stochastic variation between samples. The broken-stick model thus enables a precise, consistently applicable, criterion for determining the number of zones in a pollen sequence. In the case of Ionnina, using the information content method of binary division, the recommendation is 18 zones.

The broken-stick model thus provides a precise recommendation of the number of significant zones. For a data set with structure, division into n zones accounts for a higher proportion of total variance than expected from the model until all the variance that is due to the structure has been accounted for. The number of significant zones is simply the last value for n that accounts for a portion of the variance greater than the model. The method, which can be applied graphically or programmed, is used as the basis for all further analysis in this paper to explore the way in which numerical zonation works in practice and the importance of the various factors that might affect its output.

Use of the broken-stick model

Here, the broken-stick model is used to examine factors that affect the number of zones determined. These include the number of samples and taxa in a data set, the zonation method used and data transformation.

Number of samples. As the number of samples analysed increases, does the number of zones increase continuously or is there some upper limit for a given data set, after which no more zones will be found, no matter how many samples are counted? This can be investigated by selecting samples evenly through existing data sets and determining how many zones would have been detected if analysis had stopped at an earlier stage than it actually did.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

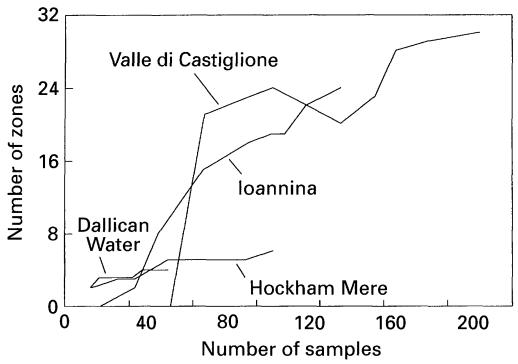

Figure 9. Variation in the number of zones considered significant as a function of the name of samples from four data sets. The significance of zones was determined with the broken-stick model. Zonation method: optimal divisive using the sum-of-squares statistic.

Figure 9 and Table 2 show the results of determining the number of zones from increasing numbers of selected samples in all four data sets. Results are typically clear-cut for the Holocene sequences. For a given number of samples, all zones are significant up to a certain number, but above that none is significant. Results for the long sequences are more complex, with variation above and below the broken-stick values as the number of zones increases. However, across all zonation methods, two patterns are evident: rapidly increasing numbers of zones as the number of samples increases in the two long sequences and a much lower rate of increase, levelling off, for the Holocene sequences. On this evidence, it seems unlikely that counting additional samples would make much difference to the number of zones detected at either Dallican Water or Hockham Mere, but additional zones might be distinguishable with more samples at Ioannina and Valle di Castiglione. The differing patterns suggest that the number of zones detected is not an artefact of the method used to help identify zones but a reflection of the nature of the underlying structure in the data. Additionally, the levelling-off of the increase in the number of zones at Dallican Water and Hockham Mere raises the possibility that these data sets may have an intrinsic structure that will generate a maximum number of identifiable zones. Figure 10 shows what happens when the number of samples is near to or fewer than the number of potentially identifiable zones. Twenty-seven samples at Dallican Water are sufficient to enable clear recognition of three zones but 28 samples at Ioannina produce a zonation pattern that is not distinguishable from the randomized samples in Figure 8. Clearly, the between-sample variation at Ioannina with this

Table 2. *Number of zones as a function of samples and zonation method*

Dallican Water									
Number of samples	80	70	64	60	53	40	27	20	
Binary sum-of-squares	4	4	4	4	4	4	3	2	
Binary information content	6	5	5	5	4	4	4	3	
Optimal sum-of-squares	4	4	4	4	3	3	3	2	
Optimal information content	6	6	4	4	4	4	4	2	
CONISS	4	4	4	4	4	4	3	3	
Hockham Mere									
Number of samples	163	142	130	122	108	81	55	41	21
Binary sum-of-squares	5	5	5	5	5	5	3	3	2
Binary information content	6	6	5	5	5	5	4	3	3
Optimal sum-of-squares	6	5	5	5	5	5	3	3	2
Optimal information content	6	6	5	5	5	5	3	3	3
CONISS	6	5	5	5	5	5	3	3	2
Ioannina									
Number of samples	217	189	173	162	144	108	73	55	28
Binary sum-of-squares	24	24	19	19	20	17	15	10	2
Binary information content	19	18	18	16	10	11	10	4	0
Optimal sum-of-squares	24	22	19	19	18	15	8	2	0
Optimal information content	18	19	17	18	17	15	9	10	0
CONISS	23	21	20	2	0	2	2	0	2
Valle di Castiglione									
Number of samples	326	285	260	244	217	163	109	82	41
Binary sum-of-squares	26	23	24	23	19	16	23	0	0
Binary information content	27	26	22	23	22	7	7	10	0
Optimal sum-of-squares	30	29	28	23	20	24	21	0	0
Optimal information content	26	19	23	22	21	17	13	15	0
CONISS	31	29	28	28	20	18	14	0	0

The broken-stick method (see text) was used to determine the maximum number of reliably recognizable zones in each case. All numbers of zones are recognized as significant up to and including the numbers shown. There might also be other significant zones above this number, but not consistently.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

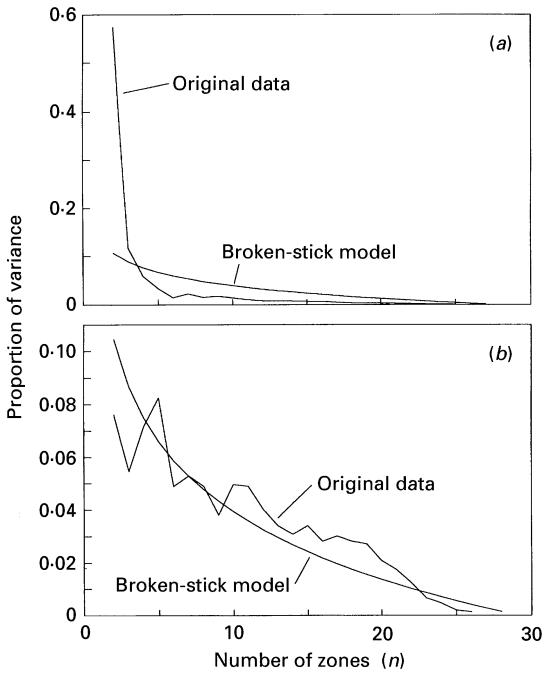

Figure 10. Variance accounted for by the *n*th zone as a proportion of the total variance compared with values from a broken-stick model: (*a*) Data set: Dallican Water, 27 samples selected evenly through the data set; (*b*) Data set: Ioannina, 28 samples selected evenly through the data set. Zonation method: optimal divisive using the sum-of-squares statistic.

Table 3. *Number of zones as a function of threshold value for selection of taxa for inclusion in the data set analysed*

Threshold (%)	0·5	1	2	5	10	20
Number of taxa	48	33	27	14	11	6
Binary sum-of-squares	6	5	5	5	5	5
Binary information content	7	6	6	6	6	6
Optimal sum-of-squares	6	6	6	6	5	5
Optimal information content	7	6	6	6	5	5
CONISS	6	6	6	6	5	5

The broken-stick method (see text) was used to determine the maximum number of reliably recognizable zones in each case. Data set: Hockham Mere.

number of samples is too great for recognition of any structure in the data set [by any zonation method (Table 2)], and this arises because the data set has no simple division into a low number of zones: it has several zones of similar status. Dallican Water, on the other hand, can be divided into a low number of zones (late glacial, early Holocene, late Holocene), with some further refinement also possible.

Number of taxa. How does the number of taxa included in the zonation analysis affect results? Birks & Berglund (1979) and Birks (1986) recommended that data sets for numerical zonation should only contain those taxa with proportions that exceed 5%, and that has been followed in all the preceding analyses. Table 3 shows the outcome of varying this threshold for the Hockham Mere data set. Lowering the threshold includes far more taxa but also increases, slightly, the number of zones recognized as significant. Increasing the threshold to 10% or even 20% changes the location of zone markers towards positions more suited to the few taxa still included. If the aim of zonation is to establish as many zones as possible, a case could be made for using a low threshold and including as many pollen types as possible.

Zonation methods. The effect of different zonation methods varies between data sets (Table 2). For example, divisive methods using sum-of-squares detect more zones at Ioannina but divisive methods using information content detect more zones at Hockham Mere and Dallican Water. CONISS apparently fails to detect zones at Ioannina for fewer than 173 samples. Examination of the output suggests that this is because the stage with three zones is not significant. Clearly, this data set does not have a tripartite structure (Figure 3), and CONISS is unable to make a significant split into three until the number of samples is large.

Data transformation. Gordon & Birks (1972) and Birks & Gordon (1985) did not use any data transformation but Grimm (1987) suggests three possible transformations: (i) standardization of variables to a mean of zero and to unit standard deviation, (ii) normalization of sample vectors to unit length and (iii) square-root transformation. Square-root trnsformation, with frequency data, upweights rare variables to abundant ones and is the one preferred by Grimm (1987) for use with CONISS. Aitchison (1986) has argued that percentage data should be transformed by a centred log-ratio method to compensate for the constraining upper limit of percentage data. Table 4 shows the effect of these four transformations on the number of zones for two contrasting data sets.

For the Dallican Water data set, there is a clear increase in the number of significant zones with square-root and standardization transformations. They increase the importance of scarcer taxa and thus have an effect in the same direction as increasing the number of taxa. However, other factors must come into play for the Valle di Castiglione data set as there is a variable response to different transformations depending on zonation method.

Summary. There is some suggestion from these results that there might be an upper limit to the number of assemblage zones that can be recognized in a sequence and that the number of zones is also a function of zonation method and the parameters

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

Table 4. *Number of zones as a function of type of data transformation*

Data transformation	None	Square-root	Standardized	Normalized	Centred log-ratio
Dallican Water					
Binary sum-of-squares	4	5	7	4	4
Optimal sum-of-squares	4	6	7	4	4
CONISS	4	6	7	4	4
Valle di Castiglione					
Binary sum-of-squares	26	25	30	27	17
Optimal sum-of-squares	30	24	33	26	20
CONISS	31	24	28	27	20

The broken-stick method (see text) was used to determine the maximum number of reliably recognizable zones in each case.

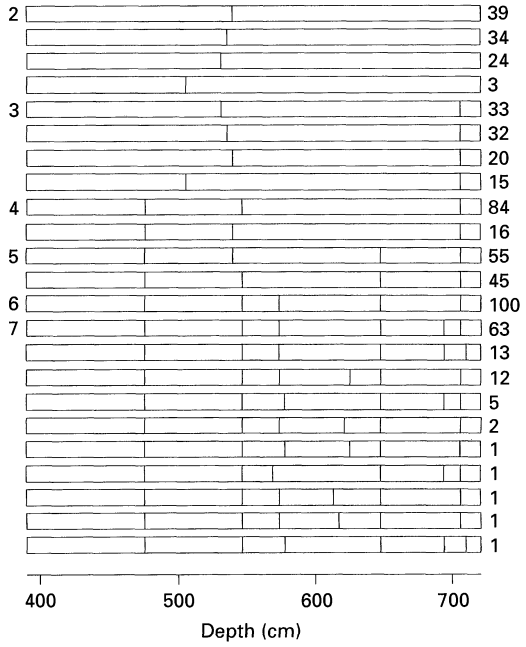

Figure 11. Patterns of zonation resulting from simulation modelling of an actual data set. The figures at the left indicate the numbers of zones, and the figures at the right are the number out of 100 simulations that produced the pattern concerned. Zonation method: optimal divisive using the information content statistic. Data set: Dallican Water.

used to define the data set used. Any statement about the significance or otherwise of a particular zonation scheme needs to consider uncertainty about which method is the most appropriate.

Factors affecting which zones are recognized

All the preceding analysis focuses on the number of zones detected by a numerical zonation scheme. However, the location and extent of zones are usually the main interest and some measure of confidence in those aspects is therefore necessary. When the output of a numerical zonation suggests a division into n zones with markers at certain points in the sequence, what confidence can we place in those markers? The numerical treatment should make the data analysis repeatable but what about the pollen counting? Re-counting pollen data sets to obtain an answer is obviously impracticable but an assessment of its repeatability can be made by simulating the data set many times (see above) and then carrying out zonations of each simulated data set to assess the repeatability of the resulting zonation scheme.

At Dallican Water, zonation by the optimal divisive method, using information content, indicated that a maximum of zones was significant (Table 2). Figure 11 shows the results of simulation modelling for the first seven zones. Four different patterns were found for splitting into each of two and three zones, two each for four zones and five zones, and one only for splitting into six zones. Splitting into seven zones, however, generated 10 different patterns and similar large numbers of patterns are found for greater numbers of splits (not shown). Given the large number of ways that 80 samples could potentially have been split into up to six zones, it is striking that not more patterns were found. It is also significant that the number of patterns begins to increase markedly for division into more than six zones and that there is consistency at six zones, as this is the level of division indicated as the maximum that would be significant by the broken-stick model.

The principle behind the simulation modelling is that any of the simulated data sets could have been obtained in practice and might be obtained by repeat counting of the same pollen samples. It indicates the sensitivity of particular zones to the counting errors in the pollen analysis. For example, the zone marker identified at 476 cm is found consistently in all patterns for division into four or more zones, and a marker at 706 cm is present in all but two patterns. By contrast, markers at around 540–550 cm, and earlier in the Holocene, are much less constant, despite being present in all patterns. Less significance should be placed on the exact location of these markers than those at 476 cm or 706 cm.

It would be expected that higher pollen counts would result in greater repeatability of zonation

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

schemes by simulation, as the confidence intervals for the individual pollen proportions become smaller. This in turn should lead to recognition of more zones, as more structure in the data set is revealed from behind the screen of noise from the variability of the original count.

DISCUSSION

Analysis

The methods and results presented here contribute to understanding of what is happening in a numerical zonation and especially the interaction between the analysis and the original data set. Although numerical analyses themselves cannot distinguish between structure in the data sets and noise introduced by the errors on pollen counts, it is possible to determine where the structure ends and the noise begins. Making this distinction then permits the analyst to assess how many zones can be confidently recognized in the sequence, where the zone markers are, how much confidence to place in each individual zone marker and whether it might be possible to distinguish more zones by counting more samples. Additionally, the analyst will have a zonation scheme which has been more objectively determined than has been possible hitherto and might be usable as a means of comparison between sites.

Applications

Stratigraphy. The approach used here enables a rigorous, consistent zonation scheme to be adopted for any sequence of biostratigraphic data. Development of stratigraphies using numerical zonation has, hitherto, been hindered by the requirement for subjective analysis of the results, particularly in terms of the number of zones that are significant. This hindrance has now been overcome. The approach not only provides recommendation for the placing of markers to separate assemblage zones but also gives a basis for assessing how many zones can be reliably defined. This has some important consequences:

(1) The approach guards against the twin errors of not making the most of the available data (by under-splitting) and of drawing conclusions that are insecurely based (through over-splitting). The latter is probably the commoner error.
(2) When working with sequences that are expected to correlate with a known stratigraphy, it will be possible to establish, with confidence, whether or nor a particular assemblage zone of interest is present or not.
(3) When correlating zones and constructing time–space diagrams of pollen assemblages zones (e.g. Cushing, 1967), the approach enables the number of zones in each sequence to be separately identified and guards against the risk of allowing subjective assessment of which zones are present to dominate the correlation process. A stratigrapher, having identified an assemblage zone in a number of sequences, might observe a similar assemblage in a new sequence and might even have it identified by a numerical zonation procedure. However, if the zone in the new sequence fails comparison with the broken-stick model, it should not be correlated with the others. It might, of course, become significant after further work.
(4) Tzedakis (1994*a*) has advocated a hierarchical approach for the description and organization of biostratigraphical data from long Quaternary sequences, such as Ioannina and Valle di Castiglione. He proposed the use of superzones as approximate equivalents to the chronostratigraphic units of the Quaternary, and formed them by combining assemblage zones, which might themselves have been divided into subzones in the conventional manner. In all the zonations carried out for this paper, the way in which the variance is distributed among zones as the analysis proceeds as a continuum, with no breaks indicative of a hierarchy. On the basis of the analysis of numerical zonation procedures, it is suggested here that the only possible basic unit for biostratigraphy is the assemblage zone, defined as being the smallest significant zone that cannot be divided into smaller significant zones. Superzones which can then be made, as necessary, by combining assemblage zones for ease of description and organization, might be especially useful for distinguishing sequences of zones all characterized by tree pollen types from sequences of zones characterized by other pollen types. The use of subzones should be reserved for subjective units that it is useful to describe or draw attention to but which are not numerically significant subdivisions of the data set. In retrospect, the description of the pollen stratigraphy at Dallican Water should have been through six zones rather than four zones and their subzones (Bennett *et al.* 1992).

Non-numerical zonations. Zones that may be considered significant in the four data sets have been identified. The approach gained helps to establish characteristics of such zones and the way that numerical schemes subdivide data sets which might be applicable in situations where zonation schemes have been established without the assistance of numerical methods.

First, sequences rarely have more than about one significant zone for every 10 samples and the figure might be nearer to one zone for every 20 samples in Holocene sequences. Denser zonation schemes

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

should be viewed with distrust without additional evidence that they are justified.

Second, some of the numerical zonation methods applied to Ioannina and Valle di Castiglione produced zones of one sample, identified as significant. It has been conventional to regard such small zones as unsatisfactory. However, if the broken-stick model is able to recognize them as significant, the best response might be to analyse further samples near to the single sample and attempt to establish the zone on a broader base.

CONCLUSIONS

The broken-stick model of MacArthur (1957) provides an effective means to help an analyst determine the number of zones that can be reliably used from the output of a numerical zonation of pollen stratigraphical data.

Simulation modelling of data sets in various ways can help to understand the broken-stick model and can contribute to understanding of the sensitivity of zonation results to the uncertainties associated with the original pollen analyses.

The number of zones recognized increases as the number of samples in the data set increases but at different rates, depending on the structure in the data set. Sequences with many zones each contributing a similar proportion of the total variance cannot be reliably split into a low number of zones and there may be no satisfactory zonation if the number of samples is too low, relative to the number of zones potentially detectable.

In real data sets, the increase in the number of zones as sample number increases does level off. This arises because the detection of zones depends on the use of the broken-stick model to exclude divisions or agglomerations of essentially random data. The source of 'randomness' is mostly the errors associated with pollen counts. Numerical analyses recognize structure in the data set coarser than this variation but once that has been removed, no more progress can be made without increasing the size of the pollen count. In effect, the size of a pollen count puts a limit on the number of zones that can be recognized, no matter how many samples are analysed.

Biostratigraphic assemblage zones can be defined as the smallest significant units found in a sequence following numerical analysis. This permits the analyst to define 'superzones' (*sensu* Tzedakis, 1994*a*) for significant higher-order groupings and 'subzones' for statistically insignificant units that might be worth drawing attention to.

ACKNOWLEDGEMENTS

I am grateful to Chronis Tzedakis for allowing me to use the Ioannina data set and to Maria Follieri, Donatella Magri and Laura Sadori for allowing me to use the Valle di Castiglione data set. I thank John Birks, Janice Fuller, Susie Lumley, Donatella Magri, Lou Maher, Chronis Tzedakis and Kathy Willis for helpful comments on early versions of the manuscript.

REFERENCES

Aitchison J. 1986. *The statistical analysis of compositional data.* London: Chapman and Hall.

Bennett KD. 1983. Devensian late-glacial and Flandrian vegetational history at Hockham Mere, Norfolk, England. I. Pollen percentages and concentrations. *New Phytologist* **95**: 457–487.

Bennett KD. 1994*a*. 'psimpoll' version 2.23: a C program for analysing pollen data and plotting pollen diagrams. *INQUA Commission for the study of the Holocene : Working group on data-handling methods, Newsletter* **11**: 4–6.

Bennett KD. 1994*b*. Confidence intervals for age estimates and deposition times in late Quaternary sediment sequences. *The Holocene* **4**: 337–348.

Bennett KD, Boreham S, Sharp MJ, Switsur VR. 1992. Holocene history of environment, vegetation and human settlement on Catta Ness, Lunnasting, Shetland. *Journal of Ecology* **80**: 241–273.

Birks HJB. 1974. Numerical zonations of Flandrian pollen data. *New Phytologist* **73**: 351–358.

Birks HJB. 1986. Numerical zonation, comparison and correlation of Quaternary pollen-stratigraphical data. In: Berglund BE, ed. *Handbook of Holocene Palaeoecology and Palaeohydrology*, John Wiley & Sons: Chichester, 743–774.

Birks HJB, Berglund BE. 1979. Holocene pollen stratigraphy of southern Sweden: a reappraisal using numerical methods. *Boreas* **8**: 257–279.

Birks HJB, Gordon AD. 1985. *Numerical methods in Quaternary pollen analysis.* London: Academic Press.

Cushing EJ. 1967. Late-Wisconsin pollen stratigraphy and the glacial sequence in Minnesota. In: Cushing EJ, Wright HE, Jr, eds. *Quaternary Palaeocology*, Yale University Press: New Haven, 59–88.

Follieri M, Magri D, Sadori L. 1988. 250,000-year pollen record from Valle di Castiglione (Roma). *Pollen et Spores* **30**: 329–356.

Follieri M, Magri D, Sadori L. 1989. Pollen stratigraphical synthesis from Valle di Castiglione (Roma). *Quaternary International* **3/4**: 81–84.

Gordon AD, Birks HJB. 1972. Numerical methods in Quaternary palaeoecology. I. Zonation of pollen diagrams. *New Phytologist* **71**: 961–979.

Grimm EC. 1987. CONISS: a FORTRAN 77 program for stratigraphically constrained cluster analysis by the methods of incremental sum of squares. *Computers & Geoscience* **13**: 13–35.

Hedberg HD. 1976. *International Stratigraphic Guide : a guide to stratigraphic classification, terminology, and procedure.* New York: John Wiley and Sons.

Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**: 2204–2214.

Kleijnen JPC. 1974. *Statistical techniques in simulation.* Part I. New York: Dekker.

Legendre L, Legendre P. 1983. *Numerical ecology.* Amsterdam: Elsevier.

Legendre P, Dallot S, Legendre L. 1985. Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. *American Naturalist* **125**: 257–288.

MacArthur RH. 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Science, USA* **43**: 293–295.

Maher LJ, Jr. 1972. Nomograms for computing 0·95 confidence limits of pollen data. *Review of Palaeobotany and Palynology* **13**: 85–93.

Mosimann JE. 1965. Statistical methods for the pollen analyst: multinomial and negative multinomial techniques. In: Kummel

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms

B, Raup D, eds. *Handbook of Paleontological Techniques*, Freeman: San Francisco, 636–673.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical recipes in C: the art of scientific computing*, 2nd edn. Cambridge: Cambridge University Press.

Tzedakis PC. 1991. *Vegetation dynamics in northwest Greece in response to Quaternary climatic cycles*. Ph.D. thesis, University of Cambridge, UK.

Tzedakis PC. 1993. Long-term tree populations in northwest Greece through multiple Quaternary climatic cycles. *Nature* **364**: 437–440.

Tzedakis PC. 1994*a*. Hierarchical biostratigraphical classification of long pollen sequences. *Journal of Quaternary Science* **9**: 257–259.

Tzedakis PC. 1994*b*. Vegetation change through glacial-interglacial cycles: a long pollen sequence perspective. *Philosophical Transactions of the Royal Society of London* **345**: 403–432.

West RG. 1970. Pollen zones in the Pleistocene of Great Britain and their correlation. *New Phytologist* **69**: 1179–1183.

This content downloaded from
154.59.124.102 on Fri, 19 Feb 2021 10:32:06 UTC
All use subject to https://about.jstor.org/terms