COMMUNITY ECOLOGY 3(2): 147-157, 2002
1585-8553 © Akadémiai Kiadó, Budapest

Probabilistic classification and its application to vegetation science

D. W. Goodall

Centre for Ecosystem Management, Edith Cowan University, Joondalup WA 6027, Australia
Fax: 61 (0)8 9400 5509, E-mail: d.goodall@ecu.edu.au

Keywords: Festucion valesiacae, Hypothesis testing, Nardo-Galion, Ordinal data, Phytosociology, Vegetation classification.

Abstract: Probabilistic classification offers various advantages in its application to vegetation studies: it can use data in the form of ordered as well as quantitative values; it can use a range of values for each attribute (species) in each relevé or group of relevés; it can use incomplete data sets; it takes account of all species, rather than only characteristic species; and it enables a null hypothesis of random distribution of species among relevés to be tested.

The procedure is here explained in some detail, and its application is illustrated, first with a classical data set from the Alps, and second with an extract from the extensive Netherlands national data base.

It is shown that the presence or absence of species is often more informative about the relationships between relevés than the quantities in which they are present. The results do not support the concept of discrete and uniform vegetation units, but rather of vegetation composition varying around centres of concentration.

Introduction

The procedure originally described by Braun-Blanquet (1928) for reporting the composition of vegetation samples ("relevés") and classifying them has come to be known as "phytosociology", and has been widely used throughout the world (Moore 1962, Mueller-Dombois and Ellenberg 1974) The classification methods used in this procedure have, however, been largely intuitive, and there is scope for introducing a more objective approach – particularly one involving the concept of a null hypothesis and the statistical significance of a departure from it. It is shown here that probabilistic classification can offer such an approach. An earlier version of the procedure was applied to scree communities in the Italian Appennines (Goodall and Feoli 1988); but probabilistic classification has been much improved in the intervening fourteen years, and it seems worth while to revisit the question.

Classification may sometimes be intended purely to serve pragmatic purposes, to facilitate discourse, or to enable particular predictions to be made about the objects classified. In the latter case, the classsification procedure is validated by the effectiveness with which this particular end is served.

Most classification procedures, however, including phytosociology, are intended to reflect underlying relations in the real world, and not simply to serve a practical purpose. The very process of classification, in this sense, implies an underlying assumption that the objects to be classified (relevés, in this case) may be regarded as falling ultimately into discrete groupings differentiated by the joint occurrence of certain attributes – in the present instance, quantities of different plant species.

The advantages that may be claimed for probabilistic classification as a tool for the analysis of vegetation data, in comparison with those more commonly applied to them, are:

- that data can be used in the original form of an ordinal scale, without conversion to percentages;
- that incomplete data sets can be used;
- that multiple values for the quantity of a species present in a relevé or group of relevés may be used as they stand, in the proportions in which they occur, rather than reducing them to a mean;
- that the procedure takes account of all species recorded rather than concentrating attention on a few "characteristic" species; and

- that probabilistic classification is based on a null hypothesis of uniformity, and provides a statistical test of this hypothesis.

Probabilistic classification

The concept of probabilistic classification was introduced in the 1960's (Goodall 1964, 1966, 1966a, 1968), but has in the past decade been much improved and extended (Goodall 1993, 1994). It may be unfamiliar to many readers, so the principles involved will first be outlined.

Probabilistic classification can, like many classification procedures, be applied to any objects ("operational taxonomic units" — OTU's) such as relevés which can be described in terms of a number of logically independent attributes such as the quantity of different species.

As a first step, for any particular attribute, all possible pairs of values are arranged in order of likeness. The ordering principle varies according to the type of attribute. For a quantitative attribute, the ordering principle is simply the unsigned difference between the two values; for an unordered qualitative attribute differing pairs are treated as uniformly unlike, but agreeing pairs are considered more alike if the value in which they agree is an uncommon one. For ordered variables, correspondingly, the likeness between any two values is the complement of the probability that any random pair of values would be equally close or closer in the ordered series. For spatial attributes, the two-dimensional continuum is divided into a number of cells, and the likeness between different cells depends on the distance between their centroids. For each possible pairing of values, the cumulative probability of a pair of values being as or more alike is then calculated. If more than one value has been recorded for an attribute in an OTU, the proportions of these values within the OTU are used as weights in calculating the likeness between different value pairs.

For any given pair of OTU's, the resemblance in respect of a given attribute may be expressed by the sum of the likenesses between each pair of values recorded, weighted by their proportions in these OTU's. These values for resemblance may then be used to order all possible pairs of OTU's in respect of this attribute. The proportion of OTU pairs resembling one another as closely or more so is then an inverse measure of the similarity in respect of this attribute – a *similarity index*.

For each attribute, a *norm* within a set of OTU's may be defined as having the overall distribution of values for the OTUs in that set, each value with its average weight.

Suppose now that one wants to consider the extent to which a particular OTU deviates from the norm of the whole set considered. The procedure is the same as for the similarity matrix, but instead the comparisons are between each OTU and the set norm. Again, the weighted sum of likenesses between each pair of values is calculated, these values are ordered, and the proportion of OTU's with resemblance to the norm greater than that of the OTU under consideration is an inverse measure of its deviation from the norm for this attribute – a *deviant index*.

Finally, one may have distinguished a subset of OTU's, and be interested in the relation of other OTU's to this subset. One calculates the norm of this subset, and then (excluding this subset from the calculations) finds the resemblance of every other OTU to this norm. The OTU's of the complementary subset are then ordered in respect of this resemblance, and the proportion with equal or greater affinity to the subset constitutes the contribution of this attribute to an inverse *affinity index.*

Once these contributions have been calculated for each attribute, the question arises of how the attribute contributions (each an estimate of probability) can be combined. For a small number of discrete probabilities, exact combination is quite practicable. But, as the number of probabilities to be combined increases, and as the number of possible values for each increases, the computing load increases exponentially, and soon becomes quite impracticable. Luckily, an approximation is available. Fisher (1934) pointed out that, for independent continuous probabilities, the combined probability could be obtained by calculating

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

where n independent probabilities p_i are to be combined. X^2 is then distributed as χ^2 with $2n$ degrees of freedom. Lancaster subsequently (1949) provided an approximation for discontinuous probabilities, such as data for attribute probabilities would always yield:

$$P = 2\sum_{i}\left[1 - \frac{p_i \ln p_i - p'_i \ln p'_i}{p_i - p'_i}\right]$$

where p'_i is the probability next smaller than p_i in the series of discrete cumulative probabilities from which p_i is taken.

Since the probability values calculated are often very small, it is convenient to convert them into logarithms. Thus, the probability P above would be converted to:

$-\log_{10} P$. The fact that this transformed figure increases with similarity, deviation or affinity, rather than being inversely related as is the case with P, makes comprehension easier.

Application to classification of vegetation

Turning now to the application of this method to phytosociological data, the OTU's here are the relevés, the attributes are most commonly the quantities of different species, usually expressed by a number on an arbitrary scale depending on the proportion of ground covered by the species. The scale is often converted to a sequence of percentage figures, in the middle of the range implied by the recorded scale number. A location is recorded, and also sometimes some environmental data such as aspect or soil type.

The three tools described – the deviant index, the similarity index and the affinity index – can be used to illuminate the relations among the relevés, in whatever form these data are recorded, as will be shown. It should be noted that the indices can be used for any types of attribute, in any mixture, all being reduced to the common terms of probability. An attribute may be recorded for only some of the OTU's, and indeterminate or unknown for others (for instance, if bryophytes have been recorded only for some of the relevés, that information can be included where it exists, while those variables are indeterminate elsewhere). Several different values may be recorded for a given species in a given relevé – if, for instance, the relevé includes subsamples, the proportions of the different values are then taken into account.

In any attempt to apply objective methods to phytoscociological data, it must be remembered that these data themselves include a strong subjective element. Selection of the sites for relevés is generally on the basis of uniformity (which can in principle be assessed objectively), but also often on conformity with a preconceived notion of the community sampled. Thus, an objective analysis of published data may well be objective as far as the analysis goes, but cannot remedy (or even determine the influence of) the subjective element in the collection of the data. At least, though, one may hope that conclusions drawn are objectively justified by the data presented, and introduce an element of hypothesis-testing which is often missing from vegetation analysis.

An advantage of the probabilistic approach to phytosociology is that it can make use of the data in their original form, as a value on an ordinal scale, rather than first converting them to percentages, which fail to reflect the great differences in precision in different parts of the scale. If the ecologist is estimating percent. cover by eye, he or she is much more certain about a difference between 3 and 5%, let us say, than between 73 and 75%. This difference in precision is reflected in the ordinal scales which the majority of phytosociologists use, but not in their percentage transforms.

The approach adopted here starts from the "null hypothesis" that the different relevés in the collection are samples from the same vegetation type, in which the values recorded for each species are randomly distributed among the relevés, and the occurrence and quantity of different species are not correlated.

Some applications of the principles of probabilistic classification to phytosociological data are offered here. The applications are intended merely to show the potentiality of the method, rather than to produce results of value in their own right. It is not suggested that this new approach should replace other more traditional approaches; but at least it offers an objective method of analysis which makes it possible to draw clear conclusions, based on an unambiguous model.

A classical set of relevés

The first set of phytosocoiological data to be analysed was collected and published by the father of the subject, Josias Braun-Blanquet (1936). It consists of some dry grassland data from the eastern Alps, which he assembled in the alliance Festucion valesiacae. There were 22 relevés; relevés 1 to 9 were placed in the association Festuceto-Caricetum supinae, relevés 10 to 13 in the Stipeto-Seselietum, and relevés 14 to 22 in the Festuceto-Poetum xerophilae. A total of 90 species were tabulated. The records were in the form of ordinal values (up to a code of 4).

A similarity matrix showed that 42 of the 231 probabilities had values (after logarithmic transformation) greater than 1.0 (against the 23 expected in random data), and of these 20 were greater than 2.0 (Table 1), the largest being 6.3 (a probability of 0.0000002); these figures justified rejection of the null hypothesis. One may note that all but one of the high similarities were between relevés attributed to the same association. The highest similarity is between two relevés of the Stipeto-Seselietum, numbers 11 and 13. These two, however, constitute the core of a small group of four relevés with high similarities, the six pair-wise comparisons within this subset of four relevés all showing similarities greater than 2.2. The chance that a set of six values from a uniform distribution between 0 and 1 should all be less than .01 is 10^{-12}; the number of ways in which a subset of four can be selected

Table 1. Similarities between relevés in the Festucion valesiacae (logarithmic scale) (only values exceeding 2.0 are listed). The data are from Braun-Blanquet (1936), where one of the tables was devoted to each Association: Table 1 Festuceto-Caricetum supinae, with relevés 1 to 9. Table 2 Stipeto-Seselietum, with relevés 10 to 13. Table 3 Festuceto-Poetum xerophilae, with relevés 14 to 22.

Relevé A	Relevé B	Similarity
1	2	2.010
	3	3.516
2	3	2.032
	4	2.586
	6	2.289
3	6	2.219
	7	3.098
5	9	4.012
9	8	2.809
10	11	5.147
	12	3.340
	13	2.259
11	12	4.496
	13	6.373
12	13	2.342
14	16	2.909
15	16	2.509
16	21	3.172
17	18	2.345
17	22	2.427

from a set of 22 is 7315, so the probability that any such set of four might appear in these data by chance is less than 10^{-8}. These four relevés are exactly those which Braun-Blanquet grouped in the association Stipeto-Siselietum.

The programs used here make it possible to see how the overall probabilities are built up. If the six similarity indices among these four relevés are examined further, it is found that the most consistent contributions are from the species *Carex nitida* and *Stachys recta* – both characteristic species for the order Brometalia rather than for the association. Very consistent contributors also, however, are *Scabiosa agrestis* and *Scorzonera austriaca*, with *Oxytropis pilosa*, *Onosma tridentinum*, *Erysimum canescens* and *Seseli varium* coming not far behind; these are the species which Braun Blanquet named as characteristic of the Stipeto-Siselietum. Two others making consistent contributions, however, are *Sempervivum tectorum* and *Verbascum austriacum*, both of which were called "Companion Species" by Braun-Blanquet. It must be remembered that the calculations here are based strictly on the particular set of data at hand, whereas Braun-Blanquet's recognition of the species as "characteristic" doubtless took account of his overall (and very extensive) experience.

Let us remove these four relevés constituting the representation of the Stipeto-Seselietum from the set, and test the affinity of the 18 others to them. The affinity of relevé 9 had a probability of .0035 – i.e., an affinity index of 2.45. Calculations of deviant indices for the whole data set had shown that this relevé differed markedly from all the rest, with a probability of .00026 (deviant index of 3.49). Let us associate this relevé with the group already separated, although Braun-Blanquet had included it in the Festuceto-Caricetum supinae rather than the Stipeto-Seselietum. The remaining 17 relevés show only two affinities greater than 1.0 with the separated group of five, which is within the bounds of expectation.

Let us now analyse further the remaining 17 relevés. The 136 similarities among them included seven values above 2.0 (only one is expected), of which one (between relevés 2 and 4) was above 3.0, showing that the data set is still heterogeneous. Moreover, there are three sets of three relevés linked in each case by similarities over 1.0. The closest of these similarity groups (relevés 1,2 & 3, with similarities of 2.77, 2.38 and 1.79) was treated as the core of a new cluster, and affinities with it were tested. No other relevés, however, showed up as closely associated with this cluster. These three relevés were all allotted by Braun-Blanquet to the Festuceto-Caricetum supinae, but so were five others which, in the probabilistic analysis, showed no particular association with them. The species making the largest contributions to their similarities showed no consistency.

If the two clusters now recognized are removed, the rest of the data still show heterogeneity as judged by the similarity matrix, with four values out of the 91 greater than 2.0. The highest of these (2.60) links relevés 14 and 16, with both of which relevé 15 is also closely associated (indices 1.64 and 1.57). Here, then, is the nucleus of a new cluster. All these relevés were placed by Braun-Blanquet in the association Festucetum-Poetum xerophilae; however, when one looks at the species making the major contributions to their similarities, they are not those characteristic of the association, but rather two characteristic of the order (*Echium vulgare* and *Seseli annuum*) and four companion species (*Aster alpinus*, *Coronilla varia*, *Lappula myosotis* and *Rosa eglanteria*). None of the remaining 11 relevés showed high affinity to these three representing the Festuceto-Poetum.

A separate analysis of this residue showed some heterogeneity (three deviant indices out of 11 and eight similarity indices out of 55 exceeding 1.0), but no affinity was shown to the relevés with the highest values in these tests, so further subdivision hardly seemed justified.

Thus, this objective analysis shows that the 22 relevés included in the Festucion valesiacae included three distinct groups representing the three associations, together with eleven relevés not showing high affinity with any one of them. If the underlying philosophy calls for any relevé within the alliance to be attached to one or other of the associations, the question arises of how these eleven should be distributed. For an answer to this, the question of significance does not arise – merely, to which of the three association clusters each of the residual relevés should be attached. One can use the affinity index to answer these questions in conformity with the probabilistic approach. Attaching to the residue the relevés already recognized as representing one of the associations, affinity indices are calculated in each of the three cases, and each residual relevé is then transferred to the grouping with which it shows closest affinity (Table 2).

It will be noted that relevés 4,5,7 and 8 show greatest affinity with the cluster C – that is, with those identified with the Festuceto-Poetum xerophilae — whereas Braun-Blanquet had placed them in the Festuceto-Caricetum supinae. On the other hand, relevés 19 and 21 show greatest affinity with cluster A (the Festuceto-Caricetum), although Braun-Blanquet placed them in the Festuceto-Poetum. Thus, though there are core groups of relevés which

Table 2. Affinity of residual relevés with core clusters.

Relevé	Attributed by Braun-Blanquet to Association	Cluster A 1, 2, 3 (Festuceto-Caricetum supinae)	Cluster B 9, 10, 11, 12, 13 (Stipeto-Seselietum)	Cluster C 14, 15, 16 Festucetum-Poetum xerophilae
4	Festuceto-Caricetum supinae	.641	.177	**1.142**
5	Festuceto-Caricetum supinae	.278	.525	**1.321**
6	Festuceto-Caricetum supinae	.000	.002	**.081**
7	Festuceto-Caricetum supinae	.001	.083	**.517**
8	Festuceto-Caricetum supinae	**.845**	.834	.467
17	Festucetum-Poetum xerophilae	.068	**.180**	.014
18	Festucetum-Poetum xerophilae	.208	**.220**	.018
19	Festucetum-Poetum xerophilae	**.866**	.054	.375
20	Festucetum-Poetum xerophilae	.477	**.501**	.090
21	Festucetum-Poetum xerophilae	**1.259**	.892	.034
22	Festucetum-Poetum xerophilae	.304	.153	**.341**

agree well with the original subjective conclusions, this concordance does not extend to the rest of the relevés recorded. In a spatial analogy, the data are not as well represented by three discrete clusters as by three smaller clusters around which the other relevés are distributed in a rather structureless cloud.

A large data set from the Netherlands data bank

The data comprise a random set of 100 relevés from those ascribed to the Alliance Nardo-Galion in the Netherlands data bank. These relevés are listed in the Appendix. The species recorded in them numbered 346. The data came into my hands converted from the ordinal data into percentages. In this conversion, which is often used as a preliminary to quantitative analyses, it should be borne in mind that the quasi-continuous scale may conceal marked discontinuities. The data as tabulated may include a number of figures of "37%", for instance, and no others between 31% and 45%, the apparently continuous percentage transformation giving a misleading appearance of precision.

The probabilistic procedure enables analyses to be done either with the original score data or with the derived percentages. Furthermore, an argument can be made that absence should be treated rather differently from the quantity if present – it is not just one end of a continuum (Goodall and Feoli 1988). Two zeroes are not to be regarded as indicating closer similarity between relevés than, say, values of 9% and 10% This point of view can be accommodated by the procedure described if each species is represented by two logically independent attributes: presence (a qualitative attribute taking the value "Present" or "Absent"), and quantity if present, which could either be an ordinal (score) or quantitative variable, in either case indeterminate if the species is absent (cf. Williams and Dale 1973).

When a similarity matrix was calculated from the data in the form of percentages (column 2 in Table 3), only five of the values showed a probability less than .01 — considerably *fewer* than is to be expected in a sample of 4950 values (the number of pairwise comparisons among 100 items) from a random sample of a uniform distribution between 0 and 1. The data were converted back to an ordinal scale similar to that used in the original observations, and a similarity matrix was again calculated. Here the results were completely different (Table 3, column 6). The number of probabilities less than .001 (i.e., with a logarithmic index greater than 3.0) was 60, as against 5 expected on the null hypothesis. Six values even exceeded 10.0.

The relatively uninformative character of the percentage data was explored further. The zero values were removed from the percentages, and a new set of qualitative attributes was introduced. Thus, as suggested above, the involvement of a species in a relevé was expressed by two variables logically independent of one another – (a) whether it was present or absent (P/A); and (b) if present, the percentage of area that it covered [if it was absent, attribute (b) was indeterminate and did not enter into the calculations for that relevé]. When this was done – using 692 attributes, instead of the 346 previously – the results were as shown in the third column of Table 3. It will be noted that the similarity values here are much closer to those obtained using the ordinal values than to those with the raw percentage data.

It was possible to analyse these results further by separating the presence/absence data and cover values in the 692-attribute data set, thus having two data sets each with 346 attrbutes. The results are shown in columns 4 and 5 of Table 3. It will be seen that the similarities based on percent. cover where present are generally rather low, but those based on presence/absence only are much closer to those where both were used together, and to those using the data in the ordinal form.

Table 3. Similarity values (logarithmic scale) among 100 relevés of Nardo-Galion. For pairings where a value in at least one of the logarithmic similarity matrices exceeds 3.00.

Relevé pair (1)	% cover incl. zeroes (2)	% cover , with P/A separate (3)	P/A only (4)	% cover where present (5)	Ordinal data (6)
3 25	2.43	13.78	11`.32	4.29	14.80
13 14	0.81	5.60	5.37	0.87	7.25
13 17	0.63	2.72	2.63	0.45	3.42
13 20	0.75	2.89	2.94	0.39	3.40
14 17	0.43	2.75	3.25	0.05	2.86
14 18	0.41	3.03	3.10	0.37	3.44
14 19	0.66	2.85	2.80	0.55	3.29
14 20	0.75	4.00	4.04	0.45	4.22
14 22	0.55	2.36	2.19	0.79	3.41
17 18	0.62	3.00	3.28	0.16	3.49
17 19	0.74	3.00	3.08	0.36	3.38
17 20	0.78	2.83	2.89	0.38	3.30
17 94	0.65	2.63	2.51	0.68	3.02
19 20	1.42	4.52	3.75	1.99	4.62
28 29	1.54	4.73	3.77	2.58	5.25
32 33	0.96	4.70	3.61	3.00	7.95
32 34	1.06	4.11	3.43	1.97	6.99
32 96	1.32	9.01	8.53	1.21	10.09
32 97	0.55	3.25	2.99	0.96	3.21
32 98	0.80	3.78	3.43	1.16	4.89
33 34	0.94	3.65	3.13	1.60	6.54
33 96	0.54	3.31	2.99	1.09	3.43
33 97	1.44	12.95	11.86	2.01	16.58
33 98	0.68	3.34	3.13	0.87	4.60
34 96	0.71	3.75	3.51	0.91	4.31
34 97	0.66	3.42	3.20	0.87	4.37
34 98	2.00	11.29	10.23	2.07	16.96
36 96	0.72	2.95	2.47	1.54	3.10
36 100	0.50	2.65	2.60	0.54	3.08
39 40	0.38	4.18	4.03	0.72	4.99
39 41	0.57	7.06	6.41	1.46	8.35
39 42	0.28	6.59	6.83	0.37	7.31
39 43	0.30	6.76	6.67	0.72	8.70
40 41	0.48	7.10	6.57	1.30	7.94
40 42	0.21	5.19	5.58	0.22	5.76
40 43	0.32	6.91	6.97	0.55	8.52
41 42	0.33	8.77	8.89	0.56	9.40
41 43	0.19	5.83	6.29	0.21	6.67
42 43	0.26	8.99	8.79	0.88	10.50
48 49	1.44	3.12	3.18	0.39	3.39
48 51	1.39	3.52	3.16	1.22	3.16
52 54	0.75	2.76	2.58	0.85	3.00
60 61	0.41	2.77	3.24	0.03	3.59
65 66	1.01	8.38	7.05	2.74	11.01
65 67	0.62	4.65	4.32	1.02	5.16
65 68	0.49	3.59	3.65	0.40	4.36
65 69	0.61	4.11	3.99	0.68	4.58
66 67	0.64	5.18	4.84	1.01	5.78
66 68	0.64	4.68	4.77	0.42	5.51
66 69	0.48	3.66	3.81	0.32	3.97
66 71	0.27	3.82	3.96	0.33	4.00
67 68	0.86	4.38	3.95	1.22	4.82
67 69	0.78	4.19	3.78	1.20	4.67
68 69	1.02	4.00	3.11	2.45	4.33
70 71	0.47	4.69	4.50	0.79	5.38
88 89	1.65	2.82	2.34	1.61	3.44
96 97	0.97	4.02	3.07	2.80	4.34
96 98	0.90	4.07	3.51	1.66	4.49
97 98	0.88	3.65	3.20	1.40	4.96
97 99	0.71	3.34	2.86	1.53	3.27
99 100	0.63	2.73	2.41	1.19	3.07

Table 4. Probability contributions of certain species to similarity indices. * Data insufficient.

Relevé Pair	Species	Data type				
		% cover	% cover, with P/A separated	P/A only	% cover where present	Ordinal data
3 & 25	52 *Juncus bufonius*	.9604	.0002	.0002	*	.0002
	54 *Radiola linoides*	.9214	.0012	.0012	.1667	.0002
	56 *Trifolium pratensis*	.9604	.0002	.0002	*	.0002
	63 *Cladonia* sp.	.9604	.0002	.0002	*	.0002
	65 *Salix repens* ssp. *argentea*	.9604	.0002	.0002	*	.0002
33 & 97	129 *Betula pendula*	.8844	.0030	.0030	.4667	.0002
	208 *Leptobryum pyriforme*	.9034	.0030	.0030	.4667	.0002
	209 *Clematis vitalba*	.9604	.0002	.0002	*	.0002
	213 *Ephemerum serratum* var.*minutis.*	9408	.0002	.0006	.3333	.0002
	214 *Isopterygium elegans*	.9802	.0002	.0002	*	.0002
34 & 98	34 *Scirpus cespitosa*	.2594	.5240	.5240	.3072	.0002
	205 *Poa angustifolia*	.8471	.0057	.0057	. 2500	.0002
	218 *Euphrasia rostkoviana*	.9604	.0002	.0002	*	.0002

One may identify the separate contributions of the different attributes to the indices in which they are combined. This has been done for the three pairwise similarities giving the highest values using the ordinal data (the pairs 3 and 25, 33 and 97, and 34 and 98). The results are shown in Table 4.

The most important contributions to the very high similarity between relevés 3 and 25, in the ordinal and other data sets, were made by species 52 (*Juncus bufonius*), 54 (*Radiola linoides*), 56 (*Trifolium pratensis*), 63 (*Cladonia* sp.) and 65 (*Salix repens* ssp. *argentea*) – in each case a probability of .0002, reflecting the fact these species occurred only in these two relevés, and that consequently these relevés were more similar, in respect of these species, than any of the other 4949 relevé pairs. The same applied to the contributions of attributes 129 (*Betula pendula*), 208 (*Leptobryum pyriforme*), 209 (*Clematis vitalba*), 213 (*Ephemerum serratum* var. *minutis*) and 214 (*Isopterygium elegans*) to the similarity between relevés 9 and 33, and to the contributions of attributes 34 (*Scirpus cespitosa*), 205 (*Poa angustifolia*) and 218 (*Euphrasia rostkoviana*) to the similarity between relevés 34 and 98. In Table 4 the contributions of these attributes to the similarities between these pairs of relevés, as calculated from the other data sets, are set out for comparison.

It is clear that contributions from the information on cover where present to the overall relationships are negligible compared with those for presence or absence of the same species.

The relative inability of the percentage-cover data to identify similar pairs of relevés may seem surprising. But it is in the nature of the similarity calculations when quantitative data are used, since the similarity depends on the absolute difference in the values observed. Take the comparison between the relevés 3 and 25 in respect of species 52 (*Juncus bufonius)*. In these two relevés, a cover value of 2% was recorded for *J. bufonius*, for all others a cover value of 0%. There was a difference in the cover value for pair 3 and 25 of zero; but the same was also true for all other pairs involving neither of these relevés. Thus, a zero difference for *J.bufonius* was recorded in 4754 of the 4950 pairs, while 196 pairs recorded a difference of 2.0%. The similarity between the two values of 2% was masked by the exactly equal similarities among the 98 values of zero.

In what follows, the data have been used in ordinal form, and the analysis of the data set proceeds exhaustively, guided only by the criterion of probability, and without reference to the sources or nature of the data. Subdivision of the data set has been pursued so long as the final groupings appear to be heterogeneous according to the probabilistic criteria used – the deviant index, and the similarity index. The earlier stages of the analysis are described at length, so as to make clear the logic of the process.

If one looks at the last column of Table 3, one can see that certain groups of relevés are linked by consistently high similarities. For instance, the 15 similarities between relevés 32, 33, 34, 96, 97 and 98 all exceed 3.0, and all but two exceed 4.0 Other groups linked in this way are: 39, 40, 41, 42 and 43; 65, 66, 67, 68 and 69; 13, 14, 17 and 20; and 14, 17, 19, and 20 (these last two groups could be combined if the threshold of acceptance is reduced to 2.32, the similarity between relevés 13 and 19).

Taking first the largest of these groups, with six relevés, they were treated as the nucleus of a possible "cluster". The affinity of all the other relevés to these six was calculated. One of the affinity indices (that for relevé 24) attained 5.455 on the logarithmic scale, with six others exceeding 2.0. Clearly these figures were far in excess of those expected on the null hypothesis (that any similarity of other relevés to those in the "cluster" was purely by chance assortment of species values), and the cluster could appropriately be increased by addition of other relevés showing high affinity to it. The first such was relevé 24. The process was repeated, and further relevés were added one by one until they totalled 20. Then, three relevés outside the cluster having about the same affinity values, they were all added to the cluster, and finally one more was added. The affinity indices at the successive stages of this process are

Table 5. Changes in affinity indices as subset size is increased.

Affinity of relevé	Orig. six	Subset Composition + 24	+35	+43	+42	+93	+38	+41	+ 40	+59	+14	+39	+60	+75	+8	+13 25 52
3	1.158	1.232	1.501	1.294	1.360	1.431	1.557	1.579	1.609	1.610	1.670	1.707	1.731	1.763	1.814	2.561
8	1.683	1.736	1.773	1.833	1.889	1.919	1.974	1.971	2.007	2.016	1.940	1.974	2.008	2.126	-	
13	1.382	1.478.	1.517	1.581	1.659	1.701	1.781	1.773	1.804	1.798	1.967	1.997	1.993	2.072	2.107	
14	2.098	2.232	2.326	2.413	2.392	2.464	2.340	2.264	2.285	2.304	-	-	-	-	-	
24	5.455	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
25	1.334	1.417	1.432	1.457	1.528	1.602	1.739	1.761	1.793	1.793	1.873	1.912	1.938	1.974	2.020	
35	3.402	3.556	-	-	-	-	-	-	-	-	-	-	-	-	-	
38	2.459	2.604	2.856	2.882	3.130	3.291	-	-	-	-	-	-	-	-	-	
39	0.926	0.968	1.032	1.325	1.596	1.665	1.997	2.152	2.247	2.237	2.336	-	-	-	-	
40	1.149	1.257	1.386	1.940	2.170	2.247	2.105	2.647	-	-	-	-	-	-	-	
41	1.617	1.686	1.558	2.009	2.344	2.357	2.511	-	-	-	-	-	-	-	-	
42	2.460	2.543	2.600	3.214	-	-	-	-	-	-	-	-	-	-	-	
43	2.703	2.855	2.934	-	-	-	-	-	-	-	-	-	-	-	-	
52	1.260	1.290	1.322	1.331	1.372	1.466	1.623	1.627	1.652	1.813	1.729	1.774	1.771	1.999	2.042	
59	1.977	2.139	2.172	2.216	2.276	2.273	2.284	2.224	2.322	-	-	-	-	-	-	
60	1.727	1.813	1.879	1.812	1.859	1.996	2.104	2.114	2.142	2.217	2.208	2.330	-	-	-	
75	1.098	1.204	1.380	1.335	1.387	1.440	1.721	1.824	2.005	1.784	1.832	1.964	2.023	-	-	
93	2.582	2.712	2.803	2.992	3.171	-	-	-	-	-	-	-	-	-	-	
Number of species	341	333	328	325	322	319	315	313	308	305	303	297	290	286	278	274

shown in Table 5. It will be noted that each relevé showed a progressive increase in affinity index as more relevés were added to the cluster, and that each except one (relevé 39) had already shown marked affinity to the original nucleus.

In Table 5, the number of species used in the calculations at each stage of the process of cluster-building is shown. There was a progressive decrease from the original 346 to 274. If a species has the same value throughout the complement of the selected subset, it cannot contribute to the calculations; thus, a species only occurring in one or more of the relevés in the selected subset, and nowhere else, is ignored in the calculations.

When this cluster A had been removed, the same process was repeated on the residue of 76 relevés. Another cluster (B) of ten relevés was identified and further divided as shown in Table 6. Repetition of the same procedure led to the recognition of ten more clusters, ranging in size from 11 to 3 relevés, before the residue of eight relevés appeared homogeneous. This residue should not, however, be regarded as a new cluster, for earlier similarity tests had shown no close relations among these eight relevés. Rather, they constitute a matrix within which each of the thirteen clusters recognized is located.

The next stage in the analysis was to check the internal uniformity of the larger clusters.

In the similarity matrix for cluster A of 24 relevés there were four clear sub-clusters – 3, 13, 14, 25 and 93; 32,33,34,96,97 and 98; 38 to 43; and 52, 59 and 60. The second sub-cluster was the clearest (in the previous stage, it had been the original nucleus of cluster A), with all fifteen indices exceeding 3.7, and three exceeding 12.0. In accordance with affinity calculations, relevé 24 was added to it. The remaining 17 relevés again showed the six relevés 38 to 43 as closely related, but no others were added to this subcluster by the affinity test. After this, the remaining 11 relevés showed extremely high similarities between relevés 3 and 25 (21.80) and between relevés 13 and 14 (13.81), both of which pairs had been in a sub-cluster recognized in the first stage of analysis of cluster A. All four were accordingly removed, leaving a residue of seven (8, 35, 52, 59, 60, 75 and 93). Among these seven relevés, a similarity matrix enables one to distinguish clearly two sub-clusters: 52,59 and 60 on the one hand, 8, 35, 75 and 93 on the other. Thus, the original cluster of 24 relevés has been divided into five sub-clusters, consisting of 7, 6, 4, 4 and 3 relevés respectively; let us call these A1, A2, A3, A4 and A5.

But the sub-cluster A1 was not homogeneous. A sub-sub-cluster A1.1 could be distinguished very clearly by the similarity and affinity calculations, consisting of relevés 24, 33 and 97. Internally, however, even this micro-cluster is very clearly distinguished into relevé 24, with a deviant index of 13.9 and relevés 33 and 97, with a similarity index also of 13.9 (at this level, they are indeed two aspects of the same test). Sub-cluster A1.2 also divided into two subsubclusters: 32 with 96, and 34 with 98. Subcluster A2 also required subdivision, the first sub-sub-cluster consisting of relevés 39 and 41, which were closely connected, with 38 associated less closely, while the second sub-sub-cluster consisted of relevés 40, 42 and 43, without further subdivision.

Subcluster A3 divided very clearly into two pairs: relevés 3 and 25, with a similarity of 13.27, and relevés 13 and 14, with a similarity of 7.24. Subcluster A4 also contained two pairs, relevés 8 and 93 on the one hand, and relevés 35 and 75 on the other. In sub-cluster A5, though relevés 52 and 59 were more similar than either to relevé 60, subdivision did not seem to be called for.

The other primary clusters were treated in the same way until similarity and deviant tests on the ultimate groupings showed no further internal heterogeneity. The results are shown in Table 6.

There may be some surprise that subdivision by this procedure continues to so fine a level, and that some of the indices remain so high even when the groups of relevés remaining are quite small. One may take sub-cluster A1.1 as an example, where among three relevés one was distinguished by a deviant index of 13.90 (i.e., a probability of 1.3×10^{-14}). If the contributions of individual attributes to this figure are examined, one finds that 47 out of the 68 species recorded in at least one of these relevés each contributes a probability of 0.333. These species were present in relevé 24 and absent from relevés 33 and 97, or absent from relevé 24 but present in both the others, or present in a larger or smaller quantity in relevé 24 than in the other two.

Following the exhaustive analysis of the 100 relevés into clusters, one would like to compare the groupings found with those recognized by traditional phytosociology. In the Netherlands data base, all these relevés were ascribed to the alliance Nardo-Galion (the criterion of selection), but divided between four associations: Galio hercynici-Festucetum ovinae, Gentiano pneumonanthes-

Table 6. Complete analysis of a sample of 100 relevés ascribed to Nardo-Galion in the Netherlands data bank. Note: The letters in parentheses indicate the Association to which the relevé is ascribed in the Netherlands data base: (A) Galio hercynici-Festucetum ovinae, (B) Gentiano pneumonanthes-Nardetum, (C) Botrychio-Polygaletum, (D) Betonico-Brachypodietum.

Cluster A	
Sub-cluster A1	
Sub-sub cluster A1.1	
Pair A1.1a	33(D) 97(D)
Single	24(C)
Sub-sub-cluster A1.2	
Pair A1.2a	32(D) 96(D)
Pair A1.2b	34 (D) 98(D)
Sub-cluster A2	
Sub-sub-cluster A2.1	
Pair A2.1a	39(D) 41(D)
Single	38(D)
Sub-sub-cluster A2.2	40(D) 42(D) 43(D)
Sub-cluster A3	
Pair A3.1	3(C) 25(C)
Pair A3.2	13(B) 14(B)
Sub-cluster A4	
Pair A4.1	8(A) 93(C)
Pair A4.2	35(D) 75(D)
Sub-Cluster A5	51(B) 59(B) 60(A)
Cluster B	
Sub-Cluster B1	
Sub-sub-cluster B1.1	
Pair B1.1a	1(B) 50(B)
Single B1.1b	100(D)
Sub-sub-cluster B1.2	65(D) 66(D)
Sub-cluster B2	
Sub-sub-cluster B2.1	67(D) 68(D) 69(D)
Sub-sub-cluster B2.2	70(D) 71(D)
Cluster C	
Sub-cluster C1	19(C) 20(C) 99(D)
Sub-Cluster C2	2(C) 7(C)
C residue	17(C) 18(C) 21(C) 22(C) 94(C) 95(C)
Cluster D	
Sub-cluster D1	36(D) 88(C) 89(C)
Sub-cluster D2	85 (C) 87(C) 92(C)
Cluster E	
Sub-cluster E1	47(B) 48(B) 49(B) 51(B)
Sub-cluster E2	
Pair E2.1	37(D) 63(A)
Residue	15(C) 61(A) 72(C)
Cluster F	
Pair F2	28(B) 29(B)
Residue	10(B) 30(B) 73(A)
Cluster G	82(A) 83(A) 84(A)
Cluster H	55(A) 56(A) 57(A) 62(A) 64(A)
Cluster I	4(A) 12(A) 53(A)
Cluster J	86(C) 90(C) 91(C)
Cluster K	5(A) 6(A) 23(C)
Cluster L	46(B) 58(A) 74(B) 76(B)
Cluster M	
Pair M1	78(A) 81(A)
Pair M2	79(A) 80(A)
Residue	16(B) 54(B)
Residue	9(B) 11(B) 26(A) 27(B) 31(B) 44(B) 45(B) 77(B)

Nardetum, Botrychio-Polygaletum and Betonico-Brachypodietum. In Table 6, the appropriate association is indicated for each relevé. In the majority of cases, the ultimate sub-clusters are of relevés attributed to the same association, and there is broad agreement even at higher levels. But there are many discrepancies. Though traditional phytosociology uses agreement in the presence and/or quantity of species in order to group relevés, the species used are only a relatively small proportion of all those recorded; probabilistic analysis, on the other hand, makes use of information from all species recorded which differ from relevé to relevé (species absent from all can contribute nothing to the analysis, and this would also be true if a species was present to the same extent in all).

Discussion

It has been mentioned that many plant ecologists have reservations about traditional phytosociology, largely because of the subjective element in the selection of sites for relevés (e.g., Poore 1955a,b, Kershaw 1964, Moore and Chapman 1986). But the methods described here would be equally applicable to samples collected with full objectivity; in any case, the method of analysis is logically quite independent of the way in which data have been collected.

The power of the probabilistic approach in separating relevé records according to their species composition is clear. One may note that the degree of subdivision indicated by this method is considerably finer than the traditional division into Associations.

The possibilities of analysing vegetation data by these techniques are not limited to data on the presence or abundance of species. Environmental data could well be included in the analyses if available – including spatial data (geographical location) or variables like aspect, which are difficult to handle in many systems. The opportunity to use data in which a range of values is recorded for each species at each location (e.g., from subsamples) could also be valuable.

No well-defined "stopping rule" has been adopted in this study. The "null hypothesis" underlying the probability calculations is that, at each stage, the species records are distributed independently among the relevés included, and on this hypothesis the probability estimates will be distributed uniformly between 0 and 1. Any test of this uniform distribution could be used, and different tests would give slightly different results, so that strict application of a conventional significance level of .05, say, might lead to different conclusions as to the inclusion of a particular relevé in a cluster. In any case, though, the choice of this or any other significance level would be an arbitrary decision.

The model of discrete vegetation types within which any "uniform" sample will fit, which underlies traditional phytosociology, is itself open to question. In this model, any sample would be from one or another of these discrete types, and would differ only randomly from other samples of the same type. If this model reflected reality, one would expect to find clusters being built up by clear-cut stepwise decisions, after which the relevés within the cluster would show no significant differences. However. the way in which clusters are repeatedly divided in this study, and the steady decline in index values as a cluster increases in size, support rather the concept of a number of centres around which relevés are concentrated, within a matrix of relevés whose composition resembles more or less remotely those represented by one or more of the centres of concentration.

The programs used in these analyses are freely available by writing to the author.

Acknowledgements: I am beholden to the ecologists at Wageningen in The Netherlands, particularly Dr. Stephan Hennekens, for making available to me a sample from their enormous data bank for Netherlands vegetation.

References

Braun-Blanquet, J. 1928. *Pflanzensoziologie. Grundzüge der Vegetationskunde.* Biologische Studienbücher 7. Walter Schoenichen, Berlin.

Braun-Blanquet, J. 1936. Über die Trockenrasengesellschaften des Festucion valesiacae in den Ostalpen. *Comm. Stat. Intern. Geobot. Med. Alp.* 49:169-189.

Fisher, R.A. 1934. *Statistical Methods for Research Workers*. 5th edition. Oliver and Boyd, Edinburgh & London

Goodall, D.W. 1964. A probabilistic similarity index. *Nature* 203: 1098

Goodall, D.W. 1966. Deviant Index: a new tool for numerical taxonomy. *Nature* 210: 216.

Goodall, D.W.1966a. A new similarity index based on probability. *Biometrics* 22: 882-907.

Goodall, D.W. 1968. Affinity between an individual and a cluster in numerical taxonomy. *Biom. Prax.* 9: 52-55.

Goodall, D.W. 1993. Probabilistic indices for classification – some extensions. *Abstracta Botanica* 17: 125-132

Goodall, D.W. 1994. The treatment of spatial data in probabilistic classification. *Abstracta Botanica* 18: 45-47.

Goodall, D.W. and E. Feoli. 1988. Application of probabilistic methods in the analysis of phytosociological data. *Coenoses* 3: 1-10. Reprinted in 1991 as Chapter 13 in E. Feoli and L.Orlóci (eds.), *Computer Assisted Vegetation Analysis.* Kluwer, Amsterdam, pp. 137-146.

Kershaw, K.A. 1964. *Quantitative and Dynamic Ecology.* E. Arnold, London.

Lancaster, H.O. 1949. The combination of probabilities arising from data in discrete distributions. *Biometrika* 36: 30-382.

Moore, J. J. 1962. The Braun-Blanquet system: a reassessment. *J.Ecol.*. 50: 761-769.

Moore, P.D. and S. B.Chapman (eds.). 1986 *Methods in Plant Ecology.* 2nd edition, Blackwell, Oxford.

Mueller-Dombois, D. and H. Ellenberg. 1974. *Aims and Methods of Vegetation Ecology.* Wiley, New York.

Poore, M.E.D. 1955a. The use of phytosociological methods in ecological investigations. 1. The Braun-Blanquet system. *J. Ecol.* 43: 226-244.

Poore, M.E.D. 1955b. The use of phytosociological methods in ecological investigations. 2. Practical issues involved in an attempt to apply the Braun-Blanquet system. *J. Ecol.* 43: 226-244.

Appendix

Identification in the Netherlands national data base of the relevés used in this analysis.

Relevé number	Netherlands identification	Cluster[1]	Association[2]
1	611	B1.1a	B
2	1431	C2	C
3	2101	A3.1	C
4	2233	I	A
5	2235	K	A
6	2359	K	A
7	2907	C2	C
8	4025	A4.1	A
9	4037	Residue	B
10	4840	F residue	B
11	7014	Residue	B
12	7982	I	A
13	9027	A3.2	B
14	9031	A3.2	B
15	9043	E2 residue	C
16	9051	M residue	B
17	9054	C residue	C
18	9055	C	C
19	9056	C1	C
20	9057	C1	C
21	9059	C	C
22	9060	C	C
23	9065	K	C
24	9105	A1.1	C
25	9106	A3.1	C
26	9114	Residue	A
27	9118	Residue	B
28	9783	F2	B
29	9784	F2	B
30	9785	F	B
31	9816	Residue	B
32	9843	A1.2a	D
33	9844	A1.1a	D
34	9845	A1.2b	D
35	10537	A4.2	D
36	10575	D1	D
37	10576	E2.1	D
38	10579	A2.1	D
39	10586	A2.1a	D
40	10587	A2.2	D
41	10588	A2.1a	D
42	10898	A2.2	D
43	10903	A2.2	D
44	12100	Residue	B
45	16302	Residue	B
46	17200	L	B
47	17207	E1	B
48	17209	E1	B
49	17210	E1	B
50	17211	B1.1a	B
51	17212	E1	B
52	17282	A5	B
53	17283	I	A
54	17707	M	B
55	17731	H	A
56	17909	H	A
57	17912	H	A
58	17915	L	A
59	17916	A5	B
60	17917	A5	A
61	17918	E2	A
62	17920	H	A
63	18008	E2.1	A
64	18009	H	A
65	18486	B1.2	D
66	18487	B1.2	D
67	18488	B2.1	D
68	18489	B2.1	D
69	18494	B2.1	D
70	18495	B2.2	D
71	18496	B2.2	D
72	19247	E2	C
73	20075	F	A
74	20076	L	B
75	20082	A4.2	D
76	23219	L	B
77	27455	Residue	B
78	27631	M1	A
79	27847	M2	A
80	27858	M2	A
81	62404	M1	A
82	63532	G	A
83	63533	G	A
84	63534	G	A
85	63579	D2	C
86	63580	J	C
87	63581	D2	C
88	63582	D1	C
89	63583	D1	C
90	63584	J	C
91	63585	J	C
92	63586	D2	C
93	75461	A4.1	C
94	75867	C	C
95	76120	C	C
96	87910	A1.2a	D
97	87911	A1.1a	D
98	87912	A1.2b	D
99	87914	C1	D
100	87915	B1.1b	D

1 See Table 6.

2 The letters indicate the Association to which the relevé is ascribed in the Netherlands data base: (A) Galio hercynici-Festucetum ovinae, (B) Gentiano pneumonanthes-Nardetum, (C) Botrychio-Polygaletum, (D) Betonico-Brachypodietum.