Hill, Mark O.; Harrower, Colin A.; Preston, Christopher D. 2013. **Spherical k-means clustering is good for interpreting multivariate species occurrence data**. *Methods in Ecology and Evolution*, 4 (6). 542-551. 10.1111/2041-210X.12038

Spherical k-means clustering is good for interpreting multivariate species occurrence data

M. O. Hill[1], C.A. Harrower, and C.D. Preston

Centre for Ecology and Hydrology, Wallingford OX10 8BB, England; moh@ceh.ac.uk

[1]present address 11 Chaucer Road, Cambridge CB2 7EB, England

Keywords

biclustering; biogeography; co-clustering; entropy; mutual information; R-mode, Twinspan

Summary

1. Clustering multivariate species data can be an effective way of showing groups of species or samples with similar characteristics. Most current techniques classify the samples first and then the species. A disadvantage of classifying the samples first is that relatively subtle differences between occurrence profiles of species can be obscured.

2. The k-means method of clustering minimizes the sum of squared distances between cluster centres and cluster members. If the entities to be clustered are projected on the unit sphere, then a natural measure of dispersion is the sum of squared chord distances separating the entities from their cluster centres; k-means clustering with this measure of dispersion is called spherical k-means (SKM). We also consider a variant in which the sum of squared perpendicular distances to a central ray is minimized.

3. Unweighted SKM is liable to produce clusters of very rare species. This feature can be avoided if each point on the unit sphere is weighted by the length of the ray that produced it. The standard SKM algorithm converges to very numerous local optima. To avoid this problem, we have developed a computationally intensive algorithm that uses multiple randomizations to select high-quality seed species.

4. The species clustering can be used to define simplified attributes for the samples. If the samples are then classified using the same technique, the resulting matrix of clustered species and clustered samples provides a biclustering of the data. The strength of the relationship between clusters can be measured by their mutual information, which is effectively the entropy of the biclustering.

5. The technique was tested on five ecological and biogeographical datasets ranging in size from 30 species in 20 samples to 1405 species in 3857 samples. Several variants of SKM were compared, together with results from the established program Twinspan. When judged by entropy, SKM always performed adequately and produced the best clustering in all datasets but the smallest.

Introduction

Methods of classifying species and samples from multivariate species occurrence data were much investigated in the 1960s and 1970s. A distinction was made between Q-mode methods, in which the samples or stands were clustered, and R-mode methods, in which the species were clustered. Occasionally, as in Lambert & Williams's (1962) nodal analysis and Hill's (1979) program Twinspan both samples and species were clustered, one after the other. By the end of the 1970s, it was accepted that the correct procedure is to classify the samples first. R-mode methods were in eclipse.

More recently, in the period 1995-2010, there has been renewed interest in numerical classification, mainly in the fields of text mining (Manning, Raghavan & Schütze, 2008) and genomics.

Along with the general increase of interest in numerical classification, two-way classification has received increased attention. Two-way classification is variously known as biclustering (Madeira & Oliveira, 2004; Gupta & Aggarwal, 2010), co-clustering (Banerjee *et al.*, 2007; Jain, 2010) or two-mode clustering (Van Mechelen *et al.*, 2004; Schepers & Van Mechelen, 2011; Hageman *et al.*, 2012). The term biclustering, used here, was apparently introduced by Mirkin (1996), who does indeed cite Twinspan as an example. There has, however, been little flow from methodologies used in text mining and bioinformatics into ecology.

A promising approach to clustering and biclustering is to treat these methods as fitting models to a data matrix. An interesting example is set out by Martella & Vichi (2012). They and several other authors (ter Braak *et al.*, 2009; Schepers and Van Mechelen, 2011) use the least-squares criterion to approximate either a raw matrix or a similarity matrix. Approximations to a raw matrix based on unweighted least squares are generally not suitable for occurrence data in ecology and biogeography. We set out a crude multiplicative model for such data, but do not use it except as a means of estimating the Akaike Information Criterion to select the numbers of clusters.

Our interest in R-mode clustering was rekindled during a study of European plant distributions (Finnie *et al.*, 2007). For this purpose, we compared species distributions with cluster centroids, using the cosine measure of similarity. This measure is widely used in text mining (Manning *et al.*, 2008). Finnie's (2007) clustering algorithm was agglomerative, building up clusters from pairs of similar individual species. It was rather complicated and had some arbitrary parameters. Therefore, in a subsequent study of British and Irish liverworts (Preston, Harrower & Hill, 2011), we used a simpler method. We called it Clustaspec. It starts by being agglomerative, and continues with a second phase in which the smallest clusters are systematically removed and their species distributed to larger ones. When Clustaspec was applied to other datasets, it usually gave good results, but it had a tendency to generate small clusters of rare species confined to special habitats. We were not entirely satisfied with it.

Both Finnie's (2007) method and Clustaspec tidied up the final clustering by means of an iterative relocation algorithm, by which each species was allocated to the nearest cluster centre, repeating the process until stability was reached. For clustering in Euclidean space, this method is known as the k-means algorithm (Krishna and Murty, 1999). Finnie's algorithm and Clustaspec defined proximity in terms of the cosine similarity measure. Their relocation algorithm was therefore a case of the spherical k-means (SKM) algorithm, whose properties have been investigated by Vinh (2008). There is, however, an important difference. In the SKM algorithm described by Vinh, the objects to be clustered are first projected on the surface of the unit hypersphere, and are thereafter clustered by the SKM algorithm. In the algorithm used by us, the unit hypersphere was not considered, the cluster centres being calculated simply as the centroids of untransformed vectors. As explained below, this amounts to weighted SKM, with weights proportional to the length of the untransformed vectors. The weights make a big difference.

In Clustaspec, we used the SKM algorithm merely for tidying up the clusters. Vinh (2008) shows that the SKM algorithm will converge to a local optimum of the SKM objective function, defined as the sum of squared chord distances between cluster centres and individual cluster vectors. He also points out that there are very many such local optima. Indeed, there are so many local optima that the quest for the global optimum can be very arduous. For this quest, we have devised an algorithm based on 'key species'. These are defined as those species that are most closely aligned to the cluster centres. Key species were used by Finnie *et al*. (2007) and Preston, Harrower & Hill (2011) to name the clusters. In the algorithm described below, they are used also to initiate the clusters.

Data and methods

DATASETS

Five datasets were studied in detail (Table 1):

1. Dune meadow data, discussed by Jongman *et al.* (1995);
2. Danube meadow data from a 25 km^2 study area east of Ulm, as discussed by Mueller-Dombois & Ellenberg (1974) and used in the manual for Twinspan (Hill, 1979);
3. The Arable bryophyte dataset analysed by Preston *et al.* (2010);
4. The Liverwort dataset used by Preston, Harrower & Hill (2011); and
5. An equivalent dataset for British and Irish native vascular plants; the dataset comprises all native records mapped by Preston *et al.* (2002), including old records as well as recent ones, but excluding records of native species from localities where they are known to be introduced.

COMPUTER PROGRAMS

The program Clustaspec was written in R by Harrower for classifying liverworts. Our program for spherical k-means was subsequently written by Hill in Fortran, using the GNU Fortran G77 v0.5.25 compiler for Windows XP (Free Software Foundation, 1999). Both Clustaspec and the new program, Spherikm (SPHERIcal K-Means), can be downloaded from the BRC website http://www.brc.ac.uk.

As in Euclidean k-means clustering, the number of clusters, *k*, has to be specified in advance. The best clustering is defined to be that which minimizes the sum of squared distances between cluster members and their centroids. Specifically, let

$$\mathbf{A} = [a_{ij}] \qquad (i = 1,...,m;\ \ j = 1,...,n)$$

be a matrix specifying the occurrence of *n* species in *m* samples; the value of a_{ij} is either the quantity of species *j* found in sample *i* or may be 1 or 0 if **A** is a matrix of presences and absences. Let $\mathbf{a}_j$ be the vector of elements corresponding to species *j*, i.e.

$$\mathbf{a}_j = [a_{ij}] \qquad (i = 1,...,m)$$

Define

$$\mathbf{b}_j = \mathbf{a}_j \,/\, \|\mathbf{a}_j\|$$

This is the projection of $\mathbf{a}_j$ on the unit hypersphere. Then the spherical k-means problem is to find a set of cluster centres

$$\mathbf{x}_1, \ldots, \mathbf{x}_k$$

on the unit hypersphere that minimize the sum of squared chord distances between the vectors $\mathbf{b}_j$ and and the cluster centres. In symbols, the criterion to be minimized is

$$D = \sum_j (\mathbf{b}_j - \mathbf{x}_h)^{\mathrm{T}}(\mathbf{b}_j - \mathbf{x}_h) = \sum_j 2(1 - \mathbf{b}_j^{\mathrm{T}}\mathbf{x}_h)$$

As $\mathbf{b}_j^{\mathrm{T}}\mathbf{x}_h$ is simply the cosine of the angle between $\mathbf{a}_j$ and $\mathbf{x}_h$, an equivalent problem is to maximize the sum of cosines between the vectors and their cluster centres. In our calculations, we have used a weighted version of the summed cosine criterion, i.e.

$$D[\mathrm{w}] = \sum \mathrm{w}_j\, \mathbf{b}_j^{\mathrm{T}}\mathbf{x}_h$$

Different weighting systems, from $\mathrm{w}_j = 1$ (standard SKM) to $\mathrm{w}_j = \|\mathbf{a}_j\|$ are compared below. In the case where weights $\mathrm{w}_j = \|\mathbf{a}_j\|$, the centroid of a cluster of weighted points $\mathbf{b}_j$ on the unit hypersphere is then exactly aligned to the cluster centroid of vectors $\mathbf{a}_j$ in the original space.

We have made much use of the spherical k-means algorithm. The SKM algorithm starts with an initial set of trial cluster centres, and derives a new set by the following two steps (Vinh, 2008).

1. The membership assignment step – each vector is assigned to the cluster of the trial cluster centre to which it is closest; and
2. The centre adjustment step – new cluster centres are located at the centroid of the members defined by step 1.

If these two steps are repeated, the algorithm converges to a local optimum.

Our algorithm, mentioned in the introduction, is based on seed vectors. The initial clusters consist of a set S of seed vectors $\mathbf{s}_1$, $\mathbf{s}_2$, ... , $\mathbf{s}_k$, selected from $\mathbf{a}_1$, $\mathbf{a}_2$, ... , $\mathbf{a}_n$. Let the local optimum derived from S by application of the spherical k-means algorithm to the seed vectors be denoted by SKM(S). In each of the clusters defined by SKM(S), there will be a best-aligned vector (a 'key species' in the terminology used above). A self-regenerating set of seeds S is one such that the key vectors in the clusters of SKM(S) are identical to the seeds. When solutions are restricted to those local optima derived from self-regenerating seeds, the search for the (restricted) global optimum is more tractable. The algorithm proceeds in three stages using random or restricted-random vectors S as seeds for SKM(S): 1 make a shortlist of suitable seeds; 2 select a list of k seeds sequentially from the shortlist by adding in the most frequently-selected key vector that is not already in the selected list; 3 adjust the list by trying out alternative seed lists in which each element of the list $\mathbf{s}_1$, $\mathbf{s}_2$, ... , $\mathbf{s}_k$ selected at stage 2 is replaced with an unselected vector from the shortlist. If any replacement seed list decreases the sum of squared deviations, select the best and repeat stage 3 with the new seed list until stability is reached. This process, consisting of stages 1 to 3, is repeated 10 times and the best solution out of the 10 replicates is retained.

PERPENDICULAR SPHERICAL K-MEANS

There are potentially two variants of the spherical k-means problem. They differ, as explained in the discussion, in how much leverage is given to aberrant cluster members. In spherical k-means as outlined above, we minimize the sum of squared chord distances. This method is abbreviated below as CSKM for chord spherical k-means. However, in principle an equally suitable criterion is the sum of squared perpendicular distances (Fig. 1). There is a small complication with this method, in that the minimum is not generally achieved by dropping perpendiculars to the ray through the centroid of the cluster. Specifically, let the minimizing ray be $\mathbf{x}$. Then, ignoring the weights, we seek to minimize

$$D = \sum_i \sin^2 (\text{angle between } \mathbf{b}_\mathrm{j} \text{ and } \mathbf{x})$$

$$= n_j - \sum_j \cos^2 (\text{angle between } \mathbf{b}_j \text{ and } \mathbf{x})$$

$$= n_j - \sum_j (\mathbf{b}_j^{\mathrm{T}}\mathbf{x})^2$$

subject to the constraint that **x** is on the unit hypersphere, i.e.

$$\mathbf{x}^{\mathrm{T}}\mathbf{x} = 1$$

To find the direction of **x**, we solve the problem with Lagrange multipliers and minimize

$$\Lambda = D - \lambda\, \mathbf{x}^{\mathrm{T}}\mathbf{x}$$

Differentiating with respect to **x**, Λ is minimized when

$$\mathrm{d}\Lambda/\mathrm{d}\mathbf{x} = -2 \sum_i (\mathbf{b}_j^{\mathrm{T}}\mathbf{x})\, \mathbf{b}_j - 2\lambda\, \mathbf{x} = \mathbf{0}$$

Therefore

$$\mathbf{x} = (-1/\, \lambda) \sum_i (\mathbf{b}_j^{\mathrm{T}}\mathbf{x})\, \mathbf{b}_i$$

This relationship allows us to solve for **x** iteratively, starting with a trial vector $\mathbf{x}_{(0)}$ which is the centroid of $\mathbf{b}_j$ and then repeating the process so that

$$\mathbf{x}_{(1)} = (-1/\, \lambda_{(1)}) \sum_i (\mathbf{b}_j^{\mathrm{T}}\mathbf{x}_{(0)})\, \mathbf{b}_j$$

and so on. The value of $(-1/\, \lambda_{(1)})$ is chosen to be the positive value that places $\mathbf{x}_{(1)}$ on the unit hypersphere. Note that because the vectors $\mathbf{b}_j$ and **x** are in the positive quadrant, all the coefficients $\mathbf{b}_j^{\mathrm{T}}\mathbf{x}$ are also positive. Once the direction of **x** is known, calculation of *D*, the sum of squared deviations, is immediate.

BICLUSTERING AND MEASURES OF CONCENTRATION

Biclustering of the data was achieved by first clustering the species, then condensing the data to account for species clusters (i.e. adding together the species vectors in each cluster), transposing, and clustering the samples by the same method. Suppose, for example, that a given sample contains species A, B, C and D, all with quantity 1, and that A, C and D belong to Species-cluster 1 and B belongs to Species-cluster 2. The composition of the sample for the purposes of the secondary clustering is Species-cluster 1 quantity 3, Species-cluster 2 quantity 1.

With presence data, a well known goodness-of-fit measure for a two-way table is the chi-squared statistic based on the sum of squared deviations between observed and expected values in cluster cells

$\sum(o\text{-}e)^2/e$. This statistic does not generalize readily to data types where the original values are quantities or are ordinal classes. A measure that generalizes better is the dimensionless (geometric) mean ratio of observed to expected values. Let *I* denote a cluster of samples and *J* denote a cluster of species. The observed value o_{IJ} is the sum of matrix elements in clusters *I* and *J*, i.e.

$$o_{IJ} = \sum_{i \in I} \sum_{j \in J} a_{ij}$$

Then the expected value is defined as

$$e_{IJ} = \sum_I o_{IJ} \sum_J o_{IJ} / \sum_I \sum_J o_{IJ}$$

Concentration can be measured by the statistic

$$K = \exp\left(\frac{\sum\sum o_{IJ} \ln(o_{IJ}/e_{IJ})}{\sum\sum o_{IJ}} \right)$$

In reporting results, *K* is called the ‘concentration ratio’ because it measures the geometric mean ratio of the observed values in the cluster cells to those that would be expected if species occurred at random. In the case where the data a_{ij} are presences and absences, *K* is effectively the *G* statistic of Sokal & Rohlf (1981) which measures the entropy (more properly the mutual information) of the biclustering. It can be argued that mathematically the best solution is that which maximizes the entropy (Banerjee *et al.*, 2007).

CLUSTER PRESENTATION AND CHOICE OF CLUSTER NUMBERS

For clarity of presentation, the clusters, once defined, were arranged by a two-stage process. First they were ordered by correspondence analysis (Hill, 1982, Jongman *et al.*, 1995). Then they were clustered hierarchically by Ward’s method (Legendre & Legendre, 1998), an agglomerative technique which at each stage unites the pair that minimally increases the total within-cluster variance. Clusters were ordered so that the hierarchy resulting from Ward’s method could be presented cleanly. In other words, when groups were united, they were placed side-by side. Correspondence analysis order was retained if there was a choice, with the cluster having minimum axis score appearing as the first in the final order. The hierarchy was printed out in Newick format for viewing in Dendroscope (Huson *et al.*, 2007). We give two examples in the Supplementary Information.

For selecting cluster numbers, the biclustering was approximated by fitting a multiplicative model with the same row totals, column totals and cluster totals

$$\hat{a}_{ij} = a_{i.}\, a_{.j}\, o_{IJ} / (o_{I.}\, o_{.J})$$

and then calculating an analogue of the concentration ratio

$$K' = \exp\left(\Sigma\Sigma a_{ij}\ \ln(a_{ij} / \hat{a}_{ij}) / \Sigma\Sigma a_{ij}\right)$$

K' measures the size of the residuals after fitting $\hat{a}_{ij}$. If the values a_{ij} were counts, then the G statistic, which is distributed as χ^2 would be

$$G = 2N \ln(K')$$

where N is the total count, i.e. $\Sigma\Sigma a_{ij}$. Let F be the number of fitted constants, k_1 the number of species clusters and k_2 be the number of sample clusters. Then in this case, ignoring a constant offset in AIC,

$$F = (k_1 - 1)(k_2 - 1) + m + n - 1$$

$$\mathrm{AIC} = G + 2F = 2N \ln(K') + 2F .$$

If this criterion is to be applied where a_{ij} are not counts then an analogue for N needs to be found. If the values a_{ij} are presences and absences (0 or 1) N can be taken to be the total $\Sigma\Sigma a_{ij}$. If a_{ij} are quantities such as species abundance values, a suitable choice of N is, in the notation of Hill (1973) the number N_2, i.e. $(\Sigma\Sigma a_{ij})^2 / \Sigma\Sigma a_{ij}^2$. The value of AIC calculated here using N_2 is called ‘quasi-AIC’, to emphasize the fact that it is not based on likelihood in a statistical model.

TESTING THE METHODS

The standard SKM analyses, for the purposes of this paper, are those in which projections of data on the unit hypersphere are weighted in proportion to $\|\mathbf{a}_j\|$, the length of their vectors. These are signified as W1 as the weights are $\|\mathbf{a}_j\|^{1.0}$. Both the chord variant CSKM and the perpendicular variant PSKM have been tested. W00 is SKM as usually understood, with species and samples projected on the unit sphere and given equal unit weight $\|\mathbf{a}_j\|^{0.0}$. Two other species weightings were considered, namely W0, in which species were weighted as in W00 but samples in the subsequent sample clustering were weighted as in W1. W0.5 is defined similarly, with species weights $\|\mathbf{a}_j\|^{0.5}$ and sample weights $\|\mathbf{a}_j\|^{1.0}$.

Datasets other than the vascular plant dataset were transposed to check whether it is better to cluster the species first and then the samples, or vice-versa. Transposed analyses, in which the samples were clustered first and the species clustered second, are denoted by Transposed W00, Transposed W1, etc.

Twinspan does not produce a specific number of clusters, but does generate a hierarchy for both species and samples. To compare it with the other methods, clusters were defined on the basis of the higher levels of the hierarchy, trying to avoid very small clusters that would give the other methods an unfair advantage. This process was not automated and clusters were selected by eye.

Results

CONCENTRATION RATIOS

Except for the dune dataset, the highest concentration ratios were found either with standard weighted CSKM or PSKM (Table 2). In the biogeographical datasets, the PSKM arrangement was the most concentrated, whereas in the Danube and Arable datasets, the CSKM arrangement was more concentrated. Twinspan produced less highly concentrated solutions. Clustaspec produced results that were rather similar to those from SKM but were somewhat less concentrated.

DUNE MEADOW DATA

The Dune dataset, the most species-poor, is small enough to be displayed in full in Fig. 2. Some samples had much bare ground. In sample 17, only *Anthoxanthum odoratum* had cover value greater than 2; its cover value 4 signifies less than 5% vegetation cover. Two biclusterings are shown. The first (Fig. 2a), with a concentration ratio of 1.51, is the standard SKM solution for 7 species clusters and 5 sample clusters. The second (Fig. 2b), with concentration ratio 1.44, shows the simplified solution with 5 species clusters and 4 sample clusters suggested by the quasi-AIC statistic. For this dataset and not the others, better results were obtained for the (7,5) case by clustering the samples first and then the species. For the preferred (5,4) case, it was better to cluster the species first.

DANUBE MEADOW DATA

The Danube dataset is displayed in Fig. 3, which shows 6 clusters for 34 species and omits the 60 species with lowest average biomass. An expanded version of the figure, colour-coded for concentration ratios and including the PSKM biclustering, is given the Supporting Information, along with the solutions suggested by quasi-AIC, which have 5 species clusters and 5 sample clusters. Concentration ratios were 1.56 for the (6,8) case and 1.47 for the (5,5) case.

OTHER DATASETS

Table 3 shows bicluster totals for the arable field dataset. The concentration ratio was 1.23 for the (9,12) case, which was investigated in detail. The minimum quasi-AIC was found with 24 species clusters and 28 sample clusters. This solution had concentration ratio 1.37, and is set out briefly in the Supporting Information.

Bicluster totals and concentrations for the liverwort and vascular-plant datasets are not shown here but are given in the Supporting Information.

Discussion

THEMES AND ALGORITHM

When differing weighting schemes W0, W0.5 and W1 were applied, it became apparent that a relatively small suite of cluster themes appeared repeatedly. An analysis of themes for liverwort analyses (see Supporting Information) revealed 16 themes from 8 analyses, each of which had 10 species clusters. Four themes, namely Southwest coast, Irish Atlantic, Calcicole montane and Eastern snowpatch, were nested within larger W1 themes. Two themes, Middle western and Rather upland, were intermediate between W1 themes.

The algorithm, based on random seeds, cannot be guaranteed to converge to the global optimum. Our use of cluster seeds is similar to the MedoidKNN procedure proposed by Kalogeratos & Likas (2011).

Our algorithm is somewhat complicated, but we found that simpler algorithms were frustratingly unable to locate really good solutions. Solutions that were close to the optimum displayed almost all the same themes. For example, the second-best solution for CSKM W1 applied to the vascular plants was found in two of the ten main replicates. It had mean cosine 0.79932 compared with 0.79939 for the best. Its 20 themes were the same. Of its key species, 13 were identical to those in the best solution, and five appeared as number 2 in order of alignment to the best solution. Of the remaining two, *Carex echinata* was 5th in order of alignment to the moorland cluster, and *Alisma plantago-aquatica* had moved from the eutrophic lowland cluster to the aquatic lowland cluster. In the best solution the cosine similarity of *A. plantago-aquatica* to the eutrophic cluster was 0.910, while its similarity to the aquatic cluster was 0.878. Clearly these are small differences, but in our judgement, the mathematically suboptimal solutions were for the most part somewhat inferior ecologically.

The algorithm is not especially quick. Typically, a solution for one of the larger problems required about 50,000 iterations of the SKM algorithm. Applied to the arable dataset, with 11,003 elements, the calculation took 27 and 40 minutes respectively for CSKM and PSKM to extract 9 species clusters and 12 sample clusters, using a desktop computer with a 2.8 GHz processor. Calculations for the vascular plant dataset, which is 140 times bigger, took about a week, partly because of the large size of the dataset and partly because more groups were sought.

We have no doubt that efficiency could be improved, but this would require either parallel processing or a more subtle algorithm.

WHAT MAKES A GOOD CLUSTERING?

From the early days of plant ecology, clustering has been used for data exploration. During the period 1950-1980 investigators sought objectivity through the use of numerical methods. The methods of Braun-Blanquet and his followers were frequently attacked as lacking objectivity. However, Goodall (1953) noted early on that Braun-Blanquet's method of 'character species' could in principle be made objective. It has much resemblance to the algorithm based on key species, used here.

In biogeographical analyses (Finnie *et al.*, 2007, Preston, Harrower & Hill, 2011), we have successfully employed R-mode methods that rely on the cosine measure of similarity. Forty years earlier, Orloci (1967) had proposed the method of 'optimal agglomeration'. This is essentially Ward's minimum variance method (Legendre & Legendre, 1998) on the surface of a hypersphere. It also uses the cosine measure of similarity but never achieved much popularity. This may well be because optimal agglomeration used unweighted vectors, i.e. the weighting scheme W0, which in our study proved less satisfactory than W1 (Table 1).

How then should we judge clustering methods? Their ability to extract clear patterns is essential. They should not pick out minor groups at the expense of the broad picture. For these reasons, the concentration ratio has all the hallmarks of a good criterion by which biclusterings can be judged. Perhaps, it should be used directly, just as maximum entropy methods are used in other applications. We do not know of a direct algorithmic approach to the maximization problem and have therefore used variants of SKM and compared them by the concentration ratio (Table 1). In principle, the 'double k-means' approach explored by Martella & Vichi (2012) could be extended from k-means to SKM using the concentration ratio as objective function. The problem of avoiding local optima would be just as severe with double SKM as with ordinary SKM, but double SKM might be useful to clean up approximate solutions derived by sequential clustering (species, followed by samples).

A good clustering should not have too many or too few clusters. For the two smaller datasets, the application of quasi-AIC to restrict cluster numbers was successful. For the Dune Meadow data, the groups (Fig. 2b) make obvious ecological sense and are: 1 dicots (and one grass) of low-nutrient permanent grassland; 2 dicots (and one moss and annual grass) of short turf; 3 competitive pasture grasses (and one dicot); 4 dune-slack margins; and 5 dune-slack centres. For the Danube Meadow data (Supporting Information, Figure S1b) the five groups are: 1 Dry calcareous grassland (Mesobromion); 2 *Poa pratensis* (dominant in one aberrant sample); 3 coarse pasture grasses (and 4 dicots); 4 wetland grasses (confined to a sample that was regularly inundated); and 5 dicots (plus two grasses and one sedge). This classification brings out themes corresponding to two main gradients, dry to wet, and high-grass to high-dicot. In addition, it distinguishes an aberrant sample.

With the arable bryophyte data, quasi-AIC suggested a substantial increase in cluster numbers from (9,12) to (24,28). The 24×28 concentration matrix is shown in the Supporting Information (Fig. S2). There is undoubtedly much structure even at this level of subdivision, but in most applications it is preferable to have a succinct overview. Indeed, Preston *et al*. (2010) recognized just six species assemblages based on detrended correspondence analysis followed by k-means clustering. Many of the clusters recognized by both CSKM W1 and PSKM W1 with 9 species clusters and 12 sample clusters are broadly similar to assemblages described by Preston *et al*. (2010).

The hierarchy derived by Ward's method (illustrated in Supplementary Information Fig. S3 and Fig. S4) also provides an overview. There is indeed no straightforward answer to what makes a good clustering. It depends on whether the investigator is looking for detail or for broad features.

COMPARISON OF METHODS

All the classifications outlined above produced recognizable patterns that can be interpreted in ecological or biogeographical terms. There was a clear progression from the more balanced W1 analyses to the W0 analyses, which generated some small but rather distinct clusters of rare species as well as some large clusters. The pattern is shown for liverwort clusters (Table 4). The two least concentrated biclusterings resulted from Twinspan and CSKM W0; here the largest sample clusters were 1402 and 790, i.e. 41% and 23% of all 3459 samples. The Twinspan classification was especially uneven, and failed to distinguish a category of montane species. In CSKM W0, the maximum cell concentration of 52.1 was for 10 Irish-Atlantic species in a cluster of 33 hectads among which 26 were in Ireland and 7 in Britain. Clearly the W0 biclusterings were too uneven to be generally suitable. The W0.5 biclusterings, on the other hand, were nearly as concentrated as the W1 biclusterings.

The PSKM W1 classification of the arable dataset produced two essentially single-species clusters, *Bryum klinggraeffii* in one cluster and *B. violaceum* in the other. This dataset has less inherent structure than the other datasets from Britain and Ireland, because it was obtained from a single, rather uniform habitat that is confined to the lowlands. The liverwort and vascular plant datasets cover the whole environment, including woods, grasslands, rivers, coasts and mountains. The CSKM and PSKM methods produced very similar results for a given weighting when applied to these data.

Apart from the fact that PSKM minimizes the sum of squared distances to rays not passing through exact cluster centroids, the main difference between CSKM and PSKM is that PSKM maximizes $\sum_j (\mathbf{b}_j^{\mathrm{T}}\mathbf{x})^2$ whereas CSKM maximizes $\sum_j (\mathbf{b}_j^{\mathrm{T}}\mathbf{x})$. This distinction underlies the main practical difference between them, namely that CSKM emphasizes overall conformity to the centroid, whereas PSKM pays less attention to species that are more deviant, emphasizing those that are well aligned. PSKM produced marginally higher-entropy biclusterings than CSKM for the two biogeographical datasets. Our analyses do not indicate that either of the two is always better. We note in passing that the most truly spherical k-means clustering would be angular spherical k-means (ASKM), which minimizes the sum of squared angles to a central ray. ASKM would take longer to compute than CSKM, because as with PSKM the position of the central ray has to be calculated by a recursion formula. ASKM would be more sensitive to poorly-aligned elements than CSKM, but we have not programmed it and do not report on its properties here.

Although the differing weightings of CSKM and PSKM produced results that differ in their concentration ratios, the selection of a preferred weighting may on occasion be better judged by the requirements of the user rather than by differences in concentration ratio. The particular choice may depend on the dataset in question. To our way of thinking, the W1 methods produced a satisfactory classification of the liverworts, which have very few ubiquitous species. When applied to vascular plants, among which widespread species are more frequent, the W1 weighting produced three groups of almost ubiquitous species, differing in the rather small areas of Britain and Ireland from which they are absent. For vascular plants, therefore, W0.5 weightings generated a more interesting set of patterns, which will be reported elsewhere.

Conclusions

Spherical k-means is shown to be a powerful clustering method, especially for R-mode analyses. It has hitherto been neglected because it tends to produce very unequal cluster sizes unless the commoner species are given greater weight. It also requires careful programming to avoid unsatisfactory local optima. There is no general answer to whether CSKM or PSKM is better; we recommend doing both and selecting the solution with higher concentration ratio. The quasi-Akaike criterion is good for selecting the number of clusters in small datasets, but in large datasets convenience is likely to be the main consideration.

Acknowledgements

This work was partly funded by the Joint Nature Conservation Committee (JNCC) and the Natural Environment Research Council under the Biological Records Centre (BRC) work programme for 2010 and 2011. We are grateful to John Birks, Cajo ter Braak and an anonymous referee for helpful comments on an earlier draft, and thank the British Bryological Society and the Botanical Society of the British Isles for permission to use their data.

References

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S. & Modha, D.S. (2007) A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research,* **8**, 1919-1986.

Finnie, T.J.R., Preston, C.D., Hill, M.O., Uotila, P. & Crawley, M.J. (2007) Floristic elements in European vascular plants: an analysis based on *Atlas Florae Europaeae*. *Journal of Biogeography,* **34**, 1848-1872.

Free Software Foundation (1999) *GNU Fortran G77 v0.5.25*. Downloaded 2010 from http://kkourakis.tripod.com/g77.htm.

Goodall, D.W. (1953) Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Australian Journal of Botany,* **1**, 39-63.

Gupta, N. & Aggarwal, S. (2010) MIB: Using mutual information for biclustering gene expression data. *Pattern Recognition*, **43**, 2692-2697.

Hageman, J.A., Malosetti, M. & van Eeuwijk, F.A. (2012) Two-mode clustering of genotype by trait and genotype by environment data. *Euphytica,* **183**, 349-359.

Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427-432.

Hill, M.O. (1979) *TWINSPAN - a FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes.* Section of Ecology and Systematics, Cornell University, Ithaca, N.Y.

Hill, M.O. (1982) Correspondence analysis. *Encyclopedia of Statistical Sciences* (eds S. Kotz, N. L. Johnson & C. Read), pp. 204-210. Wiley, New York.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M. & Rupp, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**, 460. doi:10.1186/1471-2105-8-460.

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651-666.

Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R. (1995) *Data analysis in community and landscape ecology*. Cambridge University Press, Cambridge.

Kalogeratos, A. & Likas, A. (2011) Document clustering using synthetic cluster prototypes. *Data & Knowledge Engineering*, **70**, 284-306.

Krishna, K. & Murty, M.N. (1999) Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics*, pp. 433-439.

Lambert, J.M. & Williams, W.T. (1962) Multivariate methods in plant ecology. IV. Nodal analysis. *Journal of Ecology*, **50**, 775–802.

Legendre, P. & Legendre, L. (1998) *Numerical ecology, 2nd English edition.* Elsevier, Amsterdam.

Madeira, S.C. & Oliveira, A.L. (2004) Biclustering algorithms for biological data analysis: A survey. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, **1**, 24-45.

Manning, C.D., Raghavan, P. & Schütze, H. (2008) *Introduction to information retrieval.* Cambridge University Press, Cambridge.

Martella, F. & Vichi, M. (2012) Clustering microarray data using model-based double K-means. *Journal of Applied Statistics*, **39**, 1853-1869.

Mirkin, B. (1996) *Mathematical classification and clustering.* Kluwer Academic Publishers, Dordrecht.

Mueller-Dombois, D. & Ellenberg, H. (1974) *Aims and methods of vegetation ecology.* John Wiley & Sons, New York.

Orloci, L. (1967) An agglomerative method for classification of plant communities. *Journal of Ecology*, **55**, 193-206.

Preston, C.D., Harrower, C.A. & Hill, M.O. (2011) Distribution patterns in British and Irish liverworts and hornworts. *Journal of Bryology*, **33**, 3-16.

Preston, C.D., Hill, M.O., Porley, R.D. & Bosanquet, S.D.S. (2010) Survey of the bryophytes of arable land in Britain and Ireland 1: a classification of arable field assemblages. *Journal of Bryology*, **32**, 61-79.

Preston, C.D., Pearman, D.A. & Dines, T.D. (2002) *New atlas of the British and Irish Flora.* Oxford University Press, Oxford.

Schepers, J. & Van Mechelen, I. (2011) A two-mode clustering method to capture the nature of the dominant interaction pattern in large profile data matrices. *Psychological Methods*, **16**, 361-371.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry.* W.H. Freeman, New York.

ter Braak, C.J.F. & Šmilauer, P. (1998) *CANOCO reference manual and user's guide to Canoco for Windows: software for canonical community ordination (version 4).* Microcomputer Power, Ithaca, NY.

ter Braak, C.J.F., Kourmpetis, Y., Kiers, H.A.L. & Bink, M.C.A.M. (2009) Approximating a similarity matrix by a latent class model: a reappraisal of additive fuzzy clustering. *Computational Statistics and Data Analysis,* **53**, 3183-3193.

Van Mechelen, I., Bock, H.-H. & De Boeck, P. (2004) Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research,* **13**, 363-394.

Vinh, N.X. (2008) Gene clustering on the unit hypersphere with the spherical k-means algorithm: coping with extremely large number of local optima. *Proceedings of the 2008 International Conference on Bioinformatics and Computational Biology (BIOCOMP), Las Vegas, 14-17 Jul 2008,* pp. 226-233.

TABLES

There are 4 tables

	Dune	Danube	Arable	Liverwort	Vascular
Area sampled	Netherlands, Terschelling	Germany, E of Ulm	Britain and Ireland	British Isles and Channel Islands	British Isles and Channel Islands
Data type	Abundance class	Biomass %	Abundance class	Presence-absence	Presence-absence
Sample units	2 x 2 m quadrats	Meadows	Arable fields	10 x 10 km squares	10 x 10 km squares
Number of species	30	94	164	300	1405
Number of samples	20	25	812	3459	3857
Number of non-zero items	197	788	11003	116973	1510290
Number of species clusters	7	6	9	10	20
Number of sample clusters	5	8	12	12	24

Table 1. Five datasets studied in detail, and the number of clusters into which they were grouped; abundance classes for the Dune and Arable datasets used the van der Maarel and DAFOR scales respectively.

Analysis type	Dune	Danube	Arable	Liverwort	Vascular
CSKM W00	1.45994	1.49856	1.11625	1.17084	
CSKM W0	1.44953	1.54587	1.14634	1.18142	
CSKM W0.5	1.50525	1.56026	1.19799	1.20960	1.16624
CSKM W1	1.51112	**1.56470**	**1.22544**	1.22135	1.17142
CSKM Transposed W00	1.42943	1.49001	1.18576	1.18150	
CSKM Transposed W0	1.51533	1.51884	1.20869	1.20149	
CSKM Transposed W0.5	1.51533	1.51999	1.21039	1.21359	
CSKM Transposed W1	1.51575	1.49826	1.21362	1.22167	
PSKM W00	1.49113	1.41471	1.15671	1.19455	
PSKM W0	1.50812	1.54737	1.18087	1.20631	
PSKM W0.5	1.50566	1.56142	1.21264	1.22513	1.16852
PSKM W1	1.51112	1.54140	1.22415	**1.22743**	**1.17321**
PSKM Transposed W00	1.42943	1.46029	1.16147	1.20285	
PSKM Transposed W0	1.51533	1.51842	1.21127	1.21437	
PSKM Transposed W0.5	1.51533	1.48263	1.20934	1.21829	
PSKM Transposed W1	1.51575	1.49668	1.20983	1.22495	
Twinspan	1.46822	1.51006	1.18407	1.15397	
Twinspan Transposed	1.36876	1.37165	1.10600	1.16533	
Clustaspec	1.50423	1.54139	1.17147	1.19633	
Clustaspec Transposed	**1.51952**	1.42099	1.19820	1.21366	

Table 2. Concentration ratios for biclustering by various clustering methods. CSKM and PSKM are chord and perpendicular spherical k-means respectively; W00, W0, W0.5 and W1 are differing weighting schemes. Maximum values are shown in bold type.

(a)

Cluster	1	2	4	6	3	7	5	8	9	11	12	10
1	469	97	174	263	51	29	7	48	49	17	11	8
5	694	541	1162	1741	233	43	67	264	1102	575	219	416
2	172	472	336	400	81	5	48	28	116	79	9	173
4	32	35	49	50	37	8	77	8	6	4	8	4
3	263	231	237	351	299	143	24	246	160	112	87	108
6	53	8	8	51	11	137	1	88	57	29	41	6
7	45	92	82	206	25	17	30	50	200	279	35	534
8	180	137	255	1026	88	81	80	363	1099	1303	489	527
9	6	1	13	50	1	15	0	19	54	93	238	17

(b)

Cluster	1	2	4	6	3	7	5	8	9	11	12	10
1	4.21	1.03	1.29	1.09	1.06	1.04	0.36	0.74	0.30	0.12	0.17	0.08
5	1.08	1.00	1.49	1.25	0.84	0.27	0.60	0.71	1.15	0.69	0.57	0.69
2	0.98	3.20	1.59	1.06	1.07	0.11	1.57	0.28	0.45	0.35	0.09	1.06
4	1.10	1.43	1.40	0.80	2.96	1.11	15.22	0.47	0.14	0.11	0.46	0.15
3	1.28	1.33	0.95	0.79	3.36	2.78	0.67	2.05	0.52	0.42	0.71	0.56
6	1.19	0.21	0.15	0.53	0.57	12.28	0.13	3.39	0.86	0.50	1.55	0.14
7	0.31	0.75	0.47	0.66	0.40	0.47	1.18	0.59	0.93	1.47	0.41	3.92
8	0.35	0.32	0.41	0.93	0.40	0.63	0.89	1.22	1.44	1.95	1.60	1.10
9	0.13	0.03	0.23	0.50	0.05	1.30	0.00	0.71	0.79	1.55	8.67	0.39

Table 3. Arable bryophyte data, showing (a) bicluster totals and (b) individual cell concentrations (observed/expected) for the standard CSKM biclustering. Rows are species clusters; columns are sample clusters. The mean concentration for the whole biclustering is 1.23, which is the weighted geometric mean of the individual cell concentrations in (b), weighted by the totals in (a).

Analysis	Concentration	Max cell concentration	CV spec	CV samp	Min spec	Max spec	Min samp	Max samp
CSKM W1	1.221	11.5	0.37	0.48	17	51	82	559
PSKM W1	1.227	11.6	0.38	0.37	13	51	74	439
CSKM W0.5	1.210	13.6	0.39	0.70	11	51	49	695
PSKM W0.5	1.225	15.4	0.25	0.61	19	40	46	635
CSKM W0	1.181	52.1	0.67	0.90	10	73	16	790
PSKM W0	1.206	53.3	0.50	0.94	10	57	21	793
Twinspan	1.154	21.4	0.65	1.70	13	77	5	1402
Clustaspec	1.196	17.1	0.56	0.38	10	66	81	526

Table 4. Liverwort cluster size in relation to concentration of biclustering; CV is coefficient of variation in cluster size, spec refers to species cluster size, samp to sample cluster size.

CAPTIONS FOR THE FIGURES

Figure 1. The two types of spherical k-means. Vector **x** is the centre of the cluster and $\mathbf{b}_i$ is a member of the cluster. In ordinary (chord) SKM we minimize the sum of squared chord distances $\sum \left\| \mathrm{BC} \right\|^2$, while in perpendicular SKM we minimize the sum of squared perpendicular distances $\sum \left\| \mathrm{BA} \right\|^2$.

Figure 2. Dune dataset, showing (a) the standard W1 solution resulting from both CSKM and PSKM (concentration ratio 1.51) and (b) the simplified solution with minimum quasi-AIC (concentration ratio 1.44). Species names are abbreviated as in ter Braak & Šmilauer (1998).

Figure 3. Danube Meadow dataset with biclustering by CSKM, concentration ratio 1.56. Values are biomass %. Species with average biomass less than 0.4% of the total have been omitted. The symbol + indicates presence but with less than 0.5% of the biomass in that sample.

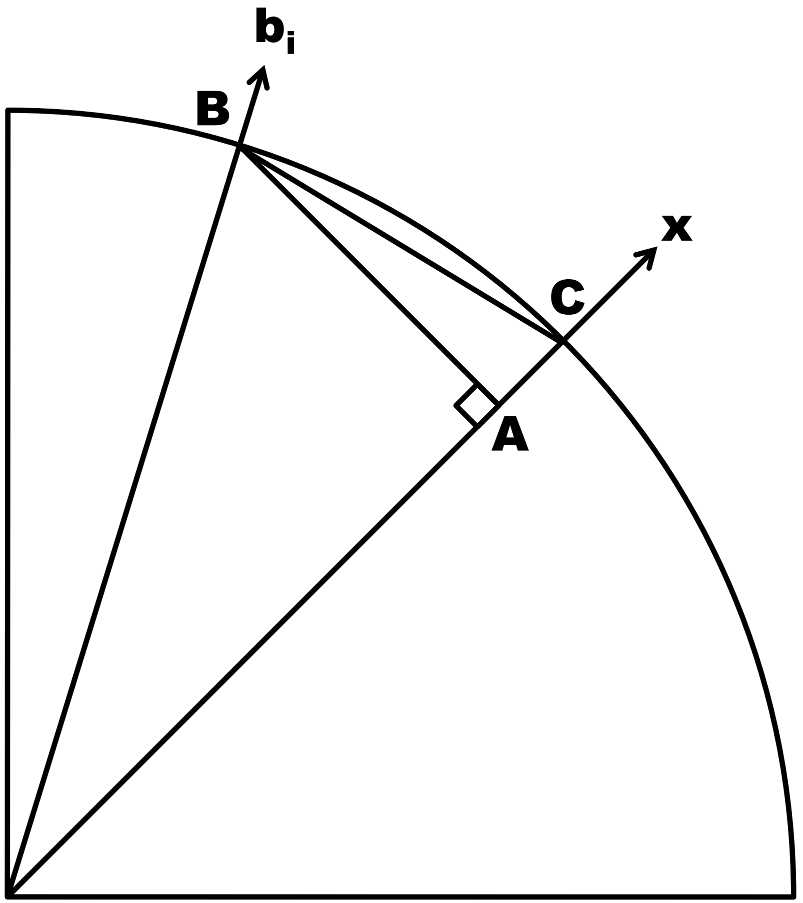


Figure 1

Figure 2

(a)

	7	6	10	5	17	2	1	18	11	19	4	9	13	3	12	15	14	20	16	8
Pla lan	5	5	3	5	2			3	3											
Rum ace	3	6		5								2			2					
Tri pra	2	5		2																
Ant odo	2	3	4	4	4					4										
Ach mil	2	2	4	2	2	3	1													
Hyp rad					2				2	5										
Air pra					2					3										
Emp nig										2										
Lol per	6	6	6	2		5	7	2	7		5	2		6						4
Poa pra	4	3	4	2	1	4	4	3	4		4	4	2	5						4
Poa tri	5	4	4	6		7	2				5	5	9	6	4				2	4
Ely rep				4		4	4				4	6		4						
Bel per			2	2		3		2			2			2						
Bro hor	2		4	2		4					3									
Leo aut	3	3	3	3	2	5		5	5	6	2	2	2	2	2	2	2	2		3
Bra rut	2	6	2	2				6	4	3	2	2		2	4	4		4	4	2
Tri rep	2	5	6	2		5		2	3	2	1	3	2	2	3	1	6			2
Vic lat			1					1	2											
Sal rep								3		3								5		
Alo gen						2					2	3	5	7	8				4	5
Agr sto											8	3	5	4	4	4	4	5	7	4
Sag pro									2	3	5	2	2		4					2
Jun buf	2											4	3		4					
Cir arv											2									
Che alb													1							
Ele pal																5	4	4	8	4
Ran fla													2			2	2	4	2	2
Jun art												4				3		4	3	4
Cal cus																	4	3	3	
Pot pal																2	2			

(b)

	10	7	5	2	6	1	18	19	11	17	9	13	4	3	8	12	15	14	20	16
Pla lan	3	5	5		5		3		3	2										
Rum ace		3	5		6						2					2				
Tri pra		2	2		5															
Ant odo	4	2	4		3			4		4										
Ach mil	4	2	2	3	2	1				2										
Leo aut	3	3	3	5	3		5	6	5	2	2	2	2	2	3	2	2	2	2	
Bra rut	2	2	2		6		6	3	4		2		2	2	2	4	4		4	4
Tri rep	6	2	2	5	5		2	2	3		3	2	1	2	2	3	1	6		
Hyp rad								5	2	2										
Sal rep							3	3											5	
Vic lat	1						1		2											
Emp nig								2												
Air pra								3		2										
Poa pra	4	4	2	4	3	4	3		4	1	4	2	4	5	4					
Lol per	6	6	2	5	6	7	2		7		2		5	6	4					
Poa tri	4	5	6	7	4	2					5	9	5	6	4	4				2
Ely rep			4	4		4					6		4	4						
Bel per	2		2	3			2						2	2						
Bro hor	4	2	2	4									3							
Alo gen				2							3	5	2	7	5	8				4
Agr sto											3	5	8	4	4	4	4	4	5	7
Sag pro								3	2		2	2	5		2	4				
Jun buf		2									4	3				4				
Cir arv													2							
Che alb												1								
Ele pal															4		5	4	4	8
Ran fla												2			2		2	2	4	2
Jun art											4				4		3		4	3
Cal cus																		4	3	3
Pot pal																	2	2		

	1	4	9	3	10	15	2	12	24	23	22	5	17	16	20	25	18	21	19	14	6	13	8	11	7
Brom erec	50	74	47	35	21	37			10																
Koel pyra	3	2	3		3																				
Fest rubr	15	2	3		4	6		2	2	+		+	2		2				+			1	+	2	
Camp rotu	1	1	+	+	1	1		1	+	1				+		+				1			+		
Fest ovin	2		1		2		1																		
Care flac	2	3				1							2	3						2					
Salv prat			2	4	5	1			4																
Poa prat	4	5	10	10	8	15	74	25	20	16	9	4	5	10	6	10	10	1	1	5	2	2	4	6	3
Gali moll	3	1	2	7	12	3	2	6	3	12	3	3	3	5	5	2	6	1	2	14	24	10	4	5	6
Ranu acri	+		+	+	3	+	+	+	+	+	2	1	+	+	1	+		+	+	1	2	2	+	1	2
Plan lanc	1	1	1	1	1	1	1	1	+	8	2	2	+	2	1	+	1	1	+	2	1	4	4	+	4
Achi mill	6	1	2	3	8	2	+	1	3	+	4	2	+	5	6	+	12	+		16		4			1
Leuc vulg	1	+	2	3	5	1	+	3	6	1	1	+	1	1	3			+	+	2	4	1	1	+	2
Tara offi	+	+		+	+	+		+	+	4	+	3	+		+	+	1			+	1	2	1		3
Cent jace	+			6		+	1	1	2	+			2	2	2			+		4	2				3
Hera spho	+			1					+	1	+	3	1	+	1	+	+				+	26			4
Arrh elat	+	2	5	2	8	15	1	10	30	26	25	4	15	15	24	35	22	25	22	12	4	10	4	9	4
Dact glom	5	5	2	15	6	6	5	32	8	8	8	12	15	15	5	18	18	18	1	10	12	15	10	12	4
Fest prat			5	3	2	6	5	10	2	2	15	20	28	5	2	3	12	15	10	2	3	2	8	10	2
Tris flav			3	5		4	2		2	4	16	8	5	10	8		5	8				6			
Vero cham	+	+	+	+	1	+		1	+	2	+	+	+	1	1	1		1	+		1	+	+	+	1
Heli pube	1		4		+		1	1	1		2	20	4	13	28		4	+			8	4			3
Trif prat	+	+		1		+	+		2	2	+	4		1	+		+	+	+		1	+	+		1
Holc lana										+		1	2			15		2	+	2	1	1	2	2	1
Geum riva								+			+		+	1	+	+	+	1	1		2	+	5	3	1
Rume acsa	+	+				+		1		+	1	+	+	1	1	1		+	1	2	2	1	3	2	1
Cirs oler										1	+	+	2		+	3		2	+	18	12	+	20	3	20
Desc cesp													10					5	2	1			2	28	11
Alop prat											10	2				1	2	15			8	4	10	6	
Care acfm																2		1	2					4	10
Dauc caro	1	+	1	1	1	+	1	+		2	+	+	+	+	1		1	+			1	2	5		1
Crep bien				+			+	1	+	6	1	2	+		1		1	+		+	1		8	+	
Glyc flui																			20						
Phal arun																			28						

Figure 3

SUPPORTING INFORMATION

Danube meadow dataset

Figure S1 (a)-(d). CSKM and PSKM biclusterings for Danube Meadow Data. Colours show the concentration of the individual biclusters, from dark green (> 3.5) to light green (>1.0) and grey (>0.5). Solutions for k_1=6, k_2=9 show the greater tendency of PSKM to pick out aberrant species and samples; here, a sample with 28% *Deschampsia cespitosa* is picked out by PSKM but not by CSKM. The best solution according to the quasi-Akaike measure is PSKM with k_1= k_2=5.

Fig. S1 (a) CSKM, k_1 = 5, k_2 =5, concentration ratio 1.46

	1	1	1	1	1	1	2	4	4	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	
	9	15	1	3	10	4	2	11	7	8	22	25	17	18	23	16	13	20	5	14	21	12	6	24	19	
Brom_erec	47	37	50	35	21	74																		10		1
Fest_rubr	3	6	15		4	2		2		+			2		+		1	2	+			2		2	+	1
Salv_prat	2	1		4	5																			4		1
Care_flac		1	2			3							2			3				2						1
Koel_pyra	3		3		3	2																				1
Poa__prat	10	15	4	10	8	5	74	6	3	4	9	10	5	10	16	10	2	6	4	5	1	25	2	20	1	2
Arrh_elat	5	15	+	2	8	2	1	9	4	4	25	35	15	22	26	15	10	24	4	12	25	10	4	30	22	3
Dact_glom	2	6	5	15	6	5	5	12	4	10	8	18	15	18	8	15	15	5	12	10	18	32	12	8	1	3
Fest_prat	5	6		3	2		5	10	2	8	15	3	28	12	2	5	2	2	20	2	15	10	3	2	10	3
Gali_moll	2	3	3	7	12	1	2	5	6	4	3	2	3	6	12	5	10	5	3	14	1	6	24	3	2	3
Heli_pube	4		1		+		1		3		2		4	4		13	4	28	20		+	1	8	1		3
Tris_flav	3	4		5			2				16		5	5	4	10	6	8	8		8			2	5	3
Achi_mill	2	2	6	3	8	1	+		1		4	+	+	12	+	5	4	6	2	16	+	1		3		3
Plan_lanc	1	1	1	1	1	1	1	+	4	4	2	+	+	1	8	2	4	1	2	2	1	1	1	+	+	3
Leuc_vulg	2	1	1	3	5	+	+	+	2	1	1		1		1	1	1	3	+	2	+	3	4	6	+	3
Hera_spho			+	1					4		+	+	1	+	1	+	26	1	3				+	+		3
Holc_lana								2		2		15	2		+		1		1	2	2		1		+	3
Cent_jace		+	+	6			1		3				2		+	2		2		4	+	1	2	2		3
Ranu_acri	+	+	+	+	3		+	1	2	+	2	+	+		+	+	2	1	1	1	+	+	2	+	+	3
Tara_offi		+	+	+	+	+			3	1	+	+	+	1	4		2	+	3	+		+	1	+		3
Trif_prat		+	+	1		+	+		1	+	+			+	2	1	+	+	4		+		1	2	+	3
Vero_cham	+	+	+	+	1	+		+	1	+	+	1	+		2	1	+	1	+		1	1	1	+	+	3
Sile_dioi									1		+	4	+			1	+			1	+	+	2		+	3
Cirs_oler								3	20	20	+	3	2		1		+	+	+	18	2		12		+	4
Desc_cesp								28	11	2			10							1	5				2	4
Alop_prat								6		10	10	1		2			4		2		15		8			4
Crep_bien				+			+	+		8	1		+	1	6			1	2	+	+	1	1	+		4
Care_acfm								4	10			2									1				2	4
Dauc_caro	1	+	1	1	1	+	1		1	5	+		+	1	2	+	2	1	+		+	+	1			4
Rume_acsa		+	+			+		2	1	3	1	1	+		+	1	1	1	+	2	+	1	2		1	4
Geum_riva								3	1	5	+	+	+	+		1	+	+			1	+	2		1	4
Ajug_rept	1						+	1	1	3	+	+	+	+			+	+	+	+	1	1	+			4
Phal_arun																									28	5
Glyc_flui																									20	5

S1 (b) PSKM, k_1 = 5, k_2 =5, concentration ratio 1.47 (selected by quasi-AIC)

	1	1	1	1	1	1	2	3	3	3	3	3	3	3	3	3	3	5	4	4	4	4	4	4	4	
	9	1	15	3	10	4	2	16	18	20	22	5	25	17	21	12	24	19	6	11	13	14	8	7	23	
Brom_erec	47	50	37	35	21	74											10									1
Koel_pyra	3	3			3	2																				1
Fest_rubr	3	15	6		4	2				2		+		2		2	2	+		2	1		+		+	1
Care_flac		2	1			3		3						2								2				1
Salv_prat	2		1	4	5												4									1
Poa__prat	10	4	15	10	8	5	74	10	10	6	9	4	10	5	1	25	20	1	2	6	2	5	4	3	16	2
Arrh_elat	5	+	15	2	8	2	1	15	22	24	25	4	35	15	25	10	30	22	4	9	10	12	4	4	26	3
Dact_glom	2	5	6	15	6	5	5	15	18	5	8	12	18	15	18	32	8	1	12	12	15	10	10	4	8	3
Fest_prat	5		6	3	2		5	5	12	2	15	20	3	28	15	10	2	10	3	10	2	2	8	2	2	3
Tris_flav	3		4	5			2	10	5	8	16	8		5	8		2	5			6				4	3
Vero_cham	+	+	+	+	1	+		1		1	+	+	1	+	1	1	+	+	1	+	+		+	1	2	3
Leuc_vulg	2	1	1	3	5	+	+	1		3	1	+		1	+	3	6	+	4	+	1	2	1	2	1	3
Achi_mill	2	6	2	3	8	1	+	5	12	6	4	2	+	+	+	1	3				4	16		1	+	3
Heli_pube	4	1			+		1	13	4	28	2	20		4	+	1	1		8		4			3		3
Trif_prat		+	+	1		+	+	1	+	+	+	4			+		2	+	1		+		+	1	2	3
Holc_lana												1	15	2	2			+	1	2	1	2	2		+	3
Sile_dioi								1			+		4	+	+	+		+	2		+	1		1		3
Glyc_flui																		20								5
Phal_arun																		28								5
Rume_acsa		+	+			+		1		1	1	+	1	+	+	1		1	2	2	1	2	3	1	+	4
Cirs_oler										+	+	+	3	2	2			+	12	3	+	18	20	20	1	4
Gali_moll	2	3	3	7	12	1	2	5	6	5	3	3	2	3	1	6	3	2	24	5	10	14	4	6	12	4
Plan_lanc	1	1	1	1	1	1	1	2	1	1	2	2	+	+	1	1	+	+	1	+	4	2	4	4	8	4
Geum_riva								1	+	+	+		+	+	1	+		1	2	3	+		5	1		4
Dauc_caro	1	1	+	1	1	+	1	+	1	1	+	+		+	+	+			1		2		5	1	2	4
Ranu_acri	+	+	+	+	3		+	+		1	2	1	+	+	+	+	+	+	2	1	2	1	+	2	+	4
Ajug_rept	1						+		+	+	+	+	+	+	1	1			+	1	+	+	3	1		4
Tara_offi		+	+	+	+	+			1	+	+	3	+	+		+	+		1		2	+	1	3	4	4
Alop_prat									2		10	2	1		15				8	6	4		10			4
Crep_bien				+			+		1	1	1	2		+	+	1	+		1	+		+	8		6	4
Care_acfm													2		1			2		4				10		4
Desc_cesp														10	5			2		28		1	2	11		4
Cent_jace		+	+	6			1	2		2				2	+	1	2		2			4		3	+	4
Hera_spho		+		1				+	+	1	+	3	+	1			+		+		26			4	1	4

S1 (c) CSKM, k_1 = 6, k_2 = 8, concentration ratio 1.56

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	
	1	1	1	2	2	2	3	4	4	4	6	6	6	6	6	6	6	6	8	5	5	5	7	7	7	
Species___	1	4	9	3	10	15	2	12	24	23	22	5	17	16	20	25	18	21	19	14	6	13	8	11	7	
Brom_erec	50	74	47	35	21	37			10																	1
Koel_pyra	3	2	3		3																					1
Fest_rubr	15	2	3		4	6		2	2	+		+	2		2				+			1	+	2		1
Camp_rotu	1	1	+	+	1	1		1	+	1				+		+				1			+			1
Fest_ovin	2		1		2		1																			1
Care_flac	2	3				1							2	3						2						1
Salv_prat			2	4	5	1			4																	1
Poa__prat	4	5	10	10	8	15	74	25	20	16	9	4	5	10	6	10	10	1	1	5	2	2	4	6	3	2
Gali_moll	3	1	2	7	12	3	2	6	3	12	3	3	3	5	5	2	6	1	2	14	24	10	4	5	6	3
Ranu_acri	+		+	+	3	+	+	+	+	+	2	1	+	+	1	+		+	+	1	2	2	+	1	2	3
Plan_lanc	1	1	1	1	1	1	1	1	+	8	2	2	+	2	1	+	1	1	+	2	1	4	4	+	4	3
Achi_mill	6	1	2	3	8	2	+	1	3	+	4	2	+	5	6	+	12	+		16		4			1	3
Leuc_vulg	1	+	2	3	5	1	+	3	6	1	1	+	1	1	3			+	+	2	4	1	1	+	2	3
Tara_offi	+	+		+	+	+		+	+	4	+	3	+		+	+	1			+	1	2	1		3	3
Cent_jace	+			6		+	1	1	2	+			2	2	2			+		4	2				3	3
Hera_spho	+			1					+	1	+	3	1	+	1	+	+				+	26			4	3
Arrh_elat	+	2	5	2	8	15	1	10	30	26	25	4	15	15	24	35	22	25	22	12	4	10	4	9	4	4
Dact_glom	5	5	2	15	6	6	5	32	8	8	8	12	15	15	5	18	18	18	1	10	12	15	10	12	4	4
Fest_prat			5	3	2	6	5	10	2	2	15	20	28	5	2	3	12	15	10	2	3	2	8	10	2	4
Tris_flav			3	5		4	2		2	4	16	8	5	10	8		5	8				6				4
Vero_cham	+	+	+	+	1	+		1	+	2	+	+	+	1	1	1		1	+		1	+	+	+	1	4
Heli_pube	1		4		+		1	1	1		2	20	4	13	28		4	+			8	4			3	4
Trif_prat	+	+		1		+	+		2	2	+	4		1	+		+	+	+		1	+	+		1	4
Holc_lana										+		1	2			15		2	+	2	1	1	2	2	1	4
Geum_riva								+			+		+	1	+	+	+	1	1		2	+	5	3	1	5
Rume_acsa	+	+				+		1		+	1	+	+	1	1	1		+	1	2	2	1	3	2	1	5
Cirs_oler										1	+	+	2		+	3		2	+	18	12	+	20	3	20	5
Desc_cesp													10					5	2	1			2	28	11	5
Alop_prat											10	2				1	2	15			8	4	10	6		5
Care_acfm																2		1	2					4	10	5
Dauc_caro	1	+	1	1	1	+	1	+		2	+	+	+	+	1		1	+			1	2	5		1	5
Crep_bien				+			+	1	+	6	1	2	+		1		1	+		+	1		8	+		5
Glyc_flui																			20							6
Phal_arun																			28							6

S1 (d) PSKM, k_1 = 6, k_2 = 8, concentration ratio 1.54

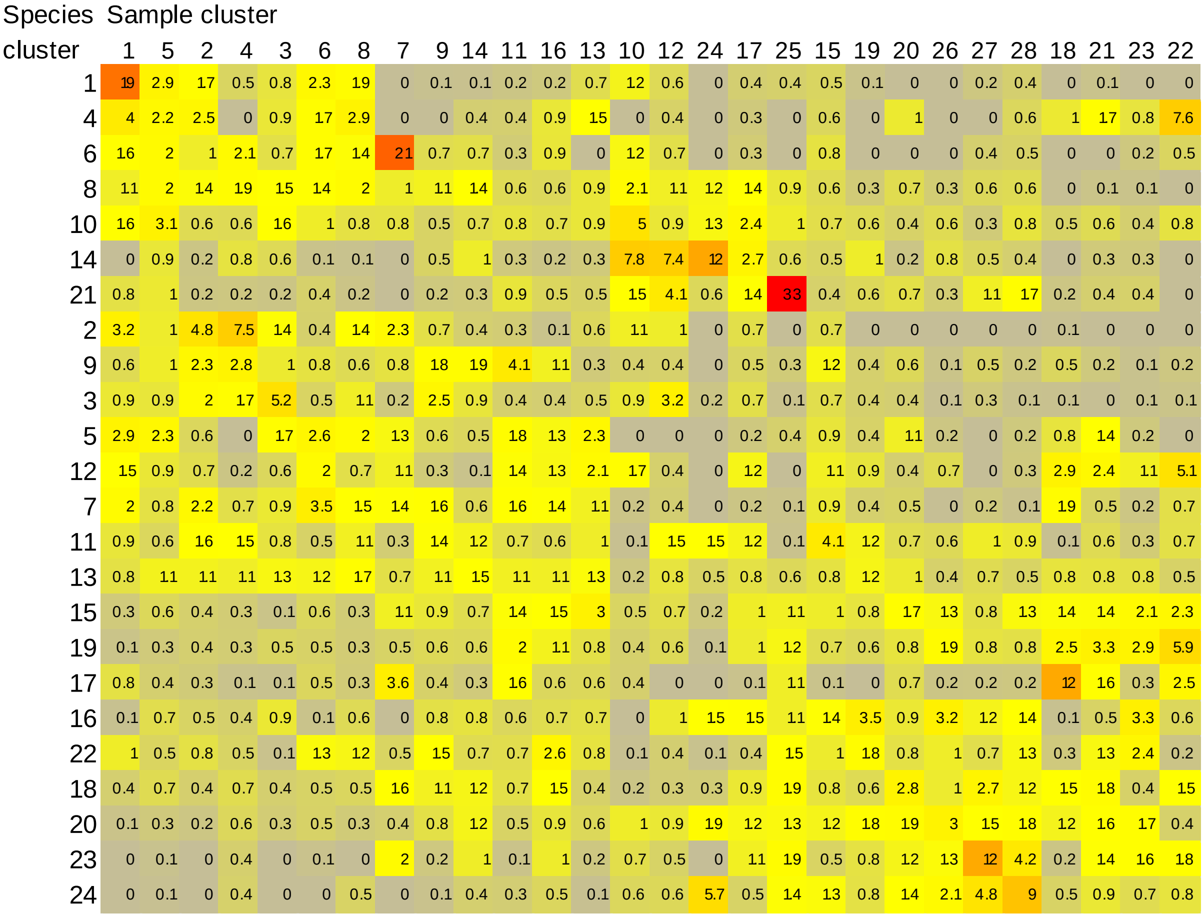
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	
	1	1	1	2	2	2	3	4	4	5	5	5	5	5	5	5	5	7	8	6	6	6	6	6	6	
Species___	1	4	9	3	10	15	2	12	24	22	5	16	25	18	20	17	21	11	19	8	6	13	14	23	7	
Brom_erec	50	74	47	35	21	37			10																	1
Koel_pyra	3	2	3		3																					1
Fest_rubr	15	2	3		4	6		2	2		+				2	2		2	+	+		1		+		1
Camp_rotu	1	1	+	+	1	1		1	+			+	+							+			1	1		1
Fest_ovin	2		1		2		1																			1
Care_flac	2	3				1						3				2							2			1
Salv_prat			2	4	5	1			4																	1
Poa__prat	4	5	10	10	8	15	74	25	20	9	4	10	10	10	6	5	1	6	1	4	2	2	5	16	3	2
Gali_moll	3	1	2	7	12	3	2	6	3	3	3	5	2	6	5	3	1	5	2	4	24	10	14	12	6	3
Rume_acsa	+	+				+		1		1	+	1	1		1	+	+	2	1	3	2	1	2	+	1	3
Cirs_oler										+	+		3		+	2	2	3	+	20	12	+	18	1	20	3
Plan_lanc	1	1	1	1	1	1	1	1	+	2	2	2	+	1	1	+	1	+	+	4	1	4	2	8	4	3
Ranu_acri	+		+	+	3	+	+	+	+	2	1	+	+		1	+	+	1	+	+	2	2	1	+	2	3
Dauc_caro	1	+	1	1	1	+	1	+		+	+	+		1	1	+	+			5	1	2		2	1	3
Leuc_vulg	1	+	2	3	5	1	+	3	6	1	+	1			3	1	+	+	+	1	4	1	2	1	2	3
Achi_mill	6	1	2	3	8	2	+	1	3	4	2	5	+	12	6	+	+					4	16	+	1	3
Tara_offi	+	+		+	+	+		+	+	+	3		+	1	+	+				1	1	2	+	4	3	3
Cent_jace	+			6		+	1	1	2			2			2	2	+				2		4	+	3	3
Crep_bien				+			+	1	+	1	2			1	1	+	+	+		8	1		+	6		3
Hera_spho	+			1					+	+	3	+	+	+	1	1					+	26		1	4	3
Desc_cesp																10	5	28	2	2			1		11	5
Care_acfm													2				1	4	2						10	5
Geum_riva								+		+		1	+	+	+	+	1	3	1	5	2	+			1	5
Arrh_elat	+	2	5	2	8	15	1	10	30	25	4	15	35	22	24	15	25	9	22	4	4	10	12	26	4	4
Dact_glom	5	5	2	15	6	6	5	32	8	8	12	15	18	18	5	15	18	12	1	10	12	15	10	8	4	4
Fest_prat			5	3	2	6	5	10	2	15	20	5	3	12	2	28	15	10	10	8	3	2	2	2	2	4
Tris_flav			3	5		4	2		2	16	8	10		5	8	5	8					6		4		4
Vero_cham	+	+	+	+	1	+		1	+	+	+	1	1		1	+	1	+	+	+	1	+		2	1	4
Alop_prat										10	2		1	2			15	6		10	8	4				4
Heli_pube	1		4		+		1	1	1	2	20	13		4	28	4	+				8	4			3	4
Trif_prat	+	+		1		+	+		2	+	4	1		+	+		+		+	+	1	+		2	1	4
Holc_lana											1		15			2	2	2	+	2	1	1	2	+	1	4
Glyc_flui																			20							6
Phal_arun																			28							6

Arable bryophyte dataset

The concentration matrix for the arable bryophyte data, classified following the quasi-Akaike criterion into 24 species clusters and 28 sample clusters, is shown in Fig. S2. It shows some marked local concentrations. The most extreme of these, species cluster 21, with key species *Didymodon luridus*, is heavily concentrated in sample cluster 25, which has only 6 samples, all in the Cotswolds (a small area of England with Jurassic limestone bedrock). The next most extreme example is species cluster 6, with key species *Didymodon insulanus*. It is concentrated in just 8 samples, from bulb fields (not arable in the ordinary sense) in the extreme southwest of England.

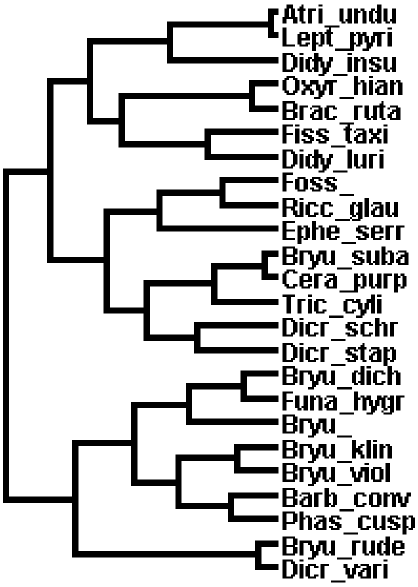
The hierarchy for the 24 species clusters is shown in Fig. S3. Each cluster is named by its key species, whose name is followed by a brief description of the cluster characteristics. The top-level division comes between clusters 13, which is a very widespread, and 15, which comprises species that are characteristic of eutrophic and calcareous soil. Note that the bottom left and top right of the concentration matrix are generally little occupied. The striking exception is cluster 4, characterized by *Leptobryum pyriforme*, in sample cluster 22, towards the top right. The distribution of *L. pyriforme* in arable fields is genuinely very odd. It occurs both in weedy, often sandy communities such as cluster 22, and on disturbed peat. Cluster 22 is visually conspicuous in the concentration matrix, but consists of only 10 samples.

Figure S2. Arable bryophyte cluster concentration ratios, with 24 species clusters and 28 sample clusters.



Species	Sample cluster																											
cluster	1	5	2	4	3	6	8	7	9	14	11	16	13	10	12	24	17	25	15	19	20	26	27	28	18	21	23	22
1	19	2.9	17	0.5	0.8	2.3	19	0	0.1	0.1	0.2	0.2	0.7	12	0.6	0	0.4	0.4	0.5	0.1	0	0	0.2	0.4	0	0.1	0	0
4	4	2.2	2.5	0	0.9	17	2.9	0	0	0.4	0.4	0.9	15	0	0.4	0	0.3	0	0.6	0	1	0	0	0.6	1	17	0.8	7.6
6	16	2	1	2.1	0.7	17	14	21	0.7	0.7	0.3	0.9	0	12	0.7	0	0.3	0	0.8	0	0	0	0.4	0.5	0	0	0.2	0.5
8	11	2	14	19	15	14	2	1	11	14	0.6	0.6	0.9	2.1	11	12	14	0.9	0.6	0.3	0.7	0.3	0.6	0.6	0	0.1	0.1	0
10	16	3.1	0.6	0.6	16	1	0.8	0.8	0.5	0.7	0.8	0.7	0.9	5	0.9	13	2.4	1	0.7	0.6	0.4	0.6	0.3	0.8	0.5	0.6	0.4	0.8
14	0	0.9	0.2	0.8	0.6	0.1	0.1	0	0.5	1	0.3	0.2	0.3	7.8	7.4	12	2.7	0.6	0.5	1	0.2	0.8	0.5	0.4	0	0.3	0.3	0
21	0.8	1	0.2	0.2	0.2	0.4	0.2	0	0.2	0.3	0.9	0.5	0.5	15	4.1	0.6	14	33	0.4	0.6	0.7	0.3	11	17	0.2	0.4	0.4	0
2	3.2	1	4.8	7.5	14	0.4	14	2.3	0.7	0.4	0.3	0.1	0.6	11	1	0	0.7	0	0.7	0	0	0	0	0	0.1	0	0	0
9	0.6	1	2.3	2.8	1	0.8	0.6	0.8	18	19	4.1	11	0.3	0.4	0.4	0	0.5	0.3	12	0.4	0.6	0.1	0.5	0.2	0.5	0.2	0.1	0.2
3	0.9	0.9	2	17	5.2	0.5	11	0.2	2.5	0.9	0.4	0.4	0.5	0.9	3.2	0.2	0.7	0.1	0.7	0.4	0.4	0.1	0.3	0.1	0.1	0	0.1	0.1
5	2.9	2.3	0.6	0	17	2.6	2	13	0.6	0.5	18	13	2.3	0	0	0	0.2	0.4	0.9	0.4	11	0.2	0	0.2	0.8	14	0.2	0
12	15	0.9	0.7	0.2	0.6	2	0.7	11	0.3	0.1	14	13	2.1	17	0.4	0	12	0	11	0.9	0.4	0.7	0	0.3	2.9	2.4	11	5.1
7	2	0.8	2.2	0.7	0.9	3.5	15	14	16	0.6	16	14	11	0.2	0.4	0	0.2	0.1	0.9	0.4	0.5	0	0.2	0.1	19	0.5	0.2	0.7
11	0.9	0.6	16	15	0.8	0.5	11	0.3	14	12	0.7	0.6	1	0.1	15	15	12	0.1	4.1	12	0.7	0.6	1	0.9	0.1	0.6	0.3	0.7
13	0.8	11	11	11	13	12	17	0.7	11	15	11	11	13	0.2	0.8	0.5	0.8	0.6	0.8	12	1	0.4	0.7	0.5	0.8	0.8	0.8	0.5
15	0.3	0.6	0.4	0.3	0.1	0.6	0.3	11	0.9	0.7	14	15	3	0.5	0.7	0.2	1	11	1	0.8	17	13	0.8	13	14	14	2.1	2.3
19	0.1	0.3	0.4	0.3	0.5	0.5	0.3	0.5	0.6	0.6	2	11	0.8	0.4	0.6	0.1	1	12	0.7	0.6	0.8	19	0.8	0.8	2.5	3.3	2.9	5.9
17	0.8	0.4	0.3	0.1	0.1	0.5	0.3	3.6	0.4	0.3	16	0.6	0.6	0.4	0	0	0.1	11	0.1	0	0.7	0.2	0.2	0.2	12	16	0.3	2.5
16	0.1	0.7	0.5	0.4	0.9	0.1	0.6	0	0.8	0.8	0.6	0.7	0.7	0	1	15	15	11	14	3.5	0.9	3.2	12	14	0.1	0.5	3.3	0.6
22	1	0.5	0.8	0.5	0.1	13	12	0.5	15	0.7	0.7	2.6	0.8	0.1	0.4	0.1	0.4	15	1	18	0.8	1	0.7	13	0.3	13	2.4	0.2
18	0.4	0.7	0.4	0.7	0.4	0.5	0.5	16	11	12	0.7	15	0.4	0.2	0.3	0.3	0.9	19	0.8	0.6	2.8	1	2.7	12	15	18	0.4	15
20	0.1	0.3	0.2	0.6	0.3	0.5	0.3	0.4	0.8	12	0.5	0.9	0.6	1	0.9	19	12	13	12	18	19	3	15	18	12	16	17	0.4
23	0	0.1	0	0.4	0	0.1	0	2	0.2	1	0.1	1	0.2	0.7	0.5	0	11	19	0.5	0.8	12	13	12	4.2	0.2	14	16	18
24	0	0.1	0	0.4	0	0	0.5	0	0.1	0.4	0.3	0.5	0.1	0.6	0.6	5.7	0.5	14	13	0.8	14	2.1	4.8	9	0.5	0.9	0.7	0.8

Figure S3. Cluster hierarchy for 24 species groups shown in Fig. S2

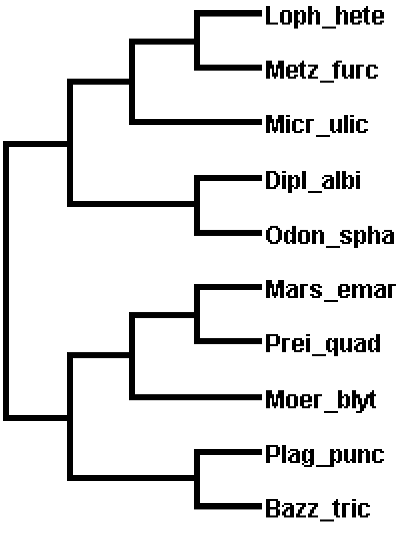


No	Key sp.	Character of group
1	Atri_undu	Peat - acid in NW
4	Lept_pyri	Peat - including fenland
6	Didy_insu	Scilly bulbfields
8	Oxyr_hian	Pleurocarp surviving plough
10	Brac_ruta	Pleurocarp, fallow and setaside
14	Fiss_taxi	Fallow, germinating from spores
21	Didy_luri	Cotswold limestone lumps
2	Foss_	Liverworts and hornworts from spores
9	Ricc_glau	Liverworts from spores, southwestern
3	Ephe_serr	Neutral soil, often clay, from spores
5	Bryu_suba	Acid from tubers
12	Cera_purp	Acid, weedy
7	Tric_cyli	Acid from tubers and spores
11	Dicr_schr	Basic and neutral from tubers
13	Dicr_stap	Very common from tubers
15	Bryu_dich	Weedy, mainly from veg propagules
19	Funa_hygr	Weedy, high nutrient
17	Bryu_	Weedy, on sandy soils, esp Norfolk
16	Bryu_klin	Very common from tubers, esp on clay
22	Bryu_viol	Very common from tubers, esp on sand
18	Barb_conv	Very common, including NW
20	Phas_cusp	Very common, not so much in NW
23	Bryu_rude	Basic soils avoiding clay
24	Dicr_vari	Basic soils including clay

Liverwort distribution dataset

The cluster hierarchy for the liverwort distribution analysis is shown in Fig. S4. This shows 10 themes that characterize liverwort distributions. Six additional themes were revealed in other analyses (Table S1). Figs S5a and S5b show the cluster totals and concentration ratios. Note the low liverwort totals in the lowland hectad clusters (to the left of the diagram in Fig. S5a).

Figure S4. Cluster hierarchy for liverwort analysis PSKM W1; each cluster is identified by a key species and a theme



Key species	Theme
Lophocolea heterophylla	Lowland
Metzgeria furcata	Ubiquitous
Microlejeunea ulicina	Southwestern
Diplophyllum albicans	Calcifuge
Odontoschisma sphagni	Bog
Marsupella emarginata	Upland
Preissia quadrata	Calcicole upland
Moerckia blyttii	Montane
Plagiochila punctata	Atlantic
Bazzania tricrenata	Northern atlantic

Theme	CSKM W1	PSKM W1	CSKM W0.5	PSKM W0.5	CSKM W0	PSKM W0	Twinspan	Clustaspec
Lowland	Loph hete	Loph hete	Loph hete		Ricc flui		Loph hete	Loph hete
Ubiquitous	Metz furc	Metz furc	Dipl albi	Metz furc	Metz furc	Metz furc	Metz furc	Pell epip
Southwestern	Phae laev	Micr ulic	Ceph stel	Ceph stel	Ceph stel	Ricc croz	Phae laev	Phae laev
Southwest coast					Ricc croz			
Calcifuge	Dipl albi	Dipl albi		Dipl albi		Dipl albi	Ceph bicu	
Bog	Odon spha	Odon spha	Odon spha	Odon spha		Odon spha		Clad flui
Middle western				Sacc viti			Leje lama	
Rather upland							Frul tama	
Upland	Mars emar	Mars emar	Trit quin	Trit quin	Mars emar	Anas orca	Mars emar	Scap undu
Calcicole upland	Colo calc	Prei quad	Colo calc				Anas minu	
Atlantic	Plag punc	Plag punc	Plag punc	Harp moll	Harp moll	Harp moll	Drep hama	Harp moll
Irish atlantic			Radu holt		Leje hibe	Radu holt		
Northern atlantic	Bazz tric	Bazz tric	Bazz tric	Anas orca	Scap orni	Scap orni		Anas orca
Calcicole montane				Jung bore	Jung bore	Trit poli		Scap dege
Montane	Moer blyt	Moer blyt	Moer blyt	Moer blyt	Mars cond	Pleu albe	Mars spha	Moer blyt
Eastern snowpatch								Mars cond

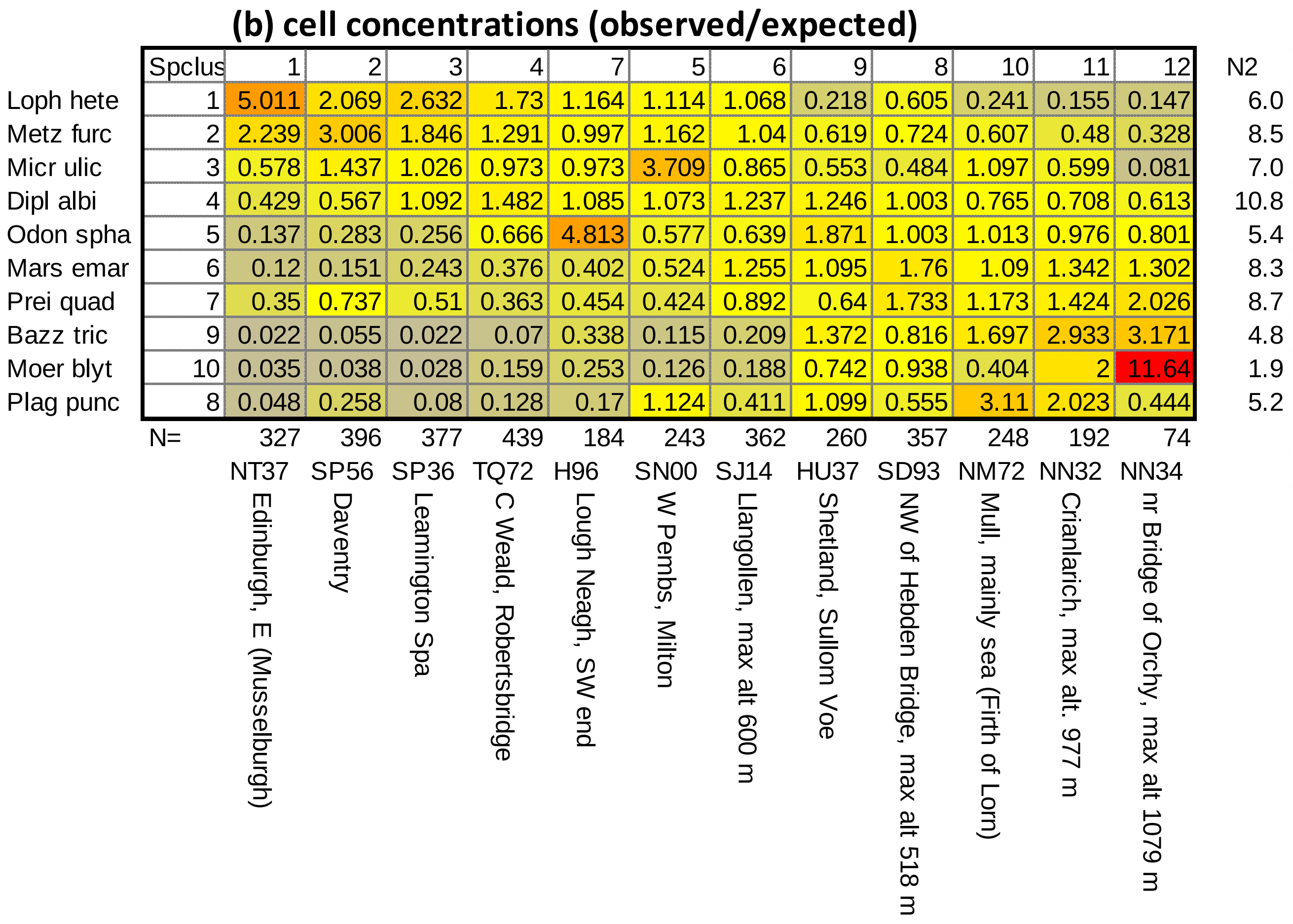
Table S1. Themes of species clusters emerging from the analyses of the liverwort dataset; full names of the key species are *Anastrophyllum minutum*, *Anastrepta orcadensis*, *Bazzania tricrenata*, *Cephalozia bicuspidata*, *Cephaloziella stellulifera*, *Cladopodiella fluitans*, *Cololejeunea calcarea*, *Diplophyllum albicans*, *Drepanolejeunea hamatifolia*, *Frullania tamarisci*, *Harpalejeunea molleri*, *Jungermannia borealis*, *Lejeunea hibernica*, *Lejeunea lamacerina*, *Lophocolea heterophylla*, *Marsupella condensata*, *Marsupella emarginata*, *Marsupella sphacelata*, *Metzgeria furcata*, *Microlejeunea ulicina*, *Moerckia blyttii*, *Odontoschisma sphagni*, *Pellia epiphylla*, *Phaeoceros laevis*, *Plagiochila punctata*, *Pleurocladula albescens*, *Preissia quadrata*, *Radula holtii*, *Riccia crozalsii*, *Riccia fluitans*, *Saccogyna viticulosa*, *Scapania degenii*, *Scapania ornithopodioides*, *Scapania undulata*, *Tritomaria polita*, *Tritomaria quinquedentata*

Liverwort - PSKM W1 Concentration ratio 1.227

(a) bicluster totals

Spclus		1	2	3	4	7	5	6	9	8	10	11	12	Total
Loph hete	1	1357	777	1781	1666	461	842	1355	170	963	282	203	52	9909
Metz furc	2	1183	2202	2437	2425	770	1713	2575	940	2248	1384	1226	226	19329
Micr ulic	3	76	262	337	455	187	1361	533	209	374	623	381	14	4812
Dipl albi	4	445	816	2833	5472	1647	3108	6020	3716	6117	3433	3555	831	37993
Odon spha	5	21	60	98	362	1076	246	458	822	901	669	722	160	5595
Mars emar	6	59	103	300	659	290	721	2900	1552	5103	2323	3201	839	18050
Prei quad	7	39	114	142	144	74	132	466	205	1135	565	768	295	4079
Bazz tric	9	4	14	10	46	91	59	180	726	883	1350	2613	763	6739
Moer blyt	10	2	3	4	32	21	20	50	121	313	99	549	863	2077
Plag punc	8	11	82	46	104	57	719	442	724	748	3080	2244	133	8390
	Total	3197	4433	7988	11365	4674	8921	14979	9185	18785	13808	15462	4176	116973
		NT37	SP56	SP36	TQ72	H96	SN00	SJ14	HU37	SD93	NM72	NN32	NN34	
		Edinburgh, E (Musselburgh)	Daventry	Leamington Spa	C Weald, Robertsbridge	Lough Neagh, SW end	W Pembs, Milton	Llangollen, max alt 600 m	Shetland, Sullom Voe	NW of Hebden Bridge, max alt 518 m	Mull, mainly sea (Firth of Lorn)	Crianlarich, max alt. 977 m	nr Bridge of Orchy, max alt 1079 m	

Figure S5(a). Bicluster totals and individual cell concentrations for Liverwort dataset, PSKM W1 analysis. Rows represent species clusters, which are named by their key species. Columns represent sample clusters, named by their key hectads, using standard naming conventions of the British and Irish National Grids. The locations of the key hectads are specified by a short phrase.



(b) cell concentrations (observed/expected)

	Spclus	1	2	3	4	7	5	6	9	8	10	11	12	N2
Loph hete	1	5.011	2.069	2.632	1.73	1.164	1.114	1.068	0.218	0.605	0.241	0.155	0.147	6.0
Metz furc	2	2.239	3.006	1.846	1.291	0.997	1.162	1.04	0.619	0.724	0.607	0.48	0.328	8.5
Micr ulic	3	0.578	1.437	1.026	0.973	0.973	3.709	0.865	0.553	0.484	1.097	0.599	0.081	7.0
Dipl albi	4	0.429	0.567	1.092	1.482	1.085	1.073	1.237	1.246	1.003	0.765	0.708	0.613	10.8
Odon spha	5	0.137	0.283	0.256	0.666	4.813	0.577	0.639	1.871	1.003	1.013	0.976	0.801	5.4
Mars emar	6	0.12	0.151	0.243	0.376	0.402	0.524	1.255	1.095	1.76	1.09	1.342	1.302	8.3
Prei quad	7	0.35	0.737	0.51	0.363	0.454	0.424	0.892	0.64	1.733	1.173	1.424	2.026	8.7
Bazz tric	9	0.022	0.055	0.022	0.07	0.338	0.115	0.209	1.372	0.816	1.697	2.933	3.171	4.8
Moer blyt	10	0.035	0.038	0.028	0.159	0.253	0.126	0.188	0.742	0.938	0.404	2	11.64	1.9
Plag punc	8	0.048	0.258	0.08	0.128	0.17	1.124	0.411	1.099	0.555	3.11	2.023	0.444	5.2
	N=	327	396	377	439	184	243	362	260	357	248	192	74	
		NT37	SP56	SP36	TQ72	H96	SN00	SJ14	HU37	SD93	NM72	NN32	NN34	
		Edinburgh, E (Musselburgh)	Daventry	Leamington Spa	C Weald, Robertsbridge	Lough Neagh, SW end	W Pembs, Milton	Llangollen, max alt 600 m	Shetland, Sullom Voe	NW of Hebden Bridge, max alt 518 m	Mull, mainly sea (Firth of Lorn)	Crianlarich, max alt. 977 m	nr Bridge of Orchy, max alt 1079 m	

Figure S5(b). Individual cell concentrations for Liverwort dataset, PSKM W1 analysis. Note the high concentration of records in Scotland and northern England (sample clusters 8-12). This is seen in the large mass of little-occupied dark cells in the bottom left of the diagram. This is much larger than the corresponding group of dark cells at the top right. Although the Scottish hectad NT37 is the key sample for cluster 1, this is a lowland cluster, with the large majority of its members in England. Three English hectads had the same mean square cosine 0.9971; in these and NT37, only two or three out of seven species in each hectad were leafy liverworts.

Figure S6 (on next page). Individual cell concentrations (observed/expected) for Native Vascular Plant dataset, PSKM W1 analysis. Rows represent species clusters, named by their key species. Columns represent hectad clusters. Hectad clusters that are predominantly Irish are marked in green at the top and bottom of the figure.

Note that the clusters with key species *Acer campestre* and *Chaerophyllum temulum* are largely absent in Ireland. In this example there are three near-ubiquitous clusters. The *Ranunculus repens* cluster is genuinely ubiquitous. The *Crataegus monogyna* cluster is largely missing from Shetland and the more mountainous parts of the Scottish Highlands. The *Conopodium majus* cluster is largely missing from the agricultural fenlands of eastern England and from NW Scotland and NW Ireland.

At this scale, some groups are geographically defined and others are ecological. The *Potamogeton crispus* cluster consists mainly of water plants with some swamp and water-margin species. The *Sagittaria sagittifolia* cluster has a similar mix of species, but with a markedly more southern distribution.

Bicluster cell concentration matrix

Samples

														R84	N56	X39	N90	M07			L67				
														Irish	Irish	Irish	Irish	Irish			Irish				
	Specclu	1	4	3	5	8	2	6	7	9	10	12	16	11	14	13	15	19	20	17	18	21	22	23	24
Cirsium acaule	1	5.58	2.6	2.19	1.1	1.11	1.56	1.1	0.59	0.99	0.19	0.35	0.25	0.24	0.25	0.17	0.06	0.03	0.05	0.09	0.09	0.13	0.08	0.03	0.05
Clematis vitalba	2	2.66	2.93	2.31	0.77	1.25	2.69	2.71	3.02	0.32	0.95	0.47	0.06	0.24	0.21	0.55	0.12	0.07	0.01	0.06	0.1	0.01	0	0	0
Sagittaria sagittifolia	4	1.5	2.66	3.11	2.2	1.83	2.13	0.74	0.35	0.62	0.18	0.47	0.26	0.76	1.62	0.24	0.31	0.06	0.06	0.1	0.05	0.05	0.01	0.01	0.02
Acer campestre	5	2.28	2.05	2.44	2.03	1.97	1.89	1.45	0.97	1.15	0.97	0.57	0.32	0.38	0.38	0.32	0.14	0.05	0.08	0.09	0.05	0.01	0.02	0.01	0.01
Erodium cicutarium ag	7	1.45	1.83	1.54	1.16	1.34	1.86	1.64	2.05	0.92	0.79	1.81	0.8	0.24	0.33	1.14	0.19	0.16	0.48	0.89	0.71	0.23	0.35	0.26	0.23
Chaerophyllum temul	8	1.62	1.47	1.51	1.41	1.59	1.08	1.15	0.65	1.74	1.28	1.22	1.46	0.32	0.49	0.41	0.21	0.17	0.87	0.58	0.16	0.14	0.31	0.26	0.3
Epilobium hirsutum	9	1.25	1.13	1.29	1.34	1.23	1.15	1.11	1.09	1.14	1.21	1.06	0.87	1.58	1.27	1.18	1.04	0.45	0.48	0.73	0.59	0.14	0.34	0.12	0.11
Potamogeton crispus	10	1.16	1.38	1.58	1.45	1.35	1.2	0.86	0.45	1.01	0.44	1.14	0.9	1.33	1.96	0.76	0.96	0.47	0.49	0.67	0.74	0.58	0.34	0.24	0.15
Parapholis strigosa	3	0.28	0.87	0.57	0.84	0.24	8.91	3.69	2.78	0.11	0.17	2.51	0.06	0.13	0.18	2.3	0.09	0.08	0.01	0.78	0.76	0.54	0.22	0	0.01
Crithmum maritimum	6	0.25	0.65	0.19	0.33	0.2	2.79	5.27	11.4	0.13	0.8	2.32	0.11	0.25	0.28	3.73	0.24	0.31	0.03	1.07	1.9	0.36	0.2	0.02	0.05
Glaux maritima	15	0.15	0.32	0.24	0.54	0.2	2.14	1.98	2.52	0.26	0.31	2.64	0.25	0.17	0.33	3.08	0.25	0.44	0.2	2.81	3.51	3.68	2.51	0.48	0.43
Crataegus monogyna	11	0.93	0.83	0.99	1.15	0.96	0.91	0.83	0.92	0.99	1.05	0.95	1.06	1.44	1.15	1.16	1.32	1.08	1.1	1.06	0.91	0.44	0.94	0.71	0.52
Oenanthe crocata	12	0.46	1.06	0.33	0.24	0.98	0.54	1.65	1.66	0.85	2.17	1.21	0.72	0.96	1.23	1.53	1.52	1.66	0.66	1.42	1.7	0.74	0.92	0.25	0.18
Conopodium majus	13	0.89	0.94	0.82	0.83	1.05	0.64	0.86	0.54	1.13	1.2	0.99	1.26	0.74	1.09	0.73	1.05	1.1	1.41	1.2	0.67	0.6	1.28	1.41	1.27
Ranunculus repens	14	0.78	0.72	0.82	0.97	0.83	0.75	0.74	0.89	0.88	0.95	0.86	0.98	1.27	1.04	1.13	1.3	1.46	1.16	1.02	1.34	1.59	1.13	1.34	1.16
Molinia caerulea	16	0.36	0.72	0.25	0.3	0.77	0.28	0.75	0.51	1.03	1.22	0.92	1.25	0.87	1.17	0.76	1.45	1.85	1.53	1.31	1.49	2.19	1.52	1.93	1.78
Littorella uniflora	18	0.1	0.59	0.21	0.19	0.4	0.19	0.7	0.48	0.41	0.71	0.79	0.84	0.69	1.73	0.62	1.26	2.46	1.27	1.79	2.59	2.97	3.14	2.69	2.53
Empetrum nigrum	19	0.18	0.27	0.08	0.12	0.33	0.08	0.36	0.16	1.09	0.78	0.88	1.79	0.37	0.79	0.42	0.69	1.34	2.37	1.86	1.31	2.92	3.06	3.52	3.97
Alchemilla glabra	17	0.09	0.25	0.1	0.32	0.62	0.07	0.18	0.02	2.23	0.77	1.25	3.03	0.25	0.96	0.25	0.74	1	3.01	1.43	0.35	0.67	1.71	2.53	3.06
Gnaphalium supinum	20	0.01	0.03	0	0.01	0.02	0.03	0.04	0.02	0.3	0.12	0.1	1.12	0.01	0.09	0.06	0.14	0.6	2.46	0.79	0.35	1.77	4.26	7.03	18.6
		Wilton, Wiltshire	Aldermarston	Milton Keynes	Pocklington, E Yorks	Newcastle-under-Lyme, Staffs	Kent coast, Faversham	N Devon, Bideford	Cornwall, Perranporth	Appleby-in-Westmoreland	Exmoor, Exford	Kirkcudbright, S of	Stirling	N of Tipperary (SW Ireland)	Castlepollard (Westmeath) - al	Dungarvan (SE Ireland) - mainl	Hollywood (Co. Wicklow) - alr	Partry Mts, S of Westport (Co.	Cawdor, and 600 m hill S of it	Mull of Kintyre, Tarbert, max a	Inishturk (Atlantic island, off C	Shetland, Mainland S of Lerwic	Skye, S of Kyle of Lochalsh	C Highlands, Strath Oykel, 500	Highlands, Fersit and Stob Coir

Concentration ratio 1.173