The Contributions of Rare Objects in Correspondence Analysis

Michael Greenacre

September 2011

Barcelona GSE Working Paper Series

Working Paper nº 571

The contributions of rare objects in correspondence analysis

Michael Greenacre

Department of Economics and Business
Universitat Pompeu Fabra
08005 Barcelona
Spain

Faculty of Biological Sciences, Fisheries & Economics
University of Tromsø
N-9037 Tromsø
Norway

E-mail: michael.greenacre@upf.edu

Abstract: Correspondence analysis, when used to visualize relationships in a table of counts (for example, abundance data in ecology), has been frequently criticized as being too sensitive to objects (for example, species) that occur with very low frequency or in very few samples. In this statistical report we show that this criticism is generally unfounded. We demonstrate this in several data sets by calculating the actual contributions of rare objects to the results of correspondence analysis and canonical correspondence analysis, both to the determination of the principal axes and to the chi-square distance. It is a fact that rare objects are often positioned as outliers in correspondence analysis maps, which gives the impression that they are highly influential, but their low weight offsets their distant positions and reduces their effect on the results. An alternative scaling of the correspondence analysis solution, the contribution biplot, is proposed as a way of mapping the results in order to avoid the problem of outlying and low contributing rare objects.

Keywords: Biplot, canonical correspondence analysis, contribution, correspondence analysis, influence, outlier, scaling.

Paper presented at the 1st International Workshop of CARME in ASSOS,
Assos, Turkey, 1-4 October 2011
http://www.carme-n.org/assos

Introduction

Correspondence analysis (CA) is a variant of principal component analysis (PCA) that is adapted to the properties of count data, and frequently used to analyze tables of species abundances or biomasses in ecology (for example, Greenacre and Vrba 1984, Greenacre 2010b). The application of CA to such ecological data is justified because of the method's links to the Gaussian model in gradient analysis and to ecological concepts such as niche theory and coenoclines (see, for example, Gauch 1982). In ecology the method is popular in its constrained form, *canonical correspondence analysis* (CCA), where the solutions for the biological data are restricted to be linearly related to explanatory environmental variables (ter Braak 1985, 1986).

In the typical ecological context of a samples-by-species abundance table, CA achieves visualizations of the samples (rows) and species (columns), with several possible styles of interpretation: for example, as a pair of multidimensional scalings of the samples and of the species in a single graphic, or as a variety of biplots (Greenacre (1993, 2007: chap. 13, 2010a: chap. 8). Each row or column is expressed relative to its respective marginal frequency, and it is these so-called profiles consisting of relative frequencies that are visualized in the classic definition of CA (an alternative version applicable to the abundances in their raw unrelativized form has been proposed by Greenacre (2010b), when the overall level of abundance in each sample should be incorporated into the visualization – for an application and comparison with regular CA, see Cochrane et al., 2011). The marginal totals of the table, that is the sample totals and species totals, also expressed as proportions of the grand total, serve two purposes in CA. First, they are surrogate estimates of the variance and are used to standardize the row and column profiles to give what are called *chi-square distances* between the profiles. Second, they are used to weight the profiles in the process of dimension reduction, so that profiles based on lower counts receive proportionally less weight in the analysis.

Several authors have criticized the use of the normalization implied in the chi-square distance, saying that it exaggerates the contribution of rare species to the analysis. For example, Rao (1995: p. 42) states: “since the chi-square distance uses the marginal proportions in the denominator, undue emphasis is given to the categories with low frequencies in measuring affinities between profiles”. Legendre (2001: p. 271) says that “a difference between abundance values for a common species contributes less to the distance than the same difference for a rare species, so that rare species may have an unduly large influence on the analysis.” The first part of Legendre’s assertion is true: as Clarke and Ainsworth (1993: p. 206) quite rightly state, “species data require a rather careful formulation of similarity, with appropriate trade-offs between the contributions of common and rarer species”. However, with respect to Rao’s definite assertion that “undue emphasis is given” and Legendre’s more guarded “may have an unduly large influence”, we aim to show that the method’s giving “undue emphasis” or “unduly large influence” to rare species is almost always not the case in practice. To do this we shall calculate influence measures and contributions to chi-square distances in correspondence analysis and canonical correspondence analysis, for both common and rare species – this will demonstrate that the influence of rare species is not excessive, that their contributions to the results are in line with their rarity and that the more common species do indeed dominate both the chi-square distances and the graphical displays. Our experience is that rare species do not excessively contribute to the results, with one exception: that is, when a species is concentrated at a sampling site and hardly any other species are observed there. This is a circumstance that hardly ever occurs in practice, and if it does occur it is clear, probably even before any data analysis is attempted, that both the species and the sample are so different from the rest of the data that one or both of them should be removed.

What is true about rare species and about the way CA maps are generally reported is that the rare species occupy outlying positions. Since outliers in statistics are associated with influential points, this gives the impression that rare species have high influence. But this ignores the fact that rare species have low weight in the CA algorithm, and influence the results much less than abundant species. The "fault" lies in the way the maps are scaled, so we propose an alternative way of presenting CA maps, called the *contribution biplot*, which shows species in positions that are related to their contribution to the results. In this alternative graphical presentation rare species are no longer outlying and are immediately seen to have a minor role in the display.

Correspondence analysis and contributions to inertia and chi-square

The theory of CA is well-known and summarized in several texts; for recent accounts see, for example, Kroonenberg and Greenacre (2005), Greenacre (2007). The method can be defined and interpreted in many different ways: for example, the method can be defined as the reconstruction, in a low-dimensional space, of the chi-square distances between the site profiles across the species, or alternatively as the weighted least-squares approximation of the abundance table itself.

Consider the data in Table 1 (only part of the complete 13-by-92 table is shown, the complete table is available online – see the table caption). These are typical marine ecological abundance data, obtained by sampling benthic species on the sea-bed as part of an annual monitoring exercise around a North Sea oil platform. The first 11 sampled stations are close to the platform in polluted areas, whereas the last 2 stations are reference stations 10kms away and regarded as unpolluted. The columns are ordered in descending order of overall abundance of the 92 species sampled, and there is almost a 1000-fold difference between the most frequent species, *Galathowenia oculata* and the rarest, for example *Modiolus modiolus.* For 36 of these 92 species the overall abundance is less than the number of sampling points. Ten species occur at only three of the 13 stations, with counts of 1 or at most 2 – these are shown as the final columns of Table 1 and we will refer to them as the "rare species group". CA visualizes the abundances in their relative form; that is, the profile of each sample (row) across the species is calculated, chi-square distances are computed between the profiles, and the profiles are then weighted by their margins in the dimension-reduction step to achieve the low-dimensional map of the sites. Species points are then typically displayed as unit profile points and the resulting map is a *biplot* – see Figure 1 (see Greenacre (2010a) for more details). But, as is well-known, when considering the table as a set of columns rather than a set of rows, species profiles (i.e., columns divided by their respective sums) have coordinates which are the same as these unit profile points, up to scaling factors along the principal axes equal to the square roots of the corresponding variances (or inertias, in CA terminology). It is evident that a species such as *Modiolus modiolus*, with most of the entries zero and just three counts of 1, will have a quite different profile compared to the "abundant" species such as *Galathowenia oculata*, which has counts in every sample.

To establish notation we summarize the algebraic definition of CA as follows, assuming the raw data matrix is denoted by $\mathbf{N}$. The row profiles visualized in Figure 1 are the rows of Table 1 divided by their respective row totals. CA is invariant to the grand total of the table, and the notation is simplified if we regard the initial data matrix as the matrix $\mathbf{P} = [p_{ij}]$ where $p_{ij} = n_{ij}/n$ and n is the grand total of the table. Let the row and column marginal totals of $\mathbf{P}$ be the vectors $\mathbf{r}$ and $\mathbf{c}$ respectively – these are the weights, or *masses*, associated with the rows and columns. Let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses. The row profiles are then rows of the matrix $\mathbf{D}_r^{-1}\mathbf{P}$ and the computational algorithm to obtain the solution, using the singular value decomposition (SVD), is as follows:

1. Center the row profiles with respect to their average $\mathbf{c}^{\mathsf{T}}$, then pre-multiply by $\mathbf{D}_r^{1/2}$ to weight the profiles by their masses, and post-multiply by $\mathbf{D}_c^{-1/2}$ to engender the chi-square metric between rows:

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2} \qquad (1)$$

2. Calculate the SVD: $\mathbf{S} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^{\mathsf{T}}$ where $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$. (2)

3. Principal coordinates of rows and columns (profiles):

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\sigma \qquad \mathbf{G} = \mathbf{D}_c^{-1/2}V\mathbf{D}_\sigma \qquad (3)$$

4. Standard coordinates of rows and columns (unit profiles):

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U} \qquad \mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V} \qquad (4)$$

One of the standard CA displays is the so-called asymmetric map, where rows (samples) are displayed by their projected profiles onto the first two principal axes, and the columns (species) by the projections of the unit profiles. This map, which is a well-defined biplot, is also known as the “row-principal” map, since rows are plotted in principal coordinates and columns in standard coordinates, as in Figure 1. With this scaling of the row and column points, each sample is at the weighted average of the species points, the weights being the relative frequencies. Hence the two reference stations, R40 and R42, are attracted to the right because they have higher than expected relative abundances of the species on the right. The species that is the most outlying on the right hand side is *Aonides paucibranchiata* (abbreviated as *Ao_pa*), which thus might be interpreted as a key species distinguishing the reference stations from the other polluted stations. But this species is one of the rare group, and happens to have two counts of 1 in both reference stations (see Table 1), which is 50% of its total count of 4. The same phenomenon is seen with *Eumida ockelmanni* (*Eu_oc*), another rare species, the most negative on the vertical axis and apparently accounting for the separation of station S15 from the rest, but in fact one of its three counts (i.e., one third of its data) is observed at this station. As we will show below, however, these rare species have very little influence on the solution, which has actually been determined by other more abundant species.

The problem with the map is that the position of species does not reflect its influence in the solution nor its “weight of evidence” in the interpretation – the position of a rare species is analogous to a mean that is calculated on a very small sample size, and thus neither reliable nor informative.

In CA, which is based on the SVD and thus on least-squares optimization, each cell of the table contributes a positive amount to the fitted solution. Such contributions can be aggregated for rows or for columns for each dimension of the solution and these provide powerful diagnostics to indicate which samples or which variables are really driving a particular result. Greenacre (2011) shows that the proportional contributions by rows or columns to each principal axis are just the squares of the elements of the corresponding singular vectors computed in step 2 of the algorithm given above. Thus the squares of the first two columns of **V** give contributions of the species to axes 1 and 2 respectively, while the contributions to the two-dimensional solution are given by these contributions weighted by the squared singular values σ_1^2 and σ_2^2 and then expressed relative to their sum $\sigma_1^2 + \sigma_2^2$:

Contribution of *j*-th column to the two-dimensional solution: $$\frac{\sigma_1^2 v_{j1}^2 + \sigma_2^2 v_{j2}^2}{\sigma_1^2 + \sigma_2^2} \qquad (5)$$

Equivalently, in terms of the standard coordinates in **Y**: $$\frac{c_j(\sigma_1^2 y_{j1}^2 + \sigma_2^2 y_{j2}^2)}{\sigma_1^2 + \sigma_2^2} \qquad (6)$$

Notice in the equivalent formula (6) in terms of the standard coordinates (which were used to plot the species as unit profiles in Figure 1) how the masses c_j of the species affect the contributions. The contributions quantify how much the species, both by their position and their mass, are driving the actual solution. Figure 2 shows a plot of the contributions of the 92 species and their overall percentage abundance in the data. This plot and similar subsequent plots are shown on a log-log scale so that the many low abundances and low contributions are separated more. There are 10 species that contribute more than average to the solution – these are labeled in red. The rare species group is labeled in blue, while the remaining species are not labeled. Contribution appears monotonically increasing with abundance, a first indication that rare species are not dominating the results. The "rarest" species (in terms of abundance) amongst the top 10 that are labeled is *Eudorellopsis deformis* (*Eu_de*), which is fact the 29th most abundant species, occurring at 8 out of the 13 stations, particularly at reference station R42, hence its position at top right of Figure 1.

Each of the 10 least abundant species contributes less than 0.5% to the solution, and could effectively be removed from the data set without any noticeable effect on the results. Their positions in Figure 1, however, tend to be outlying, as illustrated previously with *Aonides paucibranchiata* and *Eumida ockelmanni*, which gives the impression that they are important to the interpretation. This will be further discussed in the next Section.

The other criticism leveled at rare species is that they overly contribute to the chi-square distance between samples. Since the (squared) chi-square distance is also a result of the sum of positive contributions due to each species, it is similarly possible to compute the contribution of each species to this distance measure. The square of the chi-square distance in the "full space" between the *i*-th and *i′*-th stations is given by:

$$d_{ii'}^2 = \sum_j \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 / c_j \qquad (7)$$

For each species *j* the contributions to all pairs (*i*,*i′*) can be summed and expressed relative to the grand total for all species. These contributions to chi-square are plotted against relative abundance in Figure 3, again showing in red the species that have more than average contributions (in this case, 18 species), and in blue the rare species group. Again, the rare species have low contributions to the distances – in fact, among the 46 species below median abundance there are none that figure in the top 18 labeled in Figure 3.

Scaling of the display

It has been illustrated by one example that rare species do not contribute excessively to the CA solution, nor to the chi-square distances (more examples will be given in the next section). We remarked previously that rare species are often situated in outlying positions in the map, giving an exaggerated impression of their importance in the analysis. An alternative scaling of the solution, called the *contribution biplot* (Greenacre, 2011) shows each species on the same biplot axis through the origin, but reduces their standard coordinates by the square roots $\sqrt{c_j}$ of their relative abundances (cf. (5) and (6), where the contribution coordinates are $v_{ij} = \sqrt{c_j}\, y_{ij}$). With this scaling the coordinates of each species are directly related to their contributions to each axis, so that species that are outlying are truly high contributors to the solution and thus important to the interpretation. The specific link between the contribution coordinates and the contributions is as follows: the squared length of the species point on a principal axis is equal to its part contribution to that axis.

The contribution biplot for the benthic data is shown in Figure 4, with just the 10 top contributors labeled (as in Figure 2), and it is immediately clear in the map which are relevant species, while the low-contributing species, including the rare group, are now close to the origin. Thus the seven species on the right are the ones mainly responsible for the separation of the reference stations from the polluted stations, while the separation of station S24 is due to the unusually high relative abundance of *Galathowenia oculata* (*Ga_oc*). *Chaetozone setosa*

(*Ch_se*) and, to a lesser extent, *Ampharete falcate* (*Am_fa*) are the species separating out stations S15, S9 and S14, which are in fact the most polluted stations.

Further examples

Since it might be suspected that the results above depend on the chosen example, we applied the same calculations of contributions to two other data sets that we are working on at the moment, both from the Barents Sea: a data set of 88 samples on 30 species, and one of 1360 samples on 55 species. Plots of contribution to the respective two-dimensional solutions and to the chi-square distances, versus relative abundance (as in Figures 2 and 3) are given in Figures 5 and 6 for the two data sets. Both sets of results show a very similar pattern to what was observed previously: the high abundance species contribute more than the low abundance ones, and the contributions of the rare species are in line with their rareness.

As a final example, the last 1360-by-55 data set was subject to a canonical correspondence analysis (CCA), with several environmental predictors, including spatial, temporal, depth and temperature variables. The contribution of the species to the two-dimensional CCA versus their relative abundance is given in Figure 7. This plot shows an even stronger monotonic relationship between contributions and abundance, which reflects our experience that the effect of rare species is reduced in the presence of constraining variables.

Discussion

By calculating the contributions of the species to the results of four different analyses, we hope to dispel the urban legend that CA is overly sensitive to rare species. We have demonstrated this by computing how much each species contributes to a particular CA or CCA solution, as well as how much each contributes to the inter-sample chi-square distances. In all the examples presented here, the contributions of rare species are low and in line with their rareness.

The popular program package CANOCO (ter Braak and Smilauer, 2002) offers options to reduce the weight of rare species in the analysis, or to actually delete them entirely (Lepš and Šmilauer, 2003). Our experience, supported by the above examples, shows that the downweighting or deletion of rare species is not necessary. The practice of logarithmically transforming the data by log(1+*x*) prior to applying CA is similarly unnecessary, since the normalization implied by the chi-square adequately balances out the contributions of rare and frequent species.

We do recognize that the positions of rare species in CA or CCA maps are often outlying, owing to their unusual profiles, but their low weights (usually not seen in the displays) offset these outlying positions in determining their influence on the results. By using an alternative scaling of the solution such as the contribution biplot, the positions of the species reflect both their directions from the center and their contributions. This alternative presentation of the map facilitates its interpretation since it concentrates the analyst on those species that are dominant and most influential in creating the result.

Acknowledgments

This research has been supported by the BBVA Foundation, Madrid, Spain. Partial support of Spanish Ministry of Science and Innovation grants MTM2008-00642 and MTM2009-09063 are also acknowledged. Data in this article include fish species abundance data collected during the former annual shrimp surveys by the Norwegian Institute of Fisheries and Aquaculture (NIFA) and the Institute of Marine Research (IMR) in the Barents Sea.

References

Clarke, K.R. and Ainsworth, M. 1993. A method linking multivariate community structure to environmental variables. Marine Ecology Progress Series 92: 205–219.

Cochrane, S., Pearson, T., Greenacre, M., Costelloe, J., Dahle, S. and Gulliksen, B. 2011. Benthic fauna and functional traits along a Polar Front transect in the Barents Sea – advancing tools for ecosystem-scale assessments. Under review for publication.

Gauch, H. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, UK.

Greenacre, M. 1993. Biplots in correspondence analysis. Journal of Applied Statistics 20: 251–269.

Greenacre, M. 2007. Correspondence analysis in practice. Second edition. Chapman & Hall / CRC Press, London. Published in Spanish translation by the BBVA Foundation, Madrid, 2008, and freely downloadable from www.multivariatestatistics.org

Greenacre, M. 2010a. Biplots in practice. BBVA Foundation, Madrid. Freely downloadable from www.multivariatestatistics.org.

Greenacre, M. 2010b. Correspondence analysis of raw data. Ecology 91: 958–963.

Greenacre, M. 2011. Contribution biplots. Working Paper 1162, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, under review for publication. Downloadable from http://www.econ.upf.edu/en/research/onepaper.php?id=1162.

Greenacre, M. and Vrba, E. 1984. Graphical display and interpretation of antelope census data in African wildlife areas, using correspondence analysis. Ecology 65: 984–997.

Kroonenberg, P. M. , and Greenacre, M. (2005). Correspondence analysis. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic (Eds.). Encyclopedia of Statistical Sciences. Wiley, New York, pp. 1394–1403.

Legendre, P. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia: 129, 271–280.

Lepš, J and Šmilauer, P. 2003. Multivariate analysis of ecological data using CANOCO. Cambridge University Press, UK.

Nenadić, O. and Greenacre, M. J. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the **ca** package. Journal of Statistical Software, 20 (3).
URL `http://www.jstatsoft.org/v20/i03/`

R Development Core Team (2011). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Rao, C.R. 1995. A review of canonical coordinates and an alternative to correspondence analysis. Qüestiió: 19, 23–63.

ter Braak, C.J.F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41: 859–873.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector method for multivariate direct gradient analysis. Ecology 67: 1167–1179.

ter Braak, C.J.F. and Smilauer, P. 2002. CANOCO reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5). Micromputer Power, Ithaca, New York.

Table 1: Abundances of 92 benthic species in samples collected at 13 stations in the North Sea. The species (columns) are ordered in descending order of total abundance, showing the 10 most abundant and 10 least abundant ones. These data, which were also used by Greenacre (2010a, 2010b), can be downloaded from www.multivariatestatistics.org.

stn id	Ga_oc	Ch_se	Am_fa	My_bi	Go_ma	Am_fi	Ti_ov	St_li	Ch_ni	Tr_sp	···	Ao_pa	Ar_si	Sa_se	Fa_cr	Ep_tr	Op_fl	Eu_sp	Sc_in	Eu_oc	Mo_mo	totals
S4	193	34	49	30	35	19	1	22	25	9	···	0	0	0	1	0	0	0	0	0	0	594
S8	79	4	58	11	39	39	78	33	21	26	···	0	0	2	0	0	1	0	1	0	0	577
S9	150	247	66	36	41	11	5	29	35	5	···	0	1	1	2	1	1	0	0	1	0	827
S12	72	19	47	65	37	38	95	26	14	30	···	2	0	0	0	0	0	0	0	0	1	658
S13	141	52	78	35	32	18	20	16	20	35	···	0	2	0	0	2	0	1	0	0	1	644
S14	302	250	92	37	45	20	0	19	30	2	···	0	1	1	0	0	0	0	1	1	0	1043
S15	114	331	113	21	41	11	7	27	17	11	···	0	0	0	0	0	0	0	0	1	0	871
S18	136	12	38	3	41	22	55	21	9	13	···	0	0	0	0	0	1	1	0	0	1	516
S19	267	125	96	20	31	30	50	22	41	5	···	0	0	0	0	0	0	1	0	0	0	888
S23	271	37	76	156	29	40	44	22	36	63	···	0	0	0	1	1	0	0	0	0	0	978
S24	992	12	37	12	64	3	3	12	11	1	···	0	0	0	0	0	0	0	0	0	0	1331
R40	5	8	0	58	32	55	0	32	14	0	···	1	0	0	0	0	0	0	0	0	0	355
R42	12	3	5	43	23	65	2	19	10	1	···	1	0	0	0	0	0	0	1	0	0	313
totals	2734	1134	755	527	490	371	360	300	283	201		4	4	4	4	4	3	3	3	3	3	9595

Figure 1: Correspondence analysis of Table 1, with rows in principal coordinates and columns in standard coordinates (row-principal asymmetric map). Total inertia = 0.7827; inertia explained in the two-dimensional map = 57.5%. The R package **ca** by Nenadić and Greenacre (2007) was used to perform the analysis.

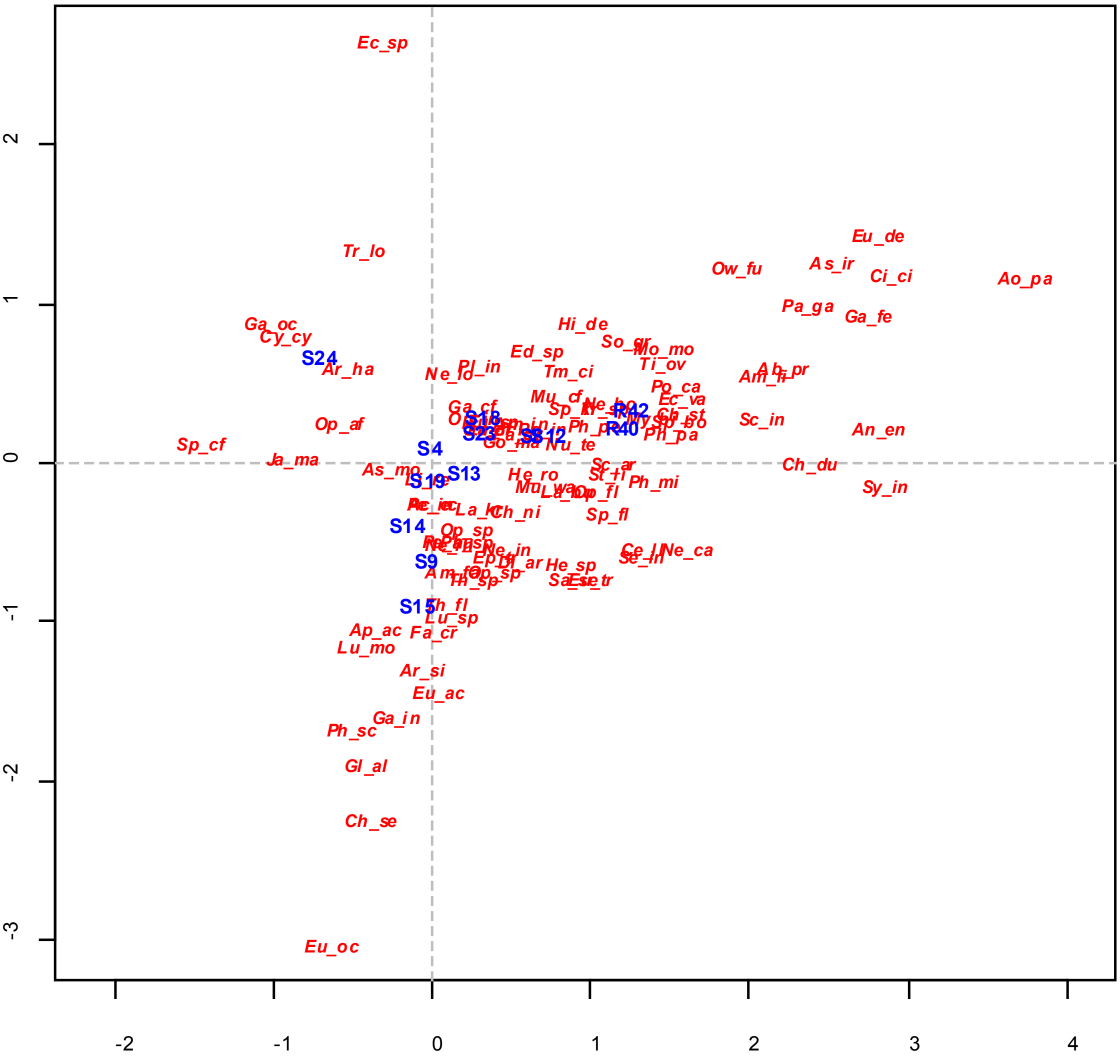


Figure 2: Plot of the relative contributions, as percentages of the inertia of the two-dimensional map of Figure 1, against the relative abundances, for the 92 species. Both axes have logarithmic scales to show more dispersion in the low abundance values. There are 10 species contributing more than the average (these are labeled in red) and they contribute 85.3% to the solution. The rare species group is labeled in blue. The Spearman rho correlation coefficient is 0.626.

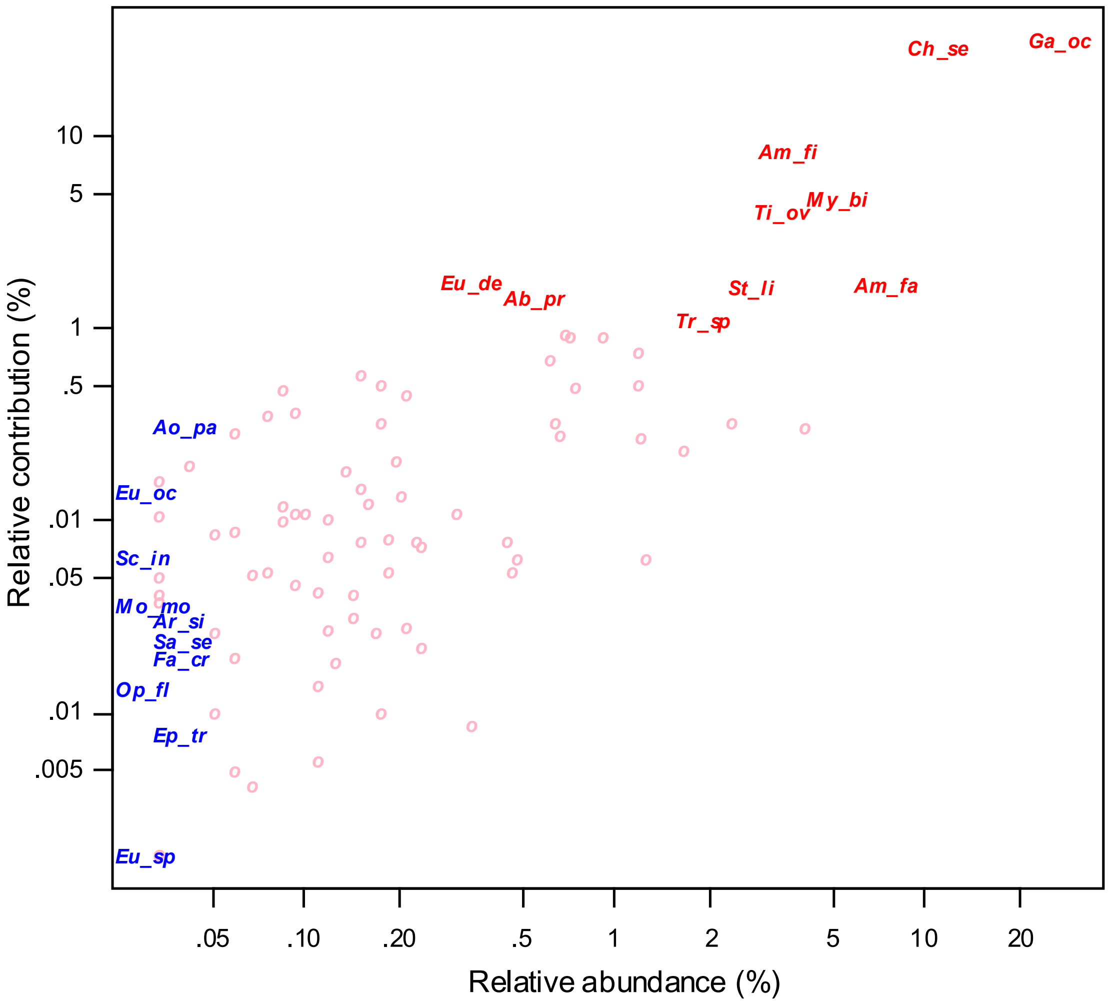


Figure 3: Plot of relative contributions to chi-square distance, as percentages of the total of all chi-square distances, against the relative abundances, for the 92 species. Both axes have logarithmic scales. There are 18 species contributing more than the average (these are labeled), and these contribute 76.2% to the total chi-square. The Spearman rho correlation coefficient is 0.657.

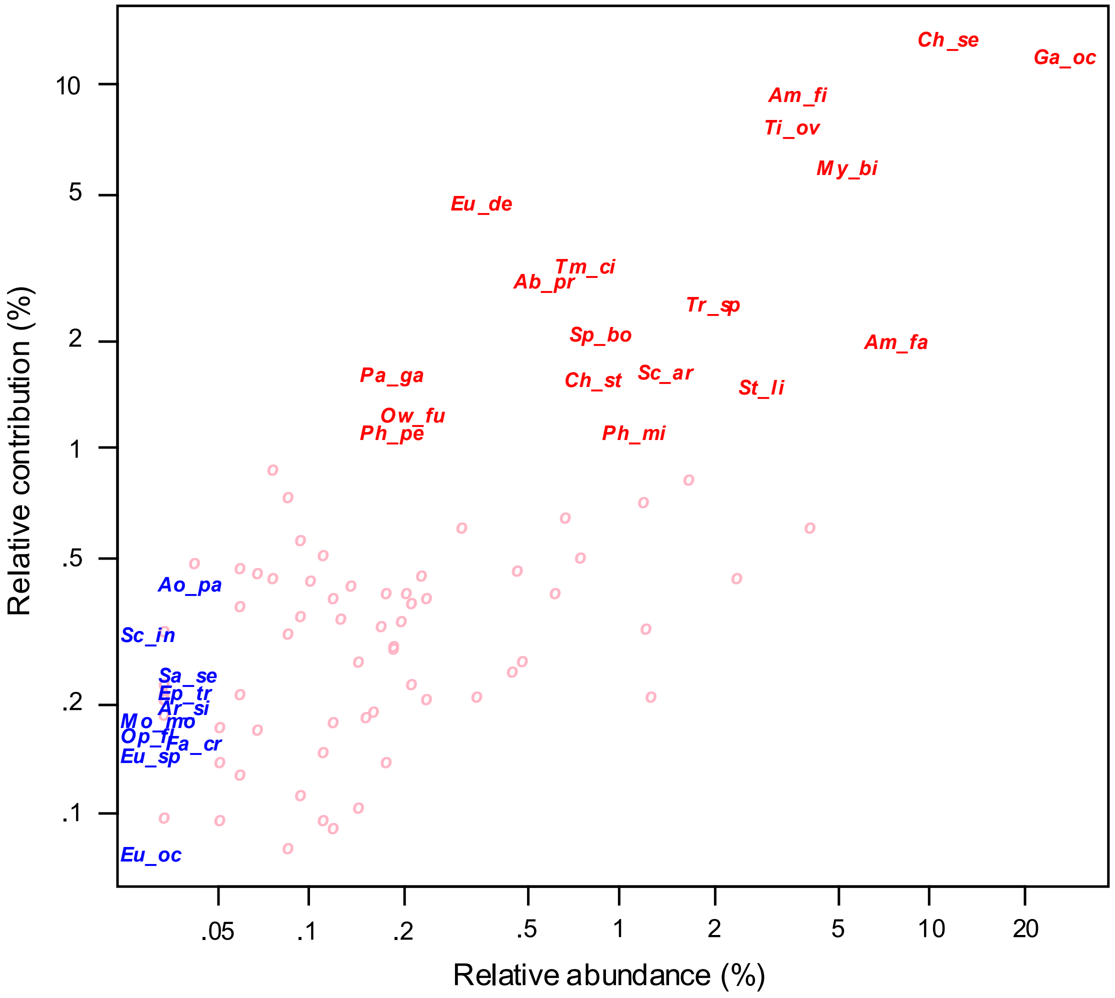


Figure 4: Same analysis as Figure 1, but with contribution biplot scaling. Species positions are indicated by lines from the origin, with labeling for the 10 species that contribute more than average to the solution. The directions from the origin are the same as in Figure 1, but here the 10 highly contributing species are clearly the points furthest from the origin.

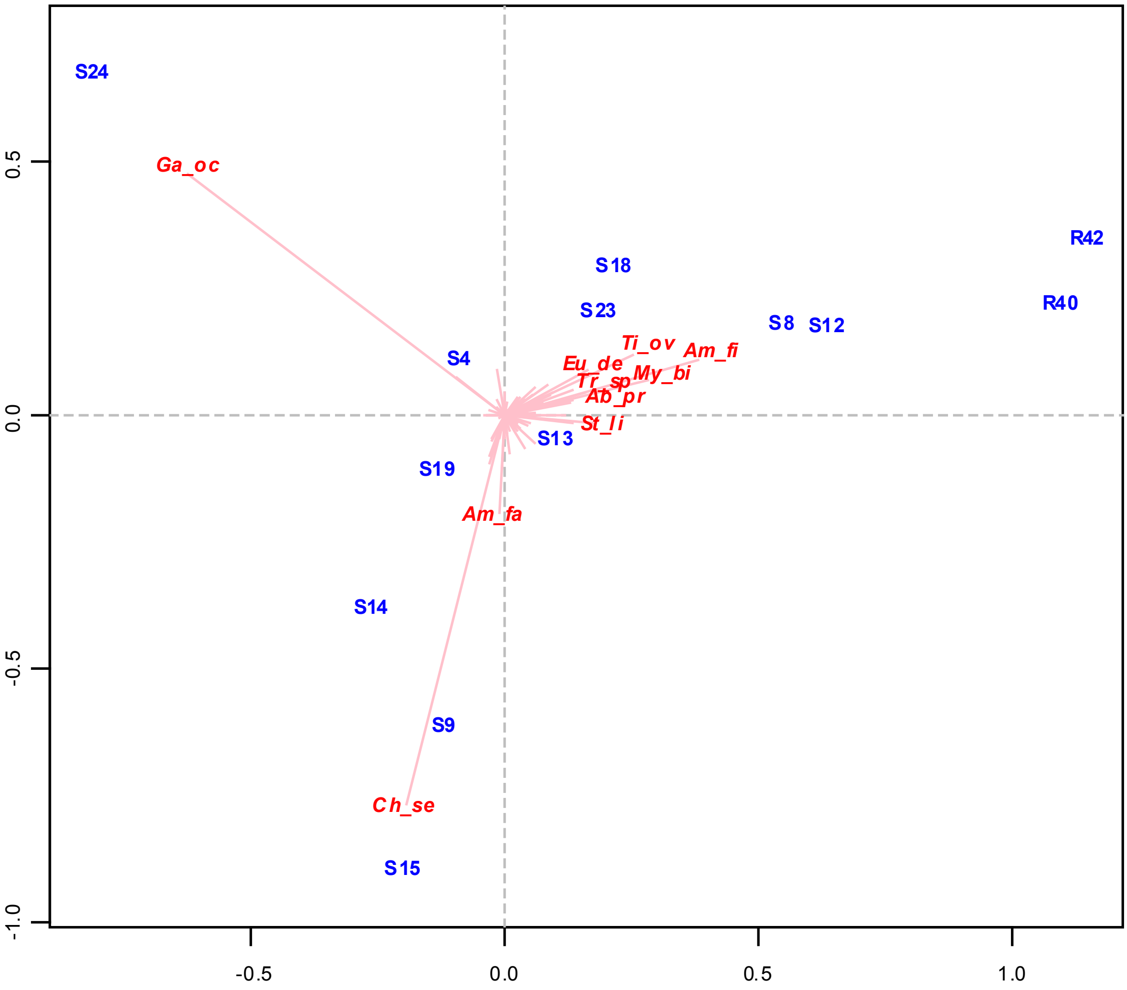


Figure 5: Contributions versus abundance plots for the 88×30 data set, on logarithmic scales. Spearman rho correlations are 0.888 and 0.911 respectively.

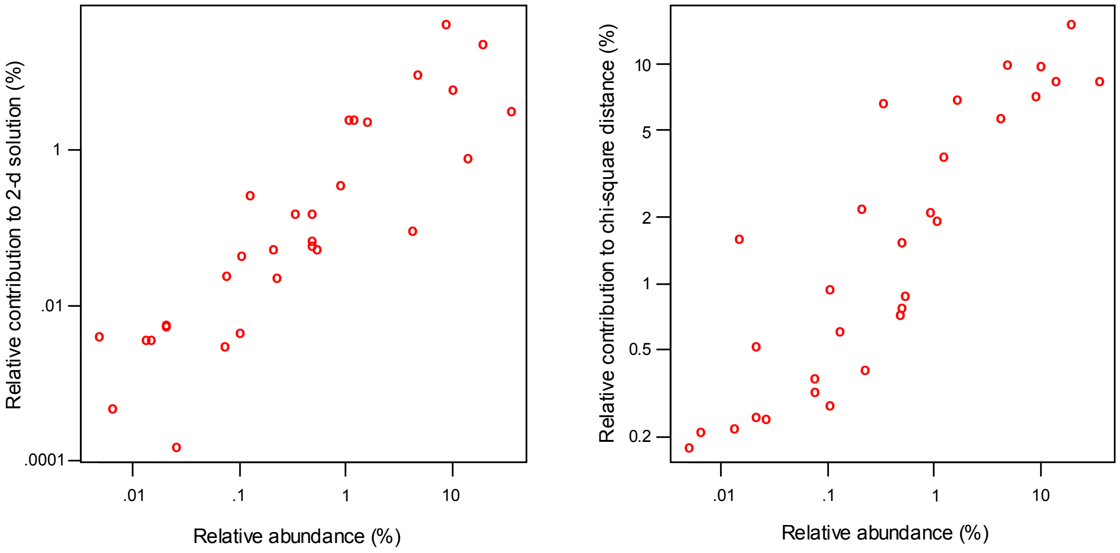


Figure 6: Contributions versus abundance plots for the 1360×55 data set, on logarithmic scales. Spearman rho correlations are 0.895 and 0.823 respectively.

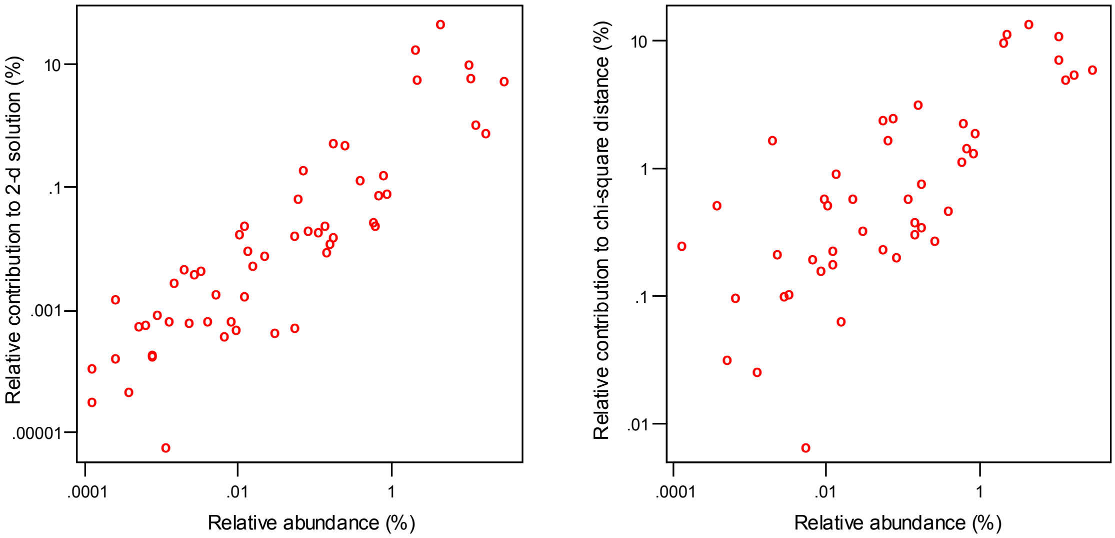
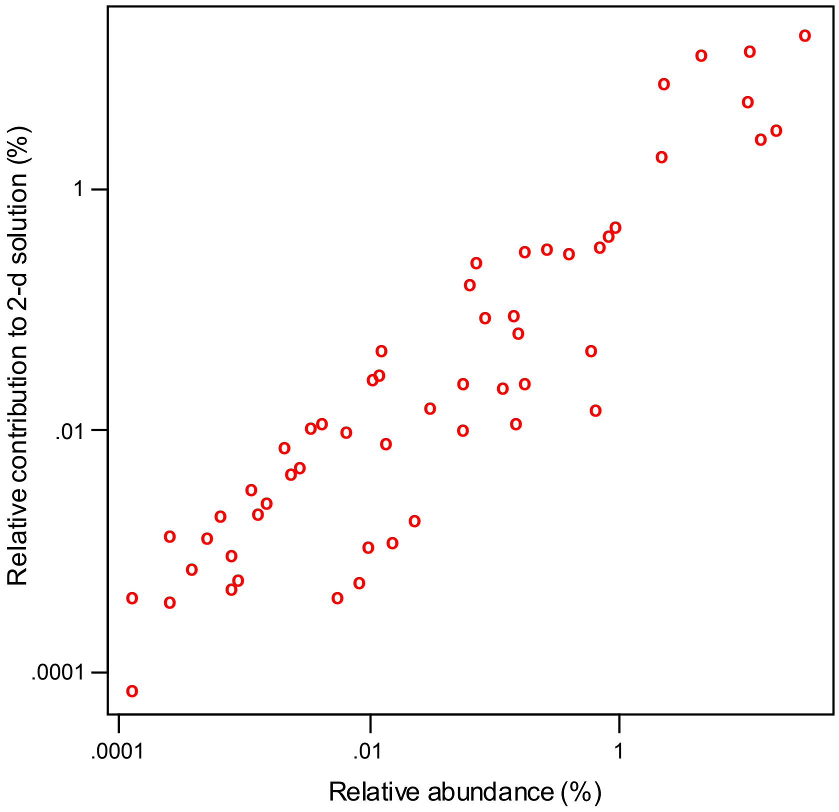


Figure 7: Log-log plot of contributions versus relative abundances for the 55 species in the 1360×55 data set, where the solution was constrained by several environmental indicators in a canonical correspondence analysis. The Spearman rho correlation is 0.904.


Relative contribution to 2-d solution (%)
Relative abundance (%)