Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 20 Offprint

Case Study 2: Functional Diversity of Fish in the Barents Sea

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

CHAPTER 20

Case Study 2: Functional Diversity of Fish in the Barents Sea

The ability of a marine ecosystem to withstand environmental changes depends on its adaptability. Biodiversity makes an ecosystem more adaptable and thereby less vulnerable to change since a high number of species can perform a wide range of ecosystem functions present in the community. Diversity can be measured at the taxonomic level, the phylogenetic level or the functional level, and it is the object of this case study to investigate the last option in the same data set studied in Chapter 19. In order to measure functional diversity, species need to be coded in terms of their functional traits. There are then two alternative ways of proceeding: either create groups of species with similar functional traits and then measure diversity of the functional groups, or use a diversity measure which depends on the particular mix of species present at a site, and how far apart they are in terms of their trait characteristics. Both these approaches will be illustrated in this case study, as well as their relationships to environmental, spatial and temporal variables.

Contents

The functional trait matrix

The starting point for a study of functional diversity is the definition of a set of attributes, called *functional traits*, that define the functioning of the species. These can be the type of feeding, movement and reproductive behaviour, for example.

[1] We are indebted to Magnus Wiedmann of the University of Tromsø for his agreement to use these data, which are part of his PhD thesis and an article in the journal *Marine Ecology Progress Series* (see Bibliographical Appendix).

Exhibit 20.1 shows a part of the trait[1] matrix for the 62 Barents Sea fish species studied in Chapter 19. The total list of traits is as follows:

Diet: three-category variable, multiple responses possible

Habitat: two-category variable

Average fecundity: continuous variable, highly positively skew

Offspring size: three-category variable

Offspring behaviour: three-category variable

Maximum size: continuous variable, highly positively skew

Shape: five-category variable

Salinity range: three-category variable

Temperature range: three-category variable

Depth range: three-category variable

Thus, there are 10 traits, 8 categorical and 2 continuous.

As can be seen in Exhibit 20.1, the categorical options are coded as zeros and ones, and the first variable (diet) can have more than one option indicated as a trait (for

Exhibit 20.1: *Part of the trait matrix coding the various functional characteristics of Barents Sea fish species*

Species		Functional traits								
		Diet			Habitat		Fecundity	Offspring		
Name	Abbrevn	benthivorous	ichthyivorous	planktivorous	demersal	pelagic	(mean)	small	medium	large ⋯
Amblyraja hyperborea	Am_hy	1	1	0	1	0	30	0	0	1 ⋯
Amblyraja radiata	Am_ra	1	1	0	1	0	26.5	0	0	1 ⋯
Anarhichas denticulatus	An_de	1	1	1	1	0	46,500	0	1	0 ⋯
Anarhichas lupus	An_lu	1	0	0	1	0	12,740	0	1	0 ⋯
Anisarchus medius	An_me	1	0	0	1	0	700	1	0	0 ⋯
Anarhichas minor	An_mi	1	0	0	1	0	19,700	0	1	0 ⋯
Artediellus atlanticus	Ar_at	1	1	0	1	0	117.5	0	1	0 ⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

example, the first species is indicated as both benthivorous and ichthyivorous for diet), whereas the others only allow one option; for example, offspring can only be in one category of small, medium or large).

Distances between species based on the traits

There are several approaches to defining a distance, or dissimilarity, between pairs of fish, based on mixed-scale data such as these. Our choice here will be to code each of the continuous variables into three fuzzy categories so that the whole trait matrix can be treated as a set of categorical data. Because the two continuous variables are highly skew, it is important to first log-transform them before the fuzzy coding. Several distance functions are possible: simply computing the sum of absolute differences between the traits of pairs of fish, or applying a distance like the chi-square distance that will normalize the traits according to their average appearance in all the fish. We chose the former approach, so that the distance between fish would not depend on the particular sample of fish included in this study (the chi-square distance would depend on the marginal trait averages). Nevertheless, to get an idea of the relationship between this set of fish and the traits, CA using the chi-square distance is still of interest, as shown in Exhibit 20.2. In the upper right corner, for example, we find fish that must have some of the following characteristics: small (ML1) and bottom dwelling (Demersal) benthivorous species, with strange shapes (*Shape_eellike* or *Shape_deep_short*), having few (FM1), medium-sized *(Medium_offspring)*, demersal eggs *(Egg_dem)* and moderate tolerance to variations in abiotic factors such as temperature and salinity.

Hierarchical clustering of fish using trait distances

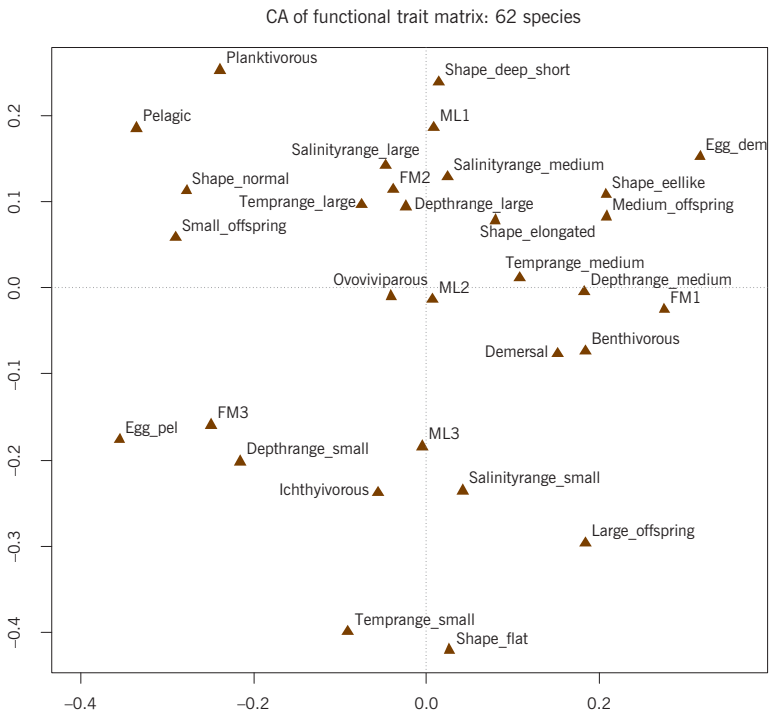
Having defined a distance between the fish, the next step is to perform a clustering of the fish into groups that are relatively homogenous with respect to the traits. Again several choices are available: complete or average linkage or Ward clustering. To ensure a certain level of compactness of the clusters we chose complete linkage – see Exhibit 20.3. Notice that the distance measure has been rescaled so that 1 equals maximum distance.

There are two approaches to defining functional diversity that we shall investigate here. The first way involves defining functional groups, using the results of the hierarchical clustering. Using the permutation test for clustering described in Chapter 17, we obtain the following estimates of p-values for significant clustering, from 2 to 12 groups:

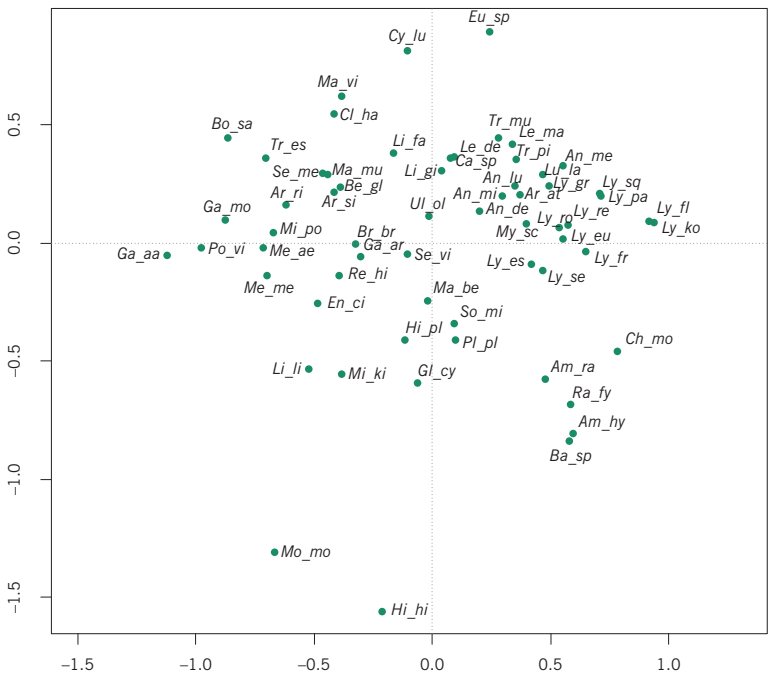
2 groups: $p = 0.989$	6 groups: $p = 0.021$	10 groups: $p = 0.048$
3 groups: $p = 0.975$	7 groups: $p = 0.177$	11 groups: $p = 0.082$
4 groups: $p = 0.354$	8 groups: $p = 0.001$	12 groups: $p = 0.019$
5 groups: $p = 0.821$	9 groups: $p = 0.006$	

Exhibit 20.2: *CA of the trait matrix, part of which is shown in Exhibit 20.1. Traits are shown in principal coordinates in (a) and the fish species in principal coordinates in (b). 27.4% of the inertia is displayed*

(a)



(b)



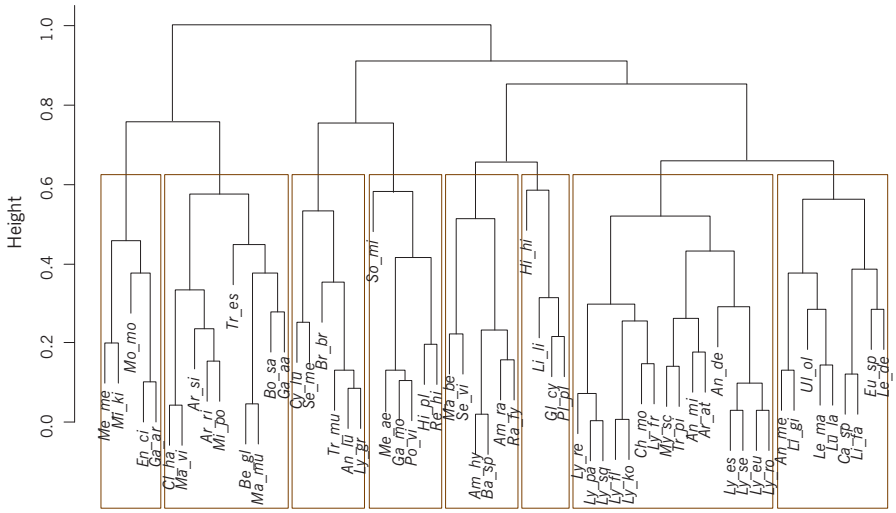


Exhibit 20.3:
Hierarchical clustering of fish based on distances between fish, showing boxes indicating eight clusters

We choose the eight-cluster solution, which is the most significant ($p = 0.001$), indicated in Exhibit 20.3. A six-group solution ($p = 0.021$) is another possibility if fewer groups are required, but we preferred more groups that are internally more homogeneous. Notice in the dendrogram that the nine-group solution ($p = 0.006$) would split off one species on its own, which is not desirable. Hence, the decision about the number of groups is based on statistical significance as a guideline, but also the nature of the dendrogram and substantive biological knowledge.

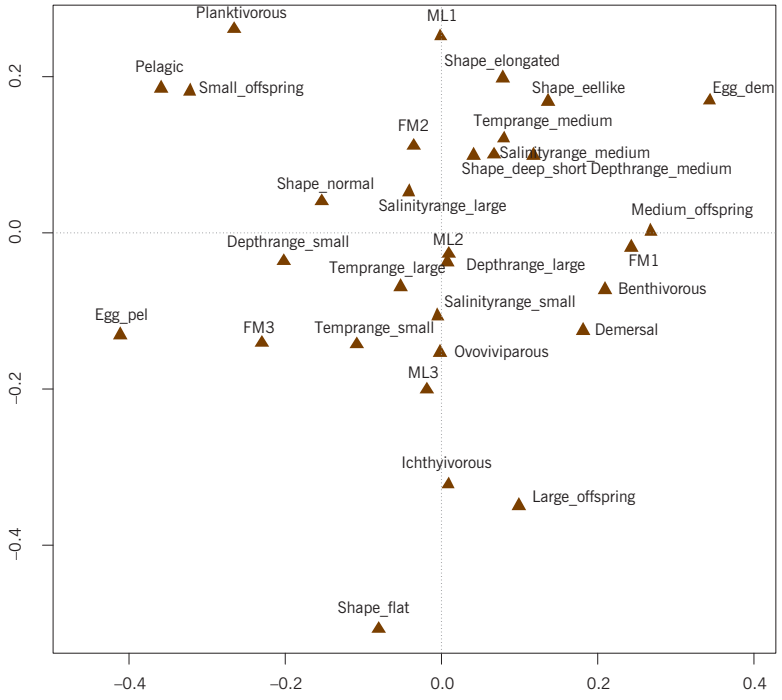
Once functional groups have been defined another CA is possible, at the group level, to interpret these definitions. The trait values for each fish group are aggregated, so we reduce the 62-row trait matrix in Exhibit 20.1 to a 8-row matrix – see Exhibit 20.4. As mentioned in Chapter 16, the CA of this aggregated matrix is a type of discriminant analysis between the fish groups, or alternatively a CCA of the original trait matrix with fish group as a constraining variable. In Exhibit 20.4(b) we show the fish group centroids as well as the convex hulls around the fish species in each group. There is some overlap because not all of the inter-group variance, contained in seven dimensions (one less than the number of groups), can be shown in the two-dimensional map.

Definition of functional diversity

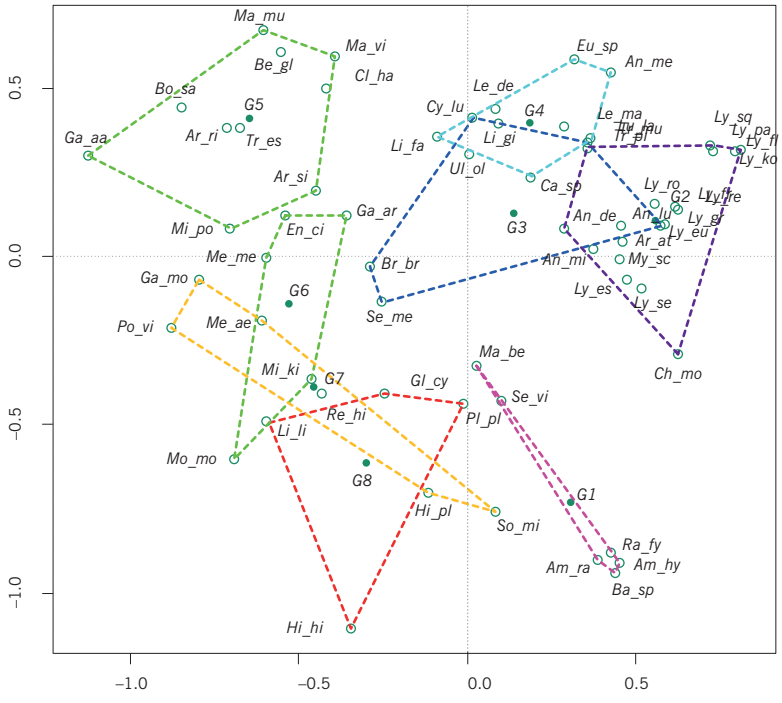
Once the tree, or dendrogram, given in Exhibit 20.3 is established, there are two ways to define functional diversity at a sampling site, one of which depends on having decided on the number of groups, as we have already done above, and the other which only needs the tree. The former is simple to understand: given a sample of fish at a site along with their abundance values, they are classified into groups and their abundance values are aggregated. Then a standard

Exhibit 20.4: *CA of the trait matrix aggregated according to the fish groups (G1 to G8) that were defined in Exhibit 20.3. The solution optimizes the group differences, although the basic configuration is similar to that of Exhibit 20.2 which optimized the fish differences. The functional traits are displayed in contribution coordinates in (a). 52.4% of the inertia between fish groups is displayed*

(a)



(b)



diversity measure is computed, for example the Shannon-Weaver index, denoted by H':

$$H' = -\sum_g p_g \log(p_g)$$

where p_g is the proportional abundance of functional group g.

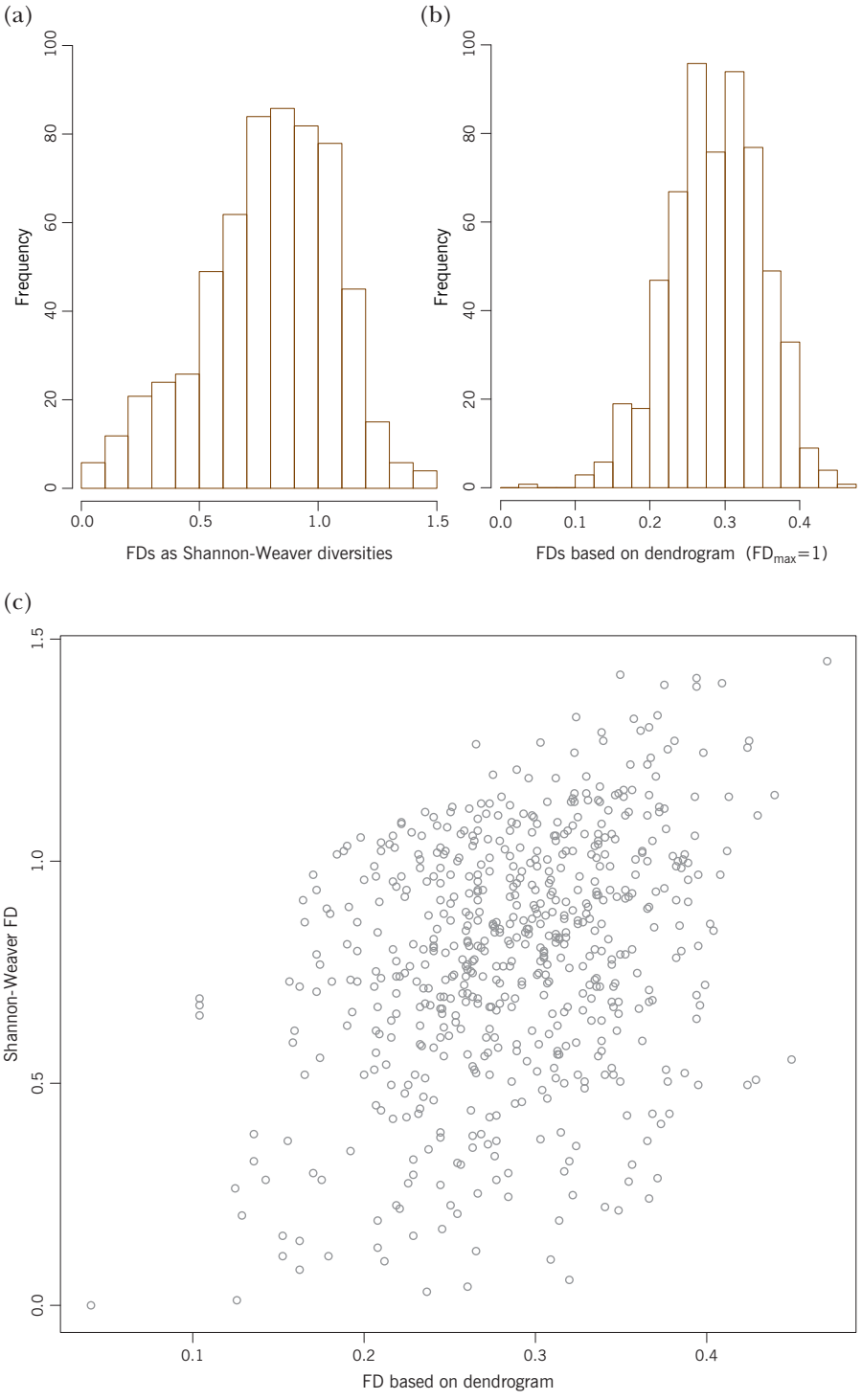
The other way is to measure diversity by summing branches on the dendrogram according to the mix of fish species found in the sample – in this case only presences of fish are used and not their abundances, although an abundance-weighted measure can be envisaged. First let us suppose that a sample contains all 62 fish, which would give the maximum diversity possible. The measure of diversity is obtained by summing all the vertical branches in Exhibit 20.3: since each of the $n - 2$ nodes of the tree has two vertical branches below it, this is the sum of $2(n - 1) = 122$ values in this example, equal here to 20.31. This value is called the *functional diversity* of the *species pool* (henceforward, we use the abbreviation FD for functional diversity).

Now for a general sample that contains only a subset of the fish species, the FD value is computed by summing all the branches linking this subset. Clearly, if the fish in a subset are "close" together in terms of trait distance, then the sum of the associated branches will be relatively low, while if there are fish species in the sample that are "far apart" with not so many common traits, then the sum of their linking branches will be relatively high. To normalize the FD measures, we shall express them relative to the maximum value of 20.31 for the species pool, so that FD will be between 0 and 1.

Using the same data as in Chapter 19, the FD values for each of the 600 sampling sites were computed in the two different ways, first as the diversity index H' taking into account the aggregated abundances, and second as the normalized value lying between 0 and 1 that only uses presences of the fish. Exhibit 20.5 shows the histograms of FD for each alternative, as well as a scatterplot of their paired values. The fairly low rank correlation of 0.3 suggests that these two measures reflect different information about the diversity.

Interestingly, it is feasible to make a permutation test on the species pool FD as an alternative test for overall clusteredness of the fish, different from testing for a particular number of groups. If the trait data are randomly permuted within each variable, e.g., within diet the three options are permuted together across the fish (and not separately), many alternative values of the species pool FD can be obtained, under a null hypothesis of no relationship between the traits. Exhibit 20.6 shows that the observed FD value is much lower than those obtained under the

Exhibit 20.5:
(a) Histogram of the group-based FDs defined as Shannon-Weaver diversities on the aggregated abundances in 600 samples for eight functional groups; (b) Histogram of the tree-based FDs using presences only and summing the branches in the dendrogram for the subset of observed species, normalized with respect to the FD of the species pool; (c) Scatterplot of the two functional diversity indices (Spearman rho correlation = 0.300)



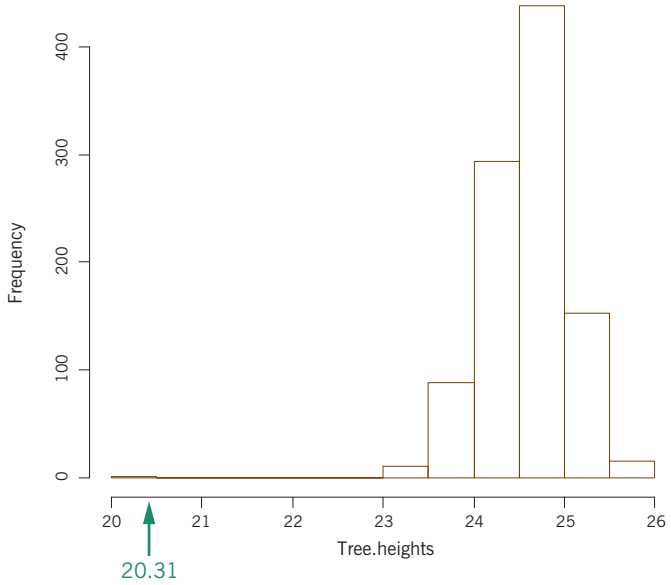


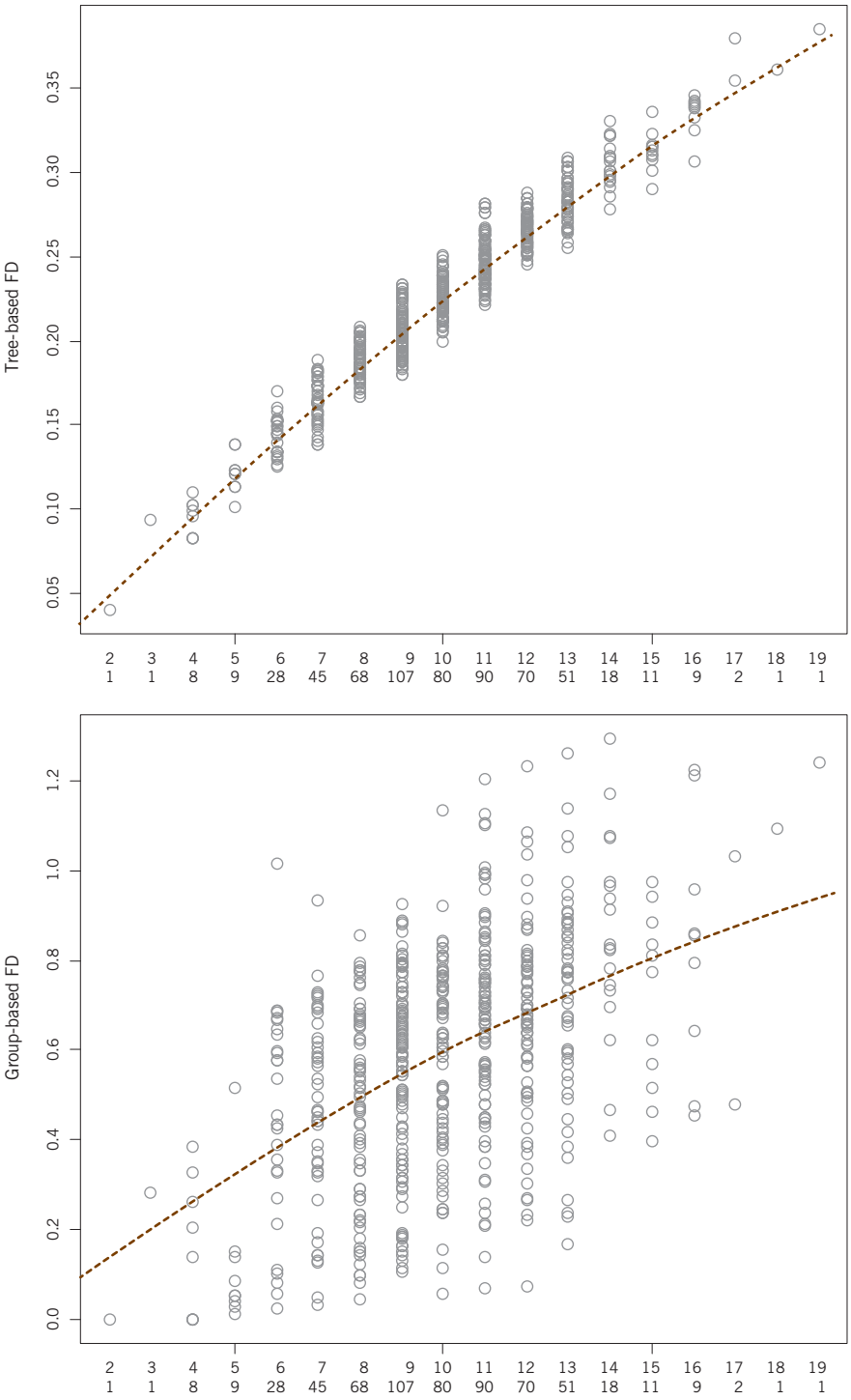
Exhibit 20.6:
Permutation distribution of the species pool FD, under the null hypothesis of no relationship between the traits. The observed value of 20.31 is the smallest and the associated p*-value, based on 1,000 permutations, is thus* p = *0.001*

null hypothesis, and the estimated p-value is $p = 0.001$, showing that there are significant similarities between the fish across the traits.

Relating functional diversity to species richness

Both measures of functional diversity will be used and compared in the remainder of this chapter, since they appear to contain different information and are defined in different ways, the common feature being the dendrogram based on the trait distances. Exhibit 20.7 shows their relationships with species richness (SR), that is the number of species in each sample. Since the tree-based FD only takes presences into account and would clearly increase with increasing number of species, as more branch lengths are summed, it is no surprise that it follows species richness very closely (Exhibit 20.7(a)). Both relationships are slightly nonlinear, with concave curves, and so we would use a quadratic function, for example, as a model for the conditional means, shown in Exhibit 20.7 (in both cases the explanatory terms SR and SR^2 are highly significant in the regressions). The deviations of the functional diversity values from that expected by their relationship with species richness are used as a measure of so-called *functional dispersion*. Higher functional dispersions at a site are associated with greater ecosystem adaptability because the number of functions displayed by the species at this site is higher than expected given the number of species present – they possess more "tools" and are thus expected to be better prepared for environmental change. On the other hand, the impact on the FD due to the loss of a species would be proportionally larger at this site since each species contributes more to the FD as compared to a site with the same SR but a lower FD (i.e., lower functional dispersion).

Exhibit 20.7:
Scatterplots of the two FD measures versus species richness (SR, the number of species in sample), showing the modelled quadratic relationships. The horizontal axis is marked with the value of SR, and below the number of sites with the corresponding value)



Relating functional diversity to space, time and environment

As a first bivariate view of associations between the two FD measures and the available covariates, Exhibit 20.8 shows the matrix of scatterplots, with Spearman rank correlations in the upper triangle and scatterplots and smooth relationhips in the lower triangle. Apart from the known features of the region, that depth is negatively correlated with longitude and temperature negatively correlated with latitude and longitude, the group-based FD is correlated with depth and the tree-based FD negatively with temperature and positively with latitude and longitude, although these last correlations are less than 0.30 in absolute value. As already seen in Chapter 19, the variable slope does not appear to have any association with any other, so we drop it from further consideration.

To show the spatial relationship latitude and longitude should be considered together along with their interaction. We can compare two ways of spatial modelling, by spatial fuzzy coding (Chapter 11) and by generalized additive modelling (GAM,

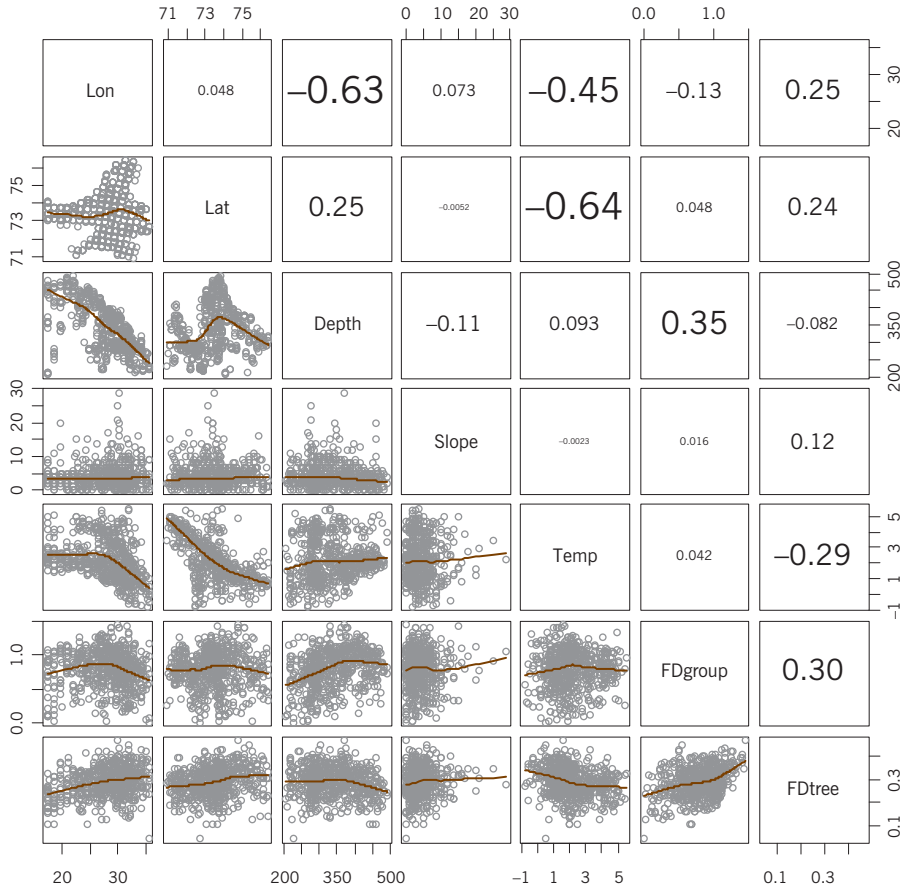
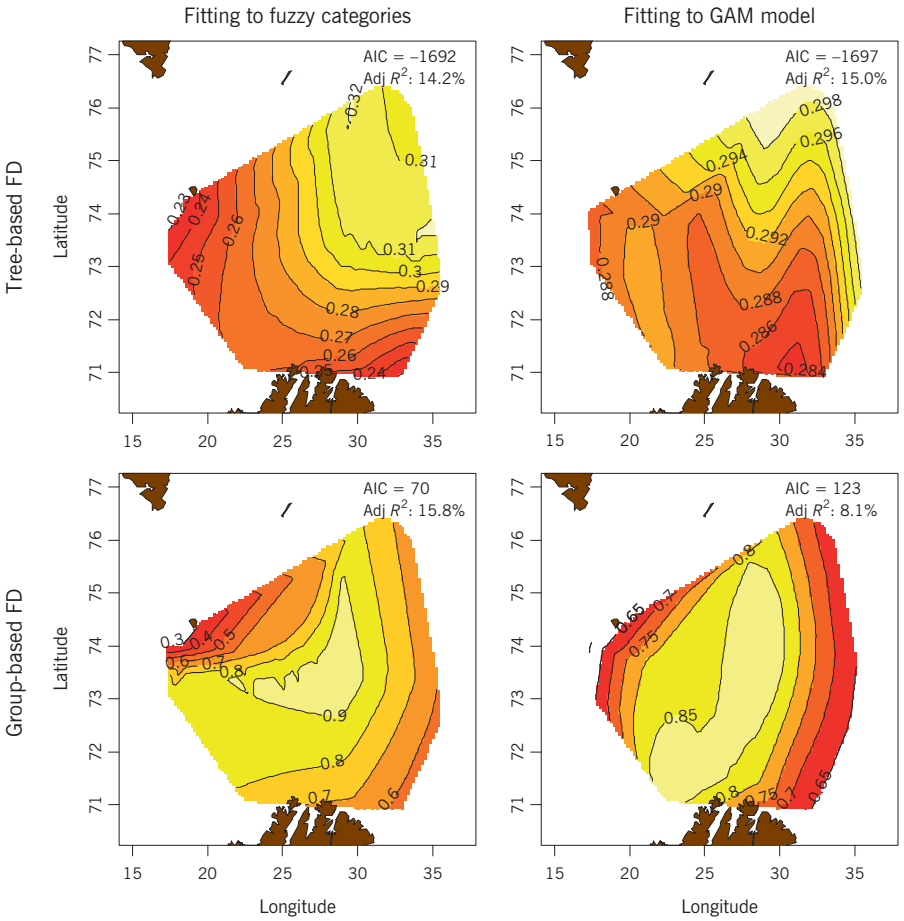


Exhibit 20.8:
Scatterplots of the variables depth, slope, temperature, longitude and latitude with one another as well as with the two measures of functional diversity, based on the functional groups (FDgroup) and on the dendrogram (FDtree). Spearman rank correlations are shown in the upper triangle, with font size proportional to their absolute values.

Chapter 18). For both models we model each FD measure on latitude and longitude interactively, and the results are shown in Exhibit 20.9 in the form of contours of predicted FD values.

The results or the two types of FD are quite different, with the tree-based FD showing a west and south to north-east gradient, with higher FD in the north-east, whereas the group-based FD shows higher diversity in the central area, falling off to the west and the south-east. Remember that the group-based FD takes the abundance values into account and the water in the central areas is warm, and species from more southern areas (e.g., Norwegian Sea) migrate into these areas (often in schools), especially in warmer years, giving a more equal spread of relative abundance values in the functional groups. Concerning the tree-based FD, the GAM fit shows a ridge in the diversity values from south to north while the fuzzy

Exhibit 20.9:
Contour plots of the spatial component of functional diversity according to the two definitions (first row is the tree-based FD, second row is group-based FD) using two modelling methods (in columns, first column is using fuzzy spatial categories, second is using GAM modelling). The northern border of Norway with Russia and the southern tip of Svalbard situate the region of interest



fit shows a wider ridge from south-west to north-east. For the group-based FD the results are similar between the two methodological approaches, but the fuzzy approach performs noticeably better according to AIC and the adjusted R^2. Another advantage of the approach using fuzzy-coded categories is that there is a p-value associated with every compass point's difference with the central category. So we can get results that for the group-based FD several sectors are significantly lower than the central (C) one: NW, E, SE and S (all with $p < 0.001$), NE ($p = 0.002$) and W ($p = 0.04$), whereas for the tree-based FD the following sectors are significantly lower than the central one: NW ($p < 0.001$), SE ($p = 0.002$) and S ($p = 0.005$).

Although the spatial variation is highly linked to the variation of environmental variables such as temperature and possibly also to temporal variation, we can study inter-year variation in the residuals from the above spatial models as well as any further relationships with the environmental variables temperature and depth. As an example, we consider the residuals of tree-based classification from the fuzzy spatial model (top left example in Exhibit 20.9), and model the residuals on year as a categorical variable, and temperature and depth either as regular continuous variables, or the four-category fuzzy versions used in Chapter 19, or as smooth functions using GAM. Both temperature and depth are found to be nonsignificant predictors of the residuals, irrespective of the coding. There is significant temporal variation, however, almost identical in all analyses, which can be plotted as in Exhibit 20.10. Remembering that these are the residuals from the spatial model, we can say that in 1999 and 2001 there were lower functional diversities compared to the spatial model (as measured by the dendrogram-based approach) and higher in 2003 and 2004. All effects are different from 0 (the mean of the residuals) and highly significant ($p < 0.0001$), apart from 2002 which is closer to 0 ($p = 0.025$).

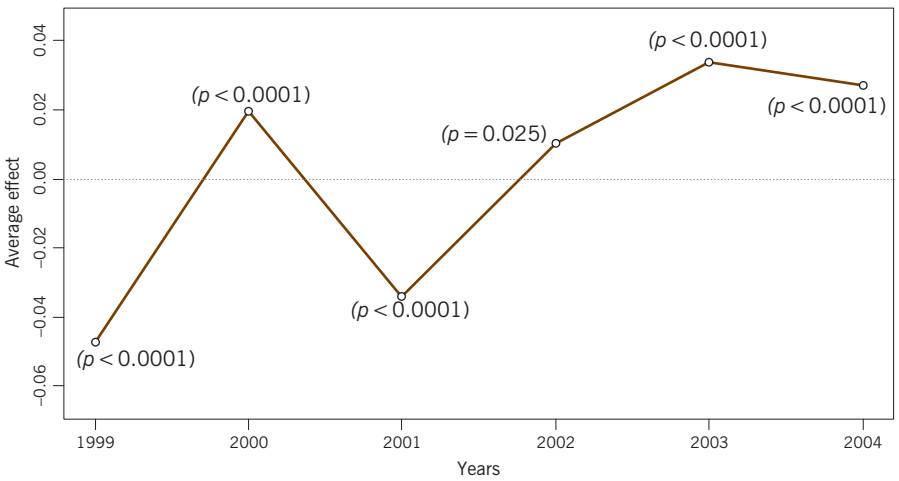
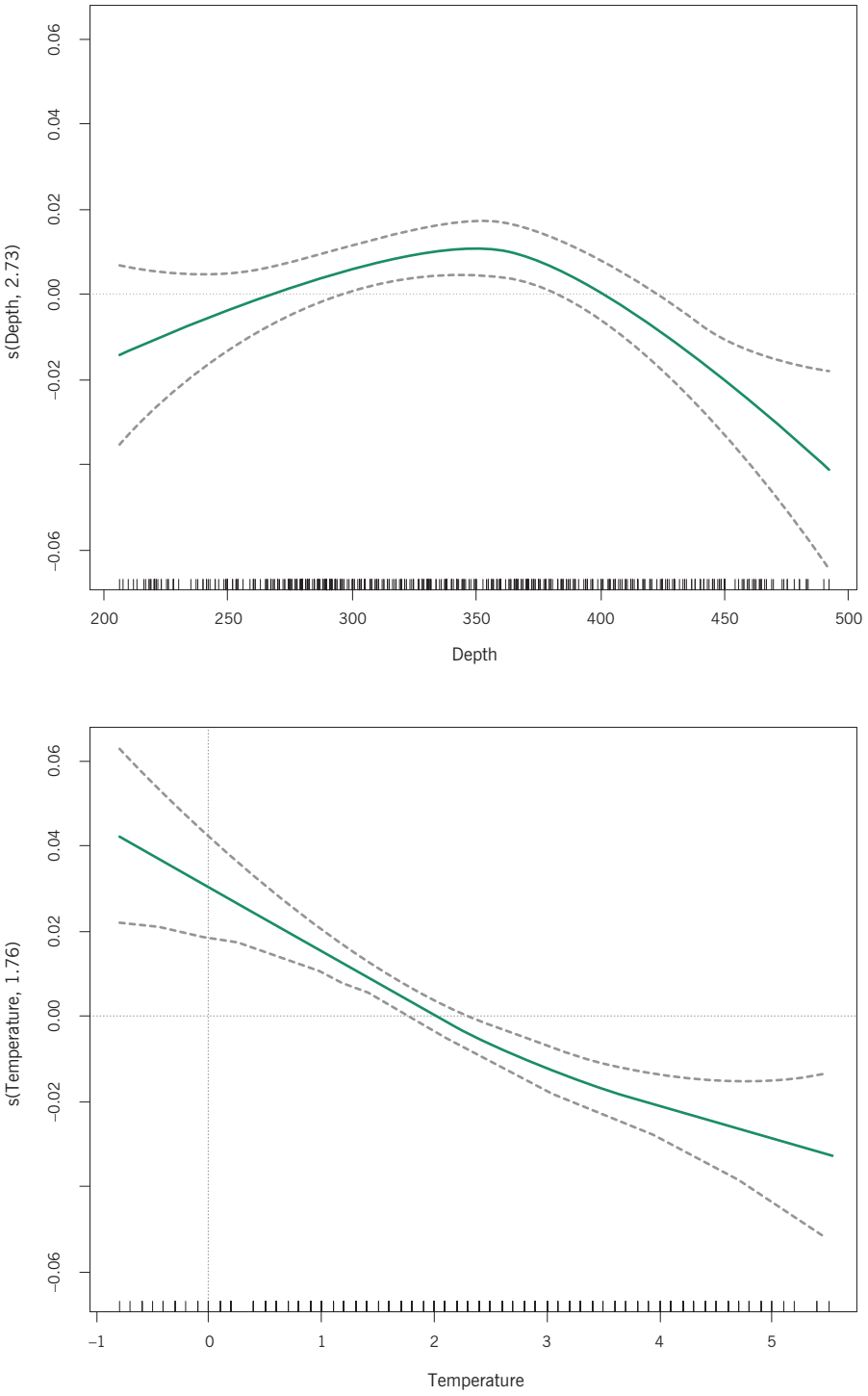


Exhibit 20.10:
Plot of regression coefficients for each year showing average estimated year effects for the residuals of (tree-based) functional diversities from the spatial model, with p*-values for testing differences compared to the zero mean of the residuals (dashed line)*

Exhibit 20.11: *GAM of tree-based FD as a smooth function of depth ($p = 0.0001$) and temperature ($p < 0.0001$). To model these effects parametrically depth would be modelled as a quadratic and temperature linear*



From the above it is clear that the effects of spatial position and of the environmental variables temperature and depth are confounded and difficult to separate. If FD is first related to temperature and depth, ignoring the spatial component, highly significant relationships are found: for example, tree-based FD goes down with increasing temperature and we find the same quadratic relationship with depth as in Chapter 18 where the response was species diversity – see Exhibit 18.7 for the analysis of only 89 sites (where temperature was nonsignificant), and Exhibit 20.11 for the present example of 600 sites. More or less the same depth value, about 350 m, is found here for maximum FD as was found before in Exhibit 18.7 for maximum species diversity. Adding the year effects gives almost exactly the same pattern as in Exhibit 20.9, with 1999 and 2001 low and the other years high. Of the FD variance, 30.2% (adjusted R^2) is explained by depth, temperature and years. Residuals from this environmental and temporal relationship, accounting for about 70% of the FD variance, can then be modelled spatially: using a GAM model as in Exhibit 20.8 there is still a significant spatial component in the residuals, although the explained variance in these residuals is only 4.3%.

In summary, temperature and depth, both of which are related to spatial position in the Barents Sea, are found to be strongly associated with functional diversity, and there are also significant differences between the years. Residuals from a model of FD as a function of these environmental and temporal variables can be explained, although to a minor extent, by spatial position.

SUMMARY: Functional diversity of fish in the Barents Sea

1. Functional diversity measures diversity in the *functional traits* (feeding, motion, reproductive behaviour, habitat preferences, etc.) among species in an ecosystem.

2. Functional groups are groups of species that share the same functional traits.

3. To measure functional diversity two approaches are considered here, both based on a dendrogram obtained by hierarchical clustering of the species according to their functional traits. They are thus both dependent on the distance/dissimilarity function used as well as the type of clustering.

4. The first way is to use the hierarchical clustering to decide on the number of clusters that are sufficiently homogeneous internally to be considered separate groups. Functional diversity (FD) at a site can then be measured by any of the usual diversity measures, for example the Shannon-Weaver diversity, which is a function of relative abundances (or biomasses) of the functional groups. We call this *group-based FD*.

5. The second way is to add up the branches of the dendrogram of the particular mix of species at the site – this takes only presences of species into account, not their abundances. We call this *tree-based FD*.

6. These FD measures are found to have monotonically increasing, slightly concave, relationships with species richness (SR). Tree-based FD is very closely related to SR because both take only species presences into account.

7. Both FD measures can be related to spatial, temporal and environmental variables in the usual way using multiple regression. Spatial coordinates are interactively coded to explain the spatial relationship. Continuous explanatory variables can be coded in their original form, possibly transformed to account for nonlinear relationships, or coded as fuzzy variables.

8. An alternative modelling strategy is to use generalized additive modelling (GAM) which produces a smooth regression relationship with the two-dimensional spatial position and the continuous variables.

List of Exhibits