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 19 Offprint

Case Study 1: Temporal Trends and Spatial Patterns across a Large Ecological Data Set

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

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

Fundación BBVA

CHAPTER 19

Case Study 1: Temporal Trends and Spatial Patterns across a Large Ecological Data Set

The examples presented in previous chapters have generally been on small- to medium-sized data sets that are good for teaching and understanding the basic concepts of the methodologies used. We conclude with two chapters detailing much larger studies that take full advantage of multivariate analysis to synthesize complex phenomena in a format that is easier to interpret and come to substantive conclusions. The two chapters treat the same set of data, a large set of samples of fish species in the Barents Sea over a six-year period, where the spatial location of each sample is known as well as additional environmental variables such as depth and water temperature. In the present chapter we shall study the temporal trends and spatial patterns of the fish compositions and also try to account for these patterns in terms of the environmental variables. But before applying multivariate analysis to data across time and space, we have to consider carefully the areal sampling across the years and reweight the observations to eliminate sampling bias.

Contents

Sampling bias 249
Data set "Barents fish trends" 251
Reweighting samples for fuzzy coded data 251
Correspondence analysis of reweighted data 253
Canonical correspondence analysis of reweighted data 255
Some permutation tests 255
Isolating the spatial part of the explained inertia 258
SUMMARY: Case Study 1: Temporal trends and spatial trends across a large ecological data set .. 260

Sampling bias

In this chapter we shall be considering samples in different regions over time over an area of interest. An important consideration is whether data have been collected in a balanced way over time in each region. This is important if one wants to summarize the data over the whole area and make temporal comparisons. If

sampling is more intense in some regions in some years and less intense in other years, this can lead to what is called *sampling bias*. Consider in Exhibit 19.1(a) the hypothetical layout of numbers of samples taken over an area divided into three regions, for three consecutive years.

Let us assume for the moment that the total number of samples over the three years is representative of the sizes (or some measure of importance in the study) of the three regions surveyed, that is region 2 is the largest, followed by region 1 and then region 3. Then, for the sampling to be balanced over the years the numbers of samples should follow the proportions 0.300, 0.433 and 0.267, as seen in the last line of Exhibit 19.1(a). Computing expected proportions for each year, in exactly the same way as one computes expected frequencies in a chi-square test, the table in Exhibit 19.1(b) is obtained. If this latter table of expected frequencies is now divided, cell by cell, by the former table of actual frequencies, a table of *weights* is obtained in Exhibit 19.1(c), reflecting the imbalances.

In Exhibit 19.1(c) the column of ones for region 1 shows that the sampling was in perfect proportion to the expected number. In contrast, region 2 is under-

Exhibit 19.1:
(a) Actual number of sample taken in three regions over a three-year period, with overall proportions of samples in each region over the whole period. (b) Expected number of samples if in each year sampling had taken place in accordance with the overall proportions. (c) The weights computed by dividing the values in table (b) by those in table (a)

(a)

	Region 1	*Region 2*	*Region 3*	*Sum*
Year 1	30	20	50	*100*
Year 2	15	30	5	*50*
Year 3	45	80	25	*150*
All years	90	130	80	*300*
Prop'n	0.30	0.433	0.267	

(b)

	Region 1	*Region 2*	*Region 3*	*Sum*
Year 1	30	43.3	26.7	*100*
Year 2	15	21.7	13.3	*50*
Year 3	45	65	25	*150*
All years	90	130	80	*300*
Prop'n	0.30	0.433	0.267	

(c)

	Region 1	*Region 2*	*Region 3*
Year 1	1	2.167	0.533
Year 2	1	0.722	2.667
Year 3	1	0.813	1.600

sampled in year 1 and over-sampled in years 2 and 3. In year 1 this region has only 20 samples whereas the expected proportion of 43.3% of 100 is 43.3. The weight of 2.167 is then used to scale up the abundances of observed species in this region. In year 2 the 30 actual samples represent an over-sampling compared to the expected value of 21.7, and these 20 samples are thus down-scaled by a factor of 0.722, and so on. Each of the 300 samples thus receives a weight in Exhibit 19.1(c) according to its year and region, some up-weighted, others down-weighted – notice that the sum of the weights allocated to the 300 samples is equal to 300.

This reweighting is not necessary if the regions are studied one by one, for example average abundances or measures of diversity can be compared within a region using the original unweighted data. However, whenever the regions are put together to estimate a value over the whole area, the weighting will be necessary. Consider, for example, if in region 3 a certain species where particularly abundant. Since this region is heavily sampled in year 1, almost twice as much compared to the expected proportion, the unweighted data in year 1 could show a difference with the other years which is due to this oversampling. Of course, we are assuming that the proportions in the last row of Exhibit 19.1(a) reflect the "population" proportions, but these can be determined by an external criterion such as the area of each region.

Data set "Barents fish trends"

In Chapter 11 the data set "Barents fish" was introduced, a relatively small data set of fish abundances of 30 species at 89 sites in the Barents Sea, during a sampling period in 1997 (Exhibit 11.2). The geographical location was handled in different ways, first by defining a spatial grouping of the samples (Exhibit 11.1), second using latitude and longitude as continuous variables (Exhibits 11.5 and 11.6) and third by defining fuzzy positions with respect to eight compass points and a central category (Exhibits 11.8 and 11.9). In this case study we extend the data set to six consecutive years of data, from 1999 to 2004, called *Barents fish trends*, thus introducing a temporal component into the study. A total of 600 samples are included. We will implement a reweighting scheme in this application, explaining how the previous argument for "crisp" regions can be extended quite naturally to our fuzzy coding of the spatial positions.

Reweighting samples for fuzzy coded data

We are going to use fuzzy coding again to code the geographical position of each sample, as described at the end of Chapter 11. If each of the 600 samples had been allocated "crisply" to one of 9, say, regions, then we would proceed as just explained by counting how many samples were in each region in each year to check if proportionally the same number of stations were sampled from year to year. The situation is hardly different for the fuzzy coding, thankfully, since we can sum the fuzzy values and not the zero-one dummy variables for the region

categories. Exhibit 19.2 shows the sums for each year and for the whole period along with the overall proportions for each region.

To balance the allocation to each region each should follow the overall proportions, that is SW should have 2.5% of the 88 samples, i.e. 2.20 (so it is slightly over-represented, since the actual value is 2.33, W should have 15.5% of 88, i.e. 13.64 (again under-represented, actual value is 11.71), and so on. If we continue computing the expected values and comparing them with the observed ones in Exhibit 19.2, the ratios expected/observed give a matrix of weighting factors in Exhibit 19.3.

Exhibit 19.2:
Sums of fuzzy-coded regional categories for each year and for all years. Columns are the eight compass points and a central region (C)

	SW	W	NW	S	C	N	SE	E	NE	*Sum*
1999	2.33	11.71	1.16	10.12	31.25	12.57	3.57	11.29	4.00	*88*
2000	4.60	18.86	1.47	14.32	39.27	11.65	2.47	11.43	2.93	*107*
2001	0.64	11.14	1.15	6.60	31.93	12.73	3.06	12.18	4.57	*84*
2002	2.46	15.83	1.41	11.83	37.44	12.50	2.22	11.01	4.30	*99*
2003	2.02	16.44	1.36	14.93	38.55	6.63	6.28	12.04	1.75	*100*
2004	2.72	18.87	1.61	15.46	45.42	14.56	4.17	14.37	4.80	*122*
All years	14.76	92.85	8.17	73.24	223.85	70.64	21.79	72.33	22.36	*600*
Prop'n	0.025	0.155	0.014	0.122	0.373	0.118	0.036	0.121	0.037	

Exhibit 19.3:
Weights for data according to year and fuzzy region

	SW	W	NW	S	C	N	SE	E	NE
1999	0.929	1.163	1.033	1.062	1.051	0.824	0.895	0.940	0.820
2000	0.572	0.878	0.991	0.912	1.017	1.081	1.573	1.128	1.361
2001	3.229	1.167	0.994	1.554	0.981	0.777	0.997	0.831	0.685
2002	0.990	0.968	0.956	1.022	0.987	0.932	1.619	1.084	0.858
2003	1.218	0.941	1.001	0.818	0.968	1.776	0.578	1.001	2.130
2004	1.104	1.001	1.032	0.963	1.002	0.987	1.062	1.023	0.947

So in category SW (south-west) 2000's samples must be downweighted by a factor of 0.572, whereas 2001's samples must be upweighted by 3.229.

Since the samples do not fall strictly into a region, how can these weights be applied? For example a particular sample in 2000 is coded spatially as follows:

SW	W	NW	S	C	N	SE	E	NE
0.125	0.862	0.000	0.002	0.011	0.000	0.000	0.000	0.000

(19.1)

i.e., it is mostly in the western section, but a bit towards south-west, and quite far from the centre (remember that you can find the exact position of this station from the fuzzy coding). Now each value that we observe in this sample, for example an abundance value of 19 for the species *Sebastes mentella* (*Se_me,* beaked redfish) is split between the fuzzy categories in the above proportions, after which the weights for the year 2000 in Exhibit 19.2 are applied. This means that we can compute a sample-specific weight as the weighted average of the weighting factors:[1]

$$0.125 \times 0.572 + 0.862 \times 0.878 + 0.002 \times 0.912 + 0.011 \times 1.017 = 0.8413 \quad (19.2)$$

Hence, all the abundance data for this sample are downscaled by the factor 0.8413, e.g., for the *Se_me* value of 19, $0.8413 \times 19 = 15.99$. Weights for all the samples are computed in the same way, and the sum of these sample weights is equal to the sample size, 600 in this case. The weights can be used to adjust the abundances as well as in computing regression relationships, using weighted regression, or in computing overall measures over the whole study area such as means and diversity measures, for example, which are then appropriately re-weighted to compensate for sampling biases in different areas.

In most applications such as this one, where the sampling is not drastically out of proportion from year to year, it is not going to make a big difference to the results of a multivariate analysis whether one uses the original abundance matrix or the reweighted one – nevertheless, reweighting is an insurance against possible sampling bias. The negative side of this approach, however, is that in a severely under-sampled region such as the south-western region in 2001 (Exhibit 19.2) there might be some unusual samples that then become up-weighted and thus over-emphasize the species (or lack of species) in that region, so we should still have a certain minimum sample size in each area and each year to avoid estimation bias. In what follows, we will consistently use the reweighted data set and can report in passing that the results are very similar when compared to those of the unweighted data.

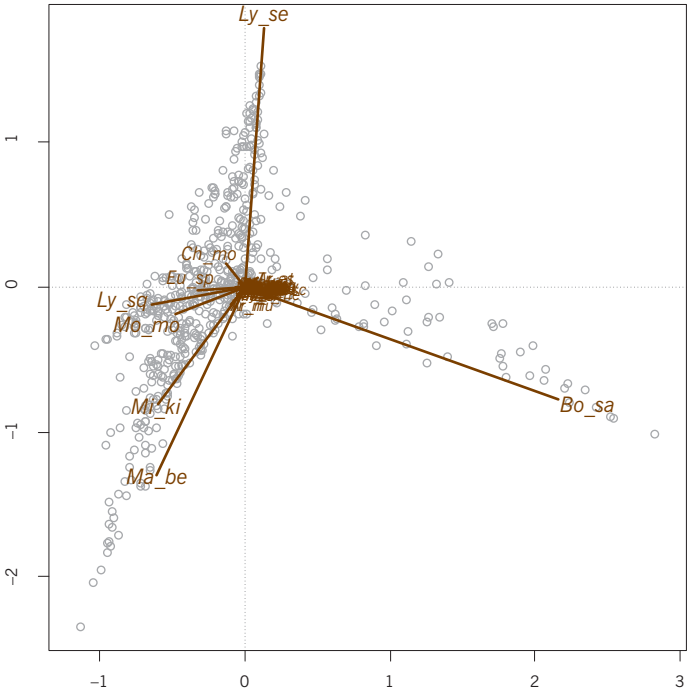
Correspondence analysis of reweighted data

Exhibit 19.4 shows the CA of the abundance matrix, first the samples (gray circles) and species (brown abbreviated labels) and then an enlargement of the central area showing the centroids of the year points and all the fuzzy categorical variables. Six species contribute more than average to the axes, shown with bigger labels. The first CA axis separates species dominating in cold Arctic waters from species found

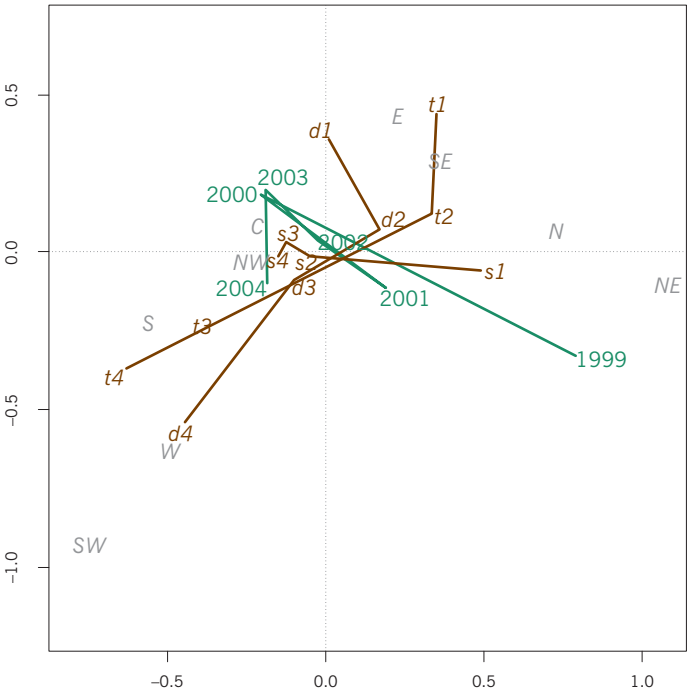
[1] There are three different weights here: (1) the fuzzy coded values in (19.1) that add up to 1, which will be used as weights in the weighted averaging (only four of them are nonzero); (2) the fuzzy-region-specific weighting factors in Exhibit 19.2; and (3) the final value of 0.8413 which is a weight to apply to the abundances of this particular sample.

Exhibit 19.4: *Correspondence analysis contribution biplot of the "Barents fish trends" data set. The upper plot shows the active data, the samples and species (high-contributing species are shown with bigger labels).The lower plot shows the centroids of all the categories, linking together categories of ordinal variables. 32.6% of the total inertia of 4.017 is explained by these two first dimensions*

(a)



(b)



in warmer Atlantic waters, in other words a latitudinal effect. The second axis separates species in shallower waters from those found in deeper waters.

Canonical correspondence analysis of reweighted data

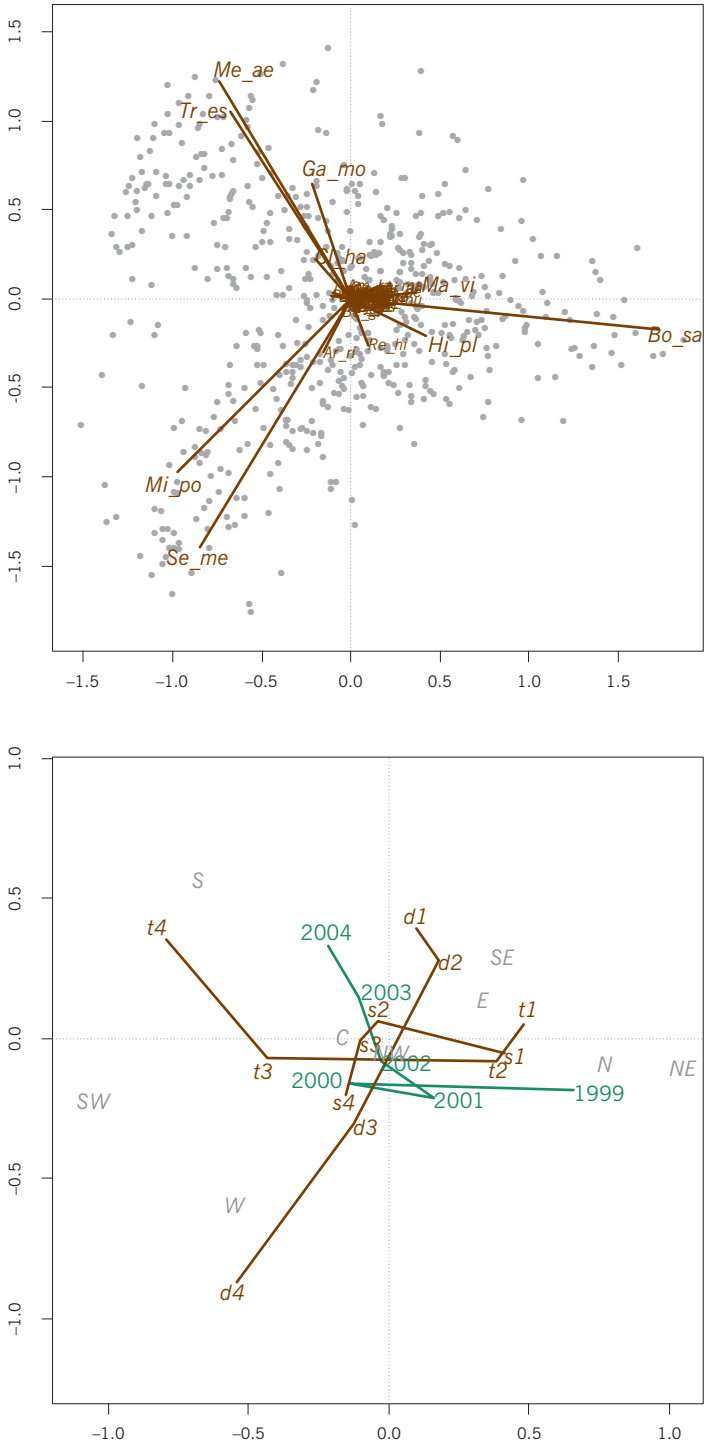
The CA will show the main dimensions of the 600 samples in the abundance matrix without specific reference to the interannual differences and differences across temperature, depth, slope and in space. The CCA will look at the dimensions of abundance that are in the space of the all these explanatory variables. Each set of dummy or fuzzy variables contributes one less than its number of categories to the dimensionality of the restricted space in which CCA operates: time (6 – 1 = 5), space (9 – 1 = 8) and depth, temperature and slope (4 – 1 = 3, each), totalling 22. This 22-dimensional restricted space contains 37.7% of the total inertia of the data, in other words there is 62.3% of the inertia that is unrelated to the explanatory variables. Exhibit 19.5 shows the result in the same format as Exhibit 19.4. If the scale of the lower centroid plot is compared to that of the centroid plot in Exhibit 19.4, it is clear that the categories are more spread out, which is the objective of the CCA to discriminate maximally between the categories. The high-contributing fish species have changed now, apart from *Bo_sa* which still maintains its important position on the first dimension, separating year 1999. The cloud of samples in upper left are associated with species extending out in that direction of the ordination, found in warm water coming from the Atlantic in the south, while the cloud of samples at bottom left is associated with deep water species found in the western area. The temporal trend is now clearer, with years 2003 and 2004 tending even more towards the warmer area of the map.

Each year points shows the centroid of all the samples for a particular year, grouping all the fuzzy regions. A trajectory for each region can be indicated as well, this time as supplementary points – that is, we fix the CCA solution and compute centroids for regional subsets of samples over the years. Exhibit 19.6 shows the regional trajectories for the categories N, E, W and S as well as their overall spatial and time centroids that were shown in Exhibit 19.5. It can be seen now that, of these four regions shown, it is mainly the southern and eastern regions that continue moving towards the "warm" region of the map in 2003 and 2004, whereas in the northern and western regions the warming trend stops from 2003 to 2004. In this way one can interpret the interaction between space and time, seeing the difference in trends between regions, or equivalently the difference in spatial patterns over time, while the six year points and nine region points show the average time trend and spatial pattern.

Some permutation tests

We can conduct various permutation tests to make conclusions about the statistical significance of the CCA results. A first test can be to confirm, as we surely believe, that the association between the abundance data and all the environmental data is significant. The environmental data set is kept fixed

Exhibit 19.5:
Canonical correspondence analysis of the "Barents fish trends" data. The format is the same as Exhibit 19.4, with the samples and species plotted in the upper biplot and an enlarged version of the category centroids in the lower plot. 58.5% of the restricted inertia is explained by these two dimensions



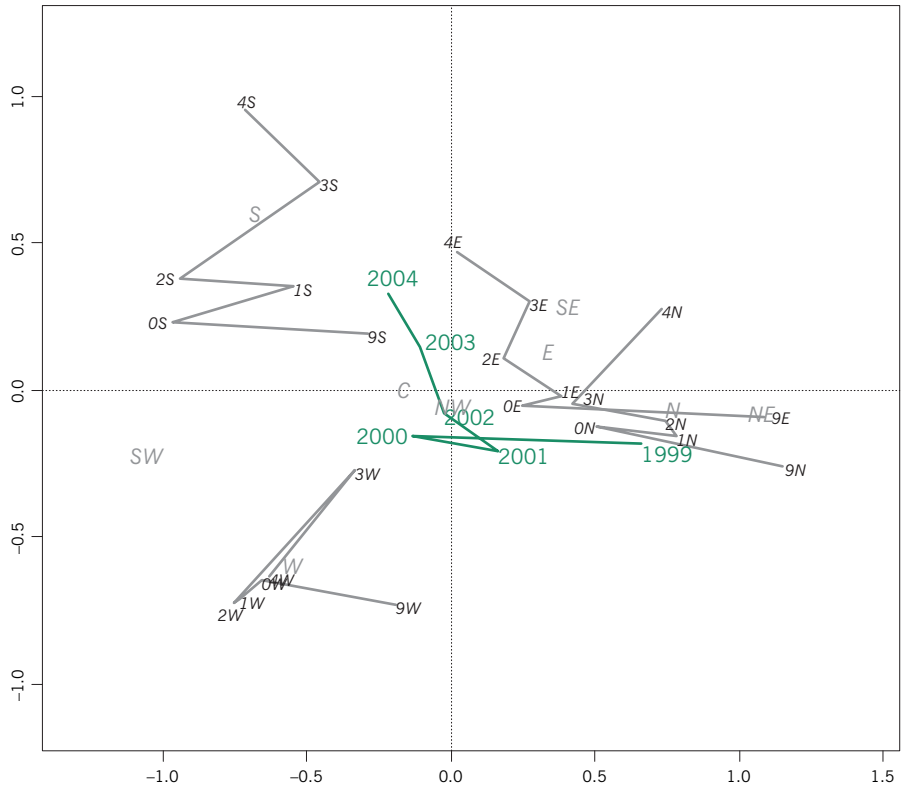


Exhibit 19.6:
Temporal trajectories in regional categories north, east, south and west. Time and regional centroids are at (weighted) averages of the corresponding category points: for example, S is at the average of the six points making up the trajectory for south, while 2004 is at the average of all the 2004 points (for all nine regions, only four shown here)

and the samples in the abundance data set are permuted many times. In 1,000 permutations the highest inertia explained is by the original data, so the significance is $p = 0.001$ at most. What is more interesting is to see the significance of individual variables. Using them one at a time as constraining variables, the associated p-values are all highly significant ($p = 0.001$) except for slope ($p = 0.12$). Ordering them by explained inertia, Exhibit 19.7 shows the percentage of variance explained, denoted by R^2 because it is the direct analogue of the coefficient of determination in regression, as well as an analogue of the

Variable	k	R^2	Adjusted R^2
Spatial	9	0.275	0.242
Temperature	4	0.119	0.104
Depth	4	0.086	0.070
Year	6	0.057	0.030
Slope	4	0.024	0.007

Exhibit 19.7:
In descending order, the proportion of inertia explained, R^2, and adjusted R^2, of the five categorical environmental variables; k *is the number of categories*

adjusted R^2 which takes into account the number of categories (see Appendix A and B for details). The spatial position of the sample has by far the most explanatory power.

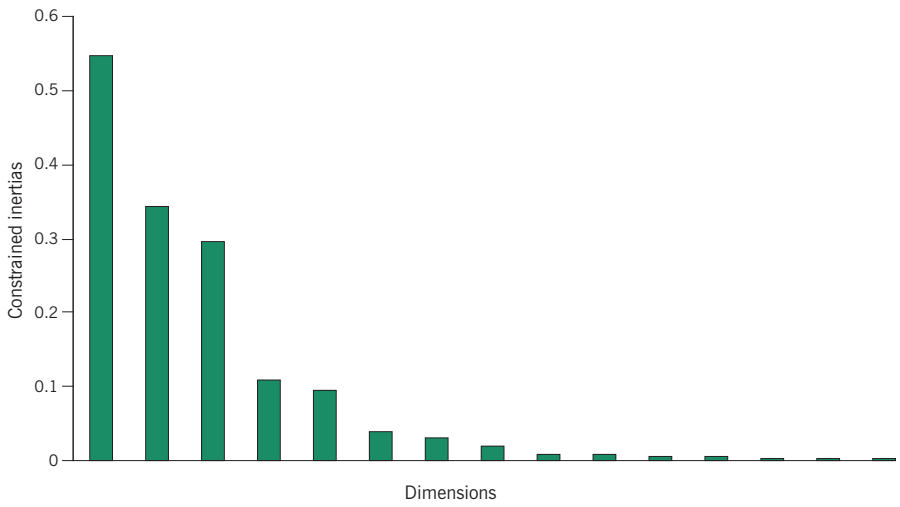
Another statistical aspect of the constrained ordination solution that requires investigation is the dimensionality of the solution. The scree plot of inertias on successive dimensions is shown in Exhibit 19.8, suggesting a three-dimensional solution. A permutation test for the percentages of constrained inertia, as described in Chapter 18, but applied to the inertias on the constrained dimensions, confirms without any doubt that there are actually three significant dimensions in the constrained space. On the supporting website of this course there is a video of the three-dimensional ordination, which gives an idea of this additional dimension and the 19.6% additional inertia it accounts for.

Isolating the spatial part of the explained inertia

Because the spatial component is intimately related to the environmental variables, especially temperature, it is possible to use CCA to isolate which part of the constrained inertia is purely due to the spatial component and not confounded with the environmental variables. A partial CCA is used, which involves first partialling out the effect of one set of variables, and then doing a CCA on the residuals using a different set of constraining variables. The steps in separating contributions to inertia of inter-correlated variables are as follows:

– Perform the CCA with all constraining variables, in this case environmental, temporal and spatial: the inertia in the constrained space is 1.4746, i.e. 36.7% of the total inertia of the abundance data of 4.0170.

Exhibit 19.8:
Scree plot of the inertias of successive dimensions in the constrained space of the CCA of the "Barents fish trends" data. The first three dimensions clearly stand out from the rest



– Perform the CCA with the environmental and temporal variables constraining: the inertia in the constrained space is now 0.9761, i.e. 24.3% of the total.

– Perform the CCA with the spatial variables constraining: the inertia in the constrained space is now 1.1044, i.e. 27.5% of the total. It is clear that there must be some confounding between the spatial variables and the others, because 24.3% + 27.5% is much higher than 36.7%, that is the total constrained inertia including all variables in a CCA.

– Perform the CCA with the environmental and temporal variables constraining after partialling out the spatial variation: the inertia in this constrained space is 0.3703, i.e. 9.2% of the total. This 9.2% of the inertia is due to the environmental variables only in a space uncorrelated with the spatial variation.

– Similarly, perform the CCA with the spatial variables constraining after partialling out the other variables: the inertia in the constrained space is now 0.4982, i.e. 12.4% of the total.

– From the last two calculations it must be that 36.7% – 9.2% – 12.4% = 15.1% is inertia due to that common effect of the spatial with the other environmental and temporal variables.

This set of inertia components can be depicted in compositional form as shown in Exhibit 19.9. This figure shows how much of the variation is unexplained and how the part that is explained is divided between the spatial and environmental predictors (where "environmental" includes the temporal trend in this case).

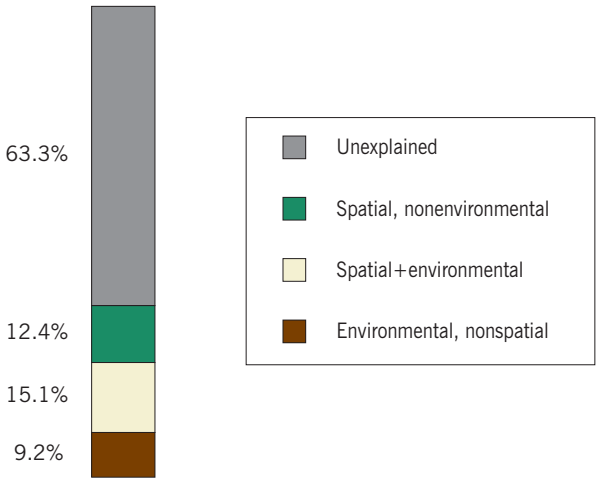


Exhibit 19.9:
Partitioning the total inertia in the abundance data into parts due to the spatial variables and other variables separately, and their part in common

Remember that, of the 36.7% explained variance, the two-dimensional CCA ordination of Exhibits 19.5 and 19.6 only accounts for 58.5%, that is 21.5% of the total variation in the fish abundances. If we take into account the third dimension (see Exhibit 19.8), this brings the explained constrained inertia up to 58.5 + 19.6 = 78.1%, which is 28.7% of the total inertia. Hence, in summary, using the available explanatory variables, depth, temperature, spatial position and year, we can give a statistically justifiable explanation of 28.7% of the variation in the species abundances.

SUMMARY: Case Study 1: Temporal trends and spatial trends across a large ecological data set

1. This case study involved a large data set of fish abundances from trawl samples taken in the Barents Sea, over a six-year period. In addition to the fish data, the environmental variables bottom depth, water temperature and slope of sea-bed were available for each sampling site, as well as latitude and longitude coordinates.

2. In studies such as these that involve sampling across a region over time it can happen that there is unrepresentative sampling in certain areas at different time periods. Conclusions about temporal trends, for example, can become biased due to these sampling imbalances.

3. Samples can be reweighted to be in line with some fixed distribution. In this study we took the distribution over the whole six-year period as the target distribution and reweighted the samples in nine fuzzy regions to be in line with this distribution, thereby eliminating bias in the estimates.

4. Sample weights can be used to reweight the abundance data, after which ordination by CA or CCA, for example, continues as before. In computing average temperatures or diversity measures across the whole region, weighted averages are used.

5. Permutation testing is useful for verifying that the relationship between the fish abundances, regarded as responses, have a statistically significant relationship with the environmental variables and to confirm temporal trends. Similarly, we can test how many dimensions in the solution are nonrandom.

6. The overall variation in the abundance data can be partitioned into a part explained by the environmental and spatial variables. The environmental and spatial predictors are confounded, however, but we can quantify the parts of variation that are purely environmental, purely spatial and a confounding of environmental and spatial.

List of Exhibits

Exhibit 19.1: (a) Actual number of sample taken in three regions over a three-year period, with overall proportions of samples in each region over the whole period. (b) Expected number of samples if in each year sampling had taken place in accordance with the overall proportions. (c) The weights computed by dividing the values in table (b) by those in table (a) 250

Exhibit 19.2: Sums of fuzzy-coded regional categories for each year and for all years. Columns are the eight compass points and a central region (C) 252

Exhibit 19.3: Weights for data according to year and fuzzy region 252

Exhibit 19.4: Correspondence analysis contribution biplot of the "Barents fish trend" data set. The upper plot shows the active data, the samples and species (high-contributing species are shown with bigger labels).The lower plot shows the centroids of all the categories, linking together categories of ordinal variables. 32.6% of the total inertia of 4.017 is explained by these two first dimensions 254

Exhibit 19.5: Canonical correspondence analysis of the "Barents fish trends" data. The format is the same as Exhibit 19.4, with the samples and species plotted in the upper biplot and an enlarged version of the category centroids in the lower plot. 58.5% of the restricted inertia is explained by these two dimensions 256

Exhibit 19.6: Temporal trajectories in regional categories north, east, south and west. Time and regional centroids are at (weighted) averages of the corresponding category points: for example, *S* is at the average of the six points making up the trajectory for south, while 2004 is at the average of all the 2004 points (for all nine regions, only four shown here) 257

Exhibit 19.7: In descending order, the proportion of inertia explained, R^2, and adjusted R^2, of the five categorical environmental variables; k is the number of categories 257

Exhibit 19.8: Scree plot of the inertias of successive dimensions in the constrained space of the CCA of the "Barents fish trends" data. The first three dimensions clearly stand out from the rest 258

Exhibit 19.9: Partitioning the total inertia in the abundance data into parts due to the spatial variables and other variables separately, and their part in common 259