28 Multivariate analyses of South-American zoobenthic species — spoilt for choice

Ieno E.N., Zuur A.F., Bastida R., Martin, J.P., Trassens M. and Smith G.M.

28.1 Introduction and the underlying questions

Defining spatial and temporal distribution patterns of a soft-bottom benthos community and its relationship with environmental factors has been a common task of many coastal marine ecologists. However, the choice of the most appropriate statistical tools for benthic data has been subject to considerable debate among researchers.

Several research programmes aimed at studying the dynamics of benthic species and their environment, have been carried out at South American estuarine and coastal areas during the last few decades (Bemvenuti et al. 1978; Ieno and Bastida 1998; Lana et al.1989; Giménez et al. 2005) focusing not only on the importance of commercial benthic species but also on the rapid habitat fragmentation and deterioration (Elías 1992b; Elías and Bremec 1994) that have resulted from different levels of human impact and that have reduced the available feeding areas for birds.

Samborombón Bay (Buenos Aires, province, Argentina) is an area of major importance in the life cycle of a large number of organisms that play key roles in the food web of the ecosystem (Ieno and Bastida 1998; Martin 2002). Bivalves, crustaceans and especially polychaetes that inhabit the inter-tidal and tidal flats represent an important link in the food chain from primary producers to predators such as resident and migratory birds and fishes. Samborombón Bay is used by migratory nearctic and austral shorebirds from September to April; the main species preying on macrozoobenthos during the annual stop over migrations are the Red Knot (*Calidris canutus rufa*), the White-rumped Sandpiper (*Calidris fuscicollis*), the Hudsonian Godwit (*Limosa haemastica*), the American Golden Plover (*Pluvialis dominica*) and the Two-banded Plover (*Charadrius falklandicus*) (Myers and Myers 1979; Blanco 1998; Ieno et al. 2004). Direct observation and fecal and gizzard analysis have shown that polychaetes along with decapod crabs are the most important items in the diet of these shorebird species during their stay at Samborombón Bay.

The data analysed here, which have been introduced in Chapter 4, come from a benthic-monitoring programme in the autumn-spring period in 1997 at 30 stations from the inter-tidal mudflats of San Clemente Channel in the south of Samborom-

bón Bay. The area is characterised by a benthos displaying high species densities and low species diversity (Ieno and Bastida 1998). The monitoring plots (transects) on San Clemente Channel were selected to represent the major macrobenthic habitats due to the overwhelming abundance of short-lived and fast-growing polychaete species (Ieno and Bastida 1998; Martin 2002). In the original study, the main goal was to determine the relationship between waders and their inter-tidal food supply. The sampling scheme was determined by the topography of San Clemente Creek as well as the feeding behaviour of the secondary consumers.

The underlying question we aim to answer with this particular data set is whether the environmental variables (sediment composition) had any effect on the macrobenthic species data. We also want to determine whether there are differences between transects and seasons and, in particular, whether the two transects close to the eastern part of the study area are different. The main advantage of these data is that only a few species (infaunal data) were monitored at a very low spatial scale. This makes the following statistical explanation and interpretation easier for the reader to understand.

The aim of this chapter

Because we have multiple species, multiple sites and multiple explanatory variables, we are in the world of multivariate analysis. This means that we have to choose from methods such as principal component analysis, redundancy analysis, correspondence analysis, canonical correspondence analysis, non-metric multidimensional scaling (NMDS), the Mantel test, discriminant analysis, etc. This choice is primarily determined by the underlying questions and the type of data. We require a method that can deal with both species and environmental data; hence RDA, CCA, or the Mantel tests are the most obvious candidates. It is then a matter of carrying out a thorough data exploration to identify the appropriate technique. One not only has to choose a particular technique, but also certain settings have to be selected within that technique.

To illustrate the thinking and decision-making process, we present two different analyses. In the first analysis, which we call the 'careless approach', we highlight the pitfalls of not thinking seriously about the underlying questions and demonstrate some of the common mistakes made by inexperienced users of statistical software packages. We also present the way (we think) the data should be analysed. Hence, the aim of this chapter is to show some of the difficulties and dangers of being spoilt for choice.

28.2 Study site and sample collection

Fieldwork was carried out at the extreme southeastern section of the Samborombón Bay, at the narrow navigation San Clemente Channel (Bértola and Ferrante 1996) (Figure 28.1). The area can be described as a typical temperate South American saltmash, characterised by the conspicuous epifaunal burrowing mud

crab, *Chasmagnathus granulata*, together with a dominant vegetation composed by *Spartina densiflora*, *S. alterniflora*, *Salicornia ambigua*, *S. virginica* and *Sirpus maritimus*. Infaunal polychaetes worms mostly dominated the inter-tidal area (Ieno and Bastida 1998; Martin 2002). Sediment texture was dominated by fine and very fine sand with a very soft and flocculent mud fraction of 15–30% (< 63 μm). A full description of the area is given in Ieno and Bastida (1998) and Martin (2002). A photograph of the sampling area is given in Figure 28.2.

Two benthic sampling programmes were carried out on the inter-tidal sediments of the San Clemente channel, one in early May and one in mid-December 1997. Three transects were selected, two in the eastern margin and one in the western margin of the channel, and 60 sediment samples were taken in total to determine infauna abundance and composition. Samples were sieved through a 0.5 mm mesh, and infauna was stored in 70% ethanol (Holme and McIntyre 1984).

The use of a small dredge limited adequate sampling and estimation of the big epibenthic grapsid *Chasmagnathus granulata*; therefore, this species was not included in the data analysis. A list of all available explanatory variables is given in Table 28.1.

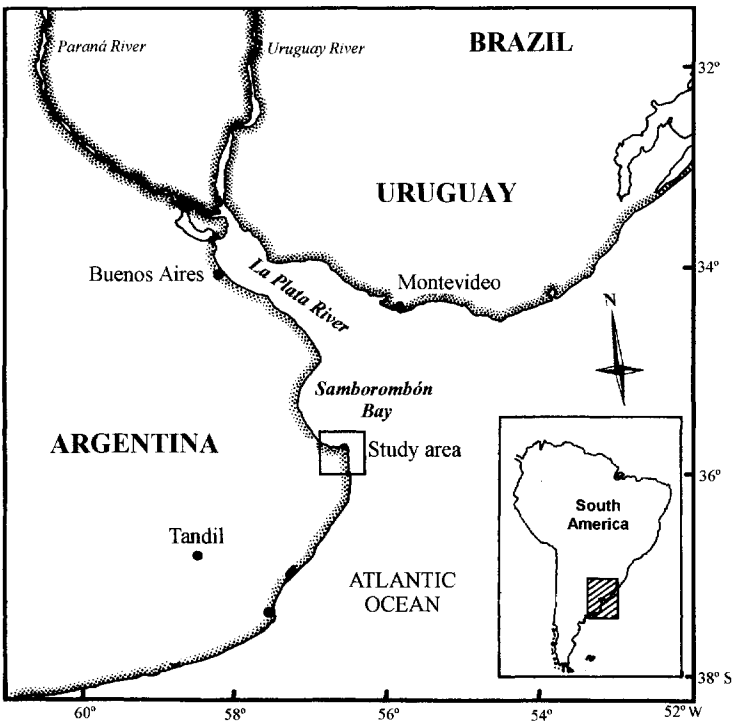

Figure 28.1. Map of the study area.

Figure 28.2. Inter-tidal flats in the study area with the so-called 'Cangrejales' (crab zone) and *Spartina* sea grass bed.

Table 28.1. List of available explanatory variables.

Explanatory Variable	Remarks
MedSand	Continuous variable measured in percentage (%). Medium sand (250–500 μm) (Wentworth 1922)
FineSand	Continuous variable measured in percentage (%). Fine and very finesand (63–250 μm) (Wentworth 1922)
Mud	Continuous variable measured in percentage (%). Particles passing the 64 μm sieve (silt-clay) fraction. (Wentworth 1922).
OrganMat	Continuous variable measured in percentage (%). Organic Matter determined by oxidation method (Walkley and Black 1965)
Transect	Nominal variable: 3 monitored transects (A, B and C), with values 1, 2, and 3.
Season	Nominal variable with values 0 (Southern hemisphere autumn, May) and 1 (Southern hemisphere spring, December) identifying the time of the year that sampling took place.
Channel	Nominal variable with values 0–1 to identify location of transects at both margins of San Clemente channel. (Transect A = Eastern sector) (B and C = Western sector).

28.3 Data exploration

The species data were already used in Chapter 4 to illustrate some of the data exploration techniques. To avoid repeating the same graphs, we will just summarise what we found there. The species data did not contain any large outliers, but two species (*U. uruguayensis* and *N. succinea*) had many observations with zero

abundance. The general impression was that a square root transformation on the species would be beneficial as it brings the species within the same range, but this also depends on which statistical method will be applied in the next step.

Except for transect and seasonal effects, we did not look at the explanatory variables in Chapter 4. A pairplot for the continuous explanatory variables shows that mud and fine sand are highly correlated (Figure 28.3). The scatter of points for mud and organic material also indicate a strong linear relationship, although this is not supported by the correlation coefficient, which has a value of only 0.55. This is probably due to one observation, which has high organic material but low mud. Anyway, the scatter of points clearly indicates that mud is collinear with fine sand and with organic material, so we decided to omit mud from any further analyses. One could even argue that fine sand and organic material are negatively related, but it is less obvious as the patterns with mud.

The last question we address in the data exploration is whether the environmental conditions differ by transect. Cleveland dotplots (Figure 28.4) show that transect 2 has considerable higher median sand values and transect 1 has a higher mean organic material. This indicates that the environmental conditions differ considerably per transect, and the implication of this will be discussed in the next section.

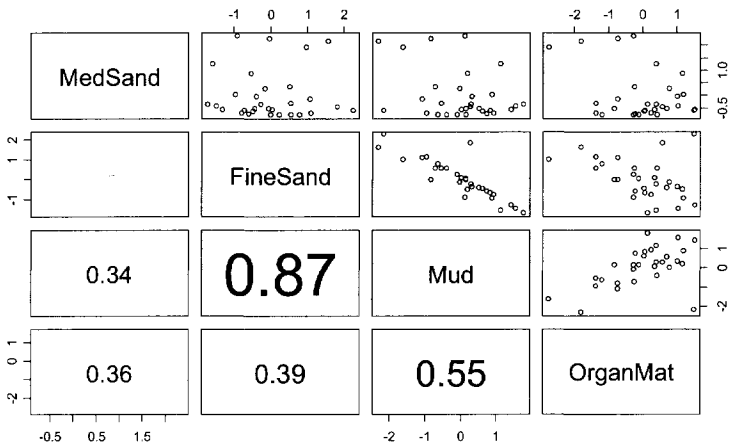

Figure 28.3. Pairplot of all continuous explanatory variables. The lower diagonal panels contain the (absolute) correlation coefficients, and their font size is proportional to the value.

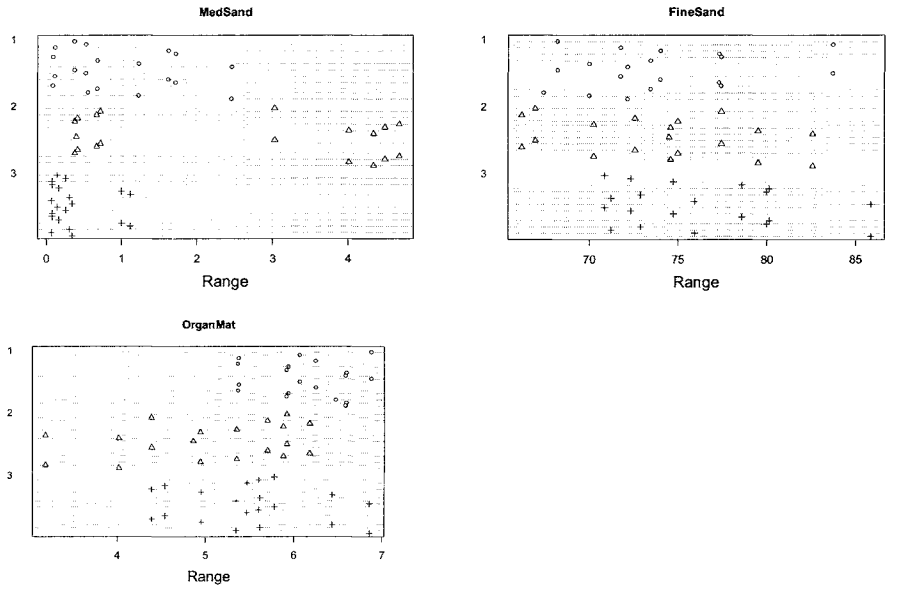

Figure 28.4. Cleveland dotplots for the three continuous explanatory variables medium sand, fine and organic material conditional on transect. The *y*-axis corresponds to an observation (grouped by transect), and the horizontal axis shows the value for each observation.

The careless approach

The data exploration indicates that two species have lots of observations with zero abundance. To avoid problems with collinearity, mud was removed from the analysis. There is also evidence that the environmental conditions differ per transect. In the careless analysis it is now time to start clicking in a software package. Due to the large number of zeros, PCA, RDA, CA and CCA are less suitable, and probably NMDS combined with the Mantel test would seem to be a good approach.

How-it-should-be-done

Thinking a bit more deeply, we can argue that based on the data exploration, only the Jaccard index or Sørensen index should be used. Owing to the environmental differences among the three transects, the permutation method in the Mantel test needs to be done conditional on transect (i.e., permutations should only be made within transects and not between transects). We also think that a special data transformation combined with an RDA (Chapter 12) might be useful as it visualises Chord distances, a measure of association that is suitable for this type of data.

28.4 The Mantel test approach

Recall from Chapters 10 and 26 that the Mantel test starts with two matrices; one for the species (**Y**) and one for the explanatory variables (**X**). It then calculates a distance matrix for the species data ($\mathbf{D_Y}$) and for the explanatory variables ($\mathbf{D_X}$). These distance matrices are of dimension 58-by-58 as there are 58 observations (two observations were omitted due to absence of all species).

The first problem is to choose a measure of association for the species data and for the explanatory variables. Let us discuss the explanatory variables first. The data matrix **X** is of the form:

	S_1	S_2	S_3		S_{58}
MedSand	2.46	1.23	0.56	.....	0.15
FineSand	72.17	70.60	67.22	.....	70.87
Orgmat	6.59	6.60	6.48	.....	5.78
Season	0	0	0	.....	1
Channel	0	0	0	.....	1

To calculate $\mathbf{D_X}$ we need to define the association between S_i and S_j for every i and j combination between 1 and 58. An obvious choice is to use Euclidean distances, but there is one problem with this; it will be dominated by fine sand as it has the largest variance. Hence, such a distance matrix $\mathbf{D_X}$ would mainly represent differences in sites due to fine sand differences, whereas we want to have a distance matrix that represents all environmental variables. To do this, we have to make sure that each explanatory variable is within the same range. The two easiest ways are either to normalise each explanatory variable (Chapter 4) or apply ranging (divide all the observations of a particular explanatory variable by its maximum observed value). Because the data also contain two nominal variables with values 0 and 1, we decided to apply ranging. As a result all explanatory variables are rescaled between zero and one and the Euclidean distance function now weights each variable in the same way.

The careless approach

In the careless approach we assume that we have access to a software package that allows us to choose from 40 different measures of association.

As to the species data, we picked 10 different measures of association, and each time, we applied the Mantel test. The choices were the Jaccard (S) and Sørensen (S) indices, a variation on these two that give triple weight to joint presence (S), the Ochiai index (S), Euclidean distance (D), Chord distance (D), and Whittaker's index of association (D), the Bray–Curtis index (D), the Chi-square distance (D) and the simple matching coefficient (S). An 'S' indicates that the measure of association is a similarity measure, and a 'D' stands for distance measure. The Mantel test works with distance matrices and the default conversion is distance = 1 − similarity. The results are given in the first two columns of Table 28.2. The R_M statistic is the Pearson correlation between elements of $\mathbf{D_Y}$ and $\mathbf{D_X}$. The permuta-

tion tests did not take into that the environmental conditions differed per transect. Imagine the following situation:

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	..	S_{58}
X_1:	1	2	3	96	97	98	1	..	2
X_2:	3	2	1	97	97	97	2	..	3
X_3:	2	2	2	92	93	94	3	..	4
T	1	1	1	2	2	2	3	..	3

Suppose that the *X*s are the environmental variables, S_1 to S_{58} are the observations and *T* identifies the three transects. If we were to permute the observations arbitrarily, the high values of transect 2 would end up in other transects. But the underlying principle of permutation methods and bootstrapping is that we generate data, which are only slightly different compared with the original data. Hence, it is better to permute the data only within transects. The same holds if we work on distance matrices and apply permutation methods to obtain significance values. If we ignore the transect effect, and if there are large differences between transects in terms of environmental conditions, the *p*-values will be too small. There may also be other dangers for the inexperienced user. For example, if the 'similarity to distance conversion method' is changed to

$$d_{ij} = \sqrt{1 - s_{ij}^2}$$

then the R_M statistic involving a similarity coefficient (Jaccard, Sorenson, etc.) becomes considerably larger (Table 28.3). This means stronger relationships! Similarly, care must be taken in quantifying the association between $\mathbf{D_Y}$ and $\mathbf{D_X}$. If instead of using the Pearson correlation coefficient, we had chosen the Spearman correlation coefficient we would have obtained slightly smaller values for R_M. We must also be aware of the importance of carrying out sufficient numbers of permutations. Most software packages permit the user to set the number of permutations. We have used 9999 permutations in Table 28.2 and Table 28.3, but if we had set it to 999, then each time we would have run it, the *p*-values would have been slightly different. If we were unscrupulous, we could have repeated the analysis 5 or 10 times and selected the smallest *p*-value.

We decided to conclude that whatever measure of association we choose, there is a significant relationship between species dissimilarities and environmental differences at the 58 sites. The only exception is the Chi-square distance matrix, but this is probably due to rare species.

Table 28.2. Results of the Mantel test using various different measures of association for the species data. The number of permutations was 9999. The similarity to distance conversion was: $d = 1 - s$, where s is the similarity and d the distance. The p-values marked with * were obtained by permuting the observations only within a transect. The test statistic is the Pearson correlation coefficient.

Index	R_M-statistic	p-value	p-value*
Jaccard	0.242	<0.001	<0.001
Sørensen	0.197	<0.001	0.001
Triple weight to joint presence	0.165	<0.001	0.003
Ochiai	0.179	<0.001	<0.001
Euclidean	0.071	0.040	0.263
Chord	0.078	0.020	0.116
Whittaker's index of association.	0.091	0.007	0.064
Bray–Curtis	0.106	0.004	0.023
Chi-square distance	0.035	0.218	0.018
The simple matching coefficient	0.177	<0.001	0.002

Table 28.3. Results of the Mantel test using various different measures of association for the species data. The number of permutations was 9999. The similarity to distance conversion was: $d = \text{sqrt}(1 - s^2)$, where s is the similarity and d the distance. The p-values marked with * were obtained by permuting the observations only within a transect. The test statistic is the Pearson correlation coefficient.

Index	R_M-statistic	p-value	p-value*
Jaccard	0.287	<0.001	<0.001
Sørensen's	0.281	<0.001	0.001
Triple weight to joint presence	0.272	<0.001	0.003
Ochiai	0.274	<0.001	<0.001
The simple matching coefficient	0.260	<0.001	0.002

How-it-should-be-done

Following the appropriate statistical analysis strategy, we decided to use only the Jaccard index. Permutations conditional on transect were applied, and we used the default conversion of similarity to distance (distance = 1 − similarity). We found $R_M = 0.242$ ($p < 0.001$) using the Pearson correlation coefficient. Table 28.2 also contains p-values obtained by permuting conditional on transect, and these are given in the column labelled p-values*. We also carried out the BVSTEP procedure. Recall from Chapter 26 that this method carries out a forward selection on the explanatory variables and it tries to find the optimal set of explanatory variables so that R_M is maximal. It gave $R_M = 0.265$ using only medium sand, channel and organic material.

As to using the BVSTEP in the careless approach, it would have been very easy to apply the procedure to all 10 measures of association and pick the one we liked best.

28.5 The transformation plus RDA approach

In the 'how-should-it-be-done' approach, we also carried out a special data transformation followed by an RDA. Recall that this approach allows one to visualise Chord distances. The triplot in Figure 28.5 shows a clear zonation of the sites by transect. *L. acuta* was abundant in transect A, *U. uruguayensis* in transect B, which has high medium sand values and the other two species are abundant in transect C. The explained variation by all explanatory variables is 30%, and the first two eigenvalues are 0.21 and 0.7. A forward selection indicated that only medium sand and Channel are important.

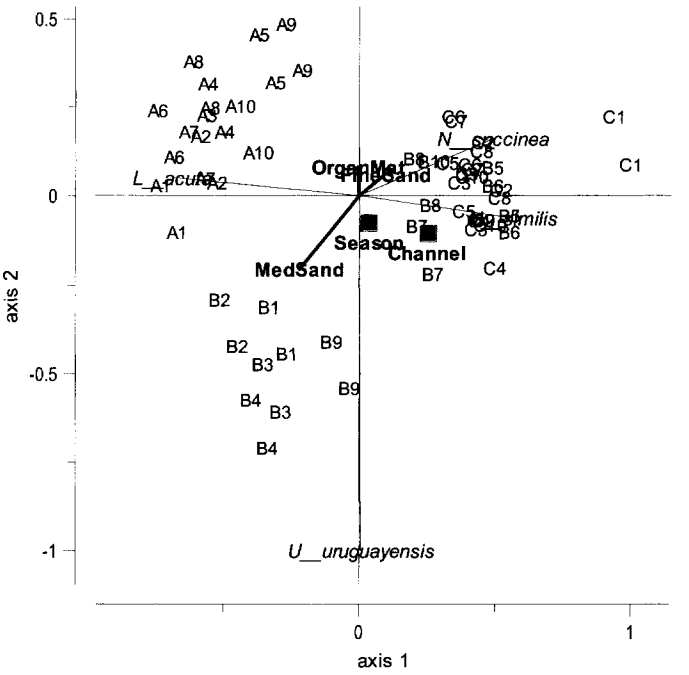

Figure 28.5. RDA triplot obtained using a special data transformation. Distances between observations are two-dimensional approximations of Chord distances.

28.6 Discussion and conclusions

In this chapter we presented two different analysis strategies. In the first we were careless and we showed how we could obtain 'nice' but not necessarily correct results. We found that there is a strong relationship between sites in terms of

species composition and environmental conditions, whatever measure of association is used. In the second ('how-it-should-be-done') analysis we carefully considered what to do. We thought carefully about the questions and decided *a priori* (or better: during the data exploration) which methods to use. We found that there was a clear difference among the three transects, and that the most important explanatory variables were medium sand and channel.

As has been highlighted, our research has focused in detail on the relationship between benthic fauna and abiotic factors. For the same transects we determined the average density in which the different birds species foraged and the density in which the different prey species occurred in the soft-bottom sediment layers. However, this complementary information has been left aside as it is not the primary goal of this case study chapter.

Nevertheless, some general ecological implications were raised from benthic food stocks and feeding opportunities. The fact that we found a clear transect effect in the species-environmental relationship may explain why most Golden Plovers feed on the medium sand patches of transect B that are characterized by the presence of the fiddler crab, *Uca uruguayensis*. Plovers defend feeding territories and rely on eyesight for prey detection. Fiddler crabs are particularly active on the surface next to their burrows during low tide and at the same time are vulnerable to predation.

We also detected a channel effect that clearly denoted a higher abundance of the ragworm, *Laeonereis acuta* in transect A, which is likely to be the major prey of the White-rumped sandpiper in the study area. Sandpipers are the most important foragers and have been shown to feed in large social flocks on medium-size *L. acuta* mainly recorded on the eastern sector of San Clemente Channel. Although the explanatory variables in the data seemed to explain part of the variation in the species composition, there are some intriguing conditions that made *H. similis* more abundant in transect C than the others. We suspect that both high turbidity due to the frequent resuspension of softer substratum and sedimentation processes underlie the environmental heterogeneity of the western sector of San Clemente Channel.

The mistakes made in the careless approach may sound silly, but the authors of this book have reviewed various submitted manuscripts in which the number of permutations was only 199 because it was the default number in the software package that was used.

Ultimately, the best way you can make sure that you have applied the most appropriated statistical tool is to think carefully about the underlying questions before applying complicated methods.

Acknowledgements

We would like to thank Fundacion Mundo Marino who kindly allowed us to collect the data and use their experimental facilities. Sergio Moron and Jorge Rebollo provided tremendous assistance in the field. This study was partially supported by AGENCIA and CONICET (Argentina). Special thanks to Barry O'Neill for commenting on an earlier draft.