27 Univariate and multivariate analysis applied on a Dutch sandy beach community

Janssen, G.M., Mulder, S., Zuur, A.F., Ieno, E.N. and Smith, G.M.

27.1 Introduction

Climate change is, beyond doubt, the most important threat facing the world's coastline and has been accompanied by intensive debate. Marine coastal ecosystems are extremely vulnerable, as they constitute the most productive and diverse communities on Earth. Coastal areas are, however, not only subject to climate change but also to many other forms of human activities. Among them, land-claim, pollution, recreation purposes and dredging activities have been threatening most of the European coasts resulting in many cases in inter-tidal habitat fragmentation and/or degradation (Raffaelli and Hawkins 1996). The consequences of these changes have been well documented in a considerable number of studies that addressed the impact and reported decreased ecosystem performance.

On a more local scale, the Dutch have fought great battles with the North Sea in order to extend their landmass as can be witnessed by the presence of dykes and sophisticated coastal defence systems. The effect of sea level rise on the ecology of the Dutch coastal system constitutes a serious issue that should not be ignored in the short term.

The Dutch governmental institute RIKZ therefore started a research project on the relationship between some abiotic aspects (e.g., sediment composition, slope of the beach) as these might affect benthic fauna. Mulder (2000) described the results of a pilot study that looked at the effects of differences in slope and grain size on fauna in the coastal zone. Using the data from this pilot study and statistical experimental design techniques, a sampling design was developed in which nine beaches were chosen by stratifying levels of exposure: three beaches with high exposure, three beaches with medium exposure and three beaches with low exposure. Sampling was carried out in June 2002, and at each beach, five stations were selected. Effort at each station was low (Van der Meer 1997).

The aim of the project was to find relationships between macrofauna of the intertidal area and abiotic variables. In this case study chapter, univariate and multivariate tools are applied in order to obtain as much information as possible. The results of the combined analyses are then used to answer the underlying question. Instead of presenting the results of the final models, we show how we got to them,

and which steps we applied (especially for the multivariate analysis). The outcome of this research will have immediate relevance for assessing and managing disturbance in marine benthic systems, with respect to degradation and biodiversity lost.

27.2 The variables

Table 27.1 gives a list of the available explanatory variables. NAP is the height of the sampling station relative to the mean tidal level. Exposure is an index that is composed of the following elements: wave action, length of the surf zone, slope, grain size and the depth of the anaerobic layer. Humus constitutes the amount of organic material. Sampling took place in June 2002. A nominal variable 'week' was introduced for each sample, which has the values 1, 2, 3 and 4, indicating in which week of June a beach was monitored. The following rules were used. Sampling between 1 and 7 June: $Week_i = 1$. Sampling between 8 and 14 June: $Week_i = 2$. Sampling between 15 and 22 June: $Week_i = 3$ and sampling between 23 and 29 June: $Week_i = 4$. The index i is the station index and runs from 1 to 45. There were nine beaches, and on each beach five stations were sampled (hence, 45 observations). Ten sub-samples were taken per station, but in this chapter we will use totals per station. $Angle_1$ represents the angle of each station, whereas $angle_2$ is the angle of the entire sampling area on the beach. Both variables were used. The variables $angle_2$, exposure, salinity and temperature were available at beach level. This means that it is assumed that each station on a beach has the same value. For $angle_2$ this assumption does not hold, and its inclusion in some of the statistical models should be done with care. A few explanatory variables contained four missing values. Most of the statistical techniques used in the analysis cannot cope with missing values, and therefore, missing values were replaced by averages.

Table 27.2 shows the Pearson correlation among the 12 explanatory variables. Only correlations significant at the 5% level are given. Except for a few variables (chalk and sorting and exposure and temperature), the correlations among the explanatory variables are relatively low, indicating that there is no serious collinearity.

As to the species, in total 75 species were measured. To simplify interpretation of graphical plots, species were grouped in the following five taxa: Chaetognatha, Polychaeta, Crustacea, Mollusca, and Insecta. Within each taxa, between 1 and 28 species were available. Species names were replaced by names taking the following form: P_1, P_2, P_3, etc. for Polychaeta species, CR_1, CR_2, CR_3, etc. for Crustacea species, M_1, M_2, M_3, etc. for Mollusca species, I_1, I_2, I_3 for Insecta, and C_1 for the Chaetognatha species *Sagitta* spec. (only one Chaetognatha was measured). A list of species and notation used in this chapter are available as an online supplement to this book.

Table 27.1. List of available explanatory variables. The columns labelled "Level" indicate whether different values are available for each station on a beach (Beach) or one value for all five stations on a beach (Station).

Number	Variable	Level	Units
1	Week	Beach	-
2	$Angle_1$	Station	-
3	$Angle_2$	Beach	-
4	Exposure	Beach	-
5	Salinity	Beach	‰
6	Temperature	Beach	°C
7	NAP	Station	m
8	Penetrability	Station	N/cm^2
9	Grain size	Station	mm
10	Humus	Station	%
11	Chalk	Station	%
12	Sorting	Station	Mm

Table 27.2. Significant correlations ($\alpha = 0.05$) among explanatory variables.

	Week	$Angle_1$	$Angle_2$	Exp.	Sal.	Temp.	NAP	Pen.	Grains	Hum	Chalk	Sort.
Week	1		−0.61		−0.31				0.41	−0.36		0.31
$Angle_1$		1	0.39						−0.32			
$Angle_2$			1	−0.45					−0.77	0.39		−0.48
Exp.				1		−0.53	−0.72		0.60	−0.32	0.40	0.56
Sal.					1			−0.40	−0.58	0.38	−0.33	−0.49
Temp.						1			−0.40			
NAP							1					
Pen.								1				
Grains									1	−0.35	0.45	0.75
Hum										1		−0.37
Chalk											1	0.79

27.3 Analysing the data using univariate methods

In this section, the species data are analysed by converting them into a diversity index (Magurran 2004). We used the Shannon–Weaver index (with base $\log_{10}$). The shape of the Shannon–Weaver index was similar to that of the species richness Because of this similarity the Shannon–Weaver index can also be seen as an indicator for the number of different species (at least for these data). Figure 27.1 shows a dotplot for the Shannon–Weaver index. Stations are grouped by beach. There are no stations with considerably larger or smaller values.

We now compare the Shannon–Weaver index with the explanatory variables. Quinn and Keough (2002) followed a similar approach and applied regression techniques on the diversity index. Table 27.3 shows the correlation between the Shannon–Weaver index and each of the 12 explanatory variables.

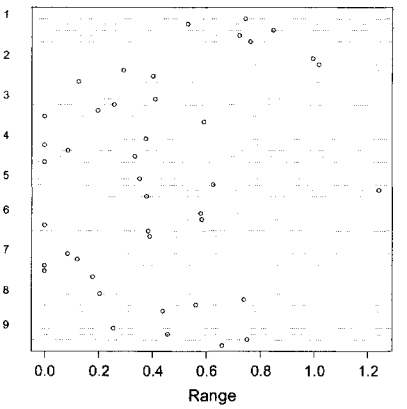

Figure 27.1. Cleveland dotplot of the Shannon–Weaver index. The horizontal axis shows the value at each site, and the vertical axis shows the 45 stations, grouped by beach.

The Pearson correlation coefficients indicate that exposure, salinity and NAP were significantly related to the Shannon–Weaver index. This correlation coefficient does not take into account that the 45 samples come from nine beaches (each beach had five stations). To take into account this two-way data pattern, between-beach and within-beach correlations can be calculated (Snijders and Bosker 1999). Results are presented in the third and fourth column of Table 27.3. The between-beach correlation is the correlation between the beach averages. Humus, chalk and sorting were significantly correlated to the diversity index, and exposure and NAP have a correlation just below the significance level. The within-beach correlation is a correlation coefficient corrected for the beach differences. NAP had by far the largest correlation, followed by grain size, chalk and sorting. For some of the variables, within-beach correlations could not be calculated because these explanatory variables had the same value at all five stations on a beach.

Figure 27.2 shows a pairplot of the explanatory variables. The variables $angle_1$, humus, chalk and sorting all have an observation with a large value. Because this might cause problems in some of the statistical analyses, a square root transformation was applied on these variables. In fact, the correlations in Table 27.3 were obtained after transforming these four variables. Using different transformations for explanatory variables is generally not recommended as it complicates the interpretation. However, in this case it seems justified as the explanatory variables represent different things, also measured in different units.

Pearson correlations between all explanatory variables were calculated and all were smaller than 0.8 in the absolute sense. However, the correlations between (i) $angle_2$ and grain size, and (ii) sorting and chalk were between 0.75 and 0.8, indicating a certain degree of collineairty.

Table 27.3. Pearson correlation, between-beach correlation and within-beach correlation coefficients between Shannon–Weaver index and each of the explanatory variables. Variables in bold typeface are significant at the 5% level. $Angle_1$, humus, chalk and sorting were square root transformed; see the text.

Variable	Cross-Correlation	Between-Beach	Within-Beach
Week	−0.08	−0.14	
$Angle_1$	−0.03	0.03	0.05
$Angle_2$	0.20	0.33	
Exposure	**−0.39**	−0.64	
Salinity	**0.34**	0.57	
Temperature	0.12	0.20	
NAP	**−0.70**	−0.66	**−0.75**
Penetrability	0.00	0.10	**−0.29**
Grain size	−0.18	−0.50	**0.34**
Humus	0.30	**0.74**	0.06
Chalk	−0.23	**−0.75**	**0.31**
Sorting	−0.21	**0.72**	**0.30**

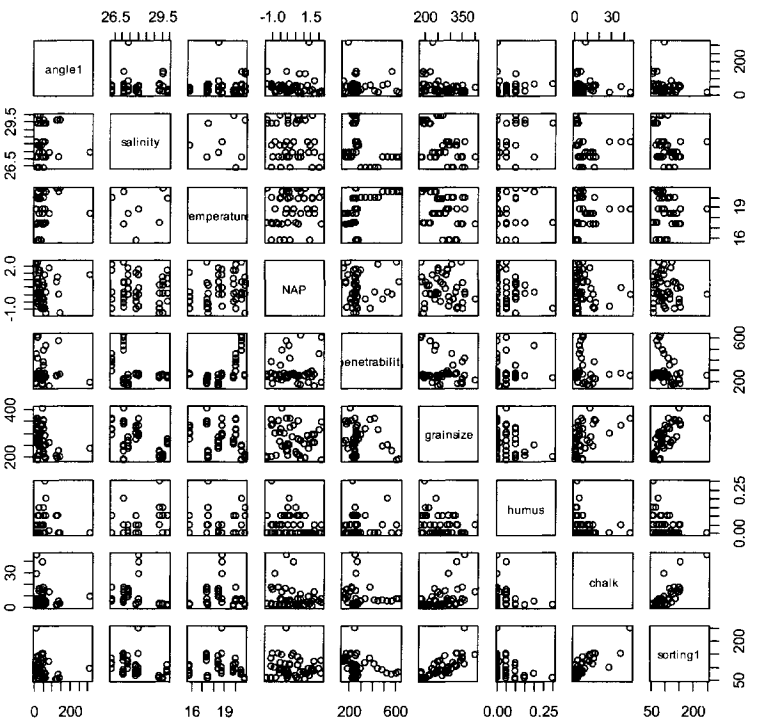

Figure 27.2. Pairplot of some of the explanatory variables.

Additive modelling

We now model the relationship between the Shannon–Weaver index and the 12 explanatory variables. One possible candidate technique is linear regression, but this means that we are imposing linear relationships on the data, whereas in ecology, most relationships are non-linear. We will use a more flexible technique, namely additive modelling. In this technique, the data will tell us the shape of the relationship, and the linear relationship is a special case (Crawley 2002). We will also use regression trees as it is even more flexible in finding relationships.

In a linear regression model, the following mathematical formulation is used to relate the response variable and explanatory variables:

$$H_i = a + b_1 X_{1i} + \ldots + b_p X_{ip} + e_i$$

where b_j are regression coefficients, X_{ij} is the value of the j^{th} explanatory variable ($j = 1,..,12$) at the i^{th} sample ($i = 1,..,45$) and H_i is the Shannon–Weaver at site i. In additive modelling the following mathematical formulation is used:

$$H_i = a + f_1(X_{1i}) + \ldots + f_p(X_{pi}) + e_i \qquad (27.1)$$

where $f_j(X_j)$ is a smoothing function (e.g., a LOESS curve) along the j^{th} explanatory variable. Because the response variable is modelled as a sum of smoothing curves, the model in equation (27.1) is called an additive model. An example of a semi-parametric additive model is:

$$H_i = a + b_1 X_{1i} + f_2(X_{2i}) + \ldots + f_p(X_{pi}) + e_i$$

X_1 is fitted parametrically whereas $f_j(.)$ are smoothing functions. We used the additive modelling approach, to find a relationship between the Shannon–Weaver index and the 12 explanatory variables. Because week and exposure are factors, the following initial model was used:

$$Y_i = a + \text{week}_i + f_2(\text{angle}_i) + \text{exposure}_i + f_4(\text{salinity}_i) + \ldots + f_{12}(\text{sorting}_i) + e_i$$

This model resulted in wide point-wise confidence bands for the smoothers, which is probably due to the collinearity we discussed earlier. Perhaps it would have been better to omit one of the explanatory variables angle and grain size. The same holds for sorting and chalk and week and exposure. Besides, in our experience it is better to do a forward selection instead of a backwards selection in additive modelling, as the technique cannot deal well with collinearity. The results of the forward selection are presented in Table 27.4. The AIC was used as model selection tool; the lower the AIC, the better the model. The second column in Table 27.4 contains the AIC values of the additive models with only one explanatory variable. To simplify the analysis, we only used four degrees of freedom for each smoother. The model with NAP was the best model containing a single explanatory variable. The third column (labelled 'NAP +...') contains the AIC value of a model with NAP and each of the remaining variables. Week and exposure were fitted as nominal variables. The model with NAP and week was the best, although differences in the AIC with the model containing NAP and exposure were small. The last column (labelled 'NAP + week +...') contains the AIC of models with

three explanatory variables; two of them were NAP and Week. No other combination gave a lower AIC than the model with NAP and Week. Hence, the most optimal model, as judged by the AIC, contained NAP and Week. Competing models are (i) NAP and exposure, and (ii) NAP, week and humus. The last model has large confidence bands for the smoother of humus and is therefore not a serious competing model. However, the model with NAP and exposure has an AIC of −11.02 whereas the model with NAP and Week has an AIC of −13.46. This is a small difference. The explained variance of both models is similar as well (71% for NAP + Week and 68% for NAP + Exposure). A model validation indicated that the model NAP + Week contained some evidence of violation of homogeneity, which was not the case for the NAP + Exposure model. The ecological interpretation of the NAP + Exposure model is easier as well. After all, the interpretation of Week is difficult. Furthermore, there is a certain amount of collinearity between them as the lowest exposure values were measured in the first week. These arguments are all in favour of the NAP + Exposure model, and we therefore present the results of this model. The model is of the form:

$$H_i = a + f_2(NAP_i) + Exposure_i + e_i$$

A useful aspect of additive modelling is that the partial fit of smoothers can be visualised (Figure 27.3). NAP showed a general downwards pattern. In fact, a cross-validation to estimate the degrees of freedom of the smoother (Chapter 7) shows that the smoother can be replaced by a parametric term. This also means that the additive model can be replaced by a linear regression. It should be noted that NAP was one of the few explanatory variables having a linear effect! Hence, starting with additive modelling was not a waste of time. Furthermore, one should also keep in mind that we are applying a model with a normal distribution. So, all the potential problems for linear regression discussed in Chapter 5 apply here as well. A model validation indicated normality and homogeneity of residuals, but if the model is used for prediction, it is theoretically possible to obtain negative Shannon–Weaver values. Switching to a Poisson distribution is not an option here as the diversity index takes on small non-integer values. The numerical output of the additive model is as follows:

	Estimate	std. err.	*t* ratio	*p*-value
Intercept	0.55	0.09	5.86	<0.001
factor(exposure)10	−0.05	0.10	−0.44	0.66
factor(exposure)11	−0.31	0.11	−2.87	0.007

Other numerical information is given by

	edf	Chi-sq	*p*-value
s(NAP)	4	53.94	<0.001
R-sq.(adj) = 0.627. Deviance explained = 67.8%.			
Variance=0.038. $n = 45$			

Table 27.4. AIC values obtained by forward selection in the additive model. The second column contains the AIC for a model with one variable. The third column contains the AIC value of a model with NAP and each of the remaining variables. The last column contains the AIC of models with three explanatory variables; two of them were NAP and Week.

Variable	One Variable	NAP +...	NAP + Week +...
Week		**−13.46**	
$Angle_1$	30.89	10.29	−6.38
$Angle_2$	25.00	−2.91	−6.06
Exposure		−11.02	−10.26
Salinity	20.99	−8.68	−5.6
Temperature	23.44	−6.51	−6.15
NAP	**3.70**		
Penetrability	32.58	10.42	5.79
Grain size	30.77	6.11	−9.51
Humus	29.70	4.23	−13.31
Chalk	30.94	−0.30	−1.58
Sorting	31.42	6.07	−9.24

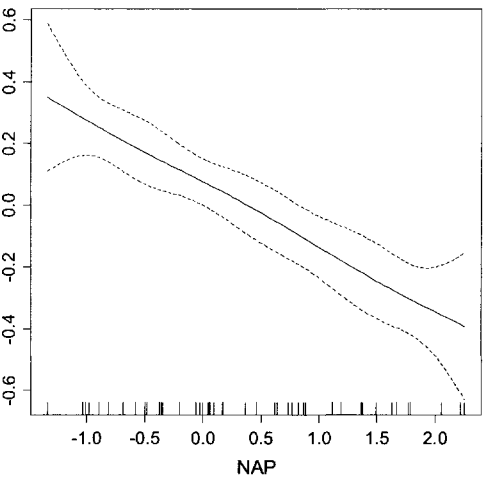

Figure 27.3. Partial fit of NAP in the additive model with exposure. Dotted lines represent 95% confidence bands. The symbols "|" indicate the value of NAP at each station. The smoother has 4 degrees of freedom although further model improvement is possible by decreasing the degrees of freedom. The horizontal axis shows the NAP gradient, and the vertical axis is the contribution of the smoother to the fitted values.

Regression trees

Regression trees (Chapter 9) are useful exploratory tools for uncovering structure in the data. They can be used for screening variables and summarising large

datasets. These models result in classification rules of the form: if $x_1 < 0.5$ and $x_2 > 20.7$ and $x_3 < 0$ then the predicted value of y is 10, where x_1, x_2 and x_3 are explanatory variables and y is the response variable. Regression trees deal better with non-linearity and interaction between explanatory variables than regression and GAM models.

The regression tree was applied on the Shannon–Weaver index values. The optimal tree after pruning (Chapter 9) is presented in Figure 27.4. In order to predict the (Shannon–Weaver) index values, one follows the path from the top (root) to the bottom (leaf). The diversity index is first split according to whether the values of NAP are larger (left branch) or smaller than 0.1685 (right branch). Hence, all stations with NAP larger than 0.1865 are in the left branch (these stations are close to the dunes), and the right branch contains the samples with NAP smaller than 0.1685 (these are close to the water line). Observations in the left branch can again be divided into two groups, namely those with very high NAP values (>1.154) and between 1.154 and 0.1685. Values at the end of the tree contain the average value of a group of observations. Observations with a high NAP value (the leftmost leaf) have an average Shannon–Weaver value of 0.051. Week 2 is important for the observations with intermediate NAP values. Observations in the main right branch have higher Shannon–Weaver values, especially if exposure is not 11 (level 3). These results are in line with those of the additive model.

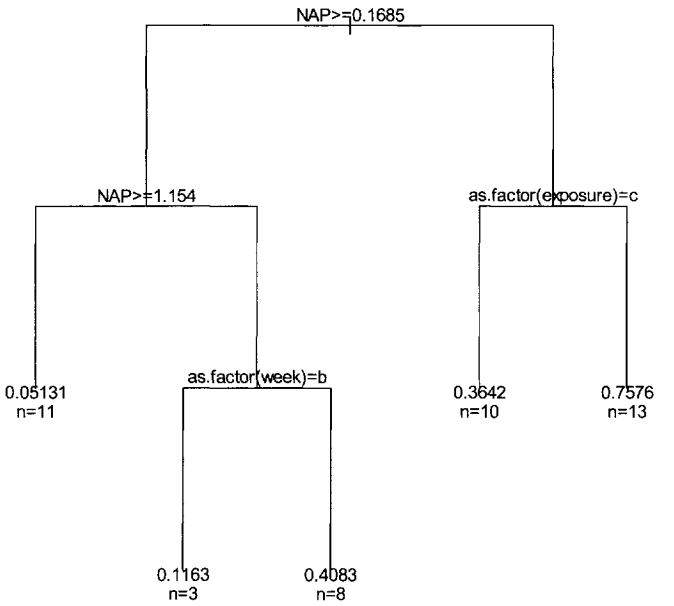

Figure 27.4. Regression tree for the Shannon–Weaver index.

27.4 Analysing the data using multivariate methods

In this section, multivariate analysis is used to analyse whether there are relationships between the species and environmental variables. The species data contain many zeros, and therefore, the application of canonical correspondence analysis and redundancy analysis is inappropriate because of patchy species and double zeros. One option is to take totals per beach and use that in a CCA or RDA. The advantage of this approach is that it reduces the percentage of zeros. But it also reduces the size of the data as there are only nine beaches. An alternative is to apply a special transformation before applying the RDA and to visualise either chord or Hellinger distances (Chapters 12, 28). Legendre and Gallagher (2001) showed that this approach is less sensitive to double zeros and therefore to the arch effect. Here, we present the results of the RDA. Various studies have

shown that the Chord distance performs well for ecological data, and therefore, we use the Chord transformation. This means that the RDA software applies a particular transformation on the species data prior to the actual RDA algorithm. As a result, the distances between observations are two-dimensional approximations of Chord distances (Chapters 10 and 12). The triplot is given in Figure 27.5. We wanted to make all species equally important in the analysis, and therefore, the correlation matrix was used (Chapters 12 and 29). A square root transformation was applied to the species data to down-weight the effect of abundant species. Note that this transformation is not related to the Chord transformation; the square root transformation reduces the influences of large values, and the Chord transformation rescales the data of each station.

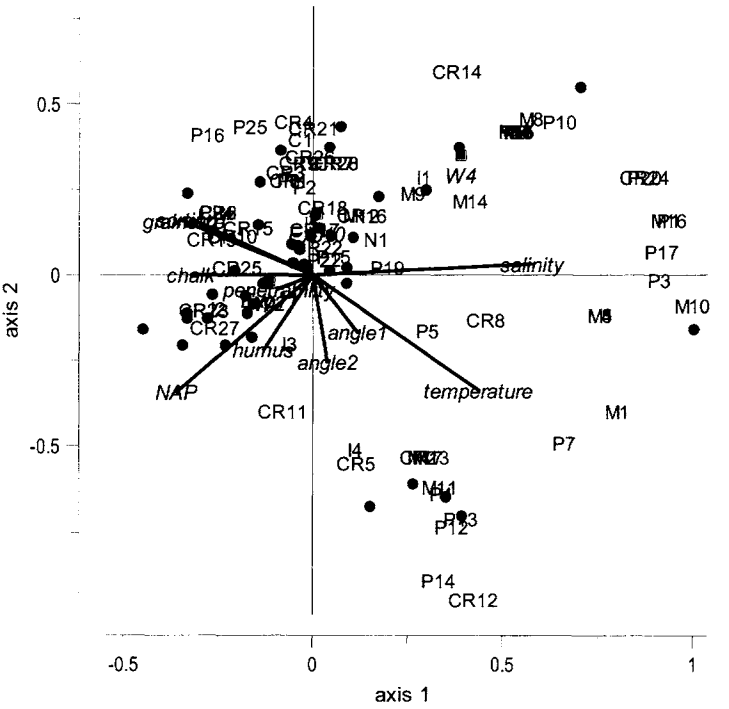

Figure 27.5. RDA triplot for the RIKZ data. The correlation matrix was used. The first two axes explain 45% of the total sum of all canonical eigenvalues (0.40) which corresponds to 18% of the variation in the species data. All explanatory variables were used.

To determine which explanatory variables are important, a forward selection procedure was applied. It uses explained variance as criteria to indicate which explanatory variable should be included/excluded in the model. The forward selection procedure indicated that week, salinity, NAP, humus, temperature and exposure are important (Table 27.5) and that all other variables can be dropped

posure are important (Table 27.5) and that all other variables can be dropped from the model. The RDA was refitted using the same Chord transformation, but now only the significant explanatory variables are used. It was felt that if one of the levels of a nominal variable was significant, all levels should be kept. Therefore, all week and exposure levels were included and not just those that were significant in the forward selection. The resulting triplot is presented in Figure 27.6. The amount of explained variance by the first two axes is similar as in Figure 27.5, namely 17%. Such a low number is common in ecological studies. Full details of the numerical output (using only the significant explanatory variables) are given in Table 27.6 and show that the first two eigenvalues are approximately similar. The first two axes explain 33% of the variation in the species data that can be explained with the six explanatory variables. The first two axes explain 52% of this, which works out as 17% of the species variation. The small difference between this percentage and the 18% using all 12 variables means that no important variables were omitted in the analysis. If the first eigenvalue is considerably larger than the second, then interpretation of the triplot should first be done along the horizontal axis. However, this is not the case here as the second eigenvalue is only marginally smaller than the first.

Table 27.5. Order of importance, F-statistic and p-values for the RDA analysis. W_4, W_2, $Expo_{10}$ and $Expo_{11}$ are nominal variables representing week and exposure levels. We had to drop W_3 as it was collinear with exposure.

Variable	F-statistic	p-value
W_4	**2.648**	**0.021**
Salinity	**1.955**	**0.004**
NAP	**2.303**	**0.007**
Humus	**1.568**	**0.064**
Temperature	**1.689**	**0.035**
$Expo_{10}$	**1.693**	**0.014**
Chalk	0.864	0.572
penetrability	0.364	0.971
Expl1	0.031	1.000
Angle1	0.682	0.713
Sortin	0.108	1.000
Angle2	0.223	1.000
Grain size	0.037	1.000
W2	0.137	1.000

The triplot indicates that NAP and humus are correlated with each other, but negatively with temperature. A group of *Crustacea* species appear at high values of NAP. Most *Mollusca* species are on the right-hand side of the triplot, which corresponds to low NAP values and high salinity and temperature and in week 4. It is possible that week, salinity and temperature have a certain degree of collinearity as temperature and salinity have the same value for all stations on a beach, and sampling on a particular beach is carried out on the same week. This can eas-

ily be verified by replacing the dots by letters A, B, C, D, E, F, G, H and J to visualise which observations are from the same beach (Figure 27.7). We avoided the letter 'I' as it was not clear in the graph. Results indicate that stations of the same beach are close to each other in the triplot. This means that these stations have similar species composition and environmental conditions. In particular beach B seems to have high temperature.

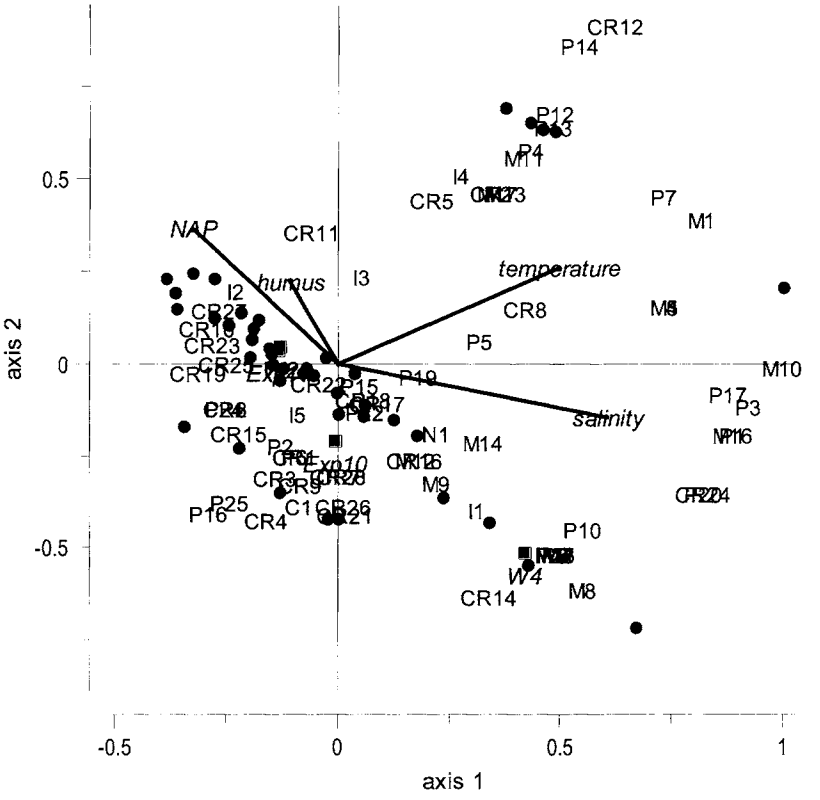

Figure 27.6. RDA triplot for the RIKZ data. The correlation matrix was used. The first two axes explain 45% of the total sum of all canonical eigenvalues (0.40) which corresponds to 18% of the variation in the species data. Only significant explanatory variables were used, namely week, salinity, NAP, humus, temperature and exposure.

Table 27.6. Numerical output for RDA using only the significant explanatory variables. The total variation is 1 and the sum of the canonical eigenvalues is 0.33.

	Axis 1	Axis 2
Eigenvalue	0.10	0.07
Eigenvalue as % of total variation	10%	7%
Eigenvalue as cumulative % of total variation	10%	17%
Eigenvalue as % sum of all canonical eigenvalues	30%	22%
Eigenvalue as cumulative % sum of all canonical eigenvalues	30%	52%

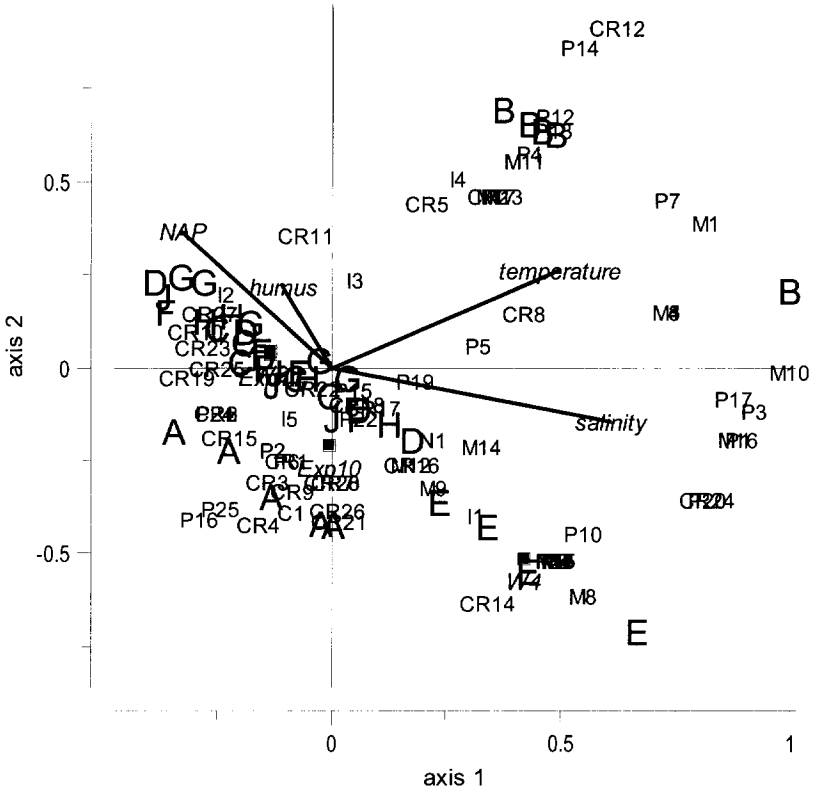

Figure 27.7. Same as Figure 27.6 except that dots are replaced by letters A, B, C, D, E, F, G, H and J for the nine beaches.

Variance partitioning in RDA

We now address the question how much variation in the species data is purely related to exposure. Borcard et al. (1992) showed how variance partitioning is used in CCA and RDA to estimate the contribution of particular groups of ex-

planatory variables. Here, one group is exposure (the two dummy variables) and the other group is defined by all other explanatory variables. Let us call them 'the others'. To estimate the pure exposure effect, the following calculations were carried out (the same transformation as above was used):

1. Apply RDA with exposure as the only explanatory variable.
2. Apply RDA with all other explanatory variables.
3. Apply partial RDA with exposure and 'the others' as the explanatory variables.
4. Apply partial RDA in which exposure is used as explanatory variable and 'the others' as covariables.
5. Apply RDA with all explanatory variables.

Using the eigenvalues of each analysis (Table 27.7), the variance partitioning can be made (Table 27.8). Results indicate that the variation in the species data purely related to exposure is 5%. The other explanatory variables explain 29%. There is also a certain degree of shared information (collinearity) between them: 6%.

Table 27.7. Results of various RDA analysis.

RDA	Explanatory Variables	Sum of All Canonical Eigenvalues
A	Exposure	0.12
B	The others	0.35
C	Exposure with 'the others as covariate	0.05
D	The others with exposure as covariate	0.29
E	All explanatory variables	0.40

Table 27.8. Variance partitioning in RDA. The total variation of the species data in RDA is scaled to 1.

Component	Formula	Variance	%
Pure Exposure	C	0.05	5
Pure 'Others'	D	0.29	29
Shared	A-C or B-D or 1-E-C-D	0.06	6
Residual	1-E	0.60	60
Total		1	100

27.5 Discussion and conclusions

In this chapter, various different statistical techniques were applied to the RIKZ data. The reason for applying a large number of statistical techniques is that each method shows something different. By focusing on only one technique, e.g., RDA, vital information is missed. Besides, one can be more confident about the biological interpretation of the results, if all (or most) methods show something similar.

The statistical methods applied in this chapter can basically be divided into two groups, univariate analysis and multivariate analysis.

Univariate analysis

The Shannon–Weaver index was calculated for all 45 observations. For these data, it can also be seen as a measure of the different number of species per station (as it shows a similar pattern).

According to beach studies, elsewhere in the world (Brown and McLahlan 1990), the most diverse stations can be expected on beaches with fine sand and flat slopes. In this study, the flat beaches are B, G and I, and beaches with fine sand are A, B, G and I. Simple graphs can be used to show that the most diverse stations were on beaches A, B and E. These are the most northern sampled beaches indicating a possible regional effect (Janssen and Mulder 2004, 2005).

Application of all statistical techniques on the diversity indices indicated that there is a significant relationship between the diversity indices and NAP. According to Beukema (2002), the maximum number of macrobenthic species is to be found in the area of the intertidal just between mean tidal level and the low water line. Beukema (2002) studied this phenomenon on the intertidal of a tidal flat area in the Dutch Wadden Sea. Janssen and Mulder (2004, 2005) suggest that this also holds for the inter-tidal of sandy beaches.

An ecological explanation can be found in the combined effects of increasing good food conditions from the high water level down to the low level line (the NAP gradient) and the increasing predation pressure, by shrimp and young fish, up from the low water line to the high water line. Furthermore, there is a negative effect of exposure (the higher the exposure, the lower the diversity) and a week effect. The effect of exposure is in line with literature on this subject (Brown and McLahlan 1990). Exposed beaches in the world show low diversity and abundances compared with sheltered beaches. There is no clear (or strong) relationship between the diversity index and any of the other explanatory variables.

Multivariate analysis

RDA using a special data transformation was applied to the square root transformed abundances of 75 species. The explanatory variables week, salinity, NAP, humus, temperature and exposure were significant at the 5% level. The first gradient represents a NAP versus temperature/salinity gradient. The distribution of the stations in the triplot indicates a major NAP gradient.

Combining the results

The results of all analyses indicated that there is a strong relationship among NAP, exposure and diversity index, and also with the multiple species data. The angle of the beach was not important in any of the analyses. This holds for both $angle_1$ and $angle_2$. Applying a square root transformation on $angle_1$ did not improve the results. Most analyses indicated that there was a strong week effect. Be-

cause sampling took place only in June, it is likely that seasonal patterns will exist if sampling were to take place in other months as well. Using the Shannon–Weaver index, the variable NAP turned out to be the most important explanatory variable. In the multivariate analysis, it was also important.

To further optimise the sampling protocol, a statistical experimental design on the 2002 data should be applied. The most obvious way to improve the design is to select beaches that are stratified on exposure (three beaches with low exposure, three beaches with medium exposure and three beaches with high exposure). However, this might be difficult for beaches along the Dutch coast. Perhaps an alternative definition of exposure could be established.

Possible further analysis

Besides the environmental variables, spatial co-ordinates of the beaches were available. Using variance partitioning, the contribution of spatial variation can be determined (and filtered out). Let us make a distinction between spatial explanatory variables and all other explanatory variables (including exposure), denoting these as 'the others'. To determine the variation in the species data that is purely related to spatial locations of the beaches, the following calculations need to be carried out:

1. Apply RDA with the spatial explanatory variables.
2. Apply RDA with the other explanatory variables.
3. Apply partial RDA with the spatial explanatory variables, and the others as covariables.
4. Apply partial RDA in which the other variables are used as explanatory variables and the spatial variables as covariables.
5. Apply RDA in which all explanatory variables are used.

The spatial variables consisted of x and y co-ordinates. These are denoted by x and y, respectively. Following Borcard et al. (1992), the following derived spatial variables can be used: x, x^2, x^3, y, y^2, y^3, xy, x^2y, and xy^2. A linear combination of these spatial variables models most spatial gradients. For example, the function xy models a diagonal (Northeast to Southwest) gradient. However, for the multivariate RIKZ data, we could not apply this analysis as some of the explanatory variables have the same value for all five stations per beach. This means that there is a strong collinearity between the spatial variables that also have the same value per beach.