20 Univariate methods to analyse abundance of decapod larvae

Pan, M., Gallego, A., Hay, S., Ieno, E.N., Pierce, G.J., Zuur, A.F. and Smith, G.M.

20.1 Introduction

This chapter illustrates how to decide between the application of parametric models (linear regression models) and non-parametric methods (additive models). The techniques applied in this chapter will use as explanatory variables some abiotic (temperature, salinity) and biotic (algal food biomass, as indicated by chlorophyll a) factors that affect the meroplanktonic larvae. We aim to provide preliminary information about some of the pre-settlement processes and the relative influences of environmental factors and variability. Also, the taxonomic identification of some decapod larvae is often difficult, and processing the samples is time consuming. By using information from the samples already analysed and the other available data, such as how many samples per year we could analyse, we may optimise the number of samples examined to achieve the best outcomes and interpretations. In this chapter we therefore also discuss how some models can be used to optimise the number of samples for further sample analysis in other years.

Decapod crustaceans are an important fishery resource in many countries around the world. In Europe alone there are approximately 22 species that are actively exploited. The fisheries for some of these species are very important economically, and their landings value represents nearly 30% of all fish and shellfish landings value in the world (Smith and Addison 2003). The most valuable species landed into Scotland (UK) is the decapod crustacean species *Nephrops norvegicus* (the Norway lobster).

Aside from their economic importance, decapod crustaceans are also of significant ecological importance. They are diverse and abundant in coastal ecosystems, with complex life histories and population dynamics. Many decapods produce pelagic larvae that remain and develop in the water column for days, weeks or months before settling as juveniles into the benthos. After metamorphosis from the pelagic stages, they feed, grow and are fed upon in the benthic community, with survivors recruited to the adult population. For species with planktonic larval stages and then benthic juveniles and adults, there are many factors, biotic and abiotic, which influence the abundance, growth and survival of a cohort. Traditionally, these factors are divided between: (i) pre-settlement processes, operating

after the larvae are hatched until they settle into the benthos and (ii) post-settlement process, operating throughout the whole benthic period (Wahle 2003).

Although there are many published studies on decapod crustacea, these are usually related to benthic adults and juveniles and knowledge of their larval phases remains poor. It is necessary to identify and understand the biology of early stages and the environmental factors that affect them, to understand fully the population dynamics of decapods. Improving our understanding of their seasonal population dynamics will contribute to increased understanding of the role of decapods in marine ecosystems. This will also contribute to future developments of stock assessment and management methods for sustainable exploitation of harvestable resources.

As part of a PhD project, this case study chapter is a first approach to ongoing investigations into the ecological role and recruitment of decapod larvae. One aim is to study differences and temporal patterns in the decapod larval communities in the east and the west coast of Scotland. These patterns we may then relate to some or all of several environmental variables known to effect larval production and development. To solve such questions we aim eventually to analyse up to five years of plankton samples from the east and west coast of Scotland. In this case study we only use the abundance of 12 decapod crustacean families studied over two years.

The two sampling points selected in this study of decapod larvae are part of an FRS (Fisheries Research Services) long-term weekly monitoring programme. Sites are located in the northeast (Stonehaven, sampled since 1997) and in the northwest of Scotland (Loch Ewe, since April 2002). Samples analysed for this case study were taken during 2002 and 2003. Both locations are ecologically different, and some of the variables that could influence the larval dynamics are as follows. Due to the shallow depth (50 m), strong tidal flows and the wind effects (Otto et al. 1990; Svendsen et al. 1991), Stonehaven (Scottish northeast coast) is a very dynamic site and the water column remains well mixed for most of the year, with some thermal stratification in calmer summer weather. The fjordic sea loch, Loch Ewe, is a more enclosed sea area and is subject to different environmental conditions and influences. The difference between these reference sites is likely to play an important role in the settlement and recruitment of decapod larvae and could influence, along with other environmental variables, the species diversity and composition. The data show that seabed features differ between locations with hard sand and rocky bottoms off Stonehaven and muddy sand in Loch Ewe.

20.2 The data

The samples were collected between April 2002 and April 2003 from two sampling stations. One is located approximately 3 km offshore of Stonehaven, northeast of Scotland, UK (56° 57.8' N, 02° 06.2' W), whereas the second is located in a sea loch in the northwest of Scotland, Loch Ewe (57°50.14'N, 05°36.61'W) (Figure 20.1).

The plankton samples analysed were taken with a bongo net of 200-μm mesh size towed vertically from near the bottom and immediately preserved. Environmental variables were also sampled: temperature and salinity, measured near the surface and near the bottom (at 1 m and 45 m in Stonehaven and 1 m and 35 m in Loch Ewe), several nutrient chemical concentrations (not used in this case study) and chlorophyll a measured fluorimetrically from surface water sampled over 0–10 m. The decapod larvae present in the samples were counted and taxonomically identified where possible (following dos Santos and Gonzalez-Gordillo (2004), Fincham and Williamson (1978), Pike and Williamson (1958, 1964, 1972), Williamson (1957a, 1957b, 1960, 1962, 1967 and 1983), among others).

The data used in this chapter consist of abundance values for families expressed as numbers of individuals per m^2 (Table 20.1). A total of 24 plankton samples was analysed in the case of Stonehaven, and 21 for Loch Ewe. This ranged from three or four samples analysed per month in the spring/summer to one or two per month in the winter, when decapod larvae are essentially absent from the plankton samples.

The explanatory variables used in these analyses (Table 20.2) were temperature, salinity and chlorophyll a values. Also used as explanatory variables were the locations where the decapod families were found (Stonehaven vs. Loch Ewe) and the collection year (2002 vs. 2003).

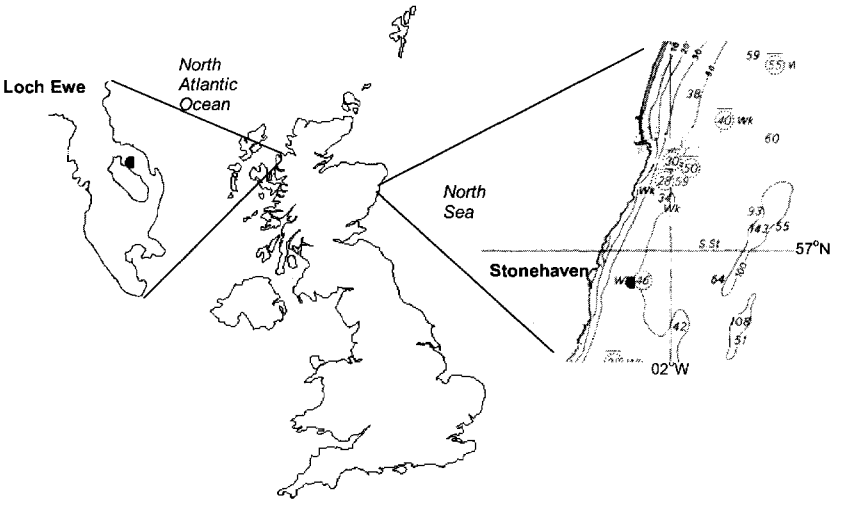

Figure 20.1. Map of the study area. The dots indicate the sampling points.

Table 20.1. List of families used to calculate the species index function.

Families	
Galatheidae	Pandalidae
Paguridae	Alpheidae
Porcellanidae	Callianassidae
Hippolytidae	Upogebiidae
Crangonidae	Laomediidae
Processidae	Nephropidae

Table 20.2. Available explanatory variables.

Explanatory Variable	Notation
Temperature at 1 m	T1m
Temperature at 45–35 m	T45.35m
Salinity at 1 m	S1m
Salinity at 45-35 m	S45.35m
Chlorophyll a	Cho.10m
Location	Location
Year	Year

Options for the analyses of these data include multivariate methods such as principal component analysis, redundancy analysis, canonical correspondence analysis, and even methods such as the Mantel test and ANOSIM. These techniques can be used to detect whether there are relationships between the abundance of the 12 decapod families and the chosen explanatory variables. Examples of such analysis are given in other case study chapters. Here, the aim is different; we want to show how to decide between using parametric models, such as linear regression and generalised linear modelling, versus using non-parametric methods such as additive modelling and generalised additive modelling. All these methods use one response variable, and describe it in terms of the explanatory variables. To obtain one response variable, we can either work on data from one family or convert the data on the 12 families into a diversity index. Obvious candidates are the Shannon–Weaver index, richness index, total abundance, Simpson index, Berger–Parker index and Macintosh index. A detailed discussion on diversity indices can be found in Magurran (2003). In practise, most diversity indices are highly correlated with each other and therefore your choice of index should be based on ecological rather on than statistical arguments. In the remaining part of this chapter, we use richness as the response variable because we are interested in the numbers of families rather than in species abundances. The list of families used to calculate richness is presented in Table 20.1. The total number of sites was 45. The first seven families were measured at 10 or more sites. These families were also more abundant. The available explanatory variables are in Table 20.2. The variables chlorophyll a, salinity at 45–35 m and temperature at 45–35 m have four missing values in the year 2002.

20.3 Data exploration

Figure 20.2 shows a pairplot of the species richness (response variable) and all explanatory variables. The first row and column shows the relationship between richness and each of the explanatory variables. The correlations between richness and explanatory variables (first column) indicate that there might be temperature and chlorophyll a effects. All other panels can be used to asses collinearity: the correlation between explanatory variables. It is clear that temperature (T1m) at the surface and at 45–35 m (T45-35m) cannot be used jointly in any analysis, as the correlation is 0.98. Because (i) T1m has a higher correlation with species richness and (ii) T1m has no missing values (T45-35m has 4 missing values) we decided to use T1m. The correlation between salinity at the surface (S1m) and at 45–35 m (S45.35m) is not as high as for temperature, but we decided to use only S1m. Note that temperature and the nominal variable year have a correlation of 0.74. If a nominal variable has more than two classes, the correlation coefficient is meaningless. However, in this case there are only two classes, which are coded as 2002 and 2003. This means that a correlation of 0.74 indicates that temperature in the second year was higher.

Month and temperature

Note that there is a clear pattern in the month-temperature panel in Figure 20.2. It shows a nice sinusoidal pattern reflecting the seasonal temperature pattern. In the summer months, temperature is higher than in the winter, but we also have higher richness when the temperature is high. This means that we can either use month as an explanatory variable or temperature, but not both, as they both reflect the same ecological effect. Because temperature has a linear effect on richness, we decided to use it instead of month, which clearly has a non-linear relationship with richness. Another reason to avoid using month is that in some months only one observation was taken, whereas in other months we have three of four observations. This would cause problems if month were treated as a nominal variable.

In practice, the situation is a bit more complicated than this as we are actually dealing with a time series. However, for the moment we ignore the time series aspect and assume that temperature represents the seasonal pattern. We return to this aspect in the discussion.

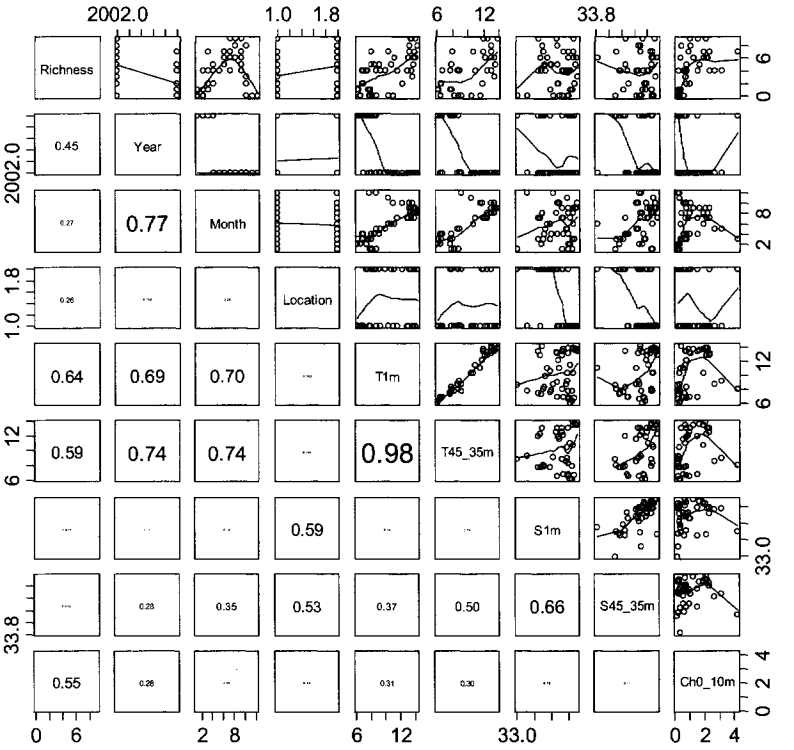

Figure 20.2. Pairplot of species richness and all explanatory variables. The lower diagonal part of the pairplot shows the (absolute) correlation coefficient in which the font size is proportional to the value of the correlation. The upper diagonal shows the scatterplots with smoothing curves.

Transformation

Cleveland dotplots (Figure 20.3) indicated that there are a few samples with slightly higher chlorophyll a values. A first round of application of statistical methods indicated that these observations were influential (e.g., resulting in wide confidence bands for the smoothing techniques), and therefore, we decided to apply a square root transformation to chlorophyll a. The original and transformed chlorophyll a data are presented in Figure 20.3.

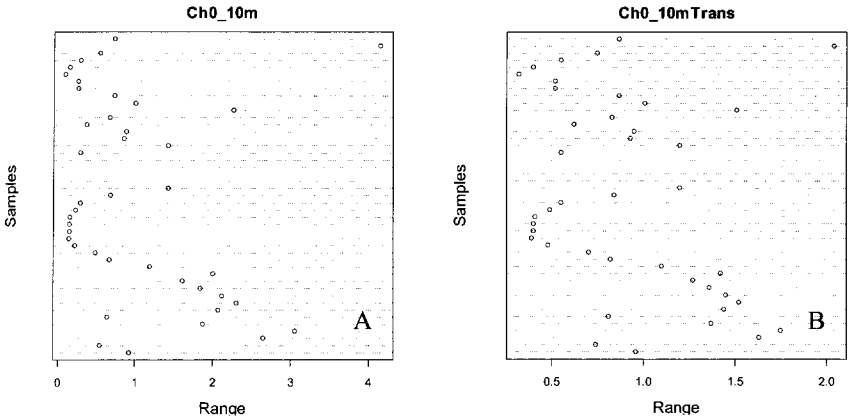

Figure 20.3. Dotplot of chlorophyll a (A) and square root transformed chlorophyll a (B).

20.4 Linear regression results

A linear regression model was applied to explain the variation in the richness index as a function of temperature, salinity, chlorophyll a (transformed), location and year. A forward and backward selection using the AIC was applied and the optimal model, as judged by the AIC, contains temperature, chlorophyll a and location. The estimated parameters, standard errors, *t*-values, etc. are given below.

	Estimate	Std. Error	*t*-value	*p*-value
Intercept)	−4.23	1.06	−3.97	<0.001
T1m	0.43	0.10	3.94	<0.001
chlorofilatrans	3.34	0.71	4.70	<0.001
factor(Location)2	1.55	0.58	2.68	0.011

Residual standard error: 1.76 on 37 degrees of freedom
Multiple R^2: 0.66, Adjusted R^2: 0.63
F-statistic: 24.19 on 3 and 37 df, *p*-value: <0.001

Temperature and chlorophyll a have a positive relationship with richness: the higher the temperature or the higher the chlorophyll a, the more decapod families are present. The results for the nominal variable location show that there are more decapod families found at Loch Ewe (location = 2) than at Stonehaven, which is the baseline value. Panel A in Figure 20.4 shows the standard graphical output of linear regression: residuals versus fitted values to verify the homogeneity assumption, a QQ-plot for normality, the scale-location plot for homogeneity and the Cook distance function for influential observations. None of the panels shows serious problems. The other three panels are residuals versus individual explanatory variables. Any patterns in these graphs are an indication of serious model misspecification. The residuals plotted versus temperature (Figure 20.4 B) show that

all samples with lower temperature have a positive residual. Meaning these samples are all under-fitted, as residuals are calculated as observed minus fitted values. Figure 20.4-C shows a graph of residuals versus transformed chlorophyll a. The samples with the six highest chlorophyll a values are all over-fitted (negative residuals). There is no clear residual pattern for location (Figure 20.4-D).

If a graph of residuals versus individual explanatory variables shows any pattern, then model improvement is required. One option is to include quadratic terms in the model, for example:

$$\text{Richness}_i = \alpha + \beta_1 \text{ Temperature} + \beta_2 \text{ Temperature}^2 + \ldots + \varepsilon_i$$

However, the linear and quadratic terms might be highly collinear, and the fit of such quadratic models might be poor. Furthermore, based on prior knowledge of the study system there is no reason to expect a quadratic relationship. Another way to improve the model is by including interaction terms. We added two-way interactions between all possible combinations, but none of the interactions resulted in a significant model improvement as judged by F-statistics of nested models and individual t-values. An alternative approach is a smoothing method like additive modelling or generalised additive modelling.

Additive modelling is based on the Gaussian distribution, and the GAM can use other distributions like the Poisson distribution. The choice of distribution to use can also be inferred from the linear regression results, namely from the plot of residuals versus fitted values; see Chapters 5–7 for further details. If an increase in spread for larger fitted values can be seen, the application of GLM or GAM with a Poisson distribution and log-link function should be considered. The choice between GLM and GAM can be made based on the graphs of residuals versus each explanatory variable. However, for these data, no clear violation of homogeneity can be seen (Figure 20.4), and therefore, the application of additive modelling seems to be appropriate. This decision process is illustrated with a flowchart in Figure 3.1.

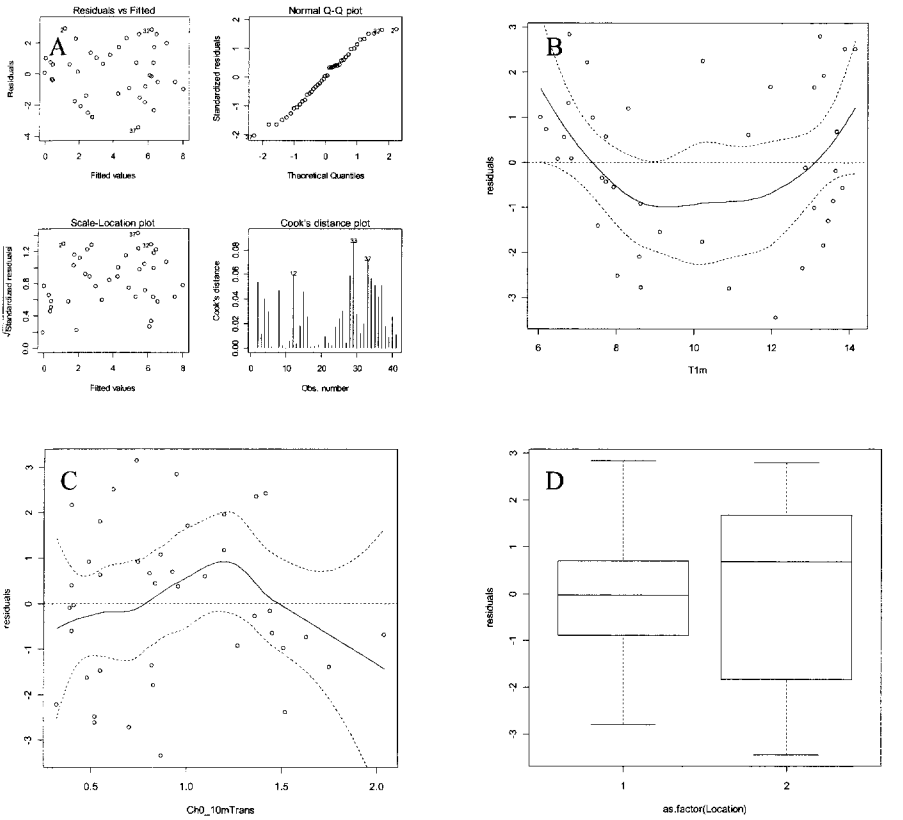

Figure 20.4. Graphical output of the linear regression model. Panel A: standard regression output: residuals versus fitted values, QQ-plot of residuals, a scale location diagram and the Cook distances. The other panels show the residuals versus fitted values for temperature (B), chlorophyll a (C) and location (D). In panels B and C Loess smoothing lines and corresponding point-wise 95% confidence bands were added to aid visual interpretation.

20.5 Additive modelling results

The following additive model (Chapter 7) was applied:

$$\text{Richness}_i = \alpha + f_1(\text{T1m}_i) + f_2(\text{S1m}_i) + \text{f}_3(\text{Ch0.10mTrans}_i) + \text{Year}_i + \text{Location}_i + \varepsilon_i$$

Year and location are fitted as nominal variables. Results of this model indicated that various smoothers can be dropped from the model. Hence a model selection is required. We used the AIC method to choose the optimal model. The AIC of the model containing all explanatory variables is 160.37 (Table 20.3).

Dropping only year gives AIC = 160.04, dropping only Location results in AIC = 165.74, etc. The lower the AIC, the better is the model. In this case, it suggests dropping the nominal variable Year as it is the least important variable. The new model is of the form:

$$\text{Richness}_i = \alpha + f_1(\text{T1m}_i) + f_2(\text{S1m}_i) + f_3(\text{Ch0.10mTrans}_i) + \text{Location}_i + \varepsilon_i$$

The process of dropping variables then repeats, but the AIC indicates that this is the optimal model (Table 20.4). However, the confidence bands and p-value ($p = 0.14$) for the salinity smoother were rather large. This is a typical example in which the AIC is too conservative and some extra model selection steps are required. Based on the width of the confidence bands we decided to remove salinity from the model. All remaining variables were significant at the 5% level.

Table 20.3. The results of the first step in the backward selection process in additive modelling. An 'X' means that the explanatory variable was used in the model. In each model cross-validation was applied to estimate the optimal degrees of freedom for each smoother. Bold type indicates the lower AIC and for hence the optimal model.

Variables	Selected Explanatory Variables					
Year	X	-	X	X	X	X
Location	X	X	-	X	X	X
T1m	X	X	X	-	X	X
S1m	X	X	X	X	-	X
Ch0.10mTrans	X	X	X	X	X	-
AIC	160.37	**160.04**	165.74	164.6	164.25	196.13

Table 20.4. The results of the second step in the backward selection process in additive modelling. An 'X' means that the explanatory variable was used in the model. In each model cross-validation was applied to estimate the optimal degrees of freedom for each smoother. Bold type indicates the lower AIC and for hence the optimal model.

Variables	Selected Explanatory Variables				
Location	X	-	X	X	X
T1m	X	X	-	X	X
S1m	X	X	X	-	X
Ch0.10mTrans	X	X	X	X	-
AIC	**160.04**	163.94	169.74	162.4	197.06

The final model is of the form:

$$\text{Richness}_i = \alpha + f_1(\text{T1m}_i) + f_2(\text{Ch0.10mTrans}_i) + \text{Location}_i + \varepsilon_i$$

In each step of the backwards selection, cross-validation was applied to estimate the optimal degrees of freedom for each smoother. The numerical output of the optimal model is given by

	Estimate	std. err.	*t*-ratio	*p*-value
Intercept	3.16	0.33	9.352	<0.001
factor(Location)2	1.78	0.54	3.265	0.002

Approximate significance of smooth terms:

	edf	chi.sq	p-value
s(T1m)	2.15	3.36	0.005
s(Ch0_10mTrans)	1.32	3.82	0.002

R-sq.(adj)=0.697. Deviance explained=73.1%. Variance=2.5872. n = 41

These results indicate that the explanatory variable Location is significantly different from 0 at the 5% level. At location two, 1.78 more families are expected compared with location one (Stonehaven). The degrees of freedom for the smoothers of temperature and chlorophyll a are rather small, indicating nearly linear relationships. The smoothers are given in Figure 20.5, and the shape of the line also indicates a nearly linear relationship. Figure 20.6 shows the residuals of the optimal additive model plotted versus the explanatory variables. As there are no clear patterns now, we can conclude that the additive model is better than the regression model. A more formal approach that comes to the same conclusion is an analysis using the *F*-test. The output is given by

Model 1: Y ~ s(T1m) + s(Ch0_10mTrans) + factor(Location)
Model 2: Y ~ T1m + Ch0_10mTrans + factor(Location)

	Resid. df	Resid. Dev	df	Deviance	*F*	*p*-value
1	35.53	91.92				
2	37.00	115.35	1.470	−23.430	6.162	0.009

Model one is the additive model in which cross-validation is applied to estimate the optimal degrees of freedom for each smoother. Model two is the (nested) linear regression model. The change in deviance is 23.43. The corresponding *F*-statistic is equal to 6.16 ($p = 0.009$) and indicates that the more complicated additive model performs better.

20.6 How many samples to take?

In the previous section, backwards selection was used to find the optimal additive model. It contained temperature, chlorophyll a (transformed) and a location effect. The model was based on 41 samples collected over two years, and only analyses part of a much larger dataset held at at the FRS Marine Laboratory in Aberdeen (UK). At the time of writing, data from several other years are still waiting to be analysed. A wide range of measurements, in addition to those discussed here, are taken for each sample (e.g., DNA information), and a full analysis will be time consuming and costly. As this is a problem common in ecology, it is important to know how many samples per year should be collected.

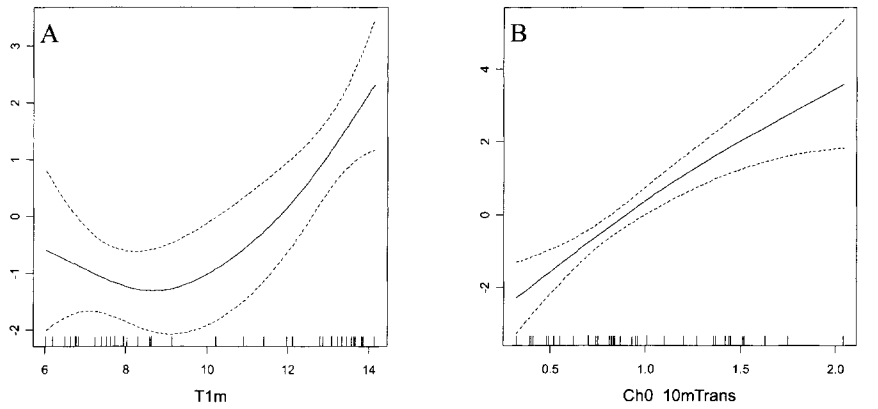

Figure 20.5. Smoothed curves obtained by the additive model. A: The temperature effect on species richness. B: The chlorophyll a effect. Dotted lines are 95% confidence bands. The horizontal axes show the environmental gradient and the vertical axes the contribution of the smoother to the fitted values.

One option is to apply a power analysis using the linear regression results. However, the regression model was not optimal, and therefore, it is less appropriate to use its results for a power analysis (Zar 1999). The use of power analysis for additive modelling is less well developed than it is for regression modelling. However, in this section, we show how we can still get a reasonable idea of the optimal sample size. One approach is to ask the question, what would happen if we had only analysed 35 samples instead of 41: a reduction in sample size of approximately 10%? Will we still find the same relationships as in Figure 20.5? If the answer is yes, then perhaps we can take 10% fewer samples for the remaining years. If the relationship is changed, then clearly we should not omit the samples. One way of assessing what happens if 10% of the data are omitted, is as follows: (i) apply the additive model using all 41 samples, (ii) remove a random 10% of the data and reapply the additive model on the remaining data, and (iii) repeat step (ii) a few times (say 20) so that the choice of the actual omitted samples does not have any influence. This process can then be repeated for 15%, 20%, 25%, 30%, etc. of omitted data. The effect of omitting data on the temperature effect is presented in Figure 20.7. The original smoothing curve and corresponding 95% point-wise confidence bands are plotted (thick lines) plus 20 smoothing curves obtained by omitting x% of the data. Figure 20.7-A shows the smoothing curves if 10% of the data are omitted. The variation among the 20 curves is relatively small compared with the 95% confidence bands of the original smoother. If 25% of the data are omitted, this variation increases, and once 50% of the data are removed, it goes beyond the original range of confidence bands. The same set of graphs is presented in Figure 20.8 for the effects on chlorophyll a results when sample sizes are reduced.

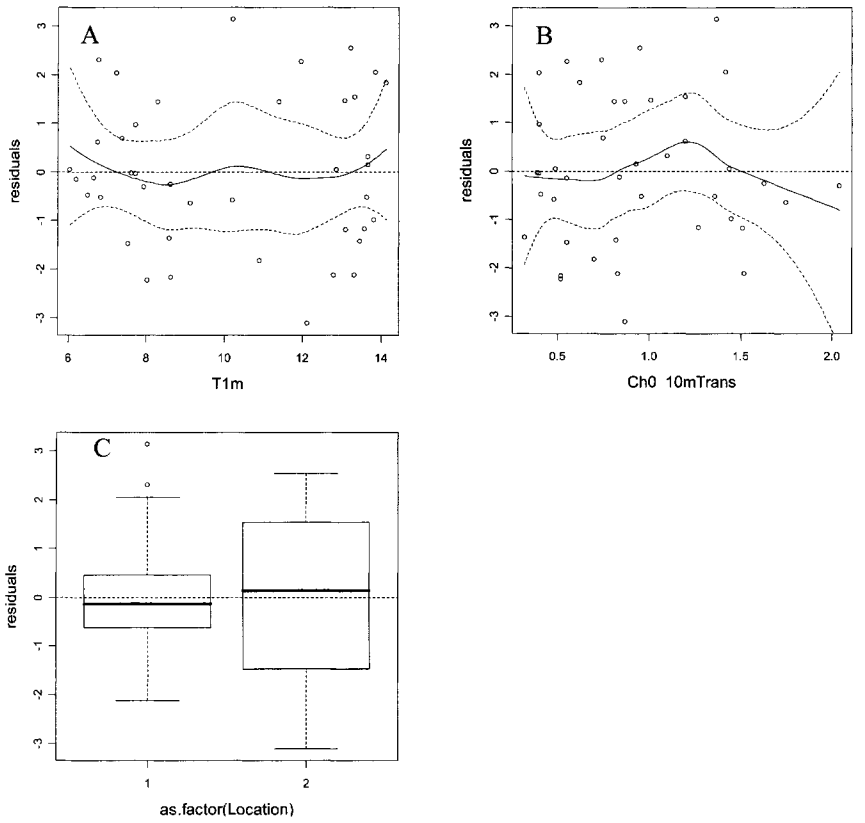

Figure 20.6. Residuals of the optimal additive model plotted versus the explanatory variables temperature (A), chlorophyll a (B) and location (C).

Both graphs seem to suggest that if 10% of the data are omitted, we end up with similar temperature and chlorophyll a relationships compared with using all data. Removing 25% still gives similar curves, but the variation between the fitted curves is higher. These results would suggest that if 90%, or perhaps even 85% of the available samples are processed and included in the analysis, similar temperature and chlorophyll a effects could be expected.

20.7 Discussion

The additive model performed better than a linear regression model and showed that all the major factors analysed (temperature, food and location) are influential

in determining the "species" richness (at a family level) of the decapod larval communities examined at the Loch Ewe and Stonehaven sites.

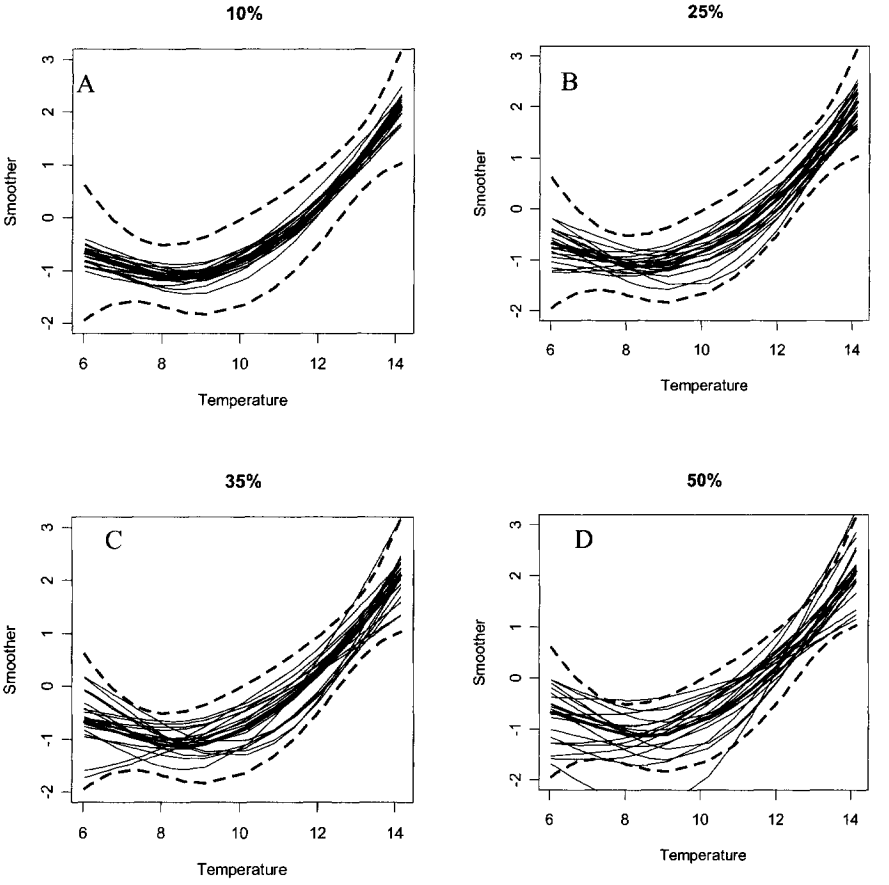

Figure 20.7. Original temperature smoothing curves and 95% point-wise confidence bands and 20 smoothing curves obtained by omitting x% of the data, where x = 15 (upper left), x = 25 (upper right), x = 35 (lower left) and x = 50 (lower right).

It should be noted however that measuring true species richness involves discrimination to species level, which may yield different results. However the results obtained here are at least compatible with intuitive expectations. The major determinant of species richness, over the time periods examined for both sites, appears to be temperature. This is not unexpected as Loch Ewe generally has a higher annual average temperature (10.4°C) than Stonehaven (9.8°C).

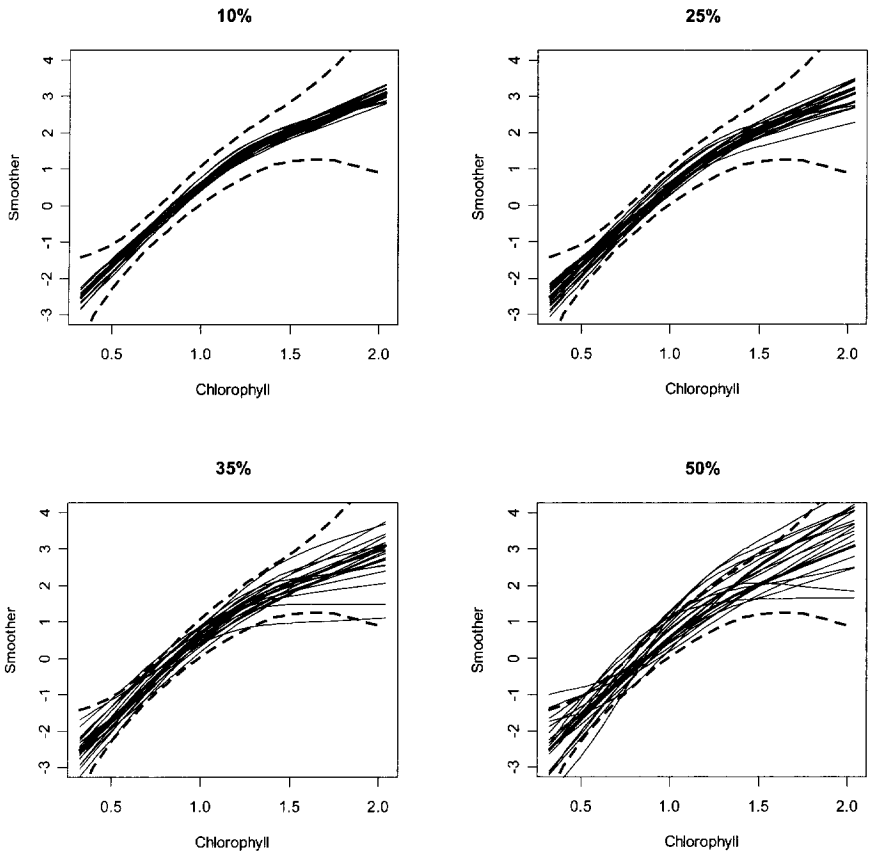

Figure 20.8. Original chlorophyll a smoothing curves and 95% point-wise confidence bands and 20 smoothing curves obtained by omitting x% of the data, where $x = 15$, $x = 25$, $x = 35$ and $x = 50$.

Many marine species groups show increased diversity with increasing temperature, at least when comparing sub-tropical with northern temperate and boreal communities along latitudinal gradients (Raymont 1980). Temperature is the prime environmental variable affecting the physiological rates of growth and reproduction. Temperature also indirectly affects the productivity of food organisms, and it has a major influence on larval development (Costlow and Bookhout 1969). Chlorophyll a is also important, and there were obvious differences between the two sites. The importance of chlorophyll a in the relationships is also expected, as it indicates the availability of decapod food availability. We can therefore conclude that higher field temperatures, increased food levels or

extended seasonal food production periods are all likely to be beneficial to the development of decapod larval communities.

The analysis here, therefore, confirms our expectations in a quantitative and objective manner, and allows the derivation of relationships to be applied to wider or more detailed studies.

The time aspect

The decapod data used in this chapter were measured monthly during a period of two years. In some months there are four observations and in other months only one. Once more data become available, time series analysis might be a more appropriate tool to analyse these data. However, this will not be easy as the data are irregularly spaced and most time series methods require equidistant observations.

In Chapter 16 we discussed how generalised least squares (GLS) is used to extend the linear regression model with an auto-correlation structure on the residuals. The same can be done for smoothing methods but confusingly, it is called additive mixed modelling. This combination of linear regression or additive modelling and auto-correlation structures on the noise is demonstrated in several case study chapters. The key to deciding when to apply linear regression and when to apply GLS is: 'do the residuals show any clear auto-correlation?' If they do, then GLS (or the smoothing equivalent) should be applied. For these data, we were on the borderline between the two approaches. Residuals of the optimal models were plotted against the explanatory variable Month (which was not used in the model selection procedure), and there was some small indication that winter months had lower residuals than summer months. We reapplied the smoothing models replacing temperature with month (as a continuous smoother), and the explained variance and model fit were similar to the additive models with temperature. We also defined a new nominal explanatory variable 'season' with values 1 in the summer months and 0 in the winter months and used it as an explanatory variable together with temperature, chlorophyll a and location. This gave a slightly better performance for the smoothing model in terms of residual patterns. But it also complicates the model's interpretation. We expect that once more data become available, GLS and/or its smoothing equivalent will be the more appropriate methods to apply on these data. GLS and its smoothing equivalent is applied in several case study chapters.

Acknowledgements

Part of this work was carried out as part of a Marie Curie Fellowship (QLK5-GH-99-50530-13). Thanks to Alain Zuur, Elena Ieno and Graham Smith for their invitation to contribute with this case study and thanks to all the people of the FRS Marine Laboratory and the University of Aberdeen who were involved in this project. Alain Zuur would like to thank Rob Fryer for a short discussion on the bootstrap approach.