Chapter 17
Additive Mixed Modelling Applied on Deep-Sea Pelagic Bioluminescent Organisms

A.F. Zuur, I.G. Priede, E.N. Ieno, G.M. Smith, A.A. Saveliev, and N.J. Walker

17.1 Biological Introduction

The oceans, with a mean depth of 3,729 m and extending to a maximum depth of 11 km comprise the largest habitat on earth. The distribution of living organisms in this vast environment is far from uniform and description of this variation in space and time is challenging, both from the point of view of sampling and of statistical analysis. Most life in the oceans is dependent on primary production in the surface layers, generally in the epipelagic zone down to a depth of 200 m, where there is sufficient solar radiation to sustain photosynthesis. Microscopic algae or phytoplankton containing the pigment chlorophyll intercept solar light and use the energy to combine CO_2 and water to produce simple sugars polysaccharides, oils, proteins, and all the other constituents of the living organism. The algae and phytoplankton are either consumed by planktonic animals or dies loses buoyancy and becomes part of the downward stream of particulate organic matter (POC) falling towards the sea floor. The primary consumers themselves produce faecal pellets that enhance the POC flux and also form the basis of the food chain in the surface layers of the oceans. Predators living at greater depths also ascend at night to feed on the surface riches and then descend during the day digesting and excreting as they go. Thus, surface-derived production is exported downwards by passive and active processes sustaining life throughout the water column down to the abyssal sea floor.

There is therefore a general pattern of decrease in species abundance and biomass with depth. There are linear and non-linear components to this decline. Pressure, which increases linearly with depth, tends to disrupt biochemical reactions so that deep living organisms have acquired specially adapted protein structures. Below the photic zone, temperatures become lower, defining a cut off at the thermocline beneath which biochemical processes are slower. In this zone, biomass consumes oxygen, which in the absence of replenishment by photosynthesis can result in an oxygen minimum zone at around 1,000 m depth where life can become impossible.

A.F. Zuur (✉)
Highland Statistics Ltd., Newburgh, AB41 6FN, United Kingdom

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_17,

Below this depth, sea water is cold, well-oxygenated, and ventilated by water originating from the sinking of cold water in the polar regions. Non-linearities can also be introduced by the presence of distinct water masses of different densities stacked on top of each other at different depths producing a layered effect in the ocean. Widder et al. (1999), for example, describes high animal abundances in a thin layer of less than a metre thick in the Gulf of Maine at a density discontinuity.

Most deep sea organisms are capable of emitting light in the form of bioluminescence. Usually blue light is produced either from discrete light organs or as luminescent secretions released into the water. This luminescence can be mechanically stimulated and is either the result of an alarm response by the organism or in the case of fragile animals, disintegration, and release of luminescent material into seawater. This is the mechanism that produces the so-called phosphorescent wake of ships and boats on calm nights in the open ocean. For scientific investigations, bathyphotometers are used. These work by pumping water through a chamber equipped with a light sensor, which counts the number of photons produced per litre of water. This system works well in the surface layers where organisms may occur at over 1,000 m^{-3} but pumping becomes impractical at depths greater than 1,000 m where organisms are rare. For investigations in deeper waters, Priede et al. (2006) developed a free-fall vehicle with a downward looking high sensitivity ISIT (Intensified Silicon Intensified Target) video camera focussed on a 0.19 m^2 mesh screen that filled the field of view of the camera. Flashes of light produced by luminescent organisms impinging on the screen as it descends at 0.6 $m \cdot s^{-1}$ through the water column are counted to estimate the number of bioluminescent organisms per m^3 of seawater at different depths. Since over 80% of deep sea organisms are capable of luminescence, this is a novel means of producing continuous vertical profiles of marine life abundances. In practise, the ISIT profiler cannot be used at depths less than 300 m, because surface light could damage the sensitive camera and obscure bioluminescent flashes. Luminescent sources are counted over 30-s time samples, which correspond to a depth interval of 18 m between readings depending on the exact descent rate of the lander, which can vary slightly from profile to profile.

Abundance of deep-sea bioluminescent organisms is also dependent on the intensity of overlying primary production that can vary considerably in different parts of the ocean. Temperate latitudes are characterised by highly seasonal peaks of primary production in the spring followed by a fall out of POC towards the seafloor during summer, whereas in the centre of tropical gyre regions, primary production is low and uniform throughout the year (Longhurst, 1998). In addition to seasonal and regional differences, primary production can be very irregular, occurring in patches such as eddies of water spinning in the vicinity of oceanic fronts.

Bradner et al. (1987), using a free-falling photomultiplier device, concluded that in the Pacific Ocean off Hawaii bioluminescence decreases exponentially with depth. The first results from the ISIT free-fall profiler in the Tropical Atlantic Ocean indicated a monotonic decline in abundance with depth, but the relationship was not truly exponential (Priede et al., 2006). These studies, however, gave no information on seasonal changes.

17.2 The Data and Underlying Questions

The data analysed in this chapter were gathered during a series of four cruises of the *Royal Research Ship Discovery* over two years (2001 and 2002) in the temperate NE Atlantic west of Ireland (Gillibrand et al., 2006); see Fig. 17.1. The primary purpose of the cruises was to investigate deep-sea fish living on the sea floor in the Porcupine Seabight and on the Porcupine Abyssal Plain. Cruises were organised in spring and late summer to collect samples before and after the seasonal downward flux of POC that occurs in June and July (Lampitt et al., 2001). Timing of the cruises could not be precisely controlled since ship allocation is determined by conflicts between requirements of different programmes and logistic considerations. In 2001, the cruises were in April and August, and in 2002, they were in March and October. The ISIT free-fall profiler was deployed opportunistically between trawling and other sampling operations. Each location where the ship stopped to launch the profiler is termed a station. Depending on the weather conditions, the crew would then prepare to launch the instrument over the stern (back) of the ship as it moved forward slowly at approximately 1 knot. Once the equipment was streaming behind the ship, it was released and allowed to fall towards the sea floor. A timer activated the recording system after a set delay; so the depth of starting the recording and the precise location of the profile depended on the promptness of deck and crane operations by the ship's crew. This introduced inevitable variation in data collection. At the maximum depth of 4,800 m, the descent would take over two hours, greater than the maximum one-hour recording capacity of the ISIT video system. The recorder was therefore set to start and stop at intervals to ensure sampling between the surface and the sea floor. Sometimes sampling was concentrated at particular depths. When

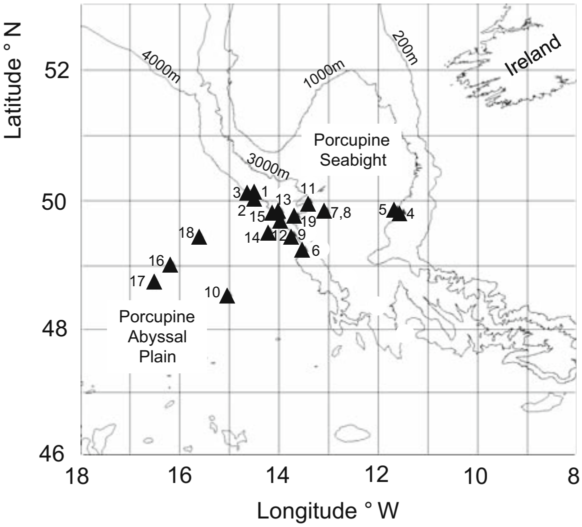

Fig. 17.1 Location of the 19 stations where measurements were taken

the vehicle had reached the sea floor an acoustic command from the ship, triggered release of ballast, and the vehicle ascended because of its positive buoyancy and was recovered back on board the *RRS Discovery*.

As there were no previous data of this kind to inform a formal sampling design, three aims influenced final sample design, which was also constrained by the ongoing ship programme: (i) to produce some replicates as close together as possible in time and space, (ii) to investigate spatial variation in waters of different depths, and (iii) to produce a balanced set of samples across the seasons. It was evident as soon as the first data were viewed that, particularly in summer and autumn, a simple paradigm of an exponential decrease with depth was inappropriate. This has resulted in the need for a sophisticated approach to the statistical analysis describing the profiles and answering questions about spatial homogeneityover the geographical sample area and about seasonal differences.

The aim of this chapter is not only to analyse the data but also to explain how to make multi-panel graphs for grouped data. We have used these graphs in nearly every chapter, but here we will use them in more detail and also make our own panel functions. A detailed explanation of these graphs can be found in Chapter 3 of Pinheiro and Bates (2000). They used specific functions from the `nlme` package to create multi-panel figures. Instead of using the Pinheiro and Bates `plot` function for grouped data, we will use the more flexible `xyplot` function from the lattice package. We show that intelligent use of graphs considerably simplifies the statistical analysis. Perhaps we should phrase this differently. Good multi-panel graphs help us to develop questions in cases when you are not 100% sure in which direction to steer the analysis. Call it a ‘hypothesis generating brainstorming session’. Therefore, we start constructing multi-panel graphs for grouped data (which can also be seen as part of the data exploration), and this will help us decide what type of statistical models to apply and how to apply them.

17.3 Construction of Multi-panel Plots for Grouped Data

Possible explanatory variables are time (of the day), month, year, station, season, latitude, longitude, and depth. Except for the last three variables, all are nominal. Figure 2.11 in Chapter 2 shows the bioluminescence sources per m^2 plotted against depth for each individual station. The graph indicates that the profiles from stations 4, 5 and 10 can be dropped, as the depth range is considerably smaller than for the other profiles. The question is now which profiles are similar and which are different. We discuss various different approaches.

17.3.1 Approach 1

Measurements took place in months 3 (March), 4 (April), 8 (August), and 10 (October). It may be the case that profiles from the same month are similar and

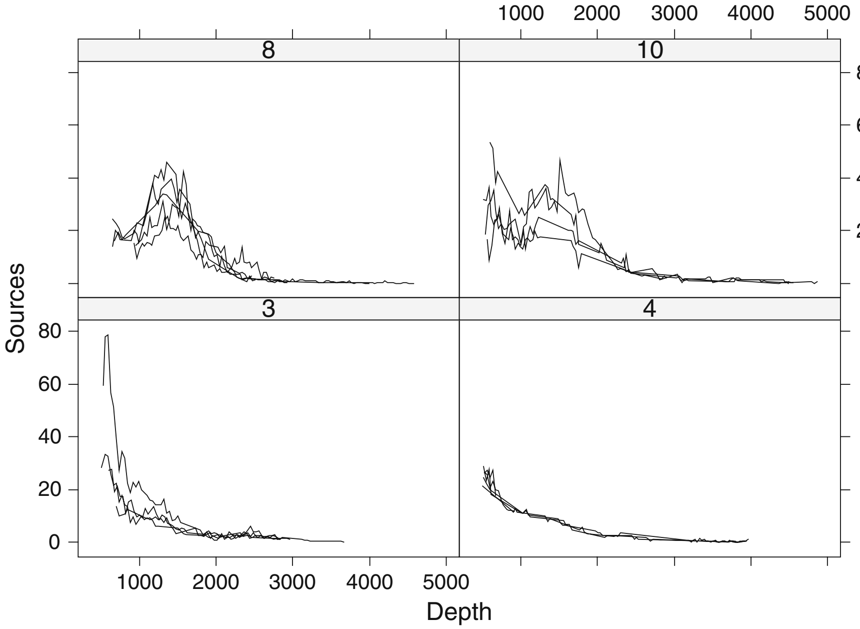

Fig. 17.2 Source–depth profiles per month. Each line represents a station. Note that the April profiles are similar. The graph may be improved by allowing for different ranges along the vertical axes

profiles from different months are dissimilar. Before applying complicated statistical methods to test this, we will draw a multi-panel graph. It has four panels. The first panel contains the profiles from month 3, the second panel from month 4, etc. The following R code accesses the data and draws the multi-panel graph with four windows (Fig. 17.2).

```
> library(AED); data(ISIT)
> ISIT$fMonth <- factor(ISIT$Month)
> ISIT$fStation <- factor(ISIT$Station)
> ISIT$fYear <- factor(ISIT$Year)
> ISIT2 < -ISIT[ISIT$fStation != "4" &
                ISIT$fStation != "5" &
                ISIT$fStation != "10" ,]
> library(lattice)
> MyLines <- function(xi, yi, ...){
     I <- order(xi)
    panel.lines(xi[I], yi[I], col = 1)}
> xyplot(Sources ~ SampleDepth | fMonth, data = ISIT2,
   groups = fStation, xlab = "Depth", ylab = "Sources",
   panel = panel.superpose,
   panel.groups = MyLines)
```

The code starts by accessing the data. The object ISIT2 is identical to ISIT (original data), except that stations 4, 5, and 10 are removed. It then creates a function called MyLines. Its task is to draw a line in the panels, while avoiding spaghetti plots. Finally, the xyplot creates a multi-panel figure with four windows. Each panel corresponds to a month. The option groups specifies that the data from the same station are grouped. The command panel = superpose ensures that lines for all stations (defined by groups) from the same month are superimposed in the same panel. Finally, the panel.groups option specifies which task should be carried out on the data defined by the groups option (station in this case). It calls our own function MyLines. The graph shows that the profiles in April are similar, but there are more differences between the profiles in other months. There is also more variation in the sources in the first 2,000 m compared to the deeper depths. This immediately indicates problems with heterogeneity. However, the April profiles do not seem to have this problem. This means that we may need models that allow for heterogeneity along depth in some months or stations, but not in all. Stations 1, 2, and 3 were all completed as close as possible to each other and all within a period of 90 hours, indicating that these can be considered as good replicate samples.

The measurements were taken in four months spread across two years, and the xyplot can easily be extended to draw a multi-panel plot with month and year information. We can also tidy up our R code. Panel labels now have a white background, the *y*-axes are allowed to have different ranges, and the label along the *y*-axis has m^{-3}, something that may take some time to work out how to do. We also divided depth by 1,000 to minimise the number of zeros in the labels along the horizontal axes. More sophisticated methods exist for this, see, for example, the labels option in the xyplot help file.

```
> xyplot(Sources ~ SampleDepth / 1000 | fMonth * fYear,
    groups = fStation, data = ISIT2,
    strip = function(bg = 'white', ...)
    strip.default(bg = 'white', ...),
    scales = list(alternating = TRUE,
                  x = list(relation = "same"),
                  y = list(relation = "free")),
    xlab = "Depth (km)",
    ylab = expression(paste(Sources," m"^{-3}, "")),
    panel = panel.superpose,
    vpanel.groups = MyLines)
```

The resulting graph in Fig. 17.3 shows that measurements in April and August only took place in 2001 and the March and October sampling only in 2002.

This makes it impossible to test for a month-year interaction, and we only use month as an explanatory variable. Similar problems exist for the explanatory variable time of the day. This reduces the explanatory variables to depth, month, station, latitude, and longitude. There is also a risk with the last three variables as each station was at a unique latitude and longitude (Fig. 17.1), a certain degree of

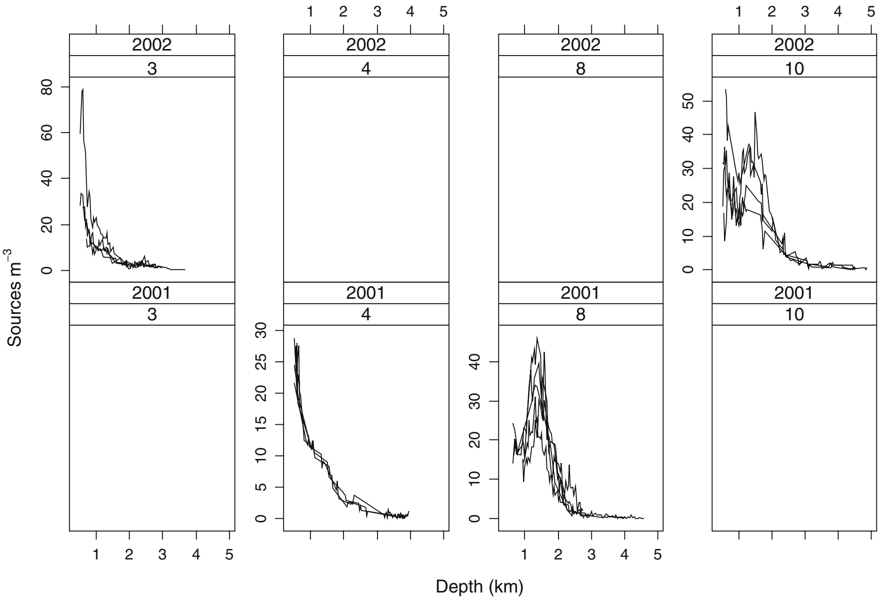

Fig. 17.3 Source–depth profiles by year and month. The *lower four panels* are for 2001 and the *upper four panels* for 2002. From *left to right* are the months

collinearity exists between station against latitude and longitude. This may become an issue if we use models that contain station as a factor, and latitude and longitude as smoothers or continuous explanatory variables.

The results so far indicate that there is a non-linear depth effect. In some months, the profiles are similar and in other months, profiles are not that similar, and there is heterogeneity between groups of profiles and within a station along depth. In the next section, we need a statistical model that describes the sources as a function of depth, station, month, latitude, and longitude. A possible model is of the form

$$S_{is} = \alpha_i + f_i(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma^2 \times |\text{Depth}|^{\delta_i}) \quad (17.1)$$

The sources at station i at depth s, S_{is}, are modelled as an intercept that differs per station, a smoothing function of depth, a month effect, and ε_{is} is the unexplained information. The smoothing function f has an index i indicating that the shape of the smoother can be different per station. This means that the source–depth relationship is allowed to differ per station. From a computing point of view, this is rather ambitious as there are 17 stations. Furthermore, the multiple panel graph in Fig. 2.11 indicates that we may expect heterogeneity along the depth gradient. Recall from Chapter 4 that there are different ways of implementing such an error structure. One option is the varPower method given in Equation (17.1). It models the residual spread for profile i in such a way that its variance is proportional to the variance covariate Depth. It is also possible to attach an index i to the variance σ^2. This allows for a different spread per station. Other alternatives exist and will be considered below. The only information that has not been used yet is spatial location. The model in Equation (17.1) assumes that residuals from different depths and stations are uncorrelated.

It may not be possible to fit the proposed model as it requires 17 smoothers with different degrees of freedom and a large number of variance components. However, the shape of the smoothers in Figs. 17.2, 17.3, and 17.4 indicate that various profiles are similar, and perhaps, we can replace these by one smoother. This means that Equation (17.1) can be changed into

$$S_{is} = \alpha_i + f_j(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma^2 \times |\text{Depth}|^{\delta_i}) \qquad (17.2)$$

The only difference is the index attached to the smoother: a j instead of an i. We are now looking for groups of profiles that can be modelled by a single smoother. Hopefully, j will take only a limited number of values. Something like $j = 3, 4, 8$, and 10 referring to the four months. In this case, profiles of each month are modelled by a single smoother for each month and we end up with a model that has only four smoothers. We could also try a model with only one smoother. However, the shape of the smoothers in Figs. 17.2, 17.3, and 17.4 indicate that things will not be as simple as this. The smoothers from month 4 (station 1, 2, and 3) are very similar and may be summarised by only one smoother, but then it becomes rather difficult to

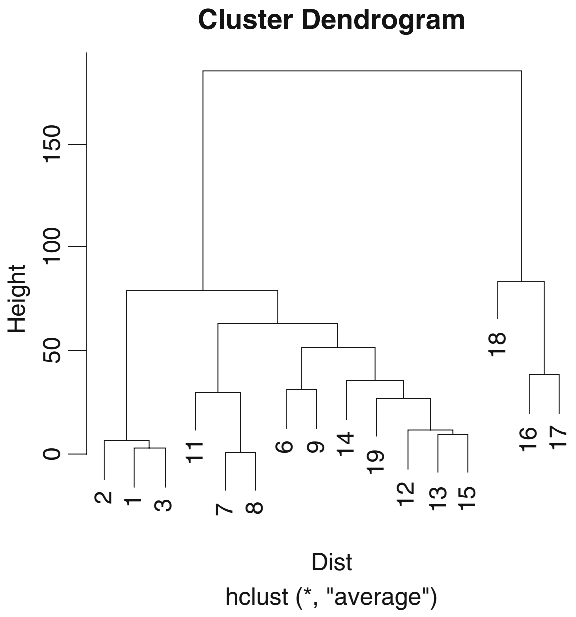

Fig. 17.4 Dendrogram representing the Euclidean distances between the geographical position of the stations. Based on the dendrogram, the following groups of stations were selected: (i) stations 1, 2, and 3, (ii) stations 6 and 9, (iii) stations 7, 8, and 11, (iv) stations 12, 13, 14, 15, and 19, (v) stations 16 and 17, and (vi) station 18

group profiles based on eyeballing. The original motivation for making the grouped multi-panel graphs was to detect groups of profiles.

17.3.2 Approach 2

Another way to group profiles is based on their geographical position. We can look at the map in Fig. 17.1 and group stations that are close to each other. A slightly less subjective approach is to calculate distances between the stations and apply clustering on the (Euclidean) distance matrix. The results can be presented in a dendrogram (Zuur et al., 2007). Judging from the dendrogram which stations should be grouped together is still subjective, but less subjective than looking at the map in Fig. 17.1. The following R code extracts the *x*- and *y*-coordinates for each station (it would have been easier to read them from a 16-by-2 ASCII file), calculates the Euclidean distances between the 16 stations, applies clustering with average linkage (Zuur et al., 2007), and presents the results in a dendrogram (Fig. 17.4). Stations that are grouped first (at the bottom) are close to each other.

```
> Xcoord <- vector(length = 16)
> Ycoord <- vector(length = 16)
> UStation <- unique(ISIT2$Station)
> for (i in 1:16) {
  Xcoord[i] <- ISIT2$Xkm[UStation[i]==ISIT2$Station][1]
  Ycoord[i] <- ISIT2$Ykm[UStation[i]==ISIT2$Station][1]
}
#Calculate a distance matrix between the 16 stations
#using Pythagoras
> D <- matrix(nrow = 16, ncol = 16)
> for (i in 1 : 16){
    for (j in 1 : 16){
      D[i,j] <- sqrt((Ycoord[i] - Ycoord[j]) ^ 2 +
                     (Xcoord[i] - Xcoord[j]) ^ 2)}}

> colnames(D) <- unique(ISIT2$Station)
> rownames(D) <- unique(ISIT2$Station)
> MyNames <- unique(ISIT2$Station)
#Apply clustering
> Dist <- as.dist(D)
> hc <- hclust(Dist, "ave")
> plot(hc, labels = MyNames)
```

The dendrogram in Fig. 17.4 suggests using the following groups of stations: (i) stations 1, 2, and 3, (ii) stations 6 and 9, (iii) stations 7, 8, and 11, (iv) stations 12, 13, 14, 15, and 19, (v) stations 16 and 17, and (vi) station 18. It should be noted that this grouping is only based on the geographical position of the stations and information on sources at these stations is not taken into account. Comparison with Fig. 17.1 shows that stations 1, 2, and 3 are very close to each other, and stations 16, 17, and 18 are the offshore stations. Between these two groups are stations within the mouth of the Porcupine Seabight.

17.3.3 Approach 3

The last approach used here to group profiles is as follows. Two profiles can be labelled as similar if they are highly correlated. The problem is that we cannot calculate a (Pearson) correlation coefficient because the data are not measured at exactly the same depths. For example, the first four measurements of station 1 are at 517, 547, 582, and 614 m. For station 2, these are 501, 865, 989, and 927 m. One way to get source values at the same depth at both stations is to apply additive modelling on each profile and predict source values at predefined depth intervals. This gives us two profiles with (predicted) values at the same depth, allowing us to calculate a correlation coefficient. If we do this for all stations, we can calculate a 16-by-16 correlation matrix. To visualise the patterns in this matrix, non-metric multidimensional scaling or clustering can be used to identify groups of profiles.

The following R code was used:

```
> MyDepth <- seq(from = min(ISIT2$SampleDepth),
                 to = max(ISIT2$SampleDepth), by = 25)
> NEWSOURCES <- matrix(nrow = 175,ncol = 16)
> NEWSOURCES[] <- NA
> library (mgcv)
> j <- 1
> for (k in MyNames){
   Mi <- gam(Sources ~ s(SampleDepth), data = ISIT2,
            subset = (ISIT2$Station == k))
   Depthi <- ISIT2$SampleDepth[ISIT2$Station == k]
   I1 <- MyDepth > min(Depthi) & MyDepth < max(Depthi)
   mynewXdata <- data.frame(SampleDepth = MyDepth[I1])
   M.pred <- predict(Mi, newdata = mynewXdata)
   NEWSOURCES[I1,j] <- M.pred
   j <- j + 1 }
> D <- cor(NEWSOURCES, use = "pairwise.complete.obs")
> colnames(D) <- unique(ISIT2$Station)
> rownames(D) <- unique(ISIT2$Station)
> Dist <- as.dist(1 - D)
> hc <- hclust(Dist, "ave")
> plot(hc, labels = MyNames)
```

The code starts by calculating the depth gradient along which we will predict source values. The variable `MyDepth` goes from the smallest to the largest

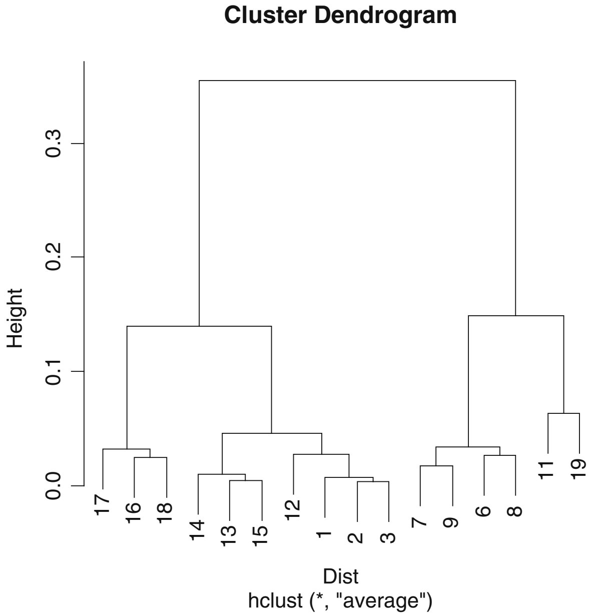

Fig. 17.5 Dendrogram obtained by applying clustering on the correlation matrix `Newsources`. Average linkage was used. The dendrogram implies the following groups of stations: (i) stations 1, 2, 3, and 12, (ii) stations 13, 14, and 15, (iii) stations 6, 7, 8, and 9, (iv) stations 11 and 19, and (v) stations 16, 17, and 18

observed depth value in the study with steps of 25 m. This vector is of length 175. A matrix `NEWSOURCES` is created. It will contain the predicted source values along the variable `MyDepth` at the 16 stations. We then start a loop, and in each iteration, data from one station are analysed using additive modelling. A new data frame is created with depth values between the lowest and highest measured depths (with steps of 25 m). Source values along these depth ranges are predicted and stored at the appropriate place in the matrix `NEWSOURCES`. Once this process is carried out for each station, this matrix contains predicted source values at the same depths. The only remaining problem is that `NEWSOURCES` has many missing values as some profiles were measured at deeper depths at some stations, or the other way around, at the less deep stations. The option `use = "pairwise.complete.obs"` ensures that the correlation matrix between the 16 profiles does not contain missing values (unless the depth ranges between the two stations were completely different as originally happened when station 10 was included). The rest of the code is identical as above and produces the dendrogram in Fig. 17.5. Using a degree of subjectivity, we can distinguish the following groups: (i) stations 1, 2, 3, and 12, (ii) stations 13, 14, and 15, (iii) stations 6, 7, 8, and 9, (iv) stations 11 and 19, and (v) stations 16, 17, and 18. Inspection shows that groups (i) and (ii) are all spring (March and April) samples. Group (v) comprises the three stations over the Porcupine Abyssal plain from October 2002. Groups (iii) and (iv) are all autumn samples from August and October within the Porcupine Seabight.

In the next section, we apply GAM models with one smoother per group.

17.4 Estimating Common Patterns Using Additive Mixed Modelling

In the previous section, several potential models were discussed. It is clear that the source–depth relationship is non-linear and we should take into account heterogeneity between and within the stations. There is also the possibility of violation of independence. This may come as a surprise, but recall in Chapters 6 and 7 we checked for temporal correlation in the data. We don't have repeated measurements in time, but we do have them along depth! The depth gradient can be seen as a spatial gradient, and this means that we may need to add a spatial (depth) correlation structure to the model.

In the previous section, we mentioned that numerical problems may be expected if 16 smoothing curves are used (one for each station). An initial analysis confirmed this problem. We therefore need to reduce the number of smoothing curves and we consider the following options.

- Use one smoothing curve for all stations.
- Use one smoothing curve for each month (four smoothers in total).
- Use one smoothing curve for each group derived from Fig. 17.4 (six smoothers in total).
- Use one smoothing group for each group derived from Fig. 17.5 (five smoothers in total).

We will set the scene with the first option and then discuss how to proceed with the other models and then judge which approach is the best.

17.4.1 One Smoothing Curve for All Stations

Instead of applying the additive mixed model in Equation (17.2), we start with a simpler model to show why we need a more complex one. Obviously, you can argue that the data exploration already indicated that we need to allow for heterogeneity, but it is always worth while formally showing why a more complex approach is required.

$$S_{is} = \alpha_i + f(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma^2) \qquad (17.3)$$

The model has one smoothing curve for all stations, a month effect (nominal variable), and a station effect (nominal variable, represented by α_i). The residuals are assumed to be independently, normally distributed with the same variance. The estimated smoothing curve is presented in Fig. 17.6A, and the residuals against fitted values in Fig. 17.6B. The latter graph confirms our suspicions; there is heterogeneity. So in spite of all the terms in the model being significant, we can bin it.

The following R code was used.

```
> library(mgcv); library(nlme)
> M1 <- gam(Sources ~ fStation + s(SampleDepth) +
            fMonth, data = ISIT2)
> E <- resid(M1)
> F <- fitted(M1)
> op <- par(mfrow = c(2, 1), mar = c(5, 4, 1, 1))
> plot(M1)
> plot(F, E, xlab = "Fitted values", ylab = "Residuals")
> par(op)
```

The mar option in par modifies the white space around the graphs. The other R code has been discussed elsewhere.

To work towards a model that can cope with heterogeneity, we consider the following series of models that increase in complexity.

$$S_{is} = \alpha + a_i + f(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma^2) \tag{17.4A}$$

$$S_{is} = \alpha + a_i + f(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma_i^2) \tag{17.4B}$$

$$S_{is} = \alpha + a_i + f(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma^2|\text{Depth}_s|^{\delta}) \tag{17.4C}$$

$$S_{is} = \alpha + a_i + f(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma_i^2|\text{Depth}_s|^{\delta}) \tag{17.4D}$$

$$S_{is} = \alpha + a_i + f(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma_i^2|\text{Depth}_s|^{\delta_i}) \tag{17.4E}$$

Instead of using a fixed intercept, we decided to use a random intercept. This is the equivalent of a random intercept mixed model (Chapter 4). So, in all mod-

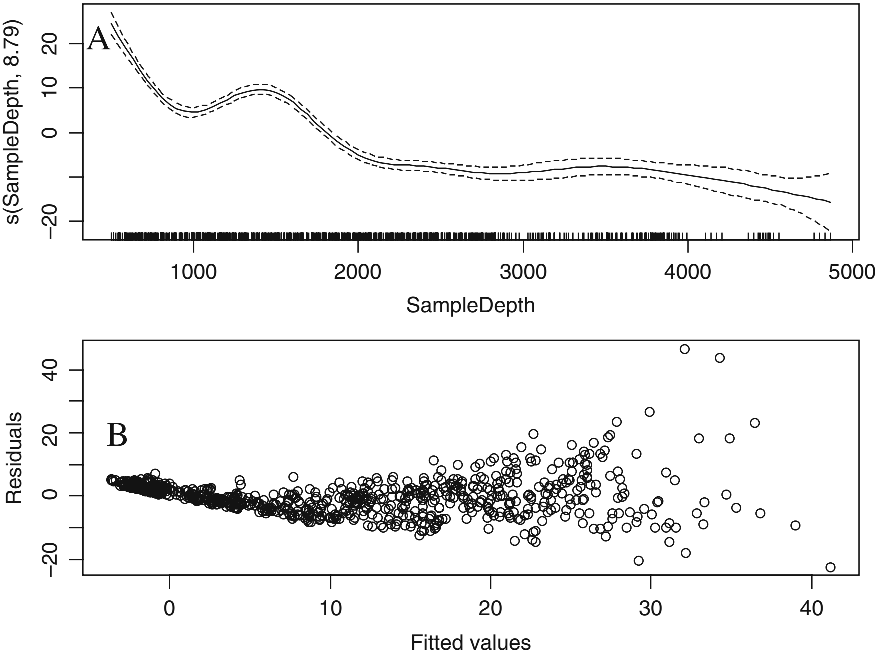

Fig. 17.6 **A**: Estimated smoothing curve for the additive model in Equation (17.3). **B**: Residuals versus fitted values showing heterogeneity. Cross-validation was used to estimate the degrees of freedom

els we assume that a_i is normally distributed with mean 0 and variance ${\sigma_a}^2$. The advantage of this approach is that instead of estimating 16 intercepts, we now only need to estimate one (α) and a variance term ${\sigma_a}^2$. The variance component a_i allows for random variation around the intercept. The model in Equation (17.4A) assumes homogeneity, and it was only added to provide a reference point. The model in Equation (17.4B) assumes heterogeneity per station but homogeneity within a station along depth, that in Equation (17.4C) assumes homogeneity between stations but heterogeneity within a station along depth (but the strength of the heterogeneity along the depth gradient is the same for each station), that in Equation (17.4D) allows for heterogeneity between stations and within stations along depth (same strength), and finally, the model in Equation (17.4E) implies heterogeneity between stations and heterogeneity within stations along depth. The crucial point in Equation (17.4E) is that the heterogeneity within stations along depth is allowed to differ between the stations. Hence, it is the most complete (and complicated) model in this set of models. It should be noted that the only difference between these five models are the random components. In the models, we apply later in this chapter, we use the same five random components. We refer to them as models A to E. The only difference between the models in Equations (17.4A–E) and the ones used later is the fixed effects structure (smoothers).

The following code applies models in Equations (17.4A–E) in R and compares them using the AIC and BIC criteria.[1] It was observed that the numerical algorithms performed better when we rescaled the depth so that values were between 0.5 and 5 (km) instead 500–5,000 m:

```
> lmc <- lmeControl(niterEM = 5000, msMaxIter = 1000)
> M17.4A <- gamm(f1, random = list(fStation =~ 1),
     method = "REML", control = lmc, data = ISIT2)
> M17.4B <- gamm(f1, random = list(fStation =~ 1),
     method = "REML", control = lmc, data = ISIT2,
     weights = varIdent(form =~ 1 | fStation))
> M17.4C <- gamm(f1, random = list(fStation =~ 1),
     method = "REML", control = lmc, data = ISIT2,
     weights = varPower(form =~ Depth1000))
> M17.4D <- gamm(f1, random = list(fStation =~ 1),
     method = "REML", control = lmc, data = ISIT2,
     weights = varComb(varIdent(form =~ 1 | fStation),
                       varPower(form =~ Depth1000)))
> M17.E <- gamm(f1, random=list(fStation =~ 1),
     method = "REML",control = lmc, data = ISIT2,
     weights = varComb(varIdent(form =~ 1 | fStation),
               varPower(form =~ Depth1000 | fStation)))
```

[1] We used R version 2.6 and mgcv version 1.3–27. More recent versions of R and mgcv require a small modification to the code; see the book website (www.highstat.com) for updated code.

```
> AIC(M17.4A$lme, M17.4B$lme, M17.4C$lme, M17.4D$lme,
      M17.4E$lme)
            df      AIC
M17.4A$lme   8 4734.141
M17.4B$lme  23 4269.503
M17.4C$lme   9 4258.752
M17.4D$lme  24 3859.231
M17.4E$lme  39 3675.986
```

The only difference between the calls to the gamm function for these five models is the weights option. See Chapter 4 for a more detailed discussion. The names of the R objects correspond to the equation numbers on the previous page. The output of the AIC command shows that the model with heterogeneity between stations and within stations is the best model (from these five!). This is model E. The estimated smoothing curve and (normalized) residuals versus fitted values are given in Fig. 17.7. Note that all the hard work earlier did help to solve heterogeneity problems! The R code that was used to create Fig. 17.7 is similar to that for Fig. 17.6 and is not given here.

There is one thing we have ignored so far and that is spatial dependence. There are two ways we can violate the independence assumption: correlation between stations and/or correlation within (groups of) stations along the depth gradient. The first form of dependence is difficult to model within the random component structure, and it is easier to use covariates for this. We could for example use more smoothers or other explanatory variables. The second form of dependence can be checked by making a variogram (Fig. 17.8) with the following two lines of R code:

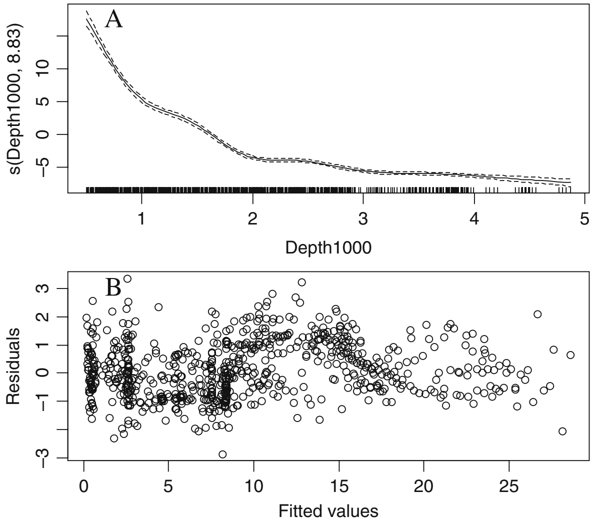

Fig. 17.7 Estimated smoothing curve (**A**) and normalized residuals versus fitted values (**B**) for the model in Equation (17.4E). Compare panel B with that of Fig. 17.6 and note how all the fancy random structures have solved the heterogeneity problem

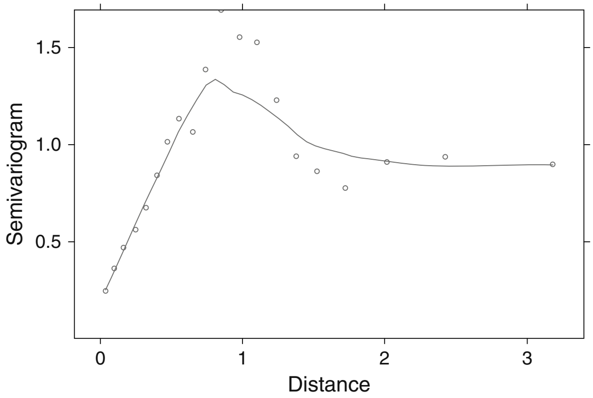

Fig. 17.8 Variogram of the normalized residuals obtained by model 4E. The spatial correlation structure is estimated within the profiles

```
> Vario17.4E <- Variogram(M17.4E$lme, robust = TRUE,
               data = ISIT2 form =~ Depth1000 | fStation)
> plot(Vario17.4E)
```

Independence of residuals expresses itself in the variogram as a horizontal band of points. In this case, the variogram shows a sharp increase during the first 1,000 m (1 km) and a small decrease thereafter.

Thus the model implies that residuals that are within a range of 1,000 m are correlated. We specifically wrote ‘the model implies’ as the most likely explanation is that the dependence in caused by an improper fixed effects structure (meaning: not enough smoothers or missing covariates).

One option is to include a correlation structure along depth within the additive modelling structure, but a better approach (to start with) is to extend the model with more smoothers (or covariates) and see whether that solves the problem.

This is done next. If it turns out that adding more smoothers or covariates does not solve the problem, then we should consider adding a correlation on the residuals within the additive mixed model. But that is a last resort.

17.4.2 Four Smoothers; One for Each Month

To solve the independence problem discussed in the previous paragraph, we extend the fixed effects part of the model by using one smoother for all stations of the same month. Just as before, we have to take into account possible violation of heterogeneity, and therefore, we consider models with similar random error structures as before. Little is lost (as can be judged by plotting residuals versus fitted values) by using different variances per month instead of station; it saves considerable computing time!

$$S_{is} = \alpha + a_i + f_j(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \quad \varepsilon_{is} \sim N(0, \sigma^2) \tag{17.5A}$$

$$S_{is} = \alpha + a_i + f_j(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \quad \varepsilon_{is} \sim N(0, \sigma_j^2) \tag{17.5B}$$

$$S_{is} = \alpha + a_i + f_j(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \quad \varepsilon_{is} \sim N(0, \sigma^2 \times |\text{Depth}_s|^{\delta}) \tag{17.5C}$$

$$S_{is} = \alpha + a_i + f_j(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \quad \varepsilon_{is} \sim N(0, \sigma_j^2 \times |\text{Depth}_s|^{\delta}) \tag{17.5D}$$

$$S_{is} = \alpha + a_i + f_j(\text{Depth}_s) + \text{Month}_i + \varepsilon_{is} \quad \varepsilon_{is} \sim N(0, \sigma_j^2 \times |\text{Depth}_s|^{\delta_j}) \tag{17.5E}$$

The differences between the models in Equations (17.4A–E) and (17.5A–E) is the index *j* attached to the smoothing function *f* and the multiple variances per month instead of station. The index *j* take the values $j = 3$, 4, 8, and 10 referring to the four months. Hence, each month is allowed to have a different depth-source profile. The following R code implements the model in Equations (17.5A), (17.5B), and (17.5E).

```
> f1 <- formula(Sources ~
          s(Depth1000, by = as.numeric(Month == 3)) +
          s(Depth1000, by = as.numeric(Month == 4)) +
          s(Depth1000, by = as.numeric(Month == 8)) +
          s(Depth1000, by = as.numeric(Month == 10)) +
          fMonth)
> M17.5A <- gamm(f1,random = list(fStation =~ 1),
      method = "REML", control = lmc, data = ISIT2)
> M17.5B <- gamm(f1, random = list(fStation =~ 1),
      method = "REML", control = lmc, data = ISIT2,
      weights = varIdent(form =~ 1 | fMonth))
> #....
> M17.4E <- gamm(f1, random = list(fStation =~ 1),
      data = ISIT2, method = "REML", control = lmc,
      weights = varComb(varIdent(form =~ 1 | fStation),
              varPower(form =~ Depth1000 | fStation)))
```

The other models can be implemented in the same way as before, and to save space, the R code is not shown here (it can also be found on the book website). Just as before, the AIC indicated that model E is the best. The estimated smoothing curves per month are given in Fig. 17.9. Note that we can see a clear distinction between the shapes in different months.

As part of the model validation, we also need to plot residuals versus fitted values to assess homogeneity (not shown here) and residuals versus each explanatory variable to asses independence. The plot of the (normalized) residuals versus depth (Fig. 17.10) shows that there is a problem as there are clear residual patterns. To aid visual interpretation, we added a LOESS smoother. It may also be useful to fit an

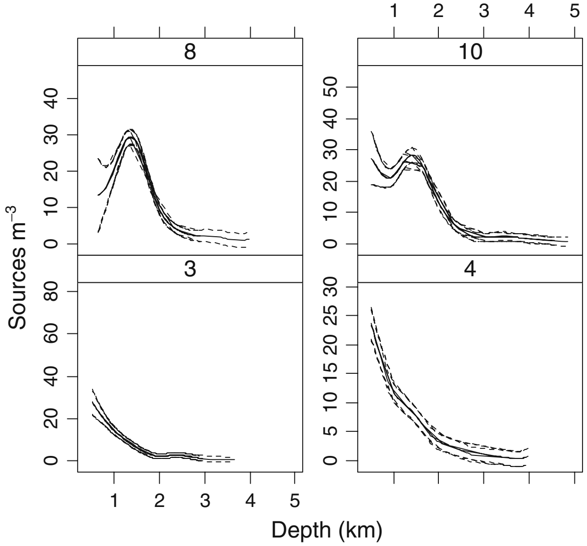

Fig. 17.9 Estimated smoothing curves and 95% point-wise confidence bands per month for model 5E. The four panels correspond to the four months. Month 3 represents stations 12–15, month 4 cruses 1–4, month 8 stations 6, 7, 8 and 11, and month 10 represents stations 16–19. The R code for this figure is presented on the book website

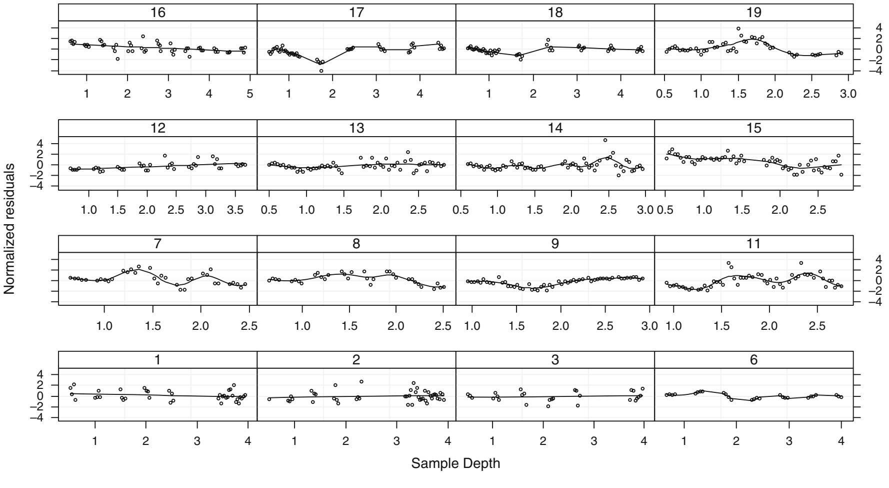

Fig. 17.10 Normalised residuals plotted versus depth for model (17.5E). Note that for some stations there is a clear residual pattern. To aid visual interpretation, LOESS curves were added. The R code for this figure is presented on the book website

additive model in each panel in Fig. 17.10 and inspect the significance levels of the smoothers. As we are fitting a smoother on residuals (as a function of depth), we should not see significant smoothers! But for stations 15 (month 3), 7, 9 and 11 (all from month 8), and 17 and 18 (month 10) we could still find a strong and significant relationship between residuals and depth.

The dependence problem is also detected if we make a variogram of the normalised residuals. It has a similar shape as in Fig. 17.8.

One option to solve this problem is to include a spatial correlation structure within the additive model, which is achieved by adding the correlation option to the gamm function:

```
correlation = corSper(form =~ Depth1000 | fStation,
           nugget = TRUE, fixed = FALSE).
```

But just as before, the residual pattern indicates that the grouping structure by months is not optimal for all stations. So, instead of adding a complicated spatial correlation structure, we should first aim to improve the fixed effects structure. This means that we have to use more covariates or a different grouping of stations.

So, to summarise this part, the fixed effect part of the model was extended from one smoother to four (one per month). Stations 12, 13, 14, and 15 are from month 3; stations 1, 2, and 3 from month 4; stations 6, 7, 8, 9, and 11 from month 8; and stations 16–19 from month 10. But the model validation showed that especially within month 8, stations are not similar. But for month 4 (stations 1–3) and month 3 (especially stations 12–14) profiles are similar! To gain further insight, we continue with a grouping structure by geographical distances.

17.4.3 Smoothing Curves for Groups Based on Geographical Distances

In Section 17.2, we discussed how to divide the 16 stations in 5 groups based on geographical distances. Our proposed grouping of stations was (i) stations 1, 2, and 3; (ii) stations 6 and 9; (iii) stations 7, 8, and 11; (iv) stations 12, 13, 14, 15, and 19; (v) stations 16 and 17; and (vi) station 18. The R code below runs the same 5 models as in Equations (17.4) and (17.5), except that the fixed structure is adjusted to take into account our new groups.

```
> G1 <- ISIT2$Station == 1 | ISIT2$Station == 2 |
        ISIT2$Station == 3
> G2 <- ISIT2$Station == 6 | ISIT2$Station == 9
> G3 <- ISIT2$Station == 7 | ISIT2$Station == 8 |
        ISIT2$Station == 11
> G4 <- ISIT2$Station == 12 | ISIT2$Station == 13 |
        ISIT2$Station == 14 | ISIT2$Station == 15 |
        ISIT2$Station == 19
> G5 <- ISIT2$Station == 16 | ISIT2$Station == 17
> G6 <- ISIT2$Station == 18
> f1 <- formula(Sources~
         s(Depth1000, by = as.numeric(G1)) +
         s(Depth1000, by = as.numeric(G2)) +
         s(Depth1000, by = as.numeric(G3)) +
         s(Depth1000, by = as.numeric(G4)) +
         s(Depth1000, by = as.numeric(G5)) +
         s(Depth1000, by = as.numeric(G5)) + fMonth)
> M.GeoA <- gamm(f1,random = list(fStation =~ 1),
     method = "REML", control = lmc, data = ISIT2)
> M.GeoB <- gamm(f1, random = list(fStation =~ 1),
     method = "REML", control = lmc, data = ISIT2,
     weights = varIdent(form =~ 1 | fMonth))
> # ...
> M.GeoE<-gamm(f1, random=list(fStation =~ 1),
     data = ISIT2, method = "REML", control = lmc,
     weights = varComb(varIdent(form =~ 1 | fMonth),
                varPower(form =~ Depth1000 | fMonth)))
```

The other models can be run with similar code. The model with heterogeneity between groups and heterogeneity along depth (with differences per group) is the best, as judged by the AIC. This is model E. Just as in the previous analysis, we made a variogram of the (normalised) residuals, and we also plotted (normalised) residuals versus depth. These graphs are not shown here, but both indicated violation of independence. Hence, grouping stations based on geographical distances does not give groups in which the profiles have similar depth profiles. We also tried small modifications of the grouping structure, but this did not solve the independence problem.

17.4.4 Smoothing Curves for Groups Based on Source Correlations

In Section 17.3, we also discussed how to calculate correlations between predicted source profiles and used these to determine a grouping structure. Recall that we determined the following five groups: (i) stations 1, 2, 3, and 12; (ii) stations 6, 7, 8, and 9; (iii) stations 11 and 19; (iv) stations 13, 14, and 15; and (v) stations 16, 17, and 18. Adjusting the R code in order to implement this grouping of stations is relatively simple. All we need is the following piece of code:

```
> G1 <- ISIT2$Station == 1 | ISIT2$Station == 2 |
        ISIT2$Station == 3 | ISIT2$Station == 12
> G2 <- ISIT2$Station == 6 | ISIT2$Station == 7 |
        ISIT2$Station == 8 | ISIT2$Station == 9
> G3 <- ISIT2$Station == 11 | ISIT2$Station == 19
> G4 <- ISIT2$Station == 13 | ISIT2$Station == 14 |
        ISIT2$Station == 15
> G5 <- ISIT2$Station == 16 | ISIT2$Station == 17 |
        ISIT2$Station == 18
```

```
> f1 <- formula(Sources ~
             s(Depth1000, by = as.numeric(G1)) +
             s(Depth1000, by = as.numeric(G2)) +
             s(Depth1000, by = as.numeric(G3)) +
              s(Depth1000, by = as.numeric(G4)) +
              s(Depth1000, by = as.numeric(G5)) + fMonth)
> M.cor4A <- gamm(f1, random = list(fStation =~ 1),
      method = "REML", control = lmc, data = ISIT2)
> # etc...
```

Other models can be fitted by using the same code as above. Again, the AIC indicated that model E is the best. We plotted the residuals versus depth and a large number of stations contained a significant (as determined by a smoother) residual–depth pattern, especially stations 12, 7–9 (entire group 2), 11, 15, 16, and 17. This means that the chosen grouping is not a good one.

17.5 Choosing the Best Model

In the previous section, we grouped stations by month, geographical distances and based on correlations between predicted source values. None of the approaches produced a grouping of stations in which residual patterns did not show any violation of independence. Well, this is not entirely true. All analysis showed that the source–depth relationship for stations 1, 2, and 3 are similar, and the same holds for stations 12, 13, and 14. So, at least we can identify these two groups. The other profiles, however, cannot be grouped so easily. In a final attempt, we decided to fit an additive mixed model with two groups of stations (1, 2, 3 and 12, 13, 14), and we used one smoother for each group. Hence, the fixed effects structure assumes that (i) stations 1, 2, and 3 have the same source–depth relationship, (ii) stations 12, 13, and 14 have the same source–depth relationship, and (iii) all other stations have different source–depth relationships. Furthermore, we used one smoother for each of the other stations. The same five random error structures described in Equations (17.4) and (17.5) were used. Some of these models, especially E, are highly complicated, and may potentially not converge. To reduce computing time, we set the degrees of freedom for each smoother to 4, and to our surprise all five models converged.

In terms of the random structure, the model that contains all the options (heterogeneity between groups, along depth but not for all groups), model E, was the best as judged by the AIC. However, the BIC indicated model C (heterogeneity along depth).

The good news is that the variogram of the normalised residuals of model C did not show a clear violation of independence. Figure 17.11 shows the estimated smoothing curve for each station. Note that the smoothers for stations 1, 2, and 3 are identical, and the same holds for those of stations 12, 13, and 14. Based on this graph, the analysis can be taken further by grouping stations 7, 8, and 19 to see whether that improves the model. The way to proceed is to (i) adopt

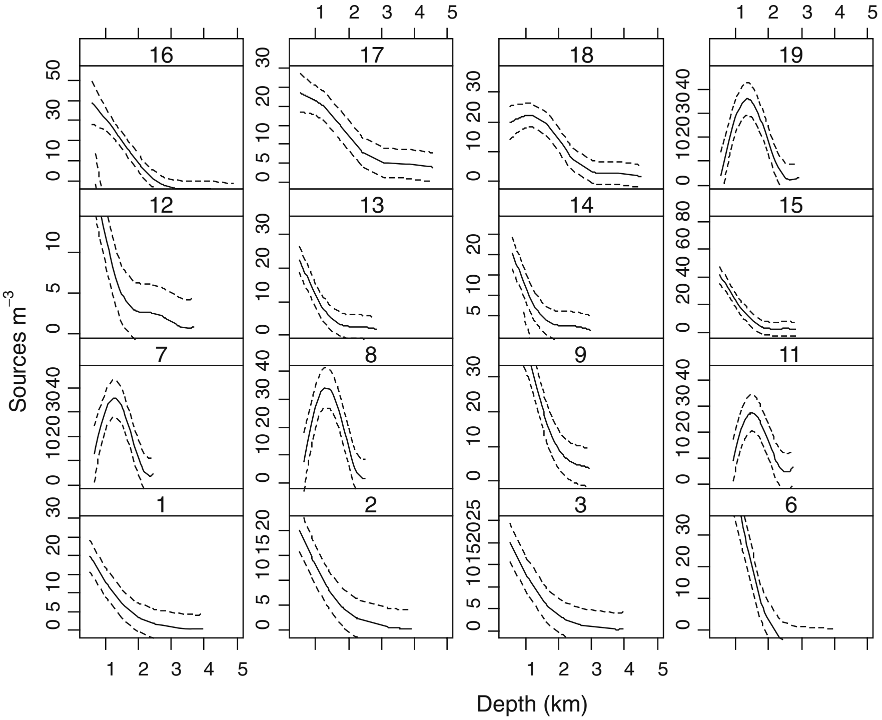

Fig. 17.11 Estimated smoothing curves for each station obtained by the model with 12 groups

the variance structure of model C, (ii) switch to maximum likelihood estimation (`method` = `"ML"`), and (iii) compare models in terms of the smoothers. We leave this as an exercise to the reader, but initial analyses indicated that there is not much to be gained by further grouping the stations.

17.6 Discussion

The data were originally published using (i) a logarithmic ($\log_{10}$) transformation on the sources, (ii) a random effect for station, and (iii) one smoother; see Gillibrand et al. (2007). The transformation solved a lot of trouble; the AIC still identified as optimal the most complicated model E, whilst the BIC indicated model C (only heterogeneity along depth). Even a visual inspection of the residuals of model A did not show any clear heterogeneity! Hence, a logarithmic transformation makes life much easier! However, there is still violation of independence; so we need to add more smoothers or covariates.

However, this is not a trivial exercise, as we showed in this chapter. The question is then: To transform or not to transform? Our opinion is that working with the original (untransformed) data gives more information on what is going on. We did not even discuss the numerical output for the optimal models; the heterogeneity parameters give a wealth of information as well. In our view, only when numerical instability of the estimation routines becomes an issue, transforming and/or standardising an option.

In Chapter 4, we mentioned that the model selection should follow a top-down approach by starting with a fixed effects structure that contains all explanatory variables and possible interaction terms. We did not follow that approach here, simply because the full model had numerical problems. So, we started by grouping stations and trying to find the optimal grouping structure. But the price we paid for this is that from the beginning, we were facing violation of independence and only when we used a close-to-optimal model, the problem became less serious.

So, what does this exercise tell us?

Firstly the close similarity and clustering of stations 1, 2, and 3 indicates that reproducibility of results is good, and there is probably no need to expend sampling effort in replicate profiles. The ISIT system has since been adapted to fit onto a standard oceanographic CTD (Conductivity, temperature, and depth) profiler (Heger et al., 2008).

Secondly, it is evident that there is a seasonal change in profiles between spring and autumn with a post-summer peak in abundance of bioluminescent sources at about 1,200 m. Simple mathematical curves such as the exponential relationship proposed by Bradner et al. (1987) are clearly inappropriate. The estimation of smoothing curves (Fig. 17.10) is very useful for the biologist since it provides an objective means of combining sets of data and producing estimates of the depth of the peak and mean number of sources m^{-3} at different depths.

Thirdly, contrary to what was stated by Gillibrand et al. (2007), there is a difference between stations in the Porcupine Seabight compared with those offshore over the Porcupine Abyssal Plain. Examining the panels in Fig. 2.11, it seems the autumn peak below 1,000 m is less strong in the offshore stations (16, 17 and 18) than closer inshore (19, 6, 7, 8, 9).

The reasons for the deep bioluminescent layer is unclear, but is probably related to two effects. The peak almost certainly represents a seasonal increase in deep biomass fed by organic matter flux from the spring bloom in surface waters. This effect is probably accentuated by accumulation in a layer of North Atlantic intermediate water at this depth, which is derived from Mediterranean water moving northwards from Gibraltar. This effect may be stronger further inshore, where there is a northward moving shelf edge current.

17.7 What to Write in a Paper

Within the field of bioluminescent research, Gillibrand et al. (2007) and this case study are two of the first texts where advanced statistical methods have been used. If you are submitting a paper in a subject area where additive mixed modelling techniques are uncommon, you will face the daunting task of convincing an entire group of scientists of the need for complicated statistical methods. The best starting point is Fig. 2.11 as it clearly shows that linear regression methods (or ANCOVA) are unsuitable. It may be an option to discuss the heterogeneity problem by showing only Fig. 17.6B. At that point, you will need to discuss why you did not apply a logarithmic transformation. Predicting values on the original scale may be a valid argument and so is the fact that a transformation changes relationships between sources and depth. In order to make the referee (and reader) of your paper happy, a non-technical explanation of additive mixed modelling and especially the variance structures is required. If you fail to do this, they will come back with the question: Why do you need all this complicated modelling?

In approach 3 (Subsection 17.3.3) we used the data to calculate Pearson correlation coefficients and applied clustering on them. We then used the same data in the GAMMs (using the results from the clustering). This approach is likely to receive (valid) criticism!

Acknowledgments We thank Dr Emma Gillibrand for permission to use data from her PhD thesis.