Chapter 15
Large-Scale Impacts of Land-Use Change in a Scottish Farming Catchment

A.F. Zuur, D. Raffaelli, A.A. Saveliev, N.J. Walker, E.N. Ieno, and G.M. Smith

15.1 Introduction

A catchment is an area of land defined by the origins and discharges of all tributary streams feeding large rivers flowing into the sea. It is therefore a natural bio-physical unit distinct from adjacent catchments and forms the obvious basis for integrated environmental management policies. In Europe, river catchments tend to be dominated by agriculture, at least at lower altitudes. In the case of the Ythan catchment (Fig. 15.1), Aberdeenshire, Scotland, where the river rises at only a few hundred metres, more than 90% of the land area is now under agricultural production. Much of this is arable crops like wheat, barley, and oil-seed rape, which demand high inputs of chemical nitrogen. The Ythan catchment also hosts large numbers of pigs and other livestock (and also some of the authors of this book).

Whilst the Ythan catchment has always been prime agricultural land, there have been major changes in land-use over the past 40 years because of market trends and drivers such as the Common Agriculture Policy. This policy encouraged growing of crops through subsidies not previously available for crops such as barley and wheat at the expense of less profitable crops such as oats. The conversion of grassland to cereals, increased application of nitrogen, and increase in animal manures and slurries over the past 40 years have inevitably affected water quality, specifically elevated levels of nitrate. These levels were so high in the 1990s that the Ythan catchment had the distinction of being the first in the UK to be designated a Nitrogen Vulnerable Zone under the European Community Nitrates Directive.

Staff at Culterty Field Station, University of Aberdeen, were able to document and describe trends in this process in great detail through a series of monitoring programmes, data analyses, and field experiments. Data on land-use were obtained from 'parish returns' – records of amounts of land under different crops and numbers of animals held for each farm in the parish that are returned to the Scottish Records Office annually. These data were extracted for all parishes (community

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_15,

Fig. 15.1 Small part of the Ythan estuary. The photograph was taken by Alain Zuur

administrative areas) within the catchment for land under oats, wheat, oil-seed rape, barley, and for numbers of pigs, cattle and sheep for the period 1960s to 1990s. Levels of nitrates (only small amounts can be attributed to sewage) were extracted from databases held by the North-East River Purification Board and supplemented by the field station's own observations. The environmental impact of high levels of nitrates is expected to be seen as blooms of algae in rivers and estuaries, where they form extensive green mats that strip the oxygen from the underlying mudflats, reducing the invertebrates available to feeding shorebirds.

Counts of shorebirds have been made every month for the past 40 years by staff and students at the field station and most of these data make up the database held by the British trust for Ornithology for this estuary. Mean counts for the winter months November–February were calculated for the most abundant waders: oystercatcher *Haematopus ostralegus*, redshank *Tringa totanus*, dunlin *Calidris alpina*, knot *Calidris canutus*, turnstone *Arenaria interpres*, bar-tailed godwit *Limosa lapponica* and curlew *Numenius arquata*.

Using these data, we were able to trace possible connections from agricultural policy and land-use change through to ecological impacts on species of high conservation importance, the shorebirds. The data sets are interesting because they are typical of those available for detecting historical trends in variables that may be linked to a current environmental impact. The time series is unusually long for ecological data, but the data were not originally collected with this specific analysis in mind (linking agriculture change with shorebird numbers) and they are imperfect in many respects, as we shall see. All too often the ecologist has to work with whatever data are available rather than what would be ideal. Unfortunately, it is impossible to collect data retrospectively, unless one has a time machine.

We analysed these data to try and answer the following questions:

1. Are there any trends in the bird time series?
2. Is there a simple and obvious relationship between agricultural change and shorebird numbers?

3. Is the relationship different for different species of shorebird?
4. Which aspects of land-use best account for changes in water quality and therefore need to be targeted for restoration programmes?

To answer these questions, we split the analysis into three stages. In the first stage, we applied a data exploration, and in the second stage, we focussed on the first question: are there any trends in the bird time series? We used GAMs (Chapters 3 and 6) for this. The reason we used smoothing techniques will become clear after we have looked at the data exploration. The 'mixed' bit is needed because the data are time series and, as always in ecology, there is heterogeneity. In the last step of the analysis, we included the information on land-use and agricultural changes.

15.2 Data Exploration

The dataset consists of average winter values for seven bird species (Oystercatcher, Turnstone, Curlew, Bar-tailed Godwit, Redshank, Knot, and Dunlin) and seven potential explanatory variables (wheat, barley, oats, cattle, sheep, pigs, and nitrate). The best way to visualise a dataset with up to 20–25 time series is a multiple panel graph made with the `xyplot` function from the lattice package (Fig. 15.2). The data file contains the variables in columns and the years in rows. To create the multiple panel graph for these data, we need to create three columns. The first column should contain all the variables we wish to plot along the vertical axes. Because there are 14 variables, the second column should contain 14 times year (we need

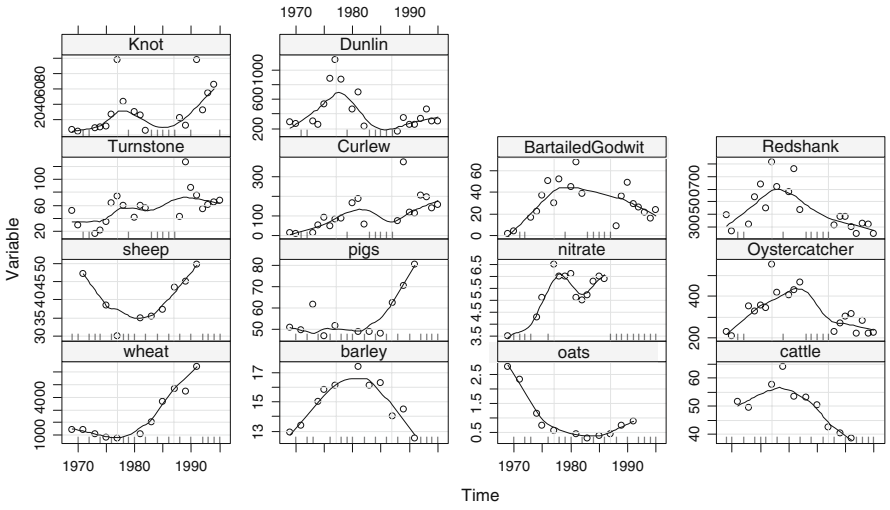

Fig. 15.2 Plot of the seven bird species and the potential explanatory variables. A LOESS smoother (with default amount of smoothing) was added to enhance visual interpretation

to concatenate Year 14 times). Finally, ID14 is the variable that tells the xyplot function which elements belong to the same variable. In the R code below, we used the levels option to ensure that the xyplot function places the panels of the birds next to each other.

```
> data(AED); data(Ythan); library(lattice)
> Birds <- as.vector(as.matrix(Ythan[, 2:8]))
> X <- as.vector(as.matrix(Ythan[, 9:15]))
> YX14 <- c(Birds, X)
> Year14 <- rep(Ythan$Year, 14)
> N <- length(Ythan$Year)
> ID14 <- factor(rep(names(Ythan[,2:15]), each = N),
    levels = c("wheat", "barley", "oats", "cattle",
        "sheep", "pigs", "nitrate", "Oystercatcher",
        "Turnstone", "Curlew", "BartailedGodwit",
        "Redshank", "Knot", "Dunlin"))
```

The code below produces Fig. 15.2.

```
> xyplot(YX14 ~ Year14 | ID14, xlab = "Time",
   ylab = "Variable", layout = c(4, 4),
   scales = list(alternating = TRUE,
                 x = list(relation = "same"),
                 y = list(relation = "free")),
   panel = function(x, y){
     panel.xyplot(x, y, col = 1)
     panel.grid(h = -1, v = 2)
     panel.loess(x, y, col = 1, span = 0.5)
     I2 <- is.na(y)
     panel.text(x[I2], min(y, na.rm = TRUE), '|', cex = 0.5)})
```

The new bit of code is the panel.text function. It plots the symbol '|' wherever there is a missing value. Although it takes a lot of complicated R code to make the xyplot graph, the results are impressive. The panels for the explanatory variables indicate serious collinearity and a large number of missing values. Most bird time series seem to have a peak around 1980, and redshanks, oystercatcher, dunlin, and bar-tailed godwit seem to follow a similar pattern over time. Some similarity between this pattern and some of the explanatory variables can also be detected. Although more difficult to see, the different ranges of the y-axis for the bird panels indicate a potential problem (heterogeneity) if we analyse all birds simultaneously.

Heterogeneity by different bird species is to be expected and can be dealt with using (i) a data transformation, (ii) standardisation, or (iii) using the varIdent residual variance structure as discussed in Chapter 4. However, as we discussed in

Chapter 2, a data transformation will be avoided as much as possible. We will return to this point later.

The data exploration indicates that we can expect problems with homogeneity and that although there may be effects on the bird numbers related to the explanatory variables, due to the large number of missing values, it may be difficult to fit a model that contains both the bird numbers and the explanatory variables. In fact, if anything comes out of the analyses of these data, we should be very happy!

15.3 Estimation of Trends for the Bird Data

The shape of the trends for the birds in Fig. 15.2 suggests using a model of the form

$$\text{Birds}_{is} = \alpha_i + f_i(\text{Year}_s) + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma^2) \tag{15.1}$$

Birds_{is} is the average number of birds species i in the winter of year s, α_i and $f_i(Year_s)$ are the intercept and smoother for bird species I, respectively, and ε_{is} is normally distributed noise with mean 0 and variance σ^2. If all the birds follow the same pattern over time, we can drop the index i from the smoother $f_i(\text{Year}_s)$. However, the shape of the smoothers in Fig. 15.2 clearly indicates that this is not the case. The ranges of the vertical axes in the same figure are rather different and suggest using

$$\text{Birds}_{is} = \alpha_i + f_i(\text{Year}_s) + \varepsilon_{is} \qquad \varepsilon_{is} \sim N(0, \sigma_i^2) \tag{15.2}$$

The only difference with the previous formula is the index i attached to the variance; it allows for heterogeneity between bird species. It makes sense to allow for this form of heterogeneity as some bird species are only ever present in low numbers while other bird species are normally found in very large numbers. This is a common problem in ecology as species of interest often occur at very different levels of abundance leading to lower and higher variances.

The following R code sets up the data for the additive mixed model in Equation (15.2) with multiple smoothers.

```
> Birds7 <- as.vector(as.matrix(Ythan[, 2:8]))
> BirdNames <- c("Oystercatcher", "Turnstone",
                 "Curlew", "BartailedGodwit",
                 "Redshank", "Knot", "Dunlin")
> ID7 <- factor(rep(BirdNames, each = N),
                levels = BirdNames)
> Year7 <- rep(Ythan$Year, 7)
> Oyst.01 <- as.numeric(ID7 == "Oystercatcher")
> Turn.01 <- as.numeric(ID7 == "Turnstone")
> Curl.01 <- as.numeric(ID7 == "Curlew")
> Bart.01 <- as.numeric(ID7 == "BartailedGodwit")
```

```
> Reds.01 <- as.numeric(ID7 == "Redshank")
> Knot.01 <- as.numeric(ID7 == "Knot")
> Dunl.01 <- as.numeric(ID7 == "Dunlin")
> f7 <- formula(Birds7 ~ ID7 +
    s(Year7, by = Oyst.01, bs = "cr") +
    s(Year7, by = Turn.01, bs = "cr") +
    s(Year7, by = Curl.01, bs = "cr") +
    s(Year7, by = Bart.01, bs = "cr") +
    s(Year7, by = Reds.01, bs = "cr") +
    s(Year7, by = Knot.01, bs = "cr") +
    s(Year7, by = Dunl.01, bs = "cr"))
```

The vector `Birds7` contains all bird data in a long vector. We also need a vector `ID7` that tells R which part of `Birds7` belongs to a certain species (N is the number of years that sampling took place). And obviously, we also need to copy and paste the variable `Year` seven times. The variables `Oyst.01`, `Turn.01`, etc., are vectors consisting of zeros and ones. For example, an element of `Oyst.01` is equal to 1 if the corresponding observation is an oystercatcher. These can be used in a GAM together with the `by` option to model interaction between year and species identity (which is ID7). As a result, a GAM gives 7 smoothers, one for each species. To reduce computing time (in some of the model that will be used later), we decided to use a cubic regression spline (`bs = cr` in R). The GAM itself is implemented with the code.[1]

```
> library(mgcv); library(nlme)
> lmc <- lmeControl(niterEM = 5000,msMaxIter = 1000)
> M0 <- gamm(f7, control = lmc, method = "REML",
             weights = varIdent(form =~ 1 | ID7))
```

As you can see, it takes more effort to prepare the data than to do the actual GAM command. REML estimation is used because we first want to find the optimal random component (Chapter 4). The `weights = varIdent(form =~ 1 | ID7)` implements the different variances per species and the `by` command in the smoother ensures that we have one smoother for each bird species i. The option `control = lmc` was used to ensure convergence.

15.3.1 Model Validation

The first validation plot we should make is residuals (normalised) versus fitted values (Chapter 4). The normalised residuals are corrected for the different variances per species. We can either plot residuals and fitted values for all species in one graph

[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.

(and use for example seven different symbols or colours) or draw them in a multiple panel plot with the `xyplot` function. We will do both. The `nlme` package has some handy tools to plot residuals versus fitted values. Obviously, we can use

```
> E0 <- resid(M0$lme, type = "normalized")
> F0 <- fitted(M0$lme)
```

and then plot residuals `E0` versus fitted values `F0` using the `plot` command or the `xyplot` function, but R can do this much faster. The following three commands each plot (normalised) residuals versus fitted values or time (Fig. 15.3).

```
> plot(M0$lme, resid(., type = "n") ~ fitted(.),
       abline = 0, col = 1)
> plot(M0$lme, resid(., type = "n") ~ Year7,
       abline = 0, col = 1, xlab = "Year")
> plot(M0$lme, resid(., type = "n") ~ fitted(.) | ID7,
       abline = 0, col = 1,
       par.strip.text = list(cex = 0.75))
```

The problem with this code is that it actually uses the lattice package and draws fancy multipanel graphs, and therefore, the `par(mfrow = c(2, 2))` tool to plot multiple graphs on the same window does not work. So, how did we create Fig. 15.3? The answer is in Sarkar (2008): Store each graph in an object, and use the `print.trellis` command (see: `?print.trellis`) to place the panels on a grid.

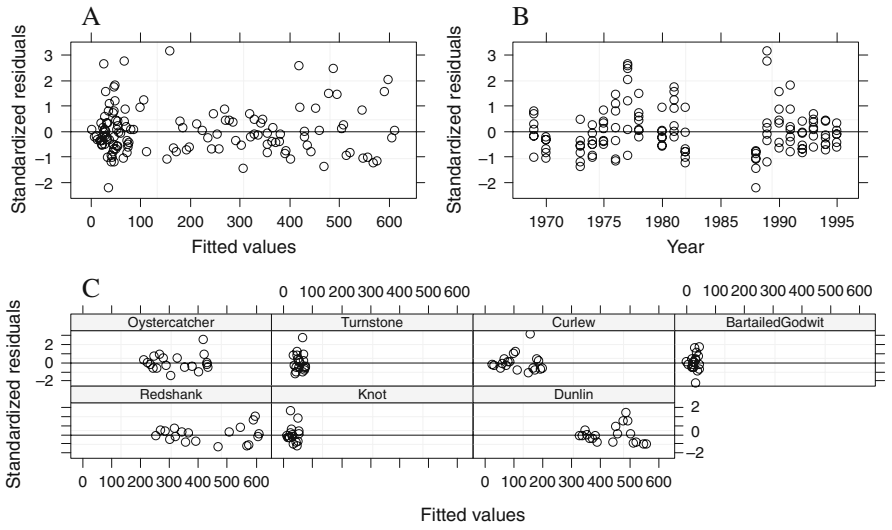

Fig. 15.3 Graphical validation of the model in Equation (15.2). **A**: Residuals versus fitted values. **B**: Residuals versus year. **C**: Residuals versus fitted values per species

```
> p1 <- plot(M0$lme, resid(., type = "n") ~ fitted(.),
            abline = 0, col = 1)
> p2 <- plot(M0$lme, resid(., type = "n") ~ Year7,
            abline = 0, col = 1, xlab = "Year")
> p3 <- plot(M0$lme, resid(., type = "n") ~ fitted(.) |
            ID7, abline = 0, col = 1,
            par.strip.text = list(cex = 0.75))
> print(p1, position = c(0, 0, 1, 1),
        split = c(1, 1, 2, 2), more = TRUE)
> print(p2, position = c(0, 0, 1, 1),
        split = c(2, 1, 2, 2), more = TRUE)
> print(p3, position = c(0, 0, 2, 1),
        split = c(1, 2, 2, 2), more = FALSE)
```

The `split` option in the `print` command tells R to divide the graphical window in a 2-by-2 grid (as determined by the last the numbers) and places each graph in a particular grid (as determined by the first two coordinates). Panel C is stretched over two grids because the `location` option specifies that `xmax = 2` (instead of 1). This is quite complicated R stuff (you could have done the same in Word with a table), but it can be handy to know. Sarkar (2008) is an excellent reference for lattice package.

What does it all tells us in terms of biology? Are we willing to assume homogeneity of variance based on Fig. 15.3A? We are hesitating a little bit as the residuals in the middle (between 100 and 400) seem to have slightly less spread. This could be a sample size issue as only a few birds have values in this range, see Fig. 15.3C. We can also argue that it looks homogeneous as by chance alone, 5% of the data can be outside the -2 to 2 interval. We also plotted residuals versus time (Fig. 15.4). Note there is an increase in residual spread for larger fitted values for some species (e.g. redshanks, curlew, and dunlin), but not for all! One option is to use a Poisson distribution, but because the data are winter *averages* and not counts, this is not the best option. Note that if we apply a generalised linear or additive model with a Poisson distribution, the average winter values are rounded to the nearest integer.

In Section 4.1, we introduced several approaches to model heterogeneity in a squid data set. The response variable was testis weight and the explanatory variable mantel length. In some months, variation in weight increased for larger length, but not in every month. We used the `varPower`, `varExp`, and `varConstPower` functions to allow for different spread along the variance covariate length per month. It seems we need a similar mechanism here to model the (potential) heterogeneity of variance. The only problem is that while we were able to use length as variance covariate for the squid data, here we do not have such a variable as the only available explanatory variables have many missing values. So instead we can use the fitted values as variance covariate. All that is needed is to adjust the `weights` option in the `gamm` function:

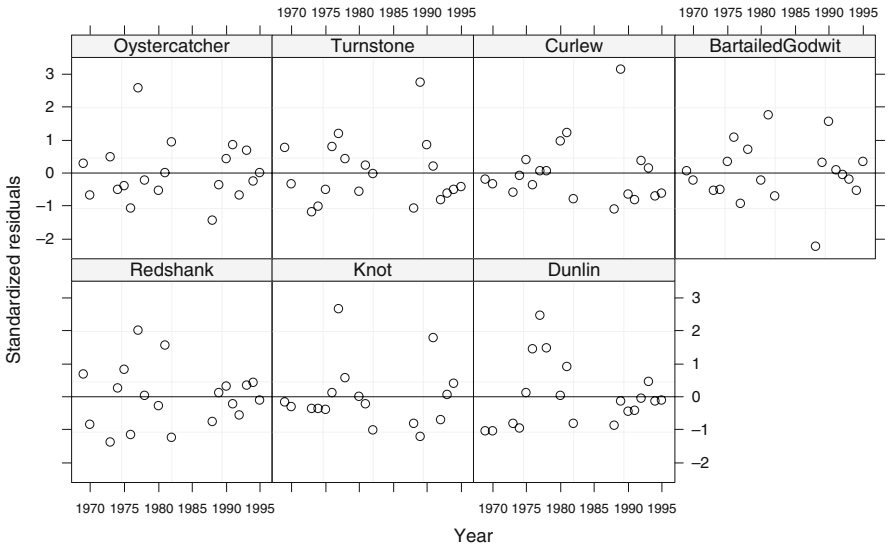

Fig. 15.4 Graphical validation of the model in Equation (15.2). Residuals versus time for each species

```
> M1<-gamm(f7, control = lmc, method = "REML",
      weights = varComb(varIdent(form = ~1 | ID7),
                varPower(form =~ fitted(.) | ID7)))
```

The variance structure creates the following additive mixed model.

$$\begin{aligned} \text{Birds}_{is} &= \alpha_i + f_i(\text{Year}_s) + \varepsilon_{is} \\ \varepsilon_{is} &\sim N(0, \sigma_i^2 |\hat{\alpha}_i + \hat{f}_i(\text{Year}_s)|^{\delta_i}) \end{aligned} \tag{15.3}$$

The variance structure for the noise ε_{is} looks rather complicated, but it is not. If δ_i is equal to 0 for all bird species i, we obtain exactly the same model as in Equation (15.2). The 'hats' above α and f_i indicate that these are estimates. If δ_i is larger than 0, the variance is proportional to the fitted values. So, this is a variance structure that allows for heterogeneity within a bird time series. The underlying principle reminds us of the Poisson distribution, where the mean equals the variance, but this *is* a normal distribution. As well as the within-bird-time series heterogeneity, we still allow for a different spread per bird using the index i attached to σ^2.

The problem with the model in Equation (15.3) is the lack of convergence. This comes as no surprise as the model contains 7 smoothers with cross-validation applied on each smoother, 7 variances ${\sigma_i}^2$, and 7 δ_is. And even more relevant, the data contains many gaps due to the missing values and time series are relatively short. We tried several options to deal with this and actually, all failed

(in terms of numerical convergence) or were considered not particularly helpful. However, we believe you can learn just as much from unsuccessful approaches as you can from successful ones. So, we now discuss some of approaches that failed.

15.3.2 Failed Approach 1

To reduce numerical computing complexity, we initially set the degrees of freedom for each smoother to 4, leaving it to decide later whether we need to increase or decrease this number. This requires modifying the `s` function:

```
s(Year7, by = Oyst.01, bs = "cr", fx = TRUE, k = 5).
```

This was done for each smoother in the model. However, this caused a new problem; the `gamm` function of the mgcv package needs either a random component or at least one smoother on which it can apply a cross-validation. Adding a random intercept has the advantage that it also allows us to automatically model the temporal correlation within the time series. For example, consider the following model:

$$\begin{aligned} &\text{Birds}_{is} = \alpha + f_i(\text{Year}_s) + a_i + \varepsilon_{is} \\ &\varepsilon_{is} \sim N(0, \sigma_i^2 \times |\hat{\alpha}_i + \hat{f}_i(\text{Year}_s)|^{\delta_i}) \\ &a_i \sim N(0, \sigma_a^2) \end{aligned} \tag{15.4}$$

Note that the intercept α no longer has an index i. Instead, there is now a random intercept a_i that is normally distributed with mean 0 and variance ${\sigma_a}^2$. This model is the smoothing equivalent of the random intercept mixed effects model discussed in Chapter 5. Recall that such a model induces the compound symmetry correlation on the time series. At this stage, it is useful to discuss this correlation structure. The model in Equation (15.4) is not fundamentally different from $\mathbf{Y}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \boldsymbol{\varepsilon}_i$, which is the hierarchical mixed model discussed in Chapter 5. Or perhaps we should write it as $\mathbf{Birds}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \boldsymbol{\varepsilon}_i$. The $\mathbf{X}_i \times \boldsymbol{\beta}$ component is the equivalent of the intercept and the smoothing curve, $\mathbf{Z}_i \times \mathbf{b}_i$ contains the random intercept and $\boldsymbol{\varepsilon}_i$ the residuals. We used a vector notation: $\mathbf{Birds}_i = (\text{Birds}_{i1}, \ldots, \text{Birds}_{i27})'$ which contains the bird data of species i for all years. Just as in Chapter 5, we can write that the marginal distribution for $\mathbf{Birds}_i$ is normally distributed with mean $\mathbf{X}_i \times \boldsymbol{\beta}$ and covariance $\mathbf{V}_i$. Equation (15.4) implies the following structure for $\mathbf{V}_i$.

$$\begin{pmatrix} \sigma_a^2 + \sigma_i^2|\hat{\alpha} + \hat{f}_i(\text{Year}_1)|^{\delta_i} & \sigma_a^2 & \cdots & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_i^2 \times |\hat{\alpha} + \hat{f}_i(\text{Year}_2)|^{\delta_i} & \cdots & \sigma_a^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_a^2 + \sigma_i^2 \times |\hat{\alpha} + \hat{f}_i(\text{Year}_{27})|^{\delta_i} \end{pmatrix}$$

This matrix has a dimension of 19 × 19 as each bird species was measured over 19 years (the other years contain missing values). The covariance between two observations of the same bird species i is ${\sigma_a}^2$, whatever the time lag between the two observations. The variance of bird species i depends on a bird-specific variance ${\sigma_i}^2$ and the fitted values in year s. (The parameter δ_i shows how strong the variance depends on the fitted values for species i.) The problem with this model is that computing time on an average computer is about 15 min and convergence problems arise. So we have to ask whether we are really interested in modelling heterogeneity between bird species as well as within each time series. The main underlying questions are related to effects of agricultural use on birds. Therefore, the first form of heterogeneity may not by of interest to this study. The easiest way to remove between-bird heterogeneity is by standardising each time series, e.g. by subtracting the mean of each time series and dividing by its standard deviation. The second form of heterogeneity requires a bit more thought.

15.3.3 Failed Approach 2

We have already established that we may not be interested in between species heterogeneity. Then why should we use seven variances to model it? Standardisation; subtracting the mean of each time series and dividing it by its standard deviation ensures that each time series is scaled in the same range. This allows us to drop the `varIdent` code to model different variances. The following R code standardises the data.

```
> Birds7 <- as.vector(as.matrix(scale(Ythan[,2:8])))
```

Note that this is nearly the same code as before, except that the `scale` function standardises each column in the selected part of the data matrix `Ythan`. We could have used:

```
> Birds7 <- c(scale(Ythan$Oystercatcher),
              scale(Ythan$Turnstone),
              scale(Ythan$Curlew),
              scale(Ythan$BartailedGodwit),
              scale(Ythan$Redshank),
              scale(Ythan $Knot),
              scale(Ythan$Dunlin))
```

Yet, a third option is to use `tapply`. We applied models (15.2) and (15.3) again, but this time we dropped the index i from ${\sigma_i}^2$ and α as all the time series are standardised (all have the same variance, and a mean of 0). R code for this is

```
f7 <- formula(Birds7 ~ 1+
    s(Year7, by = Oyst.01, bs = "cr") +
```

```
    s(Year7, by = Turn.01, bs = "cr") +
    s(Year7, by = Curl.01, bs = "cr") +
    s(Year7, by = Bart.01, bs = "cr") +
    s(Year7, by = Reds.01, bs = "cr") +
    s(Year7, by = Knot.01, bs = "cr") +
    s(Year7, by = Dunl.01, bs = "cr"))
M2 <- gamm(f7, method = "REML", control = lmc,
      weights = varPower(form =~ fitted(.) | ID7))
```

Again, the model with heterogeneity within the time series did not converge.

15.3.4 Assume Homogeneity?

Having tried everything we can think of, it is now time to acknowledge that for these data, we cannot easily model the heterogeneity within a bird time series. Probably, the data are just too short for this. There are now three options: (i) give up, (ii) transform the data and make statements on the transformed data, or (iii) assume homogeneity over time in Fig. 15.4. We decided to go for option 3. If the heterogeneity was more obvious, we would go for option (ii). The problem with option (ii) is that for other data sets, we saw that a transformation changed the shape of the trends. We readdress this issue in Section 15.5.

Before we can go into a discussion what graphs tell us in terms of biology, there is one last issue to discuss: independence over time.

15.4 Dealing with Independence

We return to the un-standardised data. We still have the potential problem of independence. The model in Equation (15.2) assumes that the residuals for bird species i in year s are independent of year s – 1, s – 2, etc. One way to verify this is the autocorrelation plot. However, due to the large number of missing values, a variogram may be a better tool to assess temporal dependence. The following R code extracts the residuals from the object `M0` (the model in Equation (15.2)) and calculates a (robust) variogram.

```
> plot(Variogram(M0$lme, form =~ Year7 | ID7,
                 maxDist = 10, robust = TRUE),
       pch = 16, smooth = FALSE, cex = 1.2)
```

The function `Variogram` is part of the `nlme` package. We used a maximum distance of 10 years as it is unlikely that birds in year s are affected by birds in year s – 10 (or over longer time lags). The `form` option specifies that the time series structure is within a bird species. If the points are scattered along a horizontal

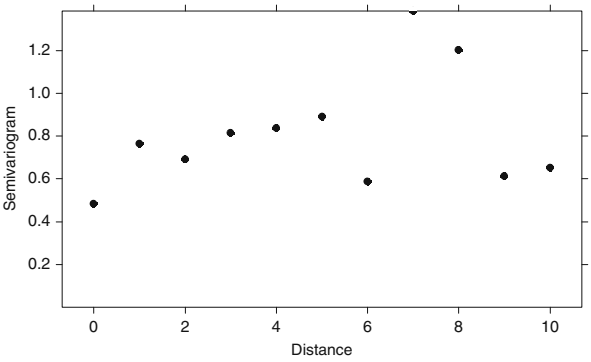

Fig. 15.5 Variogram for the residuals obtained by the additive model in Equation (15.2). The horizontal axis represents distance between years and the vertical axis the value of the variogram

line in the variogram, independence of the residuals may be assumed. Figure 15.5 indicates that this may be a valid assumption.

A more formal way of assessing dependence over time is to include a time series correlation structure in the model and then test with the likelihood ratio test (if the models are nested) or compare the models with a tool like the AIC or BIC.

Adding a temporal correlation to the model in Equation (15.2) is relatively easy.

```
> M0A <- gamm(f7, method = "REML",
     control = lmc, weights = varIdent(form=~1|ID7),
     correlation = corSpher(form =~ Year7 | ID7,
                   nugget = TRUE, fixed = FALSE))
```

The only new code is the correlation bit. It implements a spherical correlation structure as discussed in Chapter 7. In fact, we can try any of the following correlation options: No correlation, `corSpher`, `corRatio`, `corLin`, `corGaus`, `corExp`, and `corAR1`. R code for these models is given on the book website. The AIC value for the model without the temporal correlation was 1342.48, and the AICs of the models with a correlation structure were 1344.61 (`corSpher`), 1344.61 (corLin), 1343.77 (`CorRatio`), 1343.76 (`CorExp`), 1343.61 (`CorGaus`), and 1342.55 (`corAR1`). This means that adding a residual auto-correlation structure does not improve the model. Hence, our 'optimal' model is still the one in Equation (15.2).

We now have a look at the numerical output of `M0`. The estimated degrees of freedom, *F*-statistics, and *p*-values for the smoothers are obtained using the `anova(M0$gam)` command and are as follows.

```
Parametric Terms:
    df     F p-value
ID7  6 146.4  <2e-16
```

```
Approximate significance of smooth terms:
                    edf Est.rank     F  p-value
s(Year7):Oyst.01 3.626    8.000 6.101 1.72e-06
s(Year7):Turn.01 1.001    1.000 7.331 0.007861
s(Year7):Curl.01 1.001    2.000 6.839 0.001587
s(Year7):Bart.01 3.171    7.000 4.018 0.000594
s(Year7):Reds.01 3.259    7.000 4.882 7.98e-05
s(Year7):Knot.01 1.000    1.000 4.413 0.037946
s(Year7):Dunl.01 1.000    1.000 1.501 0.223209
```

All smoothers are highly significant, except for the Knot and Dunlin smoothers. The fitted values (= smoother plus intercept) are given in Fig. 15.6. Three species (oystercatcher, redshanks, and godwit) have high values around 1980 followed by a decrease. Knot, curlew, and turnstone show a nearly linear increase since the early 1970s. Confidence bands around the smoother for dunlin are rather larger and you should avoid drawing any conclusions for this species.

The following R code was used:

```
> P0 <- predict(M0$gam, se = TRUE)
> Isna <- is.na(Birds7)
> F <- P0$fit
> Fup <- P0$fit + 1.96 * P0$se.fit
> Flow <- P0$fit - 1.96 * P0$se.fit
> xyplot(F + Fup + Flow ~ Year7[!Isna] | ID7[!Isna],
   xlab = "Time", ylab = "Fitted values",
   lty=c(1, 2, 2), col = 1, type = c("l", "l", "l"),
   scales = list(alternating = TRUE,
                 x = list(relation = "same"),
                 y = list(relation = "free")))
```

Section 5.3.1 in Sarkar (2008) contains full details on the second part of this R code. First, we predict values from model `M0`. Because there are missing values, we have to remove them from the `Year7` and `ID7` vectors inside the `xyplot`. The variables F, Fup, and Flow contain the fitted values, upper confidence band, and lower confidence band, respectively. The `F + Fup + Flow ~ Year7` bit means that each vector is plotted versus `Year7`; it is not adding them up! Information on line type `(lty)` and `type` options are given in Sarkar (2008). Alternatively, just change the values and see what happens.

The shape of the trends in Fig. 15.6 makes one wonder whether we could summarise the oystercatcher, redshanks, and bar-tailed godwit with one trend and the turnstone, curlew and knot with another trend. To verify whether this is indeed the case, we can fit a model with seven trends and a model with three trends, use ML estimation in both models, and compare them using the AIC. The following code fits the model with three trends and with seven trends and compares them.

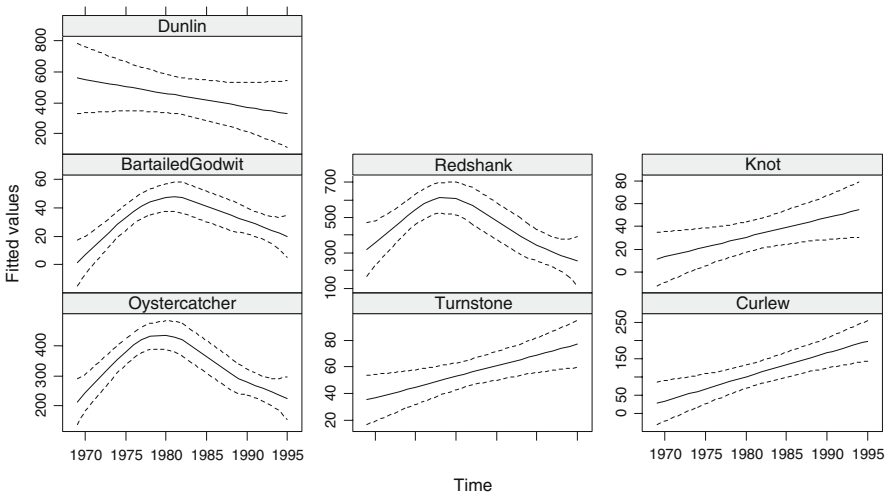

Fig. 15.6 Smoothing curves *(solid line)* obtained by the model in Equation (15.2). *Dotted lines* are 95% point-wise confidence bands

```
> ORB.01 <- as.numeric(ID7 =="Oystercatcher" |
                       ID7 =="Redshank" |
                       ID7 =="BartailedGodwit")
> TCK.01 <- as.numeric(ID7 == "Turnstone" |
                        ID7 == "Curlew" |
                        ID7 == "Knot")
> D.01 <- as.numeric(ID7 == "Dunlin")
> M0.3 <- gamm(Birds7 ~ 1 +
    s(Year7, by = ORB.01, bs = "cr") +
    s(Year7, by = TCK.01, bs = "cr") +
    s(Year7, by = D.01, bs = "cr"),
   method = "ML", control = lmc)
> M0.7 <- gamm(f7, control = lmc, method = "ML",
               weights = varIdent(form =~ 1 | ID7))
> AIC(M0.7$lme, M0.3$lme)
         df      AIC
M0.7$lme 28 1452.157
M0.3$lme 20 1460.026
```

The AIC shows that the model with seven trends is better than the model with three trends. Perhaps, we were fooled in Fig. 15.6 by the different ranges along the vertical axes. If they are all the same, the smoothers look rather different from each other.

15.5 To Transform or Not to Transform

Initially, we were frustrated with the analysis of these data and after a couple of failed approaches, we convinced ourselves that we were not interested in the heterogeneity within a time series. We tried several transformations, and by trial and error, we found that the square root transformation stabilised the within-bird-time series variance. The problem is that a transformation not only removes heterogeneity, but it may also changes the shape of the trends (and therefore the conclusions). Applying the transformation is simple, just use

```
> Birds7 <- as.vector(as.matrix(sqrt(Ythan[, 2:8])))
```

The rest of the code is identical. The `varIdent` variance structure was needed and adding a residual auto-correlation did not improve the models. The predicted trends for these data are given in Fig. 15.7. Except for dunlin, the shapes of the trends are similar compared to those in Fig. 15.6. The only differences are that for the square-root-transformed data, we can safely assume homogeneity, but the smoothers are on the square root scale.

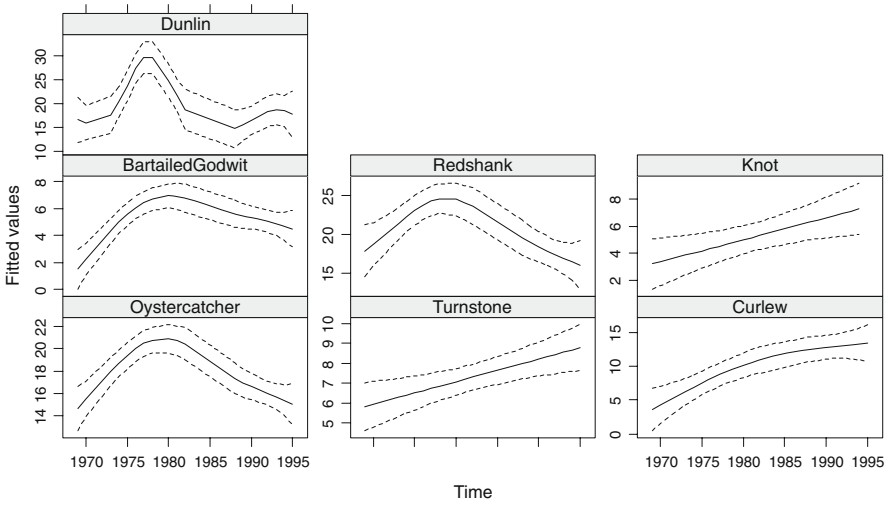

Fig. 15.7 Smoothing curves *(solid line)* obtained by the model in Equation (15.2). Square-root-transformed bird data were used. *Dotted lines* are 95% point-wise confidence bands

15.6 Birds and Explanatory Variables

In the previous section, we applied additive mixed models and found that 6 of the 7 birds could be divided into two groups. The oystercatcher, redshanks, and bartailed godwit follow a non-linear pattern over time with the highest values around

1980. The turnstone, curlew, and knot follow a linear and increasing pattern over time. Note that this is a visual observation; the actual trends are different from each other. No pattern for dunlin could be found. The question is now how to link the explanatory variables to either of the bird time series. The problem is that there are only 6 years in which both the birds and the explanatory variables were measured. The algorithm for additive modelling will use only these six years! So, there is no way we can add the explanatory variables into the additive mixed models used in the previous section. The other problem is that the shape of the smoothers in Fig. 15.2 indicate serious collinearity between nearly all explanatory variables. One of the few things that we can do is to plot the smoothers for the explanatory variables and the smoothers for the birds in one graph and see which ones are similar (Fig. 15.8). Because the explanatory variables had missing values, we predicted these values to avoid an erratic curve. But we did not predict values before the first year or beyond the last year of observation per explanatory variable. We put the curves that looked similar close to each other.

The concluding question is to decide which explanatory variable is best related to the two bird trends. Or perhaps we need to rephrase the question to: Which one is not related to the oystercatcher, redshanks, and bar-tailed godwit trend? The pigs, cattle, and sheep trends are remarkably similar to the oystercatcher, redshanks and bar-tailed godwit trend. Note that pigs and wheat follow similar patterns over time, but none of these clearly match any of the bird trends. Unfortunately, there are not enough observations in nitrate to say anything sensible, except that when the nitrate

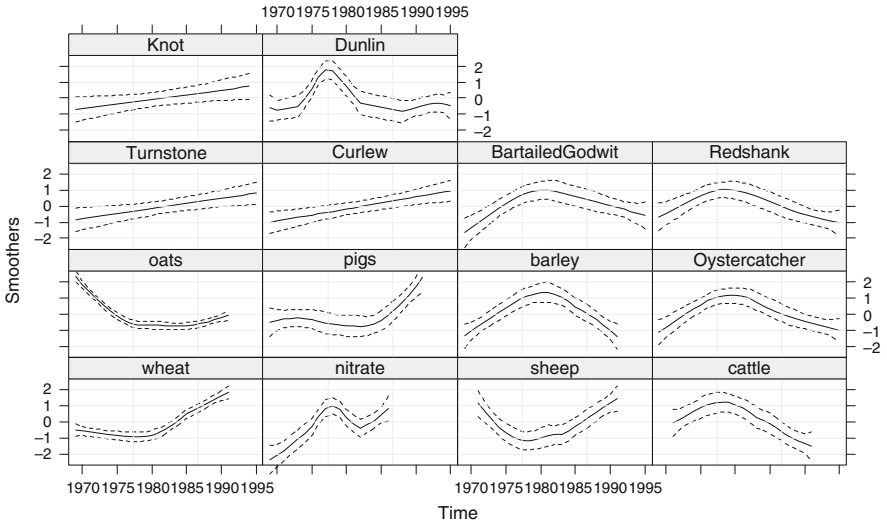

Fig. 15.8 *Smoothing curves* for each explanatory variable and the estimated smoothers from Fig. 15.6. The R code for this graph is presented on the book's website

trend went down in the late 1970s, the oystercatcher, redshanks, and bar-tailed godwit trends followed. However, this is rather speculative and based on a subjective observation.

15.7 Conclusions

The main statistical conclusions are that the seven bird time series seem to follow two patterns. The oystercatcher, redshanks and bar-tailed godwit all follow a similar (though not identical!) non-linear pattern over time, with the highest values around 1980. The turnstone, curlew, and knot all follow a linear (increasing) pattern over time. No pattern for dunlin could be found. The oystercatcher, redshank, and bar-tailed godwit trend seems to match the pigs, cattle, and sheep trends. When nitrate patterns changed, this bird trend changed as well. However, the data on the explanatory variables are too sparse to go beyond giving a nice multipanel graph where the bird trends and the trends for the explanatory variables are plotted.

These outcomes are interesting both ecologically and statistically. There were two main groups of trends over time in the bird data; one hump backed species group and one monotonic species group indicate that there are different ecological processes at work in this system: higher levels of nutrient enrichment seems good for some species, but bad for others. This may be at least in part explained by the ways in the elevated levels of nitrate are known to affect the invertebrates the birds feeding on. As nitrate levels increase in this system, the growth of mats of fast growing green seaweeds (henceforth termed 'alal mats') is stimulated, but the spatial distribution of these mats is very patchy. Underneath these patches, few of the invertebrates on which birds feed survive, but between the patches of weed, the same species of invertebrate thrive in the enriched conditions. So at low nutrient levels, the overall productivity of the estuary will increase, even though invertebrates are excluded from the patchy algal mats and the estuary can 'carry' more birds. At high levels of nutrients, however, the enriching effects on invertebrate numbers and biomass are markedly reduced as the algal mats spread into previously unaffected and enriched areas. Under this extensive cover of algal mats, the invertebrates on which birds feed virtually disappear over much of the estuary and shorebirds decline. One of the few species which is not reduced by the algal mats is the tiny mud snail *Hydrobia ulvae*, whose numbers may even increase within the mats. However, the tangled filamentous structure of the mats substantially reduces the foraging efficiency of the shorebirds so that there is no compensation for the loss of other invertebrates. One would therefore expect a non-monotonic trend in shorebirds over time with increasing nutrient run off as shown by oystercatcher, redshank, and bar-tailed godwit. Our analysis indicates that other factors are at work with respect to curlew, turnstone, and knot: Either they do not respond to algal mats in the same way as the other species or the continually increasing numbers on the estuary are a reflection of demographic processes happening outside the system, perhaps on the breeding grounds many thousands of kilometres distant. The statistical approaches used here

have crystallised this in a way which was not apparent in previous analyses, such as the likely different causes of changes in different shorebird species and allowed the framing of new research questions that can be explored in this system.

15.8 What to Write in a Paper

The first question we have to ask is as follows: Can we write a paper about the results presented in this chapter? The data analysed cover the life span of a scientific career, yet the data are too sparse to analyse bird data and agricultural variables in the same model. Having said this, fancy methods are not essential to compare trends in birds with agricultural variables. The link between nitrate and the oystercatcher, redshanks, and bar-tailed godwit trend is completely speculative and requires far more study before anything sensible can be said. This is something that should be made clear in the discussion of the paper! However, if you were to submit this chapter for publication, we would include the following.

1. An introduction describing the questions.
2. A Data and Methods section explaining how the data were collected and a few paragraphs on additive mixed modelling. Because this is a relatively new statistical method, half a page may be needed. You should explain the need for trying to add residual temporal correlation and heterogeneity structures. The referee may ask why you needed additive modelling, rather than just applying a transformation. Also justify why you used the Gaussian model and not the Poisson GLM or GAM.
3. It is tempting to present Fig. 15.2, but there will be a certain repetition with the graphs showing the final results.
4. You then need to summarise Section 15.3. Present the starting model, intermediate models, and the final model, (if these were not already presented in the Methods section), give AIC tables and likelihood ratio tests, and validate the optimal model. Present the F-values and p-values for the smoothers of the optimal model. Include the smoothers (or fitted values) for the optimal model (this is Fig. 15.6).
5. Make clear that due to the sparseness of the data, it is not possible to analyse bird data and agricultural variables in the same model. The only sensible thing to do is to present Fig. 15.8.
6. In the discussion, be sure not to say that sheep, barley or cattle are *driving* the mean winter values of oystercatcher, redshanks, and bar-tailed godwit. If anything, sheep, barley, or cattle are a measure of farming intensity, and increase use of fertilisers together with waste from livestock may drive nitrate concentrations in the Ythan. From this point onwards, the story becomes speculative, but interesting!