Chapter 6
Violation of Independence – Part I

This chapter explains how correlation structures can be added to the linear regression and additive model. The mixed effects models from Chapters 4 and 5 can also be extended with a temporal correlation structure. The title of this chapter contains 'Part I', suggesting that there is also a Part II. Indeed, that is the next chapter. In part I, we use regularly spaced time series, whereas in the next chapter, irregular spaced time series, spatial data, and data along an age gradient are analysed. We use a bird time series data set previously analysed in Reed et al. (2007). In the first section, we start with only one species and show how the linear regression model can be extended with a residual temporal correlation structure. In the second section, we use the same approach for a multivariate time series. In Section 6.3, the owl data are used again.

6.1 Temporal Correlation and Linear Regression

Reed et al. (2007) analysed abundances of three bird species measured at three islands in Hawaii. The data were annual abundances from 1956 to 2003. Here, we use one of these time series, moorhen abundance on the island of Kauai, to illustrate how to deal with violation of independence. A time series plot is given in Fig. 6.1. We applied a square root transformation to stabilise the variance, but strictly speaking, this is unnecessary as methods discussed earlier (Chapter 4) can be used to model the heterogeneity present in the original series. However, we do not want to over-complicate matters at this stage by mixing different concepts in the same model. The following R code imports the data and makes a plot of square-root-transformed moorhen numbers.

```
> library(AED); data(Hawaii)
> Hawaii$Birds <- sqrt(Hawaii$Moorhen.Kauai)
> plot(Hawaii$Year, Hawaii$Birds, xlab = "Year",
     ylab = "Moorhen abundance on Kauai")
```

Note that there is a general increase since the mid 1970s. Reed et al. (2007) used a dummy variable to test the effects of the implementation of new management

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_6,

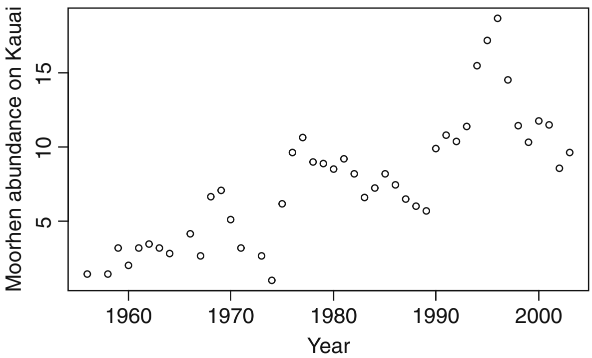

Fig. 6.1 Time series plot of square-root-transformed moorhen abundance measured on the island of Kauai

activities in 1974 on multiple bird time series, but to keep things simple, we will not do this here. The (transformed) abundance of birds is modelled as a function of annual rainfall and the variable Year (representing a long-term trend) using linear regression.

This gives a model of the form

$$\text{Birds}_s = \alpha + \beta_1 \times \text{Rainfall}_s + \beta_2 \times \text{Year}_s + \varepsilon_s \tag{6.1}$$

An alternative option is to use an additive model (Chapter 3) of the form:

$$\text{Birds}_s = \alpha + f_1(\text{Rainfall}_s) + f_2(\text{Year}_s) + \varepsilon_s$$

The advantage of the smoothers is that they allow for a non-linear trend over time and non-linear rainfall effects. Whichever model we use, the underlying assumption is that the residuals are independently normally distributed with mean 0 and variance σ^2. In formula we have

$$\begin{aligned} &\varepsilon_s \sim N(0, \sigma^2) \\ &\text{cov}(\varepsilon_s, \varepsilon_t) = \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{else} \end{cases} \end{aligned} \tag{6.2}$$

The second line is due to the independence assumption; residuals from different time points are not allowed to covariate. We already discussed how to incorporate heterogeneity using variance covariates in Chapter 4. Now, we focus on the independence assumption. The underlying principle is rather simple; instead of using the '0 else' in Equation (6.2), we model the auto-correlation between residuals of different time points by introducing a function *h(.)*:

$$\text{cor}(\varepsilon_s, \varepsilon_t) = \begin{cases} 1 & \text{if } s = t \\ h(\varepsilon_s, \varepsilon_t, \rho) & \text{else} \end{cases}$$

The function *h(.)* is called the correlation function, and it takes values between –1 and 1. Just as Pinheiro and Bates (2000), we assume stationarity. This means we assume that the correlation between the residuals ε_s and ε_t only depends on their time difference $s - t$. Hence, the correlation between ε_s and ε_t is assumed to be the same as that between ε_{s+1} and ε_{t+1}, between ε_{s+2} and ε_{t+2}, etc. The task of the analyst is to find the optimal parameterisation of the function *h(.)*, and we discuss several options in this and the next chapter. We assume the reader is familiar with the definition of the auto-correlation function, and how to estimate it from sample data; see for example Chatfield (2003), Diggle (1990), and Zuur et al. (2007), among others.

Before applying any model with a residual auto-correlation structure, we first apply the linear model without auto-correlation so that we have a reference point. In a preliminary analysis (not presented here), the cross-validation in the additive model gave one degree of freedom for each smoother, indicating that parametric models are preferred over smoothing models for this time series.

```
> library(nlme)
> M0 <- gls(Birds ~ Rainfall + Year,
            na.action = na.omit, data = Hawaii)
> summary(M0)
```

We used the `gls` function without any correlation or weights option, and as a result it fits an ordinary linear regression model. The `na.action` option is required as the time series contains missing value. The relevant output produced by the `summary` command is given below:

```
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: Hawaii
       AIC      BIC    logLik
  228.4798 235.4305 -110.2399

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -477.6634  56.41907 -8.466346  0.0000
Rainfall       0.0009   0.04989  0.017245  0.9863
Year           0.2450   0.02847  8.604858  0.0000

Residual standard error: 2.608391
Degrees of freedom: 45 total; 42 residual
```

The summary table shows that the effect of rainfall is not significant, but there is a significant increase in birds over time. The problem is that we cannot trust these p-values as we may be violating the independence assumption. The first choice to test this is to extract the standardised residuals and plot them against time (Fig. 6.2). Note that there is a clear pattern in the residuals.

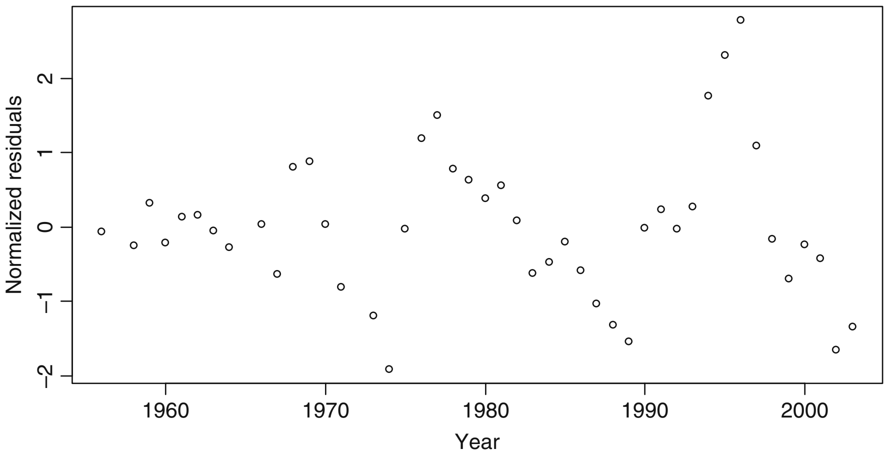

Fig. 6.2 Normalised residuals plotted versus time. Note the pattern in the residuals

A more formal visualisation tool to detect patterns is the auto-correlation function (ACF). The value of the ACF at different time lags gives an indication whether there is any auto-correlation in the data. The required R code for an ACF and the resulting graph are presented below. Note that the auto-correlation plot in Fig. 6.3 shows a clear violation of the independence assumption; various time lags have a significant

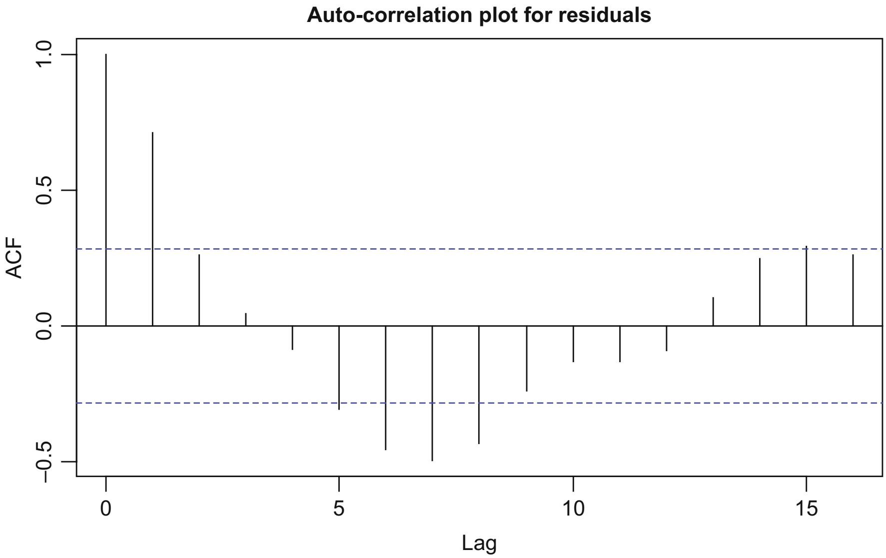

Fig. 6.3 Auto-correlation plot for the residuals obtained by applying linear regression on the Bird time series. Note that there is a clear indication of violation of independence

correlation! The ACF plot has a general pattern of decreasing values for the first 5 years, something we will use later in this section.

The R code for the ACF is given below.

```
> E <- residuals(M0, type = "normalized")
> I1 <- !is.na(Hawaii$Birds)
> Efull <- vector(length = length(Hawaii$Birds))
> Efull <- NA
> Efull[I1] <- E
> acf(Efull, na.action = na.pass,
    main = "Auto-correlation plot for residuals")
```

The function `residuals` extracts the normalised residuals. If there are no missing values, then you can just continue with `acf(E)`, but it is not that easy here. The time series has two missing values and to ensure that the correlation function is correctly calculated, we need to insert the two missing values in the right place. This is because the `gls` function is removing the missing values, whereas the `acf` function assumes that the points are at the right time position. Once this is done, we can calculate the auto-correlation function and the resulting graph is presented in Fig. 6.3.

Figure 6.3 shows the type of pattern you do not want to see if you were hoping for a quick analysis; these data clearly contain residual correlation. As a result, we cannot assume that the *F*-statistic follows an *F*-distribution and the *t*-statistic a *t*-distribution.

An alternative approach to judge whether auto-correlation is present and one that does not depend on a visual judgement of the auto-correlation plot is to include an auto-correlation structure into the model. Then compare the models with and without an auto-correlation structure using the AIC, BIC, or if the models are nested, a likelihood ratio test. However, you should not spend too much time trying to find the optimal residual auto-correlation structure. Citing from Schabenberger and Pierce (2002): 'In our experience it is more important to model the correlation structure in a reasonable and meaningful way rather than to model the correlation structure perfectly'. Similar statements can be found in Diggle et al. (2002), and Verbeke and Molenberghs (2000). We agree with this statement as differences in *p*-values for the *F*- and *t*-statistics obtained by using similar correlation structures tend to differ only marginally.

In Chapter 5, we used a slightly different mathematical notation compared to Equation (6.1); but if we use it here, the time series model for the birds in Equation (6.1) can be written as

$$\mathbf{Birds} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The vector **Birds** contains all 58 bird observations, **X** is a matrix of dimension 58×3, where the first columns consists of only ones, the second column the rainfall data, and the third column the years. The vector $\boldsymbol{\beta}$ is of dimension 3×1, and contains α, β_1, and β_2. Finally, $\boldsymbol{\varepsilon}$ is equal to a vector of length 58 with the elements $(\varepsilon_{1958}, \ldots, \varepsilon_{2003})$. Just as in Chapter 5, we can write Birds $\sim N(\mathbf{X} \times \boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V}$ is the covariance matrix of $\boldsymbol{\varepsilon}$. It is of the form

$$\mathbf{V} = \mathrm{cov}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \mathrm{var}(\varepsilon_{1958}) & & & & \\ \mathrm{cov}(\varepsilon_{1959}, \varepsilon_{1958}) & \mathrm{var}(\varepsilon_{1959}) & & & \\ \mathrm{cov}(\varepsilon_{1960}, \varepsilon_{1958}) & \mathrm{cov}(\varepsilon_{1960}, \varepsilon_{1959}) & \ddots & & \\ \vdots & \vdots & \cdots & \ddots & \\ \mathrm{cov}(\varepsilon_{2003}, \varepsilon_{1958}) & \mathrm{cov}(\varepsilon_{2003}, \varepsilon_{1959}) & \cdots & \mathrm{cov}(\varepsilon_{2003}, \varepsilon_{2002}) & \mathrm{var}(\varepsilon_{2003}) \end{pmatrix}$$

Under the independence assumption, $\mathbf{V}$ is a diagonal matrix of the form $\sigma^2 \times \mathbf{I}$, where $\mathbf{I}$ is a 58×58 identity matrix. The easiest auto-correlation structure is the so-called compound symmetry structure. We have already met this correlation structure in Chapter 5. It assumes that whatever the distance in time between two observations, their residual correlation is the same. This can be modelled as

$$\mathrm{cor}(\varepsilon_s, \varepsilon_t) = \begin{cases} 1 & \text{if } s = t \\ \rho & \text{else} \end{cases} \tag{6.3}$$

Hence, the correlation structure in Equation (6.3) is implying the following correlation matrix for $\boldsymbol{\varepsilon}$.

$$\mathrm{cor}(\boldsymbol{\varepsilon}) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}$$

This corresponds to the following covariance matrix $\mathbf{V}$, where $\rho = \theta/(\theta + \sigma^2)$.

$$\mathbf{V} = \mathrm{cov}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \theta + \sigma^2 & \theta & \theta & \cdots & \theta \\ \theta & \theta + \sigma^2 & \theta & \cdots & \theta \\ \theta & \theta & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \theta \\ \theta & \theta & \cdots & \theta & \theta + \sigma^2 \end{pmatrix}$$

Pinheiro and Bates (2000) mention that this correlation structure is often too simplistic for time series, but may still be useful for short time series. It can be implemented in R using the following code.

```
> M1 <- gls(Birds ~ Rainfall + Year,
            na.action = na.omit, data = Hawaii ,
            correlation = corCompSymm(form =~ Year))
```

The residual correlation structure is implemented using the `correlation` option in the `gls` function. The argument `corCompSymm` is the compound symmetry auto-correlation structure. The `form` argument within this argument is used to tell R that the order of the data is determined by the variable Year. However, due to the nature of the correlation structure, the `form` option is not needed (yet). Results of the `summary` command are not presented here, but give AIC = 230.47, BIC = 239.16, $\rho = 0$, and the estimated regression parameters and p-values are the same as for the ordinary linear regression model. So, we have made no improvements in the model.

The next structure we discuss is the AR-1 auto-correlation. This cryptic notation stands for an auto-regressive model of order 1. It models the residual at time s as a function of the residual of time $s - 1$ along with noise:

$$\varepsilon_s = \rho\varepsilon_{s-1} + \eta_s \tag{6.4}$$

The parameter ρ is unknown, and needs to be estimated from the data. It is relatively easy to show that this error structure results in the following correlation structure:

$$\text{cor}(\varepsilon_s, \varepsilon_t) = \begin{cases} 1 & \text{if } s = t \\ \rho^{|t-s|} & \text{else} \end{cases} \tag{6.5}$$

Suppose $\rho = 0.5$ and $t = s + 1$. The correlation between residuals separated by one unit in time is then 0.5. If the separation is two units in time, the correlation is $0.5^2 = 0.25$. Hence, the further away two residuals are separated in time, the lower their correlation. For many ecological examples, this makes sense. To emphasise the imposed correlation structure, we show the correlation matrix for $\boldsymbol{\varepsilon}$ again.

$$\text{cor}(\boldsymbol{\varepsilon}) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{57} \\ \rho & 1 & \rho & \ddots & \ddots & \vdots \\ \rho^2 & \rho & 1 & \ddots & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & \ddots & \rho & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & 1 & \rho \\ \rho^{57} & \cdots & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

The following code implements the AR-1 correlation structure.

```
> M2 <- gls(Birds ~ Rainfall + Year,
            na.action = na.omit, data = Hawaii,
            correlation = corAR1(form =~ Year))
> summary(M2)
```

The only thing that has changed compared to the compound symmetry structure is the correlation argument `corAR1`. The form argument is now essential as R needs to know position of the observations over time. The `na.action` option is also required due to the missing values. The relevant output obtained by the `summary` command is

```
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year. Data: Hawaii
       AIC      BIC   logLik
  199.1394 207.8277 -94.5697

Correlation Structure: ARMA(1,0)
 Formula: ~Year
 Parameter estimate(s):
     Phi1
0.7734303

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -436.4326 138.74948 -3.145472  0.0030
Rainfall      -0.0098   0.03268 -0.300964  0.7649
Year           0.2241   0.07009  3.197828  0.0026

Residual standard error: 2.928588
Degrees of freedom: 45 total; 42 residual
```

The parameter ρ is equal to 0.77. This means that residuals separated by one year have a correlation of 0.77; by two years it is $0.77^2 = 0.59$. This is rather high, but seems to be in line with the pattern for the first few years in the auto-correlation function in Fig. 6.3. The AIC indicates that the AR-1 correlation structure is a considerable model improvement compared to the linear regression model. In general, you would expect ρ to be positive as values at any particular point in time are positively related to preceding time points. Occasionally, you find a negative ρ. Plausible explanations are either the model is missing an important explanatory variable or the abundances go from high values in one year to low values in the next year.

6.1.1 ARMA Error Structures

The AR-1 structure can easily be extended to a more complex structure using an auto-regressive moving average (ARMA) model for the residuals. The ARMA model has two parameters defining its order: the number of auto-regressive parameters (p) and the number of moving average parameters (q). The notation ARMA(1, 0) refers to the AR-1 model described above. The ARMA(p, 0) structure is given by

$$\varepsilon_s = \phi_1 \varepsilon_{s-1} + \phi_2 \varepsilon_{s-2} + \phi_3 \varepsilon_{s-3} + \cdots + \phi_p \varepsilon_{s-p} + \eta_s \tag{6.6}$$

The residuals at time *s* are modelled as a function of the residuals of the *p* previous time points and white noise. In this case, the function *h(.)* does not have an easy formulation, see Equation (6.27) in Pinheiro and Bates (2000). The ARM(0,*q*) is specified by

$$\varepsilon_s = \theta_1 \eta_{s-1} + \theta_2 \eta_{s-2} + \theta_3 \eta_{s-3} + \cdots + \theta_q \eta_{s-q} + \eta_s \tag{6.7}$$

And the ARMA (*p*, *q*) is a combination of the two. You should realise that all these *p* and *q* parameters have to be estimated from the data, and in our experience, using values of *p* or *q* larger than 2 or 3 tend to give error messages related to convergence problems. Even for $p = q = 3$, it already becomes an art to find starting values so that the algorithm converges. Obviously, this also depends on the data, and how good the model is in terms of fixed covariates (year and rainfall in this case). The ARMA(*p*, *q*) can be seen as a black box to fix residual correlation problems.

The implementation of the ARMA(*p*, *q*) error structure in R is as follows.

```
> cs1 <- corARMA(c(0.2), p = 1, q = 0)
> cs2 <- corARMA(c(0.3, -0.3), p = 2, q = 0)
> M3arma1 <-gls(Birds ~ Rainfall + Year,
              na.action = na.omit,
              correlation = cs1, data = Hawaii)
> M3arma2 <- gls(Birds ~ Rainfall + Year,
               na.action = na.omit,
               correlation = cs2, data = Hawaii)
> AIC(M3arma1, M3arma2)
```

This code applies the ARMA(1,0) and ARMA(2,0) error structure. We chose arbitrary starting values. For larger values of *p* and *q*, you may need to change these starting values a little.

Finding the optimal model in terms of the residual correlation structure is then a matter of applying the model with different values of *p* and *q*. But remember the citation from Schabenberger and Pierce (2002) given at the start of this section; there is not much to be gained from finding the perfect correlation structure compared to finding one that is adequate. We tried each combination of $p = 0, 1, 2, 3$ and $q = 0, 1, 2, 3$, and each time we wrote down the AIC. Because not all the models are nested, we cannot apply a likelihood ratio test and have therefore based our model selection on the AIC. The lowest AICs were obtained by the ARMA(2,0) and ARMA(2,3) models and were 194.5 and 194.1, respectively. Both AICs differed only in the first decimal, and we selected the ARMA(2,0) model as it is considerably less complex than the ARMA(2,3) model. Recall that the linear regression model without a residual auto-correlation structure had AIC = 228.47, and the AR-1 structure gave AIC = 199.13. So, going from no residual correlation to an AR-1 structure gave a large improvement, while the more complicated structures gave only a marginal improvement. The estimated auto-regressive parameters of the ARMA(2,0) model were $\varphi_1 = 0.99$ and $\varphi_2 = -0.35$. The value for φ_1 close to 1 may indicate a more serious problem of the residuals being non-stationary (non-constant mean or variance). Note that the auto-correlation function in Fig. 6.3 becomes positive again for larger time lags. This suggests that an error structure that allows for a sinusoidal pattern may be more appropriate.

The correlation structure can also be used for generalised additive models, and it is also possible to have a model with residual correlation and/or heterogeneity structures.

6.2 Linear Regression Model and Multivariate Time Series

Figure 6.4 shows the untransformed abundances of two bird species (stilts and coots) measured on the islands Maui and Oahu. These time series form part of a larger data set analysed in Reed et al. (2007), but these four series are the most complete. Again, we use annual rainfall and year as explanatory variables to model bird abundances. Preliminary analyses suggested a linear rainfall effect that was the same for all four time series and a non-linear trend over time. Hence, a good starting model is

$$\text{Birds}_{is} = \alpha_i + \beta \times \text{Rainfall}_{is} + f_i(\text{Year}_s) + \varepsilon_{is} \tag{6.8}$$

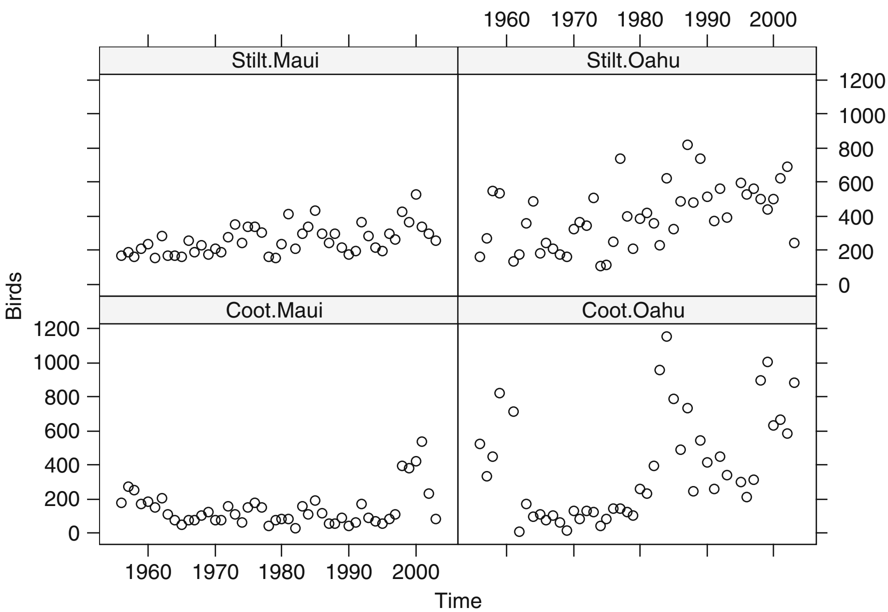

Fig. 6.4 Time series of (untransformed) silt and coot abundances on the islands of Maui and Oahu

Bird_{is} is the value of time series i ($i = 1, \ldots, 4$) in year s ($s = 1, \ldots, 48$). For the moment, we treat the time series for the two species and two islands as different time series. The intercept α_i allows for a different mean value per time series. An extra motivation to use no rainfall–species or rainfall–island interaction is that some intermediate models had numerical problems with the interaction term. Year_s is the year and $f_i(\text{Year}_s)$ is a smoother for each species–island combination. If we remove the index i, then all four time series are assumed to follow the same trend.

The range of the y-axes in the lattice plot immediately indicates that some species have considerably more variation, indicating violation of homogeneity. The solution is to allow for different spread per time series.

The following code (i) imports the data into R, (ii) creates the lattice graph in Fig. 6.4, and (iii) applies the model in Equation (6.8).

```
> library(AED); data(Hawaii)
> Birds <- c(Hawaii$Stilt.Oahu, Hawaii$Stilt.Maui,
             Hawaii$Coot.Oahu, Hawaii$Coot.Maui)
> Time <- rep(Hawaii$Year, 4)
> Rain <- rep(Hawaii$Rainfall, 4)
> ID <- factor(rep(c("Stilt.Oahu", "Stilt.Maui",
                     "Coot.Oahu", "Coot.Maui"),
                   each = length(Hawaii$Year)))
> library(lattice)
> xyplot(Birds ~ Time | ID, col = 1)
> library(mgcv)
> BM1<-gamm(Birds ~ Rain + ID +
        s(Time, by = as.numeric(ID == "Stilt.Oahu")) +
        s(Time, by = as.numeric(ID == "Stilt.Maui")) +
        s(Time, by = as.numeric(ID == "Coot.Oahu")) +
        s(Time, by = as.numeric(ID == "Coot.Maui")),
        weights = varIdent(form =~ 1 | ID))
```

The first line imports the data. The next line stacks all four time series and calls it 'Birds'. Obviously, we also have to stack the variables Year and Rainfall, and the `rep` command is a useful tool for this. Finally, we need to make sure we know which observation belongs to which time series, and this is done using the variable 'ID'. The familiar `xyplot` command from the lattice package draws Fig. 6.4. The interested reader can find information on how to add gridlines, connect the dots, etc., in other parts of this book. The model in Equation (6.8) is an additive model with Gaussian distribution. The `weights` option with the `varIdent` argument was discussed in Chapter 4. Recall that it implements the following variance structure:

$$\varepsilon_s \sim N(0, \sigma_i^2) \qquad i = 1, \cdots, 4 \tag{6.9}$$

Each time series is allowed to have a different residual spread. The `by = as.numeric(.)` command ensures that each smoother is only applied on one time series. The same model could have been fitted with the `gam` command instead of the `gamm`, but our choice allows for a comparison with what is to come.

The numerical output for the smoothing model is given below.

```
> summary(BM1$gam)
Family: gaussian. Link function: identity
Formula: Birds ~ Rain + ID +
                  s(Time, by = as.numeric(ID == "Stilt.Oahu")) +
                  s(Time, by = as.numeric(ID == "Stilt.Maui")) +
                  s(Time, by = as.numeric(ID == "Coot.Oahu")) +
                  s(Time, by = as.numeric(ID == "Coot.Maui"))
Parametric coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  225.3761    20.0596  11.235  < 2e-16
Rain          -4.5017     0.8867  -5.077 9.93e-07
IDCoot.Oahu  237.7378    30.3910   7.823 5.06e-13
IDStilt.Maui 117.1357    14.9378   7.842 4.53e-13
IDStilt.Oahu 257.4746    27.1512   9.483  < 2e-16
Approximate significance of smooth terms:
                                          edf Est.rank      F  p-value
s(Time):as.numeric(ID == "Stilt.Oahu") 1.000        1 13.283 0.000355
s(Time):as.numeric(ID == "Stilt.Maui") 1.000        1 20.447 1.14e-05
s(Time):as.numeric(ID == "Coot.Oahu")  6.660        9  8.998 4.43e-11
s(Time):as.numeric(ID == "Coot.Maui")  2.847        6  3.593 0.002216
R-sq.(adj) =  0.813  Scale est. = 26218    n = 188
```

The problem here is that the *p*-values assume independence and because the data are time series, these assumptions may be violated. However, just as for the univariate time series, we can easily implement a residual auto-correlation structure, for example, the AR-1:

$$\varepsilon_{is} = \rho\varepsilon_{i,s-1} + \eta_{is} \tag{6.10}$$

As before, this implies the following correlation structure:

$$\operatorname{cor}(\varepsilon_{is}, \varepsilon_{it}) = \begin{cases} 1 & \text{if } s = \text{t} \\ \rho^{|t-s|} & \text{else} \end{cases} \tag{6.11}$$

The correlation between residuals of different time series is assumed to be 0. Note that the correlation is applied at the deepest level: Observations of the same time series. This means that all time series have the same ρ. The following R code implements the additive model with a residual AR-1 correlation structure.

```
> BM2 <- gamm(Birds ~ Rain + ID +
       s(Time, by = as.numeric(ID == "Stilt.Oahu")) +
       s(Time, by = as.numeric(ID == "Stilt.Maui")) +
```

```
        s(Time, by = as.numeric(ID == "Coot.Oahu")) +
        s(Time, by = as.numeric(ID == "Coot.Maui")),
        correlation = corAR1(form =~ Time | ID ),
        weights = varIdent(form = ~1 | ID))
> AIC(BM1$lme, BM2$lme)
```

The only new piece of is the `correlation = corAR1 (form = ~Time | ID)`. The `form` option specifies that the temporal order of the data is specified by the variable Time, and the time series are nested. The auto-correlation is therefore applied at the deepest level (on each individual time series), and we get one ρ for all four time series. The AIC for the model without auto-correlation is 2362.14 and with auto-correlation it is 2351.59, which is a worthwhile reduction. The `anova(BM2$gam)` command gives the following numerical output for the model with AR-1 auto-correlation.

```
Parametric Terms:
     df     F  p-value
Rain  1 18.69 2.60e-05
ID    3 20.50 2.08e-11

Approximate significance of smooth terms:
                                         edf Est.rank      F  p-value
s(Time):as.numeric(ID == "Stilt.Oahu") 1.000    1.000 27.892 3.82e-07
s(Time):as.numeric(ID == "Stilt.Maui") 1.000    1.000  1.756    0.187
s(Time):as.numeric(ID == "Coot.Oahu")  6.850    9.000 22.605  < 2e-16
s(Time):as.numeric(ID == "Coot.Maui")  1.588    4.000  1.791    0.133
```

The Oahu time series have a significant long-term trend and rainfall effect, whereas the Maui time series are only affected by rainfall. The `plot(BM2$gam, scale = FALSE)` command produces the four panels in Fig. 6.5. Note that the smoothers in panels B and D are not significant. Further model improvements can be obtained by dropping these two smoothers from the model.

The long-term trend for stilts on Oahu (panel A) is linear, but the coots on Oahu show a non-linear trend over time. Abundances are increasing from the early 1970s onwards. The results from the `summary(BM2$gam)` command are not shown, but indicate that the rainfall effect is negative and highly significant ($p < 0.001$). The adjusted R^2 is 0.721. The `summary(BM2$lme)` results are not shown either, but give $\rho = 0.32$, large enough to keep it in the model.

The normalised residuals are plotted versus time in Fig. 6.6. The stilt residuals at Maui show some evidence of heterogeneity over time. It may be an option to use the `varComb` option to allow for heterogeneity per time series (as we have done here) but also along time, see Chapter 4. We leave this as an exercise for the reader. If you do attempt to apply such a model, it would make sense to remove the square root transformation. Figure 6.5 was created using the following R code.

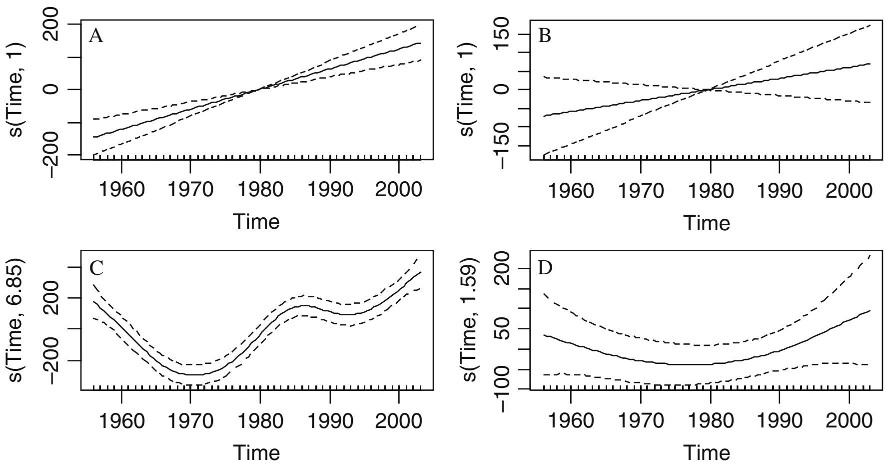

Fig. 6.5 **A**: Significant smoother for stilts in Oahu showing a linear increase over time. **B**: Non-significant smoother for stilts on Maui. **C**: Significant smoother for coots on Oahu. **D**: Non-significant smoother for coots on Maui. The four panels were created with the `par (mfrow = c (2,2))` command before the plot command

```
> E2 <- resid(BM2$lme, type = "normalized")
> EAll <- vector(length = length(Birds))
> EAll[] <- NA
> I1 <- !is.na(Birds)
> EAll[I1] <- E2
> library(lattice)
> xyplot(EAll ~ Time | ID, col = 1, ylab = "Residuals")
```

The only difficult aspect of the R code is dealing with missing values. Our approach is to create a vector `EAll` of length 192, fill in missing values, and fill in the matching values of the residuals E_2 at the right places (i.e. where we do not have missing values).

We need to investigate one last aspect. The model we applied above assumes that residuals are normally distributed with a variance that differs per time series and allows for auto-correlation within a time series. But, we also assume there is no correlation of residuals for different time series. This assumption could be violated if birds on one island are affecting those on other islands. Or there may be other biological reasons why the residual patterns of different time series are correlated. Whatever the biological reason, we need to verify this assumption. This is done by calculating the correlation coefficients between the four residual time series. If these correlation coefficients are reasonably small, we can assume independence between residuals of different time series. The following code extracts the residuals per time series, calculates an auto-correlation function, and a 4-by-4 correlation matrix.

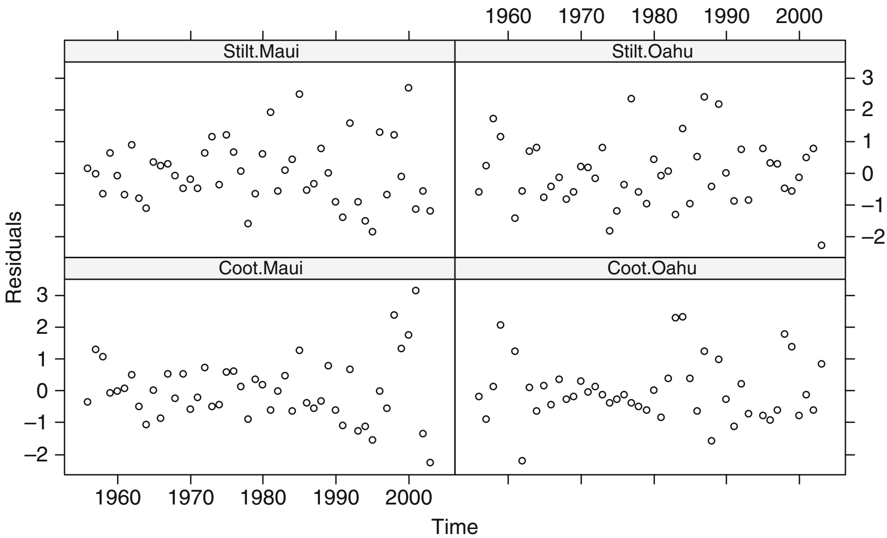

Fig. 6.6 Normalised residuals obtained by the additive model that allows for heterogeneity and an AR-1 residual error structure. The residual spread for the stilt at Oahu are perfect, but the residual spread for the stilts at Maui show a clear increase. One can argue about the interpretation of coot residual patterns

```
> E1 <- EAll[ID == "Stilt.Oahu"]
> E2 <- EAll[ID == "Stilt.Maui"]
> E3 <- EAll[ID == "Coot.Oahu"]
> E4 <- EAll[ID == "Coot.Maui"]
> par(mfrow = c(2, 2))
> acf(E1, na.action = na.pass)
> acf(E2, na.action = na.pass)
> acf(E3, na.action = na.pass)
> acf(E4, na.action = na.pass)
> D <- cbind(E1, E2, E3, E4)
> cor(D, use = "pairwise.complete.obs")
```

Results are not presented here, but all correlation coefficients are smaller than 0.2, except for the correlation coefficient between stilts and coots on Maui ($r = 0.46$). This may indicate that the model is missing an important covariate for the Maui time series. The three options are (i) find the missing covariate and put it into the model, (ii) extend the residual correlation structure by allowing for the correlation, and (iii) ignore the problem because it is only one out of the six correlations, and all p-values in the model were rather small (so it may have little influence on the conclusions). If more than one correlation has a high values, option (iii) should not be considered. You could try programming your own correlation structure allowing for spatial *and* temporal correlation.

6.3 Owl Sibling Negotiation Data

In Section 5.10, we analysed the owl sibling negotiation data. The starting point was a model of the form:

$$\begin{aligned} LogNeg_{is} = \alpha + \beta_1 \times SexParent_{is} + \beta_2 \times FoodTreatment_{is} + \\ \beta_3 \times ArrivalTime_{is} + \beta_4 \times SexParent_{is} \times FoodTreatment_{is} + \\ \beta_5 \times SexParent_{is} \times ArrivalTime_{is} + \varepsilon_{is} \end{aligned}$$

$LogNeg_{is}$ is the log transformed sibling negotiation at time s in nest i. Recall that we used nest as a random intercept, and therefore, the compound correlation structure was imposed on the observations from the same nest. We can get the same correlation structure (and estimated parameters) by specifying this correlation structure explicitly with the R code:

```
> library(AED) ; data(Owls)
> library(nlme)
> Owls$LogNeg <- log10(Owls$NegPerChick + 1)
> Form <- formula(LogNeg ~ SexParent * FoodTreatment +
                  SexParent * ArrivalTime)
> M2.gls <- gls(Form, method = "REML", data = Owls,
           correlation = corCompSymm(form =~ 1 | Nest))
```

You will see that the `summary(M2.gls)` command produces exactly the same estimated parameters and correlation structure compared to the random intercept model presented in Section 5.10. The `summary` command gives an estimated correlation of 0.138. Hence, the correlation between any two observations from the same nest i is given by

$$cor(\varepsilon_{is}, \varepsilon_{it}) = 0.138$$

It is important to realise that both random intercept and compound correlation models assume that the correlation coefficient between any two observations from the same nest are equal, whether the time difference is 5 minutes or 5 hours. Based on the biological knowledge of these owls, it is more natural to assume that observations made close to each other in time are more similar than those separated further in time. This sounds like the auto-regressive correlation structure of order 1, which was introduced in Section 6.1, and is given again below.

$$cor(\varepsilon_{is}, \varepsilon_{it}) = \rho^{|t-s|}$$

There are two 'little' problems. The numbers below are the first 12 lines of the data file and were obtained by typing

```
> Owls[Owls$Nest=="AutavauxTV",1:5]

         Nest FoodTreatment SexParent ArrivalTime SiblingNegotiation
1  AutavauxTV      Deprived      Male       22.25                  4
2  AutavauxTV      Satiated      Male       22.38                  0
3  AutavauxTV      Deprived      Male       22.53                  2
4  AutavauxTV      Deprived      Male       22.56                  2
5  AutavauxTV      Deprived      Male       22.61                  2
6  AutavauxTV      Deprived      Male       22.65                  2
7  AutavauxTV      Deprived      Male       22.76                 18
8  AutavauxTV      Satiated    Female       22.90                  4
9  AutavauxTV      Deprived      Male       22.98                 18
10 AutavauxTV      Satiated    Female       23.07                  0
11 AutavauxTV      Satiated    Female       23.18                  0
12 AutavauxTV      Deprived    Female       23.28                  3
```

The experiment was carried out on two nights, and the food treatment changed. Observations 1 and 2 were made at 22.25 and 22.38 hours, but the time difference between them is not 13 minutes, but 24 hours and 13 minutes! So, we have to be very careful where we place the auto-regressive correlation structure. It should be within a nest on a certain night. The random intercept and the compound correlation models place the correlation within the same nest, irrespective of the night.

The second problem is that the observations are not regularly spaced, at least not from our point of view; see Fig. 6.7. However, from the owl parent's point of view, time between visits may be regularly spaced. With this we mean that it may well be possible that the parents chose the nest visiting times. Obviously, if there is not enough food, and the parents need a lot of effort or time to catch prey, this is not a valid assumption. But if there is a surplus of food, this may well be a valid assumption. For the sake of the example, let us assume the owls indeed chose the times, and therefore, we consider the longitudinal data as regularly spaced. This basically means that we assume that distances (along the time axis) between the vertical lines in Fig. 6.7 are all the same. A similar approach was followed in Ellis-Iversen et al. (2008). Note that this is a biological assumption.

In this scenario, we can consider the visits at a nest on a particular night as regular spaced and apply the models with an auto-regressive correlation structure, e.g. the `corAR1` structure. The following R code does the job (the first few lines are used for Fig. 6.7):

```
> library(lattice)
> xyplot(LogNeg ~ ArrivalTime | Nest, data = Owls,
     type = "h", col = 1, main = "Deprived",
     subset = (FoodTreatment == "Deprived"))
> M3.gls <- gls(Form, method = "REML", data = Owls,
            correlation = corAR1(form =~ 1 |
                                     FoodTreatment))
```

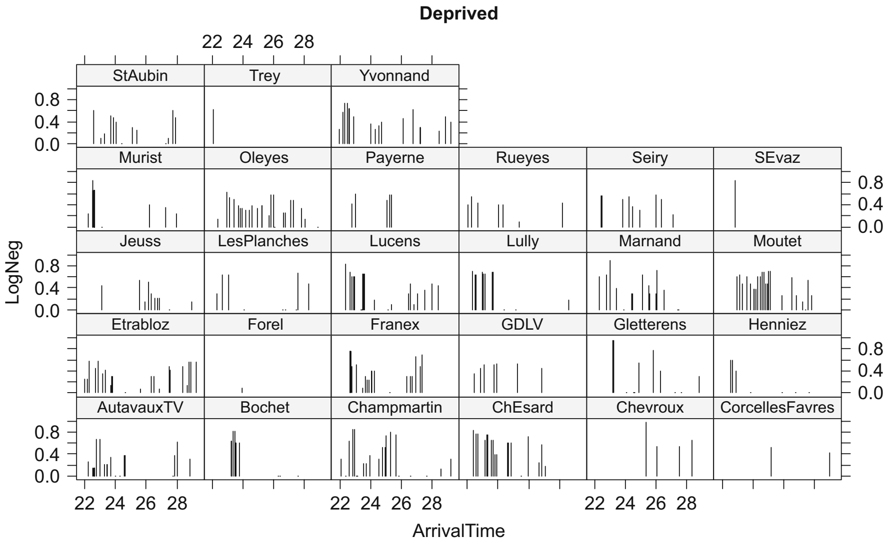

Fig. 6.7 Log-transformed sibling negotiation data versus arrival time. Each panel shows the data from one nest on a particular night. A similar graph can be made for the satiated data. R code to make this graph is given in the text

The variables `FoodTreatment` and `Nest` identify the group of observations from the same night, and the correlation is applied within this group. As a result, the index *i* in the model does not represent nest, but night in the nest. The `summary(M3.gls)` command shows that the estimated auto-correlation is 0.418, which is relatively high. The whole 10-step protocol approach can now be applied again: first chose the optimal random structure and then the optimal fixed structure. You can also choose to model arrival time as a smoother, just as we did in Section 5.8. This gives a GAM with auto-correlation.

The model with the auto-regressive correlation structure assumes that observations from different nests are independent and also that the observations from the same nest on two different nights are independent. It may be an option to extend the model with the AR1 correlation structure with a random intercept nest. Such a model allows for the compound correlation between all observations from the same nest, and temporal correlation between observations from the same nest *and* night. But the danger is that the random intercept and auto-correlation will fight with each other for the same information. These types of models are also applied in Chapter 17, where station is used as a random intercept and a correlation structure is applied along depth, but *within* the station.