Chapter 14
Estimating Trends for Antarctic Birds in Relation to Climate Change

A.F. Zuur, C. Barbraud, E.N. Ieno, H. Weimerskirch, G.M. Smith, and N.J. Walker

14.1 Introduction

The earth's climate is changing rapidly and these changes are expected to affect the structure and functioning of ecosystems. It is now clearly established that recent climate changes have impacted on living organisms. Several studies have demonstrated changes in population abundance, geographic distribution, and even microevolutionary changes in relation to climatic fluctuations (Parmesan, 2006).

Perhaps the best documented and most spectacular responses of living organisms to climate change are changes in phenology, which is the timing of seasonal activities of biological events such as the sprouting of plants. The vast majority of studies from the Northern Hemisphere that have analysed the relationships between long-term phenological and climate data sets have reported an advance in spring activities. For example, the earlier arrival and reproduction of migratory birds or earlier breaking of leaf buds since the mid-20th century in response to increasing temperatures. Some studies have also reported early onset of autumn activities such as grape-harvesting dates. However, due to the scarcity of long-term data sets, phenological changes are poorly documented in the Southern Hemisphere, particularly in Antarctica. Nevertheless, it is crucial to know whether, and to what extent, phenological changes have also occurred in the Southern Hemisphere for at least two reasons: (i) climatic changes between both hemispheres are different and (ii) we need to understand and eventually predict the impact of future climatic changes on species and ecosystems.

Permanent human occupation of the Antarctic continent is very recent compared to the other continents, and the landmark for scientific studies in Antarctica is the International Polar Year 1957–58 when most of the existing permanent research stations were built. In Terre Adélie, East Antarctica, the Dumont d'Urville research station was established during the mid 1950s, and since then, ornithologists have over wintered almost every year recording arrival and laying dates of Antarctic seabirds as part of long-term studies on Antarctic marine top predators (Barbraud

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_14,

Fig. 14.1 Emperor Penguin. The photograph was taken by C. Barbraud

and Weimerskirch, 2006). Fortunately, all but one of the Antarctic seabird species breed close to the research station and records of phenological data have been collected over a 50-year period with quasi-annual frequency.

Here, we use arrival and laying dates of three of these bird species to estimate trends and determine the effects of possible explanatory variables.

The Emperor Penguin *Aptenodytes forsteri* is the largest of the existing penguins (males weigh up to 45 kg) and breeds in winter on solid sea ice (Fig. 14.1). Males and females arrive on the breeding area from mid to late March. During the next two months, pairs form and the female lays the single egg to her male partner, who will incubate during the next two months in the heart of winter. Then, as with most seabirds, males and females alternate foraging trips at sea to feed, and to bring back food for the growing chick at the colony. The chicks leave the colony in early December at the onset of summer.

The Adelie Penguin *Pygoscelis adeliae* is a medium-sized penguin (c. 4.5 kg), breeding during the austral summer con rocky islands or on coastal nunataks (ice-free areas of the Antarctic continent). Adelies arrive and start building their nest just after mid-October and lay their eggs in mid November. The chicks leave the colonies in early February, just before the winter.

The Cape Petrel *Daption capense* is a small (c. 400 g) Procellariiform species and breeds during the austral summer on rocky islands. The breeding period is relatively short because birds arrive in mid October, lay their egg in late November, and the chicks are fledged in early March.

During the breeding period, the three species feed directly on krill or on fish that heavily depend on krill. Krill abundance and distribution are closely related to sea ice. After winters with extensive sea ice, adult krill survival and krill recruitment is high; therefore, krill abundance is higher after winters with extensive sea ice compared to winters with poor sea ice.

14.1.1 Explanatory Variables

During the breeding period, satellite tracking and diet studies have shown that all three species are more or less associated with sea ice. Both penguin species use sea

ice floes as resting platforms and forage within the pack ice. And although Cape Petrels do not forage directly in sea ice habitats, they feed in areas of open water that are covered by sea ice during winter. Consequently, you might hypothesise that sea ice extent can affect the breeding ecology of these species, either indirectly through an impact on the abundance of their food resources or directly through food resources availability. Therefore, we used sea ice extent as a candidate explanatory variable for trends in arrival and laying dates. Because our phenological data starts in the early 1950s, and sea ice extent data derived from satellite observations are only available from the early 1970s, we used a proxy of sea ice extent recently developed for East Antarctica. Methanesulphonic acid (MSA) is a product of biological activity in surface ocean water whose production is heavily influenced by the presence of sea ice in the Southern Ocean. An ice core from East Antarctica has reported a significant correlation between MSA and satellite-derived sea ice extent, and this calibration applied to longer term MSA data has permitted to reconstruct sea ice extent since the mid-nineteenth century.

The other candidate variable considered here is the Southern Oscillation Index (SOI), which represents the El Niño Southern Oscillation conditions. High, positive values of SOI indicate La Niña conditions and low negative values indicate El Niño conditions. Many studies have shown that the El Niño Southern Oscillation (and therefore SOI) impacts on demographic rates and food resources of many animals, including seabirds. In addition, contrary to the proxy of sea ice extent, SOI is a large scale climate index that may affect seabirds, both during the breeding and non-breeding season.

At present, very little is known about the at sea distribution during the non breeding period of the three seabird species considered here. Anecdotal observations suggest that both penguin species migrate north of their colonies, but remain within Antarctic waters close to the pack ice, and that Cape Petrels migrate in sub-Antarctic and sub-tropical waters. During the breeding period, both penguin species forage within the pack ice up to 150 km from the colonies but nothing is known for the Cape Petrel.

The aim of this case study is to (i) estimate trends in the arrival and laying dates in the three bird species, (ii) analyse the differences between arrival and laying dates, and (iii) determine the effects of possible explanatory variables (e.g. ice cover and the Southern Oscillation Index).

14.2 Data Exploration

Figure 14.2 contains an `xyplot` from the lattice package, showing the patterns over time in arriving and laying dates for the three bird species. To aid visual interpretation, we added a LOESS smoothing curve (Chapter 3) with a span width of 0.5 in each panel. The question addressed in this chapter is whether there is a significant trend in each series. The shape of the LOESS smoothers suggests that something is going on, but there are two main problems for these data. In principle, we have time series data; the timing of arrival in a certain year may depend on the timing in the previous year. The same holds for laying dates. This means that we should take

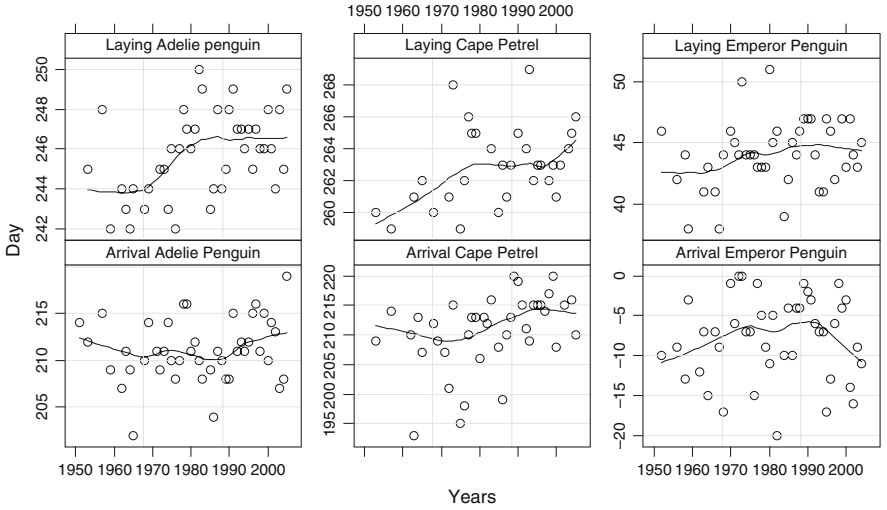

Fig. 14.2 Time series of arrival dates, laying dates, and the difference between laying and arrival dates of three bird species. A LOESS smoother with a span width of 0.5 was added to aid visual interpretation

into account the auto-correlation in the data. Preliminary graphs using the auto-correlation function (Chapter 6) showed that some time series have a significant, albeit weak, auto-correlation with a lag of 1 year. One example is given in Fig. 14.3, which shows that laying dates of the Adelie Penguin in year s is weakly related to those in year $s - 1$ (the auto-correlation with time lag 1 is significantly different from 0 at the 5% level). The other issue with the LOESS smoother is the span width (Chapter 3). If we increase it, we get a less smooth curve and decreasing the span width means that we end up with a more rapidly changing trend.

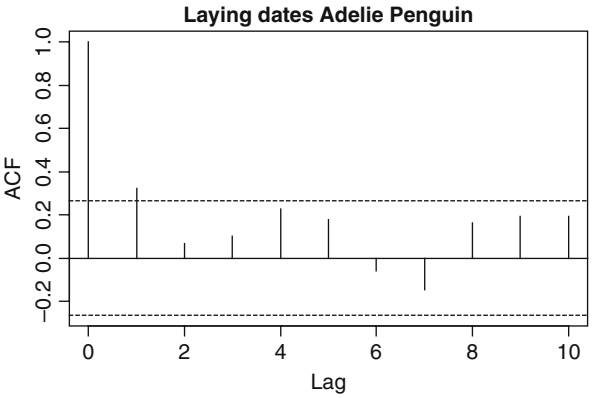

Fig. 14.3 Auto-correlation function of the laying dates of the Adelie Penguin. The horizontal axis shows the time lags and the vertical axis the correlation. The *dotted line* represents the 95% confidence bands

The smoothers in Fig. 14.2 suggest that the laying dates of the Adelie Penguin and Cape Petrel have increased since the mid-1970s. The question is now whether this is indeed the case or whether the smoother is misleading. As explained in Chapter 3, the smoother can be misleading in two ways: (i) by using the wrong amount of smoothing and (ii) by ignoring potential auto-correlation structures. The aim of this chapter is to eliminate these problems.

The following R code was used to generate Fig. 14.2.

```
> library(AED); data(Antarcticbirds)
> ABirds <- Antarcticbirds #saves some space
> library(lattice)
> Birds <- c(ABirds$ArrivalAP, ABirds$LayingAP,
             ABirds$ArrivalCP, ABirds$LayingCP,
             ABirds$ArrivalEP, ABirds$LayingEP)
> AllYears <- rep(ABirds$Year, 6)
> MyNames<-c("Arrival Adelie Penguin",
       "Laying Adelie penguin", "Arrival Cape Petrel",
       "Laying Cape Petrel", "Arrival Emperor Penguin",
       "Laying Emperor Penguin")
> ID1 <- factor(rep(MyNames, each=length(ABirds$Year)),
        levels = c(MyNames[1], MyNames[3], MyNames[5],
                   MyNames[2], MyNames[4], MyNames[6]))
> xyplot(Birds ~ AllYears | ID1,  xlab="Years",
      ylab = "Day", layout = c(3, 2), data = ABirds,
      strip = function(bg = 'white', ...)
      strip.default(bg = 'white', ...),
      scales = list(alternating = TRUE,
                    x = list(relation = "same"),
                    y = list(relation = "free")),
      panel = function(x, y){
       panel.xyplot(x, y, col = 1)
       panel.loess(x, y, col = 1, span = 0.5)
       panel.grid(h = -1, v = 2)})
```

This is a rather intimidating piece of code, but the results in Fig. 14.2 make it worthwhile. Let's go over it step by step. The `library` and `data` commands were discussed in Chapter 2. The ASCII file with the data contains seven columns of data: the year and the arrival and laying times for the three species. Each series contains 55 observations. To make the `xyplot` from the lattice package, we need to store the arrival and laying dates in a single vector of length $6 \times 55 = 330$. We could have done this data editing in a spreadsheet program like Excel, but it is much easier to do this in R. We also need a vector of length 330 that contains the year of each observation. Again, we could have copied and pasted the column year six times under each other in the spreadsheet, but it is much easier in R with the `rep` command.

So, now we have the original six blocks of data in a single column. To let the xyplot function know the identity of the blocks, we made a nominal variable ID1 that contains the names of the variables in the six blocks. The levels option in the factor command was then used to ensure that each time series of the same birds were under each other in the graph. Again, we made use of the rep function. The rest of the code is the same as we used in Chapter 2: we called the xyplot function and specified what should be plotted along the *x* and *y* axes in each panel, labels, etc. The scales option ensured that each panel has a different range along the *y*-axis. Although intimidating, it is all code that we have used before.

Making the auto-correlation function in Fig. 14.3 is actually much easier. The only thing to take care of is the na.action option. Details of the acf function were discussed in Chapter 6.

```
> L.AP <- acf(ABirds$LayingAP, lag.max = 10,
              na.action = na.pass,
              main = "Laying dates Adelie Penguin")
```

Another useful data exploration tool is the pairplot (Fig. 14.4). This addresses whether changes in the arrival and laying dates of the same birds and of different birds are similar. The following code was used to create it.

```
> pairs(ABirds[,2:7], upper.panel = panel.smooth,
        lower.panel = panel.cor)
```

The arguments upper.panel and lower.panel in the pairs command call functions. We took these from the pairs help file in R; see ?pairs. We needed to make some small modifications to the R functions panel.smooth and panel.cor, because we had missing values and preferred black lines for the smoothing lines. Our modified panel functions are available in the AED package. Each panel above the diagonal in Fig. 14.4 shows a scatterplot of two variables. A LOESS smoothing curve with a span of 0.66 was added. The panels below the diagonal contain the (Pearson) correlations coefficients between two variables.

The font size of a correlation is proportional to its value. The panels in the pairplot show that there is no strong correlation between arrival and laying dates of the same and different birds. Pairplots are also useful for identifying outliers and extreme observations, which are not present in these data. So we can avoid transformations.

With the ongoing debates on climate change in mind, it is useful to look at the difference between arrival and laying dates for each bird and its relation to the explanatory variables MSA and SOI. Figure 14.5 shows these differences in arrival and laying for the three bird species plotted against time. We have added the two explanatory variables MSA and SOI. The problem with MSA is that it has a clear trend over time, and it has a relative large number of missing values towards the end of the 1990s. To avoid collinearity problems, it is perhaps better not to use year and MSA as explanatory variables in the same model (the correlation between them

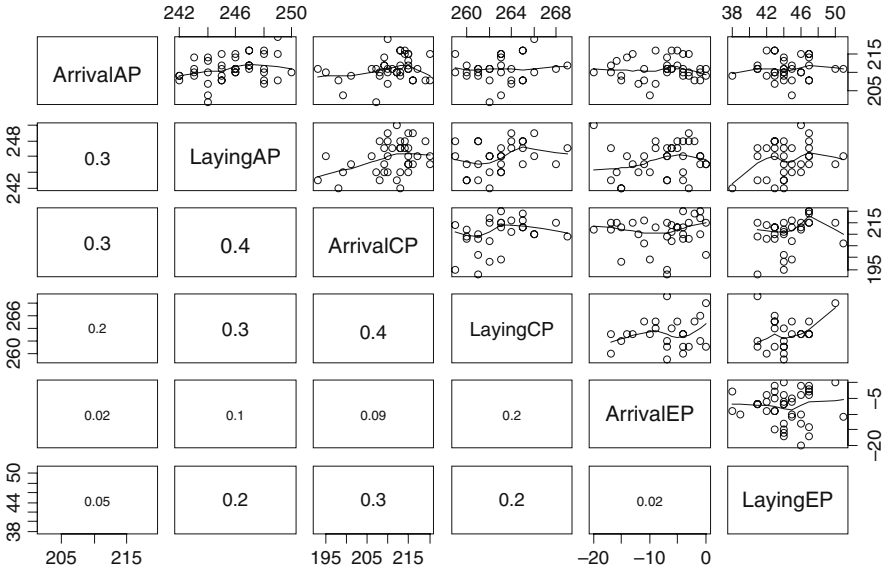

Fig. 14.4 Pairplot for the arrival and laying dates. The *lower panels* show the Pearson correlation coefficients, and the font of the correlation is proportional to its value. The *upper panels* show pair-wise scatterplot and a smoothing curve (LOESS) was added

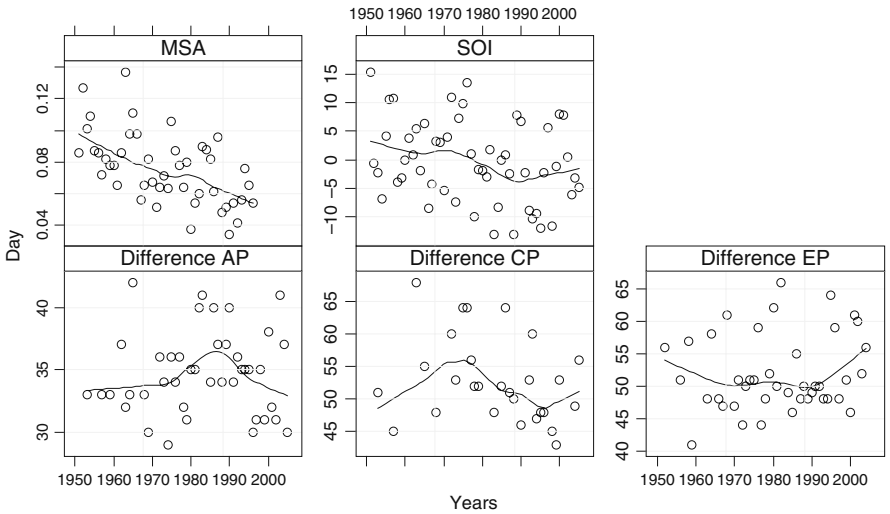

Fig. 14.5 Differences between arrival and laying dates against time for each species. AP, CP, and EP stand for Adelie Penguin, Cape Petrel and Emperor Penguin respectively. The *upper two panels* show the MSA and SOI time series; both are potential explanatory variables

is –0.57). The same holds for MSA and SOI, but now the motivation for not using them together is that the variations in MSA are driven by SOI. There is no clear trend over time in the 'difference time series' for the species.

The R code used to create Fig. 14.5 is given below. It follows the same steps as for Fig. 14.2.

```
> ABirds$DifAP <- ABirds$LayingAP - ABirds$ArrivalAP
> ABirds$DifCP <- ABirds$LayingCP - ABirds$ArrivalCP
> ABirds$DifEP <- ABirds$LayingEP - ABirds$ArrivalEP
> AllDif <- c(ABirds$DifAP, ABirds$DifCP,
              ABirds$DifEP, ABirds$MSA, ABirds$SOI)
> AllYear <- rep(ABirds$Year, 5)
> IDDif <- rep(c("Difference AP", "Difference CP",
             "Difference EP", "MSA", "SOI"), each = 55)
> xyplot(AllDif ~ AllYear | IDDif, xlab = "Years",
      ylab = "Day", layout = c(3, 2),
      strip = function (bg = 'white', ...)
      strip.default(bg = 'white', ...),
      scales = list(alternating = TRUE,
                    x = list(relation = "same"),
                    y = list(relation = "free")),
      panel = function(x, y){
        panel.xyplot(x, y, col = 1)
        panel.loess(x, y, col = 1, span = 0.5)
        panel.grid(h = -1, v = 2)})
```

14.3 Trends and Auto-correlation

The smoothing curves in Fig. 14.2 indicate the presence of long-term trends in some of the arrival and laying time series. However, we quite arbitrarily chose a span width of 0.5 for the LOESS smoother, and choosing a different value may give a different message. To estimate the optimal amount of smoothing we can use cross-validation (Wood, 2006; Chapter 3) and we can also allow for auto-correlation (Chapter 6). We now compare the models with and without auto-correlation and test which one is better. The reason for investigating whether we need a residual auto-correlation structure is that if we falsely omit it, *p*-values may be seriously inflated. Zuur et al. (2007; Chapter 23) used a bird data set in which the model with and without auto-correlation nearly resulted in different conclusions on the importance of different management variables.

Cross-validation and/or adding a residual auto-correlation structure to a smoothing model requires the use of the `gamm` function in the `mgcv` package in R. The model we apply on each time series is

$$\begin{aligned} Y_s &= \alpha + f(\text{Year}_s) + \varepsilon_s \\ \varepsilon_s &\sim N(0, \sigma^2) \\ \text{cor}(\varepsilon_s, \varepsilon_t) &= h(\rho, d(\text{Year}_s, \text{Year}_t)) \end{aligned} \tag{14.1}$$

The first part specifies that the time series is modelled as an intercept plus a smoothing function plus residuals. These residuals are assumed to be normally distributed, but not independent of each other. We allow for a certain dependence structure using the function $h()$, which depends on an unknown parameter ρ and a function $d()$ which is a function of time (or better: the difference between time points). The trick is now to find an appropriate structure for $h()$ and we can compare different forms using the AIC or likelihood ratio tests. As discussed in Chapters 6 and 7, we have a series of options to model the function $h()$. The one of interest here is the auto-regressive moving average (ARMA) serial correlation structure. We could also apply correlation structures from spatial data analysis methods (Chapter 7), which is especially useful if the time series are irregular spaced. However, the time series are regular spaced, albeit with missing values, and therefore we do not need to use any spatial correlation structures.

If the trend is linear, then we can use a model of the form

$$\begin{aligned} Y_s &= \alpha + \beta \times \text{Year}_s + \varepsilon_s \\ \varepsilon_s &\sim N(0, \sigma^2) \\ \text{cor}(\varepsilon_s, \varepsilon_t) &= h(\rho, d(\text{Year}_s, \text{Year}_t)) \end{aligned} \tag{14.2}$$

The first two parts of the equation are the familiar linear regression model. And, if we assume independence of the residuals, it *is* a linear regression model. But we have not made this assumption here, and the last part of Equation (14.2) allows for residual dependence. On the other hand, the cross-validation will help decide whether we need the model in Equation (14.1) or (14.2). Indeed Equation (14.2) is a special case of Equation (14.1), so we might as well focus in first instance only on the model in Equation (14.1). This is because a straight line (linear regression) is a special case of a smoothing curve (Fox, 2000). The model in Equation (14.1) can be fitted in R with the `gamm` function using a Gaussian distribution and identity link (Chapter 3), and Equation (14.2) is fitted using the `gls` function in the `nlme` package. The following R code fits the additive model with an ARMA error structure (Chapter 6) on the arrival dates of the Adelie Penguins.

```
> library(mgcv)
> library(nlme)
> B1 <- gamm(ArrivalAP ~ s(Year), data = ABirds,
     correlation = corARMA(form =~ Year, p = 1, q = 0))
> AIC(B1$lme)
```

The option `corARMA (form =~ Year, p = 1, q = 0)` specifies the auto-regressive residual ARMA structure of order (p, q). The notation for this is

ARMA*(p, q)*. The AIC of this model is 237.83. To choose the optimal ARMA structure, we used all combinations for p and q from 0 to 2. For the combination $p = q = 0$, you need to omit the correlation option. Hence, this is just an ordinary GAM without a correlation structure. To assess which combination of p and q results in the 'best' model, we used the AIC. The lower the AIC, the better the model.

The notation *s(Year)* means that a smoother is applied on Year and cross-validation is used to estimate the optimal amount of smoothing.

This modelling approach was applied on all six arrivals and laying date time series. All six time series gave results where the optimal residual error structure was a ARMA(0,0), meaning that no correlation structure was needed. This means that we are back to using ordinary smoothing (or regression). For all six time series, the amount of smoothing was 1 degree of freedom, meaning that each trend is a straight line. This allows us to apply the linear regression model in Equation (14.2) without the auto-correlation structure. The slope of the trend was only significantly different from 0 for the laying time series of the Adelie Penguin ($p = 0.003$) and for both arrival ($p = 0.009$) and laying ($p = 0.029$) Cape Petrel time series. For the other three series, the slope was not significantly different from zero.

14.4 Using Ice Extent as an Explanatory Variable

In this section, we consider models of the form

$$\begin{aligned} Y_s &= \alpha + f(\text{MSA}_s) + \varepsilon_s \\ \varepsilon_s &\sim N(0, \sigma^2) \\ \text{cor}(\varepsilon_s, \varepsilon_t) &= h(\rho, d(\text{Year}_s, \text{Year}_t)) \end{aligned} \tag{14.3}$$

Y_s is the arrival or laying date in year s and MSA_s is the Methanesulfonic acid concentration (μM) in year s, representing the sea ice extent. Again, we can use cross-validation to estimate the amount of smoothing, and if it turns out that the estimated degrees of freedom is equal to one, we will end up with the model

$$\begin{aligned} Y_s &= \alpha + \beta \times \text{MSA}_s + \varepsilon_s \\ \varepsilon_s &\sim N(0, \sigma^2) \\ \text{cor}(\varepsilon_s, \varepsilon_t) &= h(\rho, d(\text{Year}_s, \text{Year}_t)) \end{aligned} \tag{14.4}$$

As in the previous section, different residual correlation structures can be applied using the correlations option in `gamm` and `gls` and the AIC is used to compare them. For all six arrival and laying time series, the optimal residual correlation structure was ARMA(0,0), which means that no correlation structure is needed. Dropping the correlation structure means we are back in the world of ordinary additive modelling or linear regression, depending on the amount of smoothing. The cross-validation method gave 1 degree of freedom for each series, indicating that we can use the linear regression model in Equation (14.4).

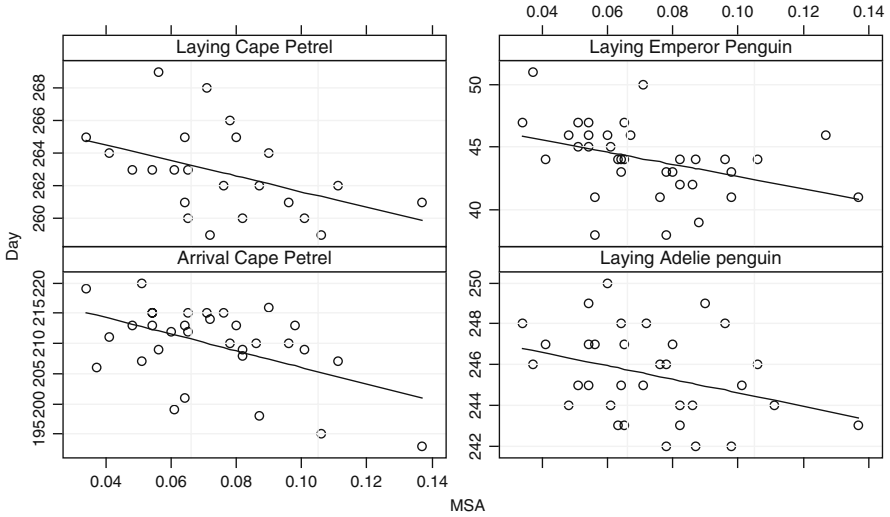

Fig. 14.6 Fitted values obtained by linear regression. Only the time series with a significant slope for MSA are shown. The R^2 for the four series are 12% (Laying Adelie penguin), 19% (Laying Cape Petrel), 15% (Laying Emperor Penguin), and 24% (Arrival Cape Petrel)

The linear regression model showed that MSA has a negative effect on laying dates of all three birds (Adelie Penguin, $p = 0.053$; Cape Petrel, $p = 0.039$; and Emperor Penguin, $p = 0.039$), and also on the arrival date of Cape Petrel ($p = 0.004$). The observed data and fitted lines for these four time series are presented in Fig. 14.6.

The following R code was used for the linear regression models.

```
> M1 <- lm(LayingAP ~ MSA, data = ABirds)
> M2 <- lm(LayingCP ~ MSA, data = ABirds)
> M3 <- lm(LayingEP ~ MSA, data = ABirds)
> M4 <- lm(ArrivalCP ~ MSA, data = ABirds)
> summary(M1); summary(M2)
> summary(M3); summary(M4)
```

This code is just the familiar linear regression and summary commands from Chapter 2 that gives the estimated values, R^2, F-statistic, t-values, and p-values. We have not reproduced all the numerical output from the `summary` commands here. The model fits are presented in the lattice graph (Fig. 14.6) using the following R code.

```
> Bird4 <- c(ABirds$LayingAP, ABirds$LayingCP,
             ABirds$LayingEP, ABirds$ArrivalCP)
> MSA4 <- rep(ABirds$MSA, 4)
```

```
> ID4 <- rep(c("Laying Adelie penguin",
         "Laying Cape Petrel",
         "Laying Emperor Penguin",
         "Arrival Cape Petrel"), each = 55)
> xyplot(Bird4 ~ MSA4 | ID4, xlab = "MSA",
     ylab = "Day", layout = c(2, 2),
     strip = function(bg = 'white', ...)
     strip.default(bg = 'white', ...),
     scales = list(alternating = TRUE,
                   x = list(relation = "same"),
                   y = list(relation = "free")),
     panel = function(x, y, subscripts, ...){
       panel.xyplot(x, y, col = 1)
       panel.grid(h = -1, v = 2)
       I1 <- !is.na(y) & !is.na(x)
       tmp <- lm(y[I1] ~ x[I1])
       x1 <- x[I1]
       y1 <- fitted(tmp)
       I2 <- order(x1)
       panel.lines(x1[I2], y1[I2], col = 1, span = 1)})
```

The first three commands create three variables containing the stacked observed data, names, and MSA values for the four series. The code for the `xyplot` should now look familiar, except perhaps applying the linear regression within the `xyplot` function. The only thing to watch for is ensuring that we deal correctly with the missing values in the data. The command `I1 <- !is.na(y) & !is.na(x)` identifies the observations for which we have an observation for both the response and explanatory variables. The linear regression is applied on these data, and the order command is used to avoid a spaghetti plot (Chapter 2).

The main problem using MSA as an explanatory variable is that we lose 16% of the data due to missing values. Recall from Chapter 2 that the *entire* row of data is omitted, even if only one variable has a missing value for that observation.

14.5 SOI and Differences Between Arrival and Laying Dates

For the last analysis in this chapter, we use SOI as an explanatory variable. A slightly different statistical approach is followed and the arrival and laying dates for a bird are analysed simultaneously. Using an interaction term, we can use this approach to make a statement on the difference of the SOI effect on arrival and laying dates. This approach is potentially invalid, but we discuss at the end of this section how to correct for this.

A simple boxplot (not presented here) shows that the variation in arrival dates is considerably larger than for laying dates for all three bird species. This means that an ordinary linear regression model (or additive model) applied on the combined

arrival and laying dates is likely to violate the homogeneity assumption. On top of this, there may be auto-correlation. An additive model applied on the individual time series using SOI as an explanatory variable showed that all trends were linear, and therefore, we will work with a linear regression model of the form:

$$\begin{aligned} &Y_{sj} = \alpha + \beta_1 \times \text{SOI}_s + \beta_2 \times ID_j + \beta_3 \times SOI_s ID_j + \varepsilon_s \\ &\varepsilon_s \sim N(0, \sigma_j^2) \\ &\text{cor}(\varepsilon_s, \varepsilon_t) = h(\rho, d(\text{Year}_s, \text{Year}_t)) \end{aligned} \tag{14.5}$$

Y_{sj} is the arrival ($j = 1$) or laying ($j = 2$) date of a particular species in year s. In R, we stack the arrival and laying dates into one vector of length 110. This vector is then modelled as an intercept plus a function of SOI, a nominal variable ID (arrival or laying), and an interaction between SOI and ID. In matrix notation we have

$$\begin{pmatrix} \text{Arrival}_1 \\ \vdots \\ \text{Arrival}_{55} \\ \text{Laying}_1 \\ \vdots \\ \text{Laying}_{55} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ 1 \end{pmatrix} \alpha + \begin{pmatrix} \text{SOI}_1 \\ \vdots \\ \text{SOI}_{55} \\ \text{SOI}_1 \\ \vdots \\ \text{SOI}_{55} \end{pmatrix} \beta_1 + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \beta_2 + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \text{SOI}_1 \\ \vdots \\ \text{SOI}_{55} \end{pmatrix} \beta_3 + \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,55} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,55} \end{pmatrix}$$

If we assume independence and homogeneity of the residuals, then this is the ordinary linear regression model with one continuous variable, one nominal variable and the interaction between them (also known as analysis of covariance, abbreviated as ANCOVA). However, we do not assume independence or homogeneity. And, as discussed in Chapter 4, we can use GLS estimation to estimate multiple variance terms, in this case ${\sigma_1}^2$ for the arrival dates and ${\sigma_2}^2$ for the laying dates. The question is then whether $\sigma_1 = \sigma_2$, or whether we indeed need two different variances.

As to the auto-correlation structure, for simplicity, we only consider the ARMA(1,0) structure. This means that the residual correlation structure takes the form (Pinheiro and Bates, 2000; Chapter 6):

$$\text{cor}(\varepsilon_s, \varepsilon_t) = \rho^{|s-t|} \tag{14.6}$$

The parameter ρ is between –1 and 1. This auto-correlation structure dictates that the larger the time period between two years, the smaller the dependence between them. The following R code applies the model in Equation (14.5) with the auto-correlation in Equation (14.6).

```
> AP <- c(ABirds$ArrivalAP, ABirds$LayingAP)
> SOI2 <- c(ABirds$SOI, ABirds$SOI)
> Y2 <- c(ABirds$Year, ABirds$Year)
> ID <- factor(rep(c("Arrival", "Laying"), each = 55))
```

```
> library(nlme)
> vf2 <- varIdent(form =~ 1 | ID)
> M5 <- gls(AP ~ SOI2 + ID + SOI2:ID, weights = vf2,
            na.action = na.omit)
> M6 <- gls(AP ~ SOI2 + ID + SOI2:ID, weights = vf2,
             na.action = na.omit,
             correlation = corAR1(form =~ Y2 | ID))
> anova(M5, M6)
```

The first command concatenates the arrival and laying dates time series of the Adelie Penguin. The second command creates a vector with corresponding SOI values, and Y2 contains the years in which an observation was taken. Finally, ID is a nominal variable identifying the two time series. The `library` command ensures we can access the `gls` function, which is needed to fit the model in Equation (14.5) in R. The `varIdent` function was discussed in Chapter 4 and allows for a different variance for each of the bird time series. We first call the `gls` function and fit a model without auto-correlation. Then we fit a model with the auto-correlation structure as specified in Equation (14.6) and store its results in `M6`. The `anova(M5, M6)` command applies a likelihood ratio test and gives

```
Model df      AIC        BIC   logLik  L.Ratio  p-value
M5     6   427.82     442.26  -207.91
M6     7   426.07     442.92  -206.03    3.744     0.05
```

These results show that there is a weak residual auto-correlation structure ($p = 0.05$), and we should keep it in the model. We can also compare a model with one variance and a model with two variances. The likelihood ratio test (not presented here) indicates that we should use two variances ($p = 0.002$). The results of the following code show that we do not need the interaction term between ID and SOI.

```
> M7 <- gls(AP ~ SOI2 + ID + SOI2:ID, weights = vf2,
            na.action = na.omit, method = "ML",
            correlation = corAR1(form =~ Y2 | ID))
> M8 <- gls(AP ~ SOI2 + ID, weights = vf2,
            na.action = na.omit, method = "ML",
            correlation = corAR1(form =~ Y2 | ID))
> anova(M7, M8)
```

As discussed in Chapter 4, if we compare two models with the same random structure, but with different fixed effect, we need to use the maximum likelihood estimation method instead of REML. The resulting test statistic is obtained by the `anova(M7, M8)` command, and it gave a test statistic $L = 0.16$ ($df = 1, p = 0.68$). This indicates that we can drop the interaction term as it is not significant.

In the model with SOI2 and ID as explanatory variables (`M8`), the `summary(M8)` command shows that only ID is significant, meaning that there is no SOI effect on Adelie Penguins. Hence, the optimal model is given by

```
> M9 <- gls(AP ~ ID, weights = vf2, method = "ML",
            na.action = na.omit,
            correlation = corAR1(form =~ Y2 | ID))
> summary(M9)
```

Its numerical output is given by

```
Correlation Structure: ARMA(1,0)
 Formula: ~Y2 | ID
 Parameter estimate(s):
     Phi1
0.26
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | ID
 Parameter estimates:
  Arrival    Laying
  1.00       0.61
Coefficients:
             Value  Std.Error  t-value  p-value
(Intercept)  211.11   0.64       329.26     <0.001
IDLaying     34.56    0.75       45.63      <0.001

Residual standard error: 3.361275
Degrees of freedom: 86 total; 84 residual
```

This output shows that the predicted arrival time for Adelie Penguin is day 211 (rounded), and the laying date is $211 + 34 = 245$. There is no effect of SOI on either arrival or laying dates. The residual standard error for the arrival dates is 3.36, but for the laying dates it is 0.6 smaller. There is also a small amount of auto-correlation as $\rho = 0.26$. This means that the residual auto-correlation between two sequential years is equal to $0.26^1 = 0.26$, and for time points that are separated by 2 years, this correlation is $0.26^2 = 0.07$.

The same analysis was carried out on the Cape Petrel time series. Using the same code, but with the first line replaced by `CP <- c(ArrivalCP, LayingCP)` and consequently AP by CP. For the Cape Petrel, we found that different variances per arrival and laying series are needed, but there was no need for an auto-correlation structure. The interaction term had a p-value of 0.09, but we decided to keep it in as it was close to the 'magic' significance level of 0.05. The SOI effect and the nominal variable ID were highly significant. The following R code was used.

```
> CP <- c(ABirds$ArrivalCP, ABirds$LayingCP)
> SOI2 <- c(ABirds$SOI, ABirds$SOI)
> Y2 <- c(ABirds$Year, ABirds$Year)
> ID <- factor(rep(c("Arrival", "Laying"),each = 55))
```

```
> vf2 <- varIdent(form= ~ 1|ID)
> M10<-gls(CP ~ SOI2 + ID + SOI2:ID, weights = vf2,
           na.action = na.omit, method = "ML")
```

The results below obtained by the summary(M10) command. Note that the spread for the laying dates is considerably lower!

```
Structure: Different standard deviations per stratum
 Formula: ~1 | ID
 Parameter estimates:
  Arrival    Laying
  1.00       0.37
Coefficients:
               Value   Std.Error   t-value  p-value
(Intercept)   210.41  0.92        226.62   0.00
SOI2          -0.39   0.13        -2.99    0.00
IDLaying      52.37   1.01        51.76    0.00
SOI2:IDLaying 0.24    0.14        1.71     0.09
```

The lines of R code below produce a scatterplot of the data, and the fitted lines obtained by the optimal model; see Fig. 14.7.

```
> plot(ABirds$SOI,  ABirds$ArrivalCP,
       ylim = c(195, 270), type = "n",
       ylab = "Arrival & laying dates")
> points(ABirds$SOI, ABirds$ArrivalCP, pch = 1)
> points(ABirds$SOI, ABirds$LayingCP, pch = 2)
> MyX <- data.frame(SOI2 = seq(from = min(ABirds$SOI),
                to = max(ABirds$SOI), length = 20),
                ID = "Arrival")
> Pred1 <- predict(M10, newdata = MyX)
> lines(MyX$SOI2, Pred1)
> MyX <- data.frame(SOI2 = seq(from = min(ABirds$SOI),
                             to = max(ABirds$SOI),
                             length = 20),
                 ID = "Laying")
> Pred2 <- predict(M10, newdata = MyX)
> lines(MyX$SOI2, Pred2)
```

For the emperor penguin, we found strong evidence to use two variance terms ($p < 0.001$) and weak evidence for an auto-correlation structure ($p = 0.07$). However, neither the interaction nor the SOI was significant. The output of the final model is given on the next page and shows there is some auto-correlation, the residual variation in laying dates is nearly half of the arrival time residual variation, and the difference between arrival and laying dates is 52 days.

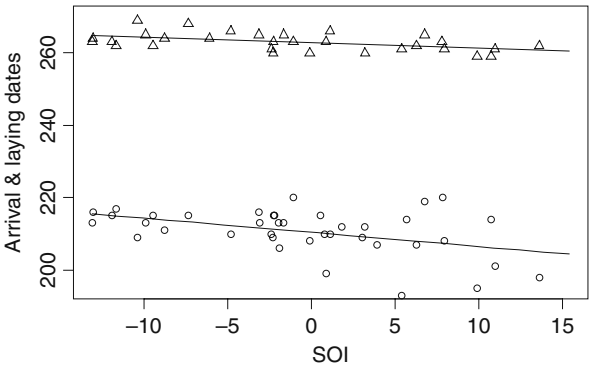

Fig. 14.7 Arrival and laying dates for the Cape Petrel. The *triangles* are the laying dates, and the *dots* the arrival dates. Due to the weak interaction, the *lines* are nearly parallel indicating that there are no strong differences between the SOI-date relationship for arrival and laying

```
Correlation Structure: ARMA(1,0)
 Formula: ~Y2 | ID
 Parameter estimate(s):
     Phi1
0.20
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | ID
 Parameter estimates:
 Arrival    Laying
1.00         0.53
Coefficients:
               Value   Std.Error  t-value  p-value
(Intercept)    -7.91   0.91       -8.62    <0.001
IDLaying       51.94   1.04       49.58    <0.001
```

At the start of this section, we mentioned that the analysis presented in this section is potentially invalid. This would happen if the residuals of the laying time series are correlated to the residuals of the arrival time series. We allowed for correlation within a series, but not between a series. In Chapter 6, we applied a similar analysis on a bird time series from Hawaii. However, in that example, because sampling was done during the breeding season on different islands, the between-series independence assumption is more plausible than it is for the time series used in this chapter. The easiest way to verify the independence assumption is to calculate the correlation between the two residual time series per species. If the correlation is not significant, we are lucky and the approach in this section is valid. In this case, the correlation is 0.29 and the associated p-value is 0.06. At the 5% level, it is not significant, but we are still not that happy with a p-value so close to the magic 5% level. Obtaining this correlation coefficient requires some rather tedious R programming (due to missing values), and the code is on the book's website.

14.6 Discussion

Wrongly ignoring dependence structures in data means a greater chance of type I errors. So, with time series data, we should always check for residual auto-correlation, and only where there is no significant auto-correlation, use methods that ignore auto-correlation. For the arrival and laying time series, there was no strong residual auto-correlation; hence, one can proceed with methods that do not include an auto-correlation structure (e.g. linear regression, additive modelling).

For all bird time series, linear regression was favoured above smoothing techniques. As well as smoothing methods, we also tried other models that allow for non-linear trends (e.g. quadratic models using Year and Year^2 as explanatory variables), but they confirmed the cross-validation results.

A detailed model validation consisting of plots of (normalised) residuals versus fitted values, auto-correlation plots, histograms, and plotting residuals versus explanatory variables was applied in each section. For nearly all models, there were no problems. We also tried models that contained both year and SOI, but these did not improve the models.

The ecological interpretation of these results shows there is a positive trend in the laying time series of the Adelie Penguin and the arrival and laying of the Cape Petrel series. This may be related to the MSA time series, but unfortunately using this variable means we lose 16% of the data. It should be noted that MSA itself is negatively related to time; there is a clear decreasing trend in MSA (Fig. 14.5). The negative relationships between MSA, laying, and arrival dates indicate that birds arrive and lay earlier when sea ice extent increases. This fits well with our knowledge of the reproductive ecology of these species because they need to build up body reserves (fat) before breeding. So, in years with extensive sea ice, food might be more abundant allowing birds to build up fat reserves quicker than years with less sea ice. Consequently, the negative trend in MSA may explain part of the positive trends in arrival and laying dates, although MSA explained at most 24% of the variability in arrival and laying dates. Other factors such as the duration of the sea season or individual characteristics such as age or experience may explain part of the remaining variability. The analysis using the combined arrival and laying dates for a species and SOI as explanatory variable showed there was a large difference in spread in arrival and laying dates for all species and that SOI has a negative effect on arrival and laying dates of Cape Petrel. We also applied the same analysis using year as the explanatory variable instead of SOI. We have not presented the results here, but they showed that the Adelie Penguin had a weak auto-correlation and large spread in arrival and laying dates (the ratio between the standard errors was 0.57) and a significant ID and year effect, but no interaction. This means that arrival and laying dates have increased over time and at the same rate. We found similar results for the Cape Petrel, except there was no auto-correlation. For the Emperor Penguin, there was only an ID effect and weak auto-correlation, but no year effect.

It is not surprising to find a large difference in spread in arrival and laying dates across the species. Unlike arrival dates, laying is closely synchronised in Antarctic seabird populations, meaning that within a population all individuals lay their egg in

a short time window every year. Therefore, laying date is probably less affected by factors such as age, experience, sex or meteorological factors than arrival date. The fact that arrival and laying dates have increased over time at the same rate for the Adelie Penguin and the Cape Petrel over such a long period suggests that the time interval separating those phenological events is relatively inflexible. This probably reflects the invariance in the timing of physiological mechanisms involved during the egg development process and to a lesser extent the time needed for birds to build their nest, courtship and pair.

Finally, an interesting result is the negative effect of SOI on arrival and laying dates of the Cape Petrel. The negative slope indicates that El Niño conditions (negative SOI) delay arrival and laying of Cape Petrels. Because Cape Petrels spend the non-breeding season at more northerly latitudes than Adelie and Emperor Penguins, they might be more affected by El Niño conditions. This result is also in accordance with previous studies on seabirds that have demonstrated that El Niño conditions usually cause a decrease in oceanic productivity and of seabird demographic parameters.

14.7 What to Report in a Paper

If we were to write a paper based on the analysis presented in this chapter, we would present an introduction describing the questions and data. We would then continue presenting the data in a multiple panel graph (e.g. Fig. 14.2), present the models (additive model and linear regression) and put emphasise on the potential auto-correlation problem. The fact that the ARMA(0,0) error structure was the most optimal structure can be presented without too much numerical output, and the same holds for the cross-validation. However, you cannot omit this information as they justify the application of the linear regression or smoothing model without auto-correlation.

Describing the approach and summarising the results that justify the linear regression model with an independent error structure does not have to be any longer than two paragraphs. The results of the linear models should be presented in a table showing the estimated parameters, t-values, p-values, and R^2 and F-statistic for all time series. You should also comment on the results of a model validation for the linear regression. (Did the residuals show any patterns in terms of homogeneity, normality, and residuals values versus year, and residuals values versus explanatory variables?) As the linear regression model was preferred over the additive model, state that non-linear patterns are unlikely to occur. The model formulation for the combined arrival and laying time series and the analysis followed may be confusing for the reader, and you may want to explain this aspect in more detail. As to what these results tell us about climate change is left for you to decide.