16 Time series analysis — Introduction

What makes a time series a time series? The answer to this question is simple, if a particular variable is measured repeatedly over time, we have a time series. It is a misconception to believe that most of the statistical methods discussed earlier in this book cannot be applied on time series. Provided the appropriate steps are made, one can easily apply linear regression or additive modelling on time series. The same holds for principal component analysis or redundancy analysis. The real problem is obtaining correct standard errors, t-values, p-values and F-statistics in linear regression (and related methods), and applying the appropriate permutation methods in RDA to obtain p-values. In this chapter, we show how to use some of the methods discussed earlier in this book. For example, generalised least squares (GLS) applied on time series data works like linear regression except that it takes into account auto-correlation structures in the data. We also discuss a standard time series method, namely auto-regressive integrated moving average models with exogenous variables (ARIMAX). In Chapter 17, more specialised methods to estimate common trends are introduced.

Questions in many time series analysis studies are as follows: what is going on, is there a trend, are there common trends, what is the influence of explanatory variables, do the time series interact, is there a sudden change in the time series, can we predict future values and are there cyclic patterns in the data? A series of different techniques are presented in this (and the next) chapter to address these questions.

16.1 Using what we have already seen before

Most of the material that has been presented in previous chapters can easily be adapted or modified so that it can be applied on time series. In Chapter 4, it was shown how time (season) can be used as a conditional variable in boxplots for the Argentinean zoobenthic data. In Chapter 4, we used the Gonadosomatic index (GSI) for a squid dataset. The boxplot in Figure 16.1 shows that there is a clear monthly pattern in the GSI index. Figure 16.2 shows how lattice graphs can be used to visualise a North American sea surface temperature (SST) time series. Mendelssohn and Schwing (2002) used the Comprehensive Ocean–Atmosphere Dataset (COADS) to generate monthly sea surface temperature time series from various grid points. COADS is a database containing data of oceanographic parameters taken mainly by 'ships-of-opportunity'. Mendelssohn and Schwing

(2002) defined 15 two-degree latitude by two-degree longitude regions 'based on a combination of ecological and oceanographic features, and data density'. The regions cover a large part of the American west coast. For each region, monthly SST time series were extracted from COADS covering the time period 1946 to 1990. The SST time series used in this book are the same as in Mendelssohn and Schwing (2002), and the interested reader is referred to their paper for further details. Figure 16.2 suggests the presence of a monthly effect, trends and differences in absolute values. Another way of visually inspecting the time series is plotting them all in one graph. Standardising the variables to ensure they all have the same range and mean value may help visual interpretation.

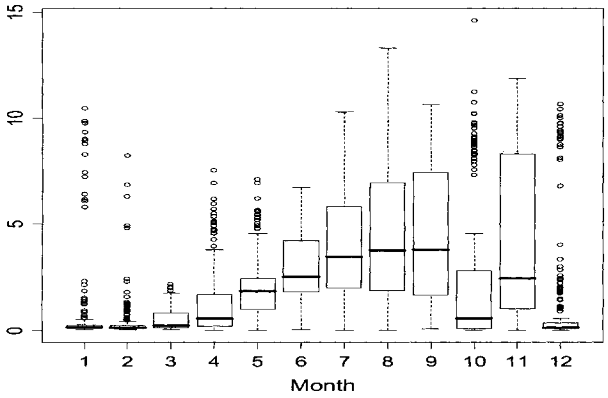

Figure 16.1. Boxplot of GSI index for squid data conditional on time. Months are represented by the numbers 1 (January) to 12 (December).

It is also interesting to enumerate all correlations or cross-correlations larger (in absolute sense) than a certain threshold value. Cross-correlations are correlations (Chapter 10) between two time series with a certain time lag *k*. How to calculate it, is discussed later in this section. Some statistical methods may produce errors if collinear variables are used. For the North American SST data, various correlations are higher than 0.9; see Table 16.1. The problem is that the high correlations are mainly due to the month effects in the time series. It may be useful to remove the month effect and look at correlations between the trends, or between deseasonalised time series (the difference between these two will be explained later in this chapter).

Figure 16.3 shows the annual time series of Nephrops catch per unit effort (CPUE) measured in 11 areas South of Iceland in the Atlantic Ocean between 1960 and 2002 (Eiríksson 1999). Most of the time series follow a similar pattern over time. The second most important step in a time series analysis is to look at auto-and cross-correlations. Carrying out this step, together with plotting the series, should give a first impression of what the data are showing.

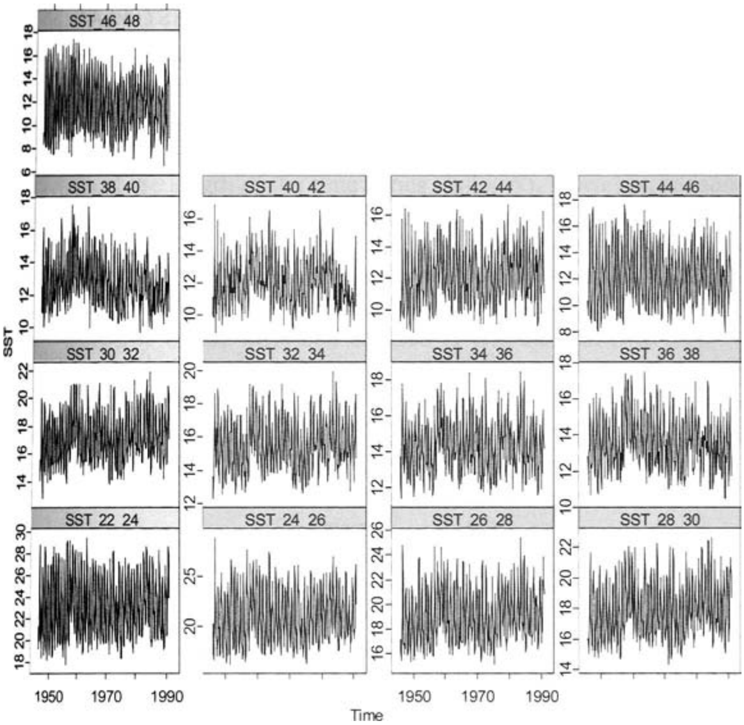

Figure 16.2. North American SST time series. The labels refer to the location.

Auto-correlation and cross-correlation

Define Y_t as the CPUE at a particular station, say station 1, in year t. The question we now address is whether there are any temporal relationships at this station. The auto-correlation function gives an indication of the amount of association between the variables Y_t and $Y_{t+k,}$ where the time lag k takes the values 1, 2, 3, etc. Hence, for $k = 1$, the auto-correlation shows the relationship between Nephrops in year t and year $t + 1$ at station 1. Data of all years are used to calculate this relationship. Formulated differently, the auto-correlation with a time lag of k years represents the overall association between time points that are k years separated.

The Pearson correlation coefficient (Chapter 10) is used to quantify this association. It is always between −1 and 1, and it is calculated (hence it is the sample auto-correlation) by (Chatfield 2003)

$$\hat{\rho}(k) = cor(Y_t, Y_{t+k}) = \frac{1}{N}\frac{\sum_{t=1}^{N-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{s_Y^2}$$

Table 16.1. Correlations larger (in absolute sense) than 0.9 for the North American SST data. The numbers 1 to 13 refer to SST-24-26, SST-26-28, SST-28-30, SST-30-32, SST-32-34, SST-34-36, SST-36-38, SST-38-40, SST-40-42, SST-42-44, SST-44-46 and SST-46-48, respectively. The numbers in the table are the variables and the associated correlation.

Variable 1				Variable 6			
	1	2	0.980		6	4	0.932
	1	3	0.946		6	5	0.941
	1	4	0.905		6	7	0.931
Variable 2				Variable 7			
	2	1	0.980		7	6	0.931
	2	3	0.973		7	8	0.932
	2	4	0.921	Variable 8			
Variable 3					8	7	0.932
	3	1	0.946		8	9	0.930
	3	2	0.973	Variable 9			
	3	4	0.961		9	8	0.930
	3	5	0.915	Variable 11			
Variable 4					11	12	0.941
	4	1	0.905		11	13	0.920
	4	2	0.921	Variable 12			
	4	3	0.961		12	11	0.941
	4	5	0.969		12	13	0.967
	4	6	0.932	Variable 13			
Variable 5					13	11	0.920
	5	3	0.915		13	12	0.967
	5	4	0.969				
	5	6	0.941				

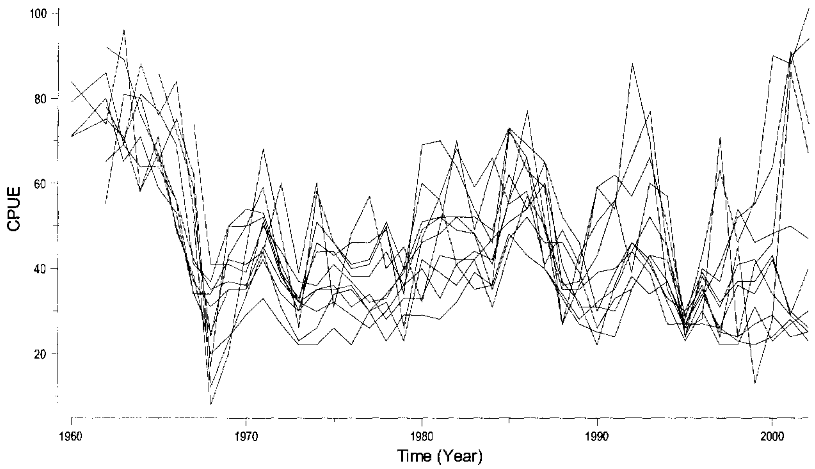

Figure 16.3. Nephrops CPUE time series.

The auto-correlation function is simply the Pearson correlation of a time series with itself, after applying a shift (lag) of k years. The term s_Y is the sample standard deviation of the time series Y_t. The notation $\hat{\rho}$ is used to indicate that we are working with the *sample* auto-correlation. One problem with the auto-correlation is that for larger time lags k, fewer points are used to calculate it. Therefore it is better to limit interpretation of the auto-correlation to the first few time lags only. But how many time lags is 'a few'? Most software packages produce time lags up to around 40% of the length of the time series (0.4 × N). However, for some datasets this may already be too large. We advise to think first about the question 'what time lags would be sensible for the data?' before calculating it. Figure 16.4 shows the auto-correlation for the Nephrops CPUE time series at station 1. The horizontal axis shows time lags, and the corresponding correlation can be read off from the y-axis. The dotted lines can be used to test the null hypothesis that the correlation is equal to 0 for a certain time lag. These dotted lines are obtained from $\pm 1.96/\sqrt{N-k}$, where N is the length of a time series. Note that by chance alone, 1 out of 20 auto-correlations may be significant, whereas in reality they are not (type I error). For station 1 there is a significant auto-correlation with time lags of 1, 2 and 3 years. This means that if the CPUE is high in year t, it is also high in year $t-1$, t – 2 and $t-3$.

The auto-correlation function is a useful tool to infer the pattern in the time series for which it was calculated. If a time series shows a strong seasonal pattern with a period of 12 months, then its values are likely to be high at time t, low at time t+6 and high again at time t+12. The same holds for $t+18$ and $t+24$, $t+30$, $t+48$, etc. The auto-correlation for such a time series shows large negative correlations for $k = 6$, $k = 18$, $k = 30$, etc., and large positive correlations for $k = 12$, $k = 24$, etc. Hence, a monthly pattern results in an oscillating auto-correlation function. The same holds for cyclic patterns. The distinction between a seasonal and cyclic component is that in the former we know the length of the cyclic period, whereas in the latter it is unknown. The auto-correlation function can then be used to get some idea on the length of the cycle. Another scenario is that the time series shows a slow-moving pattern. In this case, if Y_t is above the average, then so is Y_{t+1}. Hence, the auto-correlation with a time lag of $k = 1$ is large. The same holds for other small values of k. For larger values of k, one would expect that if Y_t is above average, then Y_{t+k} would be below average (or vice versa); hence we would obtain a negative correlation. Therefore, an auto-correlation that shows a slow decrease may indicate a trend. For example, Figure 16.4 is the auto-correlation function for the CPUE Nephrops time series at station 1. The shape of the function indicates the presence of a trend.

Now let us assume that Y_t and X_{t-k} are the CPUE values at station 1 in year t and at station 2 in year $t-k$ respectively. The question we address is whether there is any association between CPUE values at these two stations. Do the two stations have high CPUE at the same time, or has station 1 high CPUE values in year t, and station 2 in year $t-k$? Or is it just the other way around: Station 1 has high CPUE values in year t, and station 2 low values in year $t-k$? In this case, it makes sense to compare cross-correlations between Y_t and X_{t-k} for positive and negative values of k. But what if X_{t-k} is the temperature in year $t - k$? Obviously, there is not much point in comparing the CPUE in year t with temperature in year $t + 1$.

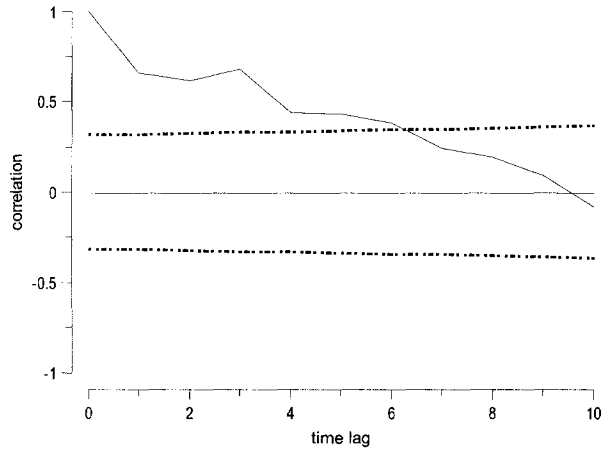

Figure 16.4. Auto-correlation for the CPUE Nephrops time series at station 1.

The cross-correlation function quantifies the association between two variables with a time lag of k years. It is again based on the Pearson correlation function, except that the second variable is shifted in time with a lag of k years. The (sample) cross-correlation is calculated by (Chatfield 2003; Diggle 1990)

$$\hat{\rho}_{YX}(k) = \begin{cases} \dfrac{1}{N}\dfrac{\sum_{t=1}^{N-k}(Y_t - \bar{Y})(X_{t+k} - \bar{X})}{s_Y s_X} & \text{if } k \geq 0 \\[2ex] \dfrac{1}{N}\dfrac{\sum_{t=1-k}^{N}(Y_t - \bar{Y})(X_{t+k} - \bar{X})}{s_Y s_X} & \text{if } k < 0 \end{cases}$$

This is the same mathematical formula as the auto-correlation except that the second bit of the formula now contains the variable X. The terms s_Y and s_X are sample standard deviations of the time series Y_t and X_t respectively. The cross-correlation can be calculated for various time lags, and the results can be plotted in a graph in which various time lags (positive and negative) are plotted along the horizontal axis and the correlations along the vertical axis. The cross-correlations for stations 1 and 2 are presented in Figure 16.5. The dotted lines can be used to test the null hypothesis that the correlation between Y_t and X_{t-k} is equal to 0. Points outside this interval are significant (at the 5% level) cross-correlations with time lag k. The same warning as for the auto-correlation holds: 1 out of 20 cross-correlations may be significant by chance only. There is a significant cross-correlation between the CPUE at station 1 and 2 at the time lags of −3, −2, .., 3.

This may indicate the presence of a slow-moving trend in both series (the same interpretation as for the auto-correlation function applies here). Table 16.2 shows the cross-correlations between CPUE time series of all 11 stations. Estimated cross-correlations in bold typeface are significantly different from 0 at the 5% significance level (Diggle 1990). Note that most correlations are significant!

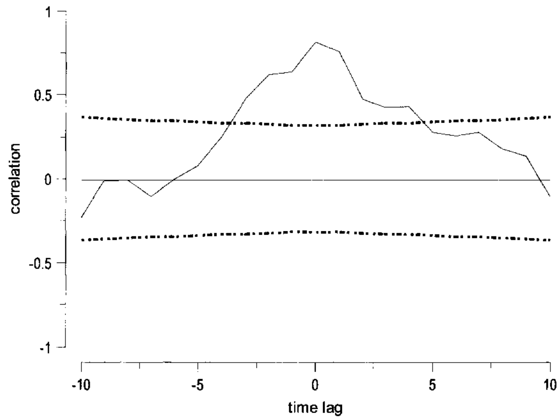

Figure 16.5. Cross-correlations for stations 1 and 2 plotted versus time lags for the CPUE Nephrops time series. Dotted lines represent the 95% upper and lower confidence bands.

Table 16.2. Correlations among 11 Nephrops time series. Values in bold typeface are significantly different from 0 at the 5% level. The numbers 1 to 11 correspond to the stations.

	2	3	4	5	6	7	8	9	10	11
1	**0.82**	**0.87**	0.29	**0.84**	**0.79**	**0.68**	**0.36**	0.28	0.08	0.27
2	1	**0.85**	**0.36**	**0.80**	**0.73**	**0.78**	**0.37**	**0.31**	0.04	0.22
3		1	**0.36**	**0.88**	**0.80**	**0.80**	**0.43**	**0.38**	0.12	0.26
4			1.00	**0.35**	**0.37**	0.22	0.09	0.14	0.11	0.24
5				1.00	**0.88**	**0.77**	**0.37**	**0.34**	0.04	0.24
6					1.00	**0.78**	**0.50**	**0.45**	0.12	**0.37**
7						1.00	**0.71**	**0.65**	0.22	**0.42**
8							1.00	**0.83**	**0.50**	**0.54**
9								1.00	**0.69**	**0.49**
10									1.00	**0.55**
11										1.00

It is also interesting to know at which time lag the maximum cross-correlation between two time series was obtained. This information is given in Table 16.3. The upper-diagonal shows the maximum cross-correlations between each of the two combinations of CPUE time series, and the lower diagonal elements show the corresponding time lag k. For example, the maximum cross-correlation between stations 1 and 2 was obtained at a time lag of 0 years. And the correlation between station 4 and station 7 had the highest value (in the absolute sense) for a time lag of $k = 2$. One can now look at the numbers in Table 16.2 (which are correlations) or Table 16.3 (which are cross-correlations) and try to find a pattern in them. Alternatively, one can visualise these association matrices using tools we have already discussed, for example multidimensional scaling (MDS). As we saw in Chapter 15, MDS can be used to graphically represent a matrix of dissimilarities. All we have to do is convert the measures of association (correlation) into a measure of dissimilarity; see Chapters 10 and 15. We leave it as an exercise for the reader to carry out this analysis.

Table 16.3. Maximum cross-correlations over time lags for the Neprophs time series data. The lower triangular part shows the time lags for which the maximum value was obtained. The numbers 1 to 11 correspond to the 11 stations. The range of the time lags was set to 25% of the length of the series.

	1	2	3	4	5	6	7	8	9	10	11
1		0.82	0.87	0.34	0.84	0.79	0.68	0.40	0.29	−0.29	0.27
2	0		0.85	0.42	0.80	0.73	0.78	0.37	0.31	−0.27	0.25
3	0	0		0.39	0.88	0.80	0.80	0.43	0.38	−0.30	0.32
4	4	−4	−1		0.41	0.44	0.45	0.31	0.36	0.25	0.56
5	0	0	0	1		0.88	0.77	0.39	0.34	−0.17	0.24
6	0	0	0	2	0		0.78	0.51	0.45	0.25	0.37
7	0	0	0	2	0	0		0.71	0.65	−0.25	0.42
8	−1	0	0	2	−1	−1	0		0.83	0.50	0.54
9	1	0	0	2	0	0	0	0		0.69	0.52
10	−3	−5	−4	1	−4	1	−4	0	0		0.55
11	0	−2	1	2	0	0	0	0	−1	0	

The portmanteau test

Instead of looking at individual time lags of the auto-correlation, it is also possible to apply a test that combines a number of time lags. The portmanteau test aggregates the first K time lags of the auto-correlation:

$$Q = N\sum_{j=1}^{K}\hat{\rho}(j)^2$$

where N is the length of the time series. If the first K (sample) auto-correlations are relatively small, Q will be small as well. Q is also called the Box–Pierce statistic (Box and Pierce 1970), and it follows a Chi-square distribution with K degrees of freedom. If the statistic is applied on residuals obtained by a model with p regression parameters, the degrees of freedom is $K - p$. As to the value of K, common values are $K = 15$, $K = 20$ and some packages give Q for a range of different K values.

A slightly 'better' statistic is the Ljung–Box (Ljung and Box 1978) statistic:

$$Q^* = N(N+2)\sum_{j=1}^{K} \frac{1}{N-j} \hat{\rho}(j)^2$$

Hence, this test takes the first K auto-correlations of the residuals, squares them, and adds them all up using weighting factors $T-j$. The statistic is typically presented for various values of K, and for $K > 15$ it can be shown that Q^* is Chi-square distributed with K degrees of freedom.

Large values of Q or Q^* are an indication that the time series (or residuals) are not white noise. White noise is defined as a variable that is normally distributed with mean 0 and variance 1. If Q is larger than the critical value, one rejects the null hypothesis that the time series (or residuals) are normally distributed. Alternatively, the p-value can be used.

The partial auto-correlation function

Suppose we have three variables Y, X and Z. In partial linear regression we removed the information of Z from both the Y and X variables and compared the residuals with each other (Chapter 5). With the auto-correlation we can do something similar. Suppose that Y_t and Y_{t-1} are highly correlated. As a result Y_{t-1} and Y_{t-2} are highly correlated as well. Because both Y_t and Y_{t-2} are highly correlated with Y_{t-1}, it is likely that Y_t and Y_{t-2} are correlated with each other! Would it not be nicer if we can calculate the correlation between Y_t and Y_{t-2} after removing the effect of Y_{t-1}? This is what the partial auto-correlation does (Makridakis et al. 1998). Technically, the partial auto-correlation a_k is obtained by applying the following linear regression and setting a_k equal to the estimated value of β_k.

$$Y_t = \text{intercept} + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_k Y_{t-k}$$

To obtain the partial auto-correlation for another time lag, k is increased in the regression model. The first partial auto-correlation is always equal to the first ordinary auto-correlation, and the same critical values can be used to assess its significance. The partial auto-correlation turns out to be particular useful in some of the techniques that are discussed later in this chapter.

Another example

An SST time series was taken from the oceanographical database COADS, available on the Internet. In this database, one can select a grid point and obtain monthly SST data. It should be noted that these data should be interpreted with care as a certain amount of smoothing and interpolation is applied. We selected the grid point with coordinates 2.5E and 57.5N (as it is close to the home of the first author), and extracted monthly time series from January 1945 until December 1992. This grid point is located east of Scotland. Various studies have compared SST with the North Atlantic Oscillation (NAO) index, which can be seen as an environmental index function. Here, we will compare the SST with the monthly

NAO index. To visualise the two time series, we used a lattice plot (Figure 16.6) and a coplot (Figure 16.7). The lattice plot shows the general increase in the SST series since the late 1980s, and the coplot indicates that there is a strong relationship between the NAO index and SST during the months January, February and March.

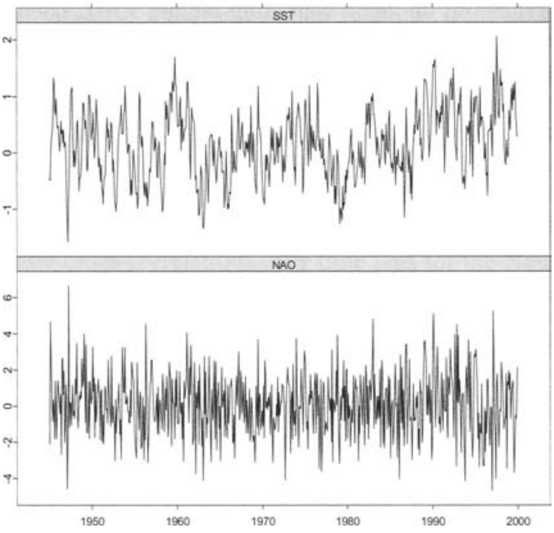

Figure 16.6. NAO and SST time series.

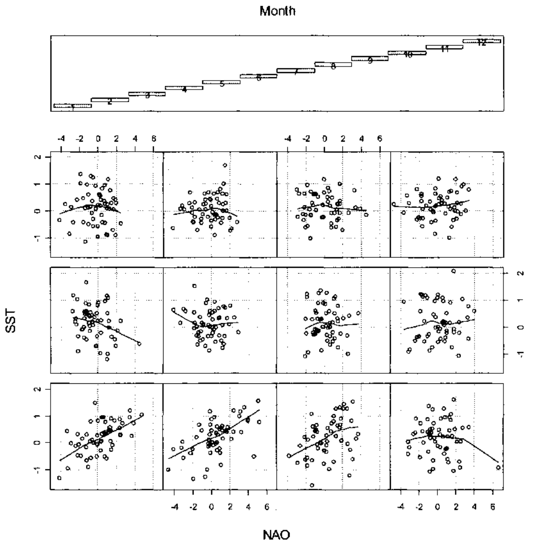

Figure 16.7. Coplot of SST and NAO index conditional on month. The lower left panel shows the relationship between the NAO and SST for all January data and

the upper right panel for the December data. A LOESS smoother was added in each panel.

The auto-correlation plot (not shown here) for the SST series shows a clear seasonal pattern. For the NAO index it is considerably weaker. However, we decided to remove the seasonal pattern from both series in order to avoid saying that they are highly related to each other just because of the seasonality. The process of removing the seasonal component is discussed later in this chapter. The deseasonalised SST and NAO series are plotted in Figure 16.8, and their cross-correlations are given in Figure 16.9. The latter graph indicates that the deseasonalised SST and NAO have similar patterns over time.

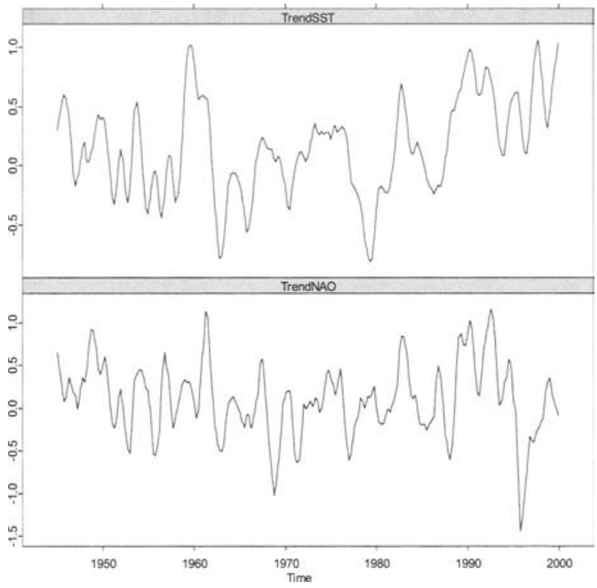

Figure 16.8. Deseasonalised SST and the NAO index.

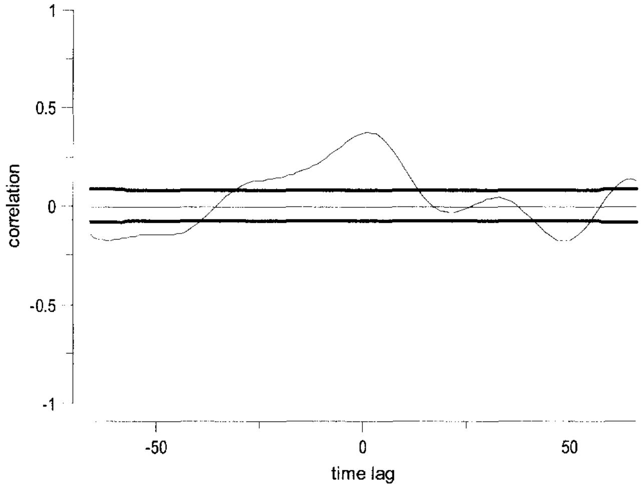

Figure 16.9. Cross-correlations between deseasonalised SST and NAO index.

Multivariate techniques

Figure 16.10 shows a time series plot of annual abundance indices derived from counts on six native duck species wintering on Scottish wetland sites (Musgrove et al. 2002). The data are expressed as a percentage of the abundance in the first year (1996 =100); see http://www.scotland.gov.uk/stats/envonline for further details and the data (source: The Wetland Bird Survey (WeBS)). Suppose that one is interested in determining which of the series are related to each other. One option is to calculate the cross-correlation matrix and inspect the numbers. However, it is also possible to apply a principal component analysis on these time series. Figure 16.11 shows the PCA correlation biplot (Chapter 12) obtained for the six time series. To simplify the interpretation, we used labels 6, 7, 8 and 9 for observations from the 1960s, 1970s, 1980s and 1990s, respectively, and 0 refers to 2000. The biplot explains 72% of the variation and shows that Pochards were abundant in the 1970s, Mallards in the 1960s, and Goosanders, Goldeneyes and Gadalls in the 1990s. All these results are in line with the time series plot (Figure 16.10). If one also has multiple explanatory variables, then redundancy analysis could be applied. To account for the time series aspect in the data we need only modify the permutation methods that we use to obtain the p-values that determine the significance of explanatory variables. Permutation techniques in RDA (or CCA) could make use of block permutation provided the series are long enough (Efron and Tibshiran 1993; Davison and Hinkley 1997; Lepš and Šmilauer 2003).

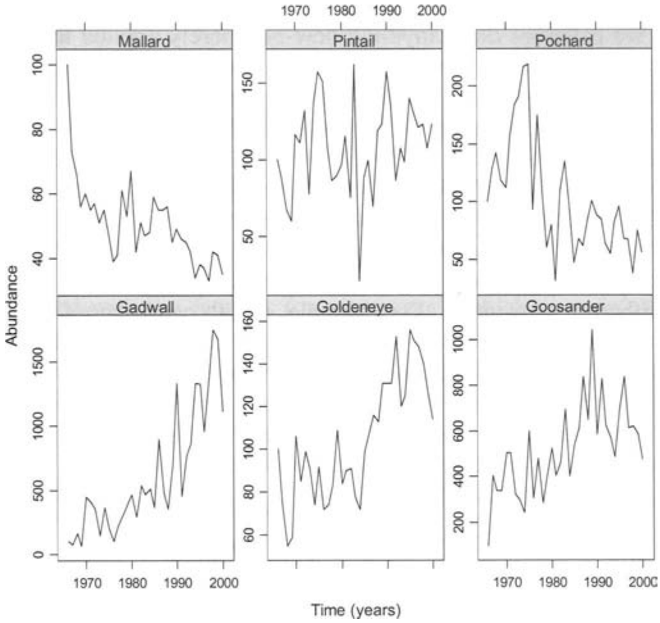

Figure 16.10. Lattice plot for the duck time series.

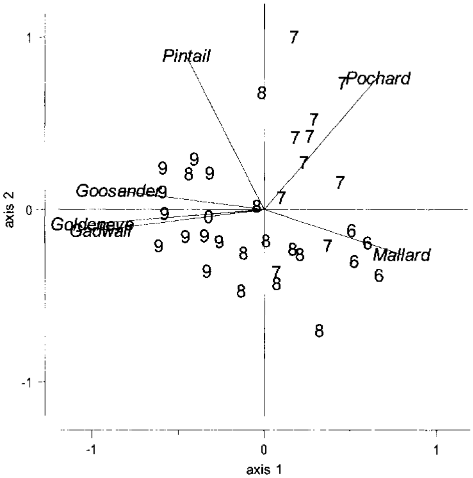

Figure 16.11. PCA biplot on six duck time series. The first two eigenvalues are 0.52 and 0.20, which means that the two axes explain 72% of the variation in the data (the eigenvalues are scaled to have sum 1). The labels 6, 7, 8 and 9 refer to observations from the 1960s, 1970s, 1980s and 1990s respectively, and 0 to 2000.

Generalised least squares

So far, we have not done anything new but merely applied tools we have discussed before. We now present the first extension. Recall that one of the assumptions in the linear regression model was independence of the data. This works it way through as independence of the residuals. For time series data, this assumption is clearly violated. It is relatively easy to show that if there is a temporal structure in the errors, then the linear regression model can seriously underestimate the standard errors of the slopes (Ostrom 1990) and this can lead to all kinds of trouble, including type I errors. So, how do we avoid this? We discuss two options. The first one is simple, add covariates such that there is no auto-correlation between residuals. For example, month can be used as a nominal explanatory variable in linear regression or as a smoother in an additive model. However, during the model validation process one must ensure that the residuals do not exhibit any temporal structure or else the model is incorrect. Chapter 20 shows an example of this approach. The second option is to extend the model in such a way that auto-correlation between residuals is allowed (Chapter 8). Recall that the underlying model in linear regression is (in matrix notation)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \text{where} \quad \boldsymbol{\varepsilon} \sim N(0,\sigma^2\mathbf{I}) \quad \text{or more generally} \quad \boldsymbol{\varepsilon} \sim N(0,\sigma^2\mathbf{V})$$

where **V** is a diagonal matrix with only ones. We formulated the underlying assumptions in terms of the underlying response variable *Y*, but as one of the assumptions states that the explanatory variables are fixed, the assumptions can also be expressed in terms of the noise. The independence assumption means that the noise at one time should not be related to noise at any other time.

The matrix **V** is the key to extending the linear regression (or additive) model so that it allows an auto-correlation structure between the residuals. Let us have a more detailed look at the structure of **V**. In linear regression it assumed to be the identity matrix:

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix}$$

The observed data are stored in a vector **Y** of the form **Y** = $(Y_1,..,Y_N)$' and the same holds for the residuals: $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_N)$'. As the variance of ε is equal to σ^2**V**, we have:

$$\text{cov}(\varepsilon_i,\varepsilon_j) = \begin{cases} 0 \text{ if i} \neq \text{j} \\ \sigma^2 \text{ if i} = \text{j} \end{cases}$$

Hence, by using the identity matrix **V** we are forcing an independence structure on the residuals. This is fine as long as the residuals are indeed uncorrelated. Ostrom (1990) gives a simple example in which linear regression is applied on a US defence expense time series. Comparing the linear regression model with a model that takes into account auto-correlation between residuals, he showed that the t-values obtained by linear regression were inflated by approximately 400%! Therefore, we better make sure that we do something appropriate with the time dependence structure if it is indeed present! The matrix **V** can be used for this. Although we will use it for time series applications, we might as well discuss it in a wider context. For example, one option is allowing groups of residuals to have a different variance. This is one way to tackle violation of homogeneity in linear regression. Suppose that we are trying to model (using linear regression) the relationship between weight and length of any species with a remarkable sexual dimorphism. A model validation may indicate that residuals of the males have a larger spread of residuals compared with the females (or vice versa), which is a violation of the homogeneity assumption. Let us assume that the first k observations in **Y** are of males and the remaining observations are of females. Instead of applying a data transformation to stabilise the variance, we could introduce two different variance components, one for the males and one for the females. Technically, this is done by using a diagonal matrix **V** with different values on the diagonal:

$$\mathbf{V} = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & v_N \end{pmatrix} \qquad (16.1)$$

In the example of the male and female we would have: $v_{male} = v_1 = v_2 = v_3 = \ldots =$ $\mathrm{v_k}$ and $v_{female} = v_{k+1} = v_{k+2} = \ldots = v_N$. As a result we have:

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & \text{if } i \neq j \\ v_{male}\sigma^2 & \text{if } i = j \quad \text{and both are males} \\ v_{female}\sigma^2 & \text{if } i = j \quad \text{and both are females} \end{cases}$$

Hence, there is no covariance between the different sexes, but males and females are each allowed to have a different variance. Specialised computer software can be used to estimate the variance components. Now suppose that sampling effort was different for the N observations. A possible scenario is if monthly averages are used but the number of observations per month differs due to a lower sampling effort during certain months (e.g., high sampling effort during spring and low during the winter). In this case, we know the sampling effort per observation and we can give more weight to observations with higher sampling effort by using the sampling effort as a weighting factor. Technically, the matrix **V** in equation (16.1) is used and the v's are set to the (known) weighting factors.

Now let us assume that the vector Y contains observations made repeatedly over time. In none of these extensions have we allowed for covariance between two observations. In the context of a time series, this means that $cov(Y_t, Y_{t+k}) = 0$ for $k \neq 0$. Now suppose that Y_t and Y_{t+1} are related to each other. We assume that the relationship between two observations that are one time unit apart is the same;

Y_1 and Y_2, Y_2 and Y_3, Y_3 and Y_4, etc. The same holds for two observations that are two units apart: Y_1 and Y_3, and Y_2 and Y_4, etc. This dependence structure implies that the covariance between ε_t and ε_{t+k} is given by

$$\text{cov}(\varepsilon_t, \varepsilon_{t+k}) = \begin{cases} \sigma^2 & \text{if } k = 0 \\ v_k\sigma^2 & \text{if } k \neq 0 \end{cases}$$

For example, the covariance for $k = 1$ is $cov(\varepsilon_t, \varepsilon_{t+1}) = v_1\sigma^2$, for $k = 2$ we have $cov(\varepsilon_t, \varepsilon_{t+2}) = v_2\sigma^2$, etc. So, all what we need is an estimate for v_1, v_2 and σ^2. It is straightforward to implement such an error covariance structure using a (positive definite) matrix **V** of the form

$$\mathbf{V} = \begin{pmatrix} 1 & v_1 & v_2 & \cdots & \cdots & v_{N-1} \\ v_1 & 1 & v_1 & \cdots & \cdots & v_{N-2} \\ v_2 & v_1 & 1 & & & v_{N-3} \\ \vdots & v_2 & v_1 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & v_1 \\ v_{N-1} & v_{N-2} & v_{N-3} & \cdots & \cdots & 1 \end{pmatrix}$$

The problem with this approach is that there are many elements in **V** to estimate. A common approach is to assume that the covariance between observations Y_t and Y_{t+k} only depends on the time lag between them; observations made close after each other have a much higher covariance than points separated more in time. This can be modelled as

$$\text{cov}(\varepsilon_t, \varepsilon_{t+k}) = v^{|k|}\sigma^2$$

Where v is between 0 and 1. The larger the time lag k, the smaller the covariance. As an example, assume that $v = 0.5$:

$$\begin{aligned} \text{cov}(\varepsilon_t, \varepsilon_{t+0}) &= v^0\sigma^2 = \sigma^2 \\ \text{cov}(\varepsilon_t, \varepsilon_{t+1}) &= v^1\sigma^2 = 0.5\sigma^2 \\ \text{cov}(\varepsilon_t, \varepsilon_{t+2}) &= v^2\sigma^2 = 0.25\sigma^2 \end{aligned}$$

Points close to each other have a much higher covariance than points with a larger time lag. This covariance structure can be modelled using a matrix **V** of the form

$$\mathbf{V} = \begin{pmatrix} 1 & v & v^2 & \cdots & \cdots & v^{N-1} \\ v & 1 & v & \cdots & \cdots & v^{N-2} \\ v^2 & v & 1 & & & v^{N-3} \\ \vdots & v^2 & v & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & v \\ v^{N-1} & v^{N-2} & v^{N-3} & \cdots & \cdots & 1 \end{pmatrix}$$

All we need is some clever software that estimates the value of v, together with σ, α and β. Hence, if we apply linear regression and use this matrix **V**, we are allowing for auto-correlation in the time series. Other correlation structures are possible (Pinheiro and Bates 2000). The auto-correlation structure can also be combined with random intercept and slope models within a mixed modelling context (Chapter 8). Generalised least squares examples can be found in Chapters 18, 19, 23, 26, 35 and 37.

16.2 Auto-regressive integrated moving average models with exogenous variables

Now that we have applied data exploration techniques, and auto-and cross-correlation functions, it is time for more advanced time series techniques. One of these methods is the auto-regressive integrated moving average model with exogenous variables, abbreviated as ARIMAX. Useful reading sources are Ljung (1987), Diggle (1990), Chatfield (2003), Brockwell and Davis (2002), among others. The text and ideas presented here owe much to Makridakis et al. (1998). The ARIMAX framework is based on the assumption that the time series is stationary, and therefore we first need to discuss this very important issue. ARIMAX consists of various building blocks, and it is perhaps easier to break them down in more simple models, namely the auto-regressive (AR) model, followed by the moving average models before dealing with the ARIMAX model itself.

Stationarity

So, what is stationarity? Well, roughly speaking it means that the time series Y_t does not contain a trend and the variation is approximately the same during the entire time span. We never said that ARIMAX models were useful for estimating trends! In fact, they cannot be used for this purpose! Stationarity is best assessed by making a time plot of the time series; it should fluctuate around a constant value, and the spread should be the same everywhere. None of the SST time series in Figure 16.2 are stationary as each series contains a seasonal component and some exhibit a long-term trend. None of the CPUE series in Figure 16.3 are stationary as they all contain a trend and variation during the 1970s is much lower for some time series compared with other periods. Neither the SST series nor the

NAO index in Figure 16.6 is stationary; the first series shows a clear trend and seasonal effect, and the NAO has changes in spread. None of the duck time series in Figure 16.10 are stationary as all contain either a trend or variation in the spread over time.

The auto-correlation can also be used to assess whether a time series is stationary. The auto-correlation function of a stationary series should drop to small values reasonably quickly. Slowly decaying auto-correlation functions (Figure 16.4) are an indication of non-stationarity, and the same holds for oscillating ones (seasonality). The partial auto-correlation function of non-stationary data tends to show spiky behaviour.

So, before moving on to ARIMAX models, we need to discuss how to get rid of non-stationarity. One option is to remove the trend and/or seasonal components and this will be discussed in Chapter 17. Within the world of ARIMAX models one often applies a different approach, namely differencing the time series. The first order difference is defined by

$$Y_t' = Y_t - Y_{t-1}$$

Note that because Y_0 does not exist, the length of the differenced series is one unit less than the length of the original series. Figure 16.12-A shows the CPUE Nephrops time series at station 3; the series is clearly non-stationary as it contains a trend and changes in spread. The time series containing the first differences is presented in Figure 16.12-B and is much closer to being stationary.

If the time series with the first differences is not stationary, the time series can be differenced a second time.

$$Y_t'' = (Y_t - Y_{t-1})' = Y_t' - Y_{t-1}' = Y_t - 2Y_{t-1} + Y_{t-2}$$

In practise, one rarely applies higher order differences as the interpretation becomes rather difficult.

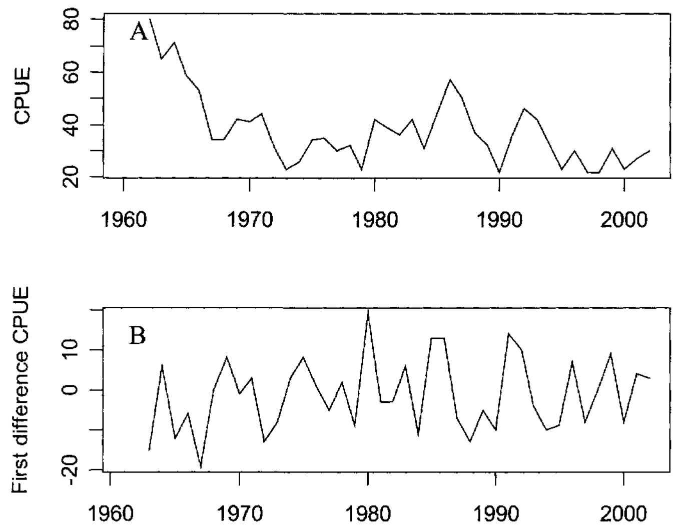

Figure 16.12. A: Non-stationary CPUE Nephrops time series at station 3 south of Iceland. B: First difference of CPUE series at station 3.

So, what do we do about monthly data? We can either analyse the first-order difference or look at seasonal differences:

$$Y_t' = Y_t - Y_{t-12}$$

This series shows the change between the same months in consecutive years. The same can be done for quarterly or even weekly data. As before, if the seasonal differenced series is not stationary, the time series can be differenced again to give

$$Y_t'' = (Y_t - Y_{t-1})' = Y_t' - Y_{t-1}' = Y_t - 2Y_{t-1}' + Y_{t-2}$$

It is also possible to difference the other way around (giving $Y_t - 2Y_{t-12} + Y_{t-24}$), but this makes the interpretation rather difficult. Figure 16.13-A and Figure 16.13-B show the time series of the Scottish SST and the auto-correlation, respectively. Both indicate clear non-stationarity. Panels C and D show the same graphs but now for seasonal differenced data. There is still strong evidence for non-stationarity. Panels E and F contain the first differences of the seasonal differenced series, and although there is still some auto-correlation with time lag 12, it is much closer to stationarity now.

One may also consider applying a transformation on the time series to stabilise the variance, if this is causing the non-stationarity.

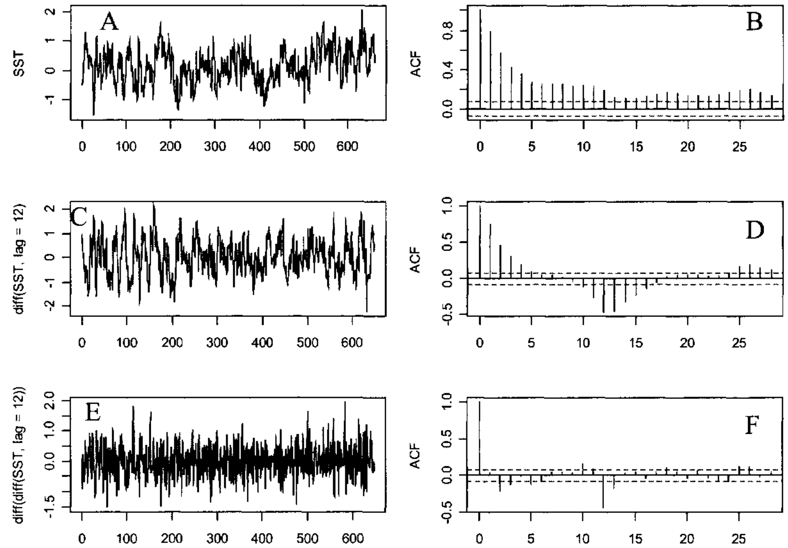

Figure 16.13. Time plot and auto-correlation for the Scottish SST series for the original data (A and B), seasonal differenced series (C and D) and first differences of the seasonal differenced series (E an F).

The AR model

In this class of models, the *stationary* time series Y_t is modelled as a function of past observations Y_{t-1}, Y_{t-2},...,Y_{t-p}, and a noise components e_t. Stationarity is rather important, and, techniques discussed in the previous paragraph should be applied if the series is non-stationary. Let us formulate an AR model for the SST time series northeast of Scotland:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots + \beta_p Y_{t-p} + \varepsilon_t$$

To ensure stationarity of the time series Y_t, we took the first-order differenced seasonal differences (see previous paragraph and Figure 16.13). The term ε_t is normally distributed noise with mean 0 and variance σ^2, and α is the intercept. The model is called an AR model of order p, also written as AR(p). Just as in linear regression we have an intercept and explanatory variables with regression slopes β_j, but the explanatory variables are now lagged response variables! The question that arises is why we are allowed to do this, and how can we obtain 'good' standard errors, t-values and p-values for the estimated regression parameters? What is the mechanism for this? Before addressing these issues, have a look at the structure of the model first. If we know the values of the intercept and all p slopes, we can easily predict the value for Y_t, Y_{t+1},Y_{t+2}, etc. Hence, for prediction purposes, this is a fantastic tool. But it is perhaps less suitable for the 'what is going on' question.

So, why are we allowed to include lagged response variables and obtain valid standard errors and p-values? The answer is stationarity and *asymptotic normality*. The last point is easy; just make sure you have at least 25 to 30 observations in time and then the central limit theory ensures that the normal distribution can be used to say that, in 95% of the cases, the population parameter β_j is within the interval given by the estimated slope $b_j \pm 1.96$ times its standard error. If the dataset is smaller, then do not apply ARIMAX or any of its subset models. Stationarity was discussed earlier in this section.

Let us work out the example for the SST series. In fact, this is not an AR(p) model, but an ARI(p) model as the series are integrated. The first question we have to address is how many lagged explanatory variables to take, or formulated differently, what is the order of the AR model? There are a couple tools we can use. First of all, based on theory, the partial auto-correlation of an AR(p) model will show spikes up from time lag 1 to p, and then drops to 0. So, we could calculate the partial auto-correlation function for the SST series and see when it drops to 0. Another tool is our best friend from Chapters 5 to 8, the AIC. It is defined in a similar way as in the linear regression model (a function of the maximum likelihood and the number of parameters). It can be calculated for different values of p, and the model with the smallest AIC can be selected as the optimal model. Just as in linear regression, the model selection procedure needs to be followed up by a model validation. In this process, the residuals need to be inspected and one should not be able to detect any patterns in it.

Table 16.4 shows the AIC values obtained by applying an AR(p) model on the differenced seasonal differenced Scottish SST data for various values of p. As can be seen, the larger the number of auto-regressive terms, the lower the AIC. In this case, one should use at least 14 AR terms. We did not try models with higher values of p as computing time becomes an issue. Instead of fully exploring the AR(14) model, we will extend the AR to ARMA models in the next two paragraphs.

Table 16.4. AIC for various values of p in an AR model for the differenced seasonal differenced Scottish SST time series.

p	AIC	p	AIC
1	1044.27	8	993.05
2	1016.95	9	995.03
3	1011.87	10	978.03
4	1012.62	11	972.78
5	994.86	12	851.29
6	989.93	13	844.88
7	991.56	14	834.39

The MA model

In an moving average model the *stationary* time series Y_t is modelled as a function of present and past error terms ε_t, ε_{t-1}, $\varepsilon_{t\ t-2}, \ldots, \varepsilon_{t\ t-q}$. A possible MA model for the SST time series northeast of Scotland is:

$$Y_t = \alpha + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \phi_3\varepsilon_{t-3} + \ldots + \phi_q\varepsilon_{t-q}$$

Just as in the AR(p) model we need to estimate the optimal number of MA parameters. The auto- and partial auto-correlation functions give some clues: The auto-correlation of an MA(q) model has spikes at lags up to q and then goes to zero, whereas the partial auto-correlation may show exponential decay or damped sine wave patterns (Makridakis et al. 1998). We could produce a similar table as for the AR(p) model in Table 16.4 but leave this as an exercise for the interested reader.

The ARIMA model

The more useful approach is the combination of the AR(p) and MA(q) model. It is of the form:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1\varepsilon_{t-1} + \ldots + \phi_q\varepsilon_{t-q}$$

This model is called an ARMA(p,q) model, and if the series are differenced, we call it an ARIMA(p,q) model and the challenge is to find the optimal values of p and q. Table 16.5 shows the AIC for various values of p and q for the Scottish SST series. We took more values of p than of q as the ecological interpretation of lagged error terms is rather difficult. The ARIMA(14,3) model has the lowest AIC. The numerical output for this model is given in Table 16.6. Note that the first auto-regressive parameter is relatively large, which may indicate that the integrated seasonal differenced series is still non-stationary. The problem with most software routines is that if one specifies an ARIMA(14,3) model, the software will estimate all time lags from 1 to 14. Based on the standard errors in Table 16.6 it may be better to omit some of the AR components as they are not significant at the 5% level (the 95% confidence band for each parameter is given by the estimated value ± 1.96 times the standard error).

Table 16.5. AIC for various values of p and q in an ARIMA(p,q) model for the differenced seasonal differenced Scottish SST time series. The five models with the lowest AIC are in bold face.

		q	
p	1	2	3
1	1041.008	940.919	942.919
2	943.294	942.919	943.635
3	942.523	944.463	894.725
4	944.315	945.794	918.098
5	990.200	943.168	894.037
6	991.451	951.403	888.152
7	993.443	924.387	936.914
8	995.047	919.169	887.076
9	995.120	932.378	882.649
10	978.820	910.351	911.431
11	932.888	907.887	897.664
12	814.875	799.124	787.017
13	**781.672**	**783.672**	**782.750**
14	**783.659**	785.665	**772.192**

Table 16.6. Estimated parameters and standard errors for the ARIMA(14,3) model for the Scottish SST series.

AR	Estimate	S.E.	AR	Estimate	S.E.	MA	Estimate	S.E.
1	−0.973	0.055	8	0.032	0.051	1	0.890	0.050
2	0.039	0.077	9	−0.009	0.048	2	−0.459	0.085
3	0.326	0.058	10	0.006	0.048	3	−0.842	0.049
4	−0.078	0.050	11	0.076	0.047			
5	0.007	0.050	12	−0.386	0.046			
6	− 0.133	0.049	13	−0.632	0.056			
7	− 0.100	0.052	4	−0.264	0.044			

Extending the ARIMA to ARIMAX models

So far, we have ignored the explanatory variables. It is relatively simple to extend the ARIMA with explanatory variables. For Scottish SST data we also have the NAO index, which could be a driving factor for sea surface temperature. A possible ARMAX model is of the form:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q} + \gamma \mathrm{NAO}_t$$

where Y_t is the SST in month t. The underlying assumption is again that the series Y_t is stationary. If this is not the case, there are two options. The first option is to take the differences of both the Y_t and the explanatory variable NAO$_t$ until the Y_t is stationary. For this specific example, we would take the integrated seasonal differences for both series. The second option is to consider the model as

$$Y_t = \alpha + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \gamma \mathrm{NAO}_t + N_t$$
$$N_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q}$$

And assume that the noise series ε_t is stationary. This brings us back to the world of generalised least squares. There is no differencing of the Y_t or NAO_t series involved. In the Scottish SST data, it may be an option to add a nominal variable month to the model:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \gamma \mathrm{NAO}_t + \text{factor(Month)} + N_t$$
$$N_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q}$$

The component 'factor(Month)' creates 11 dummy variables with zeros and ones identifying in which month a measurement was taken. This process is identical to linear regression (Chapter 5). Things can be made even more complex if we have models with lagged explanatory variables. An example is given by

$$Y_t = \alpha + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \gamma_1 \mathrm{NAO}_t + \gamma_2 \mathrm{NAO}_{t-1} + \gamma_3 \mathrm{NAO}_{t-2} + N_t$$
$$N_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q}$$

The SST in year t is modelled as a function of SST in the past (the AR components), the NAO, the NAO in the past (lagged terms), noise, and noise from the past. Indeed, a complicated model, but we never said that it would be easy. This set of models is called a dynamic regression model, and a more detailed discussion can be found in Chapter 8 of Makridakis et al. (1998). It is also possible to use models in which the regression parameters are allowed to change over time, and these will be discussed in Chapter 17.

ARIMAX models are useful for prediction but not for understanding what goes on. This statement holds especially if one starts to take differences. If ARIMAX models are applied on non-stationary data, standard errors of estimated parameters should be interpreted with great care. ARIMAX models are typically applied on univariate response variables. Recently, multivariate extensions of these models have been developed and are called vector AR models (Shumway and Stoffer 2000; Lütkepohl 1991). Specialised software and considerable statistical knowledge are required to apply these multivariate AR models, and they are not discussed here.