7 Time series analysis

In Chapters 16, 17, 33 – 36 in Zuur, Ieno and Smith (2007), various time series techniques are discussed. Applying these methods in Brodgar is straightforward, and most choices are self-explanatory. This chapter provides extra information on some of the methods.

7.1 Time series techniques

Some of the time series techniques discussed in Zuur et al. (2007) are available from the exploration or multivariate menus. Others can be obtained by clicking on the main menu time series button (Figure 7.1). The panels show further time series techniques available in Brodgar from the same menu button. There are four main sections, namely (i) trends, (ii) sudden change, (iii) relationships and (iv) cycles.

General time series methods are auto-and cross-correlations, time lags plotting ARIMAX modelling and spectral analysis. These methods make use of R, and applying them in Brodgar is similar to the linear regression menus (Chapter 5). These methods are available from the lower left and right panels in Figure 7.1.

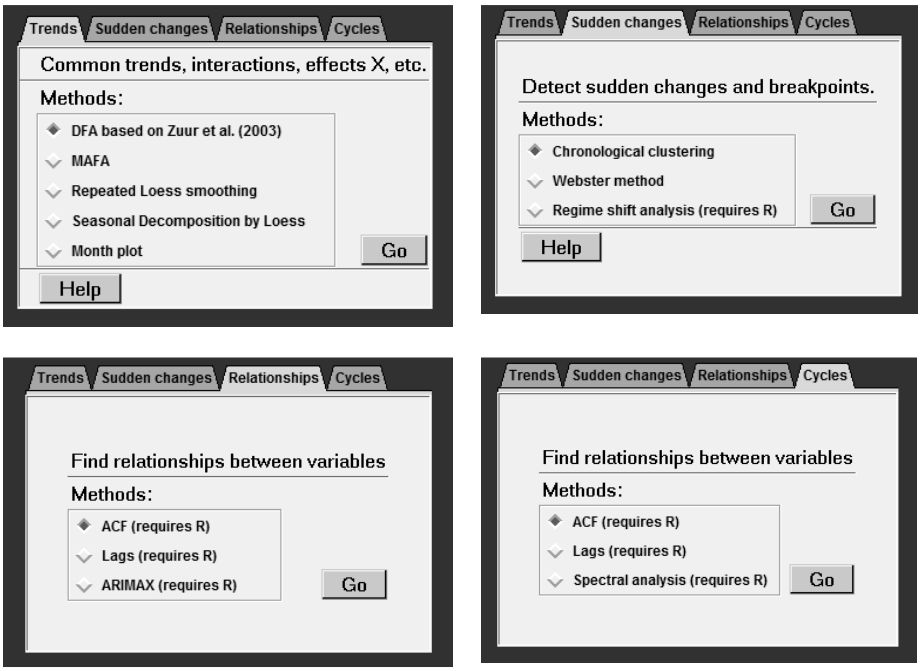

Figure 7.1. Time series analysis menus.

7.2 Dynamic factor analysis

Running dynamic factor analysis in Brodgar is simple; select 'DFA based on Zuur et al. (2003)' in the upper left panel in Figure 7.1, and click the 'GO' button. The upper left window in Figure 7.2 will appear.

In the upper left window, one can select the model, the number of common trends, the type of error covariance matrix and the response variables. The following DFA models are available in Brodgar:

1. data = M common trends + noise
2. data = M common trends + explanatory variables + noise
3. data = N univariate trends + noise
4. data = N univariate trends + explanatory variables + noise
5. data = 1 common trend + noise
6. data = 1 common trend + explanatory variables + noise
7. 1 time series = trend + noise
8. 1 time series = trend + explanatory variables + noise

We now discuss each of these models.

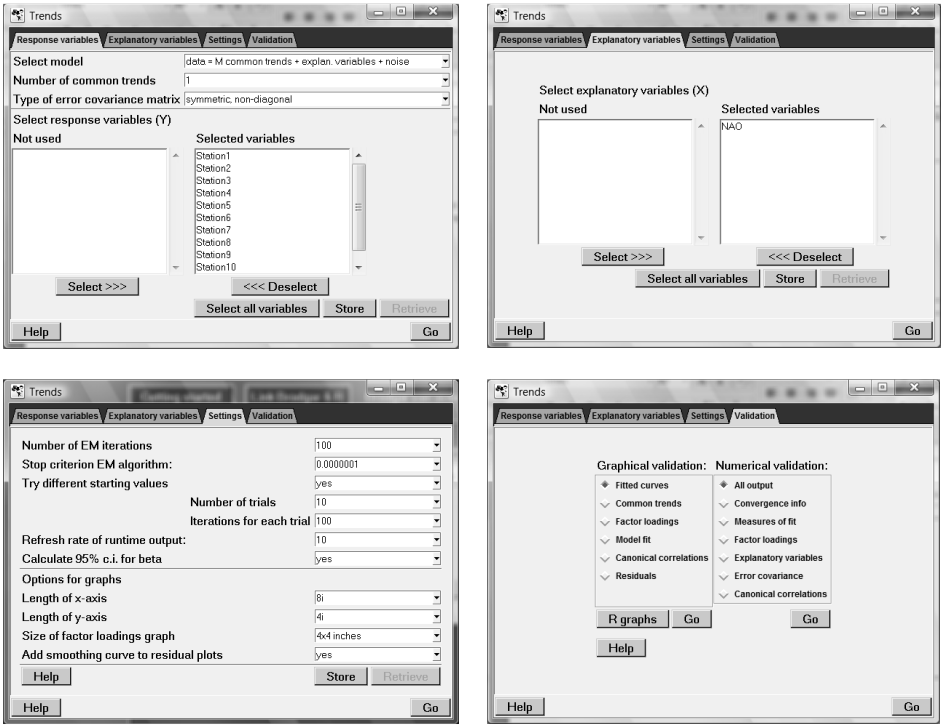

Figure 7.2. Options for the DFA (Zuur et al. 2003) method.

Model groups 1 & 2

These models are discussed in detail in Zuur et al. (2007), and Zuur et al. (2003[ab], 2004). The DFA code in Brodgar can deal with scalar explanatory variables, i.e. explanatory variables that have only one value at time *t*. However, the model structure is such that an explanatory variable can have a different effect on each of the response variables. This is probably best explained with help of a formula. Suppose we have a dynamic factor model with 5 time series, 2 common trends, and 1 explanatory variable. The mathematical formulation is given by:

$$
\begin{aligned}
y_{1t} &= a_{11} \times z_{1t} + a_{12} \times z_{2t} + b_1 \times x_t + e_{1t} \\
y_{2t} &= a_{21} \times z_{1t} + a_{22} \times z_{2t} + b_2 \times x_t + e_{2t} \\
y_{3t} &= a_{31} \times z_{1t} + a_{32} \times z_{2t} + b_3 \times x_t + e_{3t} \\
y_{4t} &= a_{41} \times z_{1t} + a_{42} \times z_{2t} + b_4 \times x_t + e_{4t} \\
y_{5t} &= a_{51} \times z_{1t} + a_{52} \times z_{2t} + b_5 \times x_t + e_{5t}
\end{aligned}
$$

where y_{it} is the value of the i^{th} time series at time t, z_{1t} and z_{2t} are the two common trends at time t, x_t is the value of the explanatory variable at time t and b_1, .., b_5 are

regression coefficients. The terms e_{1t}, .., e_{5t} are noise components and it is assumed that $e_t = (e_{1t}, .., e_{5t})'$ is normally distributed with expectation 0 and covariance matrix **R**. For simplicity, we have ignored the constant level term. In matrix notation, the model can be written as:

$$\mathbf{Y}_t = \mathbf{A} \times \mathbf{z}_t + \mathbf{b} \times \mathbf{x}_t + \mathbf{e}_t$$

The elements of **A** are called factor loadings and indicate which common trends are important for which of the N response variables. The parameters b_1, .., b_5 are regression parameters. The relationship with `ordinary' factor analysis becomes clear by writing down the expression for the covariance matrix of $\boldsymbol{Y}_t$. For simplicity, we have omitted the 'explanatory variable' component in the model.

$$\text{Cov}(\mathbf{Y}_t) = \mathbf{A} \times \text{Cov}(\mathbf{z}_t) \times \mathbf{A}' + \mathbf{R}$$

For identification purposes, the covariance matrix of the common trends is set to the identity matrix, which means that we get:

$$\text{Cov}(\mathbf{Y}_t) = \mathbf{A} \times \mathbf{A}' + \mathbf{R}$$

This is a similar covariance model as in factor analysis, except that our factors (or: common trends) are required to be smoothing functions over time. In factor analysis, the error covariance matrix R is usually taken as a diagonal matrix. Consequently, the off-diagonal elements of the covariance matrix of the response variables are modelled entirely as a function of the factor loadings. However, there is no reason why the user should not use a symmetric, non-diagonal matrix for R in dynamic factor analysis. Our experience using such a matrix is positive. The various DFA models are presented in Table 7.1. To decide which model to use, the AIC criterion can be used. This criterion is a trade-off between measure of fit, and the number of parameters in the model.

The advantage of using a non-diagonal matrix for R is that the number of common trends needed for an adequate model fit is smaller; instead of using a common trend for 2-way interactions, one single parameter can be used. The disadvantage is that the number of parameters increases drastically.

The implementation of DFA in Brodgar can cope with missing values in the response variables, but not with missing values in the explanatory variables. To avoid a fatal crash, Brodgar replaces these missing values by the average of the explanatory variable. We suggest standardising the explanatory variables, as this makes comparison of the estimated regression parameters easier. The use of highly correlated explanatory variables (multi-collinearity) should also be avoided. Occasionally, it happens that a common trend gives an exact fit for one of the response variables. This behaviour might also occur in `ordinary' factor analysis and is called a Heywood case. Using a non-diagonal matrix for R solves this problem. A possible explanation is that the time series are too noisy (or fluctuate too much), and therefore DFA, which is basically a smoothing technique, might be inappropriate.

Table 7.1. Various DFA models.

Model	matrix **R**	Interpretation
$\mathbf{Y}_t=\mathbf{A}z_t + \mathbf{e}_t$	Diagonal	The N time series are modelled as a linear combination of M common trends. These common trends represent the joint signal in a group, or all the series. The diagonal elements of **R** indicate the amount of information that cannot be explained by the common trends.
$\mathbf{Y}_t=\mathbf{A}z_t + \mathbf{e}_t$	non-diagonal	As above. Additionally, 2-way interactions (if present) between the series are modelled by the off-diagonal elements of R.
$\mathbf{Y}_t=\mathbf{A}z_t + \mathbf{b}x_t + \mathbf{e}_t$	Diagonal	The N time series are modelled as the effects of measured explanatory variables plus a linear combination of the common trends (a common pattern in the series which cannot be explained by the explanatory variables), plus a certain amount of information (time series specific) that cannot be explained by any of the other components.
$\mathbf{Y}_t=\mathbf{A}\mathbf{z}_t + \mathbf{b}x_t + \mathbf{e}_t$	non-diagonal	As above. Additionally, 2-way interactions are modelled via the off-diagonal elements of R. This might mean that less common trends are needed for an adequate model fit compared to the model above.

We have used Brodgar to analyse data sets up to 20 time series (response variables) and up to 5 explanatory variables. We used 1 to 5 common trends and both the diagonal and non-diagonal matrices R. Results for different starting values were nearly identical, indicating that the numerical estimation procedure indeed obtained the global optimal solution. For larger data sets, this might not always be the case and a local optimum might be found.

Model groups 3 to 6

In these models a trend is estimated for each response variable. It is still possible to use a symmetric, non-diagonal matrix R, resulting in so-called seemingly unrelated time series models Harvey (1989).

Model groups 5 & 6

In these models, one common trend is used, and all factor loadings are set to 1. Hence, the N time series are modelled as one common pattern plus a level parameter (see below) plus noise. Basically, this is a dynamic factor model with only one common trend and all factor loadings set to 1. The common trends are shifted up and down via the constant level parameter.

Model groups 7 & 8

These models are for univariate series only. They model the time series as:

1 time series = trend + noise.

The trend is modelled as a so-called random walk. If there is more than one response variable, a multivariate model is needed. A sensible approach is to start with N univariate trends (trend for each Y), where N is the number of time series, followed by a model containing one common trend, two common trends, three common trends, etc.

The models with common trends (groups 1, 2, 5, 6) all have constant level parameters, see Zuur et al. (2003[a,b],2004). Please note that if a model with explanatory variables is selected, you should have specified explanatory variables in the data import step.

If a model with explanatory variables is selected, make sure that the appropriate variables are selected in the upper right window in Figure 7.2.

Settings for DFA

The lower left panel in Figure 7.2 shows the default settings for DFA. Various items can be changed via this window, namely:

- Number of EM iterations. Parameters in the dynamic factor model are estimated with the so-called EM algorithm. Here, you can set the upper limit of the number of iterations the EM algorithm will carry out. For reasonably fast computers (800 MHz and above), it is advisable to use 1500 EM iterations. If computing time takes too long (either because of a slow computer or a large data set), it can be decreased to 500 iterations in preliminary analyses. If the number of EM iterations is changed, Brodgar stores it as the new default value.
- Stop criterion EM algorithm. If changes in the maximum likelihood function become smaller than this criterion, the EM algorithm will stop. We advise 0.00001. If a different value is used, please note that Brodgar does not store the new value. Hence, you will need to change this each time Brodgar is started.
- Try different starting values. Most optimisation routines rely on good starting values. If 'yes' is selected, the EM algorithm starts x-times with different starting values (which are chosen at random) and carries out y-iterations. After x-runs the starting values that resulted in the lowest AIC (or highest value of the likelihood function) are used as starting values in the final run. The user can

choose the number of runs (either 5 or 10) and the number of EM iterations in each run (50, 100 or 200). We recommend 5 loops with 100 iterations each.

- Refresh rate of runtime output. The graphical user interface of Brodgar was written in a language called Tcl-Tk. The statistical routines were programmed in FORTRAN. Once the parameter estimation process is started, the FORTRAN code will save results to a file every jth iteration and a signal is given to the graphical user interface to present these results in a window. This allows the user to monitor the progress of the estimation process. The value of j can be changed.
- Calculate 95% c.i. for beta. These are the confidence intervals for the parameters corresponding to the explanatory variables. Depending on the number of response variables and explanatory variables, estimating the confidence intervals might be time consuming.
- Error covariance matrix. Here the user can choose between a diagonal matrix or a symmetric, non-diagonal matrix for the error matrix R.

Output during runtime

The estimation process is started by clicking the 'GO' button in Figure 7.2. During the estimation process a new window will appear, see Figure 7.3. It shows the intermediate results during runtime. In particular, it shows:

- at which iteration the algorithm is,
- the estimated value of the log likelihood function,
- the change in the log likelihood function (compared to the previous iteration),
- the AIC, BIC and CAIC values (model selection tools)
- changes in the trends, factor loadings (Z), noise component (R), and parameters for the explanatory variables (beta),
- the iteration for the starting values (if selected),
- and most importantly, the common trends plotted versus time.

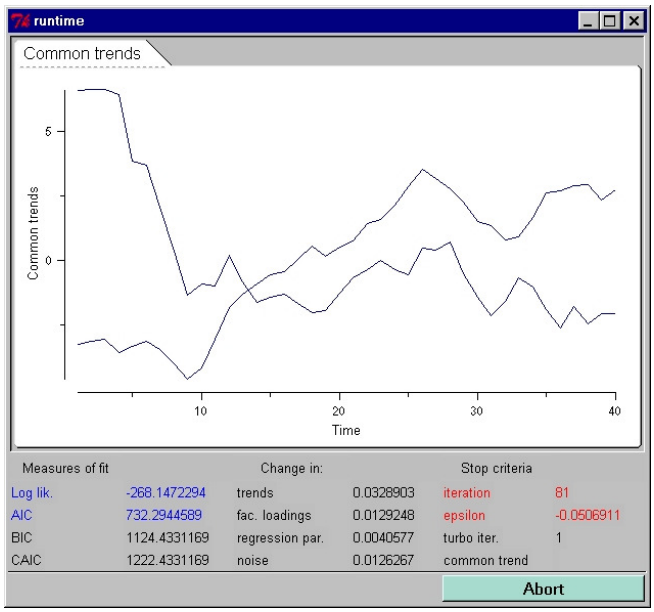

Figure 7.3. Output during runtime.

Recall that the dynamic factor model has a constant level parameter. This is modelled via a dummy explanatory variable (taking the value of one). For this reason, there will always be a change in beta, even if you have not selected explanatory variables. Changes in the parameters (the elements of Z, beta and R) are defined as the sum (over all elements in a matrix) of absolute differences of estimated values in the current iteration and the previous iteration.

Validation of the DFA model

In this paragraph, model validation of DFA is discussed. A distinction has been made between a graphical and numerical validation. Select ‘Time Series’ from the main menu, and click the ‘Validation’ tab (after convergence of the algorithm). This gives the lower right panel in Figure 7.2. This panel is also given in Figure 7.4. The options on the left correspond to the graphical tools, those on the right to numerical information. Because all information is stored in the project directory, one can easily access the validation information without running the DFA algorithm.

Graphical output

Once Brodgar has estimated the parameters in the dynamic factor model, the user can use various graphical tools to assess whether the chosen model is the most optimal one. This process follows similar lines as in linear regression. The following tools are available:

- Plot all fitted curves in one graph.
- Plot all common trends in one graph, and in separate graphs.
- Plot the factor loadings.
- Plot the model fit and observed values versus time in one graph for each response variable.
- Plot canonical correlations.
- Plot the residuals versus time.
- Plot histograms of the residuals.

We discuss these tools below. The CPUE Nephrops data set was used. Time series were standardised.

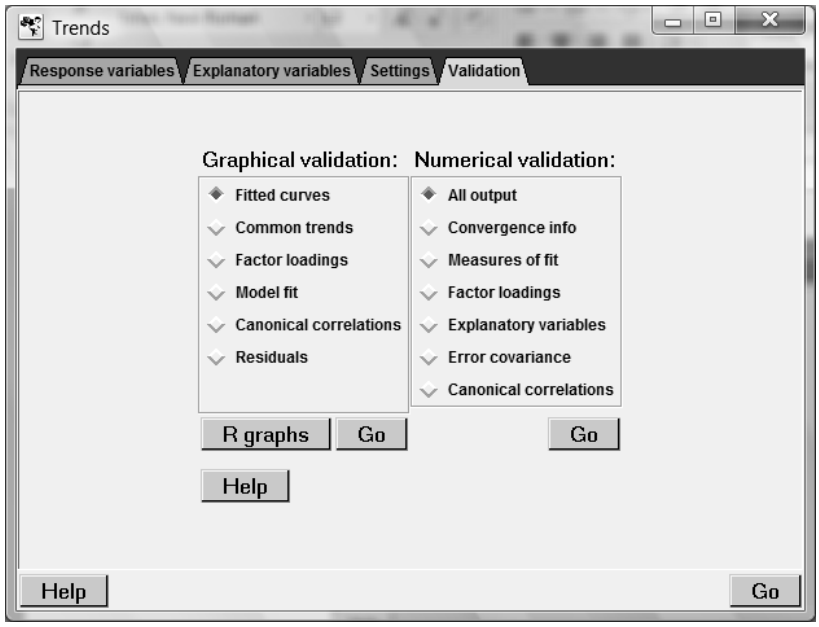

Figure 7.4. Validation options.

Clicking on the button labelled 'Fitted curves' in Figure 7.4, pops up a new window containing all fitted curves in one graph, see Figure 7.5. Using the legend, colours of the lines can be changed to blue. Comparing fitted curves with each other, and with the observed time series ('Plot data' in Data exploration), might provide very useful information. Clicking the 'Common trends' button in Figure 7.4 shows the estimated common trends, see Figure 7.6. Again, by clicking on names on the legend, the color of the common trends can be changed. The second and third tabs in Figure 7.6 contain the individual common trends and 95% c.i.

The estimated factor loadings for each axis are plotted as vertical lines from 0 to their estimated values, see Figure 7.7. Furthermore, loadings of axis i are plotted versus loadings of axis j for every combination of i and j. By clicking on the dotted line under the name of a tab, the corresponding graph will 'jump' out and can be compared with other graphs from the same window. Via the 'Settings' menu in Figure 7.2, the size of the graphs in Figure 7.6 and Figure 7.7 can be changed.

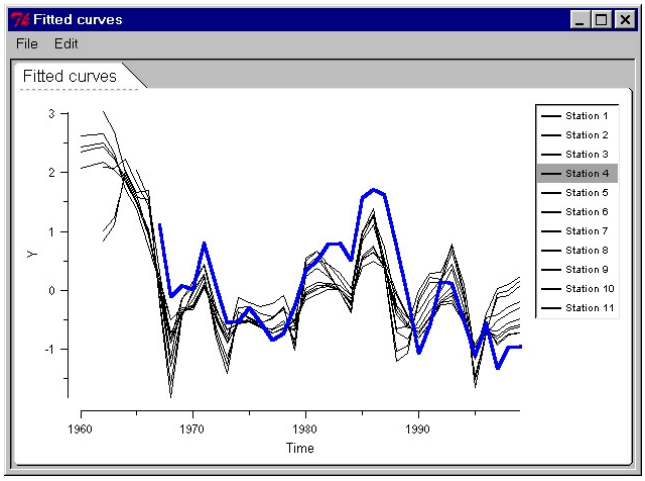

Figure 7.5. Fitted curves for Icelandic CPUE data.

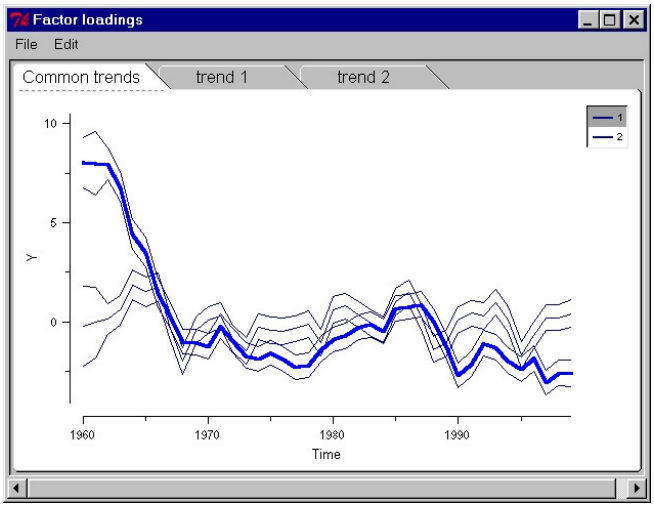

Figure 7.6. Estimated common trends and 95% confidence intervals.

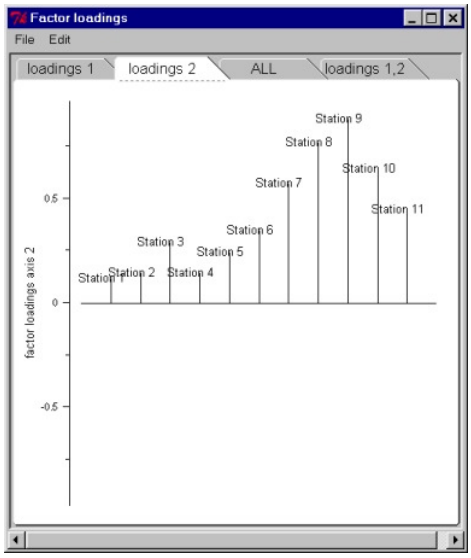

Figure 7.7. Factor loadings axis 2.

Results in Figure 7.7 indicate that the second axis is important for the stations 9, 8, 10, 8 and 11. The fit of the dynamic factor model for each response variable can be viewed via the 'plot model fit' button, see Figure 7.8.

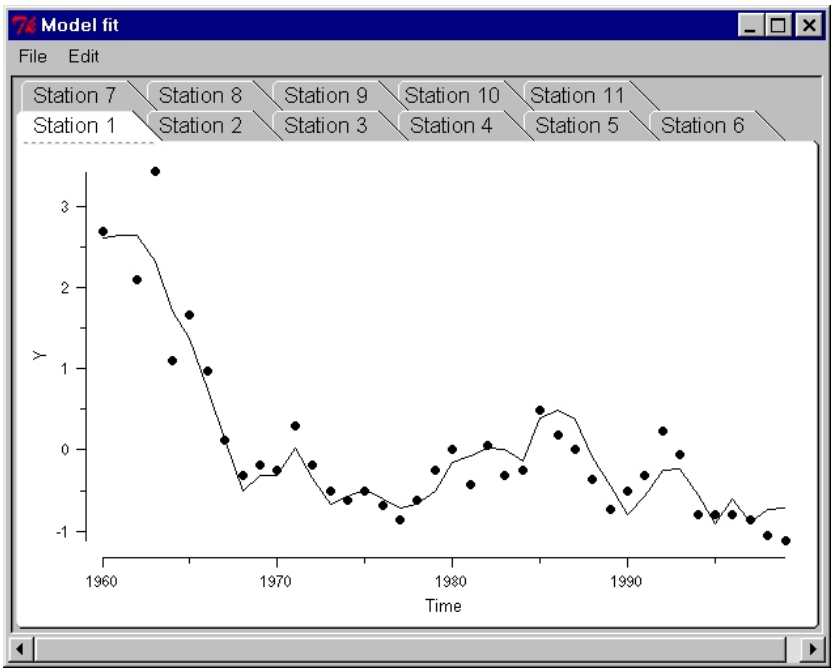

Figure 7.8. Model fit for each response variable.

Dynamic factor analysis is a dimension reduction technique. It is unlikely that every response variable is fitted well if a small number of common trends is used. This is the same with techniques like principal component analysis and correspondence analysis. In these techniques, points close to the origin are generally not fitted well. The advantage of dynamic factor analysis is that is shows the fit, whereas PCA and CA do not. Dynamic factor analysis is also a smoothing technique. In contrast to ordinary smoothing techniques, dynamic factor analysis estimates the amount of smoothing automatically. Occasionally, dynamic factor analysis produces fitted lines that are an exact fit. This means that the amount of smoothing is too small, which is an indication that the underlying model is inappropriate. In our experience, switching to a symmetric, non-diagonal matrix for the error covariance matrix solves this problem. Another option might be a transformation. Also inspect the data for outliers.

The residuals are calculated as the observed values minus the fitted values. Fitted values are the values obtained by the Kalman smoothing algorithm (Zuur et al. 2003[a]). The residuals can be plotted versus time. To enhance visual inspection of the residuals, a smoothing line can be added.

Figure 7.9 shows how graphical tools can be used in practise. Various windows can be open simultaneously.

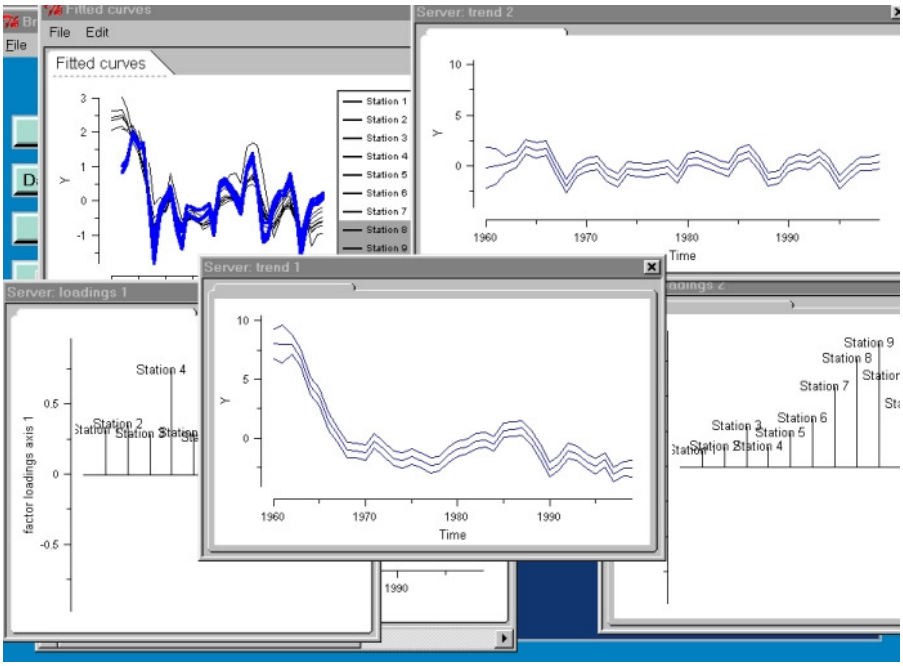

Figure 7.9. Validation tools.

Canonical correlations are correlations between the original time series and the common trends. If a canonical correlation is large, then this indicates that the corresponding response variable follows the pattern of the common trend. If it is low, then the response variable is not related (in a linear context) to the common trend. These correlations provide similar information as the factor loadings. However, in some multivariate methods (e.g. discriminant analysis and canonical correspondence analysis), there is a discussion on the reliability of weights (factor loadings).

Numerical output

The numerical output can be obtained from the options on the right side in Figure 7.4. There are two important buttons, namely 'Measures of fit' and 'Explanatory variables'. Clicking the 'Measures of fit' button will give the AIC, log likelihood value and number of parameters, among others. For the example data, the following output is obtained via this button:

```
Measures of fit
LOG Likelihood =   -313.760
AIC  =    713.520
BIC  =    885.580
CIAC =    928.580
Number of parameters =  43
```

The output was obtained for a dynamic factor model, with two common trends and a diagonal matrix for the noise covariance matrix. The AIC is the most important quantity. Various models can be tried, and the model with the smallest AIC value is likely to be the best model. The BIC and CIAC are alternative measures of fit, but in our experience the AIC is more useful.

The button 'Explanatory variables' will give the estimated values of the constant level parameters and the explanatory variables (if selected) plus standard errors and t-values. For the model above, the following output was obtained.

```
EXPLANATORY VARIABLES
  Used model:
   y(t) = GAMMA * alpha(t) + D x(t) + epsilon(t)
   alpha(t) = alpha(t-1) + eta(t)

x(t) contains the explanatory variables. The first col-
umn of D contains the constant level.

1. Constant level parameters and standard errors (first
column of D)

   Index, estimated value, standard error and t-value
      1      0.07      0.52      0.13
      2      0.15      0.58      0.26
      3      0.06      0.51      0.12
      4      0.93      1.24      0.75
      5      0.06      0.50      0.12
      6      0.05      0.51      0.10
      7      0.10      0.71      0.14
      8      0.04      0.97      0.04
      9      0.03      1.12      0.03
     10      0.25      0.81      0.31
     11      0.36      0.62      0.59
```

```
2. Estimated regression parameters
No explanatory variables were used.
```

The constant level parameter is modelled with help of a dummy variable, which takes the value of one for each response variable (Zuur et al. 2003[a]). The second column (containing the values 0.07, 0.15, etc.) contains the estimated values for the constant level parameters. The third column contains the standard errors and the fourth column the t-values. Although these *t*-values should be interpreted with care, t-values larger than 3 (in absolute sense) indicate a strong relationship between the explanatory variable and the response variable. In this case, all estimated constant level parameters are non-significant different from zero. This did not come as a surprise because the time series were standardised prior to the analysis. If besides the dummy variable, explanatory variables are used, results are presented in the same way.

The output of Brodgar can also be found in ascii files, under the output directory of Brodgar. Recall that this is the directory which has the same name (and directory path) as your project name except for the *.brd extension. The relevant files are: results.txt, Res_ci.txt, Res_covm.txt, Res_expl.txt, Res_facl.txt and Res_mf.txt. The text in these files explains what is printed.

Other numerical information (e.g. starting values of the trends, estimated error covariance matrix, convergence information) is available and follows the notation in Zuur et al. (2003a).

Copy numerical output to the clipboard

Numerical output of most techniques in Brodgar can be copied directly to the windows clipboard. From there, you can paste it into Microsoft Excel. Just click one of the 'Copy to clipboard' buttons in Brodgar, and press Control-V in Excel. The interpretation of the output is discussed below. For DFA, the output is as follows.

Factor loadings

An ascii file containing the loading can be found in: \YourProjectName\facload.txt. The first column contains the factor loadings of the first axis, the second column the loadings of the second axis, etc.

Trends

An ascii file containing the trends can be found in: \YourProjectName\at.txt. Interpretation of the columns in this file are:

- First column: time index.
- Second column: the first common trend.
- Third column: the lower 95% confidence band of the first common trend.
- Fourth column: the upper 95% confidence band of the first common trend.

- Fifth column: the second common trend (if selected).
- Sixth column: the lower 95% confidence band of the second common trend.
- Seventh column: the upper 95% confidence band of the first common trend.
- Eight column: third common trend (if selected)
- etc.

7.3 Repeated Loess smoothing in Brodgar.

Repeated Loess smoothing is available from the 'Trend' menu under the main menu button 'Time Series'. First, the user can (de-)select response variables. Brodgar will apply repeated Loess smoothing on each selected time series.

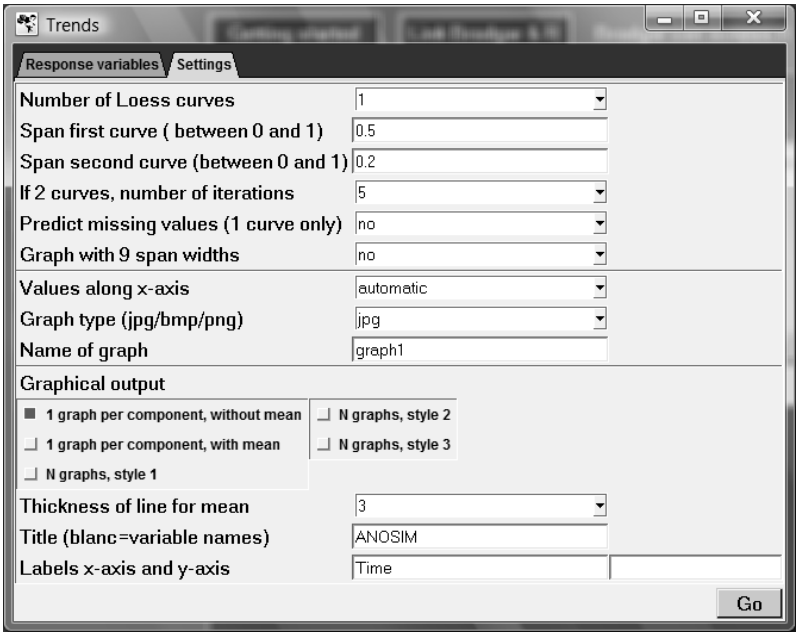

Figure 7.10. Settings for repeated Loess smoothing in Brodgar.

The following options are available (Figure 7.10).

Number of Loess curves. Either 1 or 2.

Span width of the first Loess curve. This should be a value between 0 and 1.

Span width of the second Loess curve (if selected in the first step). This should also be a value between 0 and 1, and smaller than the span width of the first Loess curve.

If 2 smoothing curves are selected, the number of iterations. This is the iteration of repeated Loess smoothing described above. In general, 5 iterations will suffice.

Predict missing values. This option, only available if one smoothing curve is estimated, will predict (using Loess) missing values in the original time series. Note that it will not predict missing values immediately at the beginning or end of the time series.

Values along the x-axis. This can be automatic (resulting in values of 1, 2, 3, etc.), or the column labels can be used. In the latter case, make sure that the column labels are numeric.

Thickness of line for the mean. This line is only plotted if the appropriate box is ticked in the graphical output menu, see above. The line width can be changed from 3 (default) to other values.

Graph type. Allows the user to select saving graphs as a JPG, BMP or PNG.

Name of graph. Brodgar will save the graph as: YourName.JPG in the project directory (assuming a JPG file was selected).

Title. This title will appear at the top of the figure.

Title x-axis. Allows the user to enter a label for the x-axis.

Title y-axis. Allows the user to enter a title for the y-axis.

The graphical output consists of:

- All smoothing curves in one graph, with or without a mean value, per component.
- The first, second and residuals on one page. The style options determine whether graphs are plotted next to each other, under each other, or in a matrix format. It is also possible to access the estimated smoothing curves, click on the 'Numerical output' button. It will give the estimated smoothing curves and data with missing values replaced by estimators (if requested). This output can be

copied and pasted manually to Excel if you wish to create graphs with a different plotting style.

7.4 Seasonal decomposition by Loess smoothing

This technique makes use of the R function 'stl', which decomposes a time series into a trend, seasonal component and residual information using Loess smoothing. Examples and details are given in Zuur et al. (2007). A related method is the month or cycle plot, and is discussed next.

7.4.1 Month plots (cycle plots)

This technique plots data of the same month in one graph. It will create a January time series, a February time series, etc. These 12 time series (of the original data) are plotted in one graph in a coplot style. Besides a month plot of the original data, Brodgar will also add a month plot for the components obtained by the seasonal decomposition by Loess smoothing, see Cleveland (1993) for details. An illustration of seasonal decomposition by Loess smoothing and month plots is presented in Figure 7.11 to Figure 7.14. Figure 7.11 shows monthly CO2 data measured on Hawaii since the late 1950s. The decomposition into a trend and seasonal component using Loess smoothing is presented in Figure 7.12 and Figure 7.13. We used a Loess window of 21 (span in lags). Figure 7.14 shows the month (or cycle) plot. Results indicate that the monthly series have a maximum in May and a minimum in October. The maximum increase in CO2 was observed in April, and the largest decrease in September.

The menu options for the seasonal decomposition using Loess smoothing are given in Figure 7.15. The options for the month plot are identical.

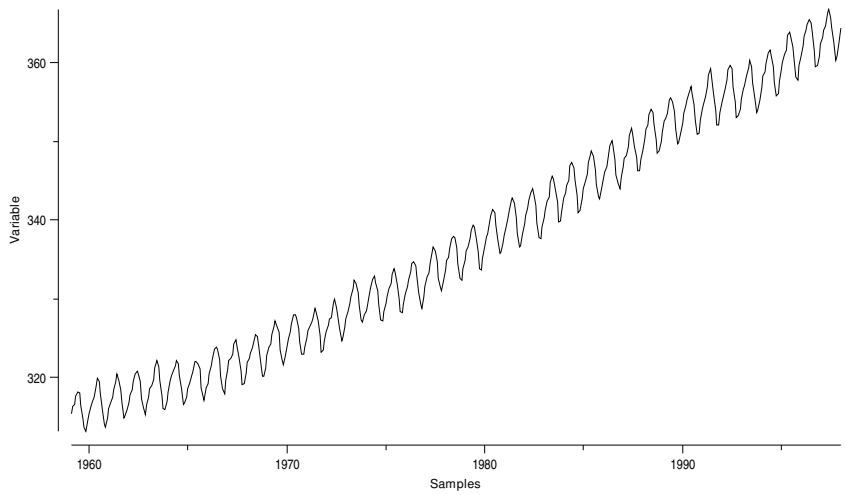

Figure 7.11. Time series plot of monthly CO2 data at Hawaii.

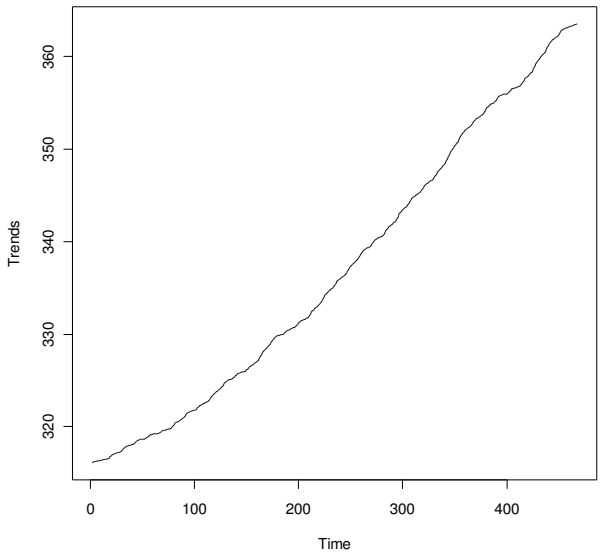

Figure 7.12. Long term trend obtained by seasonal Loess decomposition. of monthly CO2 data at Hawaii.

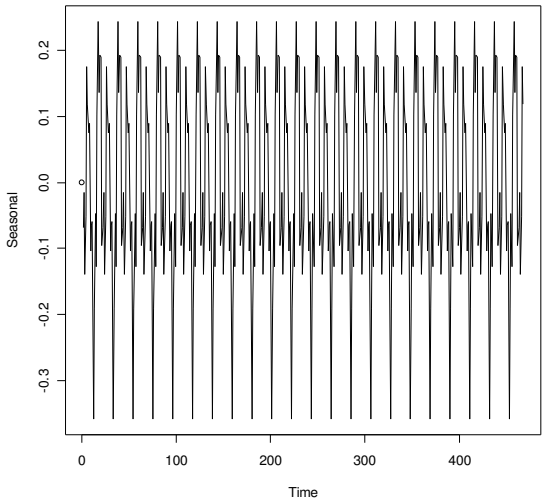

Figure 7.13. Seasonal component obtained by seasonal Loess decomposition. of monthly CO2 data at Hawaii.

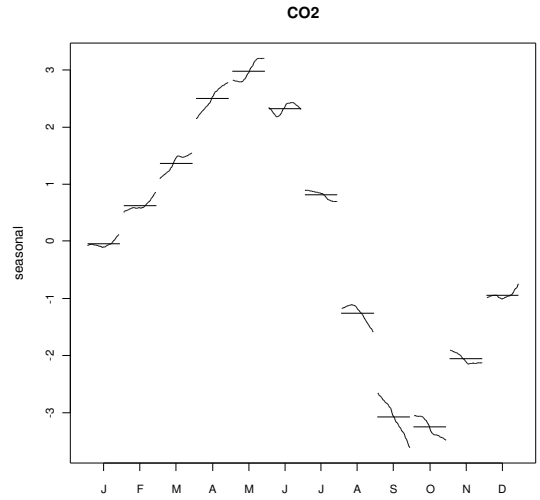

Figure 7.14. Month plot of seasonal component obtained by seasonal Loess decomposition. of monthly CO2 data at Hawaii.

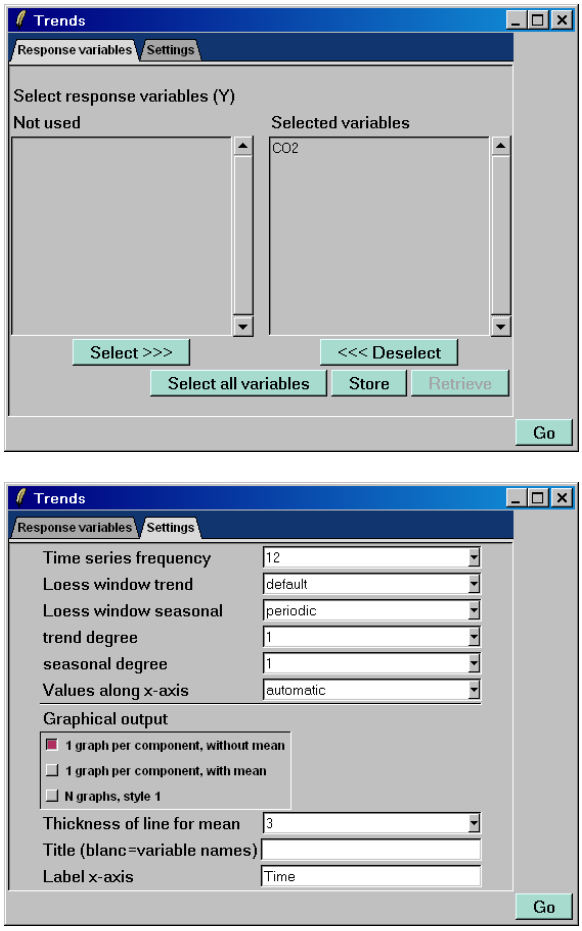

Figure 7.15. Options in Brodgar for the seasonal decomposition using Loess.

In the left panel, a response variable can be selected. The right panel shows the settings.

Time series frequency. This is the frequency of the time series. For monthly time series, this will be 12, and for quarterly time series it is 4.

Loess window trend and Loess window seasonal. The values determine the amount of smoothing that is used in the sub-series. The default value for the months is 'periodic' and means that mean values per month are taken. Alterna-

tively, the span, in terms of the lag, can be chosen. In Figure 7.13, we used a span of 21 for the seasonal component, and the default value for the trend.

Trend degree and seasonal degree. This value determines whether the Loess algorithm should use a linear regression model or polynomial model to obtain the smoothing values. See also the 'stl' help files of R. Setting it to 0 should result in smoother curves.

Graphical output. This can be 1 graph per component (with or without a mean curve), and N individual graphs (resulting in Figure 7.12 to Figure 7.14).

Thickness of line for the mean. The higher this value is, the thicker the line for the mean values.