3 Import data in Brodgar

In this chapter, we discuss how to import data into Brodgar, deal with missing values, save all your selections into a project, open such a project at a later stage, deselect/select response and explanatory variables, select a transformation and/or standardization, and potential problems and error messages.

Two important points to consider are: (i) different statistical techniques treat missing values in different ways, and (ii) nominal (categorical) variables need to be defined and imported differently for the methods in the R interface. Details are given in this chapter.

3.1 Import a data set for the first time

Start Brodgar and press on the main menu button labelled “Import data”. Main menu buttons are the menu buttons on the left hand side and top of the main window of Brodgar, see Figure 3.1. They are labelled as “Import data”, “Exploration”, “Univariate”, “Multivariate” and “Time series”.

After pressing the “Import data” button, a new window will appear with various tabs, see Figure 3.2. The tabs are labelled as:

- From spreadsheet
- Open project
- Info Y & X
- From ascii file
- Demo data

Y corresponds to response variables, and X to explanatory variables. By default, the window corresponding to the “From spreadsheet” tab is shown, see Figure 3.2.

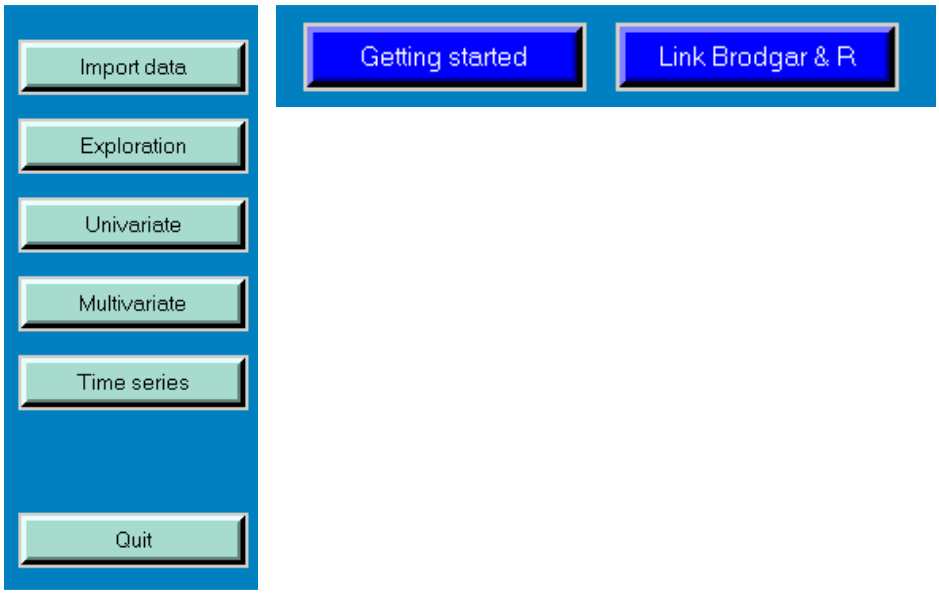


Figure 3.1. Main menu buttons are on the left, and general information buttons are at the top. Brodgar versions 2.3 and higher also have a 'View Data' button allowing the user to view (but not change!) the data.

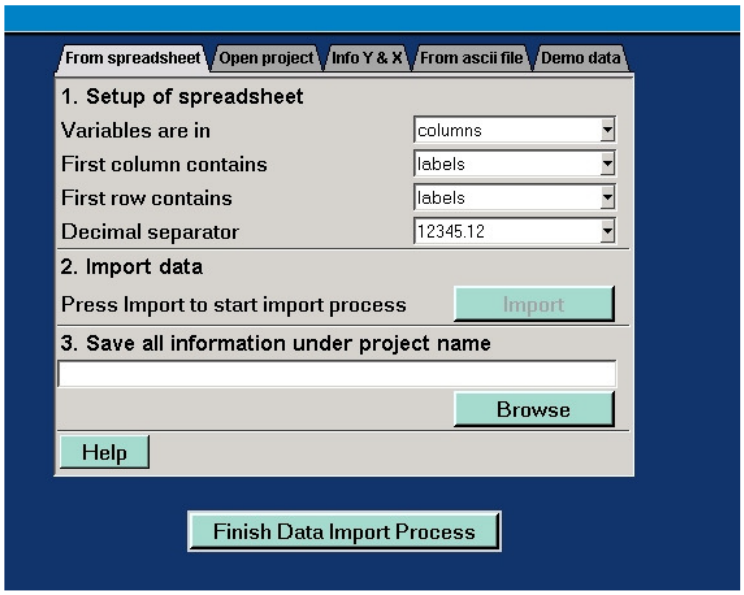


Figure 3.2. Window corresponding to the "From spreadsheet" tab.

Data can be imported in 2 different ways:

1. Copy data from a spreadsheet program to the clipboard and paste it in to Brodgar.
2. From an ascii file.

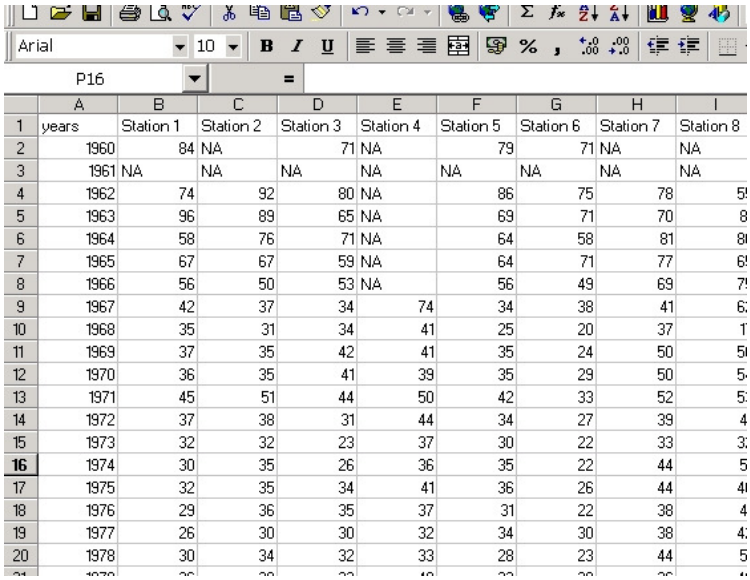
Each of these options is discussed next.

3.1.1. Import data from a spreadsheet

Suppose that the data are in an Excel file. See for example Figure 3.3, which shows part of an Excel spreadsheet containing numbers (catches) of a particular lobster species measured at 11 stations (or: sites), in the Atlantic Ocean south of Iceland between 1960 and 1999. The last column contains the NAO index, which is an explanatory variable. NAO stands for North Atlantic Oscillation index and represents the ratio of air pressure between Iceland and the Azores. Zuur et al. (2003b), among many other authors, used it as an environmental index function. The first row contains the names of all the variables (both response and explanatory variables), and the first column contains the samples (years). The value NA represents a missing value. For this data set, the stations are treated as response variables and the NAO as explanatory variables. Hence, the spreadsheet contains 13 columns and 40 rows. The first row contains the names of the response variables (11 stations) and the explanatory variable. The first column contains the 40 sample names (years). The data are in a sample-by-variable format, where variable represents both response variables and explanatory variables.

The process of importing data from a spreadsheet into Brodgar consists of three steps (see also Figure 3.2). In step 1, Brodgar needs to know whether (i) response variables are in the columns or rows of the spreadsheet, (ii) the first column contains labels or data and (iii) the first row contains labels or data. Default settings are: columns/labels/labels and can be changed by putting the mouse on top of the so-called combo-boxes in Figure 3.2 and clicking the left mouse button.

Internally, Brodgar stores the data in a sample-by-variable format, and it works with numbers of the format 12345.67 (a point as decimal separator). If are using a spreadsheet package with a comma as decimal separator (e.g. Spanish settings), please change Brodgar's default setting in the fourth combo-box of the first step (see Figure 3.2). Do not use data of the form 123,456.12. For large data sets it is faster to import data from a spreadsheet in the sample-by-variable format, and to use the point as decimal separator.



	A	B	C	D	E	F	G	H	I
1	years	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8
2	1960	84	NA	71	NA	79	71	NA	NA
3	1961	NA	NA	NA	NA	NA	NA	NA	NA
4	1962	74	92	80	NA	86	75	78	5
5	1963	96	89	65	NA	69	71	70	8
6	1964	58	76	71	NA	64	58	81	8
7	1965	67	67	59	NA	64	71	77	6
8	1966	56	50	53	NA	56	49	69	7
9	1967	42	37	34	74	34	38	41	6
10	1968	35	31	34	41	25	20	37	1
11	1969	37	35	42	41	35	24	50	5
12	1970	36	35	41	39	35	29	50	5
13	1971	45	51	44	50	42	33	52	5
14	1972	37	38	31	44	34	27	39	4
15	1973	32	32	23	37	30	22	33	3
16	1974	30	35	26	36	35	22	44	5
17	1975	32	35	34	41	36	26	44	4
18	1976	29	36	35	37	31	22	38	4
19	1977	26	30	30	32	34	30	38	4
20	1978	30	34	32	33	28	23	44	5

Figure 3.3. Part of an Excel spreadsheet containing lobster catches per year.

Once the appropriate choices have been made, go to the second step by clicking the button labelled “Import”. The window in Figure 3.4 will appear.

Before the data are copied, there are three important points to consider:

1. Missing values must be represented by NA. Brodgar transforms these to the value 999. Please check that your data set does not contain entities equal to 999, because these will also be considered as missing values.
2. If labels for both variables and samples are imported, the cell A1 must contain a value. In Figure 3.3 we used: years.
3. Do **not leave any cells empty within the data matrix.**

Highlight the data in the spreadsheet program (e.g. Excel), and copy it to the clipboard by pressing on CTRL-C or Edit|Copy. Please note that explanatory variables and response variables are imported at the same time. If the spreadsheet only contains the data, the entire spreadsheet can be highlighted. In Excel this can be done with one mouse click. Alternatively, entire rows or columns can be highlighted.

Once the data are copied from your spreadsheet program, return to Brodgar (Figure 3.4). Clicking on the “Cancel” button in Figure 3.4 will abort the data import process (it leads back to Figure 3.2). Click on “Continue” in Figure 3.4. This gives the window in Figure 3.5, which shows Brodgar’s own spreadsheet viewer. It allows you to check whether the data have been imported correctly. Please check carefully whether all (response and explanatory) variables and samples were selected. It is possible to edit, delete or change the values of the cells. Note that

variables will be in the columns, even if they were originally in rows. Click “Continue” if everything is correct. Press “Cancel” or “Previous” if not all variables, rows or columns were selected.

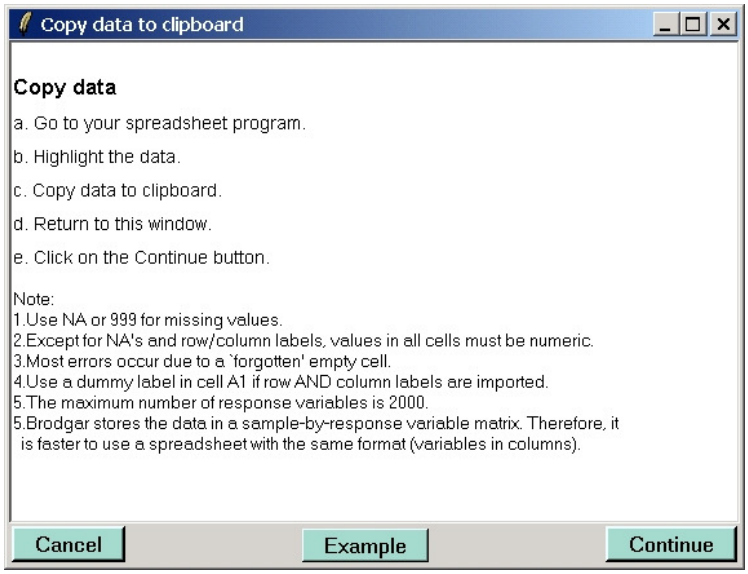
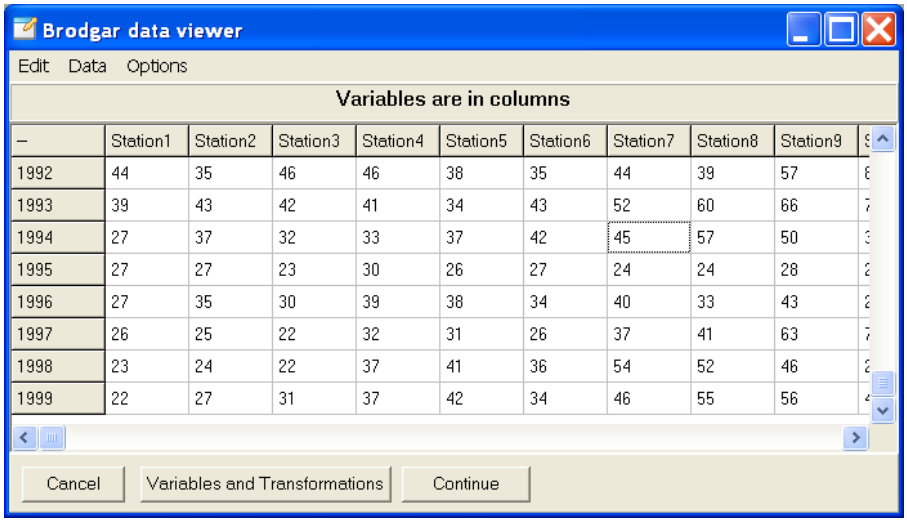


Figure 3.4. Copy data to the clipboard in Excel.



–	Station1	Station2	Station3	Station4	Station5	Station6	Station7	Station8	Station9
1992	44	35	46	46	38	35	44	39	57
1993	39	43	42	41	34	43	52	60	66
1994	27	37	32	33	37	42	45	57	50
1995	27	27	23	30	26	27	24	24	28
1996	27	35	30	39	38	34	40	33	43
1997	26	25	22	32	31	26	37	41	63
1998	23	24	22	37	41	36	54	52	46
1999	22	27	31	37	42	34	46	55	56

Figure 3.5. Brodgar’s spreadsheet viewer. You can also change row and column names and data values, and remove or delete columns in this spreadsheet.

After clicking the “Continue” button in Figure 3.5, we obtain Figure 3.2 again. The import process is now nearly finished. In the third and last step, the name of the project needs to be specified. Clicking on the “Browse” button gives not only the opportunity to specify the project name, but also the directory where to save it. Alternatively, just type in a project name. In this case, Brodgar stores the project name under the default working directory (the default working directory can be changed via: tools | set working directory). We used: c:/Mydata/Brodgar/fish.brd. If the extension .brd is not added, Brodgar will do this. Please do not use excessive long directory and project names. The maximum total length of directory and project name (together) should not exceed 300 characters. It is important not to omit the ‘:’ In ‘c:/Mydata/..’, or else one might not be able to obtain the numerical output, or view graphs.

Once a project has been specified, press the button labelled “Finish Data Import Process”. At this point, Brodgar will import the data, generate a directory c:/Mydata/Brodgar/fish, and store all information in this directory. If the first row and/or column contains data, Brodgar will generate row and column names. For the response variables and explanatory variables, these will be Y1, Y2, Y3, etc. For the observations, these will be 1, 2, 3 etc.

There are a few things that can go wrong:

1. When Brodgar reads the data from the clipboard, it assumes that the values are tab separated. If you copy data from Excel or any other spreadsheet package, this should be no problem. However, it may cause problems if data are copied from an ascii text file, since sometimes values are separated by spaces which are not tabs.
2. Missing values are represented by NA’s and Brodgar transfers these to the value 999. Please make sure that the original data do not contain 999 values because these will also be considered as missing values.
3. The button “Finish Data Import Process” was not clicked.
4. Make sure that the appropriate part of the spreadsheet is highlighted in the copy and paste process.
5. Make sure that there are no empty cells within the data. If you encounter any strange error messages, please check this point.

3.1.2 Import data from an ascii file

Please read Section 3.1.1. The import process via an ascii file is similar to the import process via the clipboard. Select the tab labelled “From ascii file” in Figure 3.2 and the window in Figure 3.6 will appear.

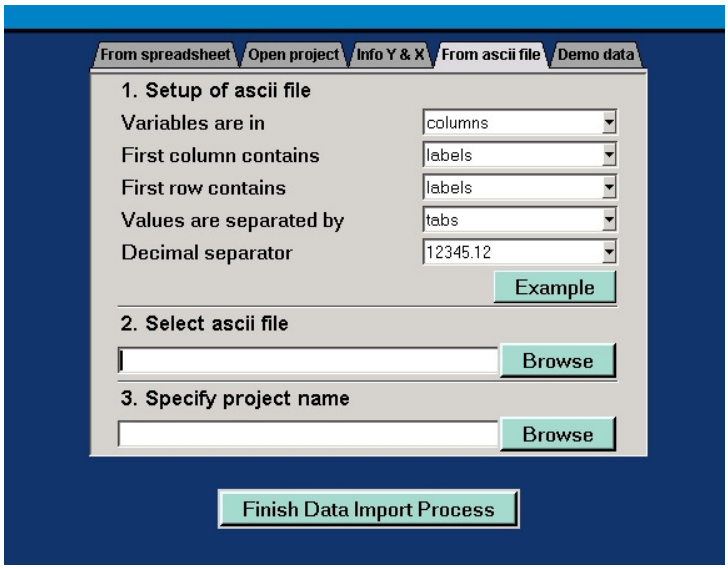


Figure 3.6. Import data from an ascii file.

First, you need to specify whether (i) response variables are in columns or rows, (ii) the first column contains labels or data, (iii) the first row contains labels or data, (iv) entities in the ascii file are separated by tabs, spaces or commas, and (v) whether the decimal separator is the point (default) or comma. The same data format as for the spreadsheet is required (NA for missing values, and if labels for both rows and columns are imported, the cell A1 needs to contain a value). In the second step, the ascii file needs to be selected. In the third step, a project name needs to be specified. Just as before, the button labelled “Finish Data Import Process” should now be clicked.

3.2 Open existing project

We will show how to retrieve an existing project. Click on the “Import data” main menu button and select the “Open project” tab. This will give the window in Figure 3.7.

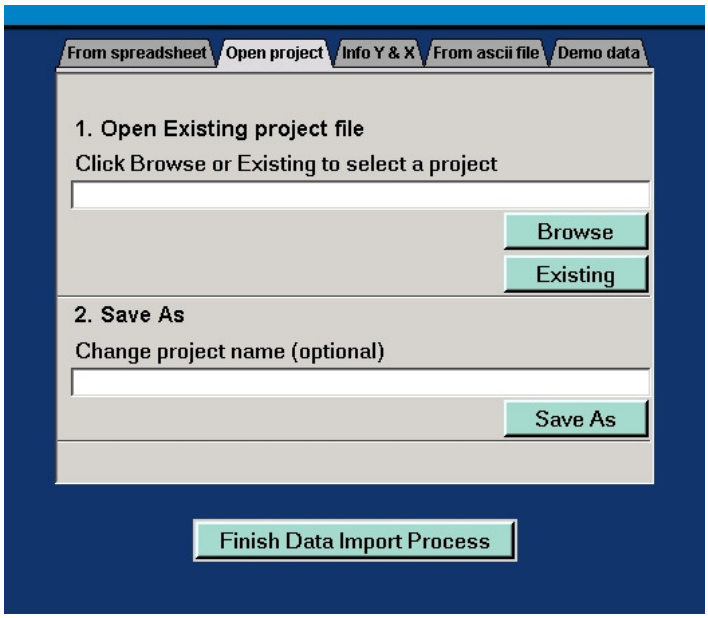


Figure 3.7. Open an existing project.

There are two steps, namely (i) select an existing project and (ii) change the project name. Step 2 is optional. Each of these steps are discussed in more detail.

If you remember the path name of the project you want to open, click on the "Browse" button. Alternatively, click on the "Existing" button. A window will pop up showing all existing projects (see Figure 3.8). Click once on the project you want to open and then click on the "Continue" button in Figure 3.8. After you click "Continue" in this window, the selected project will be opened. Brodgar will show the original data in its spreadsheet viewer. Just as before, have a quick look to see whether all variables are indeed present. Brodgar automatically retrieves the selected variables, transformation and standardisation. The delete button in Figure 3.8 allows you to remove projects. All imported data and results for the selected project will be deleted.

The second step, changing the project name, is optional. It is useful if you want to use the same data but with a different transformation and/or standardisation, or if you want to avoid overwriting the results of previous statistical techniques. If the only aim is to retrieve a project, ignore step 2. After steps 1 and/or 2, click the "Finish Data Import Process".

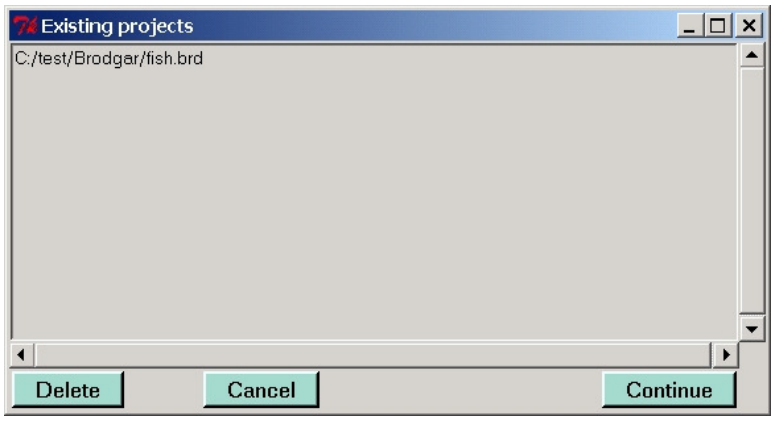


Figure 3.8. All existing projects.

3.3 The “Info Y & X” tab

Selecting the “Info Y & X” tab in Figure 3.2 gives the window in Figure 3.9. In step 1, transformations and standardisation of the variables can be selected. Y and X are the response variables and explanatory variables respectively. If the data do not contain explanatory variables, then ignore the choices for X.

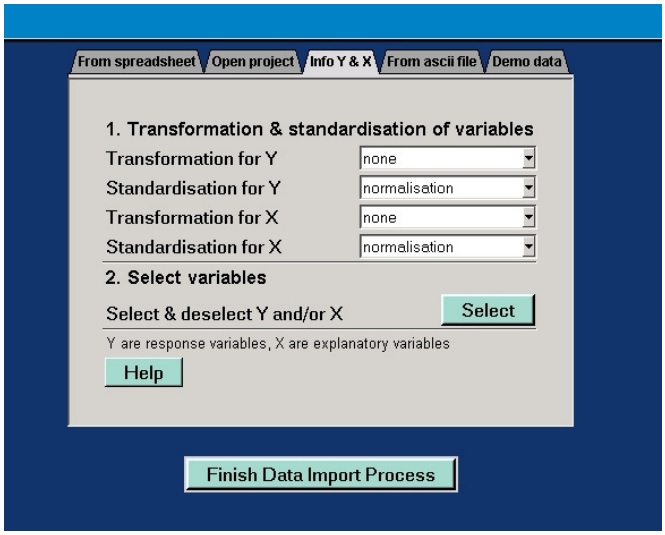


Figure 3.9. “Info Y & X” window.

Brodgar contains various transformations and standardisations, namely log10(Y), log10(Y+1), ln(Y), ln(Y+1), square root, arcsine, and ranking. If the original data takes the values: 2 7 4 9 22 40, the ranked transformed data will be: 1 3 2 4 5 6.

Brodgar contains various useful tools to help choosing a transformation. Further details when to choose which transformation can be found in Chapter 4 in Zuur et al. (2007).

Standardisation

If more than one variable is analysed, standardisation might be an option, though this also depends which statistical technique is applied. Standardising the response variables is sensible if for example dynamic factor analysis is applied, but methods like canonical correspondence analysis and redundancy analysis will apply their own standardisation (and therefore no standardisation should be applied by the user). One option is to centre all variables around zero:

$$Y_i^{new} = Y_i - \overline{Y}$$

Where *i* is the mean and Y_i the value of the i^{th} observation. The most commonly used standardisation is given by:

$$Y_i^{new} = (Y_i - \overline{Y}) / \sigma_y$$

where σ_y is the sample standard deviation. The transformed values Y_i^{new} are now centred around zero, have unit variance and are unit-less. This transformation is also called normalisation. Other, less used transformations are:

$$Y_i^{new} = Y_i / Y_{max} \quad \text{and} \quad Y_i^{new} = (Y_i - Y_{min}) / (Y_{max} - Y_{min})$$

These two transformations are called ranging 1 and ranging 2 respectively, and they rescale the data between zero and one.

Selecting variables

Selecting response variables and explanatory variables in Brodgar is done in step 2. By default, all imported variables are considered as response variables. Click on the button “Select” in Figure 3.9, this gives the window in Figure 3.10. To deselect or select a variable, click on any cell that has the value “Yes” or “No”. Brodgar will also ensure that variables are not a response variable and explanatory variable at the same time. To select (or de-select) multiple variables, keep the left mouse button pressed as you move the pointer.

Once a selection has been made, click on the “Continue” button in Figure 3.10 to return to Brodgar. You need to press the “Finish Data Import Process” button to invoke all changes.

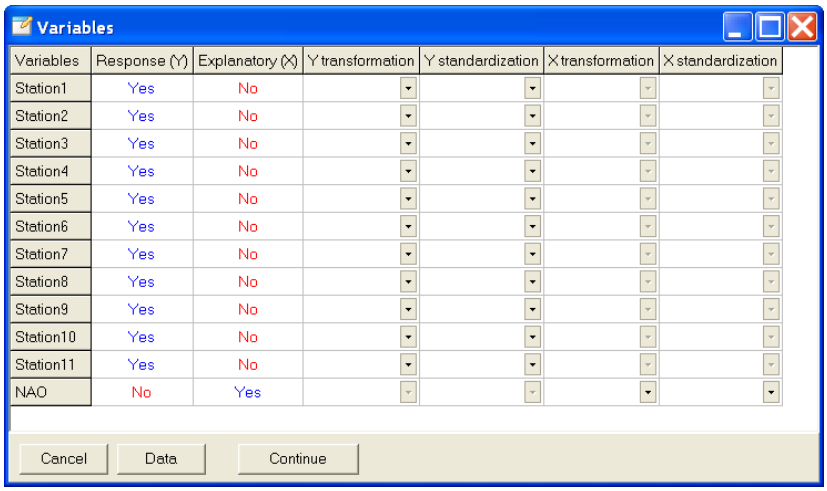


Figure 3.10. (De-)selecting response and explanatory variables. You can also select different transformations and standardisations per variable.

Whichever technique is used in Brodgar, it will use the selected response variables and explanatory variables that were obtained in this step. If you want to apply a technique with a different set of response variables or explanatory variable (e.g. PCA on explanatory variables), the changes need to be made in the Import data process.

Figure 3.10 also shows that it is possible to apply transformations and standardisations on individual variables, instead of the same transformation on all variables. Fortunately, it will not apply two transformations on the same variable, should you have selected one in Figure 3.9 and in Figure 3.10

3.4 Demo data

There are various demo data sets in Brodgar. These can be accessed by clicking on the “Demo data” tab in Figure 3.2. Click the combo-box under “Select a demo data set“ and select one of the data sets. Once a selection has been made, the associated project file will appear automatically under step 3 (3. Specify project name). Click the “Load data” button and the spreadsheet viewer will show the selected data set. Click the “Continue” button in the spreadsheet viewer, followed by the “Save Changes and Finish Import Data Process” button. Transformations and standardisations are applied automatically.

Brodgar contains various demonstration data sets, namely:

- A lobsters time series data set.
- The hunting spider data.

- A zoobenthic data set from Argentina. Zoobenthic species live in the soil of intertidal areas.
- A zoobenthic data set from The Netherlands.
- A wedge clam data set.
- A squid data set from different ICES areas.
- Fisher's Iris data.
- Harbour porpoise fatty acid data.

These data sets can be used to explore all the techniques available in Brodgar. A short description of these data sets is presented below.

Lobster time series data

This data set consists of time series of catches of a particular lobster species in 11 areas in the Atlantic Ocean between 1960 and 1999. The data were available on an annual basis. Dynamic factor analysis can be used to detect common trends in the time series, interactions between the series, and relationships with explanatory variables like the NAO index. The full analyses of these data are presented in Chapter 17 in Zuur et al. (2007).

Spider data

This is a data set used by many ecologists and statisticians to demonstrate how canonical correspondence analysis and related techniques can be used to detect species-environmental relationships. The data set concerns the distribution of wolfspiders in a dune area. The data were originally published in van der Aart & Smeenk-Enserink (1975). Since then these data have been used in various publications to illustrate ordination techniques. The original data consist of counted abundances of twelve species captured at 100 sites (pitfall traps) in a dune area in the Netherlands. Six environmental variables were measured at 28 sites. These were water content, bare sand, moss cover, light reflection, fallen twigs and herb cover. We use the observed abundances at the 28 sites where environmental variables were monitored. Species data are transformed by taking square roots.

Zoobenthic data from Argentina

These data were used in Chapters 4 (data exploration) and 28 (Mantel test) in Zuur et al. (2007). The data consist of measurements on four zoobenthic species in three transects. Each transect contained ten sites, and all sites were measured in Autumn and Spring, resulting in a 60-by-4 data matrix for the species data. There are also a couple of explanatory variables like mud content, etc.

Zoobenthic data from The Netherlands (RIKZ)

We called these data “the RIKZ” data in Zuur et al. (2007; 2009), and used them to illustrate linear regression, GAM and linear mixed effects modelling. The response variable is species richness (the different number of species present at each site), and all other variables are covariates.

Wedge clam data

We use this data set as a first (and basic) exercise for linear regression analysis in our courses. Biomass (AFD) is modelled as a function of length and month. There is one outlier, and to linearize the relationship, and stabilize the variance, biomass and length need a log transformation. Make a scatterplot of AFD versus length, and use the row number as labels (see Scatterplot - Settings) to identify the identity of the outlier. Log-transform AFD and length.

Squid data

We used these data in Section 7.7 in Zuur et al. (2007) to illustrate cross-validation in GAMs. GSI is modelled as a smoothing function of month, and year, location and sex as categorical explanatory variables.

Fisher’s Irish data

From the R help file on this data set: This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*. Apply discriminant analysis on these data.

Harbour porpoise data

Blubber fatty acid composition data in harbour porpoises were used in Chapter 29 of Zuur et al. (2007) to illustrate principal component analysis.

You are now familiar with all data import functions in Brodgar. There are various ways to proceed from here and to start with, data exploration is discussed in the next chapter.

3.5 Missing values and nominal variables

Most methods available from the R interface can handle missing values. Formulated differently, all cells in the spreadsheet that have an NA are dealt with by the statistical methods in R. In most cases, this means that rows with missing val-

ues are omitted (e.g. in GAM). In the ordination methods, missing values are replaced by mean values per column. Further details can be found in the help files, which are available from push buttons in Brodgar.

Nominal variables are variables of the form: yes/no or yellow/green/blue, or transect 1/transect 2/transect3, or observer 1/observer 2/observer 3. Other examples are month, year (if the data are not time series data), etc. First of all, these variables need to be transformed to numerical values. The first example is the easiest; let a sample have the value 1 for yes, and 0 for no. Hence, we end up with an extra column containing only zeros and ones. The same principle holds if the variable has three classes; use a 1 for yellow, 2 for green and 3 for blue (or 1 for January and 12 for December). Hence, we have a new column with ones, twos and threes. Recall that Brodgar can only import numerical data, therefore the column with yellow/green/blue needs to be removed. All this data manipulation should be done in for example Excel. The same can be done for the other examples. If methods from the R interface are used, no further Excel data manipulations are required; the program will allow you to identify nominal variables before analysis are carried out.

However, all the other statistical methods (most of the multivariate and time series techniques) require further manipulation in Excel or else the new variables will be considered as quantitative variables. Table 3.2 shows an example for colour. The first column contains the sample number, the second the colours. Table 3.3 shows the coding required for the R-interface techniques. Note that all data are numeric. All methods in the R-interface can cope with these data. Table 3.4 shows how to prepare the data (e.g. in Excel) so that (non R-interface) ordination and time series methods can be applied. The new columns are called dummy variables. If colour is an explanatory variable, two of the three classes need to be selected as explanatory variables during the data import process, and one should not be selected. This is because the ordination and time series methods will consider yellow, green and blue as different explanatory variables; but they are linearly correlated. It does not matter which one is not selected. Because it is likely that both non R-interface and R-interface methods will be applied, we advise to import data that contains both the factor and the dummy variables, see Table 3.4. This table shows the data, as we would advise to import it into Brodgar. Once these data have been imported, variables can easily be deselected and selected depending on the statistical technique. Depending on which technique is used, a particular class of colour can be deselected.

Table 3.2. Artificial data to illustrate coding of nominal variables.

Observation	Colour	Other variables
1	Green	…..
2	Yellow	…..
3	Blue	…..
4	Blue	…..
5	Green	…..
6	Yellow	…..

Table 3.3. Artificial data to illustrate coding of nominal variables. The numbers 1, 2 and 3 represent yellow, green and blue respectively.

Observation	Colour	Other variables
1	2	…..
2	1	…..
3	3	…..
4	3	…..
5	2	…..
6	1	…..

Table 3.4. Artificial data to illustrate coding of nominal variables.

Sample	Colour	Yellow	Green	Blue	Other variables
1	2	0	1	0	…..
2	1	1	0	1	…..
3	3	0	0	1	…..
4	3	0	0	1	…..
5	2	0	1	0	…..
6	1	1	0	0	…..

3.6 Problems and error messages

Here are some common problems encountered by people who just started to use Brodgar.

Problem 1

Changes are made to the status of response variables and explanatory variables (e.g. a variable was changed from response to explanatory under Data Import – Info Y & X), but in all the techniques it is still enumerated as a response variable.

Solution: The reason for this is that the user did not click on the “Finish Data Import Process” button.

Problem 2

The user imports the data without any problems, but despite the fact that Brodgar can access R, none of the methods using R is working.

Solution: It is highly likely that row or column labels are of the form:

- Myname%1
- Myname$2
- Myname#3

- Myname-4
- Myname*1

The use of long names and names containing characters of the form *, &, #, etc. should be avoided. The name Y1 should be avoided as well. Brodgar replaces spaces in the variable names by dots.

Problem 3

The data are copied in Excel, and all other steps of the data import process are carried out correctly in Brodgar. But if the "Finish Data Import Process" button is clicked, the following error message appears:

```
can't set "Yout1(…)": syntax error in expression "…"
   while executing
"set Yout1($i) $DATA($s,$i)"
etc.
```

Solution: It is highly likely that the data contain alpha-numerical values, or empty cells. See Section 3.1 for further details on how to solve this.

Problem 4

The user imports the data correctly, but all the FORTRAN techniques (e.g. dynamic factor analysis, chronological clustering, PCA, RDA, CCA, CA) crash.

Solution: This happens if excessively long values are used, e.g. a number of the form 12345678901234567,1234. Brodgar stores data in the FORTRAN format F15.5. This means that each value is allowed to contain 15 numbers before the decimal separator, and up to 5 decimal values. Hence, the value 0.12345678 is stored as 0.12345. The value 1234567891234567.0 will result in a crash because it contains more than 15 numbers before the decimal separator.

Problem 5

The data are imported without any problems and the R techniques work fine, but the principal component analysis crashes.

Solution: This happens if there are variables with always the same value (variance is 0). You will need to remove them from the analysis.

Problem 6

The user can load demo data without any problems, but most other buttons under Data Import menu are disabled.

Solution: The (capital sensitive) licence codes were not typed in correctly, or a demo version is being used.

Problem 7

The package works fine for the demo data sets, and most of the statistical techniques work fine for data sets containing up to a couple of thousand variables and samples. However, for data sets containing more than 50000 samples and variables, the package becomes very slow.

Solution: Unfortunately, there is no solution to this problem. Brodgar was designed to analyse small and intermediate data sets. Occasionally, someone buys Brodgar and wishes to analyse very large data sets (e.g. 10000 time series measured every minute for a couple of months). Analysis of such data requires FORTRAN or C++ programming. And some of the techniques can cope better with larger data sets than other techniques.

Problem 8

A regression analysis, GAM or GLM (or any other R technique) is applied, Brodgar displays the graphical output, but nothing happens if the 'Numerical output' button is clicked.

Solution: Brodgar gives the operating system the command to open a txt file. In principle, your operating system associates a *.txt file with notepad. If this is not the case, start Windows Explorer, and select: Tools - Folder options - File types - New, and follow the instructions. Alternatively, go to your project directory using Windows Explorer, and open the most recent generated *.txt file. On Windows XP, make sure that the project name is of the form: "C:/MyDirectory/Etc/…" and not "C:MyDirectory/Etc/…". For all operating systems, make sure that the length of the complete path name of the project is not too long. If none of these solutions solve the problem, then it might help to copy and past the Windows program notepad.exe in the Brodgar installation directory. This is especially the case for Windows XP-Professional.