

CHAPTER 3

3. Importing data

One of the first things you probably need to do is to import data into S-PLUS. Many formats used by other (statistical) packages can be imported directly, which may stimulate a migration to S-PLUS.

3.1 Import

S-PLUS can import and export the following formats directly. See Section 5.6 for a detailed description of how to export data.

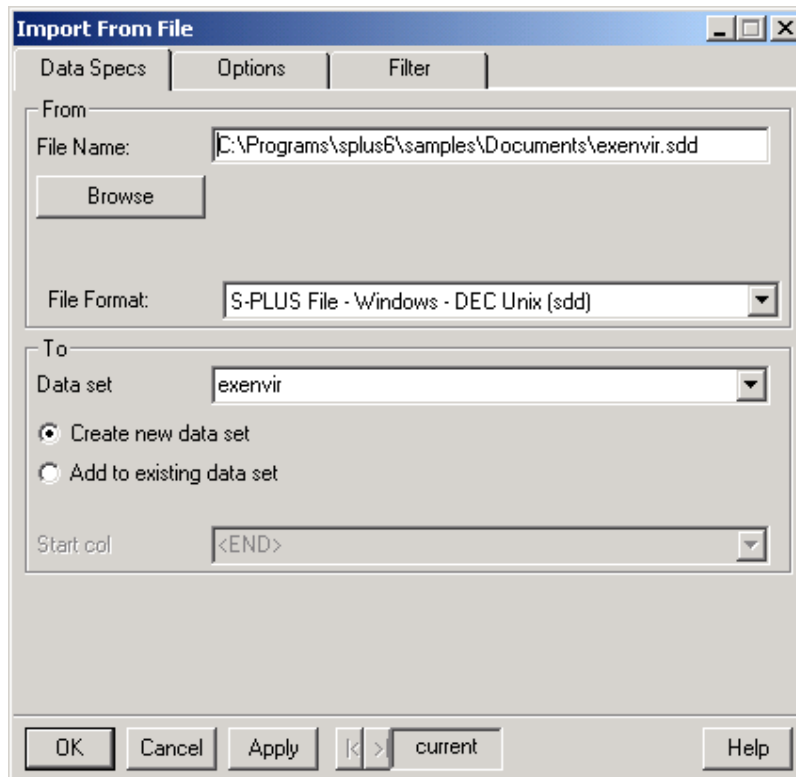
- ASCII-files
- Microsoft Excel and Quattro Pro spreadsheets
- Lotus sheets, Paradox
- dBase files
- FoxPro files
- Systat files
- SPSS files (and SPSS Export files)
- SAS files (and SAS Transport files)
- Microsoft Access files
- Matlab files
- S-PLUS transport files
- STATA files
- Gauss files (.DAT automatically reads the related DHT)

Financial data from FAME, LIM and Bloomberg, and data from databases that support the ODBC standard can also be imported directly. S-PLUS supports ODBC version 2.0 and 3.0.

The following example describes how to import the S-PLUS file called `exenvir.sdd`. This file can be found in the directory `Samples\Documents`, which is located in the installation directory of S-PLUS. Proceed as follows:

1. Go the **File** menu and choose **Import Data? From File**. An import dialog with with three tabs will appear.

3. Importing data



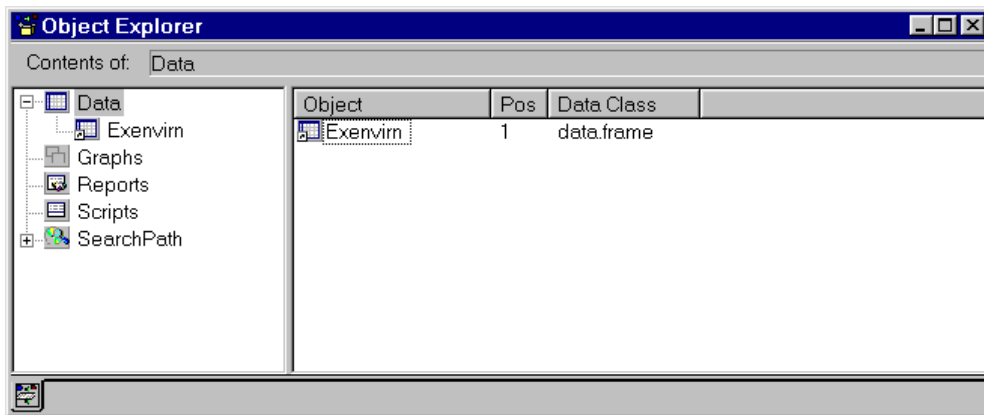
2. Select 'S-PLUS File - Windows' in the field 'File Format'.
3. Select the file 'exenvir.sdd'.
4. Importing a data file in S-PLUS will result in an S-PLUS data frame. Enter a name for the data frame in the field 'Data set'. The default name for the data frame is that of the data file, without extension, so in this example 'exenvir'.
5. Click 'OK' to import the file.

	1	2	3	4	5
	Ozone	Radiation	Temperature	Wind	
1	41.00	190.00	67.00	7.40	
2	36.00	118.00	72.00	8.00	
3	12.00	149.00	74.00	12.60	
4	18.00	313.00	62.00	11.50	
5	23.00	299.00	65.00	8.60	
6	19.00	99.00	59.00	13.80	
7	8.00	19.00	61.00	20.10	
8	16.00	256.00	69.00	9.70	
9	11.00	290.00	66.00	9.20	
10	14.00	274.00	68.00	10.90	
11	18.00	65.00	58.00	13.20	
12	14.00	334.00	64.00	11.50	
13	34.00	307.00	66.00	12.00	
14	6.00	78.00	57.00	18.40	
15	30.00	322.00	68.00	11.50	

If the file is successfully imported, the data is stored in an S-PLUS data frame and is displayed in a data window.

A data frame, one of the S-PLUS data structures, is a very convenient structure for data analysis. Usually the rows of a data frame correspond to the different observations and the columns correspond to the different variables. Data frames will be discussed in detail in Chapter 5.

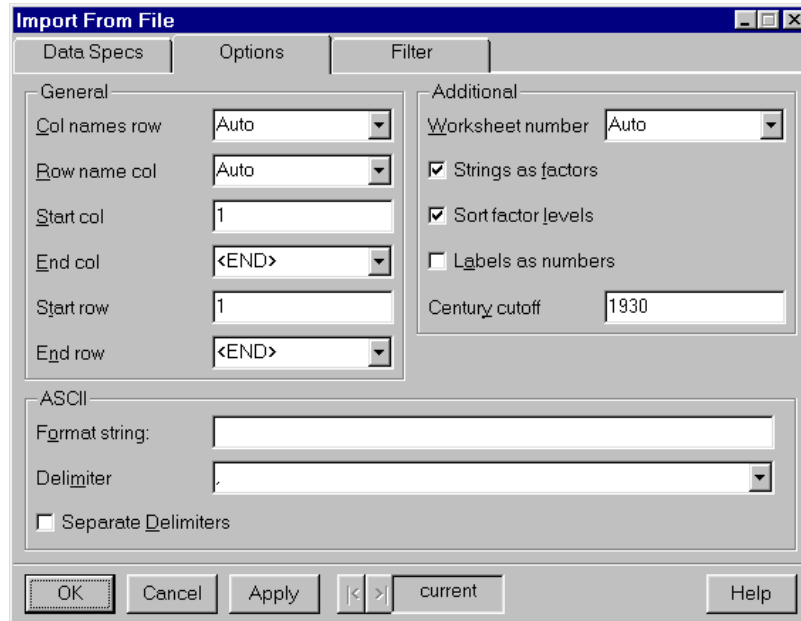
By default, S-PLUS shows the imported data in a data window. For instructions how to change this see Appendix B. The ‘exenvirn’ data frame in S-PLUS is stored in the *working database*. It has become a separate object in S-PLUS without any link to the original ‘exenvirn.sdd’ file: changing the S-PLUS ‘Exenvirn’ data frame does not affect the original data file. The new object will be shown in the Object Explorer, as in the next figure. We will discuss the working database and the Object Explorer in the next chapter.



If you make changes to the ‘Exenvirn’ data frame and then close the data window, S-PLUS asks if you want to save the changes to a file. Choosing ‘No’ doesn’t mean that the changes are not saved; they are stored as an internal S-PLUS object, which you can use later in S-PLUS. Choose ‘Yes’ if you want to keep an “external” data file that you can share with another S-PLUS user. (You can deactivate this prompt if you wish. See Appendix B for details.)

3.2 Import options

The default import options are adequate in most cases, but it may be necessary to set specific import options. To do this, use the Options Tab of the Import dialog.



3.2.1 Column and row names, data block location

Use the options 'Start Col', 'Start Row', 'End Col' and 'End Row' to specify a block of data to be imported. This is not only convenient when importing only a smaller part of the data, but also to prevent S-PLUS from importing 'unwanted' data. Consider the following Excel sheet.

	A	B	C	D	E
1					
2					
3		garbage here, don't import			
4					
5		july 780	august 875		
6					
7			var1	var2	var3
8			12.45	1.04	m
9			34.56	1.03	f
10			53.56	1.09	m
11					
12					

To import the data correctly in S-PLUS you need to type '7' in the field 'Col names row', '3' in the field 'Start col' and '8' in the field 'Start row'.

3.2.2 Delimiters

When importing text files, you have to specify a delimiter character which tells S-Plus how to separate the text into columns. Default delimiters are commas, spaces or tabs (\t). But if the data in your text file is separated by semi-colons, you type the ';' symbol in the 'Delimiter' field in the Option tab of the Import dialog.

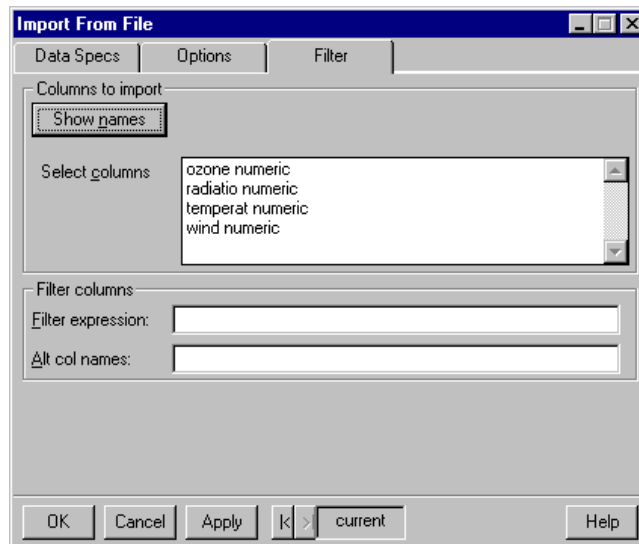
3.2.3 Importing characters

If a data file has a column with text, S-PLUS will import this text column as a factor column by default. A column with "male", "female", "male", for instance, will be imported as a factor column with two levels: "male" and "female". If you want a text column to behave as a character column in S-PLUS, you need to uncheck the checkbox 'Strings as factors'. Chapter 8 discusses the difference between factor columns and character columns.

3.3 Filtering data

By default, the complete data set is imported into S-PLUS. When you do not need all the data, you can import only a part of the data into S-PLUS by setting a filter. You can select which columns you want to keep and you can specify a filter expression to import only those rows that satisfy the filter expression.

Let's have another look at the SAS file Exenvirn.sd2. Having selected this file, we select the 'Filter' tab of the Import dialog and click the button 'Show names' to have an overview of the columns (see next figure). We then select the columns to import, for example 'ozone' and 'wind'.



The column list also indicates the data type of each column (numeric or factor). See section 5.2.

3.3.1 Filter expressions

A filter expression is an expression that consists of variable names and logical operators. You can use the following logical operators:

==	is equal to;	!=	is not equal to;
<	is smaller than;	>	is larger than;
<=	is smaller than or equal to;	>=	is larger than or equal to;
&	and;		or.
!	not;		

The variable names in the filter expression must be known in the data file. Examples of filter expressions:

```
ozone > 2.4
ozone > 1.3 & ozone < 3.2
```

```
Sex == "Male"
Sex == "Female" & Bloodtype != "A"
```

Note When a data file does not have column names, S-PLUS will generate default column names (like Col1, Col2, Col3 etc). These names must be used in the filter expressions.

3.3.2 Block reads and writes

A new feature in S-PLUS 6 is the ability to process data sequentially, using block reads and writes. Instead of importing all the rows of a data file one after another, it is possible to read the first 10,000 rows, process these, then read the next 10,000, process these etc. This requires some knowledge of the S language, since the block reads and writes are not integrated in the menu system. We will give some examples in section 9.5.

