

GRAPHICAL AND NUMERICAL ANALYSIS OF SPATIAL DATA WITH A GUI AND R



R. DUTTER

Vienna University of Technology

Vienna, AUSTRIA


e-mail: R.Dutter@tuwien.ac.at

Abstract

The paper is on the basis of DAS+R¹, a package of  where the emphasis is on a user friendly menu, a graphical user interface (GUI) enabling the call of different -functions which otherwise would afford a cumbersome typing in of possibly complicated commands for sophisticated methods of statistical data analysis. The package is still under development, nevertheless the actual state of the program system already enables the user to start easily with standard methods and continue with more complicated methods if the clickable commands already exist or by switching to the command line modus. We emphasize the repeatability of the generation of commands in several ways to intensify the speed of obtaining senseful results. A special view is put on the analysis of spatially depending uni- or multivariate data, particularly on problems of geochemical data.

We describe in short the important features of data analysis in this field with realized functionalities with its clickable icons in DAS+R. We also give short illustrations on practical examples with geochemical, spatial data.

1 Introduction

This paper is on the basis of DAS+R, and formally a short description of it, a package of  still under development using a graphical user interface which should ease the application of more or less sophisticated methods.

The basis of the graphical user interface comes from the R Commander (see Fox, 2004). It uses Tcl/Tk programming tools (see Welch and Jones, 2003). The emphasis is on the analysis of spatially depending uni- or multivariate data, particularly on problems of geochemical data.

Three special properties of DAS+R should be stressed:

- Interactive definition of data subsets (numerically or graphically) together with set operations. Usage of these subsets in almost all graphics and computations.
- Intensive use of possible relations between the geographical information with the values of data in the statistical and graphical analysis.
- The strong requirement of fast reproducibility and repeatability with small variations in the analysis.

The main part of the talk will be with life illustrations:

¹<http://www.statistik.tuwien.ac.at/StatDA/DASplusR/>

1. Very basic statistical examples which could also be used successfully with very little effort, in elementary courses.
2. Graphical illustrations and the corresponding analysis.
3. Spatial data and their relations, e.g. applying kriging.

The main window of the GUI is shown in Figure 1. It consists of (from top to bottom) a menu bar, a toolbar, a script window, an output window, and a messages window. Most of the R-commands are entered via a menu in the menu bar. In Figure 1 the entry **Advanced** is clicked such that further submenu buttons show up.

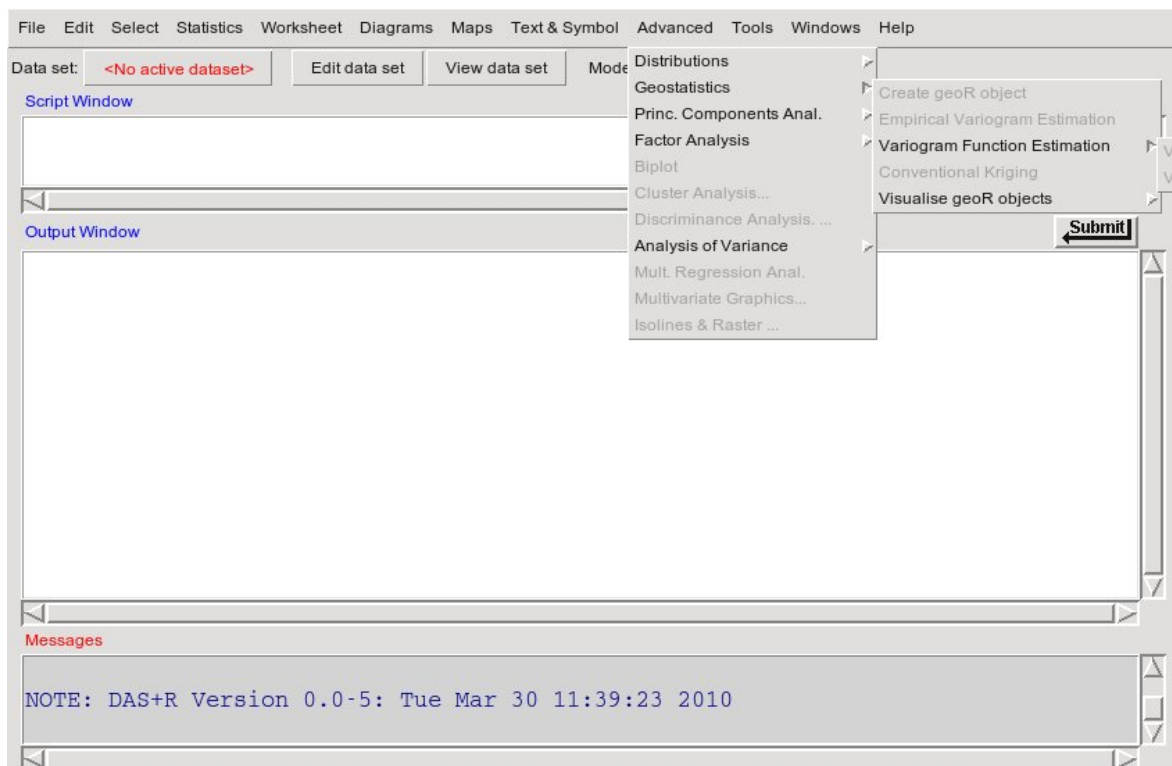


Figure 1: Main Graphical User Interface of DAS+R



2 Important Features in the Implementation of Data Analysis

Here we describe in short the contents of the main entries in the pull down menu in Figure 1.

2.1 Useful Data Structures and Entering Data

The following discussion on user oriented data structures is concerned about the entry ‘file’ in the pull down menu of DAS+R and is by no means complete or exhaustive but shows some useful features of data coming from people working in applied fields like geosciences.

Data to be analyzed may be entered in different ways:

- By numbers and/or characters to fill a table. (This case will not be very frequent because usually, the data will be prepared beforehand on some storage median in some sort of format.)
- From an spreadsheet, usually in a csv-format.
- A data set is already available in the -workspace.
- A data set previously stored in a .R-dump file.
- A data set stored in another package of .
- A data set stored in a file with a known ‘foreign’ format like SPSS, Minitab, STATA, etc. .

We only comment on data from spreadsheets: Besides getting numbers or character strings for ‘data values’ in a sort of matrix of values, it is very interesting (and usual) to attribute to each column (variable) a unit, the sort of generation of the data value, and some other comments for each variable. This can be made available by using some common specific keywords in the rows of the spreadsheet. We can see an illustration in Figure 2.

	A	E	F	G	H	I	J	K	L	M	N	O
1	HEADER						KOLA PROJECT, regional sampling 1995 (Finland (FIN), Norway (NOR))					
2	COMMENT DATASET						C-Horizon of Podsol profiles, air dried, fraction <2 mm, nylon screen					
3	SAMPLE IDENTIFIER											
4	COORDINATES											
5	COMMENT VARIABLES							all <DL				
6	EXTRACTION						Aqua Regia	TOTAL	Aqua Regia	TOTAL	TOTAL	Aqua Regia
7	METHOD						GF-AAS	INAA	ICP-AES	XRF	XRF	GF-AAS
8	UDL											
9	LDL						0	5	10	300		0.1
10	UNIT	mg/kg			cm		mg/kg	mg/kg	mg/kg	mg/kg	wt.-%	mg/kg
11	VARIABLE	ELEV	COUN	ASP	TOPC	LITO	Ag	Ag_INAA	Al	Al_XRF	Al2O3	As
12		135	FIN	NW	35	20	0.01	2.5	10200	75100	14.19	0.3
13		140	RUS	SW	52	4	0.01	2.5	3540	74400	14.06	0.2
14		255	FIN	N	52	31	0.02	2.5	16100	80600	15.24	0.4
15		240	RUS	NE	40	20	0.02	2.5	12000	79300	15	0.3
16		80	NOR	N	50	10	0.02	2.5	9850	72800	13.76	2.2

Figure 2: Excerpt of Spreadsheet of Data

2.2 Editing

The entry ‘Edit’ allows specification of editing of the data set, the variables and cases, and all kind of parameters. Furthermore, subsets can be generated (numerically or graphically) and/or edited.

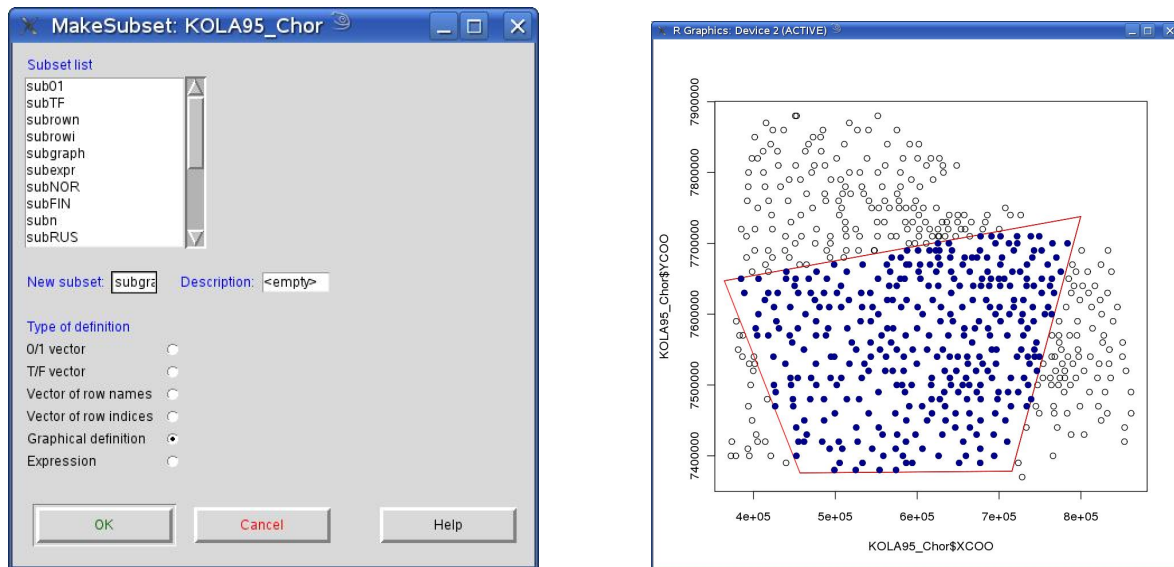


Figure 3: Graphical Generation of a Data Subset

We illustrate this by a simple definition of a subset graphically. Figure 3 on the left side shows the widget for specifying necessary attributes like subset name, type of definition, optionally description, etc.

The second part of Figure 3 shows this graphical definition by clicking at the points of a polygon line which encloses the subset.

2.3 Selecting

From a complete (probably huge) data set, variables, cases and/or subsets can be selected on which will be specially concentrated in the further analysis.

2.4 Statistics

Different statistics, e.g. all kind of summaries of the data and simple tests, are available.

2.5 Worksheet

The main idea is that some graphics should be nicely plotted on a ‘worksheet’ with size to be specified. A ‘worksheet’ is splitted into frames in which a graphical output of R can be placed. These frames can have any size and can be placed anywhere in the worksheet (also in an overlapping manner) (see Figure 4 for a possible, final result).

2.6 Diagrams (Graphics)

Different standard graphics can be produced: histograms, density plots, cdf’s, boxplots, xy-plots, ternary, ...

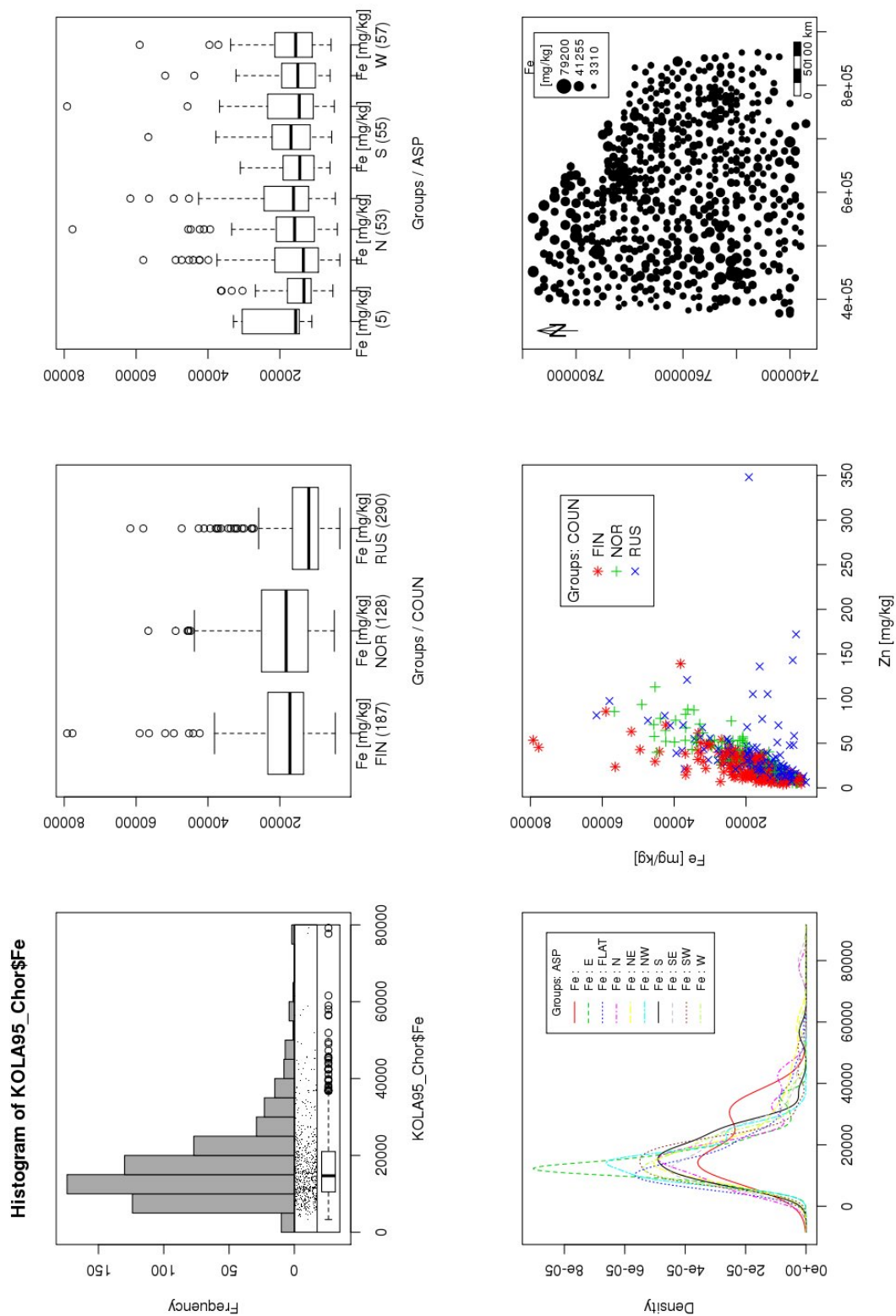


Figure 4: Worksheet with 6 Frames

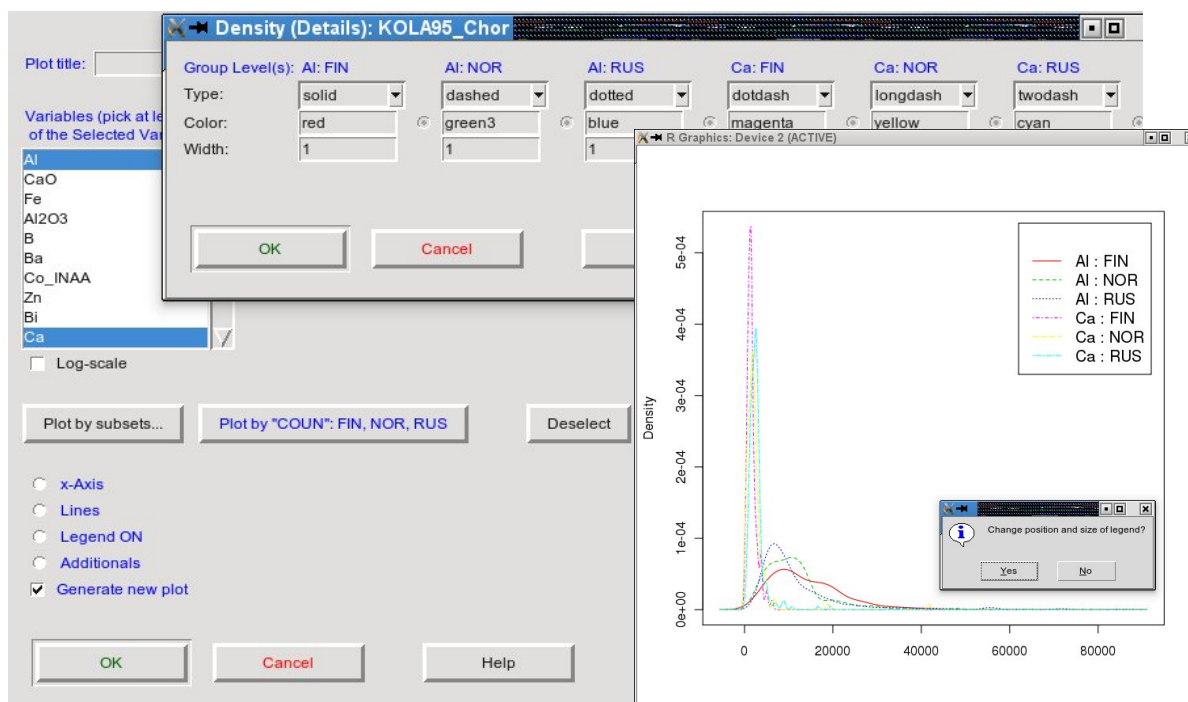


Figure 5: Widgets for Density Plots

An illustration is given in Figure 5 with 3 widgets for plotting density traces: The deepest one shows the specification of two variables (Al and Ca) and grouping of the data by the variable (factor) COUN with 3 levels. The next higher widget shows the default specifications for all the combinations of the 2 variables and the 3 group levels FIN, NOR and RUS. Finally, a graphics window with the 6 requested density traces is shown together with a legend and a question if its position should be changed (interactively).

2.7 Maps

Background maps can be imported and used, data plotted in form of maps via different 'symbols', proportional dots, surface maps using simple interpolation or even kriging.

2.8 Text & Symbols

Text and/or symbols may be added to a graphic.

2.9 Advanced

The entry 'Advanced' opens the following menus.

- Distributions (random numbers!)
- Geostatistics (spatial data!)

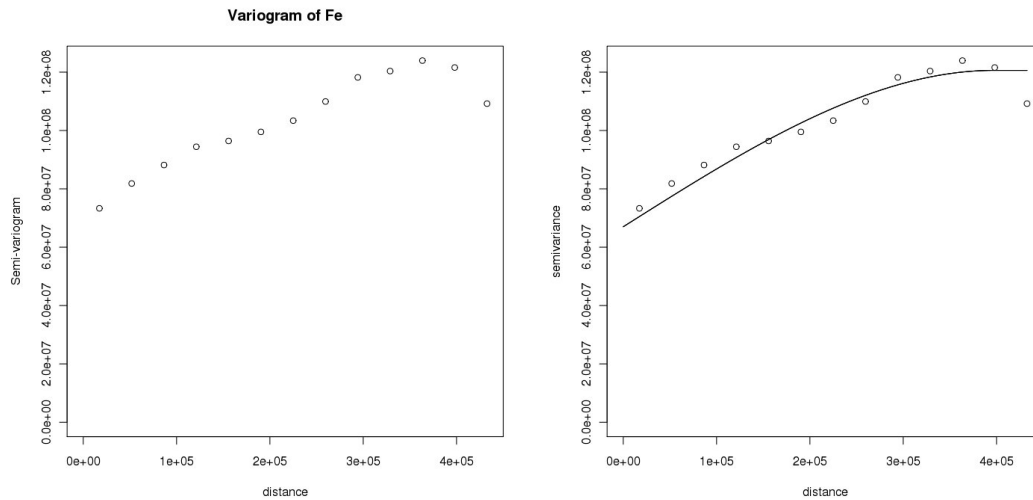




Figure 6: Empirical Variogram and Estimated Model

- Princ. Components Anal.
- Factor Analysis
- Biplot
- Cluster Analysis
- Discriminance Analysis
- Analysis of Variance
- Mult. Regression Anal.
- Multivariate Graphics
- Isolines & Raster

We only comment on the entry **Geostatistics**.

2.9.1 Geostatistics

Here the functions of the -package **GeoR** are used. From the active dataset, first a **GeoR**-object is created. Then the structural analysis of random field may be performed by computing empirical variograms with the following estimation of theoretical models. The estimation can be made by a numerical procedure or interactively controlled by the -function **eyefit** which allows continuously moving the parameters and visualizing the result simultaneously (see Figure 6 for a result).

Using the estimated variograms a kriging procedure may be invoked for inter- and extrapolation. Finally all **GeoR**-objects may be visualized. Figure 7 shows e.g. the kriged values of **Fe** in form of a raster plot with isolines.

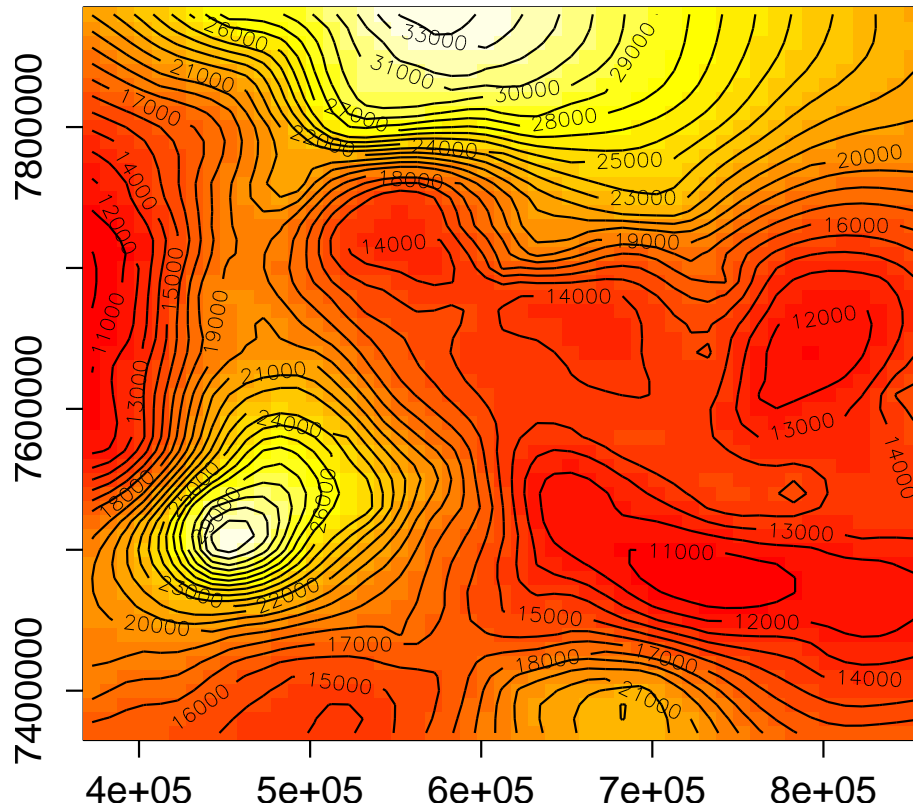


Figure 7: Kriged Values of Fe

2.10 Tools

Here the user gets some auxiliary tools as for loading available packages, showing useful objects like colorsets or the available symbolsets, etc.

2.11 Windows and Help

These are the usual entries for arranging the windows, setting up the toolbar, etc. **Help** is the usual button for calling some help about the function assembly.

References

- [1] Fox J. (2004). Getting Started with the R Commander: A Basic-statistics Graphical User Interface to R. 'useR 2004' Conference, May 20-22, 2004, Vienna University of Technology, Austria.
- [2] Reimann C., Filzmoser P., Garrett R.G., Dutter R. (2008). *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. Wiley, New York.
- [3] Welch B.B., Jones K. (2003). *Practical Programming in Tcl and Tk*. Prentice Hall PTR, New York.