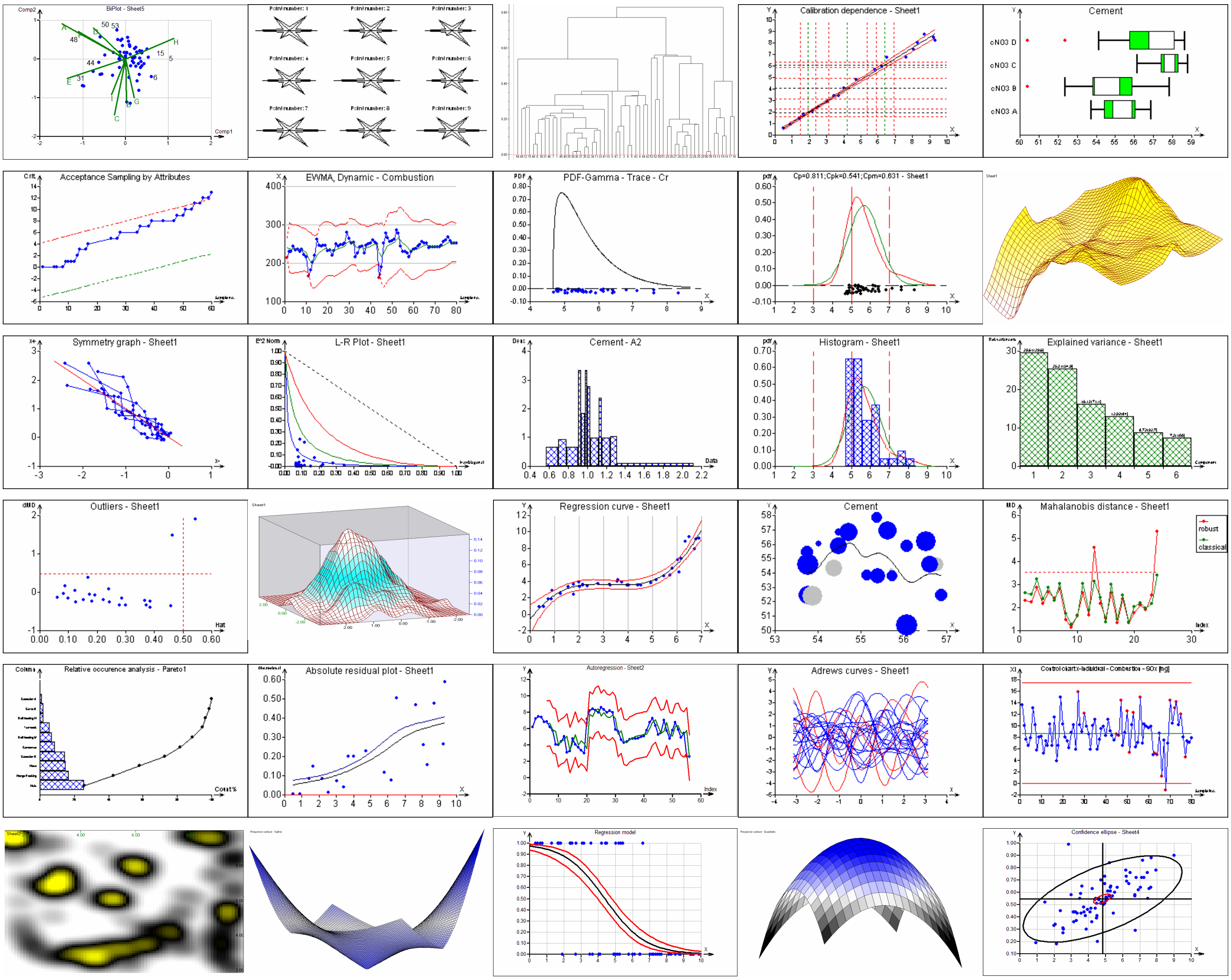
Advanced Interactive Data Analysis in Technology and Research

QC.Expert®

PROFESSIONAL

VERSION 3.1

User's Manual



YOUR DATA TURNED INTO INFORMATION

TriloByte® Ltd.
STATISTICAL SOFTWARE

TriloByte STATISTICAL SOFTWARE

Advanced Interactive Data Analysis in Technology and Research

QC.Expert®

ADSTAT

VERSION 3.1

User's Manual

TriloByte STATISTICAL SOFTWARE
St. Hradiste 300 21, 533 52 Pardubice, Czech Republic

phone +420 466 615725, fax +420 466 615735, mail info@trilobyte.cz, http://www.trilobyte.cz

1. Contents

2. Introduction

Data visualization and nontrivial graphical diagnostic checks are very efficient and useful tools of interactive data analysis, therefore we stress their use within the program. We implemented methods which are known to provide substantial amount of objective information. These methods are recommended by international standards ISO 9000, 14000, QS9000, VDA and other standards and quality control manuals (TQM, for example). The methods were carefully selected to help the user to analyze data effectively. With their help, the user should be able to discover both useful information contained in data and possible problems and challenges posed by data.

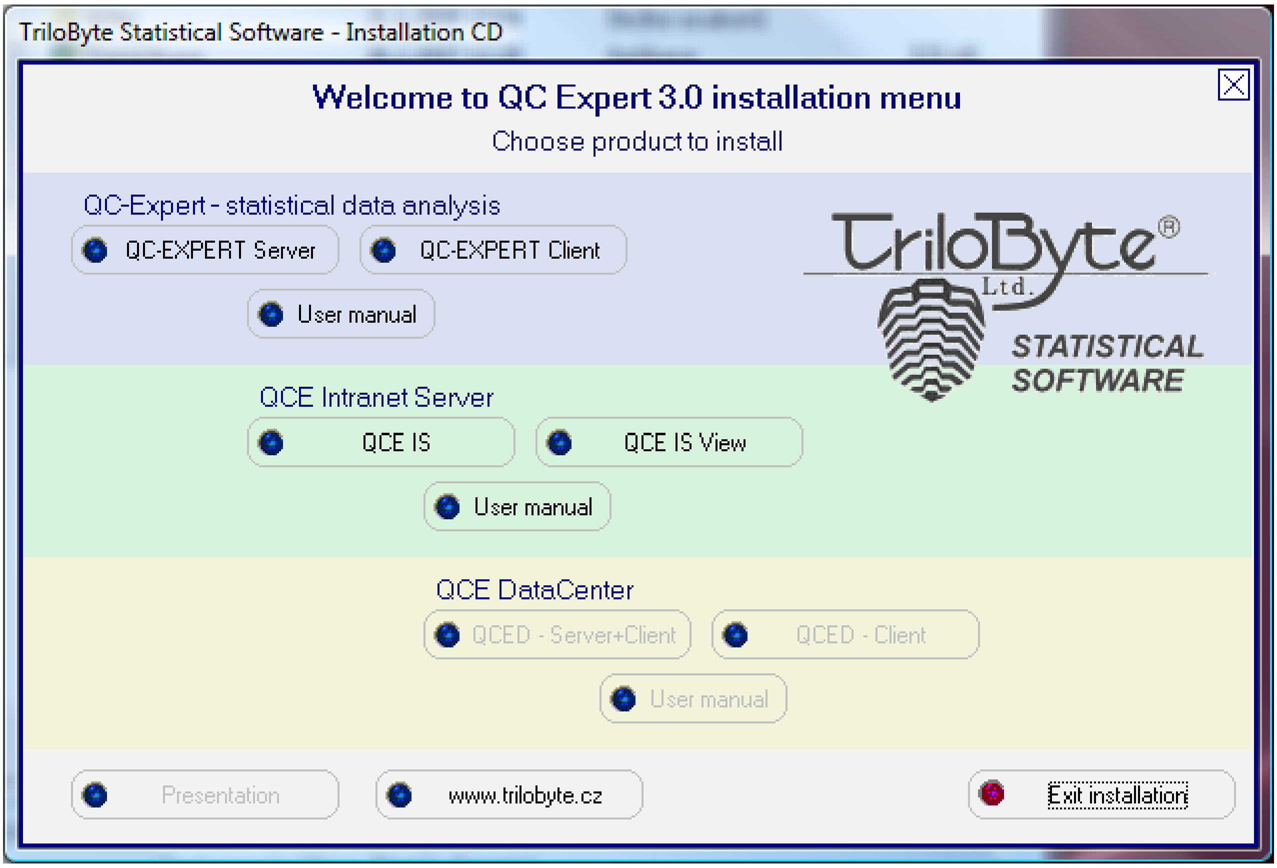
This manual serves as a basic description of the software. It shows how to install and execute the program, edit data and parameters, etc. The manual does not describe mathematical and statistical methods in any detail. It is not intended as a textbook of data analysis methodology. An interested reader is referred to the book M. Meloun, J. Militký: Chemometrics for Analytical Chemistry, vol. 1&2, Ellis Horwood 1992, also available from TriloByte Statistical Software. The book contains description of methods implemented in the software as well as the detailed instructions for their application and interpretation, accompanied by practical examples.

3. Installing the program

Congratulations for your purchase of QC.Expert, a statistical quality control system! The QC.Expert can be installed on a PC with the following minimal configuration: 486 or higher, 512 MB RAM, 50 MB hard disk space, running under operating system MS Windows 2000, XP, Vista. The QC.Expert is available on a CD-ROM. When installing the software, please follow the instructions below.

Installation steps:

1. When installing from 3.5” disks, make sure that the disks are write protected (both notch windows should be open). Then the data on the disk cannot be accidentally overwritten, damaged or erased.
2. When installing from a CD-ROM, insert the CD-ROM into your CD-ROM drive. In case that an installation window (see the picture below) does not appear automatically on the screen, execute the program SETUP.EXE manually. When installing from 3.5” disks, run the program SETUP.EXE and follow the instructions on the screen. Unlike the CD-ROM, the 3.5” disks contain neither documentation files nor Acrobat Reader. The following instructions are relevant for CD-ROM installation only.



3. Choose Install QCExpert Server and follow the instructions on the screen.
4. If you have not installed Acrobat Reader yet, choose Install Acrobat Reader. This program will enable you to read documentation saved in the PDF extension files on your installation CD-ROM in the \DOC directory. You can read the documentation upon selecting “CD-ROM documentation” in the main installation window. The documentation files can be printed.
5. Installation program creates a program group from which you can run the QC.Expert.
6. When running QC.Expert for the first time after installation, input the serial number printed on your disk. Open the demonstration examples data file Ex_q3e.vts. Data from this file are used in examples throughout the electronic documentation.

The electronic PDF documentation file (saved in \DOC directory on the installation CD-ROM) (TriloByte reserves the right to modify or update the documentation files):

- **QC.Expert Manual (manual.pdf)** Electronic version of this manual.

TriloByte would like to thank you for using the QC.Expert.

4. Basic description of the QC.Expert system

The QC.Expert program is intended for advanced users in quality control departments as well as for users of various other backgrounds, who need to analyze data and processes. The program follows general recommendations for correct application of statistical methods. Application of modern statistical techniques (robust estimation, power transformations, dynamical graphics) improves result interpretation and increases quality control efficiency. It is assumed that the QC.Expert user is familiar with basic principles of statistics and control charts construction to the extent covered e.g. by the recommended textbook or by the basic course offered by TriloByte.

4.1. Elementary commands

The QC.Expert runs under MS Windows 2000, XP, Vista.. The program is run interactively, using menus, toolbars and dialog panels.

4.1.1. Menu

Menu bar is a basic tool for running the QC.Expert. It appears at the top of the application window. Available menu items change, depending on the currently active window. Next, we will describe all menu items for the Data, Protocol, Graphs and Interactive graph windows.

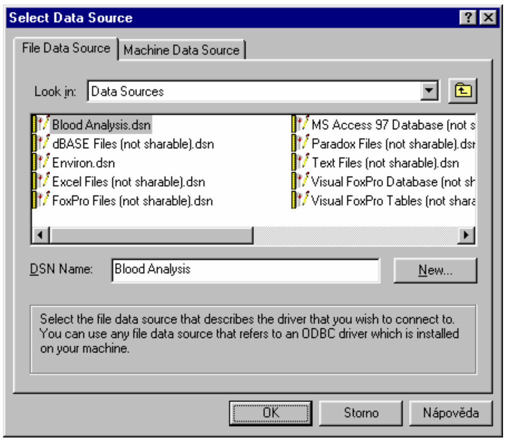
Menu for the **Data** window:

File	Edit	Format	QC.Expert	Window	Help

4.1.1.1. File

- *Open* – opens an XLS (Excel), TXT (text) or VTS (QC.Expert) file.
- *Save* – saves a spreadsheet in the XLS or VTS format. When the TXT format is selected only the active sheet is saved.
- *New* – Deletes any data currently present in the Data window.
- *Serial Port* - Activates the module PCom for on-line data communication through the serial interface (COM1-COM4) using RS232 protocol. This function is designed for real time data acquisition from electronic devices with RS232 numeric data output. The data can be stored in active sheet and used to draw on-line control charts. The communication module PCom is described in a separate chapter.
- *ODBC Data Source* - This is an optional way to read data into the QC.Expert sheet using an existing ODBC (Open DataBase Connectivity) driver with an SQL query. This option is very convenient when data analysis is to be carried out on a part of a larger existing database. The basic ODBC drivers are part of the Windows system and must be installed on your computer, for details on installing ODBC please refer to your MS Windows manual or help. Submenu items: *Connect Data Source*: Defines and opens a new data source or opens an existing source (see figure below), *Disconnect Data Source*: Closes the opened ODBC data source leaving last content of the sheet in QC.Expert unchanged, *Define New Query*: (This item is not active unless an ODBC connection is open) Edit the current SQL query of the active ODBC connection, which allows the data in sheet to

be altered any time; before editting the SQL query you are prompted to confirm whether or not to delete all current data before performing new query (this is recomended) , *Update (Alt-Ctrl-A)*: (This item is not active unless an ODBC connection is open) Perform the last SQL query. This is useful when we expect new data rows have been added in the database.



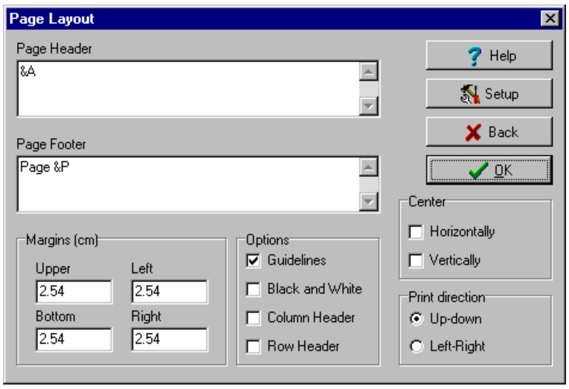
Look in: Data Sources

Blood Analysis.dsn
dBASE Files (not sharable).dsn
Environ.dsn
Excel Files (not sharable).dsn
FoxPro Files (not sharable).dsn
MS Access 97 Database (not s
Paradox Files (not sharable).dsr
Text Files (not sharable).dsn
Visual FoxPro Database (not sh
Visual FoxPro Tables (not shar

DSN Name: Blood Analysis

Select the file data source that describes the driver that you wish to connect to. You can use any file data source that refers to an ODBC driver which is installed on your machine.

- *Settings* – sets the header, protocol and graph formats (see Chap. 4.3.).
- *Page setup* – sets a page before printing a spreadsheet, see the following figure:



Page Header: &A

Page Footer: Page &P

Margins (cm): Upper 2.54, Left 2.54, Bottom 2.54, Right 2.54

Options: ☑ Guidelines ☐ Black and White ☐ Column Header ☐ Row Header

Center: ☐ Horizontally ☐ Vertically

Print direction: ● Up-down ○ Left-Right

Page Header: enter a text you want to appear as a header on all printed output pages. &A is a shortcut for a sheet name, &P is a shortcut for page number, &D prints date, &T prints time.
Page Footer: enter a text you want to appear at the bottom of all printed output pages. &A is a shortcut for a sheet name, &P is a shortcut for page number, &D prints date, &T prints time.
Margins: page margins width in centimeters.
Options:

Guidelines: prints a table with horizontal and vertical lines
Black&White: any color is printed as black. When this selection is turned off, different colors are mapped to different shades of gray on a B&W printer, which might cause that some printout parts are more difficult to see.
Column Header: prints a table with column headings.
Row Header: prints a table with row names.

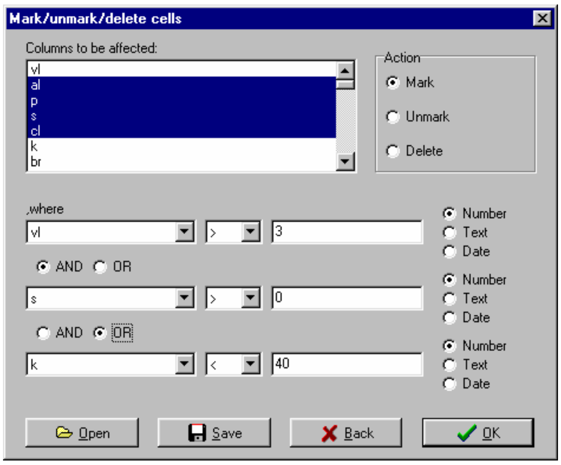
Center: table can be centered both horizontally and vertically. When both fields are checked, the table is printed in the middle of a page.
Print direction: determines how to split a table before printing, when it does not fit a single.

- *Print* – prints the active sheet.
- *Exit* – exits the QC.Expert system.

4.1.1.2. Edit

- *Cut* – cuts a selected text into the clipboard

- *Copy* – copies a selected text into the clipboard
- *Paste* – pastes the clipboard to the insertion point (the clipboard may contain text or a picture in WMF, i.e. Windows Metafile format)
- *Paste values* – pastes data as values only, ignoring formulas
- *Mark* – marks selected cells in red. The cells can then be selectively included or excluded from a computation, or they can be used to identify points on an interactive graph.
- *Mark query* - Marks cells in *Columns to be affected* which fulfil the *where* conditions. Instead of marking, cells can also be unmarked or deleted according to the selection in *Action* group. Maximum of 3 conditions can be used. The conditions are evaluated one by one without any priority of the *OR* and *AND* operators. The comparisons are made in *Number*, *Text* or *Date* formats. (Note: in text format it holds "1000" < "20").



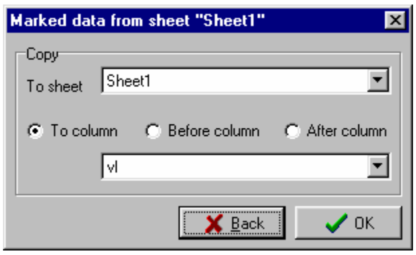
Mark/unmark/delete cells

Columns to be affected:
vl
al
p
s
cl
k
br

Action
● Mark
○ Unmark
○ Delete

where
vl > 3 ● Number ○ Text ○ Date
● AND ○ OR
s > 0 ● Number ○ Text ○ Date
○ AND ● OR
k < 40 ● Number ○ Text ○ Date

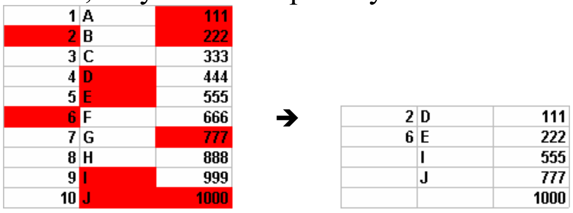
- *Copy marked cells* - Marked cells are copied column-wise to a new column in the same or a different sheet.



Marked data from sheet "Sheet1"

Copy
To sheet: Sheet1
● To column ○ Before column ○ After column
vl

If the marked cell are discontinuous, they will be copied by columns as shown below.



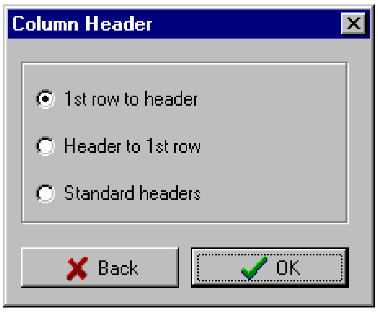
1	A	111
2	B	222
3	C	333
4	D	444
5	E	555
6	F	666
7	G	777
8	H	888
9	I	999
10	J	1000

➔

2	D	111
6	E	222
	I	555
	J	777
		1000

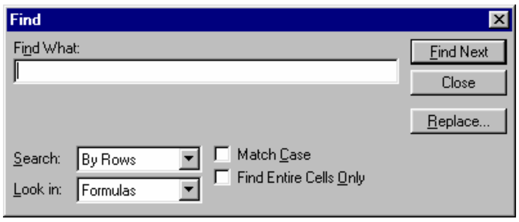
- *Unmark* – cancels marking made previously in current the selection.
- *Unmark all* – cancels all marking in a sheet. When a sheet contains a lot of data, unmarking may take some time and the program offers an accelerated deselection. The accelerated deselection cancels some of text formats (font size and type, text color etc.).
- *Column names* – specifies how to create column headings
- *1st row to header*: uses first row of the table as column names. These names can be used later when performing calculations.
- *Header to 1st row*: puts column names to the first row of a table (original table rows are shifted down). This is useful when exporting the data to other programs.

- *Standard headers*: sets standard column names (alphabet characters).



● 1st row to header
○ Header to 1st row
○ Standard headers

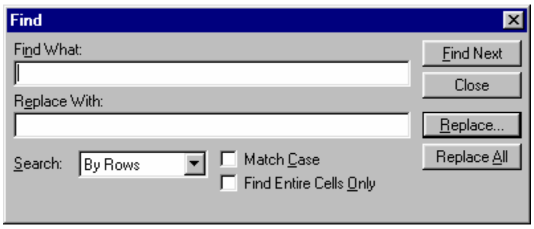
- *Go to* – jumps to a specified cell. Cell address is entered in the format [column character][row number] with no space between the two parts, e.g. D100.
- *Find* – looks for a character string in a selected part of the table. When no cells are selected, all cells are searched. The *Replace* button switches to the replacement mode, see the next item.



Find What:

Search: By Rows
Look in: Formulas
□ Match Case
□ Find Entire Cells Only

- *Replace*- finds a specified string and replaces it by another string.

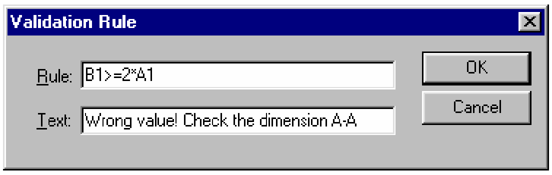


Find What:

Replace With:

Search: By Rows
□ Match Case
□ Find Entire Cells Only

- *Validation* – a rule for keyboard data entry. It is entered for a selected block, using relative column character and relative row number (i.e. with respect to the selected block). For instance, when the entries in the second column are supposed to be at least twice as large as the corresponding entries in the first column, we select the block and enter the rule: B1>=2*A1. Next, we enter text to appear when the rule is violated to the *Text* field. Logical operators *and*, *or* are entered as binary functions. For example, when we want to check whether a value is within 0 to 100 range, we define (for a selected block) the rule and((a1>0);(a1<100)). Warning: the rule is checked only when entering data manually via keyboard. The data already present in the table are not checked!



Rule: B1>=2*A1
Text: Wrong value! Check the dimension A-A

4.1.1.3. Format

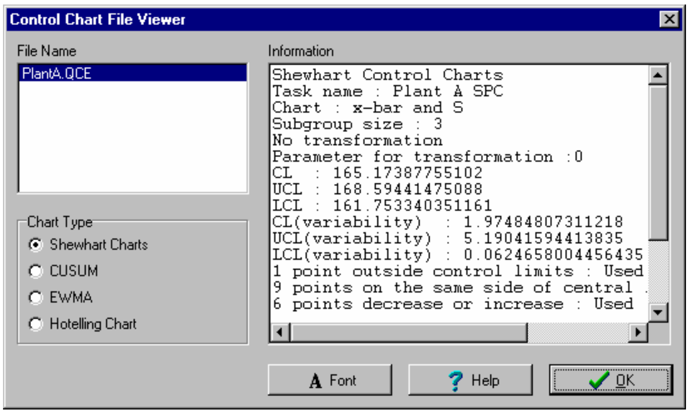
- *Font* – type, size and color specification applied to selected cells.
- *Number* – number/date/time format specification applied to selected cells.
- *Sheet* – *Append* after the last sheet, *Insert* before the active sheet, *Erase* the active sheet, set the active sheet as *Read only*.

- *Row – Height* sets row height, *Automatic height* – row height is adjusted automatically, with respect to the text size, *Insert* – inserts*, Delete* – deletes a row.
- *Column - Width* – sets column width, *Automatic width* – column width is chosen automatically, with respect to the text length, *Insert* – inserts, *Delete* –deletes a column.
- *Freeze/Unfreeze* – sets a separator beyond which no cells are affected by scrolling. This is useful e.g. for a header specification. The separator is removed by a repeated selection of the item. Separator is always set according to the cursor's position in the table.
- *Time axis*
  Automatic -specifies a column to be used as a time variable. A column of specified length with given start and endpoints is generated.
  Manual – an alternative way how to enter a time variable. The time values are entered individually into the table.

4.1.1.4. QC.Expert

All computations relate to the active sheet only. Detailed description of all modules

- *Basic data analysis* – runs the Basic data analysis module
- *Acceptance sampling* – runs the Acceptance sampling module
- *ANOVA* – runs the ANOVA module
- *Correlation* – runs the Correlation analysis module
- *Transformation* – runs the Data transformation module
- *Simulation* – runs the Simulation module (simulation, error propagation)
- *Response surface methodology* – runs the RSM module
- *Shewhart control charts* – runs the Shewhart control charts module
- *Extended* – runs the Extended diagrams module (EWMA, CUSUM, Hotelling)
- *Pareto chart* – runs the Pareto analysis module
- *Linear regression* – runs the Linear regression module
- *Nonlinear regression* – runs the Nonlinear regression module
- *Multivariate analysis* – runs the Multivariate analysis module
- *Calibration* – runs the Calibration module
- *Viewer* – Opens a panel where *.QCE files can be selected for reading. The QCE files contain control charts parameters saved previously by pressing the *Save parameters* button while creating a chart, see figure below.

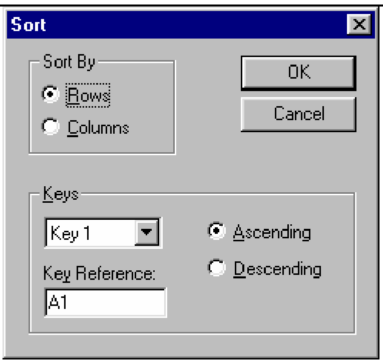


PlantA.QCE

Chart Type
● Shewhart Charts
○ CUSUM
○ EWMA
○ Hotelling Chart

```
Shewhart Control Charts
Task name : Plant A SPC
Chart : x-bar and S
Subgroup size : 3
No transformation
Parameter for transformation :0
CL  : 165.17387755102
UCL : 168.59441475088
LCL : 161.753340351161
CL(variability)  : 1.97484807311218
UCL(variability) : 5.19041594413835
LCL(variability) : 0.0624658004456435
1 point outside control limits : Used
9 points on the same side of central .
6 points decrease or increase : Used
```

- *Sort* – sorts a selected block by rows or columns. *Key 1, Key 2* etc. specify the indexes on which to sort. Only the data within a selected block will be sorted. Warning: empty cells act as having the smallest value, so that it is not recommended to sort data containing missing cells or columns of unequal length.



Sort

Sort By

● Rows

○ Columns

Keys

Key 1

Key Reference:

A1

● Ascending

○ Descending

4.1.1.5. Window

- *Tools* – switches the tools panel on/off.
- *Status line* – switches the status line (appears at the bottom of the screen) on/off.
- *Cascade* – cascade arrangement of all windows.
- *Tiles horizontal* – horizontal arrangement of all windows.
- *Tiles vertical* – vertical arrangement of all windows.
- *GRAPHS* – shows the Graphs window.
- *PROTOCOL* – shows the Protocol window.
- *DATA* – shows the Data window.

4.1.1.6. Help

- *QCExpert* – invokes help from QCExpert.HLP
- *About application* – shows the title window.

Menu for the **Protocol** window:

File	Edit	Format	Window	Help

4.1.1.7. File

- *Save* - saves the Protocol window contents in the XLS or VTS format. When the TXT format is selected, only the active sheet is saved.
- *New* – deletes any data present in the Data window.
- *Set* – sets header and format for protocol and graphical output, see 4.3.
- *Page* – sets page appearance for protocol printing:

Title: enter a text you want to appear as a header on all printed output pages. &A is a shorthand for a sheet name, &P is a shorthand for page number, &D prints date, &T prints time.
Footnote: enter a text you want to appear at the bottom of all printed output pages. &A is a shorthand for a sheet name, &P is a shorthand for page number, &D prints date, &T prints time.
Margins: page margins width in centimeters.
Options:

> *Grid*: prints a table with horizontal and vertical lines
> *Black&White*: any color is printed as black. When this selection is turned off, different colors are mapped to different shades of gray on a B&W printer, which might cause that some printout parts are more difficult to see.
> *Column names*: prints a table with column headings.
> *Row names*: prints a table with row names.

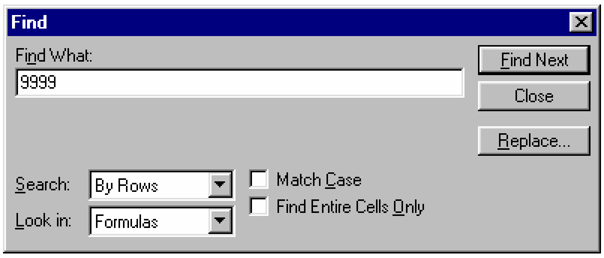
Placement: table can be centered both horizontally and vertically. When both fields are checked, the table is printed in the middle of a page.
Print direction: determines how to split a table before printing when it does not fit a single page.

- *Print* – prints the active sheet.

- *Exit* – exits the QC.Expert system.

4.1.1.8. Edit

- *Cut* – cuts a selected text into the clipboard
- *Copy* – copies a selected text into the clipboard
- *Paste* – pastes the clipboard to the insertion point (the clipboard may contain text or a picture in WMF, i.e. Windows Metafile format)
- *Paste values* – pastes data as values only, ignoring formulas
- *Find* – looks for a character string in a selected part of the table. When no cells are selected, all cells are searched. The *Replace* button switches to the replacement mode, see the next item.



Find

Find What: 9999

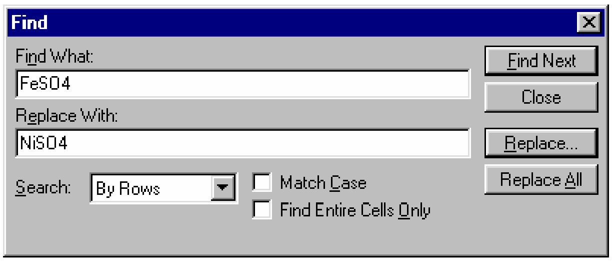
Search: By Rows

Look in: Formulas

□ Match Case

□ Find Entire Cells Only

- *Replace*- finds a specified string and replaces it by another string.



Find

Find What: FeSO4

Replace With: NiSO4

Search: By Rows

□ Match Case

□ Find Entire Cells Only

4.1.1.9. Format

- *Font* – type, size and color specification applied to selected cells.
- *Number* – number/date/time format specification applied to selected cells.
- *Sheet* – *Append* after the last sheet, *Insert* before the active sheet, *Erase* the active sheet, set the active sheet as *Read only.*
- *Row* – *Height* sets row height, *Automatic height* – row height is adjusted automatically, with respect to the text size, *Insert* – inserts, *Delete* – deletes a row.
- *Column* - *Width* – sets column width, *Automatic width* – column width is chosen automatically, with respect to the text length, *Insert* – inserts, *Delete* –deletes a column.
- *Separator* – sets a separator beyond which no cells are affected by scrolling. This is useful e.g. for a header specification. The separator is removed by a repeated selection of the item. Separator is always set according to the cursor's position in the table.

4.1.1.10. Window

- *Tools* – switches the tools panel on/off.
- *Status line* – switches the status line (located at the bottom of the screen) on/off.
- *Cascade* – cascade arrangement of all windows.
- *Tiles horizontal* – horizontal arrangement of all windows.
- *Tiles vertical* – vertical arrangement of all windows.
- *GRAPHS* – shows the Graphs window.
- *PROTOCOL* – shows the Protocol window.

- *DATA* – shows the Data window.

4.1.1.11. Help

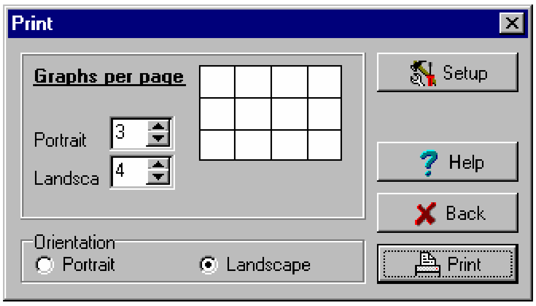
- *QCExpert* – invokes help from QCExpert.HLP
- *About application* – shows the title window.

Menu for the **Graphs** window:

File	Graph	Window	Help

4.1.1.12. File

- *Save graph* – saves graphs in a selected format. You can choose either WMF or BMP formats in the dialog window for saving files. Additional choices are:
  Current graph: saves the current graph (red framed).
  All graphs (current sheet): saves all graphs from the current graph sheet.
  All sheets: saves all graph sheets currently in the Graph window. Each graph sheet is saved in a separate file, whose name is derived from the sheet name. All sheets must have distinct names.
- *Print* – prints all graphs present in the current sheet. Number of graphs per page and their layout are selected in the following dialog window.



Graphs per page

Portrait 3

Landsca 4

Orientation

○ Portrait ● Landscape

- *Set* – sets header, format for protocol and graphical output, see 4.3.
- *Exit* – exits the QC.Expert system.

4.1.1.13. Graph

- *Copy* – copies the currently active graph (red framed), or all graphs in a current graph sheet into the clipboard. Graphical resolution is the same as the resolution of the original on the screen, therefore *Normal size* is recommended. The clipboard copy of a graph from the Graphs window is always a bitmap which cannot be modified. An editable copy can be obtained either by saving the graph in the WMF format or by copying corresponding interactive graph to the clipboard.
- *Normal size* – sets graph size to the default.
- *Fit to window* – sets graph size so that all of them fit into the window.
- *Delete sheet* – deletes the current graph sheet, permanently removing all graphs it contains.
- *Delete all* – deletes all sheets currently present in the Graph window.

4.1.1.14. Window

- *Tools* – switches the tool panel on/off.
- *Status line* – switches the status line (located at the bottom of the screen) on/off.
- *Cascade* – cascade arrangement of all windows.
- *Tiles horizontal* – horizontal arrangement of all windows.
- *Tiles vertical* – vertical arrangement of all windows.
- *GRAPHS* – shows the Graphs window.
- *PROTOCOL*– shows the Protocol window.
- DATA – shows the Data window.

4.1.1.15. Help

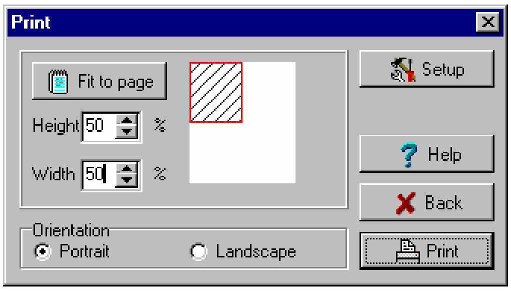
- *QCExpert* – invokes help from QCExpert.HLP
- *About application* – shows the title window.

Menu for the **Interactive graphics** window:

File	Graph	Window	Help

4.1.1.16. File

- *Save* – saves graph in WMF or BMP format.
- *Print* – prints graph. Graph size and page layout are set in the following dialog window:
  Fit to page: graph uses whole page, preserving scale approximately.
  Height: percentage of page height, used by the graph
  Width: percentage of page width, used by the graph



Print

Height 50 %

Width 50 %

Orientation

● Portrait ○ Landscape

- *Exit* – exits the QC.Expert system.

4.1.1.17. Graph

- *Copy as* – copies a graph to the clipboard in BMP or WMF format.
- *Zoom* – when this selection is checked, a detail can be zoomed by dragging a mouse. When it is not checked, points on the graph can be selected.
- *Normal size* – restores automatic scaling after zooming a detail.
- *Horizontal lines* – adds horizontal lines.
- *Vertical lines* – adds vertical lines.
- *Grid* – adds a grid of horizontal and vertical lines.
- *Select points (Graph->Table)* – the points, selected in the graph are marked in the corresponding table.
- *Select points (Table->Graphs)* – the points, selected in the table are marked in the corresponding graph.

4.1.1.18. Window

- *Tools* – switches the tools panel on/off.
- *Status line* – switches the status line (located at the bottom of the screen) on/off.
- *Cascade* – cascade arrangement of all windows.
- *Tiles horizontal* – horizontal arrangement of all windows.
- *Tiles vertical* – vertical arrangement of all windows.
- *GRAPHS* – shows the Graphs window.
- *PROTOCOL*– shows the Protocol window.
- *DATA* – shows the Data window.

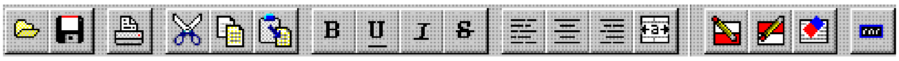
4.1.1.19. Help

QC.Expert system help.

4.1.2. Toolbars

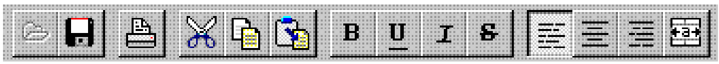
The toolbars appear at the top of the QC.Expert program window. They serve as a shorthand for some of the menu commands. The toolbar appearance changes, depending on the currently active window. This is analogous to the menu changes described earlier. To make the orientation easier, we will describe all toolbars in the following paragraphs. Various menu items are described in detain in 4.1.1.

Toolbar for the *Data* window



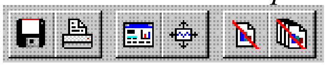
- Open file (inactive)
- Save file
- Print
- Cut
- Copy
- Paste
- Bold
- Underline
- Italic
- Special font
- Left alignment
- Center
- Right alignment
- Center in selected cells
- Mark data in selected cells
- Unmark data in selected cells
- Unmark all
- Title

Toolbar for the *Protocol* window



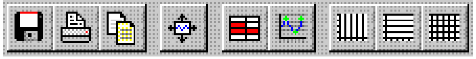
- Open a file (inactive)
- Save a file
- Print
- Cut
- Copy
- Paste
- Bold
- Underline
- Italic
- Special font
- Left alignment
- Center
- Right alignment
- Center in selected cells

Toolbar for the *Graphs* window



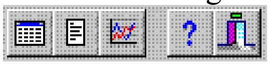
- Save graphs
- Print graphs
- Normal size
- Fit to window
- Delete active sheet
- Delete all sheets

Toolbar for the *Interactive graphics* window



- Save graphs
- Print graphs
- Copy into clipboard
- Restore normal size
- Select points (Graph->Table)
- Select points (Table->Graph)
- Draw vertical lines
- Draw horizontal lines
- Draw grid/draw a finer grid

The toolbar located in the upper right corner of the program window is common for all windows, it does not change when active windows are switched.



- Switch to the Data window
- Switch to the Protocol window
- Switch to the Graphs window
- Help
- Exit

A more detailed explanation of listed items can be found in the paragraphs, describing the corresponding menu commands.

4.1.3. Dialog panels

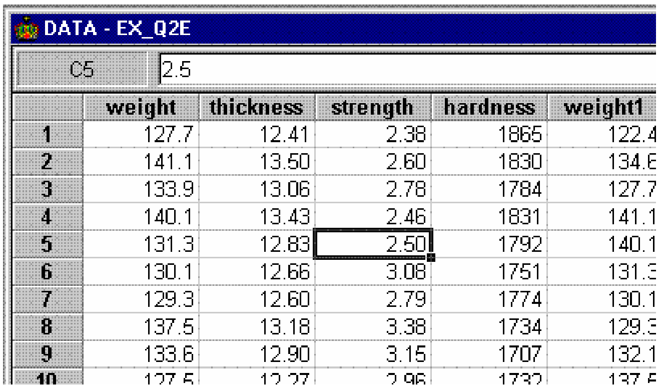
The dialog panels are used to customize the QC.Expert session, to read and save data, to input the values, needed for computations or for other purposes. Dialog panel input is usually finished by pressing the “OK” button. The requested action follows – e.g. parameter values are saved or computations are started.

There are at least three windows on the screen at any time. These windows: DATA, PROTOCOL, GRAPHS cannot be closed. The QC.Expert consists of the text output (protocol) and graphical output (graphs). Graphs and protocols are outputted to the appropriate windows on the screen. These windows might contain several sheets to capture different outputs. Sheets that are no longer needed can be deleted at any time. Selected output can be printed using the menu command *File-Print*.

4.2. Communicating with the program – Windows

4.2.1. Input Data

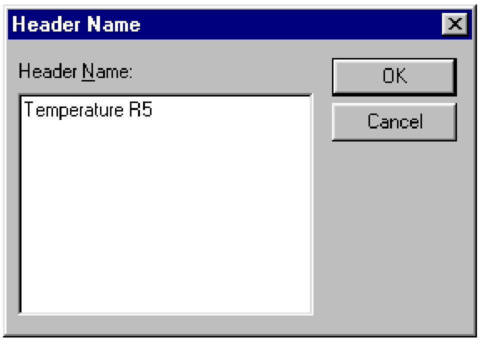
Data obtained by measurement or by other means are always arranged in a table in the *Data* window. Any computations are performed with the active sheet data. Any data for an analysis are expected in table form, with distinct variables entered as columns, their individual observations in different rows. Most of the program procedures tolerate missing values in the form of empty cells. Data must not contain any text entries. There should be no empty columns in the data sheet. The column names can appear as the column headings (see Fig. 1). The names are used when a calculation is requested. They appear in the output as well. The column names can be entered after clicking on the column headings (Fig. 2). Maximum allowable number of data cells depends on the character of requested computations and the available computer resources (memory, disk space). As a general rule, the data should not be extremely large in order to keep the output readable. The maximum physical row number is 65526 (i.e. 2^{16}), the maximum physical column number is 256 (i.e. 2^{8}).



C5 2.5

	weight	thickness	strength	hardness	weight1
1	127.7	12.41	2.38	1865	122.4
2	141.1	13.50	2.60	1830	134.6
3	133.9	13.06	2.78	1784	127.7
4	140.1	13.43	2.46	1831	141.1
5	131.3	12.83	2.50	1792	140.1
6	130.1	12.66	3.08	1751	131.3
7	129.3	12.60	2.79	1774	130.1
8	137.5	13.18	3.38	1734	129.3
9	133.6	12.90	3.15	1707	132.1
10	127.5	12.27	2.96	1732	137.5

Fig. 1 A Sheet with column headings.

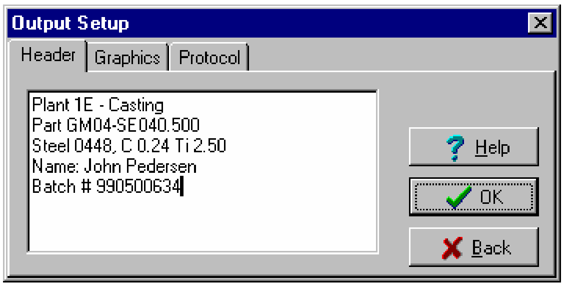


Header Name: Temperature R5

Fig. 2 Column header dialog

4.3. Output

The output, available after requested computations are finished, consists of text and graphical parts which appear in the *Protocol* and *Graphs* windows. The *File-Settings* menu can be used to specify the output sheet organization, see paragraphs 4.3.1. and 4.3.2. For a detailed identification, several lines of text can be entered using the *File-Settings-Header* menu, see Fig. 3. The text appears on each protocol.



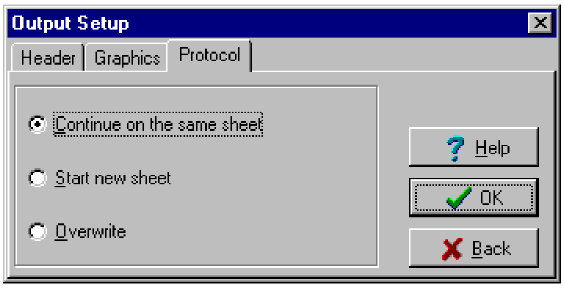
Plant 1E - Casting
Part GM04-SE040.500
Steel 0448, C 0.24 Ti 2.50
Name: John Pedersen
Batch # 990500634

Fig. 3 Setting a header

4.3.1. Protocol

The *Protocol* window contains results arranged in a table form. The results of different statistical characteristics are organized into paragraphs, stressing good readability of the text. Numerical results of statistical tests are commented in words. Warnings about possible violation of the basic assumptions about data behavior are given in red bold print. Other important conclusions are stressed in the same way. It is mainly the lack of normality, strong dependencies in the data (autocorrelation and correlation), differences among tested means (ANOVA), violation of statistical stability in control charts, etc. More details can be found in the paragraphs, describing various statistical methods. The output is organized in a table, which can be saved, printed or copied to other application (Lotus, Excel, Word, etc.) using the clipboard. Selected parts of the output (e.g. smoothed data, simulated values) can be copied into a table in the *Data* window and used for further computations. One of the three protocol arrangements can be selected using the *File-Settings-Protocol* menu, see Fig. 4:

Continue – protocols are written to the same sheet, a new protocol is appended after the previous text,
New sheet – each protocol is written to a separate sheet (the maximum sheet number is 32),
Overwrite - new protocol overwrites all text previously present in the current sheet.



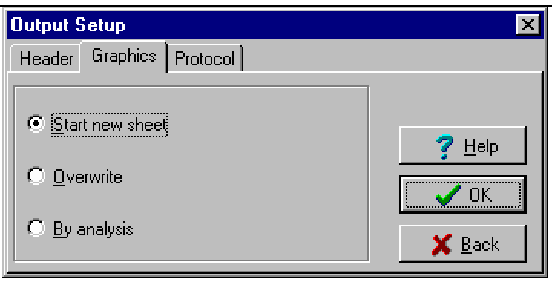
● Continue on the same sheet
○ Start new sheet
○ Overwrite

Fig. 4 Protocol settings

4.3.2. Graphs

After an analysis is finished, the *Graphs* window contains resulting graphs and charts. They are used mainly for data diagnostics, to check validity of analyses performed and as an appendix to the text output in the protocol. The number and type of the resulting graphs depends on the analysis performed and on selections made by the user.

Graphs can be written to the *Graphs* window in one of the three ways, depending on the selection made in the *File-Settings-Graph* menu, see Fig. 5:
New sheet – each graph is written to a separate sheet (the maximum sheet number is 32),
Overwrite - the new graph overwrites all graphical output previously present in the current sheet.
By analysis type – when a particular type of statistical analysis is requested, a new sheet is created and named after the analysis type. Any graphics produced by the analysis is sent to this sheet. When a sheet of the same name exists, it is overwritten.



Output Setup

Header | Graphics | Protocol

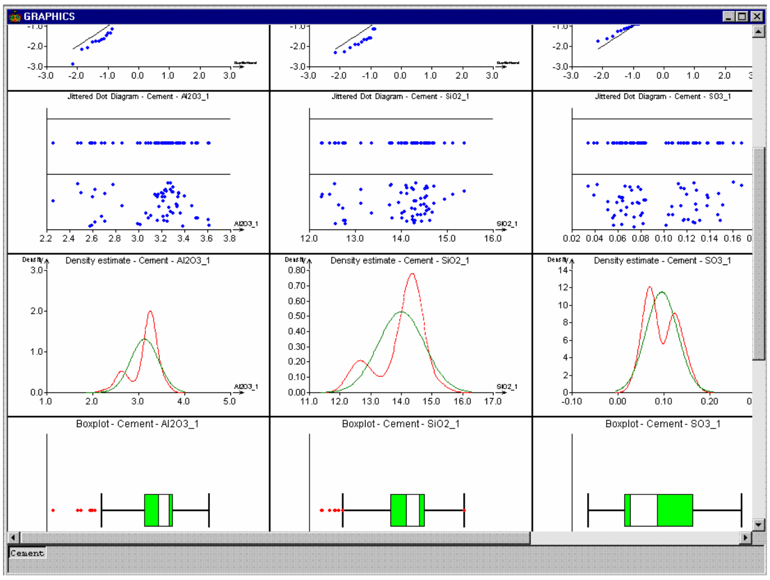
● Start new sheet

○ Overwrite

○ By analysis

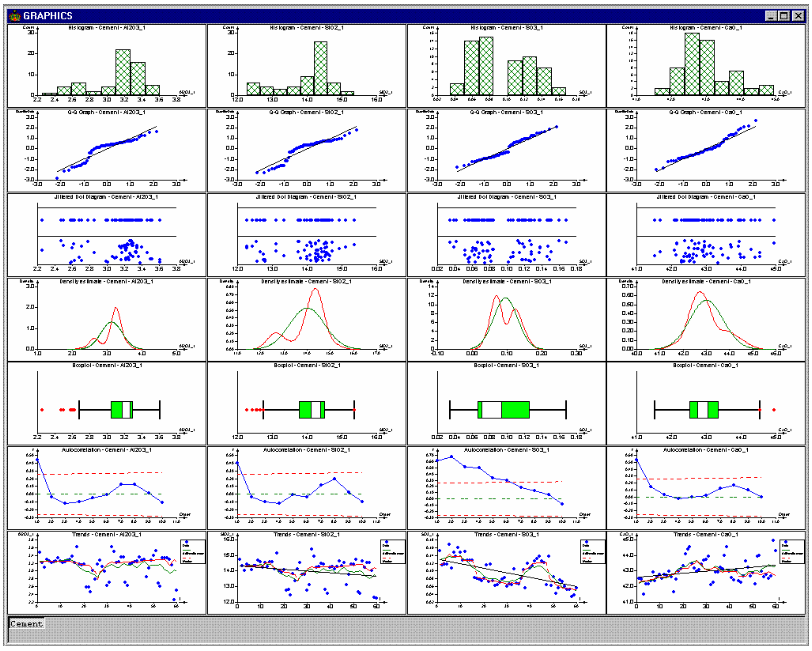
Fig. 5 Graph settings

Graphs can be of normal size (Fig. 6), or they can be viewed in a more detail using the *Graphs* window toolbar. The individual graph size can be changed so that all graphs can be seen in one window. Such an arrangement improves orientation in the graphical output, text on individual graphs might not be legible however. The two viewing modes can be switched using the *Graph-Normal size*, *Graph-Fit to window* menus respectively.



GRAPHICS
Jittered Dot Diagram - Cement - Al2O3_1
Jittered Dot Diagram - Cement - SiO2_1
Jittered Dot Diagram - Cement - SO3_1
Density estimate - Cement - Al2O3_1
Density estimate - Cement - SiO2_1
Density estimate - Cement - SO3_1
Boxplot - Cement - Al2O3_1
Boxplot - Cement - SiO2_1
Boxplot - Cement - SO3_1
Cement

Fig. 6 Normal graph size in the *Graphs* window



GRAPHICS
Histogram - Cement - Al2O3_1
Histogram - Cement - SiO2_1
Histogram - Cement - SO3_1
Histogram - Cement - CaO_1
Q-Q Graph - Cement - Al2O3_1
Q-Q Graph - Cement - SiO2_1
Q-Q Graph - Cement - SO3_1
Q-Q Graph - Cement - CaO_1
Jittered Dot Diagram - Cement - Al2O3_1
Jittered Dot Diagram - Cement - SiO2_1
Jittered Dot Diagram - Cement - SO3_1
Jittered Dot Diagram - Cement - CaO_1
Density estimate - Cement - Al2O3_1
Density estimate - Cement - SiO2_1
Density estimate - Cement - SO3_1
Density estimate - Cement - CaO_1
Boxplot - Cement - Al2O3_1
Boxplot - Cement - SiO2_1
Boxplot - Cement - SO3_1
Boxplot - Cement - CaO_1
Autocorrelation - Cement - Al2O3_1
Autocorrelation - Cement - SiO2_1
Autocorrelation - Cement - SO3_1
Autocorrelation - Cement - CaO_1
Trends - Cement - Al2O3_1
Trends - Cement - SiO2_1
Trends - Cement - SO3_1
Trends - Cement - CaO_1
Cement

Fig. 7 All Graphs fitted to the *Graph* window

To select a graph, click the mouse on it. Borderlines of a selected graph are highlighted. When selected, a graph can be copied to the clipboard in a bitmap format by pressing Ctrl-C. Resolution of the copy remains the same as the resolution of the original graph on the screen, therefore it is recommended to copy normal size graphs. The WMF format can be used for an interactive graph (see the next paragraph). All graphs within an active sheet can be copied by *Graph-Copy-All*, this menu command works regardless of graph selection made previously. The resolution follows the same rules as the resolution of a selected graph copy described previously. Double clicking on a graph in the *Graphs* window opens an interactive graph in a new window.

4.3.3. Interactive graphics

An interactive graph window can be opened by double clicking mouse on a graph in the *Graphs* window. The interactive graph mode allows more detailed analysis of a selected graph. Several interactive graph windows can be opened simultaneously, only one of them is active at a time, however.

The following tasks can be done in an interactive graph window:

Cursor coordinates can be read at any time from the interactive graph title bar.
Points selection can be done by clicking on a selected point. Drag mouse holding left button to select a rectangular region. Selected points are always marked green.
Selection can be cancelled by clicking the right mouse button in the area below x-axis or left from y-axis, see Fig. 9.

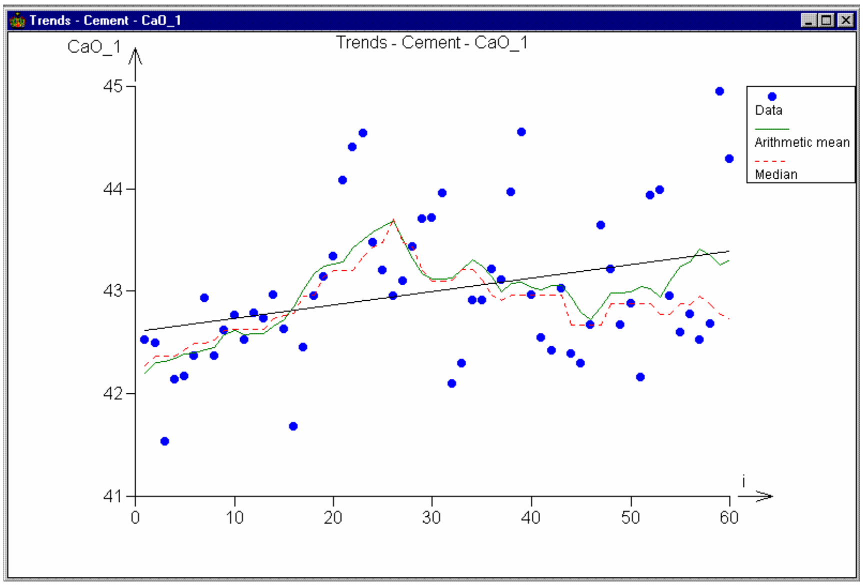


Fig. 8 Interactive graph window

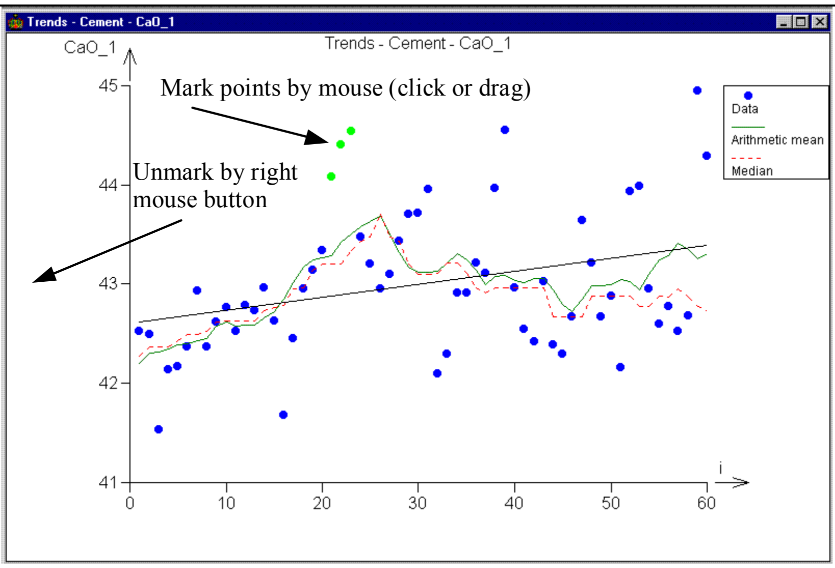


Fig. 9 Interactive graph with selected points

Switching sparse and dense grid on/off is done by three buttons on the toolbar located at the top of QC.Expert program window, see the Fig. 10. Alternatively, the *Graph/Horizontal lines/Vertical lines/Grid* menu can be used. A grid is useful when one wants to locate a graphical object more precisely. The following figures illustrate sparse/dense grid appearance.

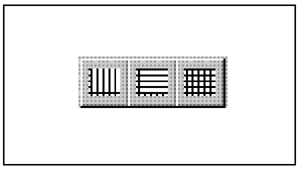


Fig. 10 Lines and grid buttons

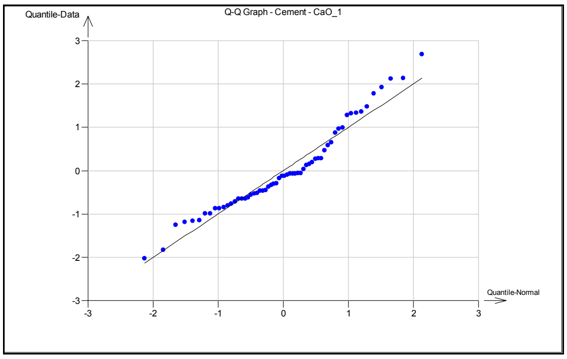


Fig. 11 Sparse grid

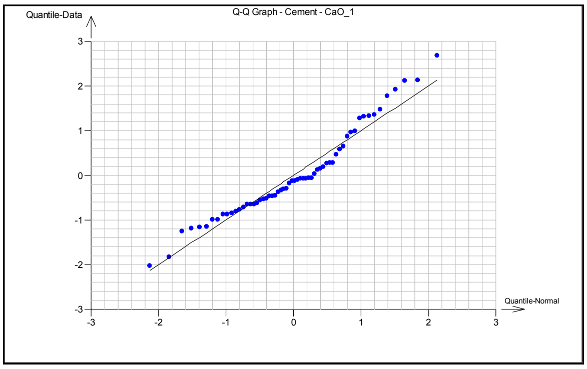


Fig. 12 Dense grid

Magnification (detail) is useful mainly for graphs containing large number of data points, when exploring individual data in detail. Drag mouse, holding left button and the SHIFT key simultaneously to achieve greater magnification. Magnification is done by dragging mouse with left button pressed and holding the SHIFT key. This procedure can be repeated several times see more details.
Detail cancellation is done by clicking the right mouse button within the graph area and holding the SHIFT key. Alternatively, the button on the toolbar can be used.

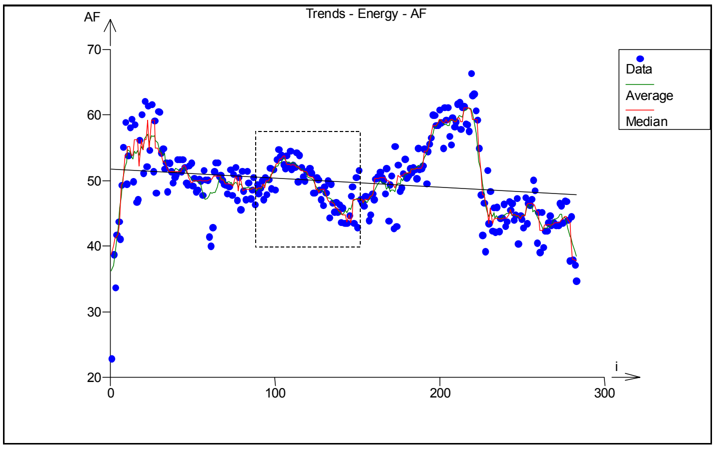


Fig. 13 Selecting an area to zoom (SHIFT+left mouse button)

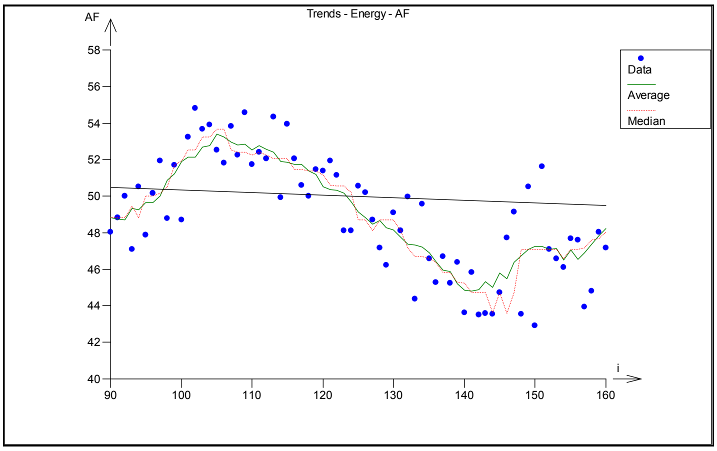
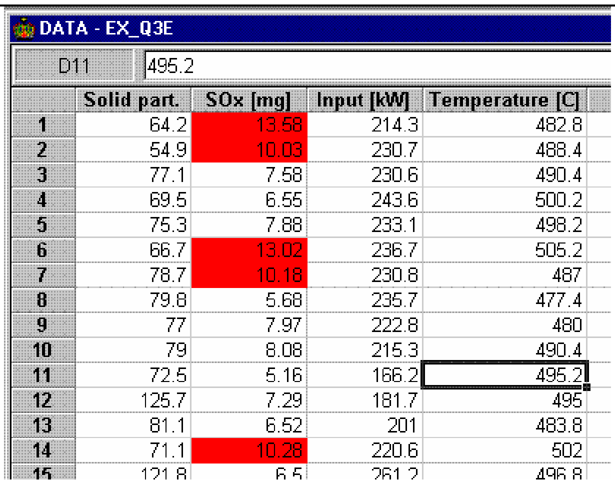


Fig. 14 Zoomed detail of previous picture

Graph-spreadsheet selection When the green marked (selected) points in the graph correspond to data values in the input data table, they can be marked using the button on the toolbar, see Fig. 15. The selected data are marked red in the table. One can specify whether such data should be omitted from subsequent calculations or whether the computations should use the marked data only. This feature is convenient when considering the influence of outliers or otherwise suspect data on analysis results. In case that points selected in the graph cannot be assigned to data, a message “Graph-table assignment is not possible” appears.



DATA - EX_Q3E

D11 495.2

	Solid part.	SOx [mg]	Input [kW]	Temperature [C]
1	64.2	13.58	214.3	482.8
2	54.9	10.03	230.7	488.4
3	77.1	7.58	230.6	490.4
4	69.5	6.55	243.6	500.2
5	75.3	7.88	233.1	498.2
6	66.7	13.02	236.7	505.2
7	78.7	10.18	230.8	487
8	79.8	5.68	235.7	477.4
9	77	7.97	222.8	480
10	79	8.08	215.3	490.4
11	72.5	5.16	166.2	495.2
12	125.7	7.29	181.7	495
13	81.1	6.52	201	483.8
14	71.1	10.28	220.6	502
15	121.8	6.5	261.2	496.8

Fig. 15 Selected data in the Data window

Spreadsheet-graph selection Data points, selected in the data table can be highlighted in the graph, using the button on the toolbar. Data selection can be used as a filter: subsequent calculations can be performed for all data, selected data or data which are not selected.

4.3.4. Data

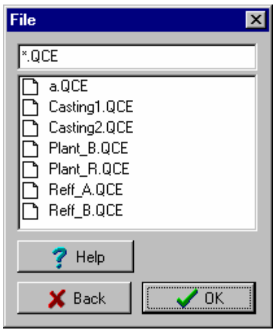
Data are entered and kept in the *Data* window. They are organized in sheets. The maximum sheet number is 256. One sheet can hold up to 16384 rows or 256 columns. The numeric, text and time (date, hour) formats are allowed. Data can be imported from MS Excel, version 5 or 7. Data from higher Excel versions or other applications need to be converted to the text (ASCII) format with columns separated by tabs (tab-delimited format), before they are imported into the QC.Expert.

Since the basic rules for the spreadsheet, mouse and clipboard use in the QC.Expert system are similar to common spreadsheets like MS Excel and the user is assumed to be familiar with such programs, no detailed description of these rules is given in this manual. When unsure about the basic rules, the user is referred to the Excel manual or TriloByte customer service. Some of the basic spreadsheet functions are available in the QC.Expert program:
Formulas are entered after the equal (=) sign, using common functions in the usual English syntax (e.g. sqrt for square root) and cell references, e.g.:

```
=sqrt(a1)                =sum(c1:c100)      =rand()
=round(sin(5*a1);2)      =average(a1:d1)    =now()
```

4.3.5. Parameters

The control chart parameters can be saved in a parameter file *.QCE. The file can contain four types of parameters: for Shewhart chart, CUSUM, EWMA and Hotelling chart. The parameters can be saved by the *Save chart* Save chart button after making changes in the dialog panel related to the appropriate chart type. Previously saved parameter values can be read into the program using the *Load chart* Load chart button from the same dialog window. Warning: when reading parameters from a file, all types of parameters saved in the file are used, overwriting values previously set within the program. For instance, a file aaa.qce can contain Shewhart X bar chart parameters only, but parameters for CUSUM, EWMA and Hotelling charts can be added and saved in the same file as well. When parameters for NP chart are saved to the same file aaa.qce, the original parameter values for X bar are overwritten.



File

*.QCE

a.QCE
Casting1.QCE
Casting2.QCE
Plant_B.QCE
Plant_R.QCE
Reff_A.QCE
Reff_B.QCE

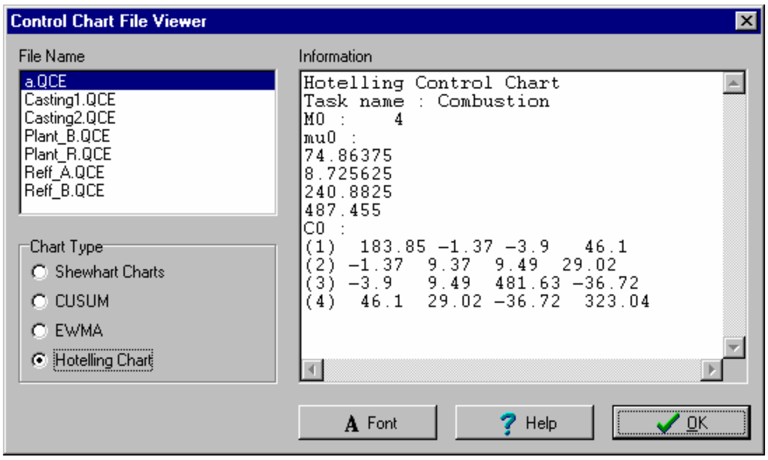
Contents of a parameter file can be checked by viewer, invoked from the menu for the *Data* window: *QCExpert-Viewer*. If parameters for a certain chart type are not set, empty spaces or zeros appear in their places. The mean vector (mu0) and covariance matrix (C0) are shown for the Hotelling chart.

```
Shewhart Control Charts
Task name : Combustion
Chart : x-individual and R
Subgroup size : 1
No transformation
Parameter for transformation :0
CL  : 74.86375
UCL : 104.748613991382
LCL : 44.9788860086183
CL(variability)  : 11.2367088607595
UCL(variability) : 36.7103278481013
LCL(variability) : 0
1 point outside control limits : Used
9 points on the same side of central line : Used
6 points decrease or increase : Used
14 alternating points : Used
2 of 3 points outside 2 sigma : Used
4 of 5 points outside 1 sigma on the same side : Used
15 points inside 1 sigma : Used
8 points outside 1 sigma : Used
```

```
CUSUM Control Chart
Task name : Combustion
K :     0.4
Limit :   4
CL :    74.86375
Sigma : 13.5590433528907
FIR   : Not used
```

```
EWMA Control Chart
Task name : Combustion
W :     0.25
A :     0.05
CL :    8.725625
Sigma : 3.06039372156797
```

```
Hotelling Control Chart
Task name : Combustion
M0 :    4
mu0 :
74.86375
8.725625
240.8825
487.455
C0 :
(1)  183.85 -1.37 -3.9   46.1
(2) -1.37  9.37  9.49  29.02
(3) -3.9   9.49  481.63 -36.72
(4)  46.1  29.02 -36.72  323.04
```


Control Chart File Viewer

File Name

- a.QCE
- Casting1.QCE
- Casting2.QCE
- Plant_B.QCE
- Plant_R.QCE
- Reff_A.QCE
- Reff_B.QCE

Chart Type

○ Shewhart Charts
○ CUSUM
○ EWMA
● Hotelling Chart

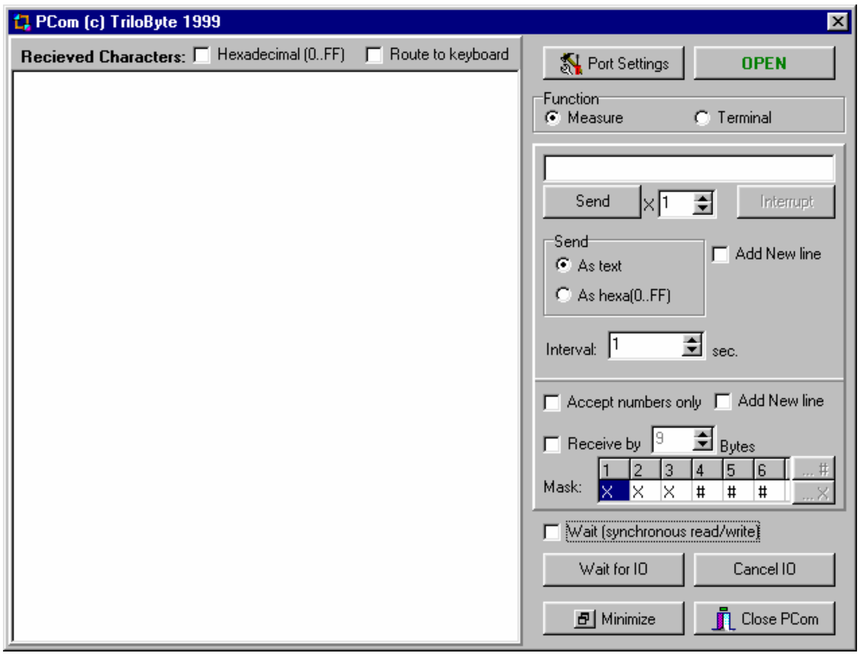
Information

```
Hotelling Control Chart
Task name : Combustion
M0 :     4
mu0 :
74.86375
8.725625
240.8825
487.455
C0 :
(1)  183.85 -1.37 -3.9    46.1
(2) -1.37   9.37   9.49  29.02
(3) -3.9    9.49  481.63 -36.72
(4)  46.1  29.02 -36.72  323.04
```

Fig. 16 File viewer

4.4. Serial Communication

QC.Expert™ can read data directly from an external device via the serial port COM using the RS232 protocol. The data can be read into the active sheet or used for constructing on-line control charts with basic graphical diagnostics. The device must provide data in ASCII code. The possibility to set parameters and a simple filter allows a wide range of devices to be connected to QC.Expert. Serial communication is controled by a stand-alone module **PCom**, which must be installed on your computer in the same directory as QC.Expert itself. (If you have bought the full system, PCom is automatically installed during installation of QC.Expert.) The PCom module is started using menu File – Serial port, see 4.1.1.1, p. 4-14 and must be running while using serial port. After starting the serial communication, the main PCom window appears, see Fig. 17. Press button „*Port Settings*“ to set the parameters of the COM port for communication. The parameters are given by the particular device connected to the PC and should be found in the device documentation or available from its manufacturer. In this *Serial Port Setup* panel (Fig. 18) the port number (COM1 to COM4) must be chosen as well as transfer rate, stop bits, data bits, parity and other parameters. Particular settings may be saved for later use by pressing *Save* button. Previously saved settings can be restored using the *Read* button. The main PCom window (Fig. 17) has two parts: on the left – a pane for monitoring the received data and on the right – the control elements for the communication. In the monitoring pane all the read and filtered data which are sent to the application will appear. This can be used for visual check of the data at any time. The following paragraphs deal with particular functions of PCom which make it sure for the characters to have valid number format required by QC.Expert.



PCom (c) TriloByte 1999

Recieved Characters: □ Hexadecimal (0..FF) □ Route to keyboard

Function: ● Measure ○ Terminal

Send ×1

Send: ● As text ○ As hexa(0..FF) □ Add New line

Interval: 1 sec.

□ Accept numbers only □ Add New line

□ Receive by 9 Bytes

Mask:	1	2	3	4	5	6	... #
	X	X	X	#	#	#	... X

□ Wait (synchronous read/write)

Fig. 17 Main window of the PCom communication module

Control elements of PCom module

Received characters

Hexadecimal – if checked, prints the received data as hxadecimal codes of the respective characters, which is useful when the input data contain non-printable characters (ASCII 0-31) like BkSp, CR, LF, TAB.

Route to keyboard – when checked, redirects the input from port to keyboard thus simulating normal keyboard. Received characters may thus be sent to any application (not only QC.Expert!) just as if they were typed on keyboard. The keyboard functionality is not affected in any way.

Port settings – opens the port settings panel mentioned before.

OPEN/CLOSE – opens/closes port. Pressing the *Open* button will start the serial communication using the current settings and the button chanes to *Close*. Presing the Close button again will stop the current communication and close the port. The port is also closed when the PCom module is closed.

Function

Measure – This is the default measuring mode of PCom module. Use this mode for all measurements.

Terminal – Alternative possible use of PCom which can here serve as a simple terminal which receives and/or sends data through the serial port. This mode may be used for communication between two computers or to receive data typed on another computer connected by a serial „Kermit“ type cable. PCom must be active on both computers with equal settings.

Send – This button starts to send repeatedly the character sequence (or „string“) typed in the field above this button to the serial port. The time period in seconds for sending the string must be specified in the field *Interval*. The number of repetitions are specified in the field „*X*“. This function is used for example when your device requires some command to send data to the computer. Check „*Send As Hexa*“ if this command contains non-printable characters (ASCII < 32). Then all the character string must be written in the hexadecimal code, i.e. two digits 0 – F. For example, the text „**AB 1<tab>Z**“ will be written as „**41422031095A**“ (<tab> stands for the non-printable tabulate character, ASCII 09; the code **20** is a space).

Add new line – A New line (CR LF, **0D0A** hexa) will be added after each character string sent to the device.

Interrupt – Interrupts sending commands to the device.

Accept numbers only – A filter which deletes all non-number characters from the received characters, leaving only numbers *0-9*, decimal point, minus and plus signs.

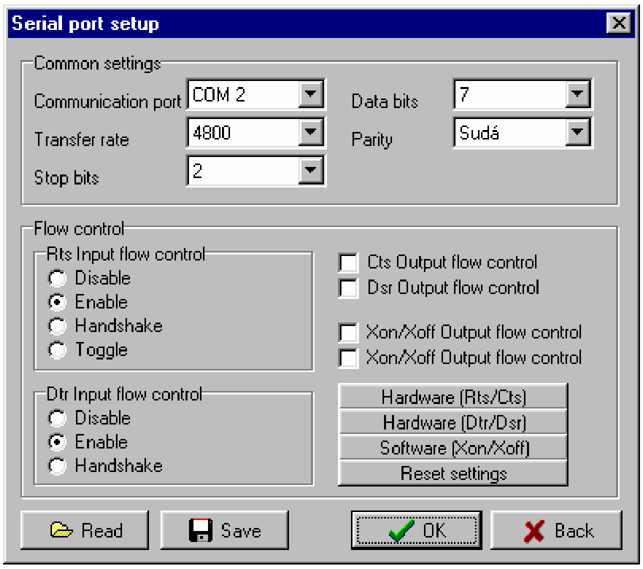
Add new line – Adds the New line (CR LF, **0D0A** hexa) at the end of each received character string.

Receive by – Receives only the first *N* characters from each character string received from the device. The rest of the characters are ignored.

Mask – Mask can be used to filter the unwanted positions in the received character string. It is active only when the „*Receive by*“ option is checked. In the numbered fields, choose „**#**“ (accept), or „**X**“ (ignore). The number of positions in mask is equal to the number of characters specified in the „*Receive by*“ option.

Minimize – Minimizes the PCom window. The eindow can be recalled by clicking on the button in the Windows panel.

Close PCom – Closes port, stops all communication and closes the PCom communication.



Serial port setup

Common settings

Communication port: COM 2

Transfer rate: 4800

Stop bits: 2

Data bits: 7

Parity: Sudá

Flow control

Rts Input flow control
- ○ Disable
- ● Enable
- ○ Handshake
- ○ Toggle

Dtr Input flow control
- ○ Disable
- ● Enable
- ○ Handshake

□ Cts Output flow control

□ Dsr Output flow control

□ Xon/Xoff Output flow control

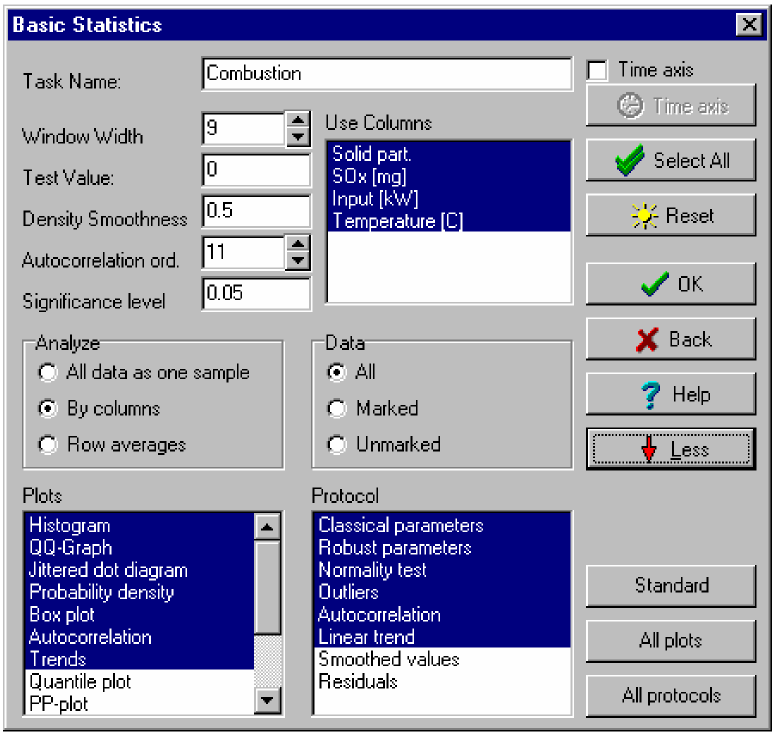
□ Xon/Xoff Output flow control

Fig. 18 Serial port setup panel

5. Basic statistics

Menu:	QCExpert	Basic statistics

The basic statistics module is useful for a preliminary data analysis, as well as for a more detailed look at the data. Various tools from this module can also be used to test whether the data are consistent with assumptions needed for a successful application of other statistical methods. Some of the basic and common assumptions about data are: normality, independence and homogeneity. Therefore, no outliers and gross errors should occur in the data.



Task Name: Combustion

□ Time axis

Window Width 9

Test Value: 0

Density Smoothness 0.5

Autocorrelation ord. 11

Significance level 0.05

Use Columns
Solid part.
SOx [mg]
Input [kW]
Temperature [C]

Analyze
○ All data as one sample
● By columns
○ Row averages

Data
● All
○ Marked
○ Unmarked

Plots
Histogram
QQ-Graph
Jittered dot diagram
Probability density
Box plot
Autocorrelation
Trends
Quantile plot
PP-plot

Protocol
Classical parameters
Robust parameters
Normality test
Outliers
Autocorrelation
Linear trend
Smoothed values
Residuals

Fig. 19 Basic data analysis dialog panel

5.1. Data and parameters

Data are organized into columns (variables). The first row always contains column names. When “All data” or “Columns” choices are selected in the Basic data analysis dialog panel, columns of different lengths are allowed. When computations are requested for “Subgroup means”, length of all columns should be the same. Minimum number of columns is 1. Minimum number of data points is 3. Columns to be analyzed can be selected in the *Columns* window inside the *Basic data analysis* panel, see Figure 17. Various other parameters can be set there as well.

Trend order determines how many consecutive data points will be used to compute moving averages and moving medians. The value should be smaller than half of the sample size.

Test the mean value The value μ_0 for a t-test is entered here. The program tests whether the true mean of the data is different from μ_0 at a specified significance level.

Density smoother The kernel width of kernel smoother is entered here. The smoother is used to estimate probability density function. High values of the parameter result in smoother probability density estimate and vice versa. The parameter has to be positive, a value about 0.5 is recommended.

Autocorrelation order gives the maximum lag for which the autocorrelation coefficient is computed. The value has to be smaller than the sample size minus 2.

Significance level gives significance level for statistical tests and confidence level for confidence intervals. It has to be positive and smaller than 0.5. The parameter multiplied by 100 gives the value in percent. A commonly used value is 0.05 (i.e. 5%).

Computations done for:
All data Data from all selected columns will be used for computations as if they came from a single column.
Columns Computations will be applied for each of the selected columns separately.
Subgroup means Row means for selected columns are computed. If columns differ in length or they have missing values, the computation is performed for complete rows only. This computation is most useful for data diagnostics and X bar control charts.
Time axis window is checked when there is a column containing time values among the data. The time column is identified by pressing the *Time axis* button.
Select all selects all columns in the active sheet for further computations.
Default values This button can be used when one is not sure whether previously entered parameter values are correct. When the button is pressed, parameters in the dialog panel are set to default values.
More/Less button specifies amount of output requested. Requested amount of both printed and graphical output is specified here. Pressing the *Standard* button produces the usual (reduced) amount of output containing the most important information only. *All graphs* and *All protocols* buttons are used to request the complete output.

Note: the size of objects produced when *Smoothed values* and *Residuals* items are selected depends on the sample size. They can fill the output sheet completely for large data sets.

5.2. Protocol

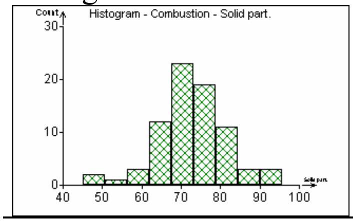
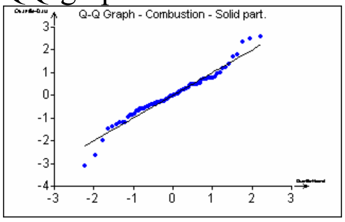
Column	Column name.
Row number	Total number of rows in the analyzed dataset.
Number of valid data points	Total number of valid data points in the dataset.
Number of missing values	Number of empty cells in the dataset.
Classical parameters	
Arithmetic mean	Mean value estimate for normal data.
Lower limit	Lower limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on arithmetic mean.
Upper limit	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on arithmetic mean.
Variance	Variance estimate.
Standard deviation	Square root of the variance estimate.
Skewness	Estimate of the third moment, skewness.
Difference from 0	Skewness for normal as well as other symmetrical distributions is zero. When the skewness is significantly different from 0, the data cannot be assumed to come from a symmetrical distribution. The test for normality (see later) is more useful.
Kurtosis	Estimate of the fourth moment, kurtosis.
Difference from 3	Kurtosis for normal distribution is 3. When the kurtosis is significantly different from 3, the data should be assumed to come from a non normal distribution. The test for normality (see later) is more useful
Half-sum	Half sum estimate, i.e. half of (maximum plus minimum)

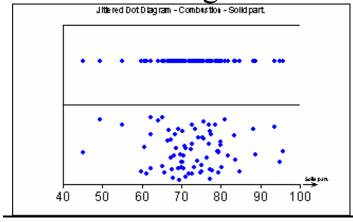
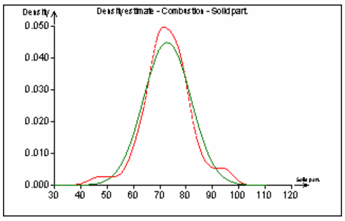
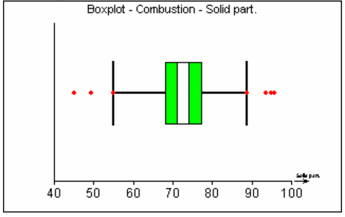
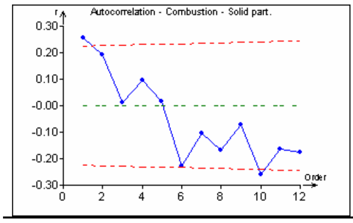
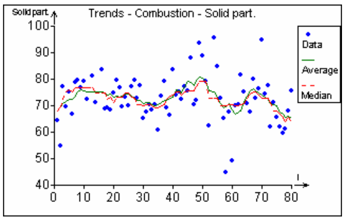
Modus	Estimate of the modus, i.e. location of the probability density function maximum.
t-test	
Hypothesized value	The value, entered to the “Test the mean value” field from the “Basic data analysis” panel.
Difference from the hypothesized value.	A comment, describing in words whether mean value is significantly different from the hypothesized value at specified significance level. The hypothesized value can be entered in the Basic data analysis panel, see Figure 17.
Calculated	Calculated test criterion.
Theoretical	Appropriate t-distribution quantile.
p-value	The smallest significance level for which the equality of true mean to the hypothesized value is rejected when using the observed data. When the p-value is smaller than a selected significance level, the true and hypothesized mean are significantly different.
Robust parameters	
Median	Estimate of the median, i.e. 50th percentile. It might be a more useful estimate of the mean value than the arithmetic mean, when normality does not hold or when outliers are present in the analyzed data.
CI lower	Lower limit of the confidence interval for the median, computed for a specified confidence level.
CI upper	Upper limit of the confidence interval for the median, computed for a specified confidence level.
Standard deviation	Median based standard deviation.
Variance	Median based variance.
10% trimmed mean	The arithmetic mean computed from symmetrically trimmed data, i.e. after omitting 5% smallest and 5% largest values. This robust estimate of the mean is recommended when outliers are expected.
CI lower	Lower limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
CI upper	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
Variance	Variance estimate based on median.
Standard deviation	Standard deviation estimate based on median.
20% trimmed mean	The arithmetic mean computed from symmetrically trimmed data, i.e. after omitting 10% smallest and 10% largest values. This robust estimate of the mean is recommended when suspecting outliers presence.
CI lower	Lower limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
CI upper	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
Variance	Median based variance.
Standard deviation	Median based standard deviation.

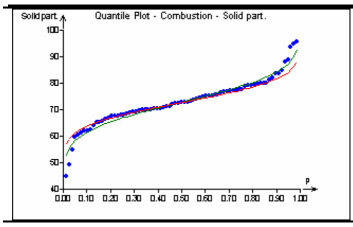
40% trimmed mean	The arithmetic mean computed from symmetrically trimmed data, i.e. after omitting 20% smallest and 20% largest values. This robust estimate of the mean is recommended when suspecting outliers presence.
CI lower	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
CI upper	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
Variance	Median based variance estimate.
Standard deviation	Median based standard deviation.
Runs test	Test for randomness in the sequence of values larger and smaller than the mean. When the larger/smaller pattern is too regular or when the differences from the mean form long runs with many consecutive values of the same sign, the data are suspect for the lack of independence. Such data are considered to be dependent.
Small sample analysis	Mean value estimate and confidence interval computed by the Horn's quantile based method. This estimator is recommended for small samples of 3 and more data points. The method usually produces more correct values than the arithmetic mean for such small sample sizes. The method should not be used for N>20.
N	Sample size.
Mean value	Estimate of the mean value.
Lower limit	Lower limit of the confidence interval.
Upper limit	Lower limit of the confidence interval.
Test for normality	A test for checking normality, based on both skewness and kurtosis. The output contains values of classical statistical characteristics
Normality	Conclusion of the test performed at a specified significance level, described in words.
Calculated	Calculated test criterion.
Theoretical	Appropriate t-distribution quantile.
p-value	The smallest significance level for which the normality is rejected using the observed data.
Outliers	A robust, quantile based test procedure to check whether outliers are present.
Homogeneity	Conclusion of the test for outlier presence, commented in words.
Number of outliers found	The number of data points, which can be considered as outliers, based on the previous test.
Lower limit	The data below this limit are considered to be outliers.
Upper limit	The data exceeding this limit are considered to be outliers.
Autocorrelation	Autocorrelation estimates and related tests performed at a selected significance level.

Autocorrelation order	Autocorrelation order.
Coefficient	Autocorrelation coefficient estimate. It is the correlation coefficient for a pair of variables, discussed in the paragraph 5.3., computed for a special choice of the two variables.
p-value	The smallest significance level for which the zero autocorrelation is rejected using the observed data. When the p-value is smaller than a specified significance level, the autocorrelation is significant.
R0Crit	Critical value for the autocorrelation test. Autocorrelation estimates larger than this value are significant.
Result	Conclusion of the autocorrelation test, described in words.
Test for a linear trend	Test for linear trend in the data. Even when the linear trend is not significant, a trend of nonlinear character might still be present in the data, e.g. periodic oscillations. Some nonlinear trends might be detected by other tests, e.g. by the runs test.
Slope	Slope of the fitted line.
Significance	Conclusion of the linear trend test in terms of slope significance, commented in words.
Theoretical	Appropriate t-distribution quantile.
p-value	The smallest significance level for which the hypothesis of no linear trend is rejected using the observed data. When the p-value is smaller than a specified significance level (usually 0.05), the trend is considered to be present (significant).
Smoothed values	Smoothed (detrended) values obtained by running means or running medians. With an appropriate choice of the running mean/median length a trend can be estimated by the smoother. Detrended data can be obtained by substracting the trend estimate from the original data. When showing a strong trend, data should be appropriately detrended before they are used in control charts.
Residuals	Residuals, i.e. the differences between observed and smoothed data.

5.3. Graphical output

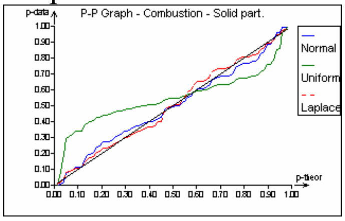
Histogram  	Frequency histogram with a constant bin width; optimum number of bins is selected automatically, with respect to the sample size. Clicking the right mouse button on a bar in the dynamical graph shows frequency and limits of the class selected.
QQ-graph  	A graphical tool for checking normality and outliers presence; for normal data without outliers the points should lie close to the line; for normal data with outliers, points in the central parts should lie close to the line the endpoints further away from the line; for data coming from a positively skewed distribution (e.g. lognormal, exponential) the shape should be nonlinear, convex ; for data coming from a negatively skewed distribution the shape should be nonlinear, concave ; for data coming from a distribution with curtosis higher than normal, i.e. those showing high concentration around the mean (e.g. Laplace), the shape should be

	concave-convex ; for data coming from a distribution with curtosis smaller than normal, i.e. those with small concentration around the mean (e.g. uniform), the shape should be convex-concave . One advantage of the QQ-graph, compared to the statistics describing skewness, kurtosis etc. is that one can visually check whether the lack of normal appearance (nonlinearity) is caused by just a few points or whether it is a general tendency shared by all data.
Jittered dot diagram  	Data are plotted along the x-axis. y-axis has no physical meaning. The y-coordinates of the plotted points are random. This technique allows for a better recognition of individual data points which might coincide if plotted just along the x-line when their observed values are similar or the same.
Probability density function  	Comparison of the normal probability density curve (solid green line) with a kernel density estimate, computed from the data (dashed red line). The kernel estimate uses the Gaussian kernel. Smoothness of the estimate is given be the kernel width, entered as “Density smoother” parameter in the dialog panel shown in Figure 18. Smaller parameter values cause more rugged shape of the estimate (following more details of the data). When the data are not homogeneous and show a clustering tendency, several local maxima of the density estimate can occur. For normal data, both curves should be close to each other. On the other hand, one should realize that with small enough smoothing parameter, local maxima occur for any data.
Boxplot  	This is a standard diagnostic tool. The large box contains 50% of the data, its upper edge corresponds to 75th percentile, its lower edge to the 25th percentile. Median is located in the middle of the white rectangle inside the green box. Width of the white rectagle inside the green box corresponds to the width of the confidence interval for the median. Two black lines correspond to the inner fence. The data points outside the inner fence are marked red. They might be considered as outliers.
Autocorrelation  	Plot of autocorrelation coefficients against lag up to the maximum lag, specified in the dialog panel, see the Figure 17. Red lines show the limits beyond which the coefficient is considered to be significant at a previously selected significance level. When some of the coefficients exceed these limits, the data should be considered to be dependent.
Trends  	Plot of the smooth trend in the data, estimated by the running mean smoother (solid line) and the running median smoother (dashed line). The “Trend order” parameter is inputted in the dialog panel shown in Figure 17. Higher values of the parameter result in smoother curves, less sensitive to local behavior of the data, showing a global trend. Small parameter values lead to curves sensitive to a local data behavior. The running median smoother is less sensitive to errors and isolated outliers in the data (it is robust), so that it is recommended in cases when such problems are expected. When the linear trend test yields a significant result, the regression line is plotted as well
Quantile plot	Shows empirical quantiles and the inverse cumulative distribution function (the quantile function, QF) of the fitted normal distribution. The green curve corresponds to the normal QF with the classical estimates of the



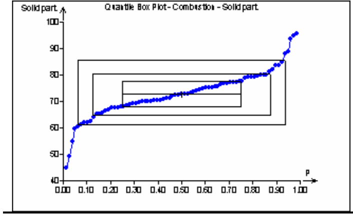
parameters (non-robust), the red curve corresponds to the median based estimates of the parameters (robust). Depending on which of the curves fits the data better, either mean or median might be chosen as an estimate of the mean value.

PP-plot



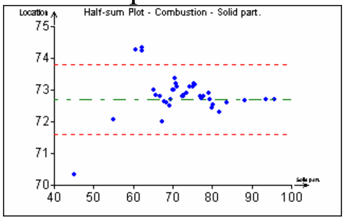
Compares data distribution with several theoretical models, using the empirical cumulative distribution function and cumulative distribution function of normal (solid blue curve), Laplace, and uniform distributions. A model which fits data well should plot approximately as the $y = x$ line. The plot can be used to distinguish among symmetrical distributions according to their kurtosis. Apparent similarity to the uniform distribution suggests that the data were truncated (both small and large values excluded).

Quantiles box plot



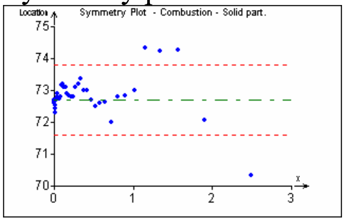
Points are plotted in the same way and have the same meaning as for the quantile graph. Relative position of the plotted rectangles show symmetry resp. asymmetry of the data distribution. The horizontal line inside the smallest rectangle corresponds to median, the vertical edge of the smallest rectangle corresponds to the confidence interval for median.

Half sum plot



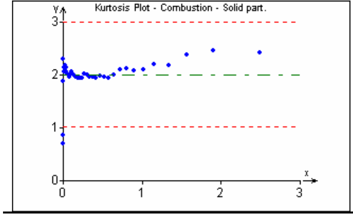
A sensitive indicator for distributional asymmetry. Ideally, the points should lie on a horizontal line. Green horizontal line corresponds to median and dashed red lines to its confidence limits. When the data distribution is asymmetric, the plot shows a clear trend (increasing for a negative skewness and decreasing for a positive skewness), going far beyond the dashed lines. Pairs of data points (first-last, second-second largest, etc.) are used when constructing the plot, so when selecting a point on the plot, two data points are marked in the data table.

Symmetry plot



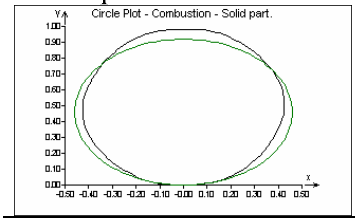
It has a similar use as the Half sum plot from previous paragraph. Slope of a trend is proportional to skewness. When the data distribution is asymmetric, the plot shows a clear trend (increasing for a negative skewness and decreasing for a positive skewness), going far beyond the dashed lines. Pairs of data points (first-last, second-second largest, etc.) are used when constructing the plot, so when selecting a point on the plot, two data points are marked in the data table.

Kurtosis plot



The meaning is analogous to the previous two plots. Slope of its trend is proportional to the difference (kurtosis-3). When the kurtosis is very different from normal, the plot shows a clear trend. Pairs of data points (first-last, second-second largest, etc.) are used when constructing the plot, so when selecting a point on the plot, two data points are marked in the data table.

Circle plot



It is used for a complex visual assessment of normality, considering skewness and kurtosis simultaneously. Green circle (ellipse) is an ideal (for a normal distribution), black “circle” is constructed from data. Both curves should be close to each other for normal data.

6. Two-sample comparison

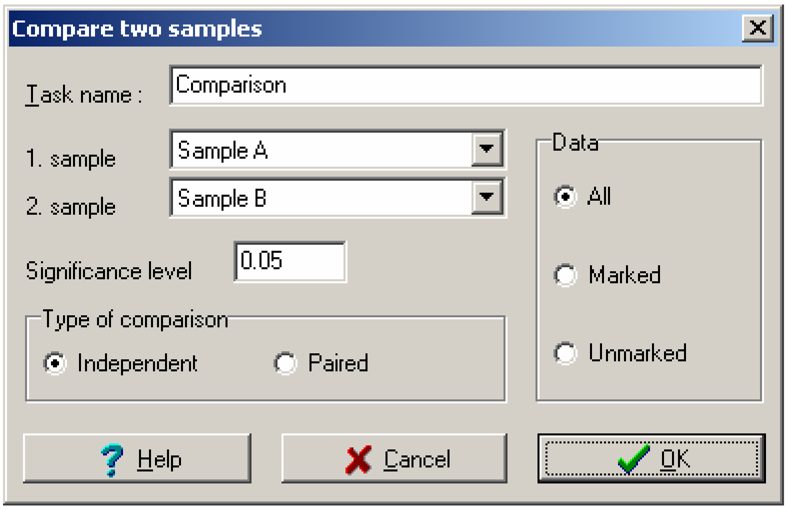
Menu:	QCExpert	Two-sample comparison

This module is intended for a detailed analysis of two datasets (two samples). The module offers two analyses: independent samples comparison, and paired samples comparison.

Independent samples *x*, *y* feature no mutual relationship. They can have different sample sizes, in general. Ordering of the elements of both samples is arbitrary and can be changed without any information loss. Main point of this analysis is to decide, whether the expected values E(*x*) and E(*y*) of the two samples are different. Weight of peanuts from two different locations can serve as an example of two independent samples. On each location, a few dozens of the peanuts are selected at random and weighted individually.

On the contrary, the paired test focuses on comparison of two related datasets, for instance on two sets of measurements, taken on the same units, under different circumstances. Measurements of each unit come in *x*, *y* pairs. The paired test can be performed to decide whether the different conditions influence measurements on the same unit. Technically, the paired comparison goes through the test of whether the expected value of the ***difference*** between first and second variable, $E(x - y)$ is significantly different from zero. For example, consider comparison of blood cholesterol levels for a group of patients, measured before and after a particular medical treatment. There have to be the same number of pre- and post-treatment measurements (patients who might dropped from the study during the treatment are omitted). Relative ordering of the pre- and post-treatment measurements is important: both measurements of the same patient have to appear on the same line.

6.1. Data and parameters



Task name : Comparison

1. sample: Sample A

2. sample: Sample B

Significance level: 0.05

Type of comparison: ● Independent ○ Paired

Data: ● All ○ Marked ○ Unmarked

Fig. 20 Dialog panel Two-sample comparison

Names of the columns holding values of the first and second variable have to be entered in the dialog panel. In the *Comparison type* part, one has to specify whether the test for *Independent samples* or *Paired samples* is requested. Although the *Significance level* is set to the 0.05 (5%) by default, it can be edited. Similarly as with all other modules, analysis can be requested either for *All* data, or *Marked* data, or *Unmarked* data.

Independent samples
Data are in two columns, whose lengths can be different. Empty cells will be omitted.

Paired samples
Data are in two columns, whose lengths should be the same. If any of the two values in the same row is missing, whole row is omitted.

6.2. Protocol

Protocol content is different when independent samples and when paired samples were tested. The same is true for graphical outputs. Both output versions are described below.

Independent samples

Task name	Project name taken from the dialog panel.
Significance level	Required significance level α.
Columns to compare	Names of the columns containing samples to compare.
Sample size	The sample size of first dataset (n_1) and second dataset (n_2).
Average	Arithmetic averages of the first and second column, $\bar{x}_1$, $\bar{x}_2$.
Standard deviation	Standard deviations of the first and second sample, s_1 a s_2.
Variance	Variance of the first and second sample, s_1^2 and s_2^2.
Correl. coeff. R(x,y)	This entry, together with the warning „Significant correlation!“ will appear only in the case that correlation between the two columns is significant (significantly different from zero) at the significance level α. In such a case, there might be a serious problem with the data and/or their collection procedures, or paired comparison might be called for. If this row is not included in the Protocol, correlation coefficient is not significantly different from zero.
Variance equivalence test	Also called Variance homogeneity test. Tests whether the two sample variances are different. The test is based on approximate normality. Specifically, the data should not contain any outliers. If that is not the case, robust variance estimates should be used instead (see below).
Variance ratio	Test statistic, max($\sigma_1^2/\ \sigma_2^2$, $\sigma_2^2/\ \sigma_1^2$)
Degrees of freedom	Degrees of freedom that are used to look up the critical value, i.e. the value of the quantile of the *F*-distribution with n_1-1 and n_2-1 degrees of freedom
Critical value	*F*-distribution quantile, $F(\alpha, n_1-1, n_2-1)$
Conclusion	Variance homogeneity test conclusion in words: „Variances are not different“, or „Variances are different“.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about variance homogeneity were rejected for the given data.
Robust variance test	Alternative variance homogeneity test for two samples. It is intended for non-normal data, mainly those coming from distributions differing from the normal distribution by skewness. The test should not be used for normal data (due to a lower power).
Variance ratio	Test statistic, max($\sigma_1^2/\ \sigma_2^2$, $\sigma_2^2/\ \sigma_1^2$).
Corrected degrees of freedom	Degrees of freedom corrected for the departure from normality.
Critical value	*F*-distribution quantile.
Conclusion	Variance homogeneity test conclusion in words: „Variances are not different“, or „Variances are different“.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about variance homogeneity were rejected for the given data.

Mean equivalence test for Equivalent varances	Test of the null hypothesis of equal means in the case of equal variances. When the variances are significantly different, *unequal* variances version of the test needs to be used, see below.
t-statistic	Test statistic.
Degrees of freedom	Degrees of freedom for the t-test.
Critical value	t-distribution quantile.
Conclusion	Test conclusion in words.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about equal means would be rejected for given data.
Mean equivalence test for Different variances	Test of the null hypothesis of equal means in the case of unequal variances. When the variances are not significantly different, *equal* variances version of the test needs to be used, see above.
t-statistic	Test statistic.
Degrees of freedom	t-test degrees of freedom.
Critical value	t-distribution quantile.
Conclusion	Test conclusion in words.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about equal means is rejected for given data.
Goodness of fit test	
Two sample K-S test	Kolmogorov-Smirnov test, comparing distributions generating the two independent samples. It is based on maximum difference between empirical distribution functions (computed from the two samples). Note that it is possible that both means and variances are not significantly different, while the KS test shows significant difference between the distributions. Typically, this is connected to a substantial difference of at least one of the distributions from normality (usually asymmetry or bimodality). Data are not suitable for the simple t-test, then.
Difference DF	Maximal empirical distribution functions difference. It is the test statistic for the KS test.
Critical value	KS-distribution critical value.
Conclusion	Test conclusion in words: „Distributions are significantly different“ or „Distributions are not significantly different“

Paired samples

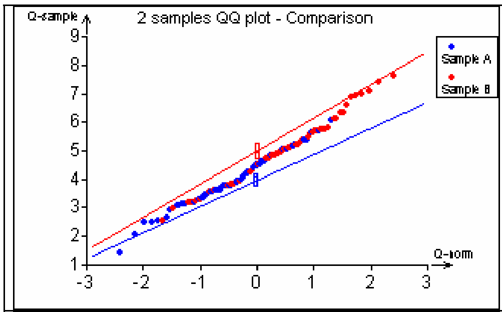
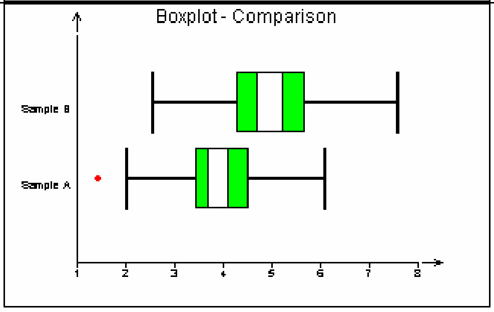
Task name	Project name taken from the dialog panel.
Significance level	Required significance level α.
Columns to compare	Names of the columns containing samples to compare.

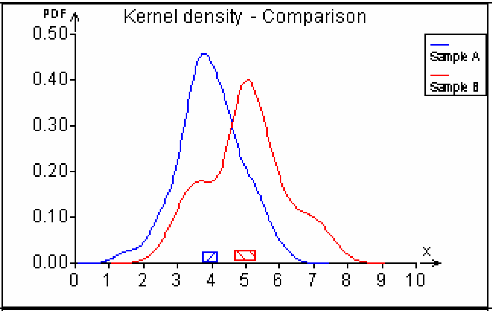
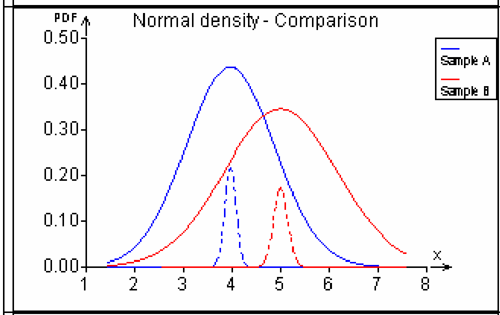
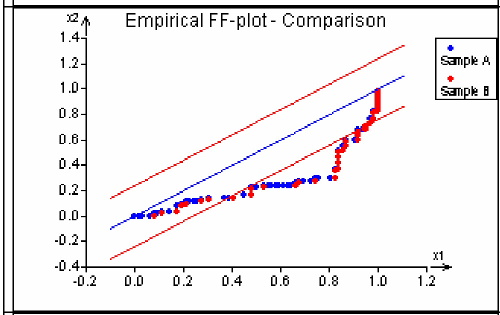
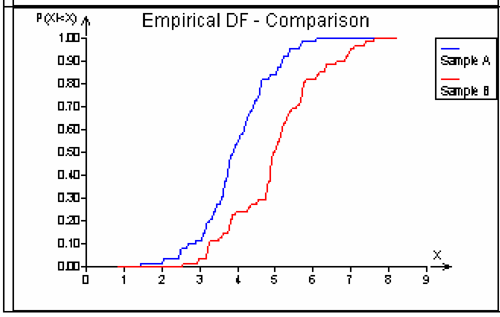
Analysis of differences	
Sample size	Number of data pairs, n.
Average difference	Arithmetic mean of the difference between the first and second variable $x_1–x_2$ (first of a pair – second of a pair), $\bar{x}_d$
Confidence interval	(1-α)% confidence interval for arithmetic mean of differences.
Standard deviation	Standard deviation of the differences, s_d.
Variance	Variance of the differences, s_d^2.
Correlation coefficient R(x,y)	Sample correlation coefficient r. It estimates correlation between the first and second data column. When the correlation is not significant, red warning will appear. Paired comparison choice is somewhat suspicious. There might be some problem with the dataset. For instance, relative ordering of the first and second columns might be distorted. Or, the x_1, x_2 pairs come from a box that is too narrow/low.
Test of difference	The test of difference between the first and second pair members.
t-statistic	Test statistic, $\bar{x}_d \cdot \sqrt{n} / s_d$.
Degrees of freedom	Number of degrees of freedom, $n-1$.
Critical value	
Conclusion	Test conclusion in words. The differences are either „NOT SIGNIFICANTLY different from zero“, or „SIGNIFICANTLY different from zero“.
p-value	p-value corresponds to the smallest significance level on which the null hypothesis about mean difference being equal to zero is rejected for given data.

6.3. Graphical output

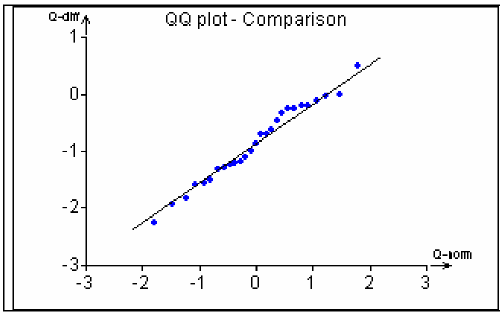
Graphical output is different, according to whether paired or independent samples comparison was choice was selected (similar to the Protocol differences).

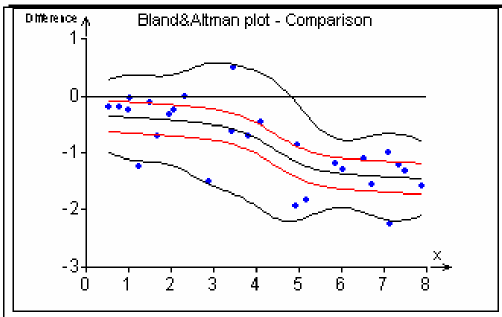
Independent samples

	Q-Q plot for all data. All data are plotted as one sample. First or second sample data are plotted in different colors (see the legend). The two sample means are marked and their confidence intervals are plotted as hatched boxes. Plotted lines' slopes correspond to standard deviations of the two samples. Hence, the steeper line corresponds to sample with a larger standard deviation.
	Boxplots help to compare the samples visually. Larger box contains inner 50% of the data. Right border of the green box corresponds to the 75th percentile. The left border of the green box corresponds to the 25th percentile. Center of the white band corresponds to the median. White band corresponds to the confidence interval for median. Two black whiskers correspond to the so-called inner fences. All data beyond the inner fences are plotted individually, as red points. They are suspicious and can be considered as outliers. Asymmetric placement of the

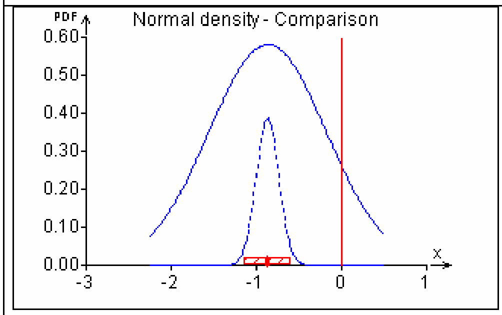
	white band in the green box shows data distribution asymmetry.
	Kernel density estimates computed for the two samples separately. Blue curve corresponds to the first sample, while the red curve corresponds to the second sample. Confidence intervals for means are plotted as hatched boxes. When these boxes do not overlap, the means are statistically different on the significance level selected.
	Gauss‘ density curves corresponding to the two samples‘ means and variances. The colors assignment is the same as on the previous plot. For comparison purposes, there are densities for the two arithmetic averages plotted as well (the y-coordinate is shrunken down).
	Joint empirical F-F plot for testing distributional differences between two independent samples. Empirical distribution function values for the first and second samples are plotted as *x* and *y* coordinates. (The empirical distribution functions are shown on the next plot.) If the two distributions are not significantly different, the points are close to the central (blue) line. If the any of the points falls beyond one of the two red lines, then the distributions differ significantly.
	Empirical distribution plot for the first and second sample. The *Y* coordinate corresponds to the distribution function value (i.e. to the probability that there is a measurement smaller than or equal to the value of the X coordinate).

Paired samples

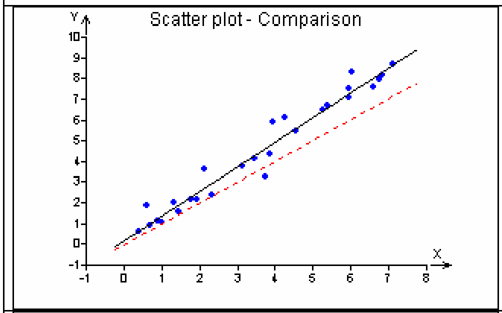
	Q-Q plot for checking normality of differences between first and second member of the pair graphically. If the points are placed close to the line, the data do not look non-normal. When this is not the case, substantial departure from normality is suggested. Information content of the tests reported in the Protocol can be seriously impaired then.



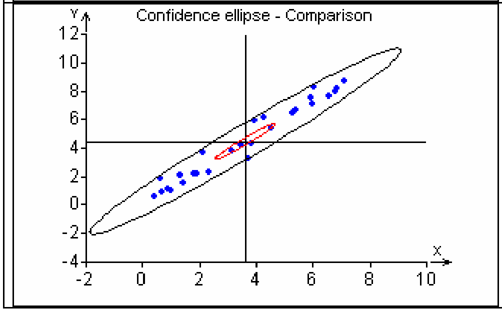
Bland and Altman plot. Average of a particular data pair is plotted on the X-axis, while the difference for the same pair is plotted on the Y-axis. This plot helps to detect possible dependence between variability (estimated by the pair members difference) and value attained. For a better orientation, horizontal zero line is added to the plot. Smoothed average of the differences is plotted as a function of average (black curve). Corresponding confidence interval of this estimate is plotted, with the two red curves. Smoothed value ±2σ, where σ is estimated nonparametrically and it is plotted as two black curves. Ideally, (when the difference is uniformly zero for any value of average), the red curves should contain horizontal zero line, and the ±2σ band should be approximately linear, parallel to the horizontal zero line.



Gaussian density curve. The parameters are estimated from the differences between pair members under the normality assumption. The inner curve corresponds to the approximate density of the arithmetic mean of the difference (*Y* coordinate is shrunken for better readability). Vertical red line corresponds to zero difference. The hatched box corresponds to the mean difference confidence interval. If zero is contained in the interval, then the mean difference between first and second sample is not significantly different from zero.



This plot is useful when judging degree of interdependence between the first and second sample. *y*=*x* line is plotted in red (dashed). It corresponds to the zero difference. The second line corresponds to the best *y*-depends-on-*x* line, fitted to the data.



The same points as on the previous plot are plotted here. The data points here are viewed as a sample from a bivariate normal distribution, however. Black ellipse corresponds to the region containing approximately100.(1-α)% of the data (under bivariate normality). The red ellipse corresponds to the border of the 100.(1– α)% confidence region for the vector of two means.

7. Probabilistic models

Menu:	QCExpert	Probabilistic models

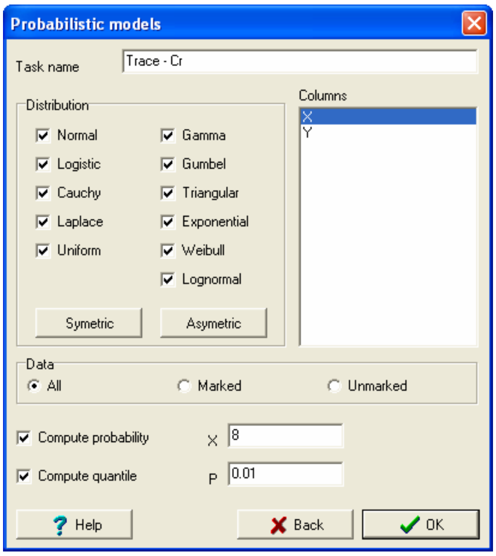
The module fits statistical models from various classes by the method of maximum likelihood (MLE, *Maximum Likelihood Estimate*) . There are 11 univariate distribution types available for use, 5 of them are symmetric, remaining 6 are asymmetric. For each distribution type, the module fits a model of a given type by looking for the best parameter values via maximization of the likelihood function. Best fitting models found in each distributional class are compared across classed using the following two criteria: likelihood function value and linearity (correlation coefficient) of the P-P plot. Within a distributional class, best fitting parameters are found by the numerical maximalization of the likelihood function, or of its‘ logarithm *L*, respectively,

$$L\left(A,B,C,\mathbf{x}\right)=\sum_{i=1}^{n}\ln f\left(A,B,C,x_i\right),$$

where A,B,C are parameters of the given distribution type, **x** is the n vector of measured data, and f is probability density of the given distribution, see below. In addition to the parameter estimates for each of the models and some further statistics, the module can produce diagnostic plots and user-required quantiles. The module can be useful in situations where data are not normally distributed. It can be used not only in the situation where the non-normal model to be used is known in advance (e.g. from a physical theory), but even if one is not sure about the distribution type that should be used to fit the data.

7.1. Data and parameters

The module is intended for analysis of the data, which come from ***symmetrical*** or ***positively skewed*** data. Negatively skewed data need to be multiplied by (-1) before the analysis. The data to be analyzed should be entered in one column, whose name is selected in the *Columns* field. In the *Distribution* field, one can choose which of the available models will be fitted. Symmetric distributions are listed on the left, asymmetric are listed on the right. By default, all 11 types are fitted. *Symmetric* and *Asymmetric* buttons can be used to enter all symmetric, respectively asymmetric distributions available. At least one distribution type has to be selected. Asymmetric distributions' fitting is generally longer. Hence, if we are not specifically interested in fitting asymmetric models, it is better not to choose them If the data are symmetric, or if they have a positive skewness, skewed distribution fitting can even fail. Such a failure is indicated by the message „Not available“, or „Error“. One can choose in the Data field whether all data, marked only, or unmarked only data should be used for the calculations. If the *Calculate probability* selection is checked, the user has to supply a value in the *X* field. Probability that a random variable with the fitted distribution is smaller than or equal to this value (i.e. the cumulative distribution function value), will be returned by the software. This probability will be listed in Protocol for all models whose fitting is requested. If the *Compute quantiles* selection is checked, the user has to enter a probability, say p (a value between 0 and 1, $0 < p < 1$), for which the quantiles are to be evaluated. The Protocol lists p-th and (1- p)-the quantiles for each of the fitted models.



Probabilistic models

Task name: Trace - Cr

Distribution

☑ Normal ☑ Gamma
☑ Logistic ☑ Gumbel
☑ Cauchy ☑ Triangular
☑ Laplace ☑ Exponential
☑ Uniform ☑ Weibull
☑ Lognormal

Columns: X (selected), Y

Data: ● All ○ Marked ○ Unmarked

☑ Compute probability X 8

☑ Compute quantile p 0.01

Fig. 21 Dialog panel for the Probabilistic models

Below, we list probability density functions for all available models. For each of them, parameter restrictions are listed (if there are any). Probability density function is generically denoted by $f(x)$, while the cumulative distribution function is denoted by $F(x)$, and the quantile function by $F^{-1}(x)$.

Normal distribution:

$$f(x)=\frac{1}{B\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-A}{B}\right)^2\right),\ A\in R, B>0$$

Uniform distribution:

$$f(x)=\begin{Bmatrix}\dfrac{1}{B-A} & pro\ A\le x\le B\\ 0 & jinak\end{Bmatrix},\ A,B\in R, A<B$$

Laplace distribution:

$$f(x)=\frac{1}{2B}\exp\left(-\frac{|x-A|}{B}\right),\ A\in R, B>0$$

Logistic distribution:

$$f(x)=\frac{1}{B}\frac{\exp\left(\dfrac{x-A}{B}\right)}{\left[1+\exp\left(\dfrac{x-A}{B}\right)\right]^2},\ A\in R, B>0$$

Cauchy distribution:

$$f(x)=\frac{1}{\pi B\left[1+\left(\dfrac{x-A}{B}\right)^2\right]},\ A\in R, B>0$$

Asymmetric distributions

Exponential distribution:

$$f(x)=\frac{1}{B}\exp\left(\frac{A-x}{B}\right),\ x\ge A, B>0$$

Gamma distribution:

$$f(x)=\frac{1}{B\Gamma(C)}\left(\frac{x-A}{B}\right)^{C-1}\exp\left(\frac{A-x}{B}\right),\ x>A, B>0, C>0$$

Gumbel distribution:

$$f(x)=\frac{1}{B}\exp\left(\frac{A-x}{B}\right)\exp\left(-\exp\left(\frac{A-x}{B}\right)\right),\ A\in R, B>0$$

Lognormal distribution:

$$f(x)=\frac{1}{C(x-A)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln(x-A)-B}{C}\right)^2\right),\ x>A, B\in R, C>0$$

Triangular distribution:

$$f(x)=\begin{cases}\dfrac{2(x-A)}{(B-A)(C-A)}\ pro\ x<C\\ \dfrac{2(B-x)}{(B-A)(B-C)}\ pro\ x\ge C\end{cases},\ x>A, B>C$$

Weibull distribution:

$$f(x) = \frac{C}{B}\left(\frac{x-A}{B}\right)^{C-1} \exp\left(-\left(\frac{x-A}{B}\right)^{C}\right), \; x > A, B > 0, C > 0$$

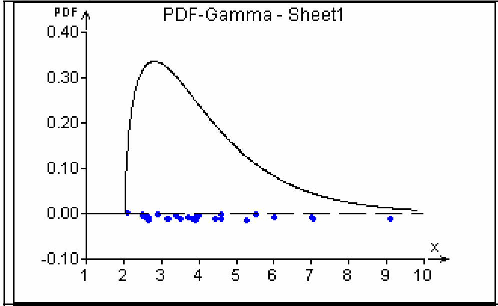
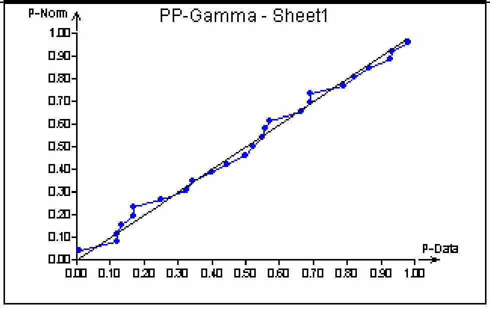
7.2. Protocol

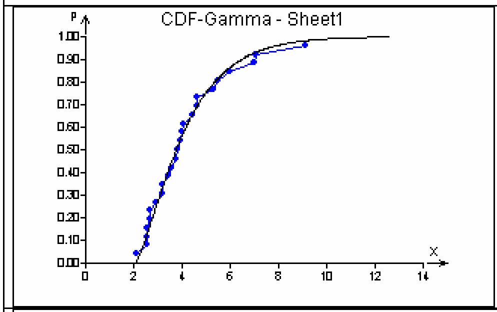
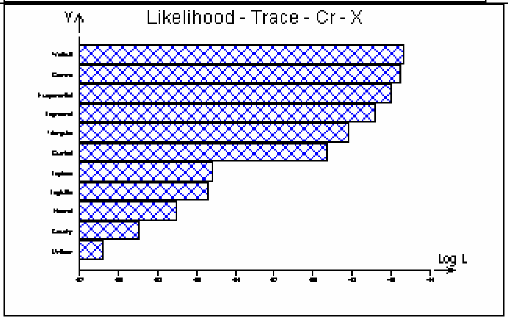
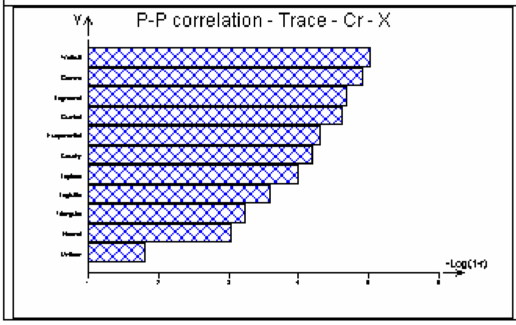
Probabilistic models, Maximum likelihood estimation (MLE)	
Task name:	Name of the spreadsheet containing data.
List of the distributions fitted	This paragraph lists all distributions, which were selected in the Dialog panel and which were fitted to the data subsequently. The list is organized into two parts: Symmetric models and Asymmetric models. For each model, loglikelihood value is listed, together with the correlation coefficient for the P-P plot, and MLE estimates of the parameters.
Loglikelihood	The value of the logarithm of the likelihood function, *L* (for a given distribution). Loglikelihood is suggested as the main criterion to look at, when comparing how well different distributional models fit the data. Large *L* values should correspond to better fitting distributions.
P-P correlation	Correlation coefficient, r_P from the P-P plot. This plot is similar to the Q-Q plot. It is true that the higher r_P (closer to one), the better is the empirical probability $i/(n+1)$ approximated by the theoretical model probability $F(x_i)$. The r_P can be looked at when comparing how good is the fit of different distributional types. This is an alternative to the criterion based on the (log)likelihood, and in general, it does not need to give the same results as before.
Parameters	Parameter estimates, obtained by the maximum likelihood method. Various parameters' meaning should be clear from the definitions of various distribution types, listed in the paragraph 7.1.
Sample moments	Various moment estimates.
Average	Arithmetic average of the data. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	Sample variance. $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Skewness	Sample skewness. $a = \frac{n}{(n-1)(n-2)}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}$
Kurtosis	Sample kurtosis. $b = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4}$
Median	Sample median, $\tilde{x} = \left(x_{(n/2)} + x_{(n/2+1)}\right)/2$ for *n* even, $\tilde{x} = x_{((n+1)/2)}$ for *n* odd.
Moments computed from the fitted models	Moments, median and mode, which are computed analytically as functions of model parameters. If no closed form is available, - appears in the table below.

Expected value	$\mu = \int_{-\infty}^{\infty} x\, dF(x)$
Variance	$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\, dF(x)$
Skewness	$g_1 = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} (x-\mu)^3\, dF(x)$
Kurtosis	$g_2 = \frac{1}{\sigma^4} \int_{-\infty}^{\infty} (x-\mu)^4\, dF(x)$
Median	$F^{-1}(0.5)$
Mode	Location of the maximum of the probability density function, $\arg\max_x f(x)$.
Quantiles and probabilities	Cumulative distribution function value at *x* (probability that a random variable with the fitted distribution is smaller than or equal to *x*). Or, p-th, (1-p)-th quantiles for a given p. These are listed in the table for each of the fitted models when *x* and/or *p* is entered in the appropriate field of the Dialog panel (see the **Chyba! Nenalezen zdroj odkazů.**).
Prob(*X*)	Cumulative distribution function value at *x* (probability that a random variable with the fitted distribution is smaller than or equal to *x*), *F*(*X*)
Quant(*p*)	p-th quantile (the value which will not be exceeded with the probability *p,* that is the value which will be exceeded with the probability (1 – *p*)), $F^{-1}(p)$.
Quant(1 – *p*)	(1-*p*)-th quantile, the value that will be exceeded with the probability *p*, $F^{-1}(1-p)$.

7.3. Graphical output

For each of the distributions selected, probability density curve is plotted, together with the P-P plot, and cumulative distribution function plot. For easier visual judgment of how good the model fit is, individual data points are plotted below the plotted curve. The last two plots help to compare quality of the fit across different models.

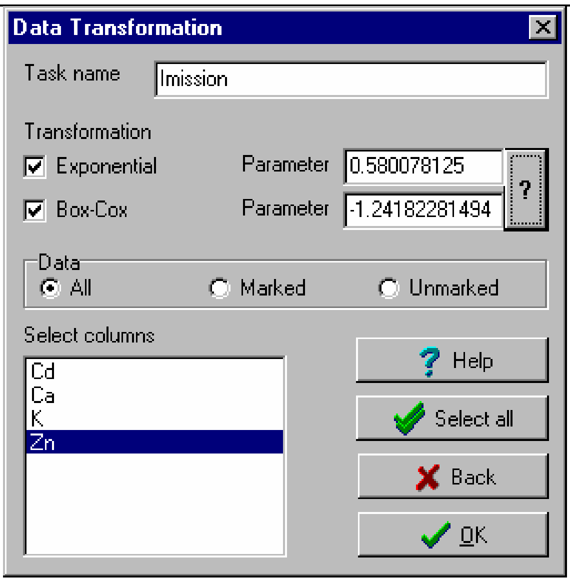
	Probability density function, *f*(*x*) plot. It is drawn for each of the models fitted. Individual data points are plotted below the x-axis. Definitions of probability density function for each of the models available are listed above. This plot has no direct diagnostic meaning.
	The P-P plot. Values of $i/(n+1)$ are plotted on the horizontal axis, while the fitted model cumulative distribution function evaluated at data points, $F(x_i)$ is plotted on the vertical axis. If the empirical and model distribution were exactly the same, then all the points would lie on the *y*=*x* line (it is plotted as well, for comparison). How good is the data distribution approximated by a particular model can be checked informally from visual inspection of how close the data are to the ideal *y*=*x* line. More formally, correlation coefficient can be used to measure this closeness. The correlation coefficient is listed in the Protocol, together with the loglikelihood maximum. The

	correlation is also listed on the *Quality of the fit* plot, see later.
	Cumulative distribution function plot, $F(x)$ for the selected distributions. The model curves are superimposed by the empirical distribution curve to judge the goodness of model fit visually. Empirical distribution function points have y-coordinates (x_i; $i/(n+1)$).
	Overall plot which helps to decide, which models fit the data well, if judged by the loglikelihood values, L. maximum L for each of the selected distribution model type is plotted in the form of the bar chart. The bars are ordered from shortest to longest, for better orientation. The larger the maximum L, the better. If two distribution types show similar maximum L values, they should be judged as being equally good. The maximum value of L is (among other things) related to the sample size and to the data variability, so that L maximums should not be directly compared for datasets of different sizes, or with different standard deviations.
	Overall plot which helps to judge which models fit the data well, if judged by the transformed correlation coefficient from the P-P plot, namely by $-\ln(1 - r_P)$. The transformation is used because the r_P values do not discriminate among various models very nicely, when judged visually. The higher $-\ln(1 - r_P)$ value, the better. Ordering of fitted distributional types according to the loglikelihood maximums and transformed correlation coefficients are generally not the same.

8. Transformation

Menu:	QC.Expert	Transformation

This module is helpful when dealing with skewed (asymmetrically distributed) data, for which a departure from normality was detected, using *the Basic data analysis* module. The *Transformation* module uses two strategies to find the best transformation. Mean, confidence interval for the mean and the table of selected quantiles are produced for the transformed data. Computed characteristics are backtransformed to the original scale. The *Exponential* transformation is based on skewness minimization, while the *Box-Cox* transformation is based on the maximum likelihood method. The transformation approach takes into account asymmetry of the data distribution and often yields better mean and quantile estimates. It is intended for systematically skewed data, but not for situations where skewness is caused by several outliers, see 5.1.3. The transformation technique is used also for asymmetric control chart limits construction, see 5.9.



Data Transformation

Task name: Imission

Transformation

☑ Exponential Parameter 0.580078125

☑ Box-Cox Parameter -1.24182281494

Data: ● All ○ Marked ○ Unmarked

Select columns: Cd, Ca, K, Zn (Zn selected)

Fig. 22 Transformation dialog panel

8.1. Data and parameters

Data are in columns. Columns can be selected in the *Columns* field of the *Transformation* dialog panel, see Figure 22. When the transformation parameter *r* is known, it can be entered in the *Parameter* window. The optimal *r* value is computed upon pressing the "?" button. Computations are started by the OK button. When several columns are selected, computations are done as if all data came from a single column.

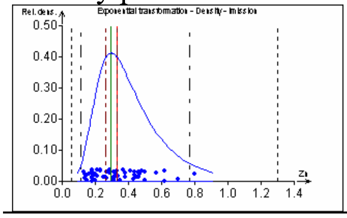
8.2. Protocol

Optimized parameter	The best transformation parameter *r* value, found by likelihood maximization.
Lower and upper confidence limit	Confidence interval for the transformation parameter *r*. The interval becomes narrower as the sample size increases. When the interval contains 1, it is not advisable to transform the data. This situation occurs either when the data are already normal or when the sample size is not large enough to rule out the possibility that they are. If the interval contains 0, the data might be considered lognormal.
Likelihood for r=1	Logarithm of the likelihood function evaluated at r=1 (corresponds to no transformation).
Likelihood maximum	Maximum of the log likelihood function , i.e. the log likelihood evaluated at the optimal *r* value.
Conclusion	Recommendation about transformation. NO means that the transformation does not lead to a substantial improvement. YES means that the transformation is recommended. It is recommended when the *r* parameter is significantly different from 1at 95% significance level.
Significance	Result of the *r*=1 hypothesis test. When the result is significant at 95% level, transformation is recommended, see the previous item.
Inputted parameter	Value of the transformation parameter, entered by the user in the Transformation dialog panel, see Figure 22. Other than optimal value can be entered.
Likelihood	Log likelihood evaluated at the selected parameter value.
Corrected mean	Arithmetic average, computed after the optimal transformation was selected by the Box-Cox transformation approach. It might be a better estimate of the mean than the

	simple arithmetic average, computed in the Basic data analysis module when the data come from a highly skewed distribution.
LCL	Suggested value for the Shewhart X-bar control chart lower limit. The number of data columns corresponds to the subgroup size, see 6.4.
UCL	Suggested value for the Shewhart X-bar control chart lower limit. The number of data columns corresponds to the subgroup size, see 6.4.
LWL	Suggested value for the lower warning limit.
UWL	Suggested value for the upper warning limit.
Corrected percentiles	Percentiles estimates which take asymmetry into account. In case of numerical problems, some of the values are not reported.
p	Percentile specification.
Lower, upper	Percentiles, corresponding to p% (lower) or to (100-p)% (upper)

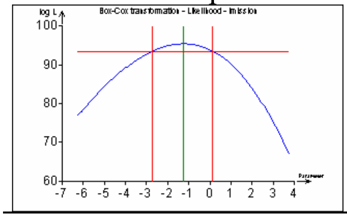
8.3. Graphical output

Density plot



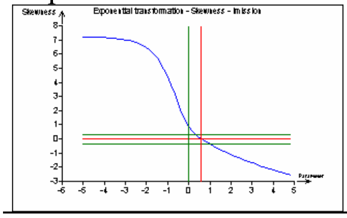
Density plot (see 5.1.3.) shows the data distribution shape as well as corrected mean (green vertical line), confidence interval for mean value (red line), ±2σ and ±3σ lines (capturing 99.73% of data).

Likelihood for the Box-Cox transformation parameter



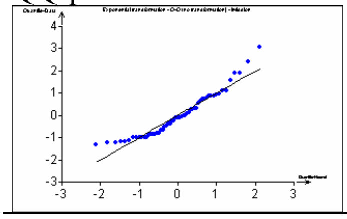
Log likelihood is plotted on y-axis, while the parameter *r* is plotted on x-axis. Maximum is achieved at the optimal *r* value. Horizontal line gives 95% confidence limit for the log likelihood maximum. Vertical lines give the 95% confidence interval for the parameter *r*. If the confidence interval contains 1, no transformation is necessary and the estimates from the Basic data analysis module can be used, or the Transformation module computations can be done with *r*=1.

Skewness plot for the Exponential transformation



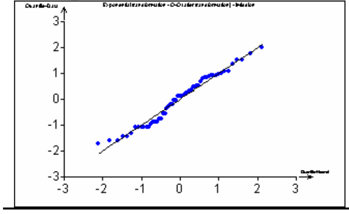
Skewness of the data plotted as a function of the transformation parameter *r*. Zero skewness corresponds to the optimal *r* value. Meaning of this plot is similar to the previous plot meaning. It helps to find the optimum transformation parameter value and to construct its confidence interval. When the vertical green line crosses the curve outside the confidence interval for skewness, (marked by the horizontal line), a transformation is recommended.

QQ-plot before transformation



The same QQ-plot as in the Basic data analysis module. Transformation is a useful remedy mainly for data showing systematically bended shape of the QQ-plot, see the left panel. The plot is useful when checking whether any nonlinearity (i.e. non-normality) detected is caused by just a few points or by a general tendency shared by all data.

QQ-plot after transformation



The plot should be more linear, when the transformation was successful. Statistics given in the protocol should be checked to judge the transformation more properly.

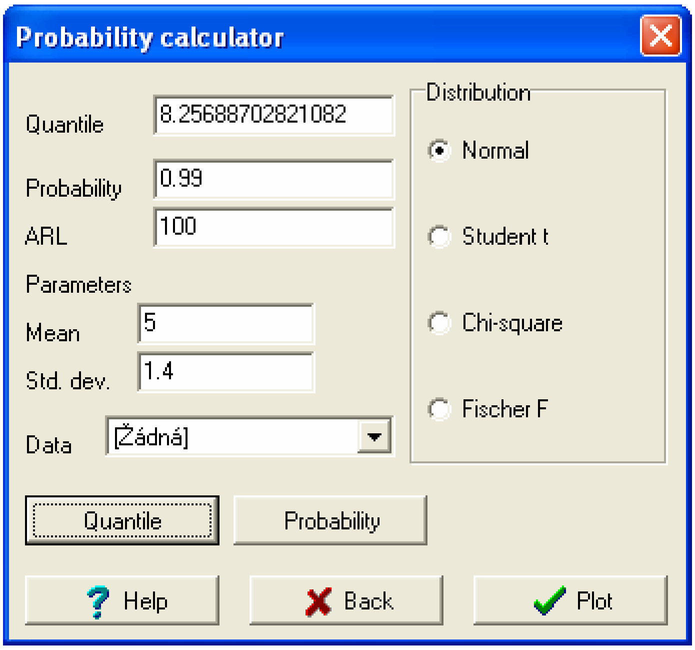
9. Probability calculator

Menu:	QCExpert	Probability calculator

This module is helpful for a quick calculation of quantiles and distribution function values for four distributions, which are often used for construction of critical values of statistical tests. Namely, they are: normal, Student-t, chi-square, Fisher-F. Various requests are entered in a dialogue panel. If needed, distribution function plot can be generated.

9.1. Data and parameters

Probability calculator has no mandatory data entry. Hence, it can be invoked even with an empty data spreadsheet. Upon invoking the dialog panel, the distribution, we want to work with, is specified. Next, its parameters have to be specified. For the normal distribution, mean and standard deviation are entered. For the Student-t, chi-squared or Fisher-F distribution, degrees of freedom ν are entered. While the t and chi-squared distributions have only one degree of freedom parameter, ν_1 (denoted as N1) the F distribution has two: ν_1 and ν_2 (denoted as N1 and N2). After entering a value to the *Quantile* field, and pressing the *Probability* button, requested probability (distribution function value) appears in the *Probability* field. On the other hand, when a value is entered in the *Probability* filed, pressing the *Quantile* button invokes calculation of the requested quantile, which appears in the *Quantile* field. When the Plot button is pressed, distribution function (for a selected distribution type, with previously entered parameters) is plotted. When working with a normal distribution, data points from the *Data* spreadsheet can be used to compute mean and standard deviation. The column containing data can be selected in the Data field of the Dialog panel. The Probability calculator can be used not only for a quick computation of probabilities or quantiles, but also for computing ARL. ARL is a function of a given probability P. When P<0.5, ARL is the reciprocal value of P, ARL=1/P, otherwise ARL=1/(1-P). ARL is the acronym for Average Run Length, that is the average number of data points counted between two consecutive occasions, when the p-th quantile is exceeded.



Probability calculator

Quantile 8.25688702821082

Probability 0.99

ARL 100

Parameters

Mean 5

Std. dev. 1.4

Data [Žádná]

Distribution

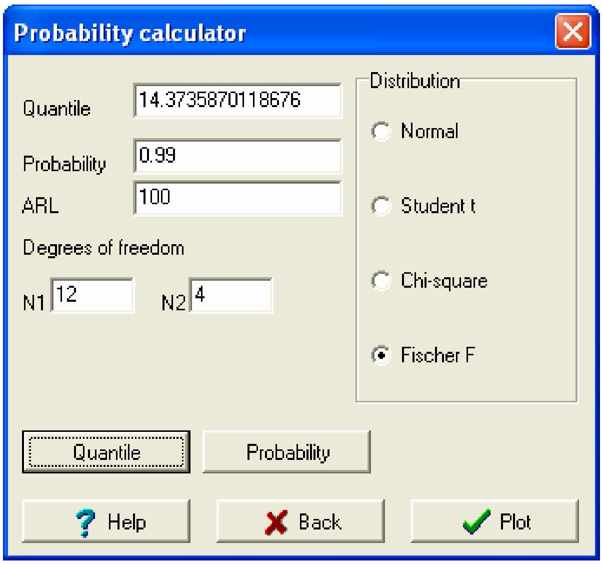
● Normal

○ Student t

○ Chi-square

○ Fischer F

Fig. 23 Dialog panel for Probability calculator (normal distribution example)



Probability calculator

Quantile 14.3735870118676

Probability 0.99

ARL 100

Degrees of freedom

N1 12 N2 4

Distribution

○ Normal

○ Student t

○ Chi-square

● Fischer F

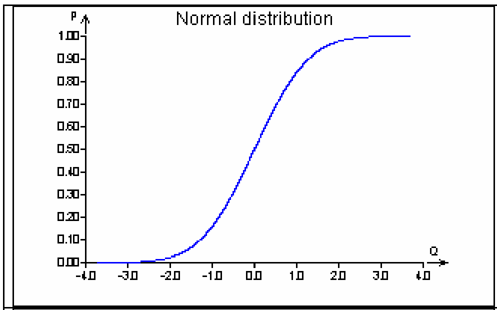
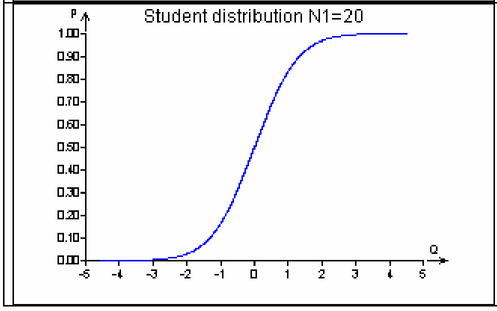
Fig.24 Dialog panel for Probability calculator (Fisher-F distribution example)

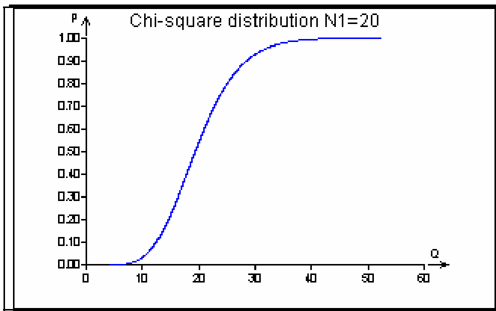
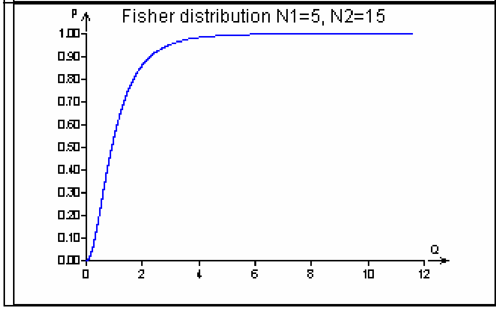
9.2. Protocol

This module produces no Protocol.

9.3. Graphical output

Distribution function plot is produced upon pressing the Plot button (after a distribution type was selected and its parameters were entered previously). Approximate distribution function values can be red from the plot when it is zoomed and plotted with the rectangular mesh. Quantile is always plotted on the horizontal axis, while the distribution function value (probability, that the random variable of interest will be smaller than or equal to the number equal to the horizontal coordinate) is plotted on the vertical axis.

	Standard normal (i.e. $N(0,1)$)distribution function plot.
	Student-t distribution with n_1=6, distribution function.

	chi-square distribution with n_1=10, distribution function.
	Fisher-F distribution with n_1=10 and n_2=12, distribution function.

10. Testing

Menu:	QCExpert	Testing

The group *Testing* consists of three modules: *Power and Sample Size*, *Tests* and *Contingency Table*. These modules are described below.

10.1. Power and Sample size

Menu:	QCExpert	Testing	Power and Sample Size

Modules in the group *Power and Sample Size* compute power of a test, required sample size an minimal difference of parameters that can be detected by the test. The tests support normal and binomial distribution. Inputs are significance level (or type I error) α, type of the test (one-sided or two-sided and theoretical (expected, specified) distribution parameter value. This parameter is the mean value for normal distribution, or probability for binomial distribution. Further, it is necessary to specify two of the following three numbers: Sample size, expected sample statistic and the power of the test 1 – β (β is the type II error). This module does not use any data from the data editor.

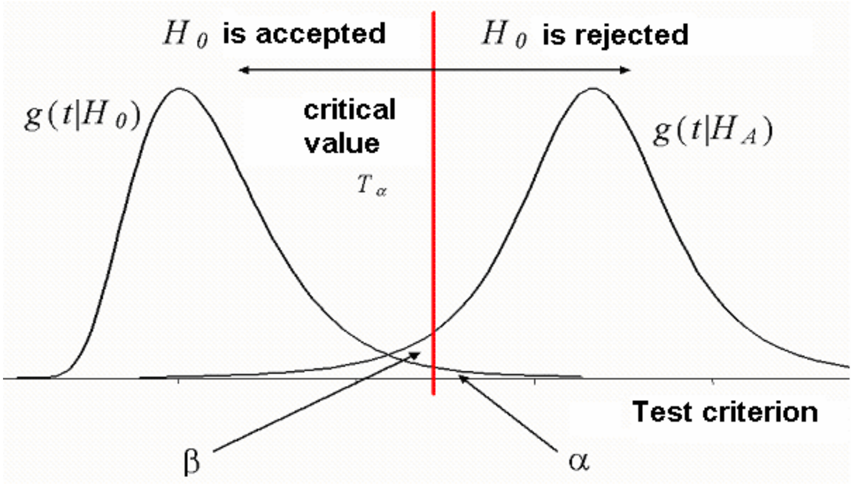


Fig. 25 Probabilities of type I (α) and type II (β) errors

Fig. 25 illustrates the types 1 and II type errors. H_0 denotes simple hypothesis to be tested, or zero hypothesis, for example equality of mean and some given number. H_A denotes alternative hypothesis, not(H_0). The result of the test is rejection or acceptance of H_0 based on comparing test criteroion T calculated from sample statistics with critical value for this test T_α.

Example: For testing equality of arithmetic average x and a given value μ we use the Student t-test, where the test criterion is $T = |x - \mu|/s$ and the critical value $T_\alpha = t_{1-\alpha/2}(N - 1)$ is 100(1–α)% quantile of the Student distribution. The two curves on the above figure illustrate the densities of distribution of the test criterion value for the cases when H_0 holds ($g(t|H_0)$) and when H_0 does not hold ($g(t|H_A)$). Two types of a mistake may happen:

The type I error, when we mistakenly reject H_0, despite the fact that H_0 is true. This will happen, when we happen to select the data from population (or measure items from a box) that all have untypically high or low value compared to other data. This will lead to too high value of T, which consenquently, compared with T_α yields refusing H_0. This situation is called the type I error ane is illustrated on . Its probability is α and can be specified by user. Usualy, we set α = 0.05, or 5%.

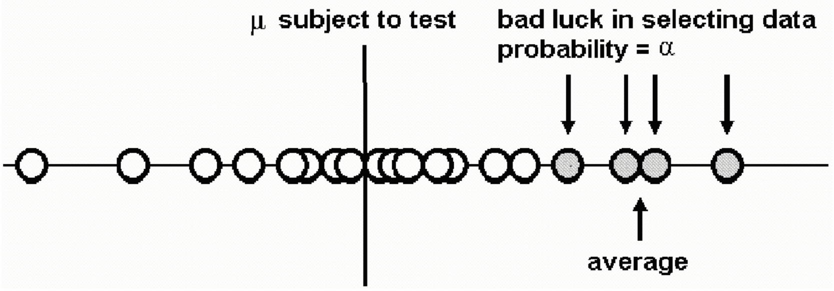


Fig. 26 Type I error, H_0 is rejected based on 4 unlucky measurements, though in fact H_0 holds.

Similarly, we can think of type II error, when we accept H_0, though it does not hold, see Fig. 27. Probability (or risk) of this situation is β. Obviously, number of data N, α, β, and difference between real and estimated parameter Δx are interdependent. When we want for example to have low both α and β, we have to take more data. When there is big Δx, we need less data. When we have available only small data set and expect small Δx, we will obtain lower „reliability“ of the test in term of high α and β, etc. All methods of Power and Sample Size have both one-sided and two-sided option. One-sided option means, that we are testing only „bigger“ or only „less“, and we don’t take into account the other possibility. By two-sided test we do not distinguish between „bigger“ or „less“. One-sided option tests always $x > \mu$ in one-sample normal tests, or $x_2 > x_1$ in two-sample normal and $P_A > P_0$ in one-sample binomial proportion tests or $P_2 > P_1$ in two-sample binomial proportion tests.

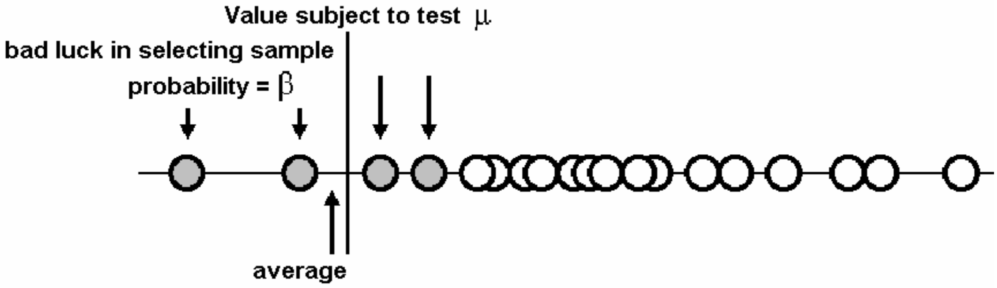


Fig. 27 The type II error, H_0 is accepted based of 4 measurements, although the true mean is not equal to the value subject to test.

The module Power and Sample Size can answer three types of questions:

- What would be the least sample size to prove the given difference between a hypothesized statistic (sample average or proportion) and a given number (or between two statistics) at a given risks α, β;

- What is the least difference difference between a hypothesized statistic (sample average or proportion) and a given number (or between two statistics) that could be proved by the test at a given sample size (or sizes) and at a given risks α, β;
- What would be the power 1 – β of a test that will prove a given difference between hypothesized statistic (sample average or proportion) and a given number (or between two statistics) at a given sample size (or sizes) and at a given risk of the type I error α.

10.1.1. Normal distribution, 1 sample

Menu:	QCExpert	Testing	Power and Sample Size	Normal distribution 1 sample

This module calculates parameters for testing arithmetic average of one normally distributed sample.

10.1.1.1. Parameters

In the dialog window (Fig. 28) specify the significance level (here also called type I error probability), the given number to be compared with average and the expected standard deviation σ. Select one- or two-sided option. Then you must specify two of the three fields: *Sample size*, *Sample average*, *Power*. At the field you want to calculate, click the coresponding button. The last calculated value will be marked in red. After calculation, the dialog window will not close, nor there is any output to the protocol. You may specify next parameters and make another calculations. Clicking *Output to protocol* will produce the output to the protocol window, the dialog window *Power and Sample Size* is closed by clicking on *Close*. The *Close* button itself does not produce an output to protocol.

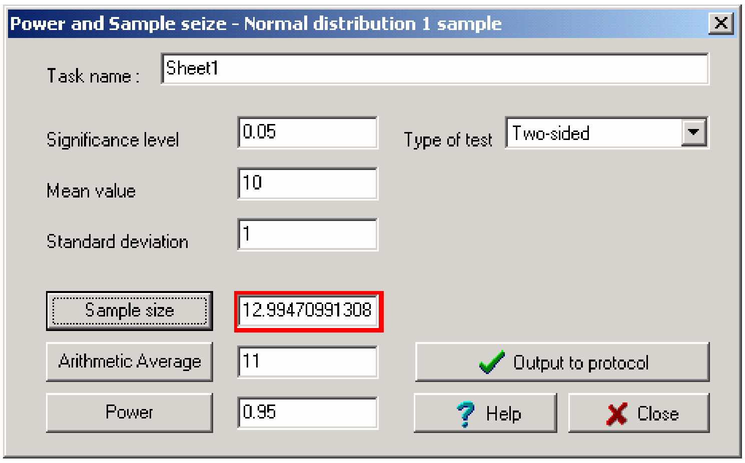
The power of the test may be calculated from *N*, μ, *X*, σ, α according to

$$1-\beta=\Phi\left(\frac{\sqrt{N}\,(\mu-X)}{\sigma}-Z_{1-\alpha/2}\right)+\Phi\left(\frac{\sqrt{N}\,(X-\mu)}{\sigma}-Z_{1-\alpha/2}\right)$$

The minimal sample size is given by

$$N=\left\{\left(\sigma\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)\right)\Big/\left|\mu-X\right|\right\}^{2},$$

where Z_α is the α-quantile of normal distribution and Φ is the distribution function (or cumulative density) of normal distribution. The unknown *X* (sample average) is calculated iteratively.



Power and Sample seize - Normal distribution 1 sample

Task name : Sheet1

Significance level 0.05

Type of test Two-sided

Mean value 10

Standard deviation 1

Sample size 12.99470991308

Arithmetic Average 11

Power 0.95

Fig. 28 Dialog window for Power and Sample size, normal distribution, 1 sample

10.1.1.2. Protocol

Power and sample size Normal dist., 1 sample	
Computation of sample size	Specifies which of the parameters was calculated
Task name	Task name from the dialog window
Significance level	Significance level α
Mean value M	Specified constant μ
Expected mean X	Specified or calculated arithmetic average.
Type of test	Specified mode: one-sided of two-sided
Null hypothesis H_0	X = M
Alternative hypothesis H_A	X <> M in case of two-sided test, X > M in case of one-sided test. It is always assumed that $x > \mu$.
Standard deviation	Specified assumed standard deviation of data
Sample size	Specified or calculated sample size, non-integer value must be always rounded to the higher integer value.
Rounded sample size	Calculated sample size rounded to the nearest higher integer.
Power of test	Specified or calculated power of the test.

10.1.2. Normal distribution, 2 samples

Menu:	QCExpert	Testing	Power and Sample Size	Normal distribution 2 samples

This module calculates parameters for testing arithmetic averages of two normally distributed samples.

10.1.2.1. Parameters

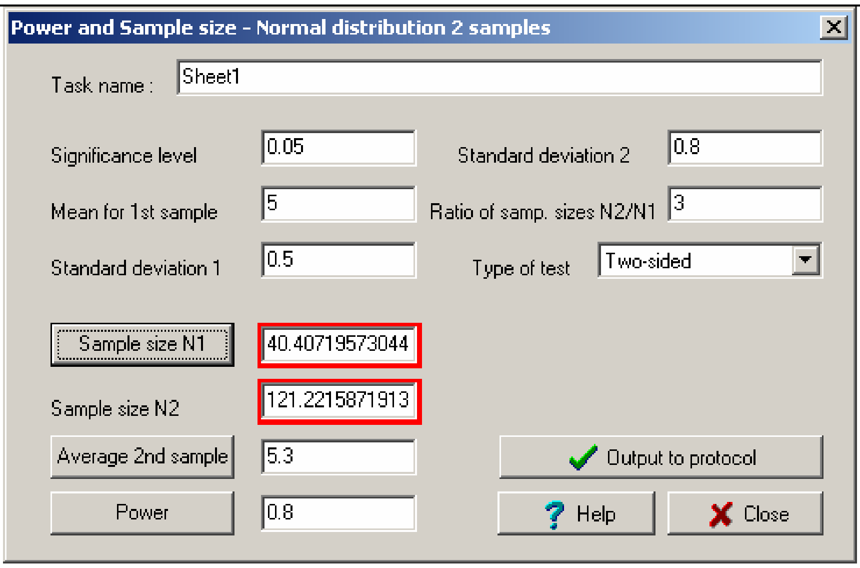
In the dialog window (Fig. 29) specify the significance level (here also called type I error probability), the average of the first sample and the expected standard deviations of the first and seconrd sample. If the Sample size is to be computed, you can also specify the ratio of the sample sizes of the first and second sample N_2/N_1. Select one- or two-sided option. Then you must specify two of the three fields: *Sample sizes* (two fields), *Second sample average*, *Power*. At the field you want to calculate, click the coresponding button. The last calculated value will be marked in red. After calculation, the dialog window will not close, nor there is any output to the protocol. You may specify next parameters and make another calculations. Clicking *Output to protocol* will produce the output to the protocol window, the dialog window *Power and Sample Size* is closed by clicking on *Close*. The *Close* button itself does not produce an output to protocol.

The minimal sample sizes N_1 and N_2 are given by

$$N_1 = \left(\sigma_1^2 + \frac{\sigma_2^2}{k}\right)\left\{\frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)}{\left|X_2 - X_1\right|}\right\}^2,$$

$$N_2 = kN_1$$

where k is ratio N_2/N_1, Z_α is α-quantile of normal distribution.



Power and Sample size - Normal distribution 2 samples

Task name : Sheet1

Significance level 0.05

Mean for 1st sample 5

Standard deviation 1 0.5

Standard deviation 2 0.8

Ratio of samp. sizes N2/N1 3

Type of test Two-sided

Sample size N1 40.40719573044

Sample size N2 121.2215871913

Average 2nd sample 5.3

Power 0.8

Fig. 29 Dialog window Power and Sample size – Normal distribution 2 samples

Note

Though the total number of measurements $N_1 + N_2$ is minimal when $N_1 = N_2$, or $N_2/N_1 = 1$, it may be profitable to force one of the two samples to be smaller, say N_1, by specifying $k > 1$, for example when the first sample is much more expensive or difficult to measure, even at a price of significant increase of N_2.

10.1.2.2. Protocol

Power and sample size Normal dist., 2 samples	
Computation of sample size	Specifies which of the parameters was calculated
Task name	Task name from the dialog window
Significance level	significance level α
Mean for 1st sample	Specified arithmetic average of the 1st sample
Mean for 2nd sample	Specified or calculated arithmetic average of the 2nd sample
Type of test	Specified mode: one-sided of two-sided
Null hypothesis H_0	$X_1 = X_2$
Alternative hypothesis H_A	$X_1 \neq X_2$ in case of two-sided test, $X_1 < X_2$ in case of one-sided test. It is always assumed that $X_1 < X_2$.
Standard deviation 1	Specified assumed standard deviation of the first sample
Standard deviation 2	Specified assumed standard deviation of the second sample
Ratio of samp. sizes N2/N1	Specified ratio of sizes of the second and first sample. This ratio is used only when calculating *Sample size*.
Sample size N1 Sample size N2	Specified or calculated sample sizes, non-integer values must be always rounded to the higher integer value.
Rounded sample sizes	Calculated sample sizes rounded to the nearest higher integer.
Power of test	Specified or calculated power of the test.

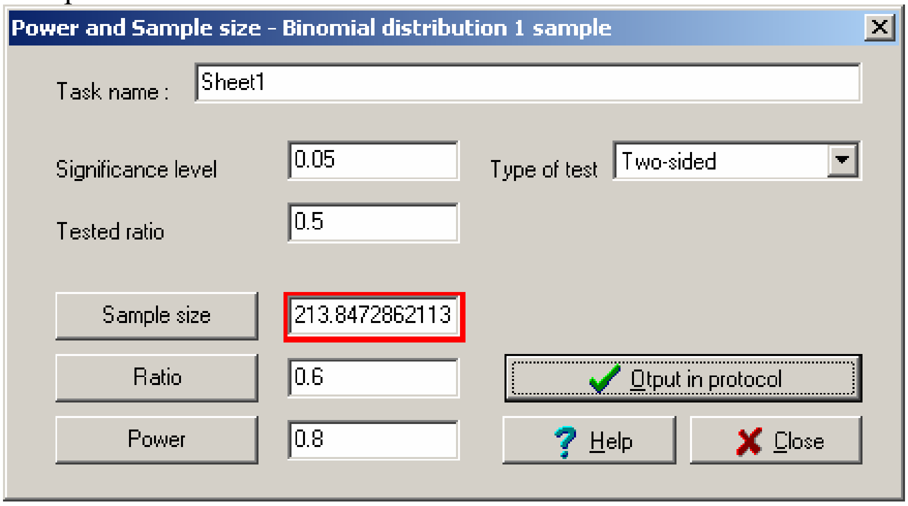
10.1.3. Binomial distribution, 1 sample

Menu:	QCExpert	Testing	Power and Sample Size	Binomial distribution 1 sample

This module calculates parameters for testing equality of probability of occurrence and a given constant (here called *Tested ratio*) based on the hypothesized ratio of observed trials and occurrencies (here called *Ratio*).

10.1.3.1. Parameters

In the dialog window (Fig. 30) specify the significance level (here also called probability of the type I error), the given number P_0 to be compared with hypothesized ratio and the expected ratio of successes and trials P_1. Select one- or two-sided option. Then you must specify two of the three fields: *Sample size N*, *Ratio*, *Power*. Then, at the field you want to calculate, click the coresponding button. The last calculated value will be marked in red. After calculation, the dialog window will not close, nor there is any output to the protocol. You may specify next parameters and make another calculations. Clicking *Output to protocol* will produce the output to the protocol window, the dialog window *Power and Sample Size* is closed by clicking on *Close*. The *Close* button itself does not produce an output to protocol.



Power and Sample size - Binomial distribution 1 sample

Task name : Sheet1

Significance level 0.05

Type of test Two-sided

Tested ratio 0.5

Sample size 213.8472862113

Ratio 0.6

Power 0.8

Fig. 30 Dialog window Power an sample size – Binomial distribution 1 sample

The sample size *N* is given by

$$N = \left\{ \frac{\sqrt{P_0\left(1-P_0\right)}Z_{1-\alpha/2} + \sqrt{P_0\left(1-P_0\right)}Z_{1-\beta}}{\left|P-P_0\right|} \right\}^2 + \frac{2}{\left|P-P_0\right|}$$

where Z_α is α-quantile of the normal distribution. Normal approximation is used, which is precise enough for $NP(1\text{-}P) > 5$. The second term is the correction for continuous approximation.

10.1.3.2. Protocol

Power and sample size Binomial dist., 1 sample	
Computation of sample size	Specifies which of the parameters was calculated
Task name	Task name from the dialog window
Significance level	Significance level α
Ratio to be tested, P_0	Specified value to be compared with *Ratio* P_A., $0<P_0<1$.
Expected ratio P_A	Specified constant value $0<P_A<1$.
Type of test	Specified mode: one-sided of two-sided
Null hypothesis H_0	$P_0 = P_A$
Alternative hypothesis H_A	$P_0 \neq P_A$ in case of two-sided test, $P_0 < P_A$ in case of one-sided test. It is always assumed that $P_0 < P_A$.
Sample size	Specified or calculated sample size, non-integer value must be always

	rounded to the higher integer value.
Rounded sample size	Calculated sample size rounded to the nearest higher integer.
Power of test	Specified or calculated power of the test.

10.1.4. Binomial distribution, 2 samples

Menu:	QCExpert	Testing	Power and Sample Size	Binomial distribution 2 samples

This module calculates parameters for testing equality of probability of occurrence in two experimens (here called *Tested ratio*) based on the hypothesized ratio of observed trials and occurrencies for both experiments (here called *Ratio*).

10.1.4.1. Parameters

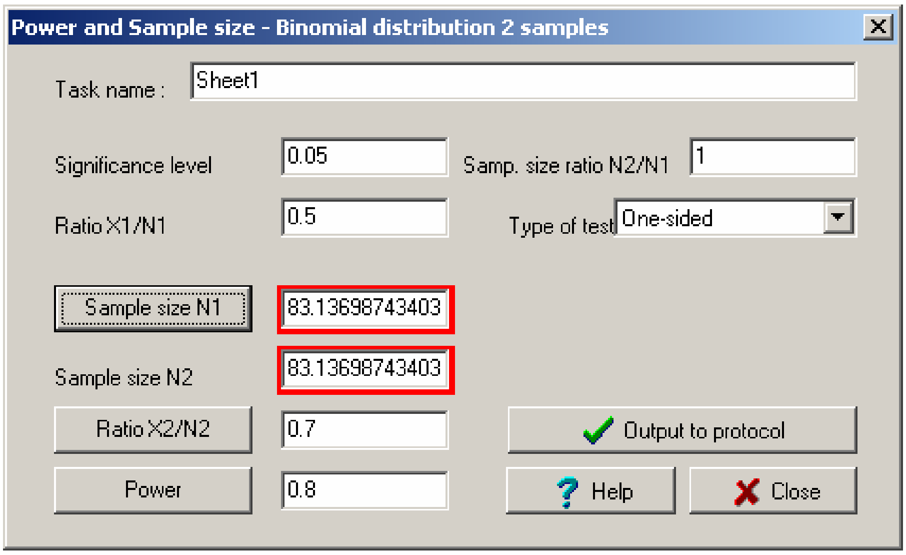
In the dialog window (Fig. 31) specify the significance level (here also called probability of the type I error), hypothesized ratios of successes and trials in the first and second experiment $P_1 = X_1/N_1$, $P_2 = X_2/N_2$. Select one- or two-sided option. Then you must specify two of the three fields: *Sample size N*, *Ratio X2/N2*, *Power*. Then, at the field you want to calculate, click the coresponding button. The last calculated value will be marked in red. After calculation, the dialog window will not close, nor there is any output to the protocol. You may specify next parameters and make another calculations. Clicking *Output to protocol* will produce the output to the protocol window, the dialog window *Power and Sample Size* is closed by clicking on *Close*. The *Close* button itself does not produce an output to protocol.

Sample sizes N_1 and N_2 are given by

$$N_1 = \left\{ \frac{\sqrt{P_1(1-P_1)+\frac{P_2(1-P_2)}{k}Z_{1-\beta}} + \sqrt{\bar{P}(1-\bar{P})+1\frac{1}{k}Z_{1-\alpha/2}}}{|P_2-P_1|} \right\}^2 + \frac{k+1}{k|P_2-P_1|},$$

where Z_α is α-quantile of the normal distribution. Normal approximation is used, which is precise enough for $NP(1\text{-}P) > 5$. The second term is the correction for continuous approximation.

$\bar{P} = \dfrac{P_1 + kP_2}{1+k}$; k is the user-specified specified ratio, $k = N_2/N_1$, so that $N_2 = k\,N_1$.



Power and Sample size - Binomial distribution 2 samples

Task name : Sheet1

Significance level 0.05

Samp. size ratio N2/N1 1

Ratio X1/N1 0.5

Type of test One-sided

Sample size N1 83.13698743403

Sample size N2 83.13698743403

Ratio X2/N2 0.7

Power 0.8

Fig. 31 Dialog window Power and Sample size – Binomial distribution 2 samples

10.1.4.2. Protocol

Power and sample size Binomial dist., 1 sample	
Computation of sample size	Specifies which of the parameters was calculated
Task name	Task name from the dialog window
Significance level	Significance level α
Expected ratio P_1	Specified value of the ratio of successes in the first experiment.
Expected ratio P_2	Specified value of the ratio of successes in the second experiment.
Type of test	Specified mode: one-sided of two-sided
Null hypothesis H_0	$P_1 = P_2$
Alternative hypothesis H_A	$P_1 \neq P_2$ in case of two-sided test, $P_1 < P_2$ P_A in case of one-sided test. It is always assumed that $P_1 < P_2$.
Ratio of sample sizes N2/N1	Specified ratio of sizes of the second and first sample. This ratio is used only when calculating *Sample size*.
Sample size N1 Sample size N2	Specified or calculated sample sizes, non-integer values must be always rounded to the higher integer value.
Rounded sample sizes	Calculated sample sizes rounded to the nearest higher integer.
Power of test	Specified or calculated power of the test.

10.2. Tests

The group *Tests* performs statistical testing for one-sample and two-sample binomial and normal distribution, for multinomial distribution and for contingency tables. Testing is based on actual experimental data, or on known statistics like average or standard deviation.

10.2.1. Binomial test, 1 sample

Menu:	QCExpert	Testing	Tests	Binomial test 1 sample

This module tests the hypothesis H_0, whether the observed number of occurrencies X in N tested trials is in accordance with a given constant probability P of occurrence of X in one (any) trial. Standard Chi-square test is employed. Assuming that the true (usualy unknown) probability P_A of occurrence is equal to the given P (the null hypothesis H_0), it would hold $P = X/N$ for $N\to\infty$. It is good to keep in mind, that not rejecting H_0 does not necessarily mean that H_0 is true. Often it only means that the number of trials is not sufficient to reject H_0.

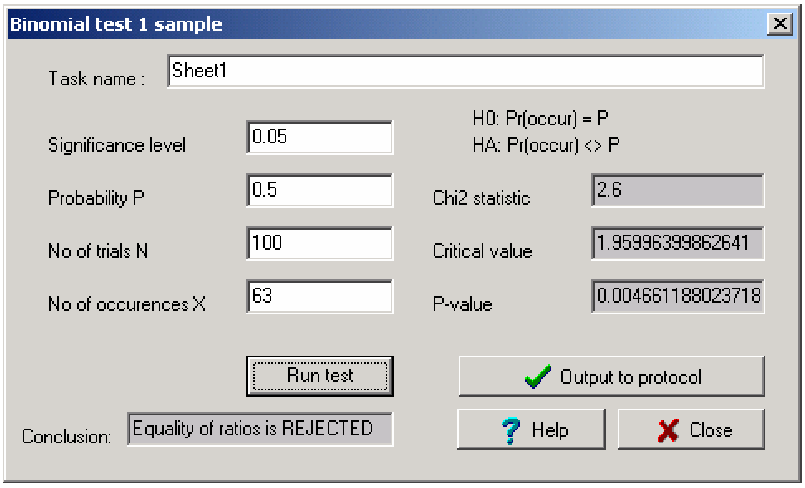
10.2.1.1. Parameters

This module does not use any data from the data spreadsheet. All information needed for the calculation are specified in the dialog window, see Fig. 32. You may modify the task name and the significance level (default value is α = 0.05). Than you must specify the given probability P to be tested, number of trials N and number of occurrencies X. After clicking the RunTest button the test is performed and results are displayed in the fields *Chi2 statistic*, *Critical value*, *p-value*. If the statistic is bigger than the critical value, the possible equality between hypothesized P and true unknown probability of success estimated by X/N is rejected. The field *Conclusion* will show verbal conclusion of the test: „*Equality of ratios is REJECTED or ACCEPTED*“. For very low number of occurrencies and/or low P, $XP < 5$, the test is less reliable.

Clicking *Output to protocol* will produce a record of the last performed test in the Protocol window, while the dialog window stil remains open. Clicking *Close* will close the dialog window. χ^2 or Chi-square test is used in this module. The χ^2 test statistic Z

$$Z = \frac{(X - NP)^2}{NP(1-P)}$$

has asymptotically distribution $\chi^2_{(1)}$. This statistic is compared to the quantile $U = \chi^2_{(1)}$ $(1-\alpha)$. If $Z > U$ then H_0 is rejected.



Binomial test 1 sample

Task name : Sheet1

Significance level 0.05

Probability P 0.5

No of trials N 100

No of occurences X 63

H0: Pr(occur) = P
HA: Pr(occur) <> P

Chi2 statistic 2.6

Critical value 1.95996399862641

P-value 0.004661188023718

Conclusion: Equality of ratios is REJECTED

Fig. 32 Dialog window for Binomial test, 1 sample

10.2.1.2. Protocol

Binomial test for equal ratio, 1 sample	
Task name	Task name from the dialog window
Overall sample size	Number of trials
Number of occurrencies	Number of occurrencies
Sample probability X/N	Calculated ratio *X*/*N*.
Probability to be tested	Given probability *P* to be tested
Significance level	Significance level α
Statistic Z	Calculated χ^2 statistic
Critical value U	Quantile of the χ^2 distribution
p-value	Biggest significance value at which H_0 would be rejected
Conclusion	Verbal test conclusion

10.2.2. Binomial test, N samples

Menu:	QCExpert	Testing	Tests	Binomial test N samples

This module generalizes the previous test. It tests simutaneously K hypotheses based on K binomial experiments, if observed numbers of occurrences X_i in N_i trials correspond to the hypothesized probabilities of this occurrencies P_i. Null hypothesis H_0 is defined as H_0: $P_i = P_{Ai}$ for $i = 1$, .. K. Chi-sqare test is employed again. Asuming that all true (unknown)probabilities P_{Ai} of occurrence of A are equal to P_i, than for $N_i \to \infty$ the probabilities P_i would be equal to X_i/N_i.

10.2.2.1. Data and parameters

This module expects data in two or three columns in the data editor. One column must contain numbers of trials for each experiment, second column must contain the numbers of successes, or occurrences of A_i, the third column may contain the probabilities P_i. The third column is not required, you may set all P_i (i = 1 .. K) to empirical average value $P_i = \Sigma X_i/\Sigma N_i$. An example of input data is in the following table for $K = 4$.

Trials	*Success*	*Probability*
200	22	0.1
200	46	0.25
100	56	0.5
250	103	0.4

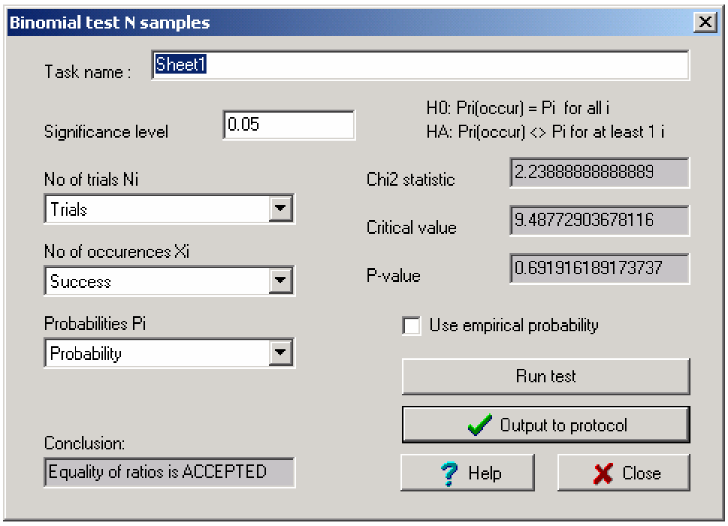
In the dialog window, see Fig. 33, you may modify the task name and the significance level (default value is α = 0.05). Then select the columns with N_i, X_i and P_i respectively. If the field *Use empirical probability* is checked, the program will use equal values $P_i = \Sigma X_i/\Sigma N_i$ and will ignore the third column, if any.

After clicking the RunTest button the test is performed and results are displayed in the fields *Chi2 statistic*, *Critical value*, *p-value*. If the statistic is bigger than the critical value, the possible equality between all hypothesized P_i and true probabilities of success estimated by X_i/N_i is rejected. The field *Conclusion* will show verbal conclusion of the test: „*Equality of ratios is REJECTED or ACCEPTED*“. For very low number of occurrencies and/or low P, $X_iP_i < 5$, the test is less reliable. Clicking *Output to protocol* will produce a record of the last performed test in the Protocol window, while the dialog window stil remains open. Clicking *Close* will close the dialog window.

Standard χ^2 test is used based on statstic C, which has asymptotic distriburion $\chi^2_{(K\text{-}1)}$.

$$C=\sum_{i=1}^{K}\frac{1}{P_i\left(1-P_i\right)}\left(X_i-N_iP_i\right)$$

This statistic is compared to the quantile $U = \chi^2_{(1)}$ (1–α). If $C > U$ then H_0 is rejected.



Binomial test N samples

Task name : Sheet1

Significance level 0.05

H0: Pri(occur) = Pi for all i
HA: Pri(occur) <> Pi for at least 1 i

No of trials Ni: Trials

Chi2 statistic 2.23888888888889

Critical value 9.48772903678116

No of occurences Xi: Success

P-value 0.691916189173737

Probabilities Pi: Probability

☐ Use empirical probability

Conclusion: Equality of ratios is ACCEPTED

Fig. 33 Dialog window for Binomial test - N samples

10.2.2.2. Protocol

Binomial test for equal ratio, N samples	
Task name	Task name from the dialog window
Number of samples K	Number of tests.

Sample sizes Ni	Numbers of trials in each test
Number of occurrences Xi	Number of occurrences in each test
Theoretical occurrences Ni*Pi	Theoretical numbers of occurrencies if H_0 were true
Actual ratios Xi/Ni	Observed number of occurrencies
Ratio to be tested Pi	Given probabilities to be tested
Hypothesis H0	All true probabilities are equal to P_i
Hypothesis HA	Alternative to HA
Significance level	Significance level α
Degrees of freedom	Degrees of freedom
Statistic Chi2	Test statistic
Critical value	Maximal acceptable value of test statistic when H_0 holds
p-value	Biggest significance value at which H_0 would be rejected
Conclusion	Verbal conclusion of the test

10.2.3. Multinomial test

Menu:	QCExpert	Testing	Tests	Multinomial test

This module tests probabilities of multinomial distribution. It is used when the trials may have more than two exclusive outputs (events) with probabilities P_{Ai}, $i = 1, .. , K$, $K > 2$ and P_A are the true unknown probabilities of occurrence of the event A_i. If we perform N trials, we recieve K frequencies, or numbers $X_1, X_2, ..., X_K$ of occurrencies of events $A_1, A_2, ..., A_K$, $\Sigma X_i = N$. Here we test H_0: $P_{Ai} = P_i$ for all $i = 1, .. , K$ based on the observed $P_{Ai} = X_i/N$, whereby $\Sigma P_i = 1$ and $\Sigma P_{Ai} = 1$. Standard Chi-squared test is used. Assuming that all true probabilities P_{Ai} of the occurrencies of A_i are equal to the given user-specified values P_i, the used statsitic C has the distribution $\chi^2_{(K-1)}$.

$$C = \sum_{i=1}^{K} \frac{\left(X_i - NP_i\right)^2}{NP_i}$$

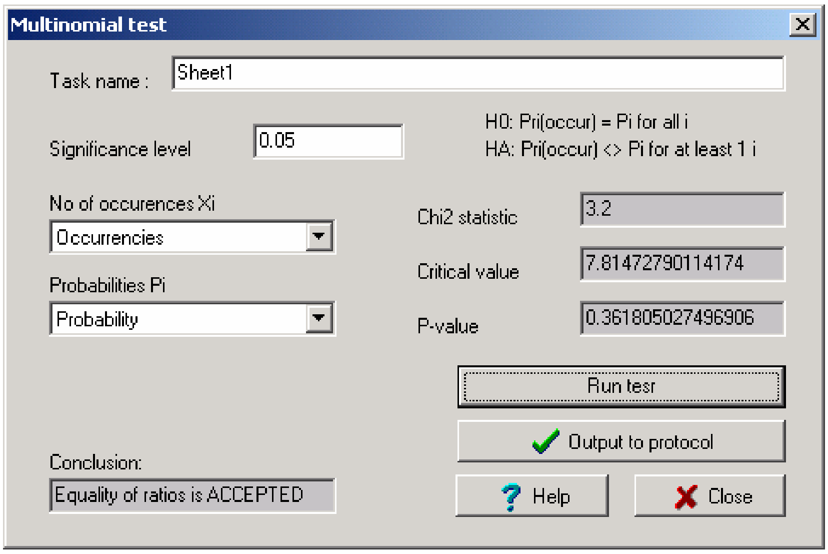
This statsitic is then compared with critical quantile $\chi^2_{(K-1)}$ (1–α). If C is bigger than critical quantile, we reject the H_0 hypothesis on the significance level α.

10.2.3.1. Data and parameters

This module expects data in two columns in the data editor. One column must contain numbers of occurrencies X_i of each event A_i. Second column must contain the expected probabilities P_i, or occurrences of A_i, the third column may contain the probabilities P_i. If the third column is missing, you may set all P_i ($i = 1 .. K$) to empirical average value $P_i = \Sigma X_i/\Sigma N_i$. An example of input data is in the following table for $K = 4$. Note that the probabilities must sum to unity, $\Sigma P_i = 1$.

Occurrencies	*Probability*
120	0.125
140	0.125
260	0.25
480	0.5

In the dialog window, see Fig. 34, you may modify the task name and the significance level (default value is α = 0.05). Then select the columns with numbers of occurrencies X_i and the probability values P_i respectively. After clicking the *Run Test* button the test is performed and results are displayed in the fields *Chi2 statistic*, *Critical value*, *p-value*. If the statistic is bigger than the critical value, the possible equality between all hypothesized P_i and true probabilities of the i^{th} event estimated by X_i/N is rejected. The field *Conclusion* will show verbal conclusion of the test: „*Equality of ratios is REJECTED or ACCEPTED*“. Clicking *Output to protocol* will produce a record of the last performed test in the Protocol window, while the dialog window stil remains open. Clicking *Close* will close the dialog window.



Multinomial test

Task name : Sheet1

Significance level 0.05

H0: Pri(occur) = Pi for all i
HA: Pri(occur) <> Pi for at least 1 i

No of occurences Xi: Occurrencies

Probabilities Pi: Probability

Chi2 statistic 3.2

Critical value 7.81472790114174

P-value 0.361805027496906

Conclusion: Equality of ratios is ACCEPTED

Fig. 34 Dialog window for multinomial test

Example

When throwing randomly a theoretical homogeneous matchbox with dimensions $a > b > c$ on horizontal plane in an homogeneous conservative gravitational field (e.g. on a table in a pub on the Earth), the matchbox may remain on either of the three sides: on the biggest one (side $A = ab$, event A) the smaller side, $B = ac$, event B, or on the smallest side, $C = bc$, event C. We want to carry out an experiment to support our hypothesis H_0 that the probabilities P_A, P_B, P_C of the positions A, B, C, see Fig. 35 are in the same ratio as the areas S_i of the sides divided by squared potential energy E_i of the respective position, $P_A : P_B : P_C = ab/c^2 : ac/b^2 : bc/a^2$, or $P_i \sim S_i/E_i^2$. The dimensions of the box are $a = 47$mm, $b = 35$mm and $c = 15$mm. Thus, the hypothesized probabilities would be $P_A = 0.89991$, $P_B = 0.07084$, $P_C = 0.02925$, since $P_A + P_B + P_C = 1$. In the experiment we received the position A in 1495 cases, position B in 115 cases and position C in 41 cases out of 1651 throws. We decided to carry out the test on the confidence level α = 0.05. The data table will have the following form:

Events	Probabilities
1495	0.8999082056
115	0.0708382553
41	0.0292535391

Open the window *Multinomial test*. Leave the *Significance level* at 0.05. Select the first and second column in *No. of occurrences* and *Probabilities* and click on *Run test*. The conclusion reads *Equality of ratios is ACCEPTED*, which means that the experiment does not contradict our theory. (On the other hand, of course, by no means this confirms it.)

Fig. 35 The three matchbox positions

10.2.3.2. Protocol

Multinomial test for equal ratio	The modul name
Task name	Task name from the dialog window
Number of classes K	Numbers of the classes *K*
Number of occurences	Number of occurrences of each event
Theoretical occurences N*Pi	Theoretical number of occurrences when H_0 holds
Actual ratios Ni/N	Observed ratios of the frequencies P_{Ri}
Ratio to be tested Pi	Given values P_i to be tested
Hypothesis H0	$P_{Ri} = P_i$
Hypothesis HA	$P_{Ri} <> P_i$
Significance level	Significance level α, usually α = 0.05
Degrees of freedom	Degrees of freedom η
Statistic Chi2	The χ^2 statistic calculated from data
Critical value	Theoretical quantile of the χ^2 distribution H_0
p-value	Calculated *p*-value of the test
Conclusion	Verbal conclusion of the test

10.2.4. Normal test, 1 sample

Menu:	QCExpert	Testing	Tests	Normal test 1 sample

This test is used to test equality between the mean value x_1 of normally distributed data and a given constant x_0. Null hypothesis is then H_0: $x_1 = x_0$ and alternative hypothesis H_A: $x_1 \neq x_0$ for two-sided test, or H_A: $x_1 > x_0$ for one-sided test.The test is based on the known arithmetic average x_1 and standard deviation *s*, that have been calculated from *n* measured samples.

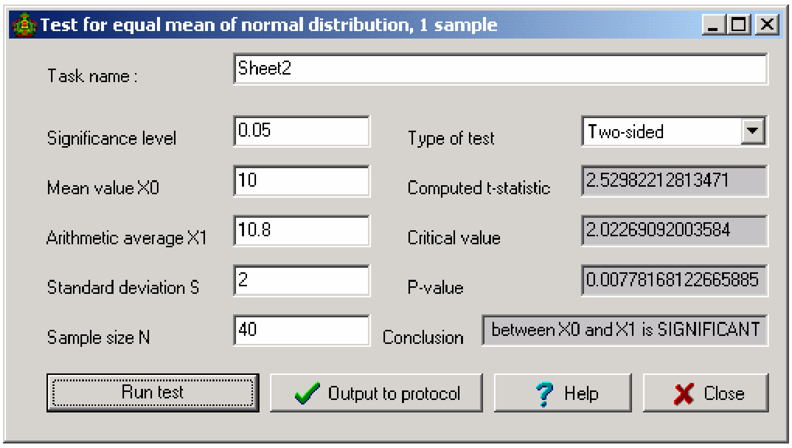
We use the one-sample *t*-test, where the t-statistic T_1 is compared with the critical value *T*:

$$T_1 = \frac{x_1 - x_0}{s}\sqrt{n}; \quad T = t_{n-1}\left(1-\alpha/2\right)$$

$t_n(\alpha)$ denotes α-quantile of the Student distribution with *n* degrees of freedom. H_0 is rejected, if $|T_1| > T$. In one-sided mode of the test the critical value $T = t_{n-1}(1-\alpha)$ is used.

10.2.4.1. Parameters

In the dialog window (Fig. 36) specify the significance level α, hypothesized mean value *X0*, measured average *X1*, standard deviation of the data *S* and sample size *N*. If only one-sided unequality is taken into account, one-sided alternative of the test should be selected in the field *Type of test*. After clicking *Run test* the values of *t*-statistic, critical value and *p*-value will be displayed in the respective fields. If the computed statistic is bigger than the critical value, then H_0 is rejected. The verbal conclusion has the form „*Difference between X0 and X1 is SIGNIFICANT*“ if H_0 is rejected or „*INSIGNIFICANT*“ if H_0 is accepted. Clicking *Output to protocol* will produce a record of the last performed test in the Protocol window, while the dialog window stil remains open. Clicking *Close* will close the dialog window.



Test for equal mean of normal distribution, 1 sample

Task name : Sheet2

Significance level 0.05

Type of test Two-sided

Mean value X0 10

Computed t-statistic 2.52982212813471

Arithmetic average X1 10.8

Critical value 2.02269092003584

Standard deviation S 2

P-value 0.00778168122665885

Sample size N 40

Conclusion between X0 and X1 is SIGNIFICANT

Fig. 36 Dialog window for normal test, 1 sample

10.2.4.2. Protocol

t-test one sample	Name of the module
Task name	Task name from the dialog window
Mean value X0	The given tested value
Sample average X1	Average of the data
Standard deviation S	Standard deviation of the data
Degrees of freedom	$n - 1$
Computed t-statistic	The value of the sample *t*-statistic T_1
Critical value T	Critical quantile $t_{(n-1)}(1-\alpha)$ of the *t*-distribution
p-value	Computed *p*-value
Conclusion	Verbal conclusion of the test

10.2.5. Normal test, 2 samples

Menu:	QCExpert	Testing	Tests	Normal test 2 samples

This test is used to test equality between two mean values of normally distributed data. Null hypothesis is H_0: $\mu_1 = \mu_2$ and alternative hypothesis H_A: $\mu_1 \neq \mu_2$ for two-sided test, or H_A: $\mu_1 > \mu_2$ for one-sided test. The test is based on the known sample arithmetic averages x_1, x_2 and standard deviations s_1 and s_2 of the samples, that have been calculated from n_1 and n_2 measurements of the first and second sample respectively. The test is based only on the 4 sample statistics x_1, x_2, s_1 and s_2, does not use original measurements and assumes normality and not too different variances of the data. If the original data are available, the module *Two samples comparison*, see chapter 6, page 6-44, is recomended to test the mean values.

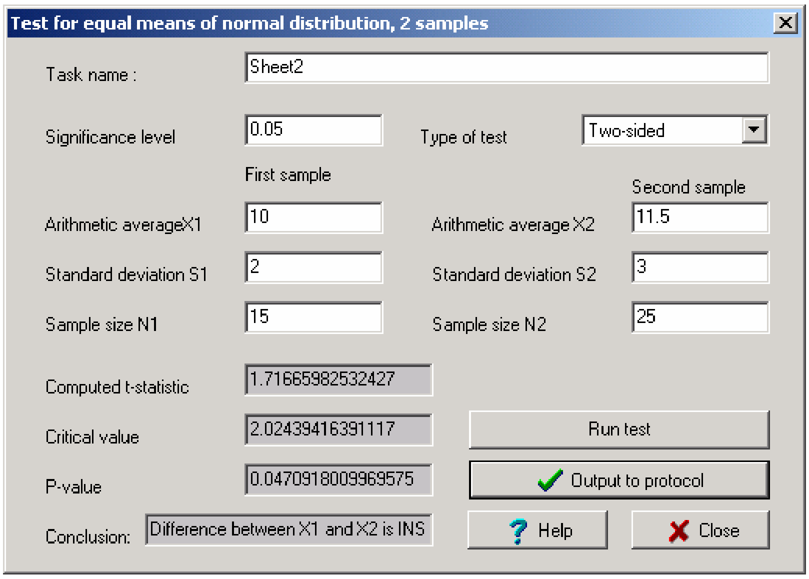
We use the two-sample *t*-test, where the *t*-statistic T_1 is compared with the critical value T:

$$T_1 = \frac{|x_2 - x_1|}{\sqrt{(n_1-1)s_1^2 + (n_2-1)s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}; \quad T = t_{n_1+n_2-2}(1-\alpha/2)$$

$t_n(\alpha)$ denotes α-quantile of the Student distribution with n degrees of freedom. H_0 is rejected, if $|T_1| > T$. In one-sided mode of the test the critical value $T = t_{n-1}(1-\alpha)$ is used.

10.2.5.1. Parameters

In the dialog window (Fig. 37) specify the significance level α, average of the first and second sample *X1*, *X2*, standard deviations of the samples *S1 and S2* and sample sizes *N1, N2*. If only one-sided unequality is taken into account, one-sided alternative of the test should be selected in the field *Type of test*. After clicking *Run test* the values of *t*-statistic, critical value and *p*-value will be displayed in the respective fields. If the computed statistic is bigger than the critical value, then H_0 is rejected. The verbal conclusion has the form „*Difference between X1 and X2 is SIGNIFICANT*" if H_0 is rejected or „*INSIGNIFICANT*" if H_0 is accepted. Clicking *Output to protocol* will produce a record of the last performed test in the Protocol window, while the dialog window stil remains open. Clicking *Close* will close the dialog window.



Test for equal means of normal distribution, 2 samples

Task name : Sheet2

Significance level 0.05 Type of test Two-sided

First sample Second sample

Arithmetic averageX1 10 Arithmetic average X2 11.5

Standard deviation S1 2 Standard deviation S2 3

Sample size N1 15 Sample size N2 25

Computed t-statistic 1.71665982532427

Critical value 2.02439416391117

P-value 0.0470918009969575

Conclusion: Difference between X1 and X2 is INS

Fig. 37 Dialog window for normal test, 2 samples

10.2.5.2. Protocol

t-test two sample	Module name
Task name	Task name from the dialog window
Sample average X1	Average of the first data sample
Standard deviation S1	Standard deviation of the first data sample
Sample average X2	Average of the second data sample
Standard deviation S2	Standard deviation of the second data sample
Degrees of freedom	$n_1 + n_2 - 2$.
Computed t-statistic	The value of the sample *t*-statistic T_1
Critical value T	Critical quantile of the *t*-distribution on the significance level α
p-value	Computed *p*-value
Conclusion	Verbal conclusion of the test

10.3. Contingency table

Menu:	QCExpert	Testing	Contingency tables

This module tests the hypothesis about independence of two categorical variables *A*, *B* based on experimentally observed occurence of combinations of particular levels of *A*, *B*. Number of levels of *A* is denoted *r*, number of levels of *B* is denoted *c*. The data are organized in the table of frequencies. Only the thick bordered box needs to be specified, the totals are calculated automatically.

	B levels			
A levels	B_1	…	B_c	Total
A_1	n_{11}	…	n_{1c}	$n_{1.}$
…	…	…	…	…
A_r	n_{r1}		n_{rc}	$n_{r.}$
Total	$n_{.1}$	…	$n_{.c}$	n

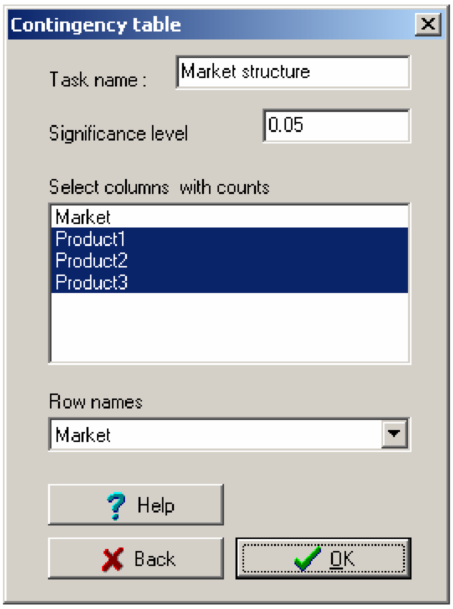
From this table we compute a test statistic *C*, which has a χ^2 distribition provided that *A* and *B* are independent.

$$C = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i.}n_{.j}} - n$$

The test statistic *C* is then compared to the critical quantile of the χ^2 distribution $\chi^2_{(r\text{-}1).(c\text{-}1)}(1-\alpha)$. If $C > \chi^2$, the hypothesis H_0 about independence of *A* and *B* is rejected. Dependence of *A* and *B* means that probabilities and numbers of occurrences in columns of the frequency table for at least one level of A are affected by levels of B. If A and B are independent, then the probability pij of observing the event $A_j \wedge B_i$ is equal of the product of marginal probabilities $p_{i.}$ $p_{.j}$. These probabilitie may be computed from the total frequencies, $p_{i.} = n_{i.}$/n, $p_{.j} = n_{.j}$ /n. The marginal probabilities sum to unity $\Sigma\, p_{i.} = \Sigma\, p_{.j} = 1$.

10.3.1. Data and parameters

The module expects data in form of frequencies shown in the thick-berdered part of the above table.The column names are taken from the header of the data table, row names may be in any column (preferably the first one) of the data table. In the dialog window, see Fig. 38, the columns of frequencies and column of row names. You may modify the task name and the significance level (usual value is α = 0.05). It is recommended that $n_{ij} \geq 5$, to ensure reliability of the test. After having selected data, run the test by clicking on the OK button. Results are written into the Protocol window and the doalog window will close.



Contingency table

Task name : Market structure

Significance level 0.05

Select columns with counts

- Market
- Product1
- Product2
- Product3

Row names Market

Fig. 38 Dialog panel for the contingency table

Example

The the following table we have volumes of products of three types 1, 2 3 sold to 4 different markets, Europe, Asia, USA, and Africa. We want to test if there is a difference in structure between products or between markets. Practically, we test if the proportion between products is the same for any market and simultaneously if the proportion between markets is the same for any product.

Market	Product1	Product2	Product3
Europe	114	25	301
Asia	68	20	225
USA	131	30	240
Africa	57	23	159

In the dialog window (Fig. 38) type the significance level (say, 0.05), select the columns with frequencies *Product1*, *Product2*, *Product3* and the column with market names, *Market*. Press *OK* to run the analysis.

Conclusion	Independence of variables is rejected
Significance level	0.05
Degrees of freedom	6
Chi2 statistic	17.11555756
Critical value	12.59158724
p-value	0.008867805134

Chi-squared statistic is 17.1, which is greater than the critical value 12.59, so the independence of the variables is rejected. That means, that the structure of the market is dependent on locality and product.

10.3.2. Protocol

Analysis of contingency table	Name of the module
Task name	Task name from the dialog window
Table of counts	
„row name“	Input frequencies (or counts) from data table. In the last column are the maginal frequencies (or sums of rows).
Theoretical	In brackets there are the theoretical frequencies based on assumption of independence.
Total	The column marginal frequencies, or sums of columns.
Table of ratios and probabilities	
„row name“	Empirical probabilities calculated from the given frequencies as $p_{ij} = n_{ij}/n$, in the last column are the marginal probabilities, or sums of rows.
Theoretical	Theoretical probabilities based on assumption of independence. calculated as $n_{i.}\ n_{.j}/n^2$, in the last column are theoretical row marginal probabilities.
Total	The column marginal frequencies, or sums of empirical probabilities in columns.

Conclusion	Verbal conclusion of the test
Significance level	Significance level of the test α ($1 - \alpha$ is called confidence). Recommended is $\alpha = 0.05$.
Degrees of freedom	$(r - 1)(c - 1)$.
Chi2 statistic	The value of the sample statistic C calculated from data
Critical value	Theoretical quantile of the χ^2 distribution $H0$ $\chi^2_{(r\text{-}1).(c\text{-}1)}(1{-}\alpha)$.
p-value	Calculated p-value of the test

11. Simulation

11.1. Simulation

Menu:	QC.Expert	Simulation	Data simulation

This module generates pseudorandom numbers with given statistical properties. Four distributions are available for random sample generation: normal, lognormal, uniform and lambda. Different parameters (mean, variance, skewness) determine the distribution within the four types. Both independent and correlated data (with a given first order autocorrelation) can be simulated. Generated data are useful when simulating real processes or when exploring properties of various statistical procedures before applying them on real data. Uniformly distributed simulated data can be used to construct table of random numbers for acceptance sampling.

11.1.1. Parameters

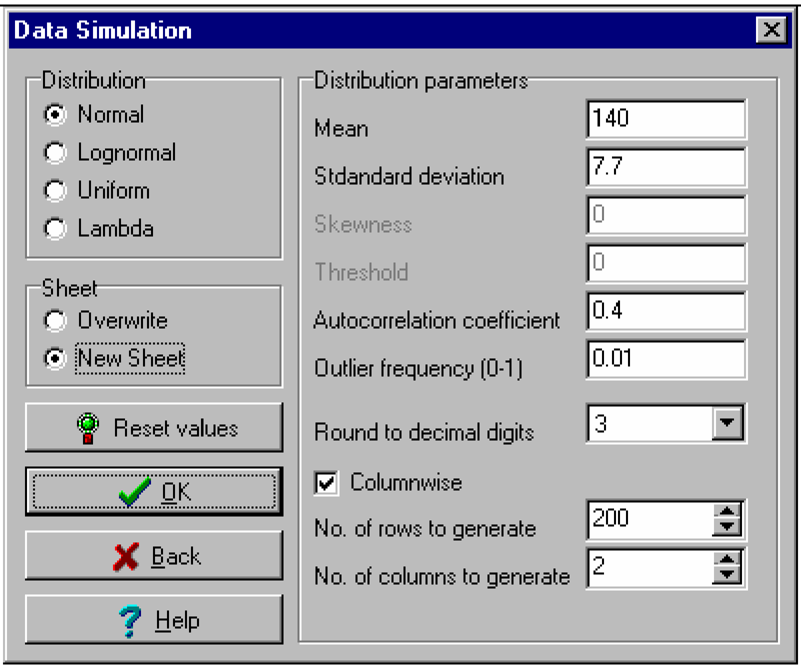
The procedure does not require any data. Parameters of various distributions are entered in the *Simulation* dialog panel, see Figure 24. The following table shows which parameters can be specified for each of the four distributions available:

Table1 Parameters to be selected

	Normal	Lognormal	Uniform	Lambda
Mean	✓	✓	✓	✓
Standard deviation	✓ positive	✓ positive	✓ positive	✓ positive
Skewness	×	×	✓	✓
Threshold	×	✓	×	×
Autocorrelation	✓-1 to 1	✓-1 to 1	✓-1 to 1	✓-1 to 1
Outliers relative frequency	✓0 to 1	✓0 to 1	✓0 to 1	✓0 to 1

× - parameter cannot be freely specified
✓- parameter can be specified, range of admissible values given

Generated data are (pseudo)random and so their sample characteristics follow the specified parameters only approximately. Nonzero skewness selected for the uniform distribution causes that data from a parallelogram shaped distribution are generated. The standard deviation parameter entered for the uniform distribution is interpreted as the range. Nonzero autocorrelation changes the resulting distributional parameters (parameter specification to the independent random variables generated before transformation to autocorrelated data), e.g. standard deviation is inflated.



Data Simulation

Distribution
● Normal
○ Lognormal
○ Uniform
○ Lambda

Sheet
○ Overwrite
● New Sheet

Distribution parameters
Mean 140
Stdandard deviation 7.7
Skewness 0
Threshold 0
Autocorrelation coefficient 0.4
Outlier frequency (0-1) 0.01
Round to decimal digits 3
☑ Columnwise
No. of rows to generate 200
No. of columns to generate 2

Fig. 39 Simulation dialog panel

11.2. Error propagation

Menu:	QC.Expert	Simulation	Error propagation

This module is useful when exploring statistical behavior of a variable which is not measured but computed from measured variables or from variables with a known statistical behavior. Using the module, range of the resulting variable, its mean and confidence interval can be estimated. The individual measured variables contributions to the resulting variable variability can be determined as well. The approach is sometimes called uncertainity evaluation or error propagation. Resulting variable properties are evaluated by the Monte Carlo simulation and by the second order Taylor expansion.

11.2.1. Data and parameters

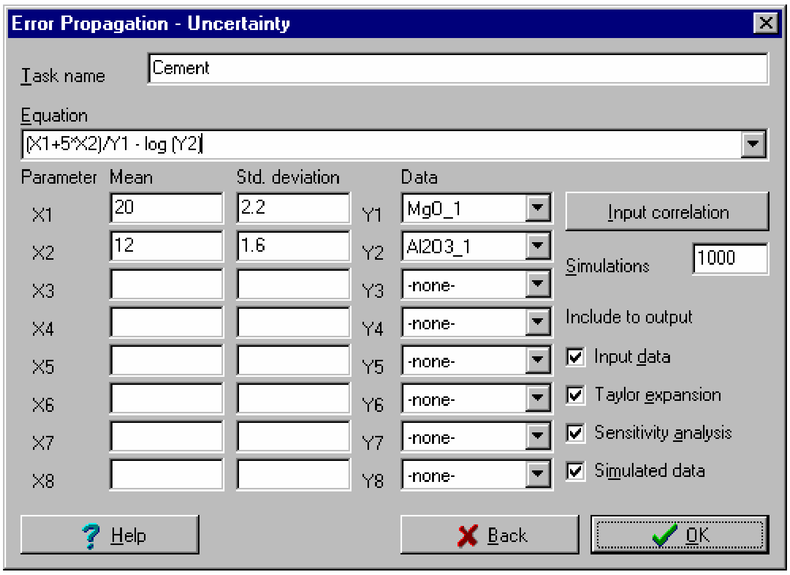
The input random variables and the function giving the resulting variable must be entered. The input variables can be given by:

a) specifying mean and standard deviation. The program generates normal random variable with the specified parameters. Such a generated random variable is denoted by X or x.

b) entering a sample of real data obtained e.g. by measurement. The program generates random sample according to the empirical distribution function computed from the data. The distribution might not be normal. This is a useful way of specifying the input distribution when a sufficient number of data points (say more than 20) is available and their distribution is not known. The inputted random variables are denoted by Y or y within the procedure. They are entered in columns, as usually.

Both specification types can be combined freely. The maximum number of X and Y variables is 20 (10+10). The Error propagation dialog panel is used to specify means and standard deviations (both parameters have to be specified) of x_i variables. y_i variables are defined by specifying a particular column from the current data sheet. The function of interest (in variables x_1, x_2, ..., y_1, y_2, ...) is entered in the usual syntax in the *Equation* window, for instance `5*x1*(1-x2)/log(3.124*y1)`. The pairwise correlation coefficients (x_i and x_j) or (x_i and y_j) variables can be entered, when they are known. The correlation coefficients values can influence mean, variance, quantile and confidence interval estimates heavily. The *Initialize* button sets all correlation coefficients to zero. When no correlations are specified, their values are set to zero. The correlation coefficients values have to lie within the (-1,1) interval.

For most situations, the number of simulations from 100 to 1000 should be satisfactory. The simulated data can be outputted to a separate sheet in the Protocol window upon request. These data can be used to analyze distribution of the resulting variable using the Basic data analysis module.



Error Propagation - Uncertainty

Task name: Cement

Equation: (X1+5*X2)/Y1 - log (Y2)

Parameter	Mean	Std. deviation		Data
X1	20	2.2	Y1	Mg0_1
X2	12	1.6	Y2	Al203_1
X3			Y3	-none-
X4			Y4	-none-
X5			Y5	-none-
X6			Y6	-none-
X7			Y7	-none-
X8			Y8	-none-

Simulations: 1000

Include to output

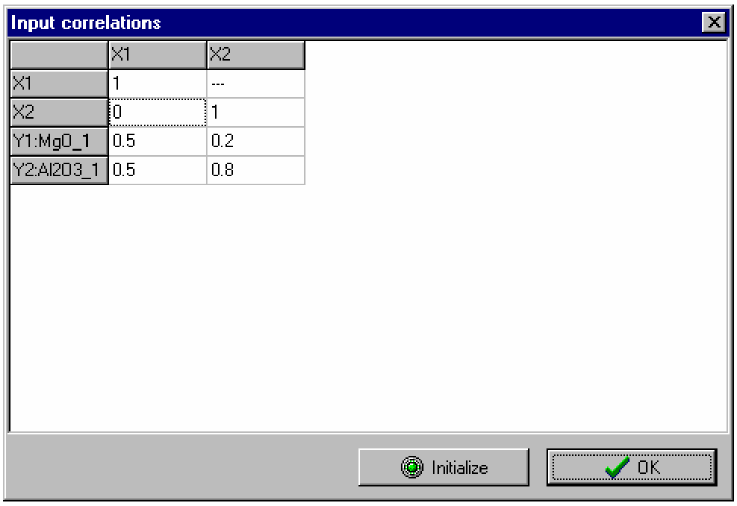
☑ Input data

☑ Taylor expansion

☑ Sensitivity analysis

☑ Simulated data

Fig. 40 Error propagation dialog panel



Input correlations

	X1	X2
X1	1	---
X2	0	1
Y1:Mg0_1	0.5	0.2
Y2:Al203_1	0.5	0.8

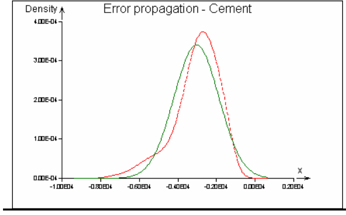
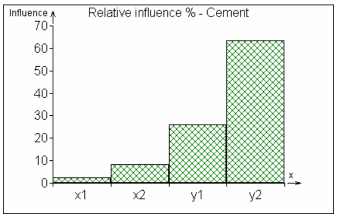
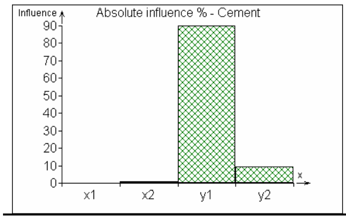
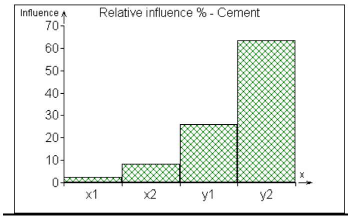
Fig. 41 Entering correlations between variables

11.2.2. Protocol

Function	Function of interest (assumed to possess continuous second order partial derivatives for the Taylor expansion).
Input variables	X variables distributions specified by their means and standard deviations.
Mean	Specified mean.
Standard deviation	Specified standard deviation.
95% interval	Confidence interval, computed via 2.5% and 97.5% percentiles.
+-3sigma	Interval containing 99.73% of data.
Inputted data	Y variables specifying simulation distribution through their empirical distribution function.
Mean	Arithmetic mean of the data as an estimate of the mean.
Standard deviation	Calculated standard deviation.

95% interval	Confidence interval computed via 2.5% and 97.5% data percentile.
+-3sigma	Interval containing 99.73% of data.
Resulting variable	Resulting variable characteristics obtained via Monte Carlo simulations.
Median	Median is often more reliable mean value estimate than arithmetic mean, when the data distribution is asymmetric or it has a high kurtosis.
Mean	Arithmetic mean as an estimate of the mean.
Standard deviation	Standard deviation estimate.
95% interval	2.5% and 97.5% percentiles of simulated data.
+-3sigma	Interval containing 99.73% of data.
Resulting variable range	Minimum and maximum of the resulting variable, generated by Monte Carlo simulation.
Sensitivity analysis	
Absolute sensitivity	Sensitivity of the resulting variable to the individual input variables. It is computed as partial derivative of the specified function with respect to a specified input variable, evaluated at the mean values of independent variables. Therefore, it is the sensitivity to a small change in the input variable. With respect to variability, it can be interpreted as the expected increase in resulting variable standard deviation when standard deviation of the input variable is increased by one unit.
Relative sensitivity	Individual input variable influence upon the resulting variable, which takes the amount of input variability into account. It is computed as the absolute sensitivity multiplied by the standard deviation of the particular input variable. When the relative sensitivity for a particular input variable is small relative to other variables, decreasing its variability or bringing it under statistical control is not very effective. If the relative sensitivity is large, decreasing the input variance or bringing the variable under statistical control can decrease resulting variability substantially.
Taylor expansion approximation	An alternative to the Monte Carlo simulations based on the second order Taylor expansion. Covariances (resp. correlations) among input variables are taken into account when computing the approximation.
Simple mean	Mean value estimate based on the first order expansion - evaluating the function at the mean of the input variables.
Corrected mean	Mean value estimate based on the second order expansion which does not take the covariances among the input variables into account.
Corrected mean (nonzero covariances)	Mean value estimate based on the second order expansion which does take the covariances among the input variables into account
Corrected standard deviation	Standard deviation estimate based on the second order expansion which does not take the covariances among the input variables into account.
Corrected standard deviation (nonzero covariances)	Standard deviation estimate based on the second order expansion which does take the covariances among the input variables into account.
95% interval	95% interval based on 2.5% and 97.5% percentiles. The computation assumes normality of the resulting variable!
Interval +-3sigma	An interval containing ca. 99.73% of data, based on 0.135% and 99.865% percentiles. The computation assumes normality of the resulting variable!
Simulated data	Simulated values of the resulting variable are saved to a separate sheet.

11.2.3. Graphical output

Probability density function plot  	Compares the simulated distribution (red) with the normal model (green). Substantial discrepancies suggest that the resulting variable is not normally distributed.
Relative influence in %  	Relative sensitivity in %, see Protocol.
Absolute influence  	Absolute sensitivity, see Protocol.
Relative influence in %  	Relative sensitivity in %, see Protocol..

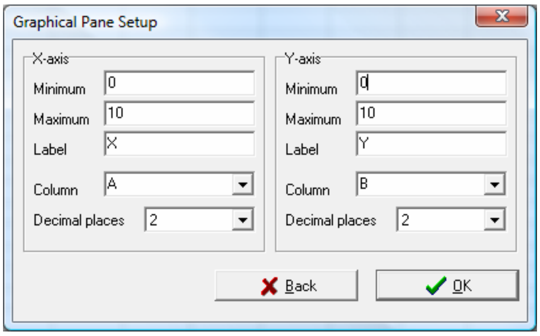
11.3. Graphical simulation

Menu:	QC.Expert	Simulation	Error propagation

This is a tool for a fast manual data simulation. Points are placed in accordance with a preconceived idea of how the data should look like using mouse. The points are then converted to numerical values and saved in two columns of the current data sheet. WARNING! When the current sheet contains other data, they might be overwritten when output columns are not selected carefully.

11.3.1. Parameters

When the Graphical simulation module is selected using the appropriate menu, the Graphical simulation dialog panel appears, see Fig. 42.

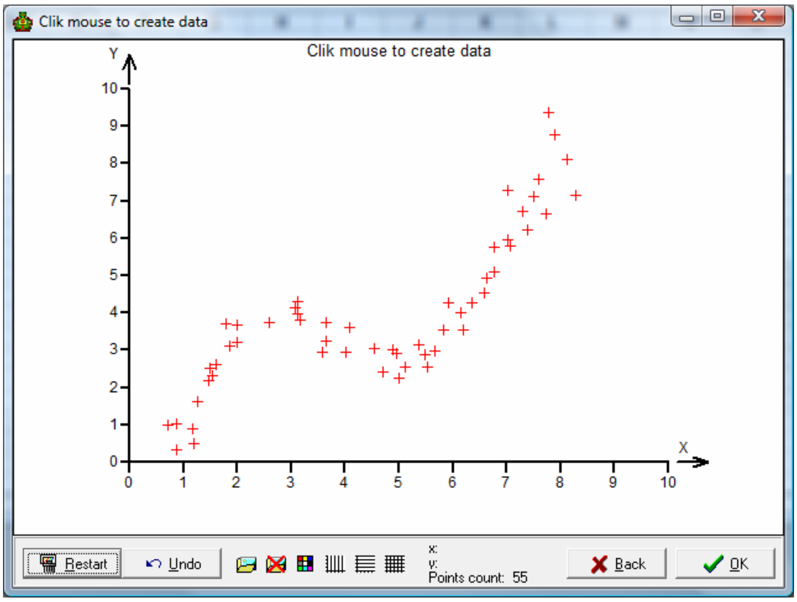


Graphical Pane Setup

X-axis
Minimum: 0
Maximum: 10
Label: X
Column: A
Decimal places: 2

Y-axis
Minimum: 0
Maximum: 10
Label: Y
Column: B
Decimal places: 2

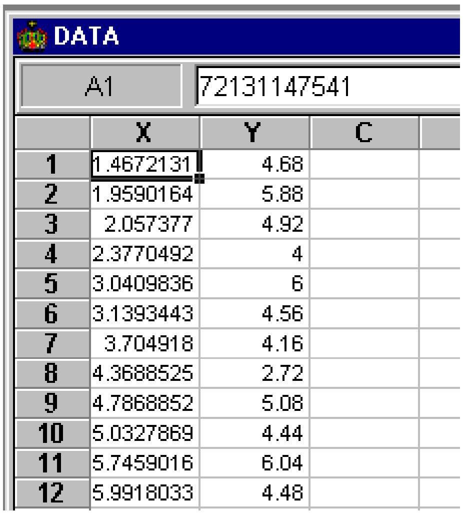
Fig. 42 Graphical simulation dialog panel



Points count: 55

Fig. 43 Graphical simulation

x- and y-axis minimum and maximum values can be selected in the panel. Column names for simulated data and columns to which the data will be outputted can be specified there as well. Upon clicking the *OK* button, an empty plot appears, in which points can be "generated" using mouse, see Fig. 43. The *Restart* button deletes all previously generated points, the *OK* button saves them in the current data sheet, (see Fig. 44) and closes the graphical window. It is recommended to start simulation with an empty data sheet.



A1 72131147541

	X	Y	C
1	1.4672131	4.68	
2	1.9590164	5.88	
3	2.057377	4.92	
4	2.3770492	4	
5	3.0409836	6	
6	3.1393443	4.56	
7	3.704918	4.16	
8	4.3688525	2.72	
9	4.7868852	5.08	
10	5.0327869	4.44	
11	5.7459016	6.04	
12	5.9918033	4.48	

Fig. 44 Simulated data from previous plot, outputted to a data sheet.

The graphical simmulation module can also be used to digitalize scanned images from literature, analog plotters etc. Select axes scales to the scanned plot, open image in jpg, gif, bmp or wmf format using the *Open file* button at the bottom of the graphical input window and use right mouse button to move and resize the image to fit in the axes, then use mouse to digitalize any point, curve or edge from the image, than press *OK* to transfer the coordinates of the points in the current data sheet. An example is given on Fig. 45 and Fig. 46.

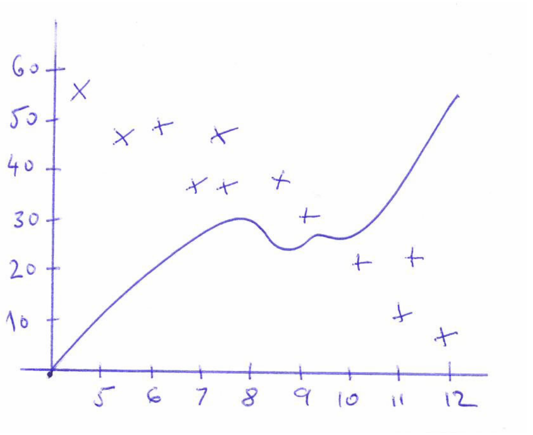


Fig. 45 Scanned jpg image to be digitalized

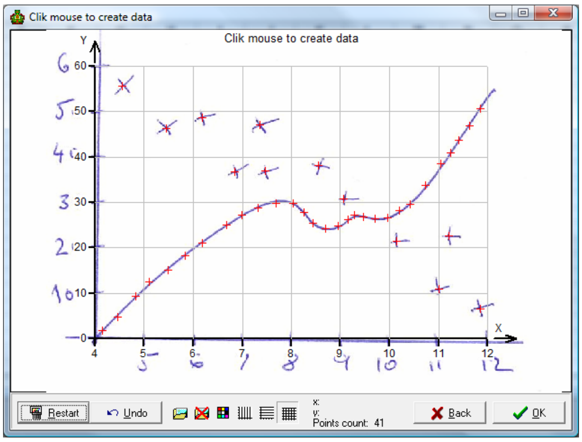


Fig. 46 Imported and digitalized image

12. ANOVA

12.1. Anova 1 factor

Menu:	QC.Expert	ANOVA

ANOVA is a common acronym for analysis of variance. The ANOVA module can be used for instance to check whether several batches of material or several material types differ with respect to a measured variable or characteristic. The groups to be compared are defined by *factor levels* (e.g. material can be a factor with levels corresponding to the individual material types). An interlaboratory comparison might serve as another example with individual laboratories corresponding to different levels of the “Laboratory” factor. Similarly, various purity levels or various production lines can be compared with respect to a response variable. The goal is to decide, whether response variable means differ among different levels of a factor. This is achieved by testing the hypothesis about equality of means among different factor levels. The ANOVA test assumes normality within groups defined by the factor levels, variance homogeneity across the groups and independence of all observations. The procedure should be applied to data free of outliers. ANOVA results are accompanied by the z-score plot. The plot shows estimated group means together with their error bars, obtained as ±3.SEM (standard error of the mean). The z-score is commonly used for interlaboratory comparisons or to compare potential suppliers.

12.1.1. Data and parameters

Scheffe’s pairwise comparisons can be requested by checking *Pairwise comparisons* within the Analysis type part of the ANOVA dialog panel. In case that the overall ANOVA test is significant and some group is found to be significantly different by the Scheffe’s test, an additional analysis is recommended. The ANOVA and Scheffe’s tests should be recomputed, omitting the observations corresponding to the differing group. It can happen that the variance is reduced and the second analysis finds a previously undetected difference. This procedure can be repeated until no significant differences are found. The z-score can be requested by checking the appropriate selection in the ANOVA dialog panel (see Figure 18). When no values are entered in the *Center* and *Standard deviation* fields, the z-score plot uses grand mean and residual standard deviation respectively. *Significance level* sets the significance level for all tests, 0.05 is a commonly used value. Data can be entered in two formats specified by either *Columnwise* or *By factor* button.

Columnwise format:
Data corresponding to different levels of a factor are entered in different columns. Number of data points in different columns/groups can be different. The minimum column number (i.e. the number of

factor levels) is 2. The minimum row number is 2. Column names should correspond to levels of the factor, e.g. Line A, Line B, Line C. Data columns can be selected in the *Columns* field of the ANOVA dialog panel, see Figure 18. The *Select all* button selects all columns of a given data sheet for the ANOVA analysis. All columns of the current data sheet are selected by default.

Data example

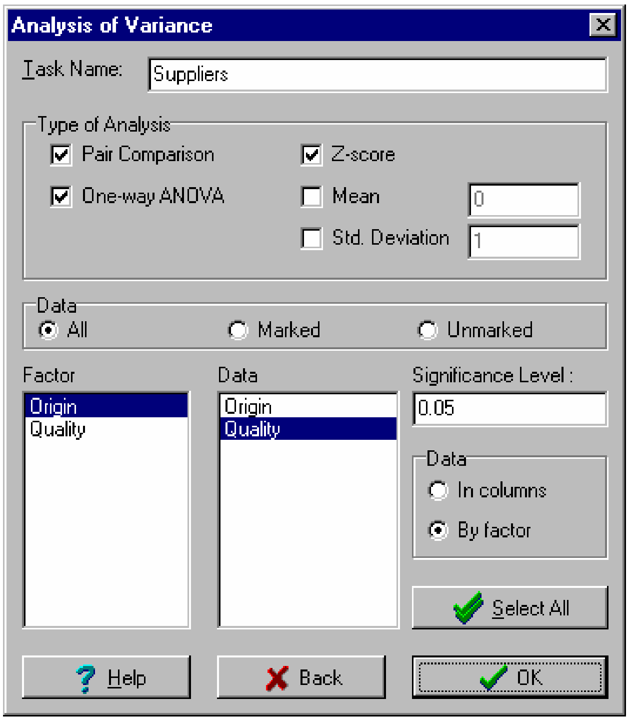
External	Lab 1	Lab 2	Lab 3
1.47	1.24	2.32	3.6
1.75	0.94	2.4	7.3
1.09	1.84	1.45	2.65
3.09	0.3	1.86	8.2

By factor arrangement*:*

Data are entered in two columns. The column chosen in the *Factor* field contains level codes. The column chosen in the Data column contains values of the response variable.

Data example:

Origin	Quality
UKR	17.17
GER	23.73
GER	23.7
ARG	24.78
BRA	27.91
SWE	23.19
BRA	26.87
ARG	24.59
UKR	19.5



Analysis of Variance

Task Name: Suppliers

Type of Analysis

☑ Pair Comparison ☑ Z-score

☑ One-way ANOVA □ Mean 0

□ Std. Deviation 1

Data: ● All ○ Marked ○ Unmarked

Factor: Origin (selected), Quality

Data: Origin, Quality (selected)

Significance Level : 0.05

Data: ○ In columns ● By factor

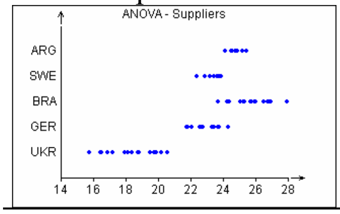
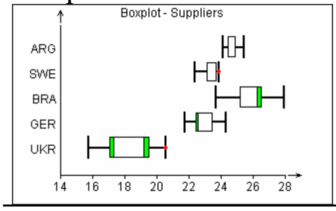
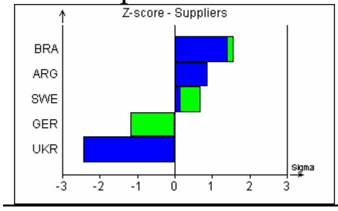
Fig. 47 ANOVA dialog panel

12.1.2. Protocol

Analysis of variance - ANOVA	
Number of levels	Number of factor levels.
Column	
Sample size	Number of valid data points within a factor level.
Level effects	Mean effect for each level of the factor. Difference between the group mean and the grand mean.
Means	Group means for each group defined by factor level.
Total mean	Overall mean, computed from all data.
Total variance	Variance computed from all data, regardless the grouping given by the factor levels. This is a legitimate variance estimate only if the factor has no effect and differences between levels are purely random.
Total mean square (corrected)	Mean of squared differences between observations and the grand mean.
Residual variance	Within group variance estimate. It is a legitimate variance estimate even if there are differences among response means for different factor levels.
Residual sum of squares	Sum of squared differences between observations and their respective group means.
Total sum of squares (corrected)	Sum of squared differences between observations and the grand mean..
Explained sum of squares	Difference between (corrected) total and residual sum of squares. It corresponds to the variability, explained by the differences among means for different factor levels.
Conclusion	Conclusion of the ANOVA test, describing in words whether the factor influences the response variable.
Factor	Indicates whether the factor has a significant effect.
Theoretical	Critical value, corresponding to the significance level chosen in the ANOVA dialog panel.
Calculated	Calculated value of the test criterion.
p-value	The smallest significance level for which the hypothesis of no factor effect is rejected using the observed data. When the p-value is smaller than a specified significance level, the factor is statistically significant.
Pairwise comparissons, Scheffe's method	All pairwise comparisons.
Pair	
Difference	Difference between group means, corresponding to different factor levels.
Conclusion	Conclusion of the test, indicating in words whether the factor is significant.
p-value	The smallest significance level for which the hypothesis of no difference among the means is rejected using the observed data. When the p-value is smaller than a specified significance level, the factor is statistically significant.
Z-score	Comparison of factor levels in terms of the standardized response group means.
Standardized value	Standardized value. A value smaller than –3 or larger than 3 might indicate that the level is different from other levels.
95% interval	Half of the 95% confidence interval length.

Difference	Limit of the 95% confidence interval.

12.1.3. Graphical output

ANOVA plot 	ANOVA plot shows individual data values versus factor levels, so that it is possible to check whether the groups defined by the factor levels differ in mean or variance. Observed values are plotted along x-axis, factor levels are plotted along y-axis. Plot inspection, together with the *Basic data analysis* module results can reveal outliers.
Boxplots 	The boxplot is described in 5.3, page 5-41. The boxplots are useful when comparing response among factor levels. Plot inspection, together with the *Basic data analysis* module results can reveal outliers.
Z-score plot 	Comparison of factor levels in terms of the standardized response group means. The reference values is 0. The vertical lines correspond to the ±3SEM (standard error of the mean) limits. Shorter bar corresponds to the standardized value. The groups having values smaller than –3 or larger than 3 are considered to be different from other groups. Longer bar corresponds to limit of the 95% confidence interval.

12.2. *ANOVA 2 factors*

Menu:	QCExpert	Anova-2 factors

Anova for 2 factors is an extension of the 1-factor anova described above. In 2-factor anova we analyze influence of two factors on a numeric response. Observed response Z_{ij} at n_X different levels of the factor X and at n_Y different levels of the factor Y may be described by Anova model for 2 factors X, Y :

$$Z_{ij} = Z_0 + \alpha_i X_i + \beta_j Y_j + \lambda_{ij} + \varepsilon_{ij}, \quad i = 1,..,n_X,\ j = 1,..,n_Y,$$

where Z_0 is absolute term (overall mean), α a β are contributions of the individual levels, elements λ_{ij} of a matrix Λ are called interactions and ε_{ij} is the random error with normal distribution and zero mean, $\varepsilon \sim N(0,\sigma^2)$. Further we define $\Sigma\alpha_i = 0$, $\Sigma\beta_j = 0$, $\Sigma\lambda_{i(j)} = 0$, $\Sigma\lambda_{j(i)} = 0$.

12.2.1. Data and parameters

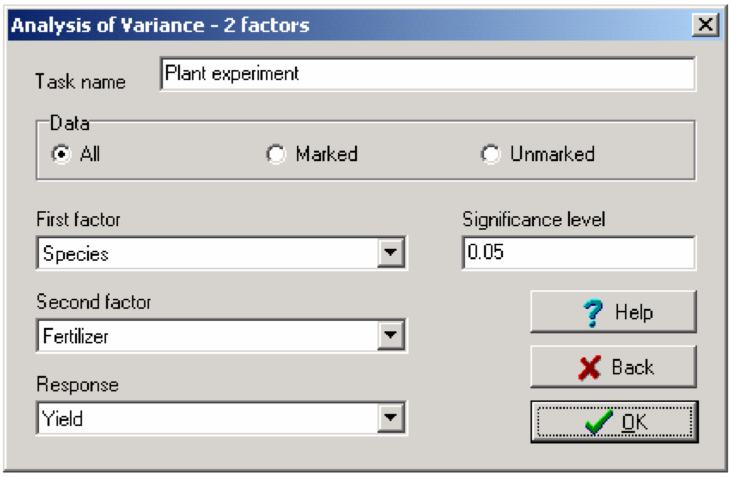
The module expects data in 3 columns. Two columns contain, levels of the two factors, in one column there are the corresponding observed response values. The combination of factors may be in arbitrary order. Factor levels are entered in form of any text string, such as RED, BLUE, GREEN, diferent string means different factor level. Both factors may have two or more levels. Each possible combination must be represented by al least one row.

The following table gives an example of data. First factor is the plant species with 3 levels: *Brazil*, *Longleaf*, *Cassablanca*, second factor is the fertilizer used, with 2 levels: *Nitrate* and *Phosphate*. The response is the observed increment in plant weight (*Yield*). There is 1 observation for each combination of levels, this experimental plan is calles plan without replications. If we had the same number $N > 1$ of observations for each combination, we would have a ballanced plan with N replications. If number of replications N_{ij} is not the same for all combinations, we have an unballanced experimental plan. Each different combination of factor levels is called *a cell*. We can distinguish 3 types of 2-factor Anova:

Species	Fertilizer	Yield
Brazil	Phosphate	14.6
Brazil	Nitrate	17.4
Longleaf	Phosphate	13.3
Longleaf	Nitrate	12.6
Cassablanca	Phosphate	17.5
Cassablanca	Nitrate	14.9

1 observations in each cell – *Ballanced ANOVA without replications*
Equal number n_0 > 1 of observations in each cell – *Ballanced ANOVA with* n_0 *replications*
Unequal number n_{ij} > 0 of observations in each cell – *Unballanced ANOVA*

In the dialog window (Fig. 48), select columns with both factors and the response. Clicking OK will run the analysis and results will be written in Protocol and Graphs windows.


Analysis of Variance - 2 factors

Task name: Plant experiment

Data: ● All ○ Marked ○ Unmarked

First factor: Species

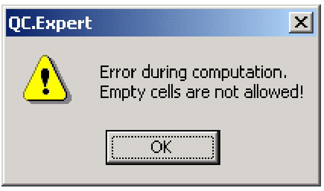
Significance level: 0.05

Second factor: Fertilizer

Response: Yield

Fig. 48 ANOVA 2 factors – Dialog window

From the data the module automatically recognizes which of the three Anova types should be used. If there is one or more cells (or combinations of factor levels) which has no observations, the message „Empty cells are not allowed“ appears.


QC.Expert

Error during computation.
Empty cells are not allowed!

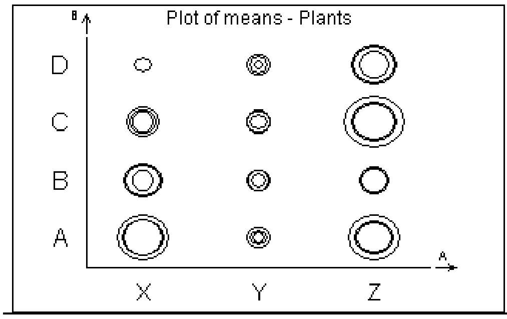
12.2.2. Protocol

Analysis of variance	Name of the module
Task name :	Task name from the dialog window
Type of model	Automatically recognized plan: - ballanced without replications, or - ballanced with replications, or - unballanced
Factors levels	Names of the factors and their levels

No of replications	In case of ballanced ANOVA: number of replications in cell n_0 In case of unballanced ANOVA: a table of numbers of replications in each cell, n_{ij}
Table of means	Table of arithmetical averages in each cell
Means for factor	Total averages for each level of one factor regardless the levels of the other
Overall mean	Total average of all responses. (This would be the estimate of mean response if no factor had any significant influence.)
Model parameters	Estimates of parameters α and β, the contributions of each level.
ANOVA Table	Summarized analysis of variance structure
Source of variability	Identification of the source (or cause) of variability
First factor	Variability caused by the first factor
Second factor	Variability caused by the second factor
Interaction	Variability caused by the interaction of factors
Residuals	Residual variability
Total	Total variability
Sum of squares	Sum of squares caused (or explained) by the source
Mean square	Mean square caused (or explained) by the source
Degrees of freedom	Degrees of freedom of the source
Std deviation	Std deviation caused (or explained) by the source
F-statistic	Computed F-statistic value for the given source
Critical quantile	Critical quantile of the F-distribution, If Critical quantile is less than F-statistic, then the parameter is a statistically significant source o variability.
Conclusion	Verbal conclusion of the significance test
p-value	The *p*-value for each test

12.2.3. Graphical output

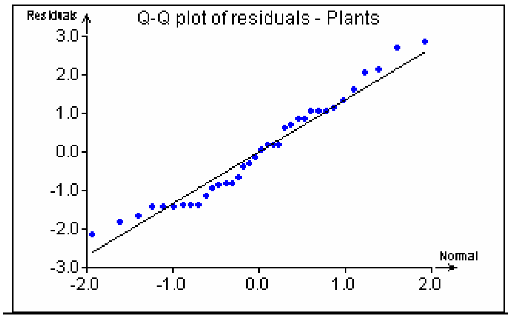
Plot of means



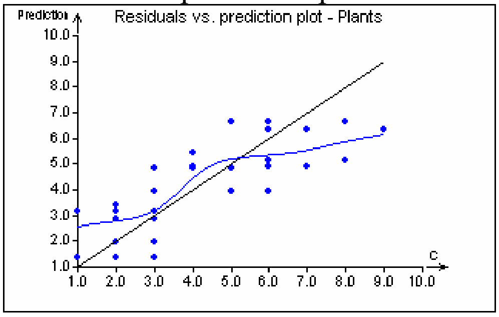
This plot visually compares the response values for each combination of factor levels (or each cell). Each circle stands for one observation. Diameter of the circles are roughly proportional to the response. Differences between cells, between levels and interactions can be seen on this plot.

Q-Q plot of residuals

Q-Q plot of residuals is to explore normality of distribution of the residuals. If the points lie roughly on the line, then the distribution is close to normal. Recall that normality of residuals is one of important assumptions of the Anova model.

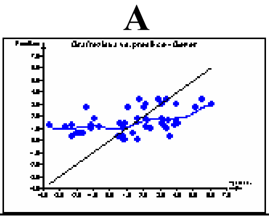


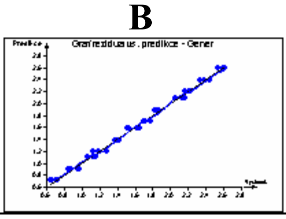
Residuals vs. prediction plot



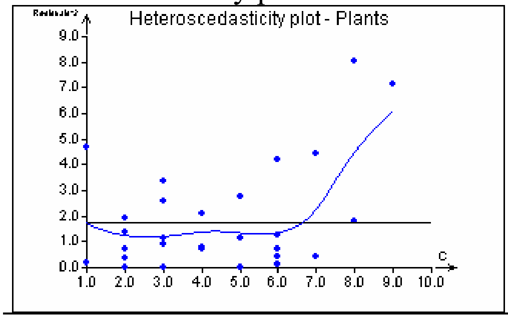
Residuals vs. prediction plot shows the quality of fit of the model. If the model is not significant (cannot explain much of the variability of response), the points lie on a horizontal line (illustration A).
The closer are the points to the line $y = x$, the more significant the model is.

Illustratoion:



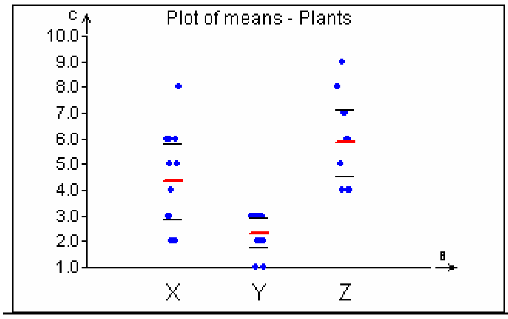


Heteroscedasticity plot



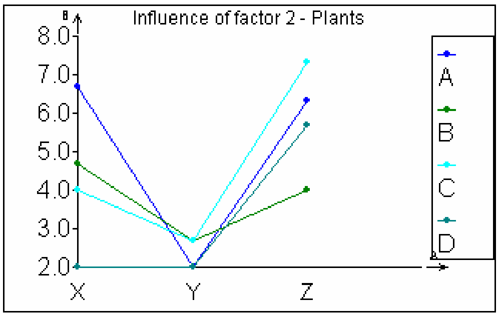
Sometimes it shows, that the variance of the response is higher for bigger values of the response. Heteroscedasticity plot shows the dependence of variance on the resonse value. As this Anova models assumes constant variance of the response, heteroscedasticity may affect its performance. The blue curve on the plot should not decline from horizontal line.

Plot of means

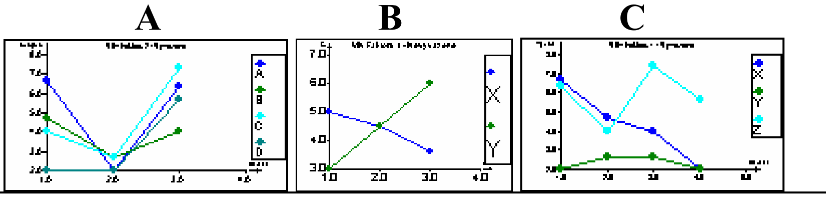
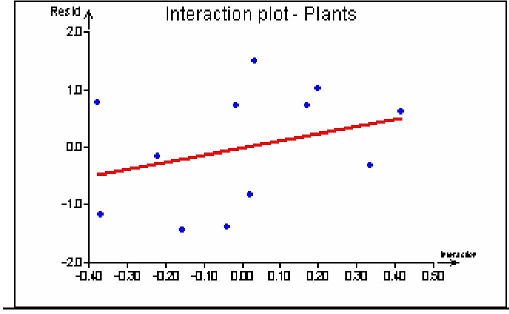


This plot shows the differences between means for individual levels of the first and second factor. Dots are the measured responses, short thick lines are the means, short black lines are the confidence intervals of the means. If the factor is statistically significant, the mean lines are red.

Influence of factor



Plot of the influence of a given factor at separate levels of the other factor. If the lines for factor *A* are similar for each level of the factor *B*, then the factor A is probably significant (illustration A). If the lines have rather opposite direction, then there is probably strong interaction between the factors (illustration B). If the lines are shaped rather randomly, then the influence and significance of the respective factor is probably low, illustration C. The plot is only qualitative visual tool an cannot fully replace the F-test.

	Illustration: 
Interaction plot  	Interaction plot visualizes the significance of the intraction term in the Anova model. If there is a significant slope in the plot, then the interaction between factors is significant. The statistical significance of the interaction is shown by red color of the line.

12.3. Generalized analysis of variance

Menu:	QC.Expert	Anova	Multi-factor

Generalized analysis of variance (GANOVA) helps to assess extent to which a selected numeric response variable is influenced by (a) qualitative factors (as name of operator, workshift, number of production line, raw material supplier, etc.) and/or (b) quantitative numeric variables (as temperature, pressure, mixer revolutions). The response variable of interest may typically be a quality parameter or any process or experimental output. As such, this module is useful for spotting how to influence or stabilise an important output parameter, to identify and prove influence of certain factors on important output and consequently eliminate or stabilise the influential factors to stabilise or improve quality.

This module is a generalization of a linear regression model with so called dummy binary variables and with use of general Moore-Penrose pseudoinversion o the characteristic matrix. Linear regression model also allows to prdict the response for a given values of predictors (a) and (b). The module analyses the influence of fixed-effect factors and continuous variables on the response. Observations Z_i at n_j different levels of predictor factor X_j and various values of the predictor variable Y_k can be described by linear regression model with unknown parameters

$$Z = \alpha_0 + \sum_j \mathbf{\alpha}_j X_j + \sum_k \beta_k Y_k + \varepsilon ,$$

where α_0 is the absolute term (overall mean value), $\boldsymbol{\alpha}_j$ is an $(n_j \times 1)$ vector of latent parameters for j-th factor and β_k is the regression coefficient for k-th variable. Random error ε_{ij} is assumed to have normal distribution with zero mean, $\varepsilon \sim N(0,\sigma^2)$. Latent parameters for factors have no direct meaning, but they can be used to test statistical significance of the factor usinf an F-test. The module *Anova – multi factor* computes $\mathbf{a}_j$, b_k, e_i, which are best estimates of the coefficients $\mathbf{\alpha}$, β, and the mesurement errors ε.

12.3.1. Data and parameters

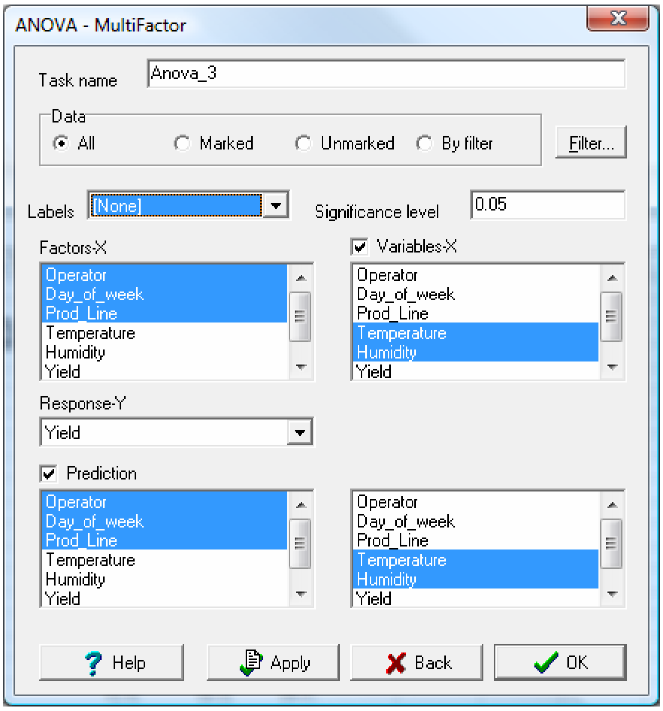
Data are organised in columns, in any order. Each column corresponds to one numerical predictor or one factor or the numerical response value. Data must contain at least one factor column and one column of the response variable. Factor columns are not numerical, factor levels are defined as texts like YES, NO, or A, B, C, D, etc. Number of different text strings define the number of levels of the factor. Each factor must have at least two levels. Variables are numerical values. Value in the response column must correspond to the combination of predictors in the respective row.

Table: Sample input. Predictors are two factors (production line and operator) and one numerical variable (temperature). For every combination of predictors there is one observed value of the response (yield).

Predictors			
Factors		*Variable*	*Response*
Line	**Operator**	**Temp**	**Yield**
A	Brown	13.3	14.6
B	Smith	16.3	17.4
C	Brown	18.7	13.3
A	Mitchel	14.5	12.6
B	Mitchel	11.1	17.5
C	Smith	16.0	14.9
.....	.....	.....	.....

If the data contain only numerical predictors, it is necessary to use linear regression module. If there is only one or two factors, it is more suitable to use the respective one- or two-factor Anova module.

In the „Anova-Multifactor“ dialog panel (see Fig. 49), select the factors and variables, select the response variable column. If there is no variable predictor, uncheck the checkbox *Variables-X*. In the field *Response-Y* select one column with the response values. Select the significance level (typically 0.05, or 5%). If you want to calculate predicted values, select columns for prediction in the field *Prediction* and check the *Prediction* checkbox. The prediction columns must have the same structure as the predictors. The predictors themselves may be selected for calculating prediction. After clicking on *OK*, the results are written in the *Protocol* and *Graphs* windows.



ANOVA - MultiFactor

Task name: Anova_3

Data: ● All ○ Marked ○ Unmarked ○ By filter

Labels: [None]

Significance level: 0.05

Factors-X: Operator, Day_of_week, Prod_Line (selected); Temperature, Humidity, Yield

☑ Variables-X: Operator, Day_of_week, Prod_Line; Temperature, Humidity (selected); Yield

Response-Y: Yield

☑ Prediction

Operator, Day_of_week, Prod_Line (selected); Temperature, Humidity, Yield

Operator, Day_of_week, Prod_Line; Temperature, Humidity (selected); Yield

Fig. 49 Anova-Multifactor dialog panel

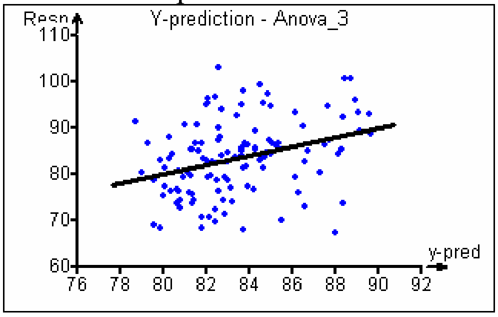
12.3.2. Protocol

Anova - Multifactor	
No of cases	Total number of cases, or rows in the data table.
Total no of predictors	Number of factors and variables.
No of factors	Number of factors.
No of variables	Number of variables.
Mean Y	Average of all responses.
Absolute term	Predicted response value with no influence of factors and zero value of all variable predictors.
Significance level	Chosen value of significance for statistical tests. Recommended value is 0.05.
No of levels	Numbers of levels for all factors in the model.
Overall ANOVA	Overall test if the predictors have any influence at the response.
Source	Source of the variability is assessed. If the model explained satisfactory portion of the response variability, then it may be assumed significant. Primary variability measure is the sum of squared residuals.
Degrees of freedom	Degrees of freedom for every factor or variable.
Sum of squares	Variability expressed as sum of squares.
Variance	Variability expressed as variance.
F-statistic	The ratio of the variance without and with the model.
p-value	If the p-value is less then the chosem significance value, then the factor is statistically significant.
Significance	Verbal result of the significance test.
Source	Variability sources
Total variability	Values for the total variability $CSC = \sum_{i=1}^{n} \left(Z_i - \overline{Z} \right)^2$
Explained variability	$CSC - RSC$
Residual variability	Residual variability $RSC = \sum_{i=1}^{n} \left[Z_i - \left(\alpha_0 + \sum_j \mathbf{a}_j X_{ij} + \sum_k b_k Y_{ik} + e_i \right) \right]$ *Note*: CSC: total sum of squares, RSC: residual sum of squares
ANOVA for individual factors	Ammount of variance explained by the predictors contained in the model, factors and variables.
Predictor	Name of factor or variable.
Parameter	Estimated value of the variable coefficient b_k, For factors this field remains empty.
Sum of squares	Sum of squares explained by this predictor.
F-statistic	Corresponding F- quantile.
p-value	*p*-value. If *p*-value is less then the required significance level (usually 0.05) this predictor is significant (it significantly influences the value of response).
Significance	Verbal result of the test: Significant or Insignificant.
Prediction	When the checkbox *Prediction*, was checked, a table of predicted values of response is included.

Prediction	Calculated (predicted) values of the response

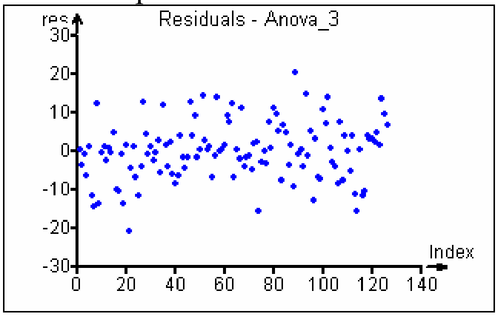
12.3.3. Graphical output

Y-Prediction plot



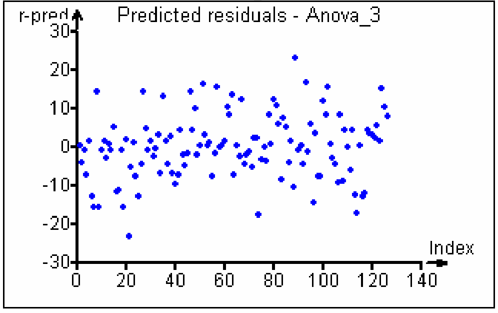
Overall fit plot. Measured values (Resp) are plotted against the predicted values (y-pred). When the points lie close to the line y=x the prediction is successful and the model describes the data well. Here, the model is assessed as a whole, separate factors and variables are assessed in the partial prediction plots, see below. This plot corresponds to the overall explained variability of the model

Residuals plot



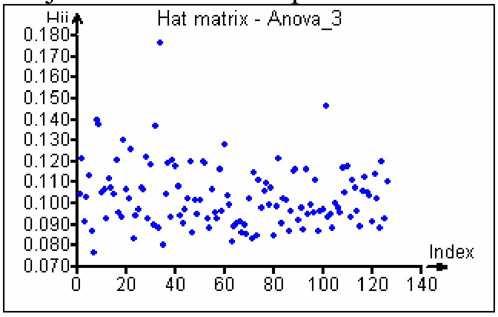
In the plot of residuals, the distances of the response from the model (or residuals) are plotted. Points that are far from the horizontal line are suspected outliers, possibly errors in the measured response of non-typical measurement.

Predicted residuals plot



Predicted residuals are similar to the residuals in the previous plot. Here, each i-th residual value is computed from data with dropped i-th measurement, so the possible outliers are usually more visible. Points far from the line y=0 are suspected outliers, or gross errors.

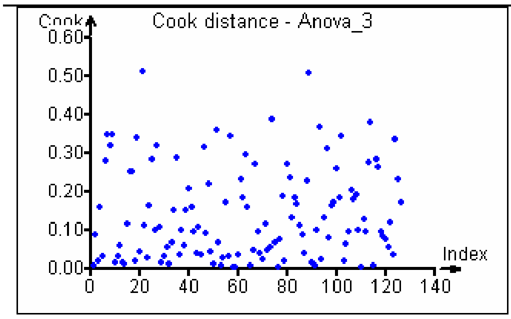
Projection Hat-matrix plot



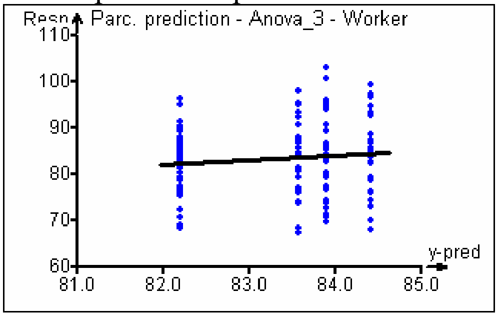
Plot of the diagonal elements of the projection matrix (or so-called hat matrix) $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Points with the high values are usually untypical in the predictor values (such as very high or low variable values or non-typical combination of factor levels). Such points are highly influential and should be paid special attention, as wrong response values could result in biased or unreliable estimates of the model parameters. On the other hand however precisely measured response in such points will significantly improve statistical properties of the model.

Cook distance plot

Value of Cook distances is another measure of influence of the measurements. Interpretation is similar as in the previous plot.

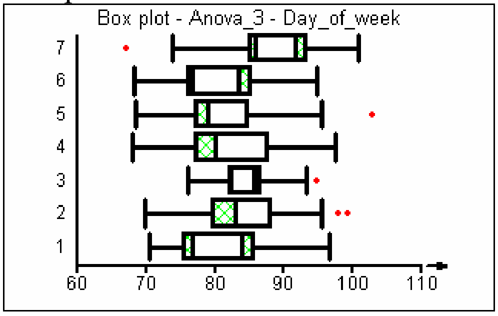


Partial prediction plot for a factor



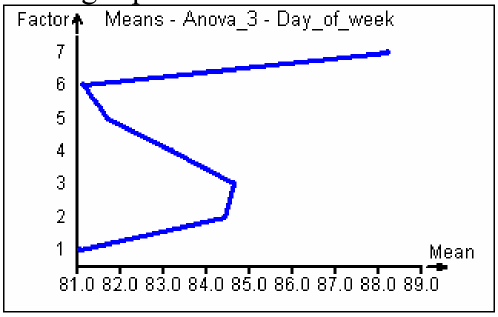
Partial prediction plot shows how separate factor levels influence the response. Strong steep dependence suggests that the respective factor has strong influence on the response. This plot corresponds to explaines variability for the factor and to test of significance of the factor.

Box plot for a factor



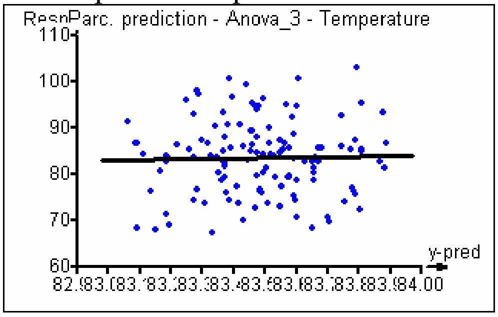
Box plot shows median and quantile characteristics of the response for every level of a factor. Limits of the box are the lower and upper quartiles (25% and 75% quantiles). Separate red points are suspected response outliers The morphology of the box plot is further described in Basic statistics, see 5.3 on page 5-41.

Averages plot



This is another representation of means of response for every level of the factor. The values on the y-axis are avarages of the response for the given factor level.

Partial prediction plot for a variable



This plot shows how much the prediction depends on a given variable. Interpretation of this plot is analogical to the partial plot for a factor.

13. Experimental design (DOE) - Design

Menu:	QCExpert	Experimental Design	Design	Full Factorial Fract Factorial

This module designs a two-level multifactorial orthogonal plan 2^{n-k} and perform its analysis. The DOE module has two parts, *Design* for the experimental design before carying out experiments which will find optimal combinations of factor levels to gain maximum information at a reasonable number of experiments and part *Analysis* described in the next chapter 14 on page 14-97, which will analyse results of the planned experiment. The main goal of DOE is to find which of the factors included in the model have considerable influence on one outcome of the experiment. The outcome is called response and it can typically be yield, energy consumption, costs, rate of non-conforming product units, blood pressure etc. Factors are variables which will set for the purpose of the experiment to two values or levels. Factors must have two states („low“ and „high“, or –1 and +1) defined naturally (night – day, male – female) or defined by the user (low temperature = 160°C, high temperature = 180°C). Each state is assigned the value –1 or +1 respectively, regardless of the sign, i.e. formally high temperature may be defined as the „low“ state (–1) and low temperature as the „high“ state with no effect to the result of the analysis. Factors may typically be night and day, cooling off/on, smoker/nonsmoker, clockwise/counterclockwise mixer rotation, etc. The user defines number of factors *n*, fraction *k* of the full experimental plan and number of replications *m* of each experiment. The module will create a matrix of the experimental plan and stores it in a new data sheet in the form of plus and minus ones. Each row in the spreadsheet represents one experiment. The number of rows is $m2^{n-k}$. Factors are named by letters of the alphabet A, B, C, …. Columns defining order of an experiment and replication are also added for referrence. The column *Response* is left empty – here the user will enter results *Y* of the carried out experiments for further analysis by the module *Design of Experiments – Analysis*. The result of the analysis will be a set of coefficients of a regression model with all linear and all mixed terms (main effects and interactions).

$$Y = a_0 + \sum a_i \mathrm{comb}\left(A, B, C...\right),$$

for example, with 3 factors *A*, *B*, *C* we have a model with $2^3 = 8$ parameters a_0 to a_7.

$$Y = a_0 + a_1A + a_2B + a_3C + a_4AB + a_5AC + a_6BC + a_7ABC$$

A, *B*, *C* are the factors, *AB*, *AC*, *BC* are second-order interactions, *ABC* is the third order interaction. The linear terms coefficients (main effects) reflect an influence of the factor level on *Y*. For example, the value $a_1 = 4$ suggests that the high level of factor *A* results in *Y* bigger by 8 units than at low level of *A*. However, to make a final conclusion about the influence of factors the statistical significance of the coefficients must be assessed either by the significance test when $m > 1$, or by the coefficient QQ-plot, see below. Coefficients at mixed terms like a_4AB are ifluences of one factor conditioned by the level of another factor (interactions). Great value of an interaction coefficient means that the factor influeces *Y* differently in dependence on the level of the other factor.

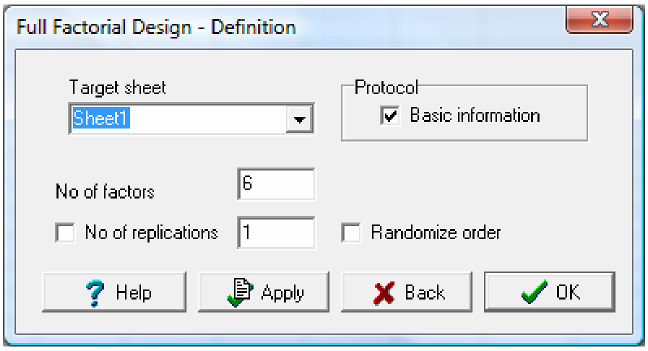
Fractional factorial designs can significantly reduce the number of experiments needed to calculate the coefficients to a fraction 2^{-k} compared to the full fractional design. The fraction *k* can be an integer, generally $0 < k < n$. The number of experiments in such a design will then be $m2^{n-k}$. The price to be paid for such a reduction of the model is aliasing. Each coefficient represents the influence of more than one term of the model, for example a_1 may stand for combination of the influences of the factor *A* and the interaction *AB*, with no possibility to distinguish between there influences. Fractional version of the above model 2^{3-1} with $k = 1$ can thus be written as

$$Y = a_0\left(1 + ABC\right) + a_1\left(A + AB\right) + a_2\left(B + AC\right) + a_3\left(C + BC\right)$$

If the interaction AB is assumed to be negligible, we can take a_1 for the main effect of A. The summation of main effects and interactions is called aliasing. The goal of fractional design is to try to create a design in which a main effect is aliased only with interaction of the highest possible order, as it is generally known that high order interacions are often not present, therefore the respective coefficient represents indeed the influence of the factor. This goal is sometimes difficult to achieve, especially for high k. This module gives the best possible predefined designs in this respect.

13.1.1. Data and parameters

Full factorial design creates a design matrix from the given number of factors n and replications m. Number of generated rows will thus be $m2^n$. Each row corespond to one experiment. Therefore this design is appropriate for lower number of factors, as the number of experiments needed may get quite high, eg. 1024 experiments for 10 factors without replications ($n = 10$, $m = 1$). In the dialog window (Fig. 50) select the target data sheet in which the design will be written. NOTE: Any contents of this sheet will be deleted, so you shoud create a new sheet (Menu: *Format – Sheet – Append*). Fill in number of factors and the desired number of replications of each experiment. If the checkbox at *No of replications* is not checked, the number of replications is ignored, $m = 1$ is taken as default. Check the box *Basic information* if you want to basic description of the design in the Protocol sheet. If the *Randomize order* box is checked, the column Order in the target sheet is randomized and after sorting the rows by this column we can obtain rows of the design in a random order, which may help to avoid possible deformation of response from the systematic sequence of similar experiments.


Full Factorial Design - Definition

Target sheet: Sheet1

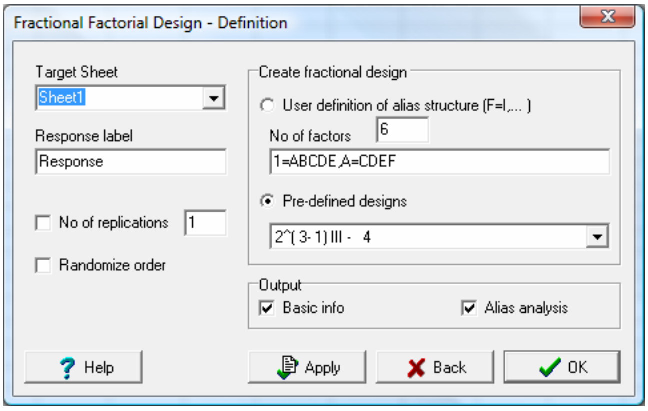
Protocol: ☑ Basic information

No of factors: 6

☐ No of replications: 1

☐ Randomize order

Fig. 50 Full factorial design dialog


Fractional Factorial Design - Definition

Target Sheet: Sheet1

Response label: Response

☐ No of replications: 1

☐ Randomize order

Create fractional design

○ User definition of alias structure (F=I,...)

No of factors: 6

1=ABCDE,A=CDEF

● Pre-defined designs

2^(3- 1) III - 4

Output: ☑ Basic info ☑ Alias analysis

Fig. 51 Fractional factorial design dialog

Fractional factorial design is derived from the full factorial design, but needs much less experiments to estimate coefficients with the drawbacks mentioned above. In the dialog window (Fig. 51) slect the target data sheet in which the design will be written. NOTE: Any contents of this sheet will be deleted, so you shoud create a new sheet (Menu: *Format – Sheet – Append*). The field Response label will be used to label the response column in the design table. If replications are required fill in the desired number of replications of each experiment. If the checkbox at *No of replications* is not checked, the number of replications is ignored, $m = 1$ is taken as default. Check the box *Basic information* if you want to basic description of the design and Alias analysis if the analysis is to be performed in the Protocol sheet. If the *Randomize order* box is checked, the column Order in the target sheet is randomized and after sorting the rows by this column we can obtain rows of the design in a random order, which may help to avoid possible deformation of response from the systematic sequence of similar experiments. The fractionation is based on the design defining relationships in the form of sufficient alias equalities. They can be written in the *User definition of alias structure* field. The number of relationships is equal to k, relationships are separated by comma. There is no easy way to find optimal design definition, as the defining relationship implies other aliases, some of which may disqualify the design. For example, if we attempt to define a 2^{4-1} design for

4 factors A, B, C, D by a defining relationship A = ABD, we will get the alias B = D and will not be able to separate influence of main effects! DO NOT use user definitions unless you are sure they are correct, otherwise they will most probably lead to an unusable or non-optimal plan, with aliases of main effects, such as A = C. It is highly recommended to use predefined designs in the drop-down list *Pre-defined designs* field. The designs are ordered by the numder of factors *n* and the fraction *k*. The design decriptions have the following meaning

2	^(	3	-	1	)	III	-	4
		Number of factors *n*		Fraction order *k*		Design resolution		Number of experiments needed

Design resolution is the information gain parameter related to the alias structure. The designs with aliases between main effect and high order interaction are more informative and have high resolution value. The design should be a compromise between the number of experiments and the design resolution.

Table 2 List of pre-defined optimal designs

No	Type of design	Fraction	Resolution	Experiments needed
1	2^{3-1}	3-1	III	4
2	2^{4-1}	4-1	IV	8
3	2^{5-1}	5-1	V	16
4	2^{5-2}	5-2	III	8
5	2^{6-1}	6-1	VI	32
6	2^{6-2}	6-2	IV	16
7	2^{6-3}	6-3	III	8
8	2^{7-1}	7-1	VII	64
9	2^{7-2}	7-2	IV	32
10	2^{7-3}	7-3	IV	16
11	2^{7-4}	7-4	III	8
12	2^{8-2}	8-2	V	64
13	2^{8-3}	8-3	IV	32
14	2^{8-4}	8-4	IV	16
15	2^{9-2}	9-2	VI	128
16	2^{9-3}	9-3	IV	64
17	2^{9-4}	9-4	IV	32
18	2^{9-5}	9-5	III	16
19	2^{10-3}	10-3	V	128
20	2^{10-4}	10-4	IV	64
21	2^{10-5}	10-5	IV	32
22	2^{10-6}	10-6	III	16
23	2^{11-5}	11-5	IV	64
24	2^{11-6}	11-6	IV	32
25	2^{11-7}	11-7	III	16
26	2^{12-8}	12-8	III	16
27	2^{13-9}	13-9	III	16
28	2^{14-10}	14-10	III	16
29	2^{15-11}	15-11	III	16

Table 3 Examples of 2^(5-2) designs

(A) Optimal design	(B) Unusable design, since A=D and B=absolute term

Design definition: D = AB, E = AC
 	
A = BD = CE = ABCDE
B = AD = CDE = ABCE
C = AE = BDE = ABCD
D = AB = BCE = ACDE
E = AC = BCD = ABDE
BC = DE = ABE = ACD
BE = CD = ABC = ADE
ABD = ACE = BCDE = 1.0
	Design definition: A = AB, B = AD
	
	A = D = AB = BD
	B = AD = ABD = 1.0
	C = BC = ACD = ABCD
	E = BE = ADE = ABDE
	AC = CD = ABC = BCD
	AE = DE = ABE = BDE
	CE = BCE = ACDE = ABCDE
	ACE = CDE = ABCE = BCDE

13.2. Protocol

Design type	Full factorial, 2^*n* or Fractional factorial 2^(*n*-*k*).

Design definition	Only for Fractional factorial, defining relationships, eg:
	E = ABC
	F = BCD

Design description	Only for Fractional factorial design 2^(*n*-*k*), resolution, number of experiments (without replications). For example „2^(3- 1) III - 4“ means 2-level factors, 3 factors in design, half – fraction of the full design, resolution III, 4 distinct experiments.
No of factors	Number of factors
No of replications	Number of replications of each experiment
No of experiments	Number of distinct experiments
Alias-structure anaylsis	Only for fractional design. Complete listing of all aliases, of grouped combinations of undistinguishable factors and interactions, Aliases described by one coeficient are on one row. For example, if the alias row contains „B AD CDE ABCE“, then the coefficient for the factor „B“ will also include effects of interactions AD, CDE a ABCE. Number „1“ represents the absolute term a_0 in the model. Aliases between factors such as A = C are undesirable, as in that case we have no information about the influence of the factors A and C.

13.3. Graphical output

This module does not generate any plots.

14. Experimental design (DOE) - Analysis

Menu:	QCExpert	Experimental Design	Analysis

This module analyses data prepared by the previous module (Experimental Design). It can analyze bothe full factorial and fractional factorial designs 2^n a 2^{n-k}, with filled in results (responses) of the experiments in the *Response* column.

The main purpose of a designed experiment analysis is to determine which of the factors have significant influence on the measured response. Based on these responses, the module computes the coefficients of the design model using the multiway ANOVA model. If the design does not contain replicated experiments, the resulting model has zero degrees of freedom. In consequence, coefficient estimates do no allow for any statistical analysis, all residuals are by definition zero and significance of factors and/or interactions can only be assessed graphically using the coefficient QQ plot. With replicated experiments the analysis is formally regression analysis, so we can obtain estimates with statistical parameters (variances) and test the significance of factors statistically. It is therefore recommended to replicate experiments where possible.

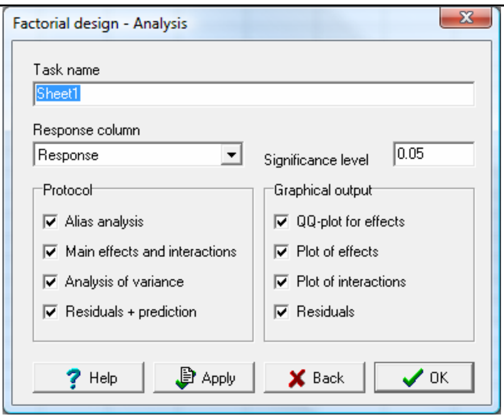
14.1. Data and parameters

An example of the data for the module Design of Experiments – Analysis is shown in Table 4. All data except the *Response* column were generated by the previous module (see chapter 13). After setting factors according the design and carrying out all 16 experimental measurements (or responses), the response values are written to the data table and whole table is submitted to analysis.

In the dialog window Factorial Design – Analysis (Fig. 52) the response column is pre-selected. The significance level is applicable only in case of replicated experiment, where statistical analysis is possible. The user can select items to be included in the text protocol output and plots in the graphical output. An advanced user can also write a design manually using the required notation: –1 for low and 1 for high factor level, first 2 columns in data sheet will be ignored, names of factor columns are ignored, factors are always named A, B, C,…, last column is expected to contain measured responses. Incorrect or unballanced designs are not accepted and may end with an error message. It is recommended however to use designs created by the Experimental design module.

Table 4 Example of data for analysis of a designed fractional factorial experiment 25-2 with 5 factors and 2 replications

Order	Replication	A	B	C	D	E	Response
1	1	-1	-1	-1	1	1	14.6
2	2	-1	-1	-1	1	1	14.5
3	1	-1	-1	1	1	-1	13.6
4	2	-1	-1	1	1	-1	13.6
5	1	-1	1	-1	-1	1	15.1
6	2	-1	1	-1	-1	1	14.7
7	1	-1	1	1	-1	-1	13.2
8	2	-1	1	1	-1	-1	13.3
9	1	1	-1	-1	-1	-1	16.4
10	2	1	-1	-1	-1	-1	16.4
11	1	1	-1	1	-1	1	15.3
12	2	1	-1	1	-1	1	15.1
13	1	1	1	-1	1	-1	14.7
14	2	1	1	-1	1	-1	14.6
15	1	1	1	1	1	1	17.1
16	2	1	1	1	1	1	16.7



Factorial design - Analysis

Task name: Sheet1

Response column: Response

Significance level: 0.05

Protocol
- ☑ Alias analysis
- ☑ Main effects and interactions
- ☑ Analysis of variance
- ☑ Residuals + prediction

Graphical output
- ☑ QQ-plot for effects
- ☑ Plot of effects
- ☑ Plot of interactions
- ☑ Residuals

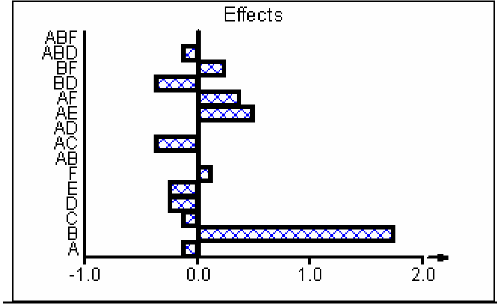
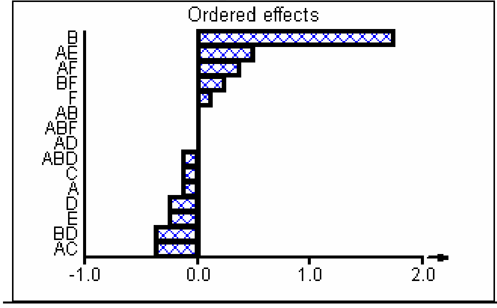
Fig. 52 Dialog window for Factorial design – Analysis

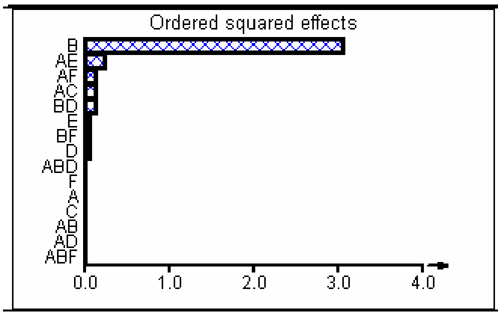
14.2. Protocol

Designed experiment analysis	
Design type	Factorial, full design, or fractional design with description in the form 2^(*n*-*k*), eg. 2^(5-2).
No of factors	Number of factors in the design
No of replications	Number of replications
No of experiments	Total number of experiments (number of data rows)
Design is / IS NOT orthogonal	Information if the design is or is not orthogonal. Orthogonality is one of the requirements for a stable and effective design. All designs generated by QCExpert are orthogonal.
Alias-structure anaylsis	Only for fractional design. Complete listing of all aliases, of grouped combinations of undistinguishable factors and interactions, Aliases described by one coeficient are on one row. For example, if the alias row contains „B AD CDE ABCE“, then the coefficient for the factor „B“ will also include effects of interactions AD, CDE a ABCE. Number „1“ represents the absolute term a_0 in the model. Aliases between factors such as A = C are undesirable, as in that case we have no information about the influence of the factors A and C.
Main effect values and interactions	Computed values of influences for factors and interactions.
Effect, interaction	Factor or interaction, remember that in fractional design, each factor or interaction listed here is aliased with one or more other intraction and the values are a sum of all aliased influences.
Coefficient	Estimates of main effects, interactions and the absolute term. The absolute term is the expected value of the response when all factors are at the low level. These coefficients are the actual effect of the factors and interactions.
Value	Estimates of parameters of the regression model. As here the factors are represented by values –1, +1, the parameter values are half the effects.
Std Deviation	Standard deviations of regression coefficients can be computed only for replicated experiments. Otherwise, the deviations are zero.

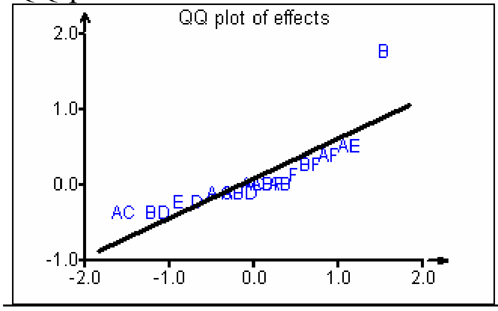
Anaysis of variance	Anaysis of variance table.
Source	Source of variability.
Total	Total variability of the response $Y - a_0$.
Explained by model	Variability explained by the model.
Residual	Residual variability not explained by the model. This variability is zero for non-replicated experiments.
Influence on variance	Separated average and variability for low (-) and high (+) levels of factors.
Source	
Average(-), (+)	Average response for low (-) and high (+) levels of factors.
Variance(-), (+)	Response variance for low (-) and high (+) levels of factors.
Ratio(+/-)	Ratio of variances at high and low level of the factors. Too high or too low value of the ratio may indicate significant influence of the given factor on response variability which can be interpreted as decrease or increase of quality if Y is the quality parameter or stability of the response variable.
Residuals and prediction	Table of predicted response and residuals. This table is applicable only for repeated experiments, otherwise responses are the same as measured responses and residuals are zero.
Response	Measured response Y.
Prediction	Predicted response Y_{pred} from the computed model.
Residual	Residuals Y - Y_{pred}.

14.3. Graphical output

Effects plot  	Plot of the computed effects sorted alphabetically and by the interaction order. Greatest values (regardless of the sign) may suggest significant influence of the respective factor or interaction. This plot shoud be compared with the Effects QQ-plot.
Ordered effects plot  	The same as the previous plot, the values are sorted decreasingly.
Ordered square effects plot	Plot of the squared computed effects sorted decreqasingly. Greatest values may suggest significant influence of the respective factor or interaction. This plot shoud be compared with the Effects QQ-plot.

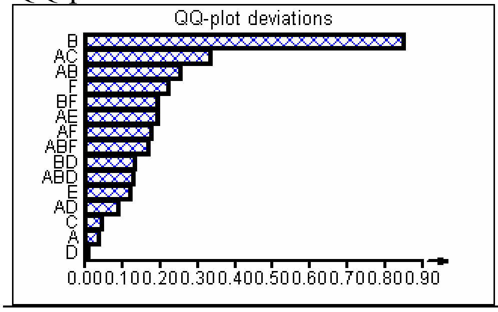


QQ-plot for effects



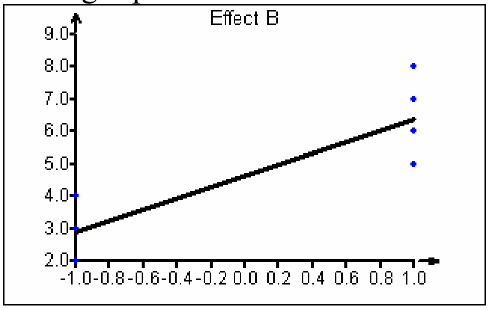
If all effects and interactions are zero, the effects distribution follow the normal distribution. In QQ-plot we can see deviations from normal distribution for individual factors. Such deviations (like factor B on the picture) can be interpreted as significant effect of the factor.

QQ-plot deviations

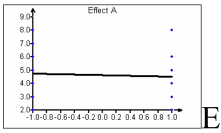


Absolute deviatiosn from the line in the QQ-plot. High values suggest significance of factors.

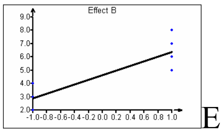
Averages plot



Averages plot gives average response for low and high level of each factor. The scale on all plots are the same so the plots can be compared.

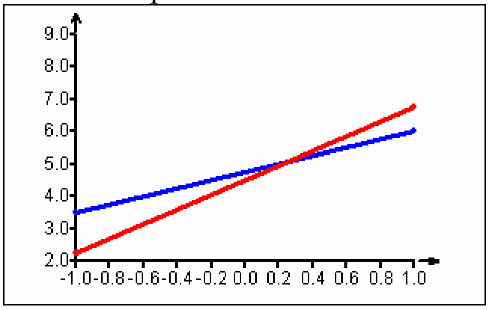


Example of small effect

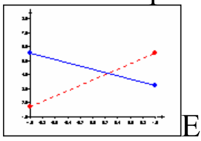


Example of high effect.

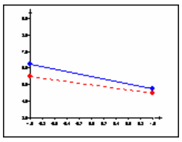
Interaction plots



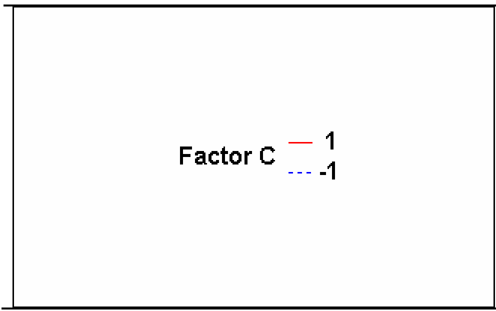
Interactions plot can reveal possible significant interactions of the first order between factors. Interaction will be diagnosed if the slopes of the blue and red line differ significantly. The scale on all plots are the same so the plots can be compared.



Example of a significant interaction



Example of an insignificant or no interaction

	Interaction of two factors, say A and B mean that a factor A influences the response differently in dependence on the level of factor B.

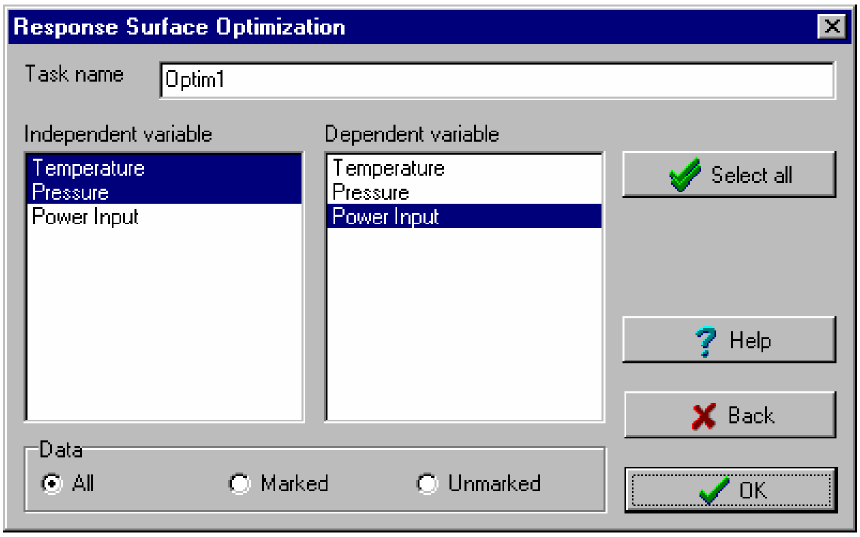
15. Response surface

Menu: | QC.Expert | Optimization

This module helps to find optimum values of technological or other (independent) variables, using observations of some output (dependent or response) variable obtained experimentally. Assumptions are that a) the optimum independent variables settings (e.g. temperature, pressure, drying time) correspond to minimum or maximum of the dependent variable mean, b) dependent variable values were observed for various settings of independent variables, c) minimum or maximum of the dependent variable mean exists and is not very far from experimental settings tried. The *Response surface* module fits a model through the experimental data – complete Taylor polynomial of second order and tries to find its extremum by looking for a stationary point (a point with zero first partial derivatives). When extremum (minimum or maximum) exists, its estimate and confidence interval are given in the protocol. When the optimized model has no extremum, stationary point corresponds to a saddle point and no optimum setting for independent variables can be found. Optimum independent variables setting can be located outside of experimental region, the estimate is less reliable in such case however.

15.1. Data and parameters

Data are organized into columns. Each column has to have the same number of data points. Dependent and independent variables are selected in the Response surface methodology dialog panel, see Figure 23.



Response Surface Optimization

Task name: Optim1

Independent variable: Temperature, Pressure (selected); Power Input

Dependent variable: Temperature; Pressure; Power Input (selected)

Data: ● All ○ Marked ○ Unmarked

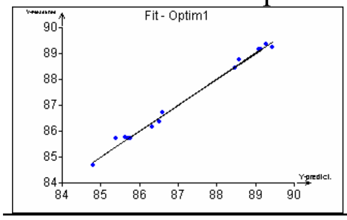
Fig. 53 Response surface methodology dialog panel

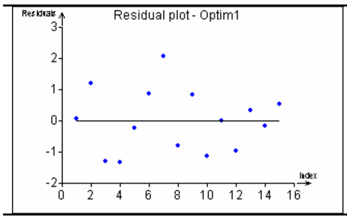
15.2. Protocol

Number of variables	Number of independent variables, i.e. variables for which optimum setting is sought.
Number of observations	Number of observations, i.e. number of data rows.
Degrees of freedom	Difference between the number of observations and the number of parameters in the quadratic optimization model. Precision and reliability of the optimum setting estimate depends on the degrees of freedom to the number of data points ratio. The number of data points should be roughly comparable to degrees of freedom.
Stationary point type	Minimum, maximum, or saddle point. Saddle point does not correspond to an extreme, so that no optimum setting can be found. A saddle point can occur ① due to narrow or otherwise improper range of independent variables in the experimental dataset, ② due to large experimental error, ③ because no optimum setting exists around the experimental points.
Stationary point	
X0	If minimum or maximum was found, the optimum setting of independent variables is given.
Lower limit	Lower 95% confidence limit for the dependent variable optimum value.
Upper limit	Upper 95% confidence limit for the dependent variable optimum value.
Estimate	Estimate of the dependent variable optimum value.
Confidence interval	Confidence interval for the dependent variable value at the stationary point.
Mean absolute error	Mean absolute difference between dependent variable observations and the quadratic optimization model.
Residual sum of squares	Sum of squared differences between dependent variable observations and the quadratic optimization model.
Residual variance	Variance of the residuals after the quadratic optimization model.
Design conditioning number	A diagnostic value useful for checking the experimental settings layout (experimental design). If it is too large, the “Ill conditioned plan” warning is issued. It means that the independent variables are almost linearly dependent and some new experimental points should be added to decrease collinearity. Collinearity can be checked in the Correlation module, see 5.3.
Correlation coefficient	Multiple correlation, summarizing how well the quadratic model fits the experimental data.
Determinant	Determinant of the X’X matrix.

15.3. Graphical output

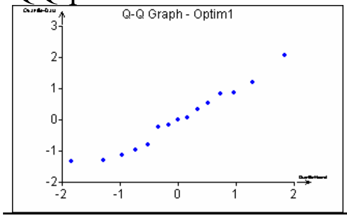
Plots are generated only if the quadratic model is chosen.

Observed vs. fitted plot  	Shows how good is the fit of quadratic model. Observed values of the response are plotted on y-axis, while fitted values are plotted on x-axis. The closer points are to the y=x line, the better is the fit. If the points plot as a random cloud, it is likely that optimum will not be found.
Residual plot	Standardized residuals, i.e. standardized differences between observations and the model fit. Residuals divided by the residual standard



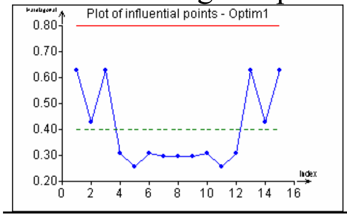
deviation are plotted on y-axis. Points, falling above 3 or below –3 horizontal lines correspond to data which are not fitted well by the quadratic optimization model. Sequential observation number is plotted on x-axis. When the points plot as a random cloud, the fit is rather good. If a trend is visible, the quadratic model does not fit data well.

QQ-plot



This plot is useful for residual diagnostics. It is a useful addition to the Residual plot. The meaning is similar to the general QQ-plot meaning, described in the Basic data analysis module, 5.1.3.

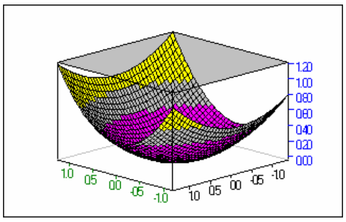
Hat matrix diagonal plot



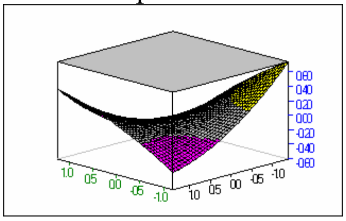
Shows points, which influence the fitted model heavily – much more than the rest of the data. Any influential points found should be checked for validity and accuracy. When the influential points cannot be checked independently, it might be advisable to exclude them.

3D quadratic response surface plot

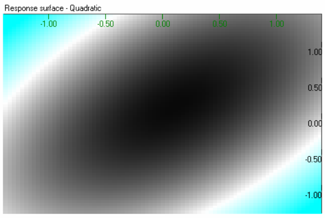
A. Minimum



B. Saddle point



C. 2D-view



In case of two independent predictors, this plot shows the shape of the estimated quadratic response surface with visible minimum, maximum (A) or a saddle point (B). Hint: Use the *2D View* option to get the contour plot (C) where the position of the extreme may be more clearly visible.

16. Multivariate Methods

16.1. Correlation

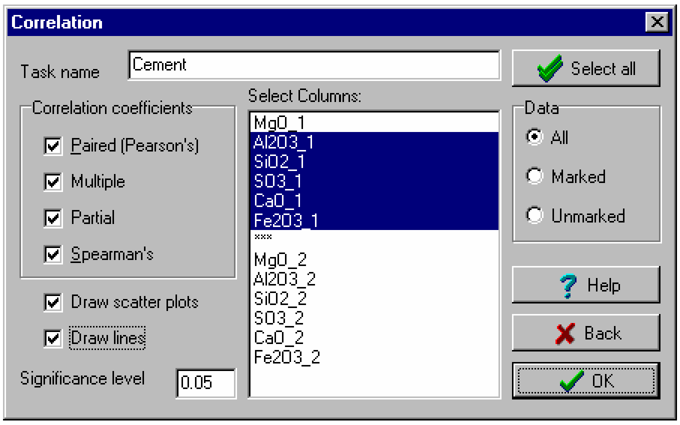
Menu:	QC.Expert	Correlation

Correlation analysis is an important tool for studying relationships among different variables. The correlation coefficient $r_{A,B}$ expresses the degree of linear dependence among variables A and B. The QC.Expert program computes three types of correlation coefficients: pairwise, partial and multiple correlation coefficients. Their meaning is discussed in detail later, in 5.3.2. The pairwise and partial correlation coefficients values lie within the (-1,1) interval. Values close to +1 or -1 correspond to a strong linear relationship. Positive $r_{A,B}$ sign means that when A increases, B tends to increase, while negative sign means that when A increases, B tends to decrease. Negative correlation coefficient sign describes a linear relationship with a negative proportionality factor (B=a-k.A, k>0). It should not be confused with a reciprocal relationship between the two variables (B=k/A, k>0). When the correlation coefficient is close to zero, it might be hard to decide whether B increases or decreases with A. The test for correlation coefficient helps to decide whether a linear relationship is significant (i.e. the correlation significantly differs from zero). Such a test is used for the autocorrelation testing in 5.1. The $r_{A,A}$ correlation between the same two variables is always 1 and hence it is not reported in the output. The A and B order does not matter, so that for each pair of variables, only one coefficient is reported. The multiple correlation coefficient expresses how strong is a linear relationship between one variable A and several other variables. When increasing the number of the variables, the multiple correlation coefficient cannot decrease (it increases or stays the same). The Spearman correlation coefficients are used for screening purposes when outliers or nonlinear monotonic dependencies are expected in the data.

Note:
When several variables are measured on one unit (e.g. a piece of product), the variables should be checked for possible correlation, before control charts are constructed. Significantly correlated variables should not be used in separate control charts (e.g. Shewhart charts). This is because the charts are constructed under the variable independence assumption so that they might not provide correct results for correlated variables. The Hotelling control chart (see 7.3) is appropriate in such cases.

16.1.1. Data and parameters

Each data column corresponds to a variable. The columns can have different number of data points. The rows containing an empty cell (missing value) are skipped during computations. The minimum column number is 2, the minimum row number is 3. Column names should correspond to variable names, e.g. Cr_amount, Mn_amount, S_amount. All columns are selected by default, specific columns of the current data sheet can be selected in the *Select columns* field of the *Correlation analysis* dialog panel, Figure 19. Depending on the items checked, pairwise, partial or multiple correlation coefficients are computed. Significance level for testing correlation coefficients is specified in the *Correlation analysis* dialog panel as well. No outliers should be present in the data to estimate correlation properly. Their presence can be checked in the Basic data analysis module. When the *Scatterplots* field is checked, the matrix of scatterplots for all variable pairs is produced. When the *Lines* field is checked, regression lines are computed and added to the scatterplots.



Task name: Cement

Correlation coefficients
☑ Paired (Pearson's)
☑ Multiple
☑ Partial
☑ Spearman's

☑ Draw scatter plots
☑ Draw lines

Significance level: 0.05

Select Columns:
MgO_1
Al2O3_1
SiO2_1
SO3_1
CaO_1
Fe2O3_1
***
MgO_2
Al2O3_2
SiO2_2
SO3_2
CaO_2
Fe2O3_2

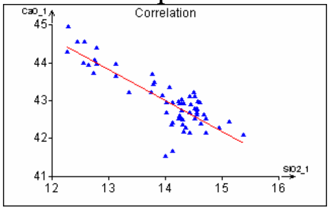
Data
● All
○ Marked
○ Unmarked

Fig. 54 Correlation dialog panel, columns A,B,C,D selected

16.1.2. Protocol

Pairwise correlations	Values of pairwise correlation coefficients. Significant coefficients are printed bold. The leftmost column contains variable names. A positive significant value of the coefficient means that when one variable increases, so does the other. A negative significant value means that when one variable increases, the other decreases. Order of the variable names can be reversed.
Partial correlation	Correlation coefficients which express strength of the net linear relationship between two variables, after filtering out the linear influence of other variables. They are meaningful only when more than two variables are considered simultaneously. The partial correlation is often a helpful tool when one is interested in studying the true relationship between two variables, which is not masked by the influence of other variables.
Multiple correlation	The coefficient describes the strength of linear relationship between one variable and several other (explanatory) variables taken simultaneously. This coefficient is larger than the largest of the corresponding pairwise correlation coefficient. The multiple correlation cannot decrease when the number of explanatory variables increases. It tends to increase even if the pairvise correlation with the newly added variable is not significant. Statistically significant coefficients are printed red bold.
Spearman correlations	Nonparametric estimates of pairwise correlations based on ranks instead of observed values. Because of their robustness, the spearman correlations are recommended when outliers are expected in the data. Statistically significant coefficients are printed red bold.

16.1.3. Graphical output

Correlation plot 	Scatterplots for all pairs of analyzed variables. They can help to detect a nonlinear relationship between variables, not captured by the correlation coefficients. When the appropriate selection is checked, regression lines are added to the plots. When correlation is significant, the line is solid red, when it is not significant, the line is dashed black.

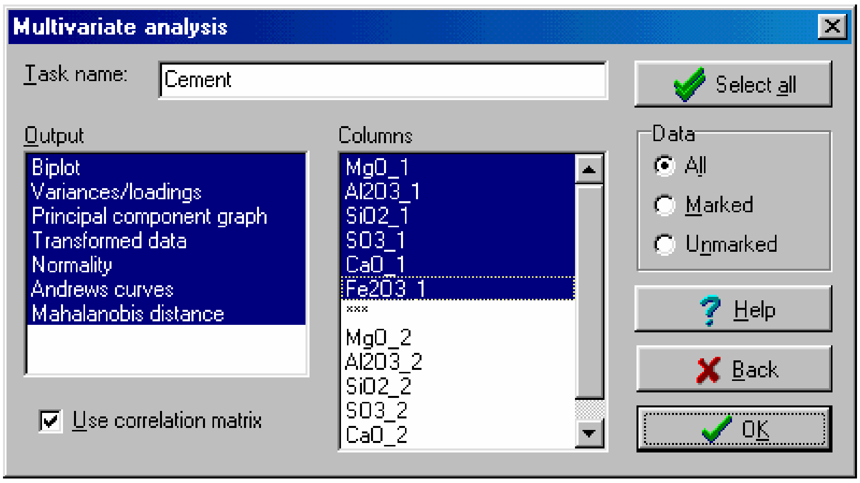
16.2. Multivariate analysis

Menu:	QCExpert	Multivariate Analysis

The Multivariate analysis module is useful for exploratory analysis of multivariate quantitative data. Further, it allows you to perform the principal components analysis. Multivariate data (formally a random sample with vector-valued observations) arise as a result of simultaneous measurement of several (m) variables on the same unit. For instance, several physical and/or chemical properties of one sample can be measured, several linear measurements can be taken on the same piece of product, or there might be several characteristics for any employee available. The number of the vector observations is denoted by n. To check whether data follow approximately multivariate normal normal distribution, a multivariate normality plot, symmetry plots can be used, together with interactive plots which can help to identify outliers. Such checks are useful for instance in connection with the Hotelling chart construction, where the multivariate normal distribution is assumed. The Andrews plot and Biplot can be used to explore data structure. The principal component analysis is based on coordinate transformation, chosen in such a way that the resulting coordinate system is orthogonal and as much of the original multivariate variability is retained by as few newly defined variables as possible. The amount of variability described by the principal components is displayed in the Screeplot. Composition of individual components in terms of the original variables is displayed in the Loadings plot. Another multivariate statistical concept is the Mahalanobis distance (MD), which is a multivariate analogue of distance between a given point and the mean, measured in standard deviation units. Such a type of distance measure has a direct probability interpretation. Large MD values occur with small probability. In the model checking context, they suggest an outlier. When outlier checking is the main goal, a robust version of MD (based on an M-estimate of the multivariate mean) might be useful.

16.2.1. Data and parameters

Data are organized in columns, each column corresponds to a variable. Number of values should be the same in all columns. Data rows with missing values in one or more variable will be omitted from computations. Minimum column number is 2. Minimum row number is 4. Column names should resemble actual variable names, e.g. Cr_content, Mn_content, Elasticity. Columns, corresponding to the variables you want to include in the analysis can be selected in the *Columns* field of the *Multivariate analysis* dialog panel, Fig. 55. All current data sheet columns are selected by default. Requested output items can be selected in the *Output* field. When the *Use correlation matrix* selection is checked, the principal component analysis is based on the correlation matrix, otherwise it is based on the covariance matrix. Covariance based analysis is recommended especially when the scale of the analyzed variables is vastly different.



Multivariate analysis

Task name: Cement

Output: Biplot, Variances/loadings, Principal component graph, Transformed data, Normality, Andrews curves, Mahalanobis distance

Columns: MgO_1, Al2O3_1, SiO2_1, SO3_1, CaO_1, Fe2O3_1, ***, MgO_2, Al2O3_2, SiO2_2, SO3_2, CaO_2

Data: ● All ○ Marked ○ Unmarked

☑ Use correlation matrix

Fig. 55 Multivariate analysis dialog panel

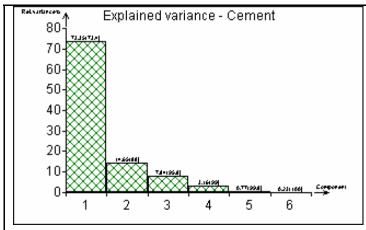
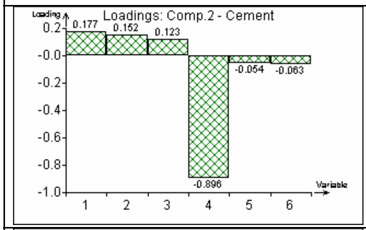
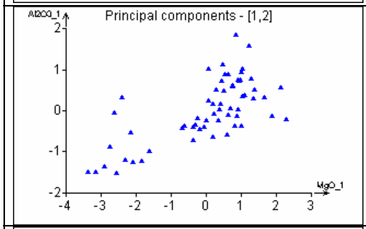
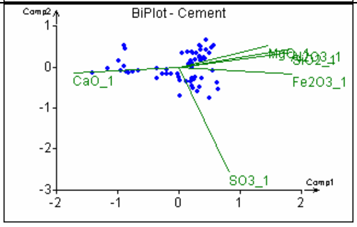
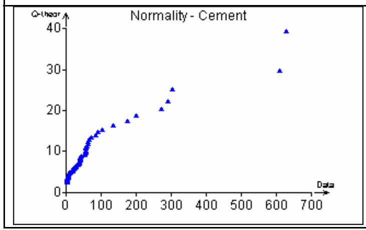
16.2.2. Protocol

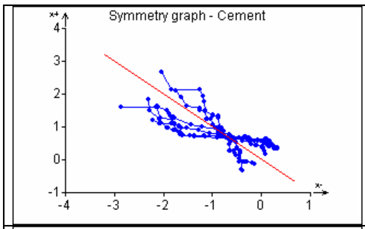
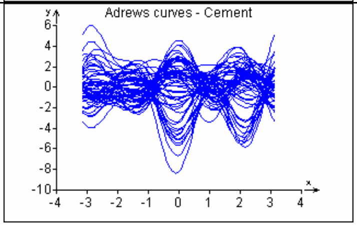
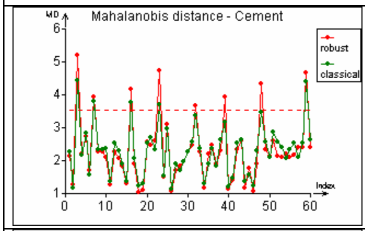
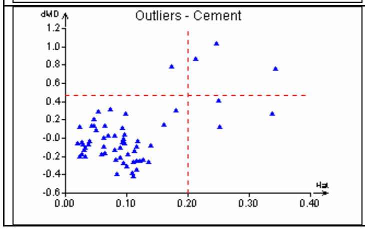
Project name	Project name, as entered in the dialog panel.
Number of variables	Number of variables (columns).
Number of multivariate data points	Number of complete data rows.
Summary statistics	
Variable	Variable name.
Mean	Arithmetic mean.
Variance	Variance.

Standard deviation	Standard deviation.
Minimum	Minimum.
Maximum	Maximum.
Correlation matrix	Pairwise correlations for all variable pairs. The diagonal consists of ones, necessarily. When the *Use correlation matrix* option is checked, the correlation matrix is used for component analysis. See also the Correlation module.
Variable	Variable name.
Covariance matrix	Covariance for all possible variable pairs, arranged in matrix form. Variances appear on the main diagonal. The covariance matrix is used for component analysis when the *Use correlation matrix* is not checked.
Variable	Variable name.
Explained variability	Four characteristics describing how much variability is explained by the principal components. The components are always sorted in descending order of importance. The first component describes most of the multivariate variability, while the last component explains the smallest variance portion.
Principal component	Principal component number.
Eigenvalue	Correlation or covariance matrix eigenvalues (depending on whether the *Use correlation matrix* option in the Multivariate analysis dialog panel is checked).
Variance	Variance explained by a particular principal component. It corresponds to the variance of the projection of original data onto the principal component.
Standard deviation	Square root of the previous characteristic.
Rel. variance,%	Variance explained relatively, in percent of the total variance.
Cumulative variance,%	Cumulative variance, explained by the current principal component and all previous ones, expressed relatively, in percent of the total variance.
Eigenvectors	Correlation or covariance matrix eigenvectors (depending on whether the *Use correlation matrix* option in the Multivariate analysis dialog panel is checked).
Data column	Original variable name.
Loadings	Eigenvectors multiplied by square root of their corresponding eigenvalue. Component loadings can help you to interpret component structure. When a component is tight only to a subset of the original variables, the loadings should be small for all the variables not in the subset.
Data column	Original variable name.
Robust M-estimates	An M-type estimate of the vector of mean values (location vector), based on iterative procedure. It is a robust alternative to the vector of arithmetic means as the classical estimate. Its use should be considered when outliers presence is suspected.
Mahalanobis distance, MD	A distance measure, having a probability interpretation under multivariate normality. It is a multivariate analogue of distance between a point and the mean, measured in standard deviation units. Two versions are available in QC.Expert.
Classical MD	A distance between the *i*-th point $\mathbf{x}_i$ and the the vector of expected vales, as estimated by the vector of means, $\mathbf{x}_p$. The distance measure respects the covariance matrix, which is estimated by **S**, it is given by $(\mathbf{x}_i - \mathbf{x}_p)^T(\mathbf{S})^{-1}(\mathbf{x}_i - \mathbf{x}_p)$.
Robust MD	A robust version of the Mahalanobis distance measure. The vector of location parameters is estimated by a robust, M-type estimator $\mathbf{x}_M$. The robust MD is then

	estimated by $(\mathbf{x}_i - \mathbf{x}_M)^T(\mathbf{S})^{-1}(\mathbf{x}_i - \mathbf{x}_M)$. Outliers have large MD value.
Transformed data	Original data expressed in the newly defined coordinates (obtained by transformation of the original coordinate system).

16.2.3. Graphical output

	Screeplot. It displays variance, explained by individual principal components, expressed relatively, in percent of the overall variance. Components are always ordered in descending order of importance, so that the most important component appears first, the least important appears last. Principal component number appears on the *x* axis, while the percentage of explained variance appears on the *y* axis. Cumulative sums of explained variance appear as numbers above each of the columns. Individual variances form which the screeplot is constructed can be found in protocol.
	Loadings plot. One plot is produced for each of the components. Component loadings can help you to interpret component structure. When a component represents only a subset of the original variables, the loadings should be small for all the variables not in the subset. Thus, for example the first component may be related mainly to mechanical properties, the second to chemical composition, etc.
	Principal component plot. One plot is produced for each pair of components. These plots might be sometimes more useful than scatterplots for original variables pairs. (Simple scatterplots are produced e.g. in the Correlation module.)
	Biplot. It is a plot in which both the observations and the variables are represented in a two dimensional space (plane). The points correspond to data lines, while the lines correspond to data columns. Data points can be selected interactively, clicking on the corresponding points on the plot. When interpreting the plot, you should keep in mind that the plot is based on an underlying 2-dimensional approximation to the original data. For each datapoint, the approximation is proportional to the result of vector multiplication of individual lines and points (taken as vectors with the (0,0) origin). From there, it follows that lines close on the plot should correspond to correlated data columns. Row vectors (points) located in the direction of a data column vector (line), should exhibit higher absolute values of the variable corresponding to that column. On the other hand, you should realize that goodness of the approximation on which the plot is based, decreases as *m* (original data dimensionality) increases. Especially for large *m*, the dimensionality reduction down to two might be too drastic and the plot might not yield much of a useful insight.
	Multivariate normality plot. It is an analogue of the univariate normal Q-Q plot, used to check multivariate normality of the data. Multivariate normal data should plot close to the line. The plot is based on the F distribution valid for Mahalanobis distance under multivariate normality. To reduce problems with dependence between individual $\mathbf{x}_i$'s and their mean $\bar{\mathbf{x}}$, jacknifing is used.

	Symmetry plot is analogous to the half-sum plot in univariate case. A point in the plot corresponds to a data point from the current data sheet, so that more than one cell is marked when one point is selected in the plot interactively. Under perfect symmetry, the points should plot on the red line $y=-x$.
	Andrews curves can be a useful tool for exploratory multivariate analysis. Each of the curves represents one datapoint (a point in *m*-dimensional space). A bunch of similar curves correspond to a cluster of data points similar to each other. Curves of diferent shape that the bulk of others suggest outliers. Individual datapoints (curves) can be selected interactively by clicking/dragging mouse. A selection can be cancelled by repeating the operation while holding the *Ctrl* key.
	Plot of both classical and robust Mahalanobis distance. The robust version (red) is more useful for detection of outliers. The classical version (green) is ploted for comparison. The points above the horizontal line (95% quantile of the appropriate null distribution) are suspect as outliers. Only the red points (obtained from the robust version) can be selected interactively.
	An alternative tool for data diagnostics. Elements of the projection matrix (see the Linear regression chapter) are plotted on the x-axis, while difference between robust and classical Mahalanobis distance are plotted on the y-axis. Points right from the vertical line or above the horizontal line are suspect.

16.3. Canonical correlation

Menu:	QCExpert	Canonical correlation

This module looks for a general linear relationship between two multivariate variables X and Y with dimensions m_1, m_2. The variables are represented by m_1, m_2 columns in the data sheet. The relationship between X and Y is expressed as canonical correlation coefficients which are tested for statistical significance. If any (at least the first) canonical coefficient is significant, then we conclude that there is a proved relationship or influence between the set of variables X and a set of variables Y.

Canonical correlation is a more general method than is pairwise and multiple correlations in Correlation module based on projections into principal components and finding a linear combination of first and second variable, which has the maximum correlation coefficient. The method provides a test of statistical significance of canonical correlation, canonical correlation coefficients, canonical variables and other results. The aim is to identify the strongest statistical relationship between groups of variables, and help users to find the real causal relationships. The result of the calculation are new pairs of univariate variables A_i, B_i. Total number of these couplesis m = min (m_1, m_2). The most important is usually the first couple

$$A_1 = a_{1,1}x_1 + a_{2,1}x_2 + \ldots + a_{m1,1}x_{m1}$$
$$B_1 = b_{1,1}y_1 + b_{2,1}y_2 + \ldots + b_{m2,1}y_{m2}$$

that is established, so that between these canonical variables A_1, B_1 was the maximum possible pair correlation coefficient. Every other canonical pair A_i, B_i is always orthogonal to all the previous canonical variables,

$$A_i^T.A_j = 0;\ B_i^T.B_j = 0 \text{ for } i \neq j.$$

If the test confirms statistical significance of correlations, it could be concluded that there is a statistically proven relatinship between groups 1 and 2 on the specified level of significance α (usually α = 0.05). Signs of canonical correlation coefficients don’t matter.

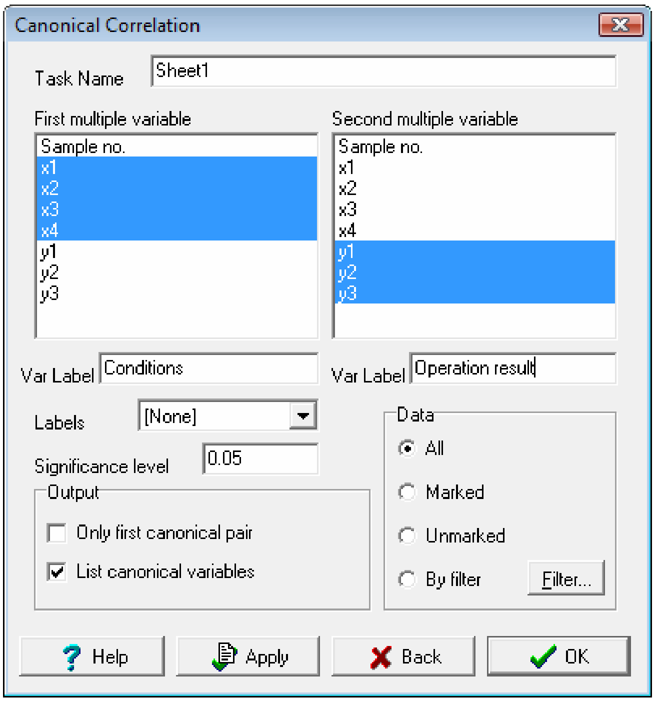
16.3.1. Data and parameters

Two multidimensional selections are analysed on the basis of data arranged in two groups of columns, as suggested in the following table. Two user-selected groups of columns usually characterized by two groups of variables, which we expect to correlate. Table 5 represents a first group of columns x1, x2, x3, x4, and a second group of columns y1, y2, y3. The number of variables m_1, m_2 in groups may be different and must be greater than 1. Here, m_1 = 4, m_2 = 3. Values in a row must always correspond to the same sample, a situation the patient, etc. All values mst be present in each row. Rows with missing values will be ignored. Typical groups parameters may be, for example, 1st Group: chemical composition, 2nd Group: physical parameters, or effect; 1st Group: feedstock parameters, 2 variable: parameters of the product; 1st group: the results of psychological tests, 2 Group: marks at school, etc.

Table 5 Data structure for canonical corelation, variable 1 = (x1, x2, x3, x4); variable 2 = (y1, y2, y3)

Sample no.	x1	x2	x3	x4	y1	y2	y3
33	8.08	2.89	500	21	6.5	28	5.24
34	8.29	4.43	600	22	6.1	32	6.51
35	8.81	3.92	600	19	5.7	33	7.91
36	8.53	3.75	700	17	5.1	38	8.15
37	9.04	3.77	600	12	3.4	32	7.02
39	7.44	2.5	500	15.5	3.8	27	6.255
44	8.83	3.46	500	14.5	4.1	28	6.555
45	7.82	3.2	600	22	4.9	33	5.94
48	8.43	3.31	500	14.5	4.1	28	6.125
15	8.02	2.9	500	21.5	5	27	5.125
16	8.91	3.08	500	21	5.2	29	6.54
17	8.95	3.14	600	17.5	4.8	33	7.855
18	8.88	3.69	600	17	4.2	32	7.64
19	7.28	4.08	600	21	5.2	32	7.51

After opening the dialog *Canonical correlations* the columns of the first and second variable are selected. The columns may not overlap! A column selected in one group may not occur in the second group. We can enter a description of the first and second group and the level of significance (usual value level of significance is 0.05, ie 5%). In addition, you can specify the contents of the output report. If the box *Only first canonical pair* is checked only the first canonical couple variables A_1, B_1. If the box *List canonical variables* is not checked the values of canonical variables will not be listed regardless of the field *Only first canonical pair*, which is advantageous when we have many rows, which would produce too long output report.



Task Name: Sheet1

First multiple variable: Sample no., **x1**, **x2**, **x3**, **x4**, y1, y2, y3

Second multiple variable: Sample no., x1, x2, x3, x4, **y1**, **y2**, **y3**

Var Label: Conditions

Var Label: Operation result

Labels: [None]

Significance level: 0.05

Output

☐ Only first canonical pair

☑ List canonical variables

Data

● All

○ Marked

○ Unmarked

○ By filter

Fig. 56 Canonical correlations dialog

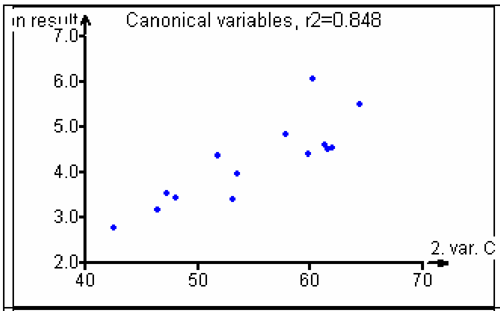
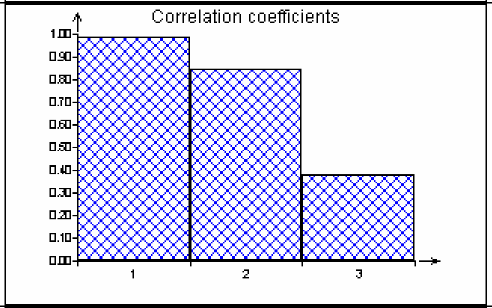
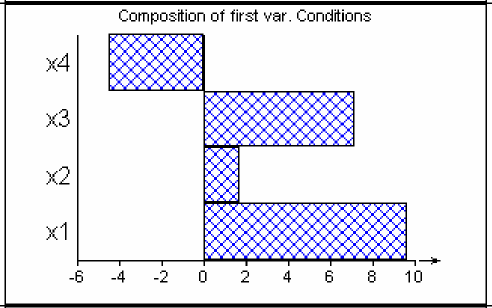
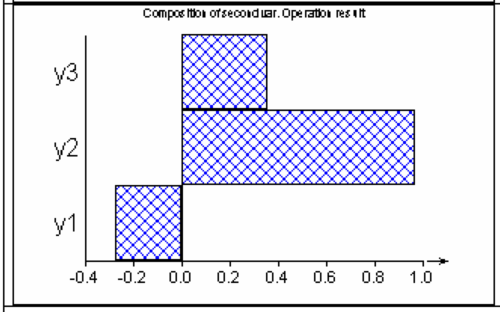
In the group *Data* we can choose a subset of data according to marked data in the data table or possibly define data by a filter. Press the *Apply* or *OK* button to run the calculation.

16.3.2. Protocol

Task name Data	Task name taken from dialog box Selected data
Basic characteristics	
First (Second) multiple variable Mean Std Deviation	Characteristics of the original variables Arithmetic averages of the columns Standard deviations of the columns
Canonical correlation coefficients Correlation *i*	Value of the i-th canonical correlation coefficient, $i = 1, \ldots, \min(m_1, m_2)$
Correlation significance	Statistical significance of the correlation coefficient at the confidence level α
Lambda Chi-squared p-value Conclusion	Test χ^2 statistic Chi-squared critical quantile p-value of the test statistic Verbal conclusion of the significance test (Significant or Insignificant)
Composition of canonical variables	Canonical coefficient a_{ij}, b_{ij}, composition of canonical variables, i-th canonical variable A_i is given by $A_i = \Sigma x_j.a_{ij}$.

First variable	Coefficients a_{ij}, for 1st canonical variable
Second variable	Coefficients b_{ij}, for 2nd canonical variable
Values of canonical variables	Values of the canonical variables A_i, B_i.

16.3.3. Graphs

	Graphic representation of all pairs of the canonical variables. The criterion of statistical significance is much tougher than for pair correlations, so even strong „looking“ correlation may not be statistically significant. The sign correlation does not matter.
	Bars chart of the absolute values of individual canonical correlation coefficients.
	Composition of the first canonical variable expressed by values $a_{1,1}$, $a_{2,1}$, …, $a_{m1,1}$
	Composition of the first canonical variable expressed by values $b_{1,1}$, $b_{2,1}$, …, $b_{m2,1}$

17. Regression

17.1. Linear regression

Menu:	QCExpert	Linear regression

Linear regression module is used to build and analyze linear regression models in their general form

$$G(y) = a_1F_1(\mathbf{x}) + a_2F_2(\mathbf{x}) + \ldots + a_mF_m(\mathbf{x}) + a_0, \qquad (1)$$

where y is a response variable, $\mathbf{x} = (x_1, x_2, \ldots x_p)$ are values of explanatory variables (written as a vector). p is the number of explanatory variables in the regression model. There are m parameters, $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ in the model. a_0 is the intercept. $F_i(.)$, i=1,…m are arbitrary functions of explanatory variables which do not involve parameters. $G(.)$ is an arbitrary function of the response variable which does not involve parameters. Individual summands $F_i(\mathbf{x})$ on the right hand side of the model equation are sometimes called model terms. Ideally, **x** is assumed to be a deterministic, i.e. non-random vector, being either purportedly set to prespecified values or its values are found out via an essentially error-free procedure. y depends on **x,** but the dependence is blurred by the presence of a random error ε. Vector of model parameters **a** can be estimated from data by various methods. Some methods are robust, some of them might not be. The (data, model, method) triplet is sometimes called the regression triplet. In order to get correct results, each of the triplet components should be given appropriate attention. Regression diagnostics and other tools offered by the QCExpert are useful in this context. There is also a wide choice of models available in the program. The user can select one of the three basic model types: simple linear model without transformation, polynomial model or general user-defined model. The selection takes place in the Linear regression dialog panel, particularly in the Transformation field:

No transformation: corresponds to a regression model of the form

$$y = a_1x_1 + a_2x_2 + \ldots + a_mx_m + a_0, \qquad (2)$$

For this model, the number of parameters, m is specified by the number of explanatory variables selected in the *Explanatory variable* window. The simplest example of such a model is a regression line, e.g.
[profit] = a_1 . [investment] + a_0,
another example, involving several explanatory variables is
[steel_strength] = a_1 . [Cr_concentration] + a_2 . [melting_time] + a_3 . [carbon_concentration] + a_0,

For instance:

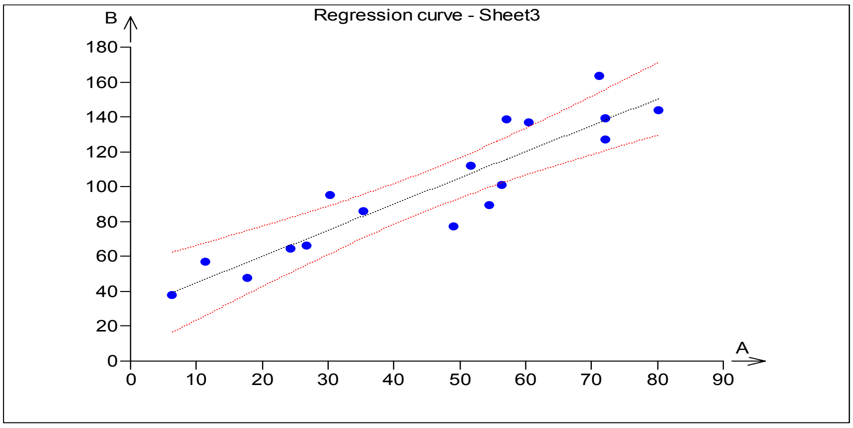


Fig. 57 Regression line

Polynomial is a model of the following form:

$$y = a_1x + a_2x^2 + \ldots + a_mx^m + a_0, \qquad (3)$$

m is degree of the polynomial. It is equal to the number of model parameters minus one. There is only one independent variable x in such a polynomial. All of its powers 1 through m appear in the model, however. Following quadratic (i.e. nonlinear) relationship can serve as an example.

[number_of_items_sold] = $a_0 + a_1$. [advertisement_costs] + a_2 . [advertisement_costs] $^2 + a_0$.

When a model involving only some powers is desired instead of the full polynomial (e.g. a model with the first and third powers only), user transformation has to be used.

For instance:

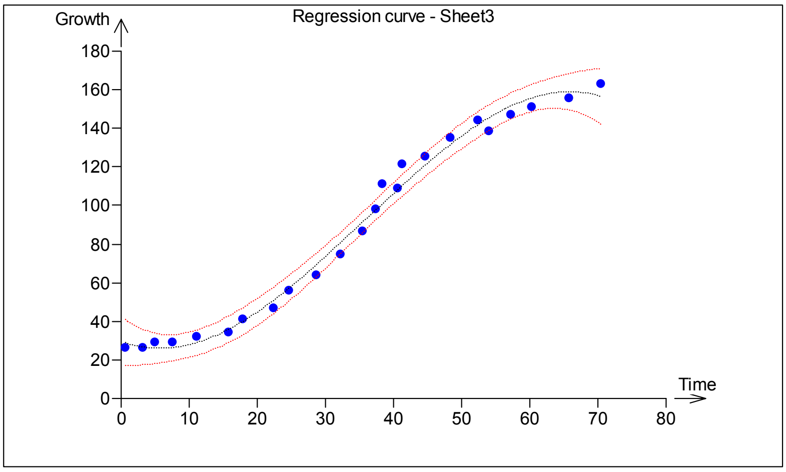


Fig. 58 Polynomial of the 3rd order

QC.Expert also allows polynomial transformation of several explanatory variables. In more than one independent variables are selected, polynomial transformation will allow for the full 2nd degree Taylor series, which is often used to fit and optimize response surfaces. Results n the protocol then include type of the stationary point (possibly optimum) and parameters of the regression model. Details are described below in paragraph 17.1.1.

User transformation: allows you to specify a linear model. It is a general formulation which includes the two special cases discussed previously (*without transformation* and *polynomial*). Earlier defined models can be selected using the appropriate selection window. A new model is specified upon choosing *Model...* after clicking the *User...* button. This action opens model specification dialog panel, see later). Individual transformation functions F_1, F_2, …, and/or G, see below can be specified there. User transformation can be used when linearizing the exponential model $y = A$. exp($B\,x$) to the the form ln (y) = $a + b\,x$, where a = ln A, $b = B$. G = ln(y), $F_I = x$, in this case, . Another example is

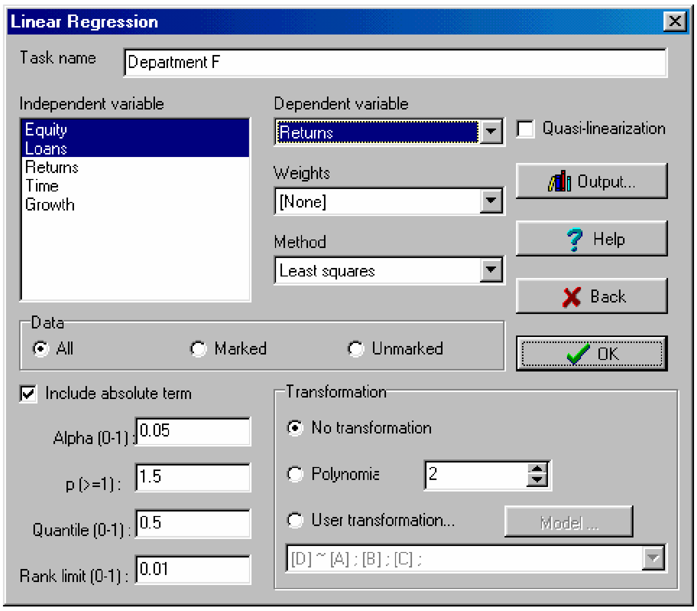
1 / [consumption] = a_1 [X1] + a_2.[X1]$^{1/2}$ + a_3 [X1].[X2] + a_4 ln[X2] + a_0,

where
G = 1/[consumption],
F_1 = [X1],
F_2 = sqrt[X2],
F_3 = [X1][X2],
F_4 = ln [X2],
[consumption] is the response variable, [X1] and [X2] are explanatory variables.
It is important to keep in mind that the transformations can involve only explanatory/response variables, they must not involve parameters a_i. Models like $y = a_1 . x^{a2} + a_0$, or $y = a_0 + a_1$. exp($a_2\,x$) cannot be specified through such transformations.

One of various (robust or classical) estimation method can be selected. This should be done in accord with the general character of the data, error behavior, or other considerations. Individual data points can be weighted differently upon inputting user-supplied weights. QCExpert allows you to inspect various potential models corresponding to all combinations of model terms by fitting all possible regression subsets. This can help you to find the most important variables, which should be included in the model, and/or their transformations.

17.1.1. Data and parameters

Unknown parameters $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ are estimated from data in the current data sheet. Each of its columns corresponds to a variable. Column header contains a variable name. The way, in which variables are selected depends on a requested transformation.



Linear Regression

Task name: Department F

Independent variable: Equity, Loans, Returns, Time, Growth

Dependent variable: Returns

□ Quasi-linearization

Weights: [None]

Method: Least squares

Data: ● All ○ Marked ○ Unmarked

☑ Include absolute term

Alpha (0-1) : 0.05

p (>=1) : 1.5

Quantile (0-1) : 0.5

Rank limit (0-1) : 0.01

Transformation: ● No transformation ○ Polynomia 2 ○ User transformation...

[D] ~ [A] ; [B] ; [C] ;

Fig. 59 Linear regression dialog panel

No transformation: when the linear regression dialog panel is open for the first time, the last column is automatically selected as the response variable column, any other data columns are selected as explanatory variables implicitly. Other choices can be made using mouse, Shift and Ctrl keys. Number of explanatory variables is not restricted at all. Exactly one response variable has to be selected. Any model specified in this way has a general form (2). The dependent variable column corresponds to y and the explanatory variables data columns correspond to $x_1, x_2, \ldots$

Polynomial: requested polynomial transformation (3) amounts to fitting a polynomial curve through data. Such a choice involves only one response and one explanatory variable. Requested polynomial order p has to be specified. Lower order polynomials are strongly preferred in general. Higher order polynomial models can be numerically unstable. Their statistical properties might be bad as well, resulting into high variability of parameter estimates and poor prediction abilities. Choice of the polynomial degree might be guided by the APSR (all possible subsets regression) results - see later. When Polynomial transformation is selected, all powers 1 through p are forced into the model. When only some of the powers are to be included in the model (e.g. 1st , 3rd, 5th), User transformation has to be selected, where each of the model terms is specified explicitly.

If more than one explanatory variable is selected, then the *Polynomial order* field stays inactive and a full quadratic model is constructed automatically. The model includes all pure terms terms of up to second order and all crossproducts of linear terms, so that its terms involve

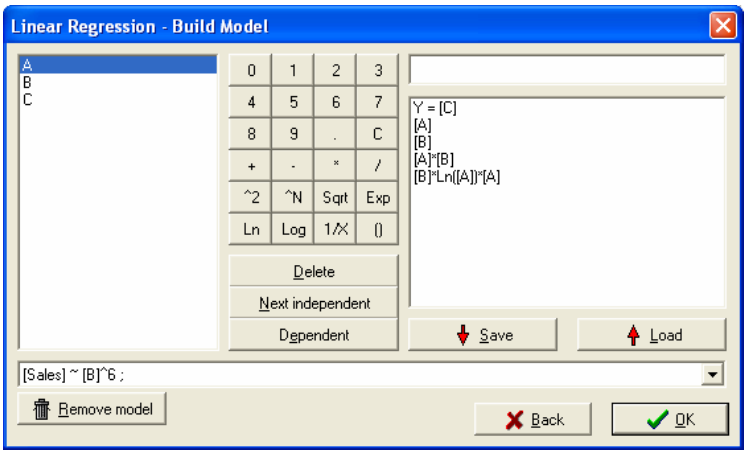
$$x_1, x_2, \ldots, x_m, x_1x_2, x_1x_3, \ldots x_ix_j, \ldots, x_mx_{m\text{-}1}, x_1^{\,2}, x_2^{\,2}, \ldots, x_m^{\,2} \qquad (4)$$

When our variables are named A, B, C, the full quadratic model with intercept will be of the form

$$y = a_0 + a_1.\mathrm{A} + a_2.\mathrm{B} + a_3.\mathrm{C} + a_4.\mathrm{AB} + a_5.\mathrm{AC} + a_6.\mathrm{BC} + a_7.\mathrm{A}^2 + a_8.\mathrm{B}^2 + a_9.\mathrm{C}^2$$

This model corresponds to an *m* dimensional quadratic surface. Such a surface can have one extreme point (minimum or maximum), corresponding to the minimum or maximum expected response. The model can be fitted in the Response surface module as well, although the output is much less detailed there (e.g. without diagnostics). One has to keep in mind that the number of data points should be larger (ideally much larger) than the number of model terms. With *m* different explanatory variables, the full quadratic model has $1.5m+m^2/2+1$ model terms. For instance, m=10 gives 66. Detailed description of the quadratic model output can be found in the *Response surface methodology* chapter.

User transformation: after selecting User transformation, the model specification panel is opened upon clicking the *Model…* button. If any models were specified previously, one of them can be selected without going through the *Model specification* panel. In such case, it is necessary to make sure that variable names in the model and in current data sheet agree. A new model is formulated using the *Model specification* panel. The current data sheet variables list is displayed in the left part of the *Model specification* panel. Only the listed variables can be used when specifying a regression model. Model input line appears in the upper right hand corner of the panel. The window located just below the input line lists dependent variable together with model terms included already.



```
Y = [C]
[A]
[B]
[A]*[B]
[B]*Ln([A])*[A]
```

```
[Sales] ~ [B]^6 ;
```

Fig. 60 Model specification dialog panel

Any function of one or more explanatory variables can be a model term. There has to be only one response variable. It is either directly the variable selected by the user from the current data sheet variables or any function of it. The response variable is denoted as Y=. For instance, let us define the model $\ln(y) = A.x + B.x^2 + C.x^{-1}$ model. The intercept can be included in the model either by checking the *Intercept*, or by including 1 (number one) as an explanatory variable during the model specification. Only one of the two possibilities should be used. (When both of them are used, an error caused by model overspecification results.)

Model specification instructions:

Double click on a variable name in the available variables list copies the name to the model input line. Variable name is always enclosed in square brackets. Function buttons can be useful when specifying more complicated models. Highlight a part of the model input line and clicking a function button subsequently to apply the function on the highlighted part as an argument. For instance, expression ln([*x*]+1) can be assembled in the following way: double click on the variable *x* (there has to be a column of this name in the current data sheet): **[x]**; write + 1 manually; highlight whole expression: **[x]+1**; click the *Ln* button, resulting into: **ln([x]+1)**. Application of ^*2*, ^*A*, *Sqrt*, *Exp*, *Log*, *1/X*, *( )* is similar. The *C* button erases model input line. Other functions have to be inputted manually (writing their name in the model input line). Available functions are listed in the table below. After specifying a term completely, another term is added by clicking on the *Next explanatory* button. The response variable is included by clicking on the *Response* button. A highlighted model term is erased when clicking the *Erase* button. There has to be exactly one response variable in any model. When finished with specification, the model is saved by the *Save* button. Then, it automatically appears in the list of previously specified models located in the bottom part of the panel. The *Read* button reads in a model from the set of previously defined models. Its terms can be edited subsequently. The *Erase model* deletes a selected model from the model list. Warning: the operation is irreversible! *OK* button finishes model specification.

You can select previously specified models from the list directly in the *Linear regression* dialog panel without opening the *Model specification* panel, be careful however: variable names in the model and in the current data sheet have to agree.

Table 6 List of available functions

Function	Value, description, restrictions	Syntax
Basic binary operators		
+	Summation	x+y
–	Subtraction	x–y
∗	Multiplication	x*y
/	Division; y ≠ 0	x/y
^	Power; for a negative x, the INTPOWER function has to be used	x^y
DIV	Integer divisor; y ≠ 0	x DIV y
MOD	Modulo; y ≠ 0	x MOD y
Functions		
TAN	Tan; x ≠ *n*π+π/2	tan(x)
SIN	Sine	sin(x)
COS	Cosine	cos(x)
SINH	Hyperbolic sine	sinh(x)
COSH	Hyperbolic cosine	cosh(x)
ARCTAN	Arc tan	arctan(x)
COTAN	Cotan; x ≠ *n*π	cotan(x)
EXP	Exponential function, base e	exp(x)
LN	Natural logarithm; x > 0	ln(x)
LOG	Decadic logarithm; x > 0	log(x)
LOG2	Base 2 logarithm; x > 0	log2(x)
SQR	Square	Sqr(x)
SQRT	Square root; x ≥ 0	Sqrt(x)
ABS	Absolute value (abs(0) = 0)	Abs(x)
TRUNC	Truncation	Trunc(x)
INT	Truncation	int(x)
CEIL	Ceiling	Ceil(x)
FLOOR	Floor	Floor(x)
HEAV	Heaviside function (indicator of a nonnegative argument, 0 for a negative argument, 1 else)	Heav(x)
SIGN	Sign (-1 for a negative argument, 0 for 0, 1 for a positive argument)	Sign(x)
ZERO	Indicator of zero (1 for zero argument, 0 else)	Zero(x)
RND	Random number from a uniform distribution on (0,x); x > 0	Rnd(100)

RANDOM	Random number from (0,1) uniform distribution. Even though it does not use any argument, a dummy argument has to be specified.	Random(0)
–	(Unary) minus before an expression	–x

Functions with two arguments		
MAX	Maximum	MAX(x,y)
MIN	Minimum	MIN(x,0)
INTPOWER	The first argument raised to the power specified by the second argument, the second argument is integer valued; it can be used even for a negative x	INTPOWER(x, –2)
LOGN	Logarithm of the first argument, using the second argument as base; x > 0, y > 1	Logn(x,3)

Relations		
GT	Greater than; if x>y then it returns 1, 0 else	GT(x,y)
LT	Less than; if x<y then it returns 1, 0 else	LT(x,y)
EQ	Equal; if x = y then it returns 1, 0 else	EQ(x,y)
NE	Not equal; if x ≠ y then it returns 1, 0 else	NE(x,y)
GE	Greater or equal; if x ≥ y 1, 0 else	GE(x,y)
LE	Less or equal; if x ≤ y 1, 0 else	LE(x,y)

Function names can be written in lowercase or uppercase letters. Relations result in 0 or 1, which can be used when specifying discontinuous functions, like le(x,0)*1+gt(x,0)*5, see also the Nonlinear regression chapter.

Further details on the Linear regression dialog panel.

Task name: A project identification (one line). It appears in the protocol and graphic output headers.

Independent variable: Select one or more explanatory variables. Use mouse (dragging, Shift-click or Ctrl-click) when selecting more than one variable. This item is not active when User transformation is selected – the variables are specified in the Model specification panel.

Dependent variable: Select one data column as a response variable. This item is not active when User transformation is selected – the variables are specified in the Model specification panel.

Intercept: When checking this option, intercept is included in the model. Do not use it when the intercept is already entered manually as the unit explanatory variable!

Alpha (0 – 1): Significance level, α which will be used for all tests and confidence intervals. It has to be larger than 0 and smaller than 1. α=0.05 is the default.

p (p ≥ 1): Coefficient *p* for L_p regression. The value is used only when the *Lp-regression* is selected (see later). *p*=1 corresponds to the least absolute differences method, *p*=2 corresponds to the least squares, $p \rightarrow \infty$ (*p*≈10 is typically taken in practice) corresponds to minimization of maximum error (minimax). When 1≤p<2 is selected, the resulting estimates are rather robust against outliers. *p*=1.5 is the default.

Quantile (0 – 1): Probability value specifying a particular quantile regression. It is used only when the *Quantile regression* is selected (see later). It has to be larger than 0 and smaller than 1. 0.5 is the default, corresponding to the least absolute differences method.

Rank limit (0 – 1): It is a restriction parameter related to the Rank correction method. Zero parameter value corresponds to the usual method of least squares (OLS). When a positive parameter is selected, the components related to small eigenvalues of the $\mathbf{X}^T\mathbf{X}$ matrix are suppressed, resulting into biased parameter estimates with smaller variance than usual estimates. Such estimates are less sensitive to an ill conditioned $\mathbf{X}^T\mathbf{X}$ matrix, which occurs typically e.g. when fitting a high degree polynomial models (see later). Value of at most 0.1 is recommended.

Quasi-linearization: When this selection is checked, quazilinearization is applied. It is useful when User transformation is selected and the response variable is nonlinearly related to one of the explanatory variables. This occurs for instance for the model `ln(y) ~ [x]; [x]^2`. Nonlinear transformation $G(y)$ linearizes the model, but it deforms error distribution and biases parameter estimates. The quasilinearization technique can eliminate the bias, to some extent. The quasilinearization is based on the idea of introducing weights $w_i=[\partial G(y)/\partial y]^{-1}$.

Weights: Select a data column w_i, you want to serve as a weighting variable. Alternatively, you van select one of the prespecified weight types: [None], [Y], or [1/Y]. The [1/Y] weights are used when the relative error for the response is constant. The weights must not be negative. Zero weight results in dropping the corresponding line from the analysis. The default is [None] – all weights are equal to one. When variances of the response in different data rows are known, say $\mathbf{S}$ = diag(w_1^{-2}, w_2^{-2}, . . ., w_n^{-2}) than the weights should be proportional to the square roots of reciprocal variances.

Method: Select one of computational methods. The selection should depend on the nature of the analyzed data.

Least squares: The basic and commonly used method. It works fine when errors are normally distributed, data are free from gross errors in both response and explanatory variables and the problem is not ill conditioned due to an unfavorable design matrix composition. The method may fail badly when some of these conditions are not satisfied.

Rational rank: A method commonly used for instance for higher order polynomials, full second order polynomials and other cases when collinearity is a problem (explanatory variables are "correlated"). Detected collinearity is indicated in the QCExpert protocol (in the *Multicollinearity paragraph*). The extent to which the rank is corrected is given by the *Restriction* parameter (value of at most 0.1 is recommended). When a positive parameter is selected, the components related to small eigenvalues of the $\mathbf{X}^T\mathbf{X}$ matrix are suppressed, resulting into biased parameter estimates with smaller variance than usual estimates. Such estimates are less sensitive to an ill conditoned $\mathbf{X}^T\mathbf{X}$ matrix.

Quantile regression: Quantile regression method, using the quantile α specified in the *Quantile* field. It corresponds to the model in which probability of the event (linear predictor<Y) is α. The method is advantageous when we are not interested in modelling changes in expected value as a function of explanatory variables, but rather in modelling changes of a more extreme tendency of the distribution, specified by a quantile. For instance, one might be interested in "minimal" strength and choose α=0.05, or in "maximal" pollution and choose α=0.95, etc. The computation method is iterative (weighted least squares method is used iteratively). Computation time depends on the number of data points. Number of data points (*n)* should be larger for more extreme quantiles (i.e. for α close to 0 or 1). *n* should be larger than 5/min(α,1−α). α=0.5 gives median regression, corresponding to the Lp

regression for p=1, i.e. to the method of the least absolute differences. Generally, the returned solution is less precise for small or large α. In some cases, the solution might not be unique.

Lp-regression: This method is based on minimization of the sum $\Sigma|e_i|^p$, amounting to a generalization of the least squares method based on Σe_i^2 minimization. Parameter p is entered in the p field *($p \geq 1$)*. p=1 gives median regression, i.e. the method of the least absolute differences. It is very useful for data whose distribution is similar to the Laplace distribution. p=2 corresponds to the least squares regression, $p \rightarrow \infty$ ($p \approx 10$ is typically selected in practice) corresponds to minimization of maximum error (minimax). It is very sensitive to outliers and it should be used only when the errors are uniformly distributed. When $1 \leq p < 2$ is selected, the resulting estimates are rather robust against outliers. p=1.5 is the default. Solution to the Lp regression might not be unique. Iterative randomized simplex optimization method is used for computations.

Least median: A modern, highly robust regression (often called LMS) method based on minimization of the median of squared differences. Iterative randomized simplex optimization method is used for computations.

IRWLS exp(-e): A robust regression method producing M-type estimates. It is based on iterative minimization of sum of squared standardized residuals $w(e_{ni})$, using weights $w(e) = \exp(-e)$. *Iteratively Re-Weighted Least Squares* are used for computations.

M-estimates, Welsch: A robust regression method producing M-type estimates. It is based on iterative minimization of sum of squared standardized residuals $w(e_{ni})$, using weights $w(e) = \exp(-e^2)$. *Iteratively Re-Weighted Least Squares* are used for computations.

BIR: Bounded influence regression. This method is robust not only against response variable outliers but also against influential observations (influence is connected to dependent variables values). It is this second robustness feature that distinguishes the method from previously discussed robust techniques. It might be useful for polynomial models when trying to suppress influence of extreme **x** points (low and high) on the fit. *Iteratively Re-Weighted Least Squares* are again used for computations.

Stepwise All: All possible subsets regression (*APSR*). This method is a useful tool for selection of important variables to be included in a regression model. The models can be compared by one of the following three measures: F-statistic (FIS), Akaike's information criterion (AIC) and mean squared error of prediction (MEP). When *APSR* is invoked, QCExpert explores all combinations of variables from the set of potential explanatory variables (model terms) supplied by the user. A regression model is fitted for each of the combinations. The results are outputted both to the protocol and to a special output data sheet *APSR* (the sheet is created automatically). The text output is further enhanced by three plots in the graph window. Warning! Maximum number of model terms allowed is 12 without the intercept, or 13, including the intercept. The restriction is common for polynomial, full quadratic and general models. Since the results are outputted to a data sheet, the restriction comes from the maximum number of data sheet rows allowed by the QCExpert. The number of all possible models gets large very quickly. For m potential model terms (including the intercept), there are $2^m - 1$ possible regressions. Ordinary least squares method is used for all computations. For further details, see the Protocol and Graphical output paragraphs.

Data: Here, you can specify which part of data you want to use in computations. You can specify all data rows, selected rows only, or the rows which are not selected.

Transformation: Data transformation is defined here, see the previous paragraphs discussing model specification.

Output: Invokes a panel allowing you to customize some of the output features, see the next paragraph for details.

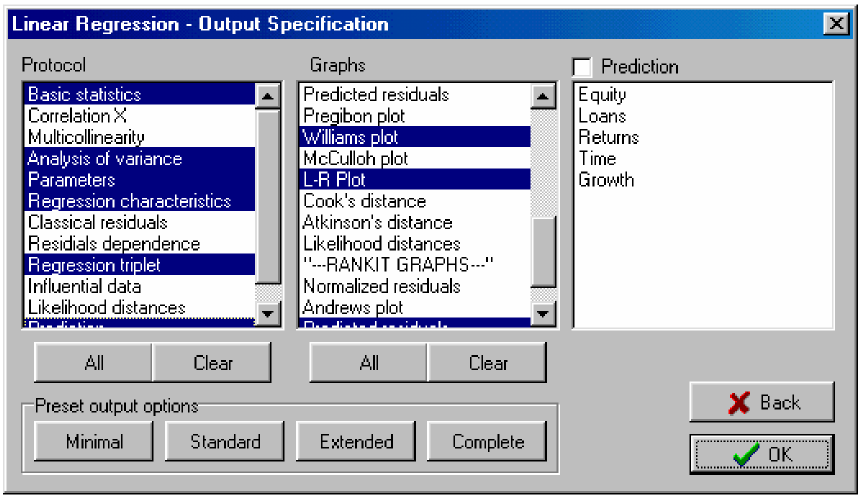
Help: Invokes help screen.

Cancel: Cancels immediately preceding operation.

OK: Runs the computations.

Output

The panel is invoked by the *Output* button in the Linear regression panel. Some of the output features can be customized here, specifying text and/or graphical items requested. There are three lists in the panel: *Protocols* (protocol items), *Graphical output* (plots or groups of plots), *Prediction* (predictions are requested for variables selected here). The *Prediction* list variables are used only when the *Prediction* item is checked. Shortcut buttons *Minimal*, *Standard*, *Extended*, *Complete*, *All*, *None* are available.



Linear Regression - Output Specification

Protocol

- Basic statistics
- Correlation X
- Multicollinearity
- Analysis of variance
- Parameters
- Regression characteristics
- Classical residuals
- Residuals dependence
- Regression triplet
- Influential data
- Likelihood distances

Graphs

- Predicted residuals
- Pregibon plot
- Williams plot
- McCulloh plot
- L-R Plot
- Cook's distance
- Atkinson's distance
- Likelihood distances
- "---RANKIT GRAPHS---"
- Normalized residuals
- Andrews plot

□ Prediction

- Equity
- Loans
- Returns
- Time
- Growth

Fig. 61 Output dialog panel

Size of some output items depends on the number of data points. Keep in mind that the output can become rather large and difficult to read when all items are requested for large datasets. Next, we will describe contents of various individual output items, both text and graphical.

Protocol **field**

Summary statistics: Basic summary statistics: mean, standard deviation. Correlation coefficient and result of its test are produced for all pairs response-explanatory variable;

Correlation X: Pairwise correlation coefficients and results of their tests for all possible explanatory variables pairs;

Multicollinearity: Eigenvalues related to the design matrix (matrix of explanatory variables), condition number κ, variance inflation factor (VIF), multiple correlation coefficients;

ANOVA: overall (arithmetic) mean of the response variable, sums of squares, mean squares for the following variability sources: (corrected) total, model, residuals (error). Results of the overall F-test for the model, observed F-statistic value, $F(1\text{-}\alpha, m-1, n-m)$ quantile;

Parameters: regression parameters estimates, followed by estimates of their standard errors, individual confidence intervals and results of their tests;

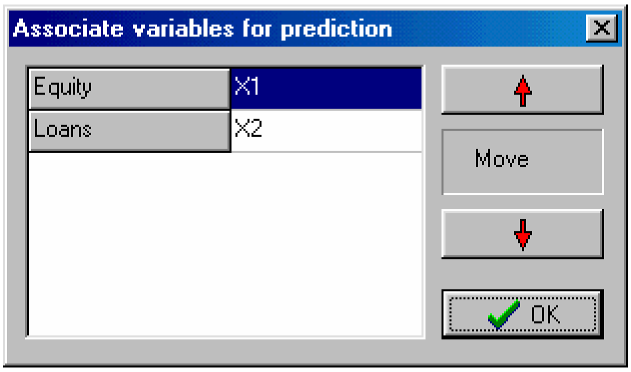
Characteristics: Multiple correlation coefficient R, coefficient of determination R^2, R_p, mean squared prediction error (*MEP)*, Akaike information criterion (*AIC)*;

Residuals: observed *Y*, predicted *Y*, standard deviation of Y, residual standard deviation, residual variance, residual sum of squares, residuals, weights, mean of absolute residuals, skewness and curtosis computed from residuals;
Residual dependence: Wald test for autocorrelation, Durbin-Watson test for autocorrelation, and sign test for lack of residual independence;
Regression triplet: Fisher-Snedecor test for the model, Scott's multicollinearity criterion, Cook-Weisberg test for heteroscedasticity, Jarque-Berr test for normality, tests for dependence;
Influential data: standard residuals, jackknife residuals, predicted residuals, projection matrix (i.e. hat matrix, **H**) diagonal, extended hat matrix ($\mathbf{H}^*$) diagonal, Cook's distance, Atkinson's distance, Andrews-Pregibon statistic, assessment of individual datapoints influence upon prediction, parameter estimates *LD*(*b*), variance *LD*(*s*), total influence *LD*(*b*,*s*);
Likelihood-related influence measure: assessment of individual datapoints influence upon parameter estimates *LD*(*b*), variance *LD*(*s*), total influence *LD*(*b*,*s*);
Prediction: Predictor values. Predictions and their confidence intervals.

Graphical output field
There are five groups of items in this field:
Regression curve;
Residuals: *Y-predicted values, Residuals vs. Predicted, Abs. residuals, Squared residuals, residual QQ-plot, Autocorrelation, Heteroscedasticity, Jackknife residuals, Predicted residuals;*
Partial regression plots: *Partial regression plots, Partial residual plots;*
Influential data: *Projection matrix, Predicted residuals, Pregibon, Williams, McCulloh, L-R Plot, Cook's D, Atkinson's distance;*
Q-Q plots: *Standardized residuals, Andrews plot, Predicted residuals, Jackknife residuals.*

Prediction field

You can select variables to be used as predictors of the response. Names of the predictors are arbitrary, but their number and order in which they appear must respect the regression model specification. When User transformation is selected, the *Variable association* panel is invoked (Fig. 62). There, you must associate selected predictor names listed on right to the model explanatory variable names listed on left. Predictors can have arbitrary number of rows (corresponding to points in which the predictions are requested). Explanatory variables used in model fitting can be used as predictors.



Associate variables for prediction

Equity	X1
Loans	X2

Fig. 62 Variable association panel

All: Selects all items
Nothing: Cancels previous selection
Minimal, Standard, Extended, Complete: Selects protocol and graphical output items according to the rules listed in the following table.

Table 7 Automatic protocol item selection

Item	Minimal	Standard	Extended	Complete
Summary statistics		o	o	o
Correlation X			o	o
Multicollinearity			o	o
ANOVA		o	o	o
Parameters	o	o	o	o
Characteristics	o	o	o	o
Residuals			o*	o*
Residual dependence				o
Regression triplet		o	o	o
Influential data			o*	o*
Likelihood related influence measure				o*
Prediction	o**	o**	o**	o**

* Size of this item depends on the number of data points!

** Depends on how the *Prediction* item is set

Table 8 Automatic graphical output item selection

Item	Minimal	Standard	Extended	Complete
Regression curve	o	o	o	o
Y-prediction		o	o	o
Residuals vs. Predicted	o	o	o	o
Abs. Residuals			o	o
Squared residuals				o
Residual QQ-plot		o	o	o
Autocorrelation			o	o
Heteroscedasticity			o	o
Jackknife residuals				o
Predicted residuals				o
Partial regression plots			o	o
Partial residual plots				o
Projection matrix	o	o	o	o
Predicted residuals				o
Pregibon				o
Williams		o	o	o
McCulloh				o
L-R plot		o	o	o
Cook's D				o
Atkinson's distance				o
Standardized residuals				o
Andrews plot			o	o
Predicted residuals		o	o	o
Jackknife residuals				o

17.1.2. Protocol

Project name	Project name, as inputted in the dialog panel.
Significance level	Inputted in the dialog panel. The level is used for all tests and confidence intervals.

Quantile t(1-alpha/2,n-m)	t-distribution quantile.
Quantile F(1-alpha,m,n-m)	F-distribution quantile.
Intercept	Is intercept included in the model?
Number of data rows	Number of complete data rows containing values for all model variables.
Number of parameters	Number of model terms, including intercept and terms created by transformations. For instance, for the 3rd order polynomial, the number of terms is 4.
Method	Computation method selected by the user.
Columns used in the model	List of variables used in the regression model.
Transformation	Transformation type selected by the user.
Summary statistics	
Variable characteristics	
Variable	Explanatory variable name.
Mean	Arithmetic average.
Std. deviation	Standard deviation.
Correlation with Y	Correlation between the response variable and the explanatory variable.
Significance	p-value from the correlation coefficient test.
Paired correlations (Xi, Xj)	Paired correlation coefficients for all explanatory variables pairs.
Multicollinearity indication	
Variable	Name of the variable related to the last column, where multiple correlations are listed (it has no relation to the other part of the output since eigenvalues cannot be, in general, directly related to individual variables).
Eigenvalues	Eigenvalues of the explanatory variables correlation matrix.
Condition number, kappa	Condition number (κ_{max}) is the ratio of largest and smallest eigenvalues (it is the maximum of condition index; l-th condition index is defined as the ratio of largest eigenvalue and the l-th eigenvalue). $\kappa_{max} > 1000$ indicates a strong multicollinearity.
VIF	Variance inflation factor, VIF > 10 indicates a strong multicollinearity.
Multiple correlation	Multiple correlation coefficient between the response and all explanatory variables.
ANOVA	
Overall Y mean	Arithmetic average of the response.
Source	Source of variability in the ANOVA table.
(Corrected) total	Response variability related to the model Y = Mean of(Y).
Model	[Total] – [Error].
Error	Residual variability, not explained by the model (i.e. the error variability).
F	F-statistic for the model. It should be larger than an appropriate theoretical F quantile. If it is larger, the actual model is significantly better than the null model Y=Mean of (Y).
Quantile F (1-alpha, m-1, n-m)	F-distribution quantile.
P-value	p-value for the test, if it is smaller than a specified significance level, the

	model is claimed to be significantly better than the null model.
Conclusion	Result of the test, stated in words.
Parameter estimates	
Variable	Variable name.
Estimate	Estimate of the regression coefficient associated with the explanatory variable.
Std. error	Standard error of the regression coefficient.
Conclusion	Result of the regression coefficient test, stated in words.
P-value	p-value for the regression coefficient test. If it is smaller than a specified significance level, significance is claimed.
Lower limit	Lower limit of the confidence interval computed with the prespecified confidence level.
Upper limit	Upper limit of the confidence interval computed with the prespecified confidence level. If zero is included in the interval, the regression coefficient is not significantly different from zero.
Characteristics of the model fit	
Multiple correlation coefficient, R	Multiple correlation coefficient characterizes how closely the model fits the data. It does not necessarily express how good the model is. R cannot decrease when a new variable is included in the model (it usually increases whenever a new variable is added)!
Coefficient of determination R^2	Square of the multiple correlation coefficient.
Predicted correlation coefficient, Rp	Predicted correlation coefficient, useful in the context of data containing outliers.
Mean square error of prediction, MEP	The ith error is the difference between actual value of the ith observation and its prediction. The prediction comes from the model based on data with the ith row omitted. MEP is a sensitive indicator of some problems, like multicollinearity and outliers. It is an important characteristics of the regression model quality.
Akaike information criterion	AIC in the regression context is related to the residual sum of squares, penalized by the model size (number of explanatory variables).
Residual analysis	
Characteristic	
Y observed	Observed response value, as it appears in the current data sheet.
Y predicted	Predicted response value.
Std. error of Y	Estimated standard error of the prediction.
Raw residual	Difference between observed and predicted response value.
Residual [%Y]	Relative residual, raw residual divided by the response value.
Weights	Weights for individual observations as inputted by the user.
Residual sum of squares	Residual sum of squares cannot decrease when a new variable is included in the model (usually, it increases).
Mean of absolute residuals	Mean of absolute residuals.
Residual standard deviation	Standard deviation estimated from residuals.
Residual variance	Variance estimated from residuals.
Residual skewness	Skewness estimated from residuals.
Residual curtosis	Curtosis estimated from residuals.

Regression triplet testing	
Fisher-Snedecor overall test	Tests whether the actual model is better than the null model including only the overall mean.
F	Computed value of the F test stastistic.
Quantile F (1-alpha, m-1, n-m)	F-distribution quantile.
P-value	p-value for the test, if it is smaller than a specified significance level, the model is claimed to be significantly better than the null model.
Conclusion	Result of the test, stated in words.
Scott‘s multicollinearity criterion	Assessment of multicollinearity („dependence“) among explanatory variables. Severe collinearity can inflate regression coefficient variances substantially.
SC criterion	Computed test statistic.
Conclusion	Result of the test, stated in words.
Cook-Weisberg test for heteroscedasticity	Tests whether the error variance is constant across values of the explanatory variables. When the heteroscedasticity is detected, use of appropriate weights should be considered.
CW criterion	Computed test statistics.
Quantile Chi^2(1-alpha,1)	χ^2-distribution quantile.
P-value	p-value for the test, if it is smaller than a prespecified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Jarque-Berr test for normality	Test for error normality based on residuals.
JB criterion	Computed test statistic.
Quantile Chi^2(1-alpha,2)	χ^2-distribution quantile.
P-value	p-value for the test, if it is smaller than a prespecified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Wald test for autocorrelation	Test for autocorrelation among errors. It is based on residuals
WA criterion	Computed test statistic.
Quantile Chi^2(1-alfa,1)	χ^2-distribution quantile.
P-value	p-value for the test, if it is smaller than a prespecified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Durbin-Watson test for autocorrelation	Test for autocorrelation among errors.
DW criterion	Computed test statistic.
Conclusion	Result of the test, stated in words.
Sign test	A nonparametric test for residual dependence. It can detect some of the model inadequacies.
Sg criterion	Computed test statistic.

Quantile N(1-alpha/2)	Normal distribution quantile.
P-value	p-value for the test, if it is smaller than a prespecified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Influence measures	
A. Residual analysis	
Characteristic	
Standardized	It is sometimes called the studentized residual. Raw residual divided by its standard error s_r.sqrt(1-H_{ii}). s_r is the residual standard deviation.
Jackknife	Jackknife residual. It is similar to the Standardized residual. Instead of s_r, the residual standard deviation for the model based on data with i-th row deleted is used for the i-th residual. This type of residual is more sensitive to outliers.
Predicted	Predicted residual, difference between the *i*-th response value and prediction obtained from the model based on data with the i-th row deleted. This type of residual is more sensitive to outliers.
Diag(Hii)	Diagonal elements of the projection matrix. A large value indicates a data point that can potentially have a high influence upon the regression estimates. Sum of the H_{ii}'s is equal to the number of parameters in the model. Potentially influential points are marked in red.
Diag(H*ii)	Diagonal elements of the H^* matrix. The matrix is obtained when the design matrix (i.e. the matrix containing explanatory variables columns) is augmented with the response variable column. A large value indicates a data point that can potentially have a high influence upon the regression estimates. Sum of the H^*_{ii}'s is equal to the number of parameters in the model plus one. Potentially influential points are marked in red.
Cook's D	Cook's distance measures influence of the i-th data point upon the regression estimates. It combines measure of potential influence with the assessment of whether the point is actually an outlier. Influential points are marked in red.
B. Influence analysis	
Characteristic	
Atkinson's statistic	Atkinson's modification of Cook's D (1985), both characteristics yield similar results usually. Influential points are marked in red.
Andrews-Pregibon statistic	Andrews-Pregibon statistic measures influence that individual data points have on the variance of the regression parameters (volume of the confidence ellipsoid). Influential points are marked in red.
Y^ influence	Relative influence of individual data points upon prediction. Influential points are marked in red.
Parameter influence, LD(b)	Relative influence of individual data points upon parameter estimates. Influential points are marked in red.
Variance influence, LD(s)	Relative influence of individual data points upon residual variance. Influential points are marked in red.
Total influence, LD(b,s)	Simultaneous influence of individual data points upon parameter estimates and variance. Influential points are marked in red.
Prediction	
Predictor value	Values of all model terms. The intercept is represented by the column of ones.
Prediction	Predicted value based on the fitted model.
Lower limit	Lower limit of the confidence interval for the predicted mean, computed for a prespecified confidence coefficient α.

Upper limit	Upper limit of the confidence interval for the predicted mean, computed for a prespecified confidence coefficient α.

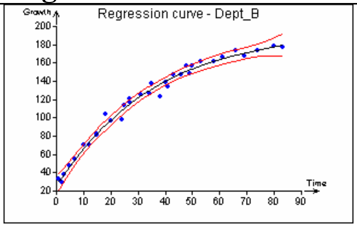
APSR regression protocol

APSR (*all possible subsets regression*) helps to find the best model according to one of the following three criteria: F-statistic, Akaike‘s information criterion (AIC) or MEP (mean squared prediction error). The APSR procedure fits all possible model terms combinations. The results are outputted both to the protocol and to a special output data sheet *APSR* (created automatically). For each possible combination of model terms, the protocol contains a paragraph indicating which terms were actually used and values of the three criteria. To save space, each of the model terms is coded by a short alphanumeric code (instead of its actual name which can be rather long and complicated). These codes are then used for each of the subsets description. The model which is the best in terms of a particular criterion can be found easily by sorting the *APSR* data sheet rows according to the criterion. Before sorting, all columns of the *APSR* sheet have to be selected, see QC.Expert – Sort. Alternatively, the point with the best value of a particular criterion can be found on the plot (part of the output, see the next paragraph, Graphical output) and selected there. A good model should have large value of F, small AIC value and small MEP value. Each of the criteria can favor different models. It is generally recommended to explore several models corresponding to very good values of a particular criteria (not only the model selected as the best). One should also keep in mind a somewhat different nature of the three criteria when interpreting *APSR* results. F is the F statistic involved in the usual F test, Akaike‘s criterion $AIC = n.\ln(RSS/n) + 2.m$ judges residual sum of squares together with a model size (the number of model terms) penalization. It was derived under much more general circumstances from information theory principles. *MEP* judges model‘s prediction abilities. There is no universally „best“ model. Selection of the model should be led by the purposes which it is intended for and subject matter knowledge of the modelled situation.

Selected columns	Variables which are considered as potential model terms. Each of them is assigned a simple code to save space and keep the output easily readable.
Model comparison	A copy of this table is saved to an automatically created data *APSR* sheet. The sheet output can be sorted according to various criteria ((*Menu – QCExpert – Sort*). Various models can be also selected graphically. The *Protocol* window output cannot be manipulated with. Output contains columns with values of the F, AIC, MEP criteria, as well as the residual sum of squares (SSE). Warning: SSE might not directly express how good the model is! The largest model has always the smallest SSE.

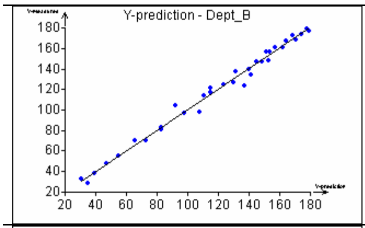
17.1.3. Graphical output

Regression curve

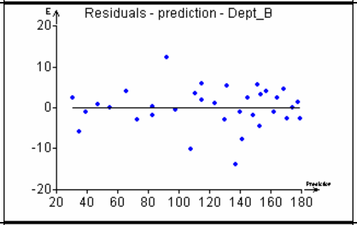


This plot is *not* produced when the model contains more than one explanatory variable. When only one explanatory variable appears in the model, the plot displays the regression curve. Red curves show the confidence band around the regression curve, computed for a prespecified confidence coefficient. It should be noted that the confidence band is realistic only when the fitted model is (approximately) correct. This is even more important when predictions further from bulk of available data points are considered. Details of the plot can be inspected upon zooming part of it. The regression curve can be inspected even outside of the interval containing explanatory variable values actually used in model fitting by inverse zooming.

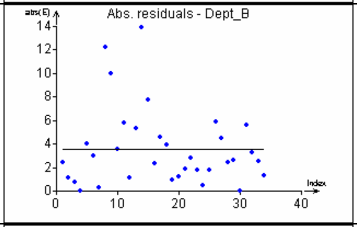
Residuals



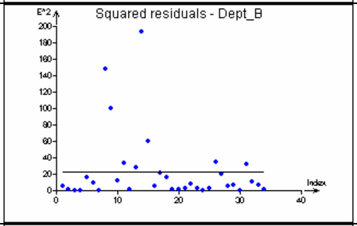
The plot shows how closely the model fits data. Predicted response values are plotted on the *X* axis, while observed response values are plotted on the *Y* axis. Vertical difference between a point and the line corresponds to a residual.



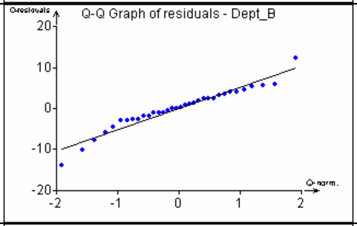
Standardized residuals plot. Predicted response is plotted on the *X* axis, while the standardized residuals are plotted on the *Y* axis. Horizontal line corresponds to the mean of residuals. When ordinary least squares are used to fit a model including intercept, the residual mean is necessarily zero.



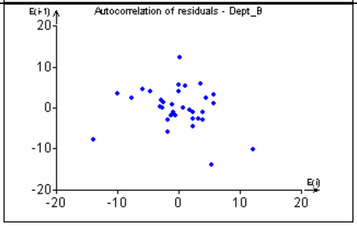
Absolute residuals. The order in which a particular data point appears in the dataset is plotted on the *X* axis. The horizontal line corresponds to the mean absolute residual.



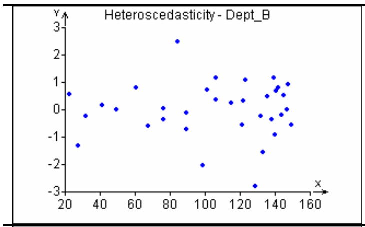
Squared residuals. The order in which a particular data point appears in the dataset is plotted on the *X* axis. The horizontal line corresponds to the mean squared residual (i.e. mean squared error estimate).



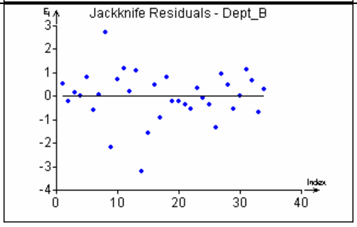
Q-Q plot for residual normality check. Approximately normally distributed (gaussian) residuals should plot close to the line. Note that the ordinary least squares tends to enhance normal appearance of the residuals (so called supernormality effect). When in doubt, one should check also the residual Q-Q plot based on some robust method.



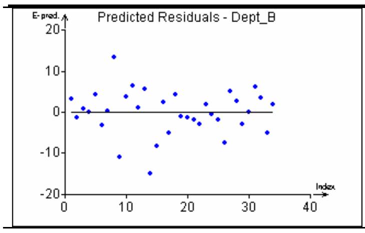
Graphical check for the first order autocorrealation in residuals. The *i* -th residual is plotted on the *X* axis, while the (*i*-)-th is plotted on the *Y* axis. When the point cloud suggests a positive slope, positive 1-st order autocorrelation is suspected. Negative slope suggests negative autocorrelation. An autocorrelation in the residuals might not always be connected to the autocorelation in errors. Residuals tend to be somewhat correlated even if the true errors are not.



Graphical check for heteroscedasticity (error variance depends on explanatory variable(s)). A non-rectangular shape of the point cloud suggests a heteroscedasticity (e.g. a fan shape).

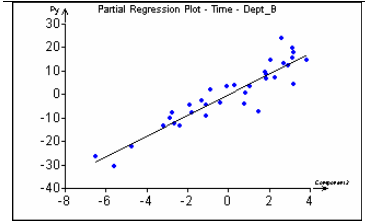


Jackknife residuals (see the Protocol paragraph) are much more sensitive to outliers in the response variable than raw residuals. Even the jackknifed residuals may fail to detect a cluster of several outliers (they mask each other).

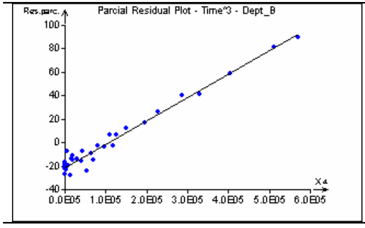


Predicted residuals are much more sensitive to outliers than the raw residuals. Even the predicted residuals may fail to detect a cluster of several outliers (they mask each other).

Partial regression plots

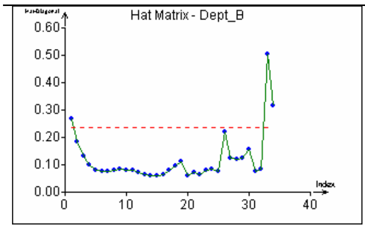


Partial regression plot displays relationship between the response and a given explanatory variable (a single model term) after the relationship has been cleared for a possible confounding caused by other variables in the model. Slope of the line corresponds to the regression coefficient for the variable in the complete model. Closeness of the linear fit on the plot is related to the significance test in the complete model.

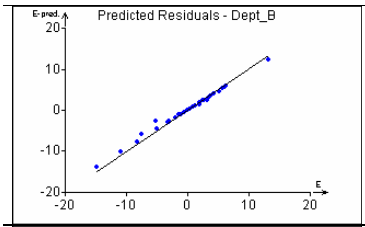


Partial residual plot. It is a modification of the partial regression plot. Nonlinear nature of the plot suggests that a term that is nonlinear in the variable just explored should be added to the model (e.g. a higher power of the variable might be tried).

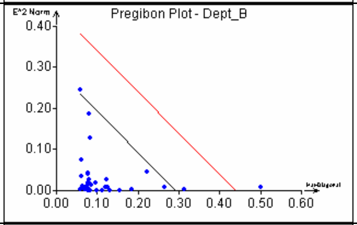
Influence



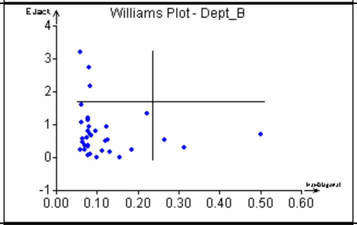
Plot of the projection matrix $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ diagonal elements. (**X** is the design matrix, i.e. the matrix containing explanatory variables as columns.) The element sizes are related to potential influence that the individual data points might have upon the regression results. The points plotted above the red horizontal line are considered to be potentially influential.



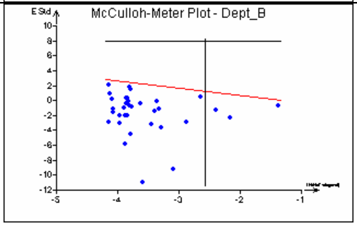
Predicted versus raw residuals plot. A large deviation from the line suggests that the corresponding observation is an outlier. The plot is very good in detection of isolated outliers. It is less sensitive to clusters of outiers which „mask“ each other.



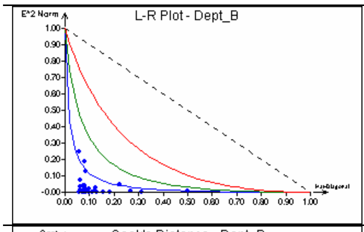
Pregibon‘s plot for simultaneous assessment of outliers and influence. The points above the lower (black) line are considered to be potentially influential, while the points above the upper (red) line are considered to be either substantially influential or outliers. Such data points should be checked carefully.



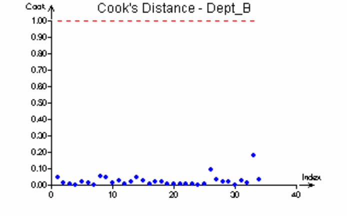
Williams‘ plot for simultaneous assessment of outliers and influence. The points located right from the vertical line are potentially influential, while the points above the horizontal line are suspected outliers.



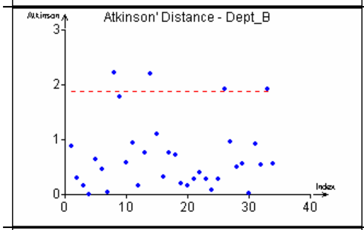
McCulloh-Meter plot for simultaneous assessment of outliers and influence. The points located right from the vertical line are potentially influential, while the points above the horizontal line are suspected outliers. The points above the red line are suspect either because they are influential or because they are outliers.



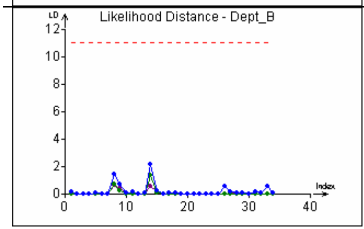
L-R plot for influence assessment. Hyperbolic curves are influence contours (connecting the points having the same influence). According to the location with respect to the three colored curves, datapoints can be classified as moderately influential, influential and substantially influential. The plot is most useful for smaller datasets.



Cook‘s distance is related to the influence data have upon magnitude (not variance) of the regression coefficients.

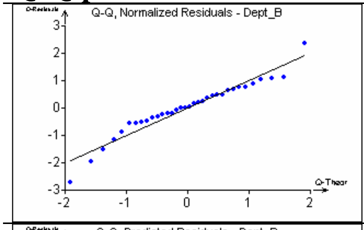


Atkinson‘s distance was derived as modification of the Cook‘s distance. Usually, the two yield similar results. Data points plotted above the horizontal line are considered to be influential.

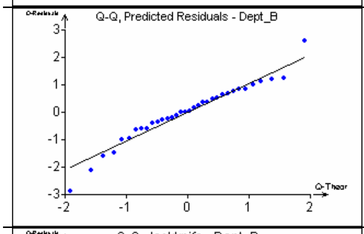


Likelihood related influence measure plot. The blue points express simultaneously the influence upon parameters and model predictions. Violet points express influence upon parameters, green points express the influence upon model predictions separately.

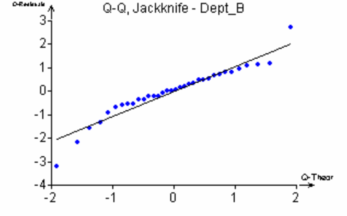
Q-Q plots



Q-Q plot of standardized residuals. It is used to assess residual normality. Approximately normally distributed residuals should plot close to the line.

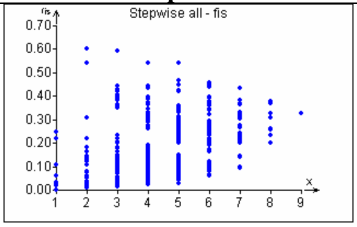


Q-Q plot of predicted residuals. It is used to assess residual normality. Approximately normally distributed residuals should plot close to the line.

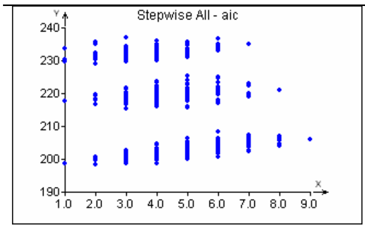


Q-Q plot of jackknife residuals. It is used to assess residual normality. Approximately normally distributed residuals should plot close to the line.

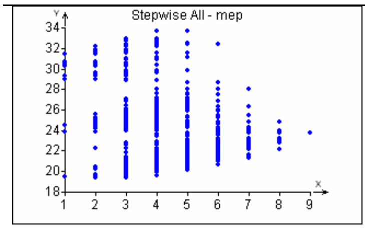
APSR related plots:



The plot is generated by the *APSR* (all possible subsets regression) procedure. It is useful when looking for the best models in terms of the F criterion. Number of variables included in the model is ploted on the *X* axis, while the F value is plotted on the *Y* axis. A good model should have a large F value. The best points (corresponding to models) can be selected interactively for further exploration (their detailed description can be found in the *APSR* data sheet, where they are selected automatically, once they are marked on the plot). It is highly recommended to choose several good looking models, explore them and select among them manually, using some subject matter knowledge.



The plot is generated by the *APSR* (all possible subsets regression) procedure. It is useful when looking for the best models in terms of the Akaike‘s criterion (AIC). Number of variables included in the model is ploted on the *X* axis, while the AIC value is plotted on the *Y* axix. A good model should have a small AIC value. The best points (corresponding to models) can be selected interactively for further exploration (their detailed description can be found in the *APSR* data sheet, where they are selected automatically, once they are marked on the plot). It is highly recommended to choose several good looking models, explore them and select among them manually, using some subject matter knowledge. Bands appear on the plot when there is a highly significant term among the potential model terms (much more important than other potential terms).



The plot is generated by the *APSR* (all possible subsets regression) procedure. It is useful when looking for the best models in terms of the mean squared error of prediction (MEP). Number of variables included in the model is ploted on the *X* axis, while the MEP value is plotted on the *Y* axix. A good model should have a small MEP value. The best points (corresponding to models) can be selected interactively for further exploration (their detailed description can be found in the *APSR* data sheet, where they are selected automatically, once they are marked on the plot). It is highly recommended to choose several good looking models, explore them and select among them manually, using some subject matter knowledge.

17.2. Nonlinear regression

Menu:	QCExpert	Nonlinear regression

Nonlinear regression module allows you to fit and analyze regression models of the general form

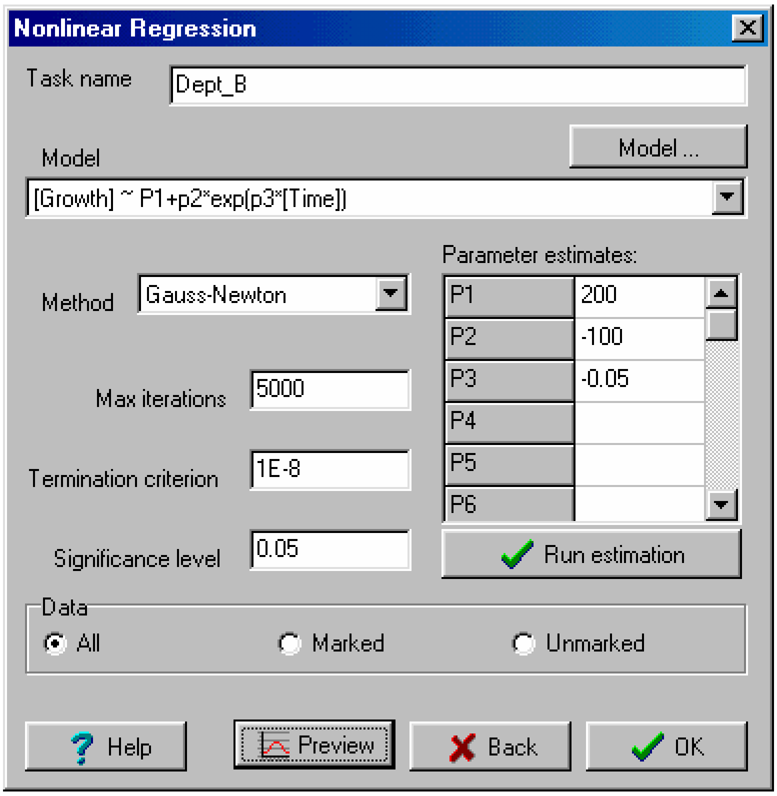
$$y = F(\mathbf{x},\mathbf{p}) \qquad (5)$$

Where y is a response variable, $\mathbf{x} = (x_1, x_2, \ldots x_q)$ are values of the explanatory variables (written as a vector). q is the number of explanatory variables in the regression model. There are m parameters, $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ in the model. $F(\mathbf{x},\mathbf{p})$ is a function of explanatory variables and parameters. Maximum number of parameters is 32, maximum number of variables is 254. Ideally, $\mathbf{x}$ is assumed to be a deterministic, i.e. non-random vector, which is either purportedly set to prespecified values or its values are found out via an essentially error-free procedure. y depends on **x,** but the dependence is blurred by the presence of a random error ε. Vector of model parameters **p** are estimated from data by the nonlinear least squares method. The user can specify a desired nonlinear model either in the Nonlinear regression dialog panel (Fig. 63) or in the *Model specification window*.

Note: If the desired model is linear with respect to the parameters, that is in the form of (1), use linear regression module instead (see the previous chapter), where the computations are of noniterative nature (no initial parameter estimates are needed). A typical example of models which are linear in parameters (albeit nonlinear in explanatory variables) are: $y = p_1x + p_2\ln(x)$ or polynomial models like $y = p_1x + p_2x^2 + p_3x^3 + p_4$. Other models might be linearized easily, for instance $y = p_1\exp(p_2x)$ can be linearized ln$y$ = lnp_1 + p_2x (quasilinearization might be needed to suppress possible error distribution distortion).

17.2.1. Data and parameters

Unknown parameters **p** = (p_1, p_2, . . ., p_m) are estimated from data contained in the current data sheet. Each column of the sheet corresponds to a variable. Names of the variables appear in the column headers. Parameters and variables are part of model declaration which can be completed either upon clicking the *Model…* button, or directly in the *Nonlinear regression* window (when the desired model was already specified previously). Detailed model specification instructions can be found later in this chapter. Once a model is specified, it appears in the *Model* window of the Nonlinear regression dialog panel.



Task name: Dept_B

Model: [Growth] ~ P1+p2*exp(p3*[Time])

Method: Gauss-Newton

Max iterations: 5000

Termination criterion: 1E-8

Significance level: 0.05

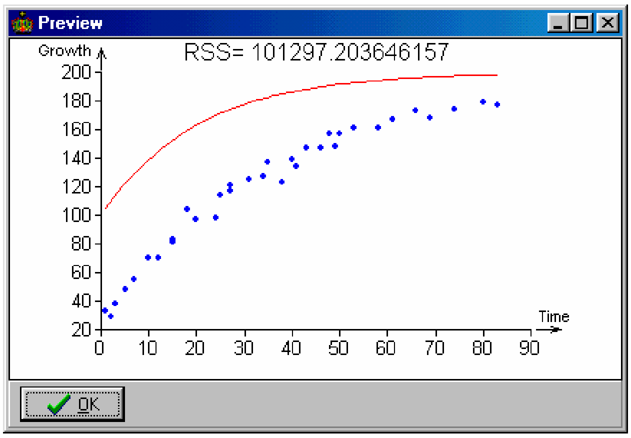
Parameter estimates:

P1	200
P2	-100
P3	-0.05
P4	
P5	
P6	

Data: ● All ○ Marked ○ Unmarked

Fig. 63 Nonlinear regression dialog panel

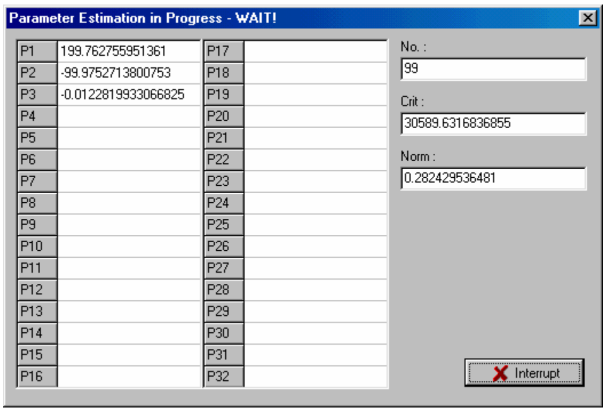
A project identification (it will appear in the header of all protocol pages and graphical output) can be inputted in the *Project name* field. Optimization method to be used by the regression procedure can be selected from the *Method* list (*Gauss-Newton*, *Marquardt*, *gradient*, *dog-leg*, *simplex* are possible choices). Maximum number of iterations is entered in the *Max iteration* field. The algorithm stops when either maximum gradient element meets a termination requirement or when the norm of the parameter change from one iteration to the next is smaller than a specified value. Alpha is the significance/confidence level used for all tests/confidence intervals. Buttons in the *Data* section of the dialog panel can be used to determine which part of the available data will be used (possible choices are: all data, selected data only, not-selected data only). Initial guess for all parameter values p_1, … p_m has to be inputted in the *Parameter estimates* field. Generally, the initial estimates should be close to the final regression estimates. Take care and time to supply as good initial estimates as possible. Rough initial estimates might produce incorrect final estimates, or it can happen that final estimates cannot be produced from rough initial values at all. Improper initial values might also lead to convergence problems resulting in very large number of iterations that the procedure needs to find final estimates (which might take considerable amount of computer time). Adequacy of the final parameter estimates can be checked by pressing the *View* button from the *Nonlinear regression* dialog panel. It displays data and model fit together with the residual sum of squares. The red (model) line should be close to data points. If the fitted model contains more than one explanatory variable, *View* produces observed versus predicted response plot together with the $y=x$ line, corresponding to an ideal fit (with no model-data discrepancy). The *View* window does not support any interactive features. It cannot be copied (using the *Ctrl-C* command) either. When finished with the plot inspection, press the *OK* button to get back to the *Nonlinear regression* dialog panel.



RSS= 101297.203646157

Fig. 64 *Preview* window

After the initial parameter estimates had been entered, computations are started upon pressing the *Compute* button. Progress of the iterative computational procedure can be monitored in the *Monitor* panel (Fig. 65) which is invoked automatically.



P1	199.762755951361	P17	
P2	-99.9752713800753	P18	
P3	-0.0122819933066825	P19	
P4		P20	
P5		P21	
P6		P22	
P7		P23	
P8		P24	
P9		P25	
P10		P26	
P11		P27	
P12		P28	
P13		P29	
P14		P30	
P15		P31	
P16		P32	

No. : 99

Crit : 30589.6316836855

Norm : 0.282429536481

Fig. 65 *Monitor* panel

The panel contains current iteration information: parameter values, iteration number, residual sum of squares and other information, depending on the optimization method. *Norm* field contains the norm of parameter vector change from one iteration to the next. The norm is compared with a prespecified number (termination criterion) and the procedure when the actual norm is smaller, procedure stops. Computations can be stopped manually at any time by the S*top* button (the procedure then returns parameter values from the last iteration).

When the computation procedure stops, the program returns to the *Nonlinear regression* panel. The *Parameter estimates* field contains parameter from the last completed iteration. When the procedure stops in a normal way (meeting the stopping criteria), returned parameters should be equal to the optimal values (nonlinear least squares estimates). When in doubt about convergence status of the procedure at termination, you should check carefully parameter values plausibility. You can use the *Preview* window to inspect the estimated model fit visually (Fig. 64). It is generally recommended to try different optimization methods in case of problems (e.g. slow convergence, divergence). Final estimates from one method can be used as starting values for the next method (or they can be somewhat edited before starting the computations). These steps might need to be repeated several times (checking the parameter estimates plausibility through *View* along the way). It might be also useful to check whether slightly perturbed final estimates used as starting values will yield the same values as before perturbation. After performing these various checks, the final estimates are accepted by pressing the *OK* button in the *Nonlinear regression* dialog panel. QCExpert then produces protocol and graphical output. ***Warning***: if *OK* is pressed before running the computations, the starting parameter estimates are accepted as the final estimates (without any optimization)!

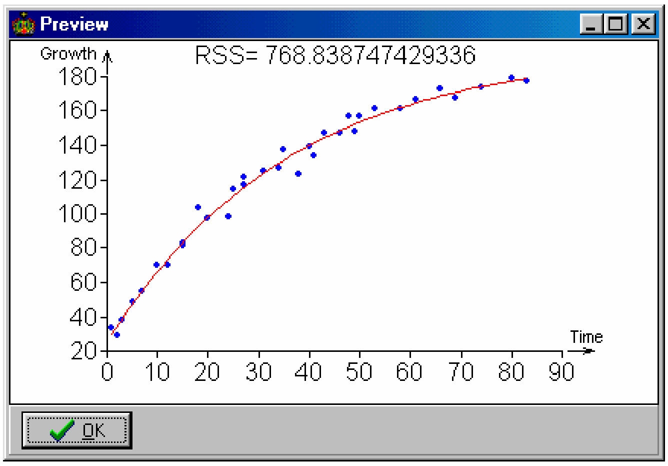
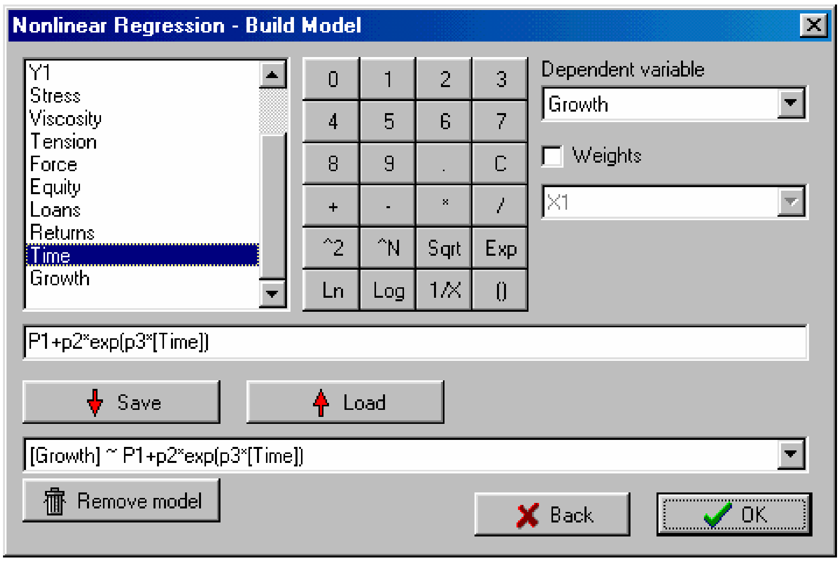

RSS= 768.838747429336

Fig. 66 *Preview* after finishing computations

Model specification: The *Model…* button opens a new panel for model specification (Fig. 67). If you specified some models previously, you can select one of them without opening the *Model specification* panel. Make sure that the variable names in the current data sheet and in the model are the same in such case. There is a list of current data sheet variables in the left part of the *Model specification* panel. These are the only variables you can use in the model building. The response variable is inputted in the upper right part of the panel. When desired, you can check the Weights option, which will enable you to specify name of a data sheet column which contains weights w_i. They correspond to coefficients by which individual residuals (not squared residuals) are multiplied. Inputted weights are automatically standardized to sum to n (number of data points). When the Weights option is not checked, unit weights, w_i=1 are used by default. There are shortcut buttons for model specification in the central part of the panel. Input line, where the model is actually specified, appears at the bottom of the panel. There you can also find a list of previously defined models. The *Save button* saves a model after it was completely specified. The newly saved model appears in the list of previously defined models in the „current“ position. The *Read* button reads a selected model in and places it in the input line, where it can be modified.


Nonlinear Regression - Build Model

Y1
Stress
Viscosity
Tension
Force
Equity
Loans
Returns
Time
Growth

Dependent variable: Growth

☐ Weights

X1

P1+p2*exp(p3*[Time])

[Growth] ~ P1+p2*exp(p3*[Time])

Fig. 67 Model specification dialog panel

Model specification instructions:

Double click on a variable name in the current data sheet variable list to copy the name to the input line. Variable name is always enclosed in square brackets. Parameters have to be represented by the P1, P2, … codes. These codes are then used in the *Parameter estimates* field in the *Nonlinear regression* panel.

Shortcut function buttons can be helpful when writing more complicated expressions. Highlighting a part of the model input line and clicking a function button subsequently, applies the function on the highlighted part as an argument. For instance, the expression ln([*x*]+1) can be assembled in the following way: double click on the variable *x* (there has to be a column of this name in the current data sheet): **[x]**; write + 1 manually; highlight the whole expression: **[x]+1**; click the *Ln* button, this action results into: **ln([x]+1)**. Application of *^2*, *^A*, *Sqrt*, *Exp*, *Log*, *1/X*, *()* is similar. The *C* button erases model input line. Other functions have to be inputted manually (writing their name in the model input line). Available functions are listed in the Linear Regression chapter. When finished with specification, the model is saved by the *Save* button. Then, it automatically appears in the list of previously specified models located in the bottom part of the panel. The *Read* button reads in a model from the set of previously defined models. Its terms can be edited then. The *Erase model* deletes a selected model from the model list. Warning: this operation is irreversible! The *OK* button finishes model specification.

A model can be defined without using mouse and shortcut buttons as well. Usual syntactic rules apply, keep in mind that variable names have to be enclosed in square brackets. Previously defined models can be selected from list of models in the *Nonlinear regression* dialog panel without going through the *Model specification* dialog panel (variable names in the data sheet and in the model specification have to agree).

Computational methods

Nonlinear regression procedure implementation can be viewed as a procedure for finding such parameter values that minimize some kind of distance between model predicted and actual response values,

$$\min_{\mathbf{p}} S(\mathbf{p}) = \min_{\mathbf{p}} D(\mathbf{y}, \hat{\mathbf{y}}), \qquad (6)$$

where *D* stands for a distance, $\mathbf{y}$ is a vector of observed response values and $\hat{\mathbf{y}}$ is a vector of model predicted response values. Euclidean distance is used most commonly for D,

$$S(\mathbf{p}) = \|\mathbf{y} - \hat{\mathbf{y}}\| = \|\mathbf{e}\| = \sqrt{\sum_{i=1}^{n} [y_i - \hat{y}_i]^2} = \sqrt{\sum_{i=1}^{n} e_i^2} \qquad (7)$$

As far as minimization is concerned, we can look at squared distance in place of the distance itself, so that we get a simple expression, without the square root sign. This is to say that the original minimization is equivalent to minimizing the sum of squared differences between data and model predictions. The minimization typically has not closed form solution so that it is performed numerically, through an iterative procedure. The procedure is based on some kind of nonlinear optimization algorithm. Different algorithms have different properties and no universally best algorithm exists. Therefore, the QC.Expert implements six different algorithms. Each of them requires a vector of initial estimates (initial guess) $\mathbf{p}_0$ to start a search for the parameter estimates $\mathbf{p}^*$ (i.e. the final or „optimal" values). First five implemented algorithms belong to derivative based methods, which use first and possibly also second derivatives of the optimized function. The derivatives are taken with respect to parameters. The sixth implemented method is the simplex method, which does not require derivatives, all it needs to evaluate is the minimized function (i.e. $S(\mathbf{p})$ or its square). The derivative based algorithm tend to be more efficient when the initial estimates $\mathbf{p}_0$ are sufficiently close to $\mathbf{p}^*$. How close they need to be, it depends mainly on how nonlinear a particular model is. When there is a strong nonlinearity and/or it is difficult to produce initial estimates $\mathbf{p}_0$ reasonably close to $\mathbf{p}^*$, the derivative based methods may fail badly. The simplex algorithm might be an alternative to try then. The simplex method might take a very long time to converge. Hence, it is sometimes useful to start with the simplex method, stop it prematurely after the estimates stabilize to some extent, and to use the returned values as the initial values for a derivative based method in the second step. The following algorithms are implemented in the *Nonlinear regression* module:

Gauss-Newton: A classical derivative based algorithm. It is built on the idea of model linearization. When the model is not very nonlinear and/or the initial estimate $\mathbf{p}_0$ is close to $\mathbf{p}^*$, it tends to converge very fast. It can diverge in less ideal situations. To reduce step length problems, the length can be reduced by a damping parameter, *Damp*≤1, which is displayed during computations. Its starting value is 1.

Marquardt: A mixed type of derivative based algorithm, which combines Gauss-Newton and gradient approach. It tends to be more reliable than any of the two methods.

Gradient-Cauchy: A derivative based method which uses direction of steepest descent direction together with Cauchy step length found by minimization in the gradient direction. The Cauchy point is determined by a heuristic approach in order to prevent the algorithm from „being locked in“ a banana shaped valley. To reduce step length problems, the length can be reduced by a damping parameter, *Damp*, which is displayed during computations. Its starting value is 1. The algorithm can be slow in a banana shaped valley.

Dog Leg: A derivative method which is, like the Marquardt method, based on a combination of gradient and linearization. It uses previous iteration history to improve Hessian (the matrix of second derivatives) approximation, see Denis Mei paper in the Literature. To reduce step length problems, the length can be reduced by a damping parameter, *Damp*, which is displayed during computations. Its starting value is 1. Two additional values *Theta* and *T* are displayed during computations.

Gradient, fixed step length: A derivative method based on the gradient of $S(\mathbf{p})$ only. The method is useful in the earlier stages of optimization. It can be very slow for strongly nonlinear models in later optimization stages (closer to the minimum). To reduce step length problems, the length can be reduced by a damping parameter, *Damp*, which is displayed during computations. Its starting value is 1.

Simplex: This method does not require $S(\mathbf{p})$ derivatives. Geometrically, it corresponds to flipping a simplex (with m+1 points) in the parametric space. The QC.Expert implementation uses a heuristic approach and so called „mutations“ when constructing the simplex. Because the method does not need derivatives at all, it is useful for strongly nonlinear models. It can be very slow, compared to derivative based methods (when the later can be applied). In the course of computations, the simplex expansion coefficient *Norm* is displayed.

17.2.2. Protocol

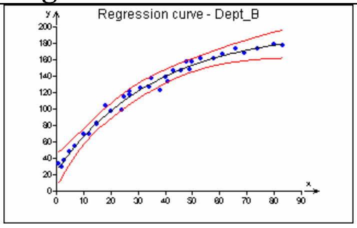
Project name	Project name as entered in the dialog panel.
Significance level	Alpha, significance/confidence level which is used for all tests/confidence intervals.
Degrees of freedom	Degrees of freedom, $n-m$ (number of data points minus the number of model parameters).
Quantile t(1-alpha/2,n-m)	t-distribution quantile.
Quantile F(1-alpha,m,n-m)	F-distribution quantile.
Method	Method used (least squares method)
Number of data points	Number of complete data rows, having information on all model variables.
Number of parameters	Number of the regression model parameters.

Method	User -selected numerical optimization method.
Explanatory variables	List of explanatory variables which appear in the regression model.
Response	Response variable.
Model	The regression model; response variable appears before the „~“ sign.
Initial values	Initial parameter values.
Computations	
Iterations	Number of iterations.
Termination	Optimization algorithm termination; the word *Convergence* is displayed when the algorithm ended in a normal way, reaching convergence; when computations were manually interrupted by pressing the *Stop* button, the word *Interrupted* is displayed; when a prespecified maximum number of iterations is exceeded without meeting termination criterion, the word *Divergence* is displayed; when no computations were performed, the words *Was not optimized* appear. **Warning**: the word *Convergence* might not necessarily mean that the returned parameter estimates are correct! You should always check adeqacy of the fitted model visually and inspect all parts of the output carefully (e.g. correlation matrix of estimates).
Computation time	CPU time (in seconds) spent by the procedure.
Max. iteration number	Prespecified maximum iteration number. When the number is exceeded without meeting stopping criteria, divergence is claimed.
Termination criterion	Norm of the parameter vector change has to be smaller than this number in order to claim convergence.
Parameter estimates	Parameter estimates found by an optimization algorithm, accompanied by the asymptotic standard errors of the estimates and asymptotic confidence intervals (using a prespecified α).
Parameter correlation matrix	Asymptotic pairwise correlations for all parameter pairs. Ones appear on the diagonal necessarily. Correlation between parameters should be expected, but when some correlations are close to +1 or –1, results are suspect. It might be useful to reparametrize the model.
Residual analysis	
Characteristic	
Y observed	Observed response value, as it appears in the current data sheet.
Y predicted	Predicted response value.
Std. error of Y	Estimated standard error of the prediction.
Raw residual	Difference between observed and predicted response value.
Residual [%Y]	Relative residual, raw residual divided by the response value.
Weights	Weights for individual observations as inputted by the user.
Residual sum of squares	Residual sum of squares cannot decrease when a new variable is included in the model (usually, it increases).
Mean of absolute residuals	Mean of absolute residuals.
Residual standard deviation	Standard deviation estimated from residuals.
Residual variance	Variance estimated from residuals.
Residual skewness	Skewness estimated from residuals.
Residual curtosis	Curtosis estimated from residuals.

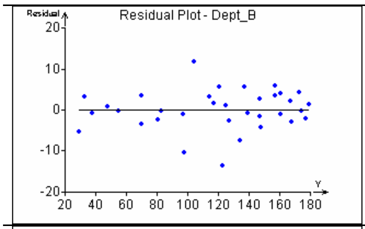
Characteristics of the model fit	
Multiple correlation coefficient, R	Multiple correlation coefficient characterizes how closely the model fits the data. It does not necessarily express how good the model is. R cannot decrease when a new variable is included in the model (usually increases whenever a new variable is added)!
Coefficient of determination R^2	Square of the multiple correlation coefficient.
Mean square error of prediction, MEP	The ith error is the difference between actual value of the ith observation and its prediction. The prediction comes from the model based on data with the ith row omitted. MEP is sensitive indicator of problems like multicollinearity and outliers. It is an important characteristics of the regression model quality.
Akaike information criterion	AIC in the regression context is related to the residual sum of squares, penalized by the model size (number of explanatory variables).

17.2.3. Graphical output

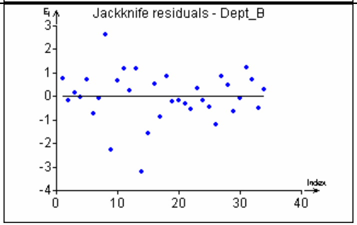
Regression curve



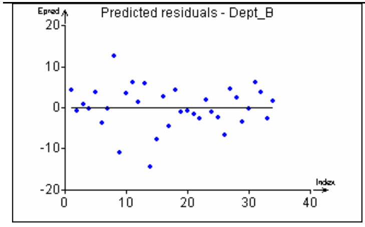
This plot is not produced when the model contains more than one explanatory variable. When only one explanatory variable appears in the model, the plot displays the regression curve. Red curves show the confidence band around the regression curve, computed for a prespecified confidence coefficient. It should be noted that the confidence band is realistic only when the fitted model is (approximately) correct. This is even more important when predictions further from bulk of available data points are considered. Details of the plot can be inspected upon zooming part of it. The regression curve can be inspected even outside of the interval containing explanatory variable values actually used in model fitting by the inverse zooming.



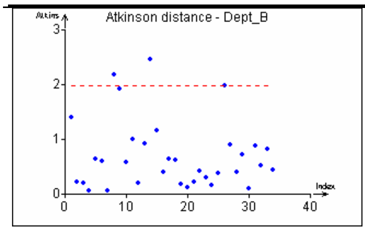
Standardized residuals plot. Predicted response is plotted on the *X* axis, while the standardized residuals are plotted on the *Y* axis. Horizontal line corresponds to the mean of residuals. Any systematic plot pattern suggests an incorrect or incomplete model, or incorrect estimates.



Jackknife residuals plot. Data index is plotted on the *X* axis, while the jackknife residuals are plotted on the *Y* axis. Horizontal line corresponds to the zero residual. Any systematic plot pattern may suggests an incorrect or incomplete model, or incorrect estimates. Outliers are detected much more precisely than on the Standardized residuals plot. Compare Linear regression.

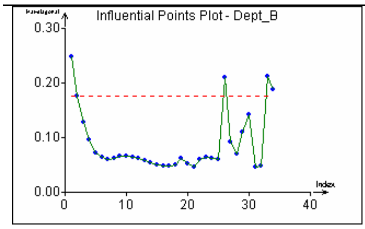


Predicted residuals plot. Data index is plotted on the *X* axis, while the predicted residuals are plotted on the *Y* axis. Horizontal line corresponds to the zero residual. Outliers are detected much more precisely than on the Standardized residuals plot. Compare Linear regression.



Atkinson distance. Data index is plotted on the X axis, Atkinson distances are plotted on the Y axis. Horizontal red dashed line corresponds to the 95% quantile of the distribution of this statistics, which is uset to detect influential data. Points above the red line are assumed highly influental.

Influence



Plot of the projection matrix $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ diagonal elements. ($\mathbf{X}$ is the matrix of the first partial derivatives w.r.t. the model parameters). The points plotted above the red horizontal line are considered to be potentially influential.

17.3. Logistic Regression

Menu:	QCExpert	Regression	Logistic

Logistic regression assumes one or more real independent variables and a binary response variable with values 0 or 1, usually representing logical value like false/true, good/bad, etc. Alternatively, the response may have a form of frequency ratio in the interval <0, 1> in case of repeated measurements. This ratio should be the number of positive results divided by number of trials n_1/n at a given value of the independent variable. Logistic regression is then used to model probability of some event in dependence on the independent variables x. It is supposed that the response is a random variable with alternative distribution with parameter π which denotes the probability of a positive outcome of a trial. Thus, the number of positive outcomes out of a fixed number n of trials have a binomial distribution Binom(n, π). This parameter depends on x monotonously and logistic regression model will be an estimate of this dependence. Applications of logistic models are wide and include diverse fields of science and technology. Typically logistic models are used to estimate risks or failures under given conditions, bank credit scoring of a client, probability of suvirval of an organism in given environment, in toxicology, pharmaceutical, medicine, ecology, reliability analysis, maket research, etc.

The probability π is modelled with a logistic model

$$\pi(x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

or after rearrangement

$$\log\frac{\pi(x)}{1-\pi(x)}=\alpha+\beta x\,.$$

The expression log($\pi(x)$ / (1 – $\pi(x)$)) is called logit. Parameters α and β are regression coefficients and their estimates a, b are computed with an iterative least squares methods. Such values of α a β are maximum likelihoods estimates. If x is univariate, logistic model may be plotted as a sigmoid-shape curve $\pi(x)$ describing the dependence of probability of a positive outcome on x. This model may then be used for prediction of the probability at any new value of x. The independent variable may be multivariate, $\mathbf{x} = (x_1, ..., x_m)$. Correspondin model for multivariate logistic regression can be expressed by

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + +\beta_m x_m)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + +\beta_m x_m)}.$$

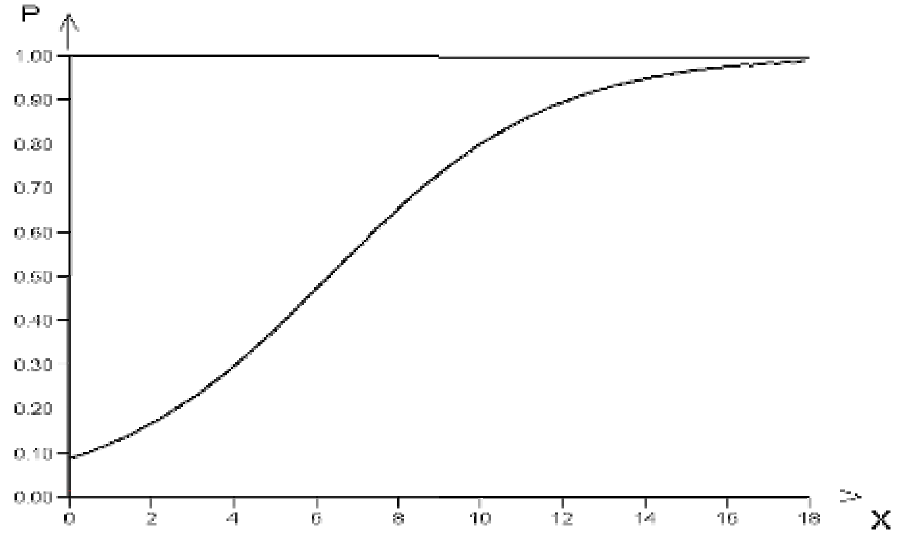


Fig. 68 Logistic regression model for one variable *x*

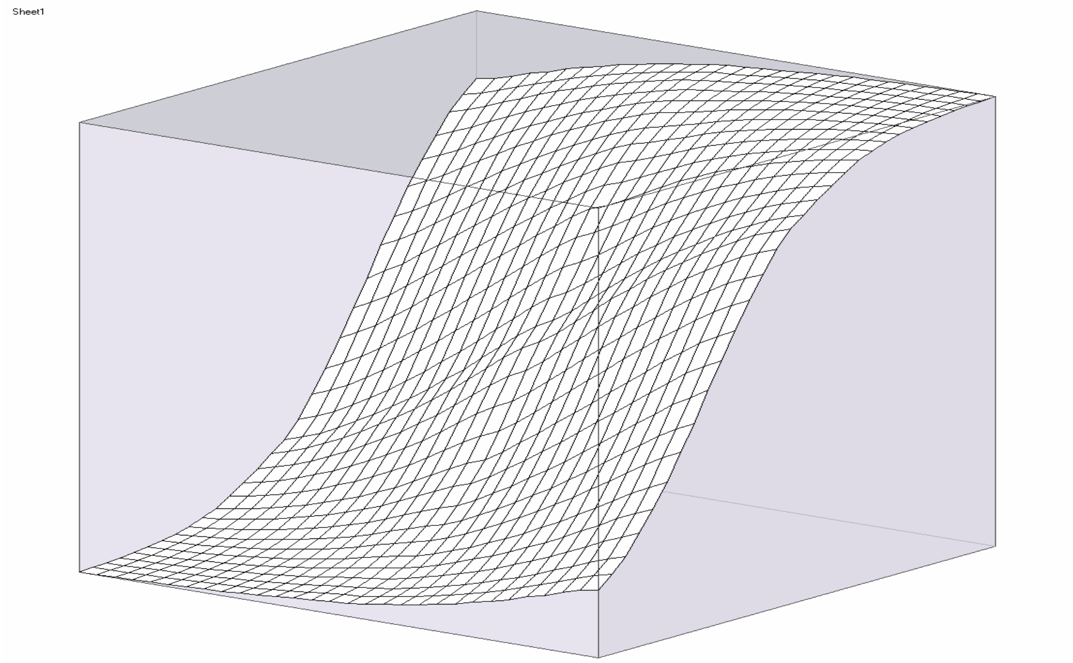
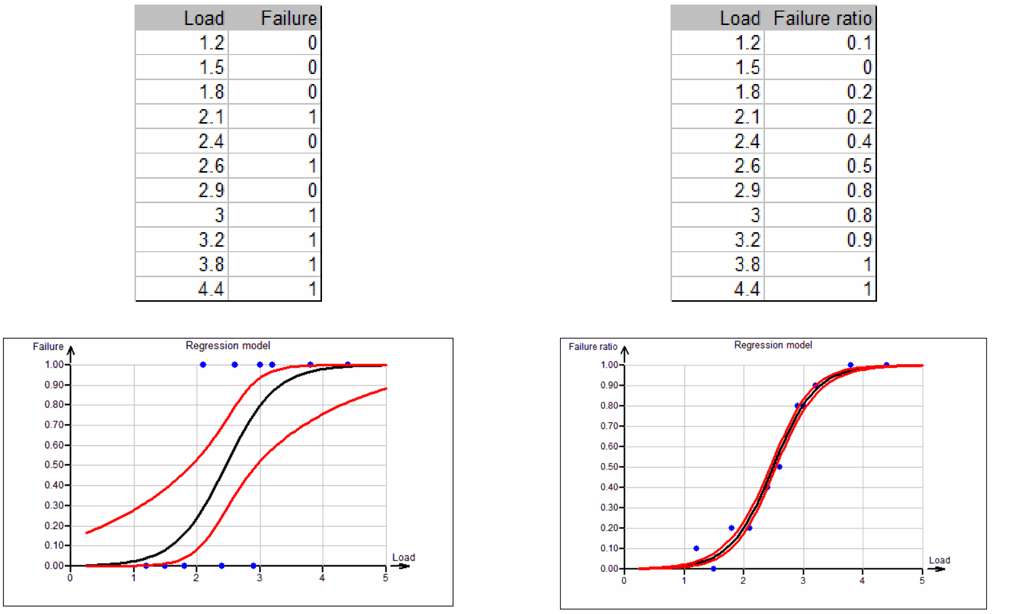


Fig. 69 Example of 3D logistic regression model for two variables x = (x_1, x_2)

Computed model can be used to predict the probability of the positive outcome of the experiment of test based on any new user-defined values of the independent variable.

17.3.1. Data and parameters

Data for this module must contain at least two columns, one of which is the independent variable, second is the test outcome (0 or 1), or the fraction of positive outcomes from a set of trials performed independently on each other at the same value of *x*. Generally, it is recomended to use the original binary data instead of summary fractions. The choice of 0 or 1 to denote the outcome is arbitrary. The dependence of probability on *x* may be ascending as well as descending. Two example of data for logistic regression are given on Fig. 70.

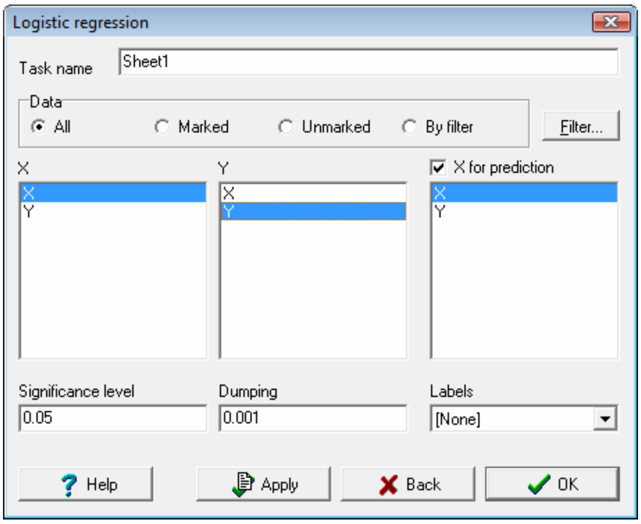


Load	Failure
1.2	0
1.5	0
1.8	0
2.1	1
2.4	0
2.6	1
2.9	0
3	1
3.2	1
3.8	1
4.4	1

Load	Failure ratio
1.2	0.1
1.5	0
1.8	0.2
2.1	0.2
2.4	0.4
2.6	0.5
2.9	0.8
3	0.8
3.2	0.9
3.8	1
4.4	1

A. Single binary responses, 11 measurements   B. Summary ratio data, 110 measurements

Fig. 70 Examples of data and resulting logistic curves



Logistic regression

Task name: Sheet1

Data: ● All ○ Marked ○ Unmarked ○ By filter

X: X (selected), Y

Y: X, Y (selected)

☑ X for prediction: X (selected), Y

Significance level: 0.05

Dumping: 0.001

Labels: [None]

Fig. 71 Dialog box for logistic regression

In the dialog panel *Logistic regression* choose one or more dependent variables X and one independent variable Y. Values for prediction can be selected in the field *X for prediction*. These values can be identical with the independent variable column. Predicted probability including confidence intervals will be computed for each value for prediction.

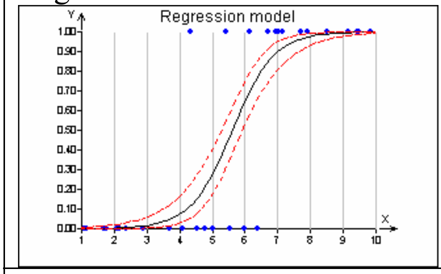
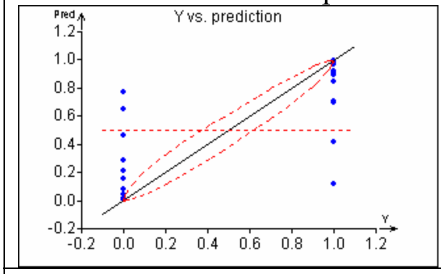
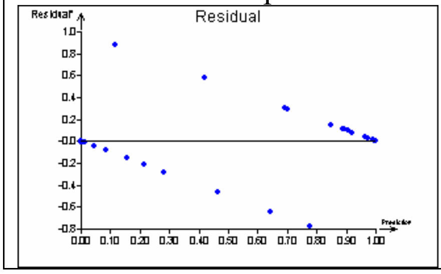
The number of variables and the variables independently for the prediction must be the same.

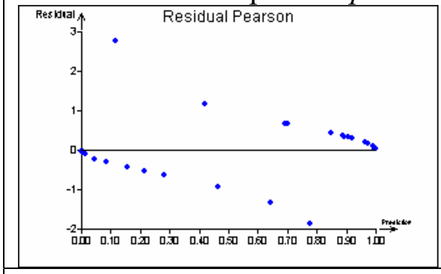
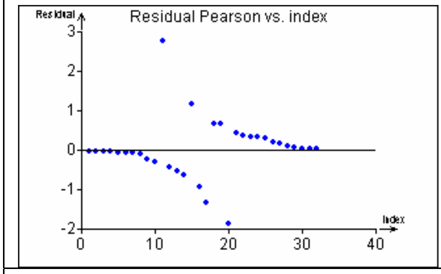
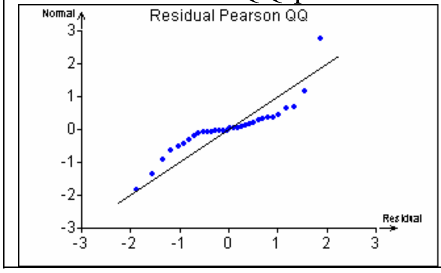
17.3.2. Protocol

No of cases	Number of rows used
Independent variables	List of independent variables

Dependent variables	Name of the dependent variable
No of iterations Max likelihood	The number of iterations used by the algorithm to calculate parameters by maximum likelihood method Logarithm of the maximal likelihood reached
Parameter estimates Parameter Estimate Std Deviation p-value	Logistic model parameter values Parameter names taken from names of the independent variables, *Abs* stands for the absolute term Parameter estimates Parameter standard deviations p-value gives the theoretical significance level at which the statistical significance of a parameter would just be rejected
Table of prediction Variable name Prediction Lower limit Upper limit	Table of predicted probability values for each value (or row) of the independent variable selected in the *X for prediction* field in the dialog box, see Fig. 71. Names of the independent variable (variables) Probability of an event as predicted by the logistic model Lower confidence limit for the predicted probability Upper confidence limit for the predicted probability

17.3.3. Graphical output

Regression model  	Logistic regression curve with its confidence band and measured data. This plot is displayed only for a single independent variable *x*. The confidence band (the upper and lower red curve) defines a confidence interval at selected confidence level 1 – α (usually 0.95, ie. 95%) of the predicted probability for each value of the independent variable. For higher precission it is advisable to zoom in the plot.
Predicted vs. measured plot  	Model versus data plot (predicted vs. observed). This plot is an analogy of the prediction plot in regression. Points far from the line y=x are not necessarily outliers.
Absolute residuals plot  	Graf of absolute residuals or a distance between data and the logistic curve. The residuals are of different nature than the residuals in a linear or non-linear regression. They do not have a normal distribution and their values are always between -1 and 1.

Pearson residuals plot vs. *p*  	Transformed Pearsonova residuals are comparable with classic residuals such as in the linear regression. They have a standard normal distribution $N(\mu = 0, \sigma^2 = 1)$ and can be used for diagnosis of outlying measurement.
Pearson residuals vs. index  	The same residuals as in the previous plot. If the data are unordered, one can better observe the distribution of residuals than in the previous plots.
Pearson residual QQ-plot  	QQ-graph of transformed residuals is an effective diagnostic tool for assessing the normality of residuals as though the data have alternative, or binomial distribution, the Poisson-transgormed residuals should be normal. Points that are significantly apart from the line are suspected outlying points, and it is appropriate to verify their validity, or conclusions from the computed model should be drawn with with caution. Ideally, points are scattered around the designated line.

18. Predictive methods

18.1. Neural Network

Menu:	QCExpert	Predictive methods	Neural network

Neural network (Artificial Neural Network, ANN, or NN) is very popular and powerful method, which is used for modelling the relationship between multivariate input variable **x** a multivariate output variable **y**. NN is generally considered to be non-linear regression model, which can make the network structure. The inspiration for the neural network structure of brain tissue was higher organisms, which neuron is connected with the so-called synapses to several other neurons. Electrical current (or information signal) flows through synapses, is processed by a neuron and transmitted by other synapses to other neurons.

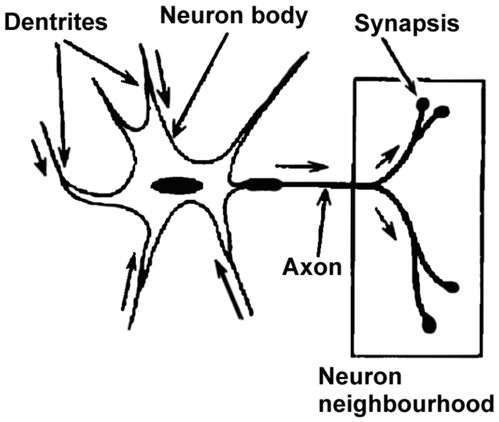


Fig. 72 Biological neuron: cell body, input dentrites, output axon, synapses-connection to other neurons

The artificial neural network tries to copy the structure and functionality of the biological neural structure and models the structure mathematically. Neuron core is represented In ANN the nodes are, by analogy, called neurons, each input variable xi entering the j-th neuronu multiplied by a weighting factor of wji. The sum of the weighted input variables $z_j = w_{0j} + \Sigma w_{ji}x_i$ is then transformed by neuron using an activation function. Activation function expresses the intensity of the neuron response to the input change. The most commonly used activation functions include logistic functions,

$$\sigma_j(z) = 1 / (1 + e^{-z}),$$

which is similar to the biological function of sensory response, for example: there is practically no difference if you tauch temperatures 50K or 150K (both are too cold) or temperatures 2000K of 4000K (both too hot). But you will very precisely distinguish between 90 and 100°F (35 and 40°C), because here is the vital information. Weights w_{ji} represent the intensity of information flow between the variable and neuron or, in the case of multi-layer networks between neurons in layers, these links are sometimes called synapses by analogy to the bio-neurons and can be interpreted as significance of variables and visualized in a plot.

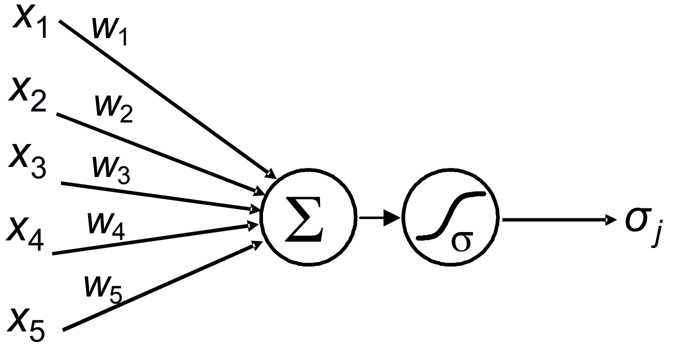


Fig. 73 Structure of an artificial neuron

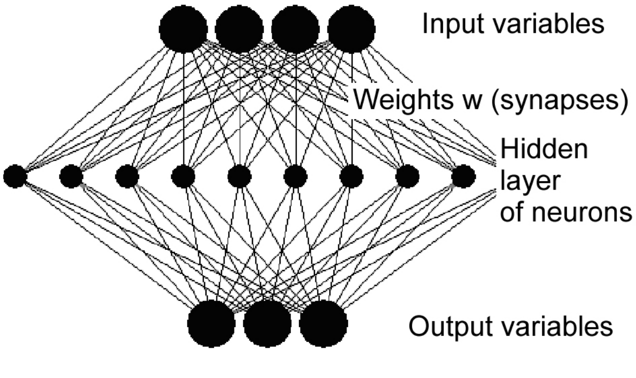


Fig. 74 Possible architecture of an ANN with 1 hidden neuron layer

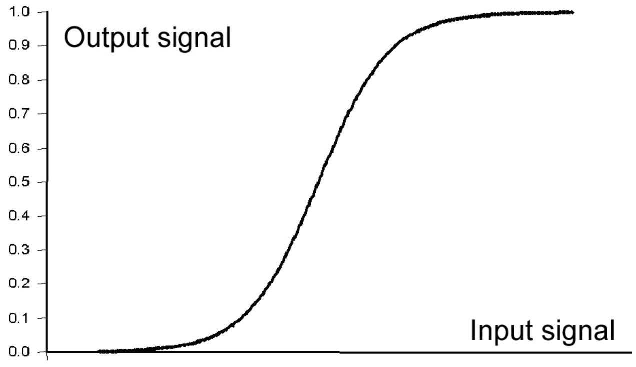


Fig. 75 Commonly used activation function of a neuron σ(*z*)

Output variables are predicted as weighted linear combination of outputs from the last hidden layer neurons, $\hat{y}_i = \sum_k w_{ik}\sigma_{Lk}$. Neural network is therefore formally a special case of multiple nonlinear regression, neural network can be practically considered non-parametric regression. If the neural network did not contain any hidden layer neurons – only input and output variables, it would be a linear regression model. Neural network is optimized to satisfy least residual squares criterion. This means that the network is set so that the squares of the differences between prediction and the measured output variables value was minimal. This is the aim of iterative optimization process, which is called learning or training neural networks by finding the best values of all weights. QCExpert uses an adaptive derivative Gauss-Newton algorithms to optimize the net. Trained network can then be used for prediction of output variables, for new specified input variables. Neural network model is local, that means that its prediction ability is sharply declining outside the range of the independent variables used to train it, see Fig. 76.

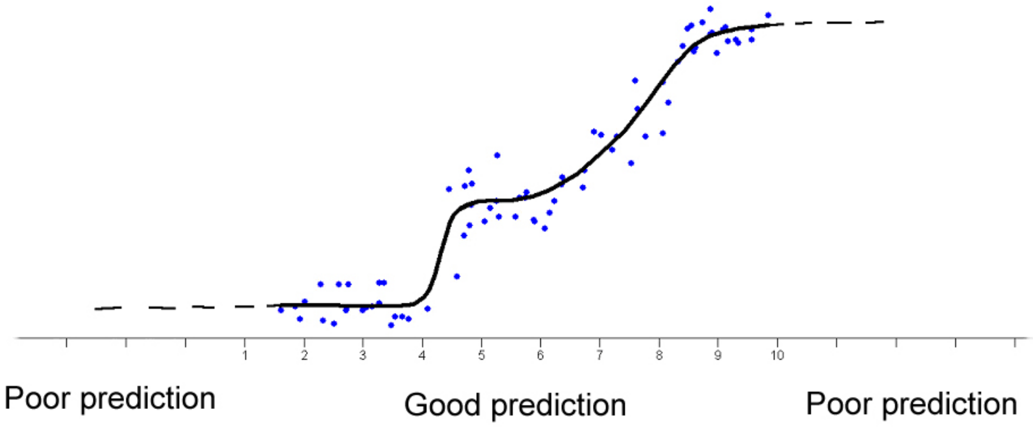


Fig. 76 Prediction capability of an ANN drops dramatically in the areas where no training data are available

A typical procedure for using neural network may be as follows.

1. Select group of predictors (independent variables, *X*) which we believe that may affect the dependent variables *Y*. Select a group of dependent variables, which should depend on the predictors. In each line must always be all values of dependent, and independent variable. Number of rows is denoted *N*.

2. Select the architecture of the neural network, the number of layers and numbers of neurons in each layer. There is no straightforward rule for the best network architecture, usually it is appropriate to use a number of neurons very roughly corresponding to the number of variables. Single hidden layer networks are recommended where we assume a linear, or slightly non-linear relationship. Two-layer network can be suitable for strongly nonlinear relationships. Using more then 3 layer networks is usually not very effective. It is necessary to keep in mind that for very complex network there is high risk of overdetermined ambiguous and unstable models or models wthich are difficult to optimize. Examples of possible architectures are given on Fig. 89. Number of data (lines) should be at least ten times greater than the number of neurons in the network, otherwise there is a risk of overdetermination and the ability of prediction may decrease. Usual architectures for middle to large-scale problems are networks with 2 to 20 neurons, and 1 to 3 layers.

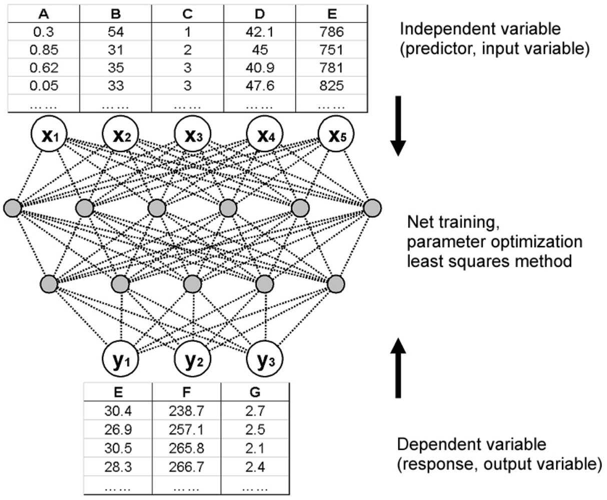
3. Optimizing parameters of the network, or the so-called "learning" neural networks. During this process, the optimization algorithm tries to find a set of weights, so that the predicted values are in the best accordance with entered dependent variables. This consistency is measured by the sum of squares, as in other least squares regression methods. In general, it can not be guaranteed that the found solution is the best possible. Therefore, it is advisable to run optimization several times to get better residual sum of squares (RSS). Optimization starts with random values of the weights, it is therefore natural that each solution found by optimizing is completely different. Even completely different combination of the weights in the network can provide virtually identical prediction model with the same RSS.

4. If an information about the reliability of prediction is required, we can use cross-validation. In this technique we select the so-called training, or learning subset of data , say $P.N$ lines ($0 < P < 1$) to be used to train the network. The rest of the data, the remaining $(1 - P).N$ lines of testing or validation data, are then used to validate network, i.e. check if the predicted values for the validation data are close to the actual dependent data.

5. The success of neural networks can be assessed according to the decrease of squares sum during the optimization process, according to fit plots of prediction and by the thickness of the lines connecting neurons (the thickness is proportional to the absolute value of the weight, which is interpreted as the intensity of information flow downward from predictors to response).

6. Prediction: A trained network may be used for predicting response variable. Put new values of the independent variables on input of the network. The structure of the variable must be the same as used to train the ANN and values should be in the same range. The network will predict the output values.

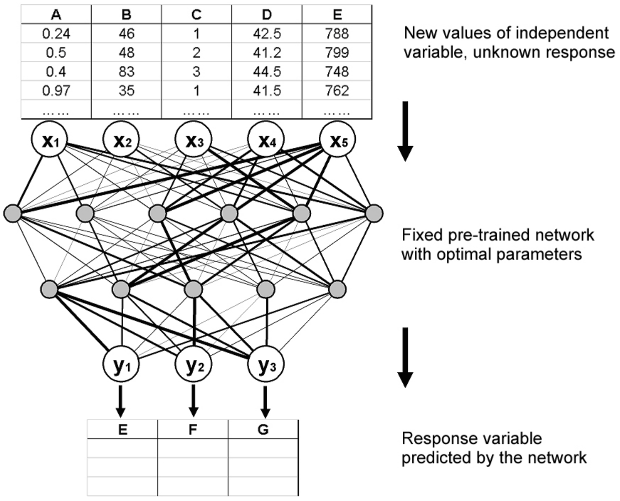
The steps are shown on Fig. 77 and Fig. 78.



A	B	C	D	E
0.3	54	1	42.1	786
0.85	31	2	45	751
0.62	35	3	40.9	781
0.05	33	3	47.6	825
……	……	……	……	……

E	F	G
30.4	238.7	2.7
26.9	257.1	2.5
30.5	265.8	2.1
28.3	266.7	2.4
……	……	……

Fig. 77 The ANN training process



A	B	C	D	E
0.24	46	1	42.5	788
0.5	48	2	41.2	799
0.4	83	3	44.5	748
0.97	35	1	41.5	762
……	……	……	……	……

E	F	G

Fig. 78 Using trained ANN for predicting unknown response

Model validation

ANN models usually do not allow the calculation of statistical parameters and diagnostics for a detailed assessment of the quality of the model (variances of the regression coefficients, F- and t-statistics for testing the relevance and significance of the model, diagnostic plots, etc.), as in the case of linear regression. It is therefore necessary to use other methods, to verify whether the model is appropriate for description of the phenomenon under study. Neural network is a very flexible instrument and can easily lead to a situation where the model will suspiciously well describe (fit) the data, but not the phenomenon (variables relationship) as a whole. This is reflected in very poor prediction of the values of the dependent variables for the new independent variables that have not yet occurred in the data, although they may be located inside the interval of training data.

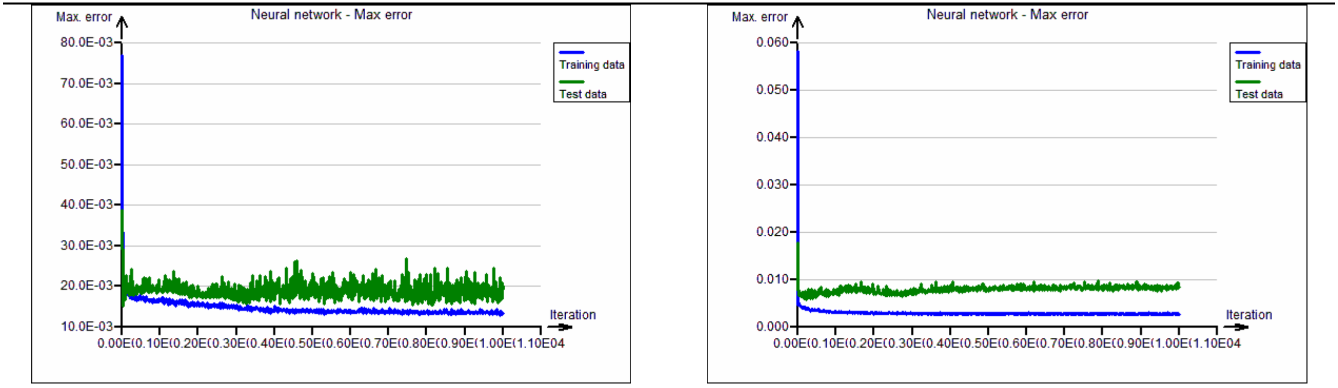


Fig. 79 ANN optimization process with cross-validation: very good prediction capability

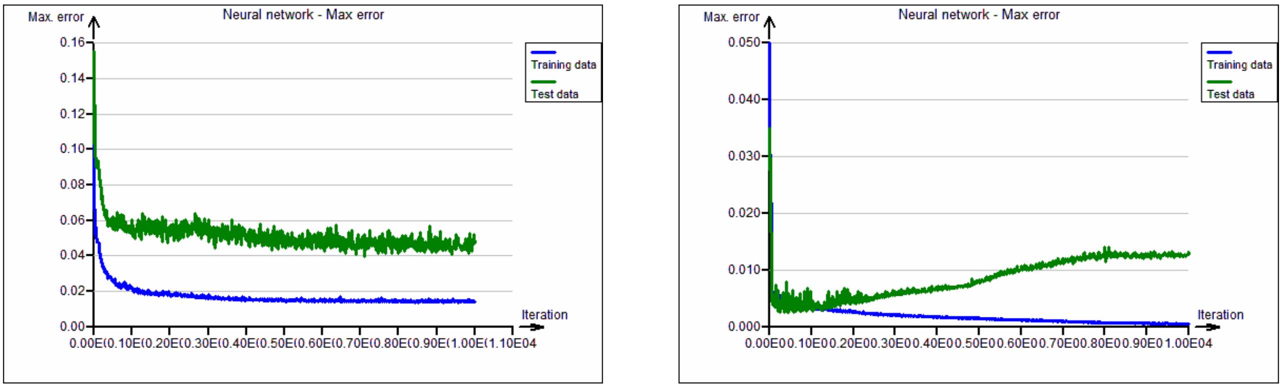


Fig. 80 ANN optimization process with cross-validation: fairly good prediction capability

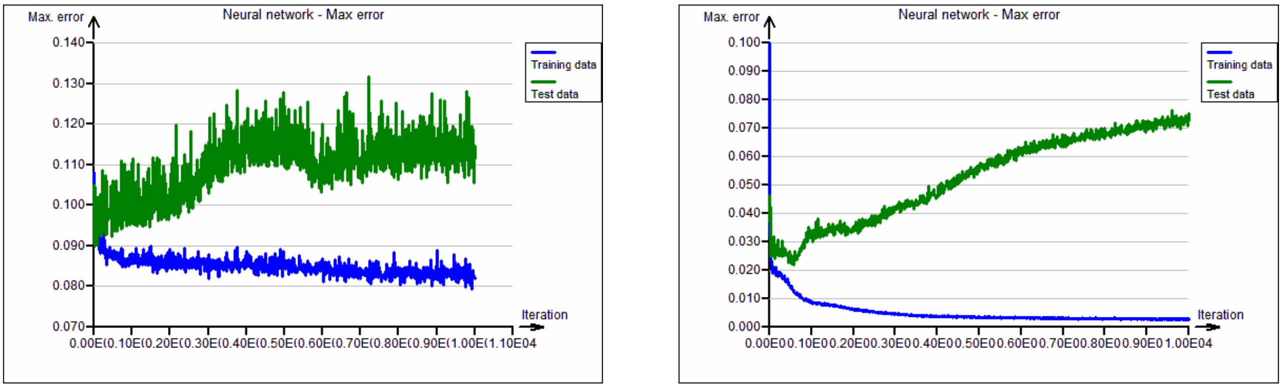


Fig. 81 ANN optimization process with cross-validation: poor or none prediction capability. The network is too complex or data size is too small, possibly there is no information in the data

To assess the prediction capabilities of the neural network validation (cross-validation) is used. Validation is based on a simple principle of training the ANN only with a certain part *P* of the data. This training part is chosen typically around *P* = 0.7 to 0.9, or 70 to 90%. The 1 – *P* rest of the data (test or validation data, not „seen“ during training the ANN) is afterwards used to calculate prediction, which which is compared to the true value of response. If this prediction agrees well with the actual response, we can confirm the ability of the ANN to correctly predict the response for data, which it has not „seen“ previously. The quality of prediction can be assessed using a graph or chart of errors during the the progress of optimization process. The Fig. 79 to Fig. 81 illustrate the use of this concept for three models, using 30% (*P* = 0.7) of randomly selected validation data. Fig. 82 through Fig. 85 illustrate different capabilities of NN-prediction models on the Data-prediction plot.

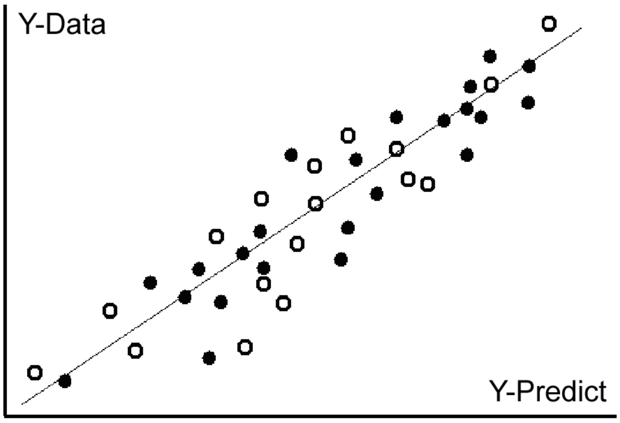


Fig. 82 Comparable quality both for training (filled) and validation (hollow) data. Good prediction capability

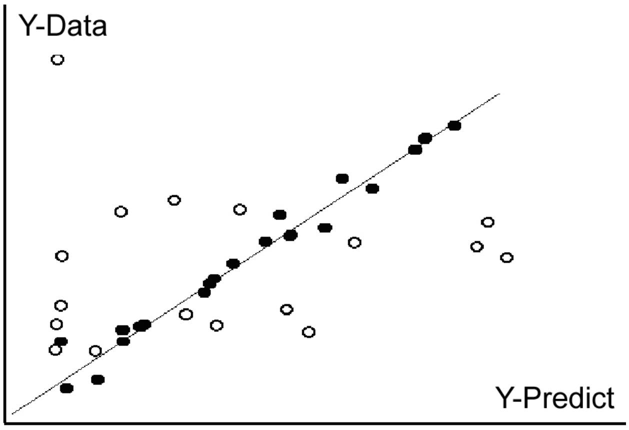


Fig. 83 Good prediction training data, but very poor for the validation data (hollow). Poor prediction capability of the network. Network probably too complex.

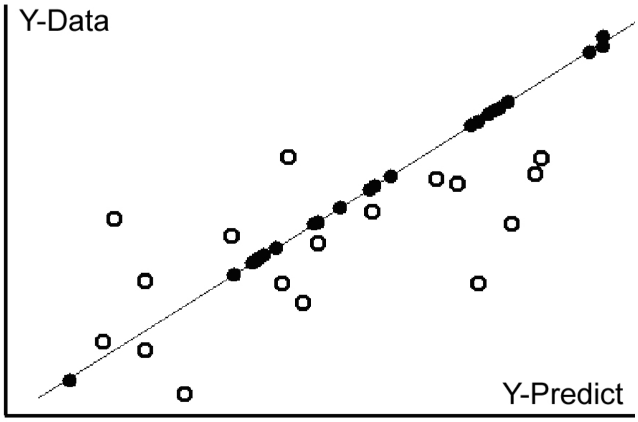


Fig. 84 Perfect prediction training data, but very poor for the validation data (hollow). Network probably too complex.

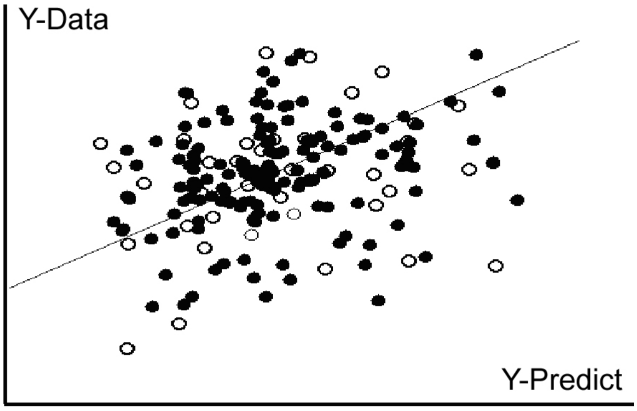


Fig. 85 Poor fit for both training and the validation (hollow) data. Try to re-run optimization, add layers or neurons, possibly there is simply no dependence between input and output.

Classification with ANN

Given the probabilistic nature of the logistic activation functions neural networks can as well be used as a modelling tool for classification when the output is a discrete variable - two-level (binary), like 0 and 1, or multi-level such as 1, 2, 3, or A, B, C. Neural network predicts the level of output variables for the given values of the independent variables, as in logistic regression, see paragraph 17.3, page 17-141. In the case of a binary response 0 – 1, prediction can be considered as the probability of occurence of „1“. The following plots on Fig. 86 and Fig. 87 illustrate the use of ANN as a classification model. On the left there is the measured response (bright point corresponds to the value of 0, dark point value 1). On the right is a shaded map of the prediction obtained by neural network. Plots were obtained by the module Graphs - 3D-Spline (see paragraph25.2.14, page 25-209).

Table 9 An example of classication data: Technological parameters X and Y (independent variable) presumably influence the result of an operation (response = OK/Fail, or 0/1). For the ANN, numerical form of the response (like 0/1) is required. The response may have more levels denoted e.g. 0, 1, 2, … or 100, 200, 300.

Parameter X	Parameter Y	Result	Result - Binary
1.5	3.1	Pass	0
1.9	2.2	Pass	0
3.5	2.8	Non-conforming	1
2.9	4.3	Non-conforming	1
2.4	2.7	Pass	0
.....	.....	.....	.....

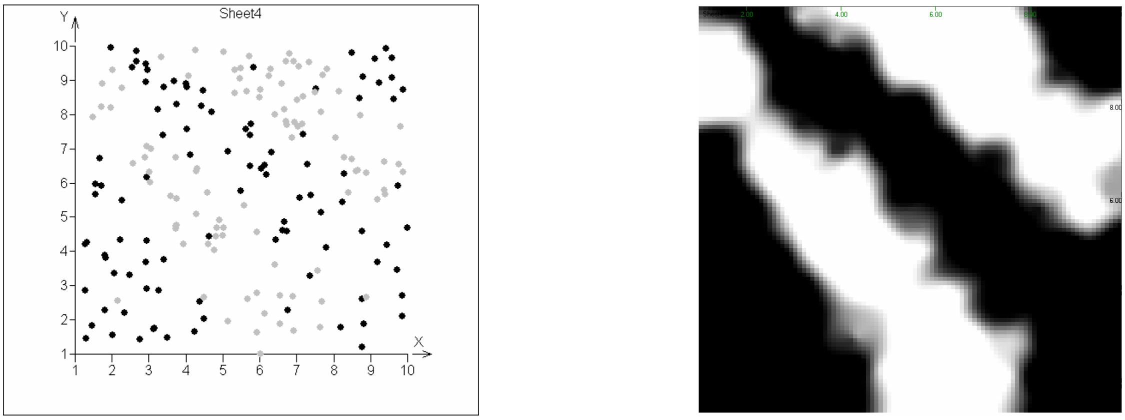


Fig. 86 Linearly separable data – One hidden layer NN with 6 neurons used as a classification model.

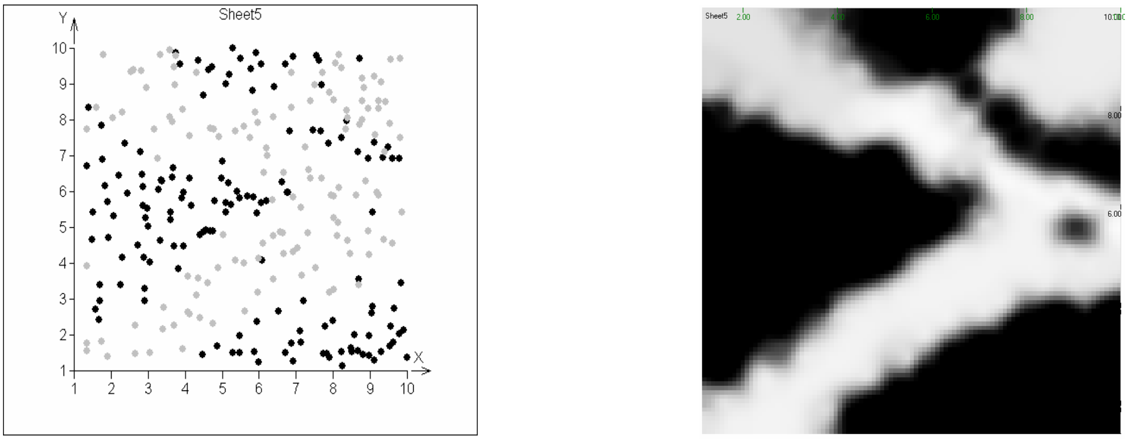


Fig. 87 Linearly inseparable data – Two hidden layers NN with 5+5 neurons used as a classification model.

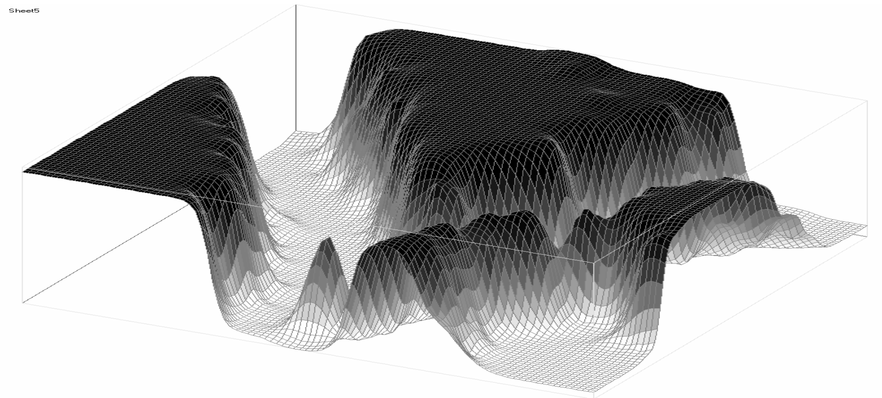


Fig. 88 3D representation of the previous example using NN prediction and 3D-Spline from the Graphs module

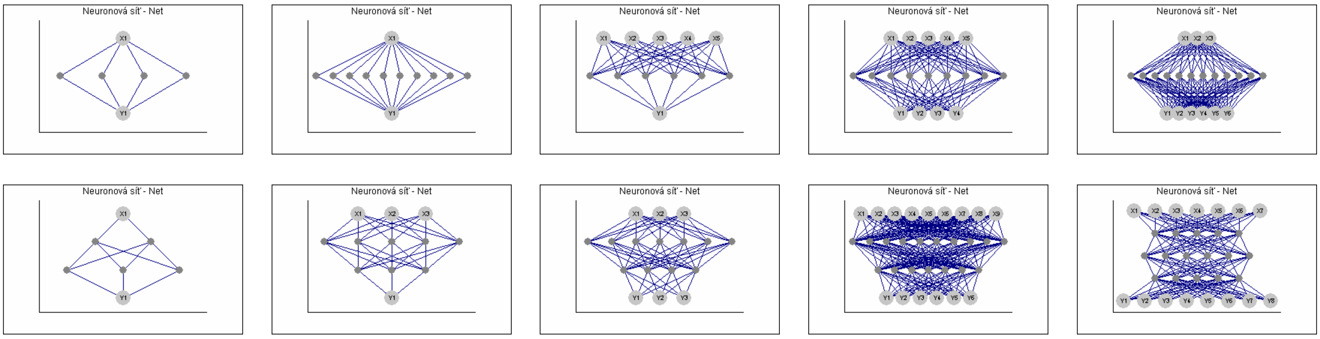


Fig. 89 Illustrative examples of suitable NN architectures. An ANN must be designed with respect to the data size and nature

18.1.1. Data and parameters

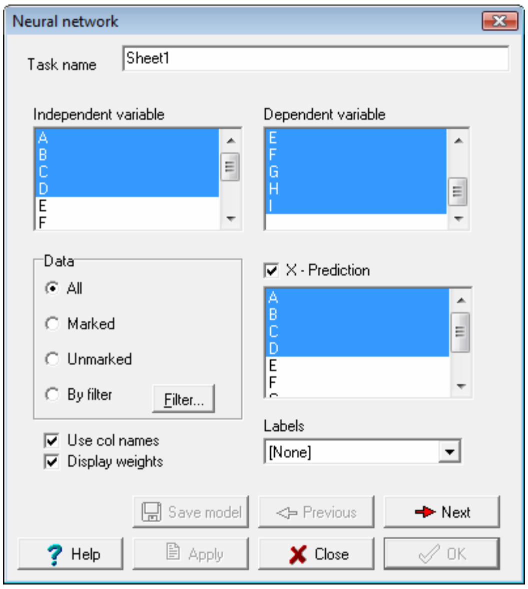
Module Neural Networks has several consecutive dialog boxes, which can set the parameters of calculation. In the first dialog box, the columns of independent and dependent variables are selected. In the Data group you can select the required data subset: all data, or just a specified subset of rows. Checking the box *Prediction* to calculate the value of prediction of the dependent variable for selected indepenent variable columns. We have to select the same number of columns, as we have selected in the field *Independent variables*, the values for prediction must also have similar values as the independent variable for reliable prediction.

The same columns can be selected as in *Independent valiable* field. If the field *Use Col names* is checked, the names of the columns are used to describe the input and output neurons in the graph. If *Display weights* is checked (recommended), the absolute value of weights are visualized as the thickness of connecting lines between neurons. The sign of the weight is represented by color (blue = positive weight, red = negative weight). Click *Next* to get to the next window. In the *Neural network architecture* dialog window we will define the network architectue – number of layers and number of neurons in the layers. Typical number if layers is 1, 2 or 3 layers. More than 3 layers may be useful only in some specific cases. The *Number of neurons in the hidden layers* field determines how many neurons to include in individual layers. The problem of how to choose a suitable architecture is discussed below. The *Number of iterations* field determines the length of the calculation in terms of number of iterations of the network optimization process, recomended default value is 10000. The *Exponent* determines exponent of the criterial function, here, the default is 2, which corresponds to least squares method. Exponents between 1 and 2 will somewhat robustify the network and are recomended when the data are suspicious for outliers or possible big errors in dependent variables.

$$s(NN) = \Sigma\, |y - y_{\text{pred}}|^k$$

Parameter *Sigmoid steepness* indicates speed (sensitivity) with which the neurons respond to change of the independent variables. The recommended value is 1. The parameters *Moment* and *Learning speed* affect the optimization algorithm. Recommended values are 0.9 and 0.1. The field *Part of training data (%)* determines what part P of the data is to be used for training the network. The rest $(1 - P)$ of data is then used for cross-validation. Cross-validation is a technique that will check stability and prediction capability of the chosen neural network model by using only part (typically 70-90% of the original rows) of the data to train the network. Then, the dependent variable of the rest of the rows is predicted from independent variable values and the predicted values are compared to the actual, measured values (never seen before by the network). To use the cross-validation, we usually choose the ratio between 60 and 90%. Choosing 100% will disable cross-validation and all the data will be used for training the network. The data for cross-validation is chosen by random number generator.

Alternatively, the user may want to select the training data manually by marking rows for cross-validation. When some rows are marked (red) check *Use unmarked data* to tell the NN to use only unmarked rows for training the network. The marked rows will then be used for cross-validation.



Neural network

Task name: Sheet1

Independent variable: A, B, C, D (selected), E, F

Dependent variable: E, F, G, H, I (selected)

Data: ● All ○ Marked ○ Unmarked ○ By filter

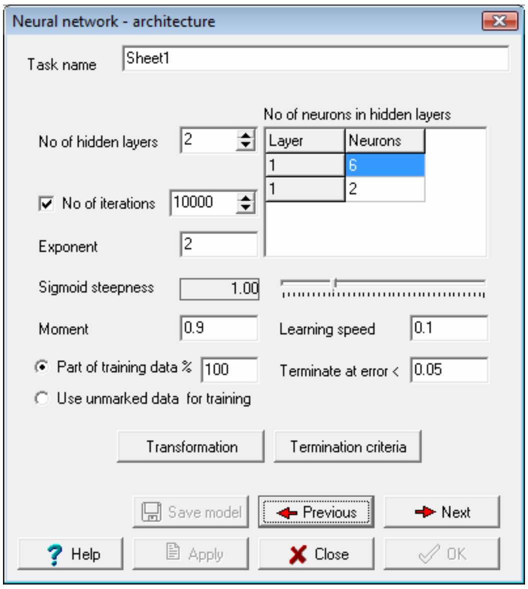
☑ X - Prediction: A, B, C, D (selected), E, F, G

Labels: [None]

☑ Use col names

☑ Display weights

Fig. 90 Step 1: Select predictors and responses, optionally choose the data for prediction. Check „Display weights“ do visualize importance of predictors and predictability of response



Neural network - architecture

Task name: Sheet1

No of hidden layers: 2

No of neurons in hidden layers:

Layer	Neurons
1	6
1	2

☑ No of iterations: 10000

Exponent: 2

Sigmoid steepness: 1.00

Moment: 0.9

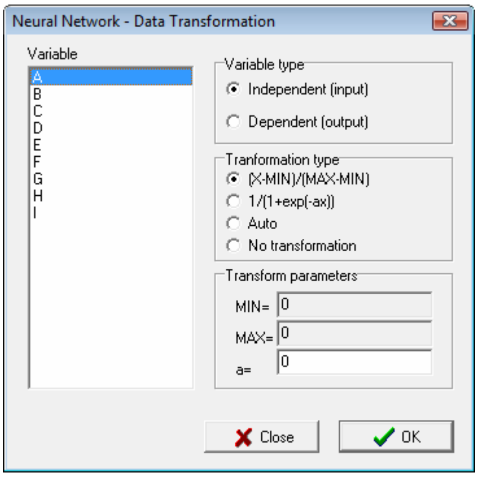
Learning speed: 0.1

● Part of training data %: 100

Terminate at error <: 0.05

○ Use unmarked data for training

Fig. 91 Step 2: Design the ANN architecture, choose number of hidden layers and number of neurons in each layer. Optionally, specify the part P of training data (rest will be used for cross-validation



Neural Network - Data Transformation

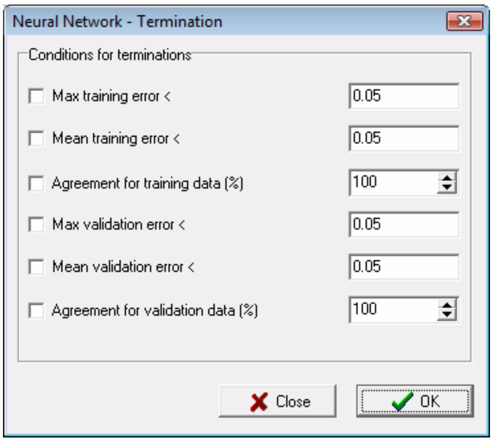
Variable: A (selected), B, C, D, E, F, G, H, I

Variable type: ● Independent (input) ○ Dependent (output)

Tranformation type: ● (X-MIN)/(MAX-MIN) ○ 1/(1+exp(-ax)) ○ Auto ○ No transformation

Transform parameters: MIN= 0 MAX= 0 a= 0

Fig. 92 Optional: Define special predictor transformation



Neural Network - Termination

Conditions for terminations

☐ Max training error < 0.05

☐ Mean training error < 0.05

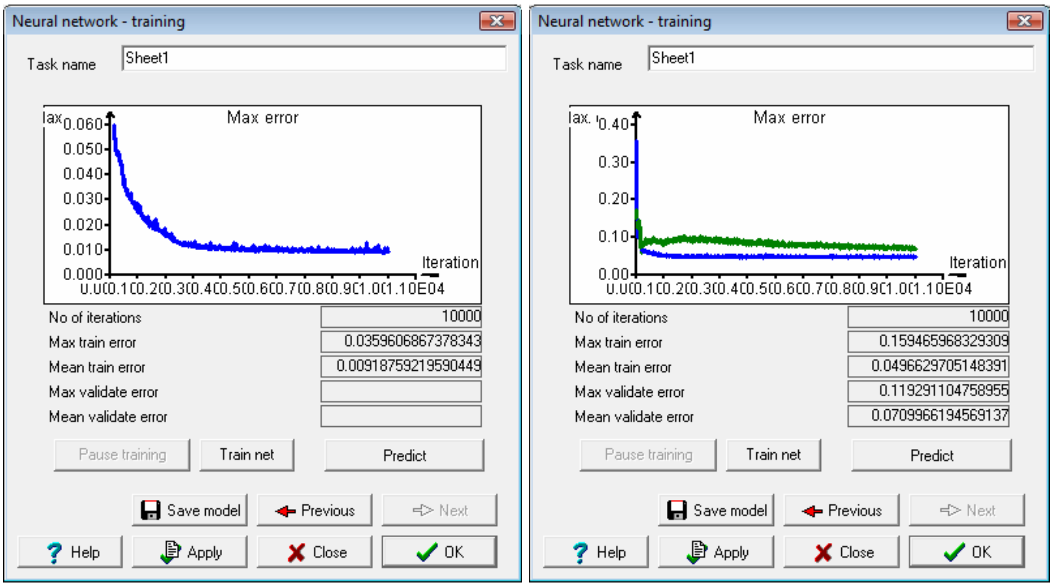
☐ Agreement for training data (%) 100

☐ Max validation error < 0.05

☐ Mean validation error < 0.05

☐ Agreement for validation data (%) 100

Fig. 93 Optional: Modify the termination conditions



Neural network - training

Task name: Sheet1

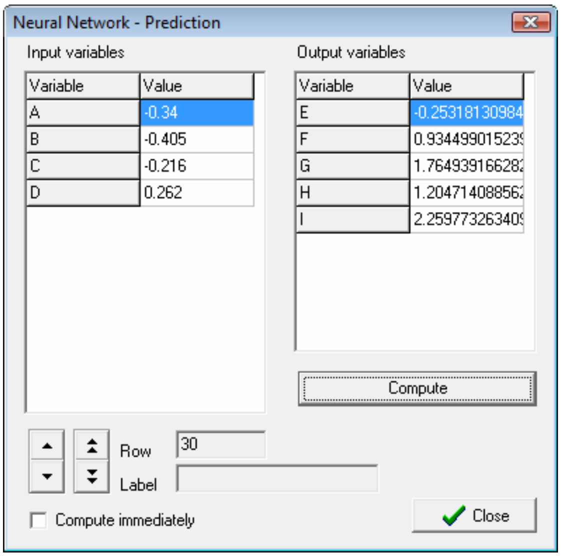
No of iterations	10000
Max train error	0.0359606867378343
Mean train error	0.00918759219590449
Max validate error	
Mean validate error	

Neural network - training

Task name: Sheet1

No of iterations	10000
Max train error	0.159465968329309
Mean train error	0.0496629705148391
Max validate error	0.119291104758955
Mean validate error	0.0709966194569137

Fig. 94 Step 3: Run the optimization (training) process and observe how the ANN is successful. Left: no cross-validation, Right: with cross-validation. Validation data errors in green. Afterwards, you may save the model for later use, train the net again with different initial weight set, run interactive Prediction panel or press OK to get the results.



Neural Network - Prediction

Input variables

Variable	Value
A	-0.34
B	-0.405
C	-0.216
D	0.262

Output variables

Variable	Value
E	-0.25318130984
F	0.93449901523!
G	1.76493916628:
H	1.20471408856:
I	2.25977326340!

Row: 30

Label:

□ Compute immediately

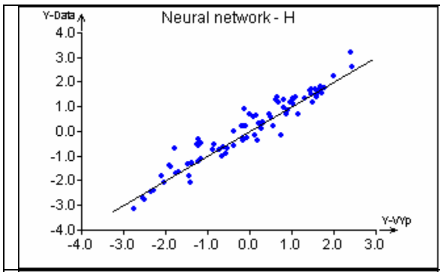
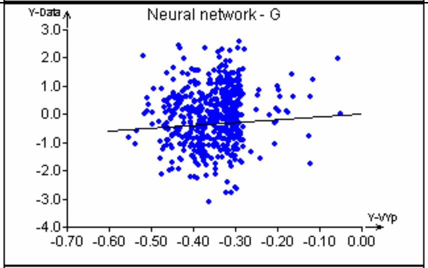
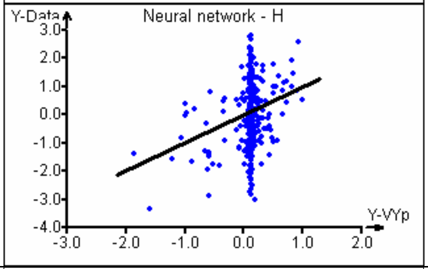
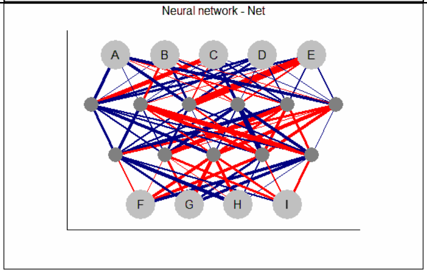
Fig. 95 Optional: Interactive prediction panel. Type new predictor values, or select a row of original data by „^“ or „v“, alter the predictor values and observe the changes in predicted response values.

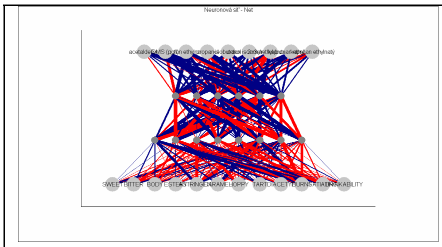
18.1.2. Protocol

Task name	Task name
Data	
Independent variable	List of independent variables
Transformation type	Type of used transformation of the independent variables
Dependent variable	List of dependent variables
Transformation type	Type of used transformation of the dependent variables
Layer, Neurons	Number of layer and number of neurons in layers
Sigmoid steepness	User defined sigmoid steepness
Moment	User defined moment parameter for optimization process
Training speed	User defined training speed

Terminate when error <	Condition for terminating optimization process
Training data (%)	Percentage or randomly selected data for training in case of cross-validation, else 100%
Termination conditions	Condition for terminating optimization process
No of iterations	Number of optimizing iterations from dialog box
Optimization report	
No of iterations	Actual number of optimizing iterations until end or user interrupt
Max training error	Minimal max training error value reached
Mean training error :	Mean training error value reached
Weights	Table of the optimized weights of the ANN
Layer / Neuron	Number or layer and number of neuron
Prediction	Table of predicted values, if chosen by the user

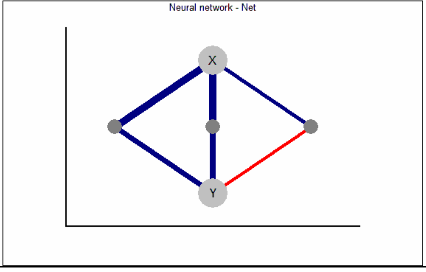
18.1.3. Graphical output

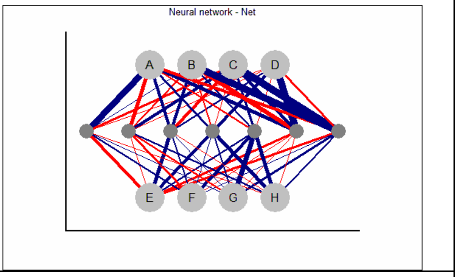
	Y-prediction plot. Plot of agreement between measured response and prediction for each of the response variables. The closer the points are scattered to the line the better the prediction of this variable. Quality of prediction usualy vary from variable to variable. This plot is an overall assessment of success of the ANN model. The plot on the left shows good quality of fit.
 	If the data do not show a clear trend (like the two plots on the left) then this response variable cannot be described well with this ANN model. This can be either due to the premature termination of the optimization before reaching the optimum weights, or too simple network that is unable to identify possible more complex dependence, or sadly (and most probably) this variable simply does not depend on the selected predictors.
 Neural network - Net A B C D E F G H I	Graphical representation of the network architecture. If the checkbox "*Display weights*" was checked (see Fig. 90 on page 18-153) the thickness of synapses (connenction lines) represent the absolute value of the corresponding weight and thus the amount of information that flows down between two neurons. From the thickness of the synapses going from the predictors we ca assess their significance (the thicker lines the more significant variable). Greater weight values on the input to response nodes (thick lines going to the predictor nodes) suggest the quality of prediction of each dependent variable. Color of synapses shows only sign of the weight (red = negative weight, blue = positive weight), which is of



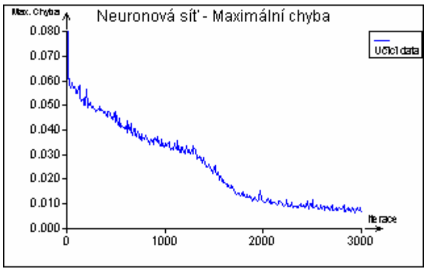
little practical interest in complicated nets, but may be of use in simple ones. Variable nodes are labelled by the column names, if the apropriate checkbox was checked.

Examples of typical arcitectures are shown on the left and below.

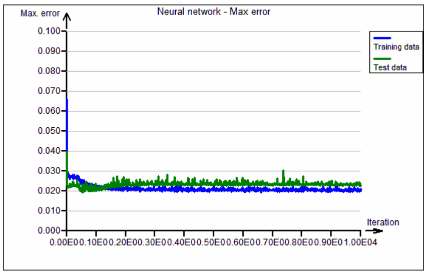




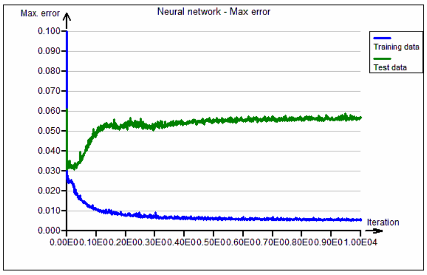
A



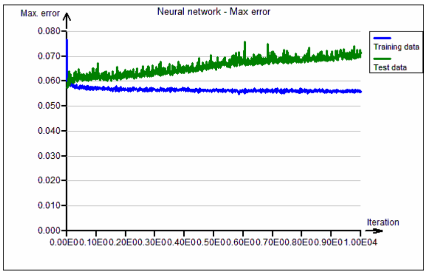
B



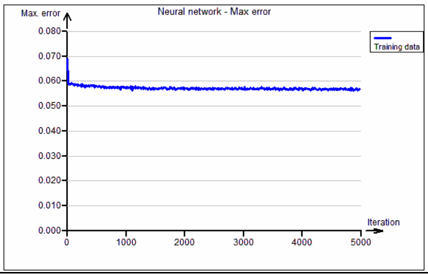
C



D



E

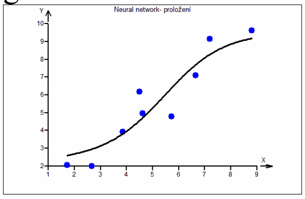


Plot of the training (network optimization) process, which decrease generally the sum of squares of differences between prediction and the actual values of dependent variable, with the number of iterations on x-axis as described above. If ths cross-validation is chosen the prediction error is ploted as well (green). According to the development of the maximum error of prediction curve we can assess the quality of the model and data.

Figure A This plot shows a typical successful training process, which gradually improves model for the specified data without validation.

Figure B The curve dropped steeply from 0.1 to about 0.02 which shows good quality of the model both in fitting the training data and in predicting the validation data, both curves are roughly on the same level. This is an optimal result if we intend to use ANN for predicting response from new data.

Figure C Plot with the same data as on Fig. B but with more complex ANN. The fit of the training data is considerably better than in the previous case (error about 0.006) but this is at the price of much worse prediction capability of the model (error 0.055). Such overdetermined model only fits the given data but provides poor prediction ability for any new data. It would be suitable just to interpolate existing data without crossvalidation. A simple example of a correct (left) and overdetermined (right) models is given below.



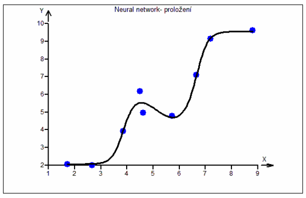


Figure D The curve shows little improvement of error (from 0.08 to 0.06). This may be due to the lack of any dependence in the data, or too simple model that is unable to explain the data, or too complex model, which failed to optimize.

Figure E A similar example shows a similar situation as Fig. D without validation.

18.2. Partial least squares

Menu:	QCExpert	Predictive methods	Partial least squares

Module PLS regression provides the user with one of the best computational tools for evaluating a pair of multidimensional variables, which is expected to have linear relationship inside one or the other multidimensional variable, and linear relationship between the two variables with each other. This computationally intensive methodology allows to explain and predict one of the variables using other group of variables. The PLS regression method found a large number of applications in the planning and management of quality in manufacturing technology, design and optimization of the characteristics of products in the development of new products, marketing studies, research in the evaluation of experiments, in clinical trials. An example might be modeling the relationship between technological parameters in the production and product quality parameters, or between the chemical composition and physical and biological characteristics. The typical questions of technological practice, which PLS can often answer include:

It has a purity of the raw material any effect on the strength of the product?
What happens if the temperature is increased in the process?
Can we increase the stability of the product by reducing the speed or rotation?
Which process parameters affect the most product strength?
How to set the value of procedural parameters to achieve the desired product characteristics?
What caused the decrease in the parameter?
In what and how subsequent production batches differ?
How to improve the stability / quality?
How to increase the strength / value / competitiveness?
Which input parameters are crucial for the quality?
Which process parameters are crucial for the quality?

Mathematical basics of the PLS regression method

Let us denote **X**(*n*x*p*) the matrix (table) of measured values of *p* variables (columns) with *n* lines and denote **Y**(*n*x*q*) the corresponding table with the same number of lines *n* but with *q* variables. Center all columns (substract column average from each column).To extract maximum information from the *p*- *q*- dimensional matrices to a lower dimension space, we decompose **X** and **Y** to the product of the orthogonal matrices **T**(*n*x*k*) and **U**(*n*x*k*), with coefficient matrices **P** and **Q**

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

while maximizing the correlation between **T** and **U**. Required dimension k, $1 < k \leq \min(p, q)$ is chosen by user, for example, on the basis of the squares sum decrease (scree plot), see below. Noise and irrelevant information contained „litter“ in every measured data is swept into residual matrices **E** and **F**. Decomposition **U** = **TB** (where **B** is a square diagonal matrix) give us a tool for computing (estimating) **Y** from **X** but also **X** from **Y**, just by switching the **X** and **Y** data because the model PLS-R is symmetric.

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{Q}^T,$$

T is calculated from the new data **X**, $\mathbf{T} = \mathbf{X}\mathbf{P}^{-}$ ($\mathbf{P}^{-}$ indicate generalized Moore-Penroseovu pseudoinversion of a rectangular matrix **P**). Furthermore, there is an internal link between **X** and **Y**. By writing $\mathbf{W} = \mathbf{B}\mathbf{Q}^{\mathrm{T}}$, we can rewrite the original pair of relations in the form

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{TW} + \mathbf{F}$$

so that the data **X** and **Y** are linked through a common scores matrix T, which is actually ortogonalized original matrix **X** in generally smaller number of dimensions, with a maximum extracted information contained in the original **X**, with removed noise (which is moved to the matrix **E**), while the maximum covariancewith the matrix **Y** is ensured. Using the relation

$$\mathbf{A} = \mathbf{P}^{-}\,\mathbf{B}\,\mathbf{Q}^{T}$$

we can reconstruct coefficients of a classical regression model with multivariate response **Y** = **AX**. Columns $\mathbf{a}_i$ of **A** contain linear coefficients (absolute terms are zero thanks to centering data) of the models $\mathbf{y}_i = \mathbf{X}\mathbf{a}_i$, where $\mathbf{y}_i$ is *i*-th column of matrix **Y**. The coefficients are not usually fully numerically identical to coefficients obtained by classical linear regression, see chapter 17.1. They are generally biased, but shrinked, which means that they have lower variances, and are generally more stable.

As mentioned above, this method is looking for a relationship between the two phenomena described by multidimensional numerical vectors. A typical example is **X** matrix containing measured technological parameters in the production of individual units or batches and matrix **Y** containing the relevant physical parameters of finished products, their deviations from the specifications, etc. Another example is a matrix **X** containing climate and chemical descriptions of the various sites and **Y** matrix with biological parameters of micro-organisms, vegetation and fauna in these locations. There are many other applications in geology, biology, toxicology, chemistry, medicine, psychiatry, behavioral sciences, pharmacology, cosmetics, food, steel industry to name just a few. With PLS prediction we can then obtain estimates of unknown quantities **Y** on the basis of known values **X**.

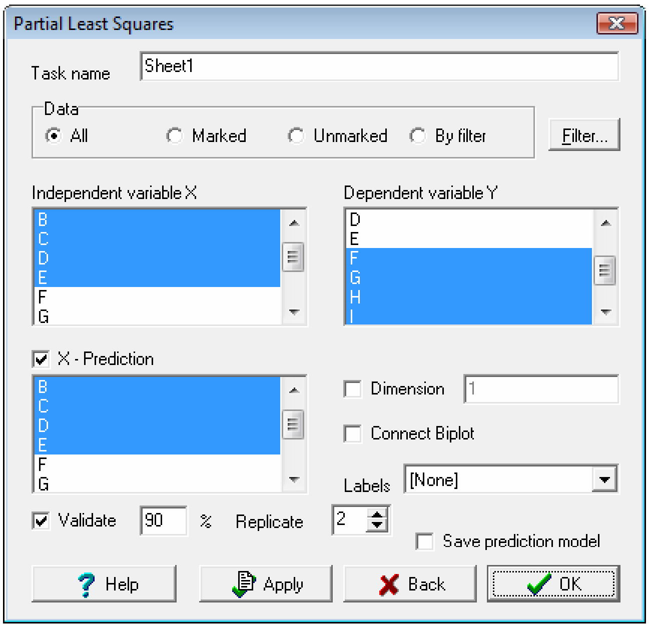
Model validation

The prediction quality of a particular PLS model can be assessed on the basis of its ability to predict the value of **y** from the value of **x**. This is used in various validation procedures, sometimes called cross-validation. The principle of validation of the model is the same as in the case of neural networks. We „hide“ part of the data before computation of the PLS model. This hidden data are called test or validation data. For the rest of the data, called training data, we calculate parameters of the PLS model. Then the validation are „unhidden“ and used to and check whether the model correctly predicts validation data. Validation must have the same nature and range of values **x**, and therefore the same model as the data used for training. For the validation data we then construct diagnostic charts, which simply conclude whether the model is appropriate for all data. If the model describes well only the training data and not validation data, this usually means that we have little data (rows), or that we have chosen is too large proportion of validation data. A proportion between 10 and 40% of the validation data is usual.

Of course, even advanced PLS regression method is not miraculous and has certain restrictions, which is mainly assumption of linearity of all relationships and normality of error distribution. Along with the ability of prediction and graphical diagnostics, however, it provides a very powerful tool for analysis and prediction of multidimensional variables. In quality control, thanks to its prediction capability, PLS regression is an ideal tool for the quality planning, design of products, optimization of technologies and applied research.

18.2.1. Data and parameters

Module PLS regression needs two data tables, matrix **X** with *p* columns and **Y** with *q* columns selected as the dialog box items *Matrix X* and *Matrix Y*, see Fig. 96. The matrix columns must contain numeric data only, the number of rows must be the same for both **X** and **Y**. Each matrix must contain at least two columns. Columns of matrix **X** must not appear in the matrix **Y**. The limiting dimension *k* can be set by user. If the box *Dimension* is not checked, the maximum dimension is set to k = min(p, q). It is recommended to perform PLS in maximal dimension first, then optionally we can determine an appropriate value of k using the scree plot (see paragraph Graphical output below) and repeat the computation again with new k. If the checkbox *Connect Biplot* is checked, consecutive points of the Biplot will be connected in the order of data in spreadsheet. This can help to follow a possible trajectory of the process. If the checkbox *X-Prediction* is checked, it is necessary to choose the same number of columns as in the field *Independent variable X*. These variables will be used to compute the predicted values of the dependent variable. The X for the prediction must have the same number of columns as the independent variable matrix **X**, but may have a different number of lines (at least 2 lines). A typical example of input matrices and data for prediction is given on Fig. 97.



Partial Least Squares

Task name: Sheet1

Data: ● All ○ Marked ○ Unmarked ○ By filter

Independent variable X: B, C, D, E (selected); F, G

Dependent variable Y: D, E; F, G, H, I (selected)

☑ X - Prediction: B, C, D, E (selected); F, G

□ Dimension: 1

□ Connect Biplot

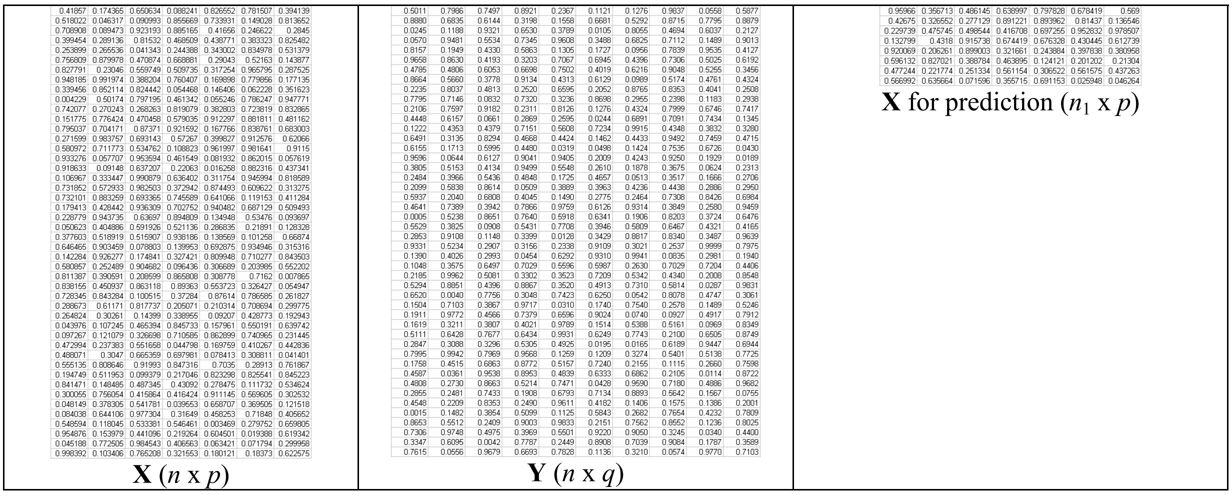
Labels: [None]

☑ Validate 90 %

Replicate: 2

□ Save prediction model

Fig. 96 Dialog box for PLS regression



X (n x p)
Y (n x q)
X for prediction (n1 x p)

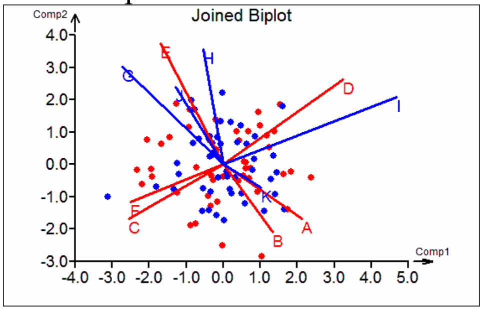
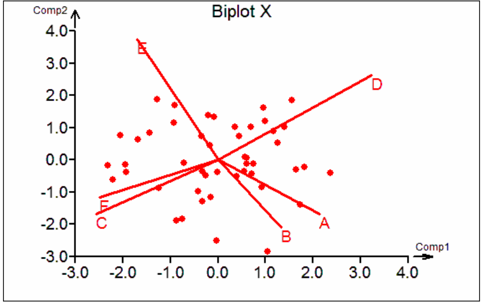
Fig. 97 Typical data for PLS regression

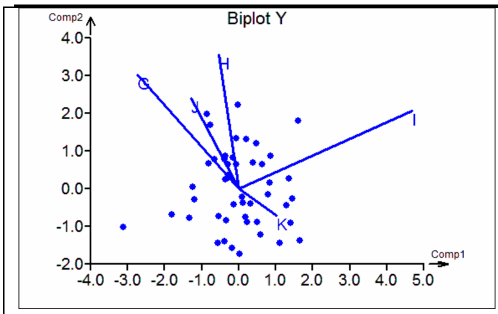
18.2.2. Protocol

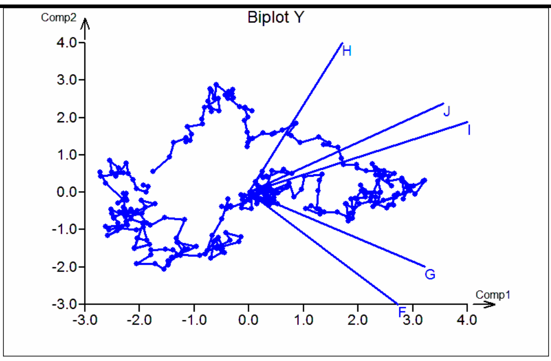
Input data	
No of rows	Number of valid rows
No of columns	Number of columns of **X** a **Y** matrices.

Columns	Column names of both input matrices.
Chosen dimension	Dimension *k* for the PLS model chosen in the input dialog window.The dimension must be less or equal to min(*p*, *q*). Scree plot may be used as an aid to select suitable *k* if required.
PLS - coefficients, B	Diagonal elements of the matrix **B**.
Explained sum of squares	Table of the squares sum of residuals with growing dimension of the model, *i* = 1, … *k*, these values are used for constructing scree plot.
No of components	Number of components (dimensions) used for the squares sum.
RSS	Residual square sum value, for 0 components the RSS is the total squares sum without a model.
Percent	% of the RSS
Explained %	(100 – %RSS).
Loadings X, P	Loadings matrix **P**.
Loadings Y, Q	Loadings matrix **Q**.
Regression coefficients, A	Matrix of regression coefficients a_{ij} formally similar to those in the separate classical miltiple linear regression models **Y** = **XA**, or $\mathbf{y}_j = \Sigma a_{ij}\mathbf{x}_i$. The coefficient values are generally different from the classical coefficients, since they are based on the orthogonal component regression and therefore they are biased, shortened (with lower standard deviations) and more stable.
Prediction	Predicted values for the data selected in the fielf X-Prediction in the PLS dialog box. This part of output is not generated unless the checkbox is checked.

18.2.3. Graphical output

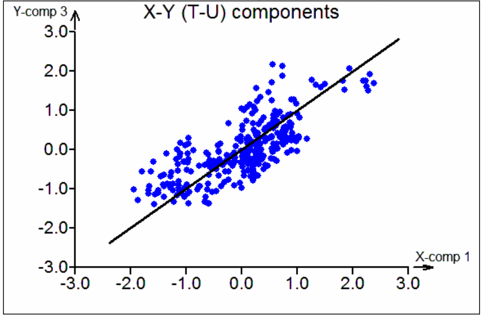
Joined Biplot  Separate Biplots  	Bi-plot for the matrices X and Y in one plot, matrix X and Y separately. Biplot is a projection of multidimensional data in the plane (the best one in terms of least squares). Points represent rows, rays correspond to columns of the original data. To identify the data rows you can use the labels of points selected in the dialog. Close vector lines (rays) are likely to be mutually correlated. Points located in the direction of a ray will have bigger value of the respective variable. You should be aware that, due to the drastic reduction in the number of dimensions, particularly for larger *p*,*q* this information and guidance is rather a global assessment of the structure and possible links and relationships in the data. If the checkbox Connect Biplot was checked, the points in the plot are linked in chronological order, which sometimes makes it possible to identify and spot trends in the time-series or non-stationary process, as shown on the illustration below.





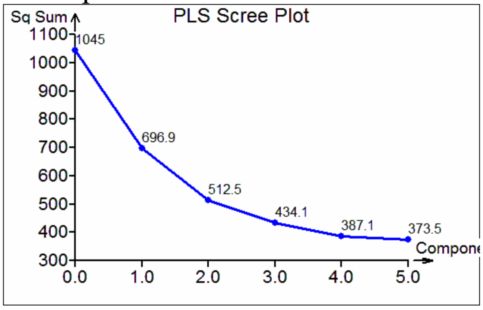
Connected biplot makes it possible to spot „wandering“ of the process in time.

X-Y Components agreement plot



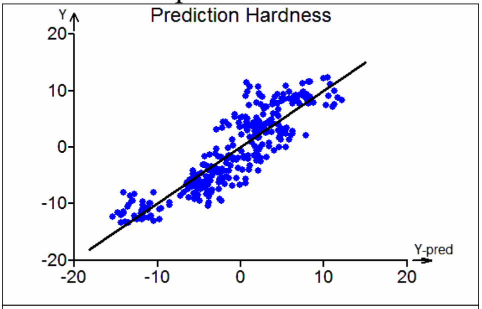
Plot of agreement between the columns of **T** and **U**. This graph shows the global success of a PLS model fiting. The closer the points to the line, the more successful a PLS model is.

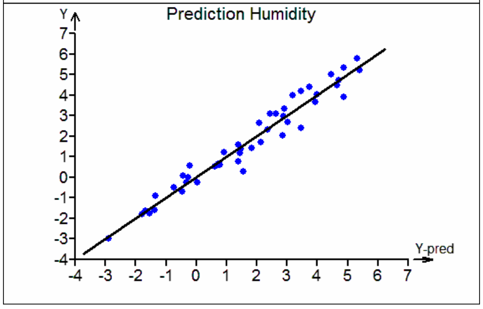
Scree plot



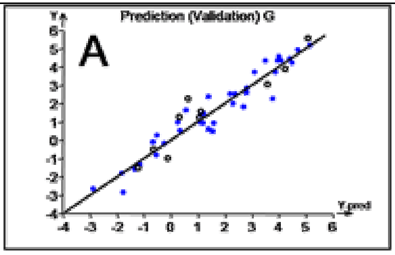
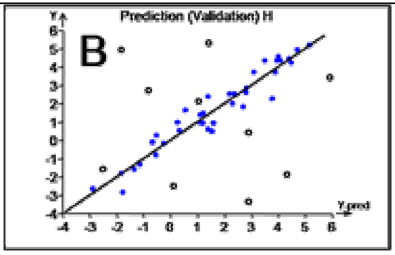
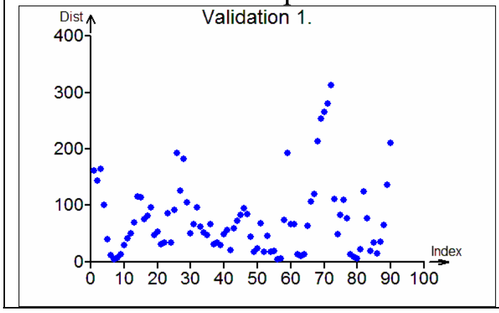
The effectiveness of the model expressed by reduction of unexplained (residual) sum of squares, depending on the number of factors included (columns of matrix **T** and **U**).

Y-Prediction plot





This plot expresses compliance of dependent variables and model prediction. The closer are the points to the line, the better the fit. This plot is created for each dependent variable. Some variables can be predicted better, others worse. If the plot does not show a visible trend, the suitable model for this vartiable was probably not found, a model is not able to predict dependent variable. If Validation checkbox was selected, the validation (test) points in the plot are marked in red (in the example below marked empty circles). In the Fig A below both the training and the validation data are fitted well, showing the model is reliable and the dependence is real. However, if the validation data strongly disagree with other data, as in the Fig B below, the PLS model may be overfitted, describing only the training data, and probably is not usable for the prediction of new values. It is advisable to try to reduce the dimension of the model by entering numbers less than min(p, q) into the field dimension.

	 Good prediction  Poor prediction
Validation residuals plot 	Plot Validation is used to assess the quality of prediction of validation data. Unique very remote points may represent outlying measurements. On the Y-axis are Eukleidian distances of the data from the model.

19. Calibration

Menu:	QCExpert	Calibration

The Calibration module is most useful for analytical laboratories and metrological departments. It contains linear and nonlinear calibration models. Automatic detection of departures from linearity can be requested. Due to the fact that the module implements weighted regression, it can be successfully used for models with heteroscedastic errors. This feature can be useful namely for analyzing low-level measurement data (e.g. trace analysis).

Simple calibration models work with two variables. First is the measured variable of ultimate interest *X*, (e.g. concentration, viscosity, temperature). Second is the measurement device response *Y*, e.g. absorbance, voltage, resistance, number of particles. Generally, the calibration problem consists of two parts: (a) calibration model construction, and (b) application of the model constructed previously. When constructing the model, one uses measurement device responses: $y_1, y_2, y_3,\ldots$ to known values of the measured variable of ultimate interest, *X* (usually administered in the form of certified standards): x_1, x_2, x_3,… . Dependence of *Y* on *X* is expressed by a regression model, which describes the relationship between known x_i and experimentally evaluated device responses y_i in the best way. QC.Expert™ uses either linear or quadratic regression model for calibration. The computations are based on direct application of either weighted or unweighted regression *Y* on *X*. Compared to inverse regression (*X* on *Y*), the direct regression is more appropriate both statistically and logically. The regression fit encompasses step (a). Fitted model is applied when one looks for "the best" estimate of the unknown value of interest, *X*, based on one or more device response records, *Y*. This estimate, based on inversion of the regression relationship should be always accompanied by some form of uncertainty assessment, e.g. in the form of $(1 - \alpha)\%$ confidence interval ($\alpha = 0.05$ is selected very often in practice). Width of the confidence interval for *X* is related to the precision with which regression parameters are estimated (and hence to the confidence band width in the regression step (a)). Further, calibration limits related to noise variability and minimum reliably measurable value are computed (critical value, detection limit and quantification limit).

19.1. Data and parameters

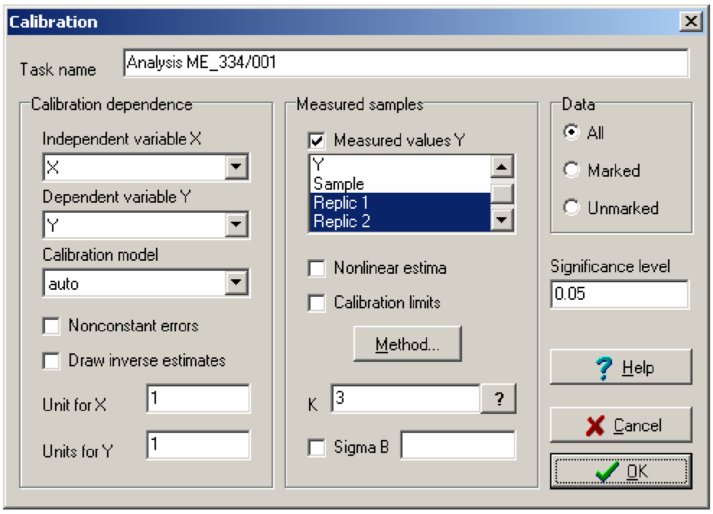
The calibration module expects data in the form of a two-column-table. In the regression terminology, one column contains *X* , the explanatory variable, while the second column contains corresponding device responses *Y* , the dependent variable. These two columns have to be specified and used subsequently when entering analysis requests. If one wants to use the calibration model fitted on these two columns in order to estimate *X* for some additional unknown samples, their recorded responses *Y* have to be entered in additional columns. Recorded responses *Y* of individual unknown samples are entered as individual rows of one column. If there is more than one recorded response per one sample, additional replications should be listed as additional cells on the same row. The following table serves as an example of the situation where we had 5 calibration standards with values *X*=1.281, 2.558, 5.430, 7.373 and 11.59. The first four standards were measured repeatedly (twice each). After the regression model was fitted, we used the resulting calibration to analyze 4 samples from Czech rivers Upa and Labe denoted as *Upa A*, *Upa B*, *Labe AE*, *Labe AR*. The first two of these samples were analyzed repeatedly (three times each, replication denoted by *Replic1* to *Replic3*), next two samples were analyzed only once. The column denoted by *Sample* is intended only to hold comments only; it cannot be selected for further operations in the dialog panel. If calibration relationship estimation is all what is needed, entered data table will consist of the first two columns only. Additional data are not required then.

Table 1 Calibration data example

X	Y	Sample	Replic 1	Replic 2	Replic 3
1.281	25.53	*Úpa A*	33.69	33.74	33.73
1.281	25.58	*Úpa B*	39.25	39.25	39.27
2.558	51.37	*Labe AE*	50.6		
2.558	51.23	*Labe AR*	57.3		
5.430	106.4				
5.430	108.7				
7.373	148.4				
7.373	146.6				
11.59	233				

Dialog panel selections for the example just described is in the Fig. 98.

Project name is a text string, originally taken from the name of the data containing spreadsheet. It can be edited. The finally selected Project name will appear as a header in the resulting protocol. The *Calibration relationship* part specifies calibration model type. Here, the user must specify explanatory variable (*X*, values of the certified standards) and the dependent variable (*Y* , the measurement response recordings). In addition, a calibration model type has to be specified as either linear (calibration relationship is linear) or quadratic. Quadratic model is the simplest model allowing for curvature (nonlinearity) in the calibration relationship. When the *auto* choice in the *Calibration model* selection is invoked, automatic linearity check will be performed. This is done as follows: quadratic model is fitted first. A statistical test is used to test whether the quadratic term is significantly different form zero. If it is statistically significant, quadratic model is used further. If it is not significant, all subsequent calculations are based on linear model. We strongly recommend using the *auto* choice, if the user is not sure about calibration model type. *Heteroscedastic errors* selection should be checked when one suspects that error variability of the *Y* reading depends on *X*. Heteroscedastic errors are quite common for instance when the calibration model is fitted across more than one order of magnitude (e.g. in trace analysis). When a heteroscedastic model is invoked, the calibration model is fitted by iteratively weighted regression procedure (IRWLS). The weights are then given as reciprocal values of the predicted residual variance, computed via nonparametric regression. The predicted variability can be inspected visually in the absolute residuals plot. Heteroscedastic model tends to give narrower confidence band in the intervals where the measuring device readings are more precise. If this increased precision occurs for *X* close to zero, heteroscedastic model based detection limits tends to be smaller.



Task name: Analysis ME_334/001

Independent variable X: X

Dependent variable Y: Y

Calibration model: auto

☐ Nonconstant errors

☐ Draw inverse estimates

Unit for X: 1

Units for Y: 1

☑ Measured values Y: Y, Sample, Replic 1, Replic 2

☐ Nonlinear estima

☐ Calibration limits

K: 3

☐ Sigma B

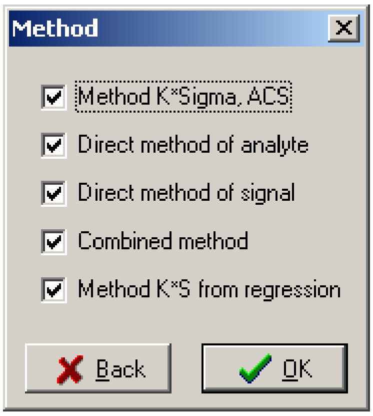
● All ○ Marked ○ Unmarked

Significance level: 0.05

Fig. 98 Calibration module dialog panel

One should mark the *Plot the inverse estimates* selection, if inverse estimates are requested in the calibration plot. You should not use this choice if there are many calibration points and/or when confidence band of the calibration model tends to be wide, since then the resulting plot is hard to read. Text appearing in the *X units* and *Y units* fields is self-explanatory, denoting measurement units of these variables. It does not influence any calculations. It appears only in the final report of results. If they are not needed, the two fields can be left blank.

In the *Measurement device reading* part, the item *New samples reading* should be selected if one wants to estimate unknown *X* from additionally recorded *Y*. After selecting this item, roll-down menu can be opened, offering names of all columns that can hold the new device response readings. Indirect (sometimes imprecisely denoted as nonlinear) estimates are specially constructed estimates of unknown *X*, derived for the situation of the so-called statistical calibration, when both *X* and *Y* are considered to be random variables and the data are viewed as a two-dimensional cloud. The *Indirect estimates* selection invokes their calculation. This method can be used both in linear and nonlinear situation. In any case, the results are rather imprecise and confidence intervals for the *X* are not computed. These estimates should not be used when the calibration relationship is strong. When you mark the *Calibration limits* selection, resulting protocol will contain the critical value, detection limit and limit of quantification. Taken together, these values are referred to as calibration limits. QC.Expert™ offers five methods to calculate these limits, based on various literature sources. When the *Method…* button is pressed, a menu appears (Fig. 99), where one can select which of the methods should be computed. At least one method has to be selected.



☑ Method K*Sigma, ACS

☑ Direct method of analyte

☑ Direct method of signal

☑ Combined method

☑ Method K*S from regression

Fig. 99 Calibration limit calculation method

In the *K* field, the *K* coefficient can be entered, which will be subsequently used for the trivial calibration limits calculation method: K*Sigma. Typically, *K*=3 is selected often in practice. Since this coefficient has the meaning of the normal distribution quantile, its value should correspond to a chosen significance level α in order to keep the results comparable with the other methods. The „*?*“ button can be used to compute the value of *K* which corresponds to a selected significance level (e.g. for $\alpha = 0.05$, we have $K = 1.96$). On the other hand, if we insist on using $K = 3$, we should change significance level to $\alpha = 0.0027$ in order to be consistent across various methods of calculation and to get comparable results. When the standard deviation of the blank σ_{blank} is known (that is the standard deviation of the measurement device signal obtained without adding any sample, i.e. when X=0), the *Sigma B* selection is marked and σ_{blank} value entered. In the *Data* part, one can (as in other modules of the software) choose whether all data, or marked row data, or unmarked data will be used for computations. Data rows can be marked, using the button in the upper bar. *Significance level* must be a value smaller than 0.5 and larger than 0. It is used for all tests, and for calculation of (1-α)% confidence limits and calibration limits.

Further, we describe briefly various methods of calibration limits calculation and list their definitions. Since this is an analytical chemistry material, we use a common chemical terminology.

Y_C … critical level of *Y*. The smallest value of *Y* that can be reliably distinguished from noise. (a *Y* value, which is exceeded by noise with probability smaller than α). Values smaller than Y_C are considered to consist of noise only, respectively to be the blank readings.

Y_D … detection limit of *Y*. Analyzed substance can be safely proved (with probability 1−α) when it gives measuring device reading above this value. Probability of obtaining the reading $y > Y_D$ under the blank measurement condition is smaller than 1−α.

Y_Q ... quantification limit of Y. The value, above which true *Y* value can be estimated with the relative error, smaller than α. Quantitative analysis should not be conducted for samples giving measurement device readings under this limit.

X_C … critical value of *X*. It is tied to Y_C through the calibration model.

X_D … detection limit of *X*. Minimum value of *X* (e.g. concentration, weight) detectable by the given method.

X_Q … limit of quantification for *X*. Minimum value of *X*, which can be estimated with the relative error smaller than α. Only the *X* values above X_Q should be estimated quantitatively.

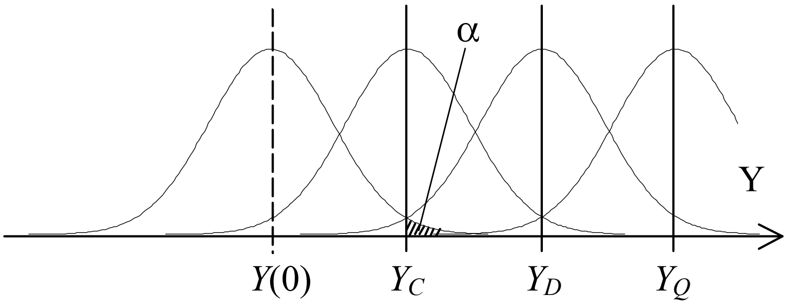


Fig. 100 Schematic draws of Y_C, Y_D and Y_Q

For K*Sigma method, see Figures Fig. 100 and Fig. 101.
$Y_C = K.\sigma$, $Y_D = 2K.\sigma$, $Y_Q = 3K.\sigma$
Sometimes, the $Y_Q = 10/3K.\sigma$ is used. For K=3, it corresponds to the 10 σ units. We suggest to choose *K* as the (1−α) – quantile of the standard normal distribution in order to keep the results comparable to the results of other methods. This method does not provide critical values of *X*. One can obtain them informally from the calibration plot, however. Estimate of σ is obtained either as the blank’s standard deviation, or as the square root of residual variance from calibration model fitting.

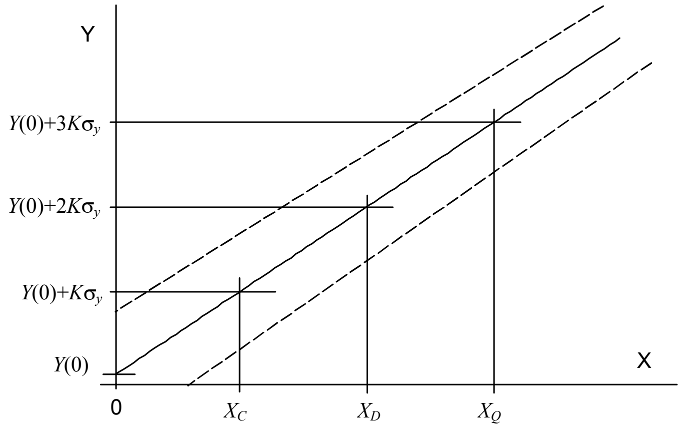


Fig. 101 Method K*Sigma

The following three methods use fully statistical properties of the calibration model and hence, they can be used to compute also critical values of *X* correctly. They are typically smaller (and hence more desirable) than those obtained from the K*Sigma method. The direct analyte method (Fig. 102) uses confidence intervals of *X* estimates.

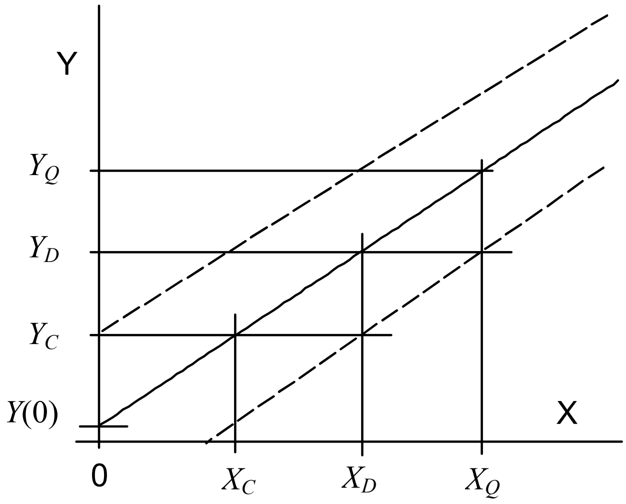


Fig. 102 Direct analyte method

The direct signal method uses confidence intervals for *Y*, Fig. 103.

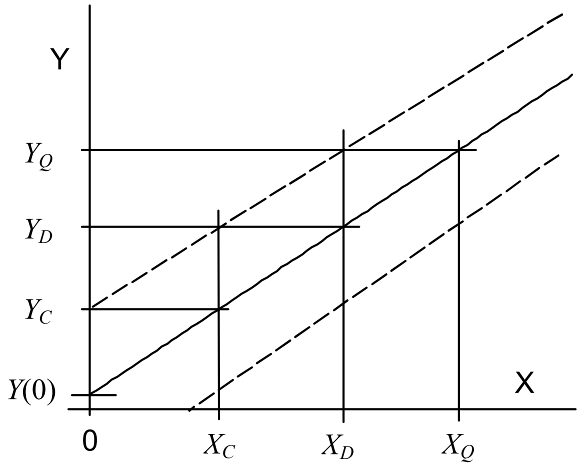


Fig. 103 Direct signal method

The Ebel and Kamm combined method combines elements of the two previous methods, Fig.104.

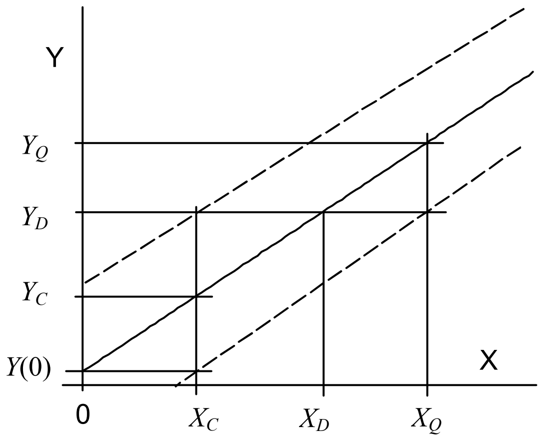


Fig.104 Combined method, Ebel and Kamm

The last method implemented, *K*Sigma from regression* is similar to the first one (*K*Sigma*) only with the distinction, that for $K\sigma$, half-width of the confidence interval for situation with x=0 is used. That is the half-width of the regression confidence band at x=0, for a given significance level.

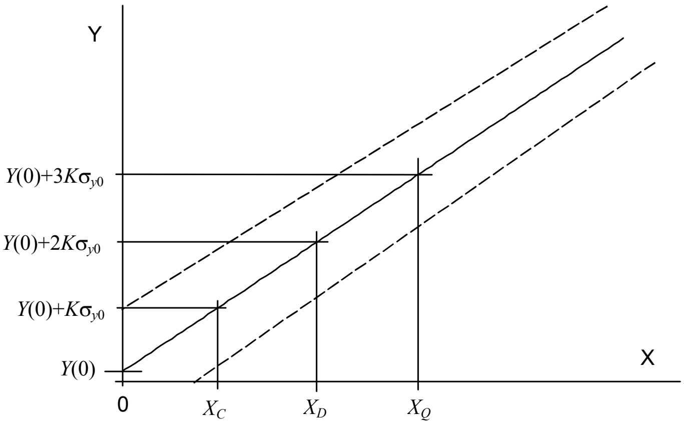


Fig.105 Method K*sigma from regression

Estimation by inversion. The main practical purpose of the calibration procedure is to be able to estimate unknown X from the recorded measurement device output Y. The required estimate is obtained by calibration relationship inversion (given by the previously estimated calibration model). One important thing to be kept in mind is the fact that, since the X is obtained from the random variable Y, it is random variable as well. Hence, it is not enough to report the point estimate itself, some measure of uncertainty should be attached. One possibility is to use the confidence interval (say the 95% confidence interval). Fig.106shows an example of the X estimate construction by the inversion of the calibration relationship. If we have some information about variability of the Y_i reading for the particular sample (obtained for instance from the repeated readings Y_{ij}, j=1,…n$_i$, then we are able to get a more realistic (even though sometimes wider) confidence interval for Y_i, and hence a more realistic estimate confidence interval for X_i. When the interval is constructed in this, more elaborate way, it reflects both variability related to the uncertainty of calibration relationship estimation and the current Y measurement variability, which is connected to a particular sample, see Fig.107. This is one of the reasons why it is so important to replicate calibration measurements if it is possible. Whenever possible, the measurement outcome should be given in the form of interval ($x_{0.025}$, $x_{0.975}$), possibly listing the point estimate x_i as well. ***Remark:*** Confidence interval for the X estimate (obtained by the inversion of the calibration relationship) is not symmetric around x_i in general.

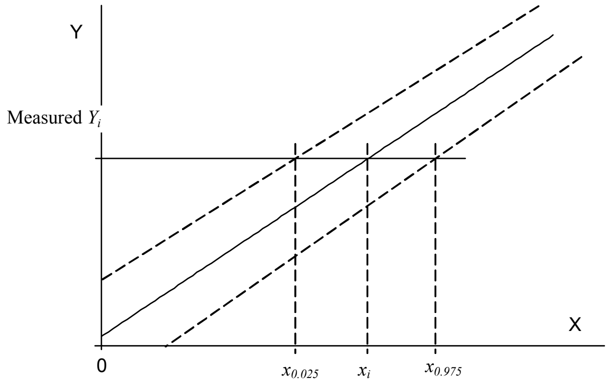


Fig.106 Estimation by inversion for one measuring device reading

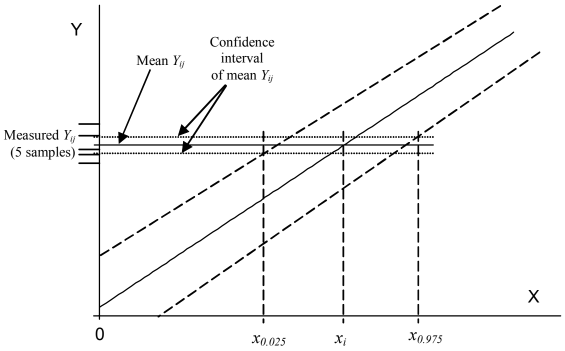

Y
Mean Y_{ij}
Confidence interval of mean Y_{ij}
Measured Y_{ij} (5 samples)
X
0
$x_{0.025}$
x_i
$x_{0.975}$

Fig.107 Estimation by inversion for repeated measuring device reading

New method validation. The Calibration module can be successfully used to validate a new method through comparison with another method, established and validated previously. For this purpose, the established (or validated, or certified) method's results are entered as *X*, while the new method's results are entered as *Y* (entered so that *X* and *Y* values corresponding to the same sample appear on the same line). The samples, on which the *X* and *Y* pairs are measured, should cover densely whole range in which the new method is to be validated. When computation request is specified in the Dialog panel, automatic calibration model type selection (or *"auto"*) should be marked. The new method is validated, if the *linear* model is selected; its intercept is not significantly different from *zero*, while it's slope is not significantly different from *one*. Appropriate slope and intercept test results can be found in the *Intercept significance* and *Slope validation* paragraphs of the Protocol. Alternatively, one can use the module *Two-sample comparison – Paired comparison*, see the respective paragraph.

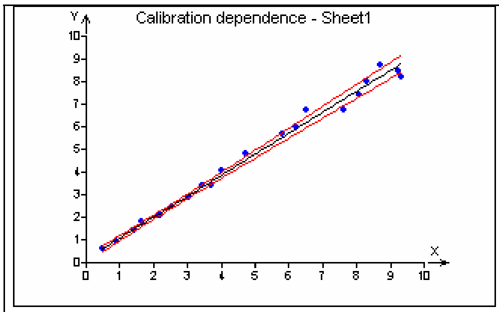
19.2. Protocol

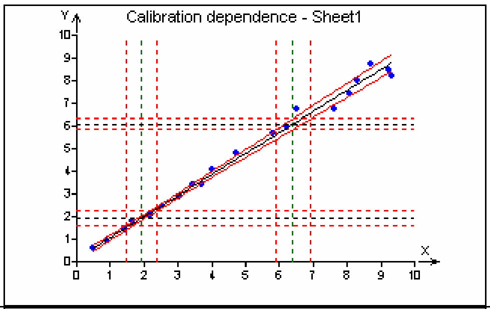
Project name	Name of the project, taken from the Dialog panel
Sample size	*n*, number of valid *X,Y* pairs used for calibration model estimation.
Significance level	Significance level α.required and entered by the user.
Calibration model selection	The requested way of calibration model selection (manual or automatic).
Calibration model type	Calibration model type used (linear or quadratic). Number of degrees of freedom ν for a given calibration model is equal to the sample size minus number of estimated parameters. For linear model, $\nu = n - 2$, for the quadratic model $\nu = n - 3$.

Is the model used feasible?	Linearity test. If we use (in the manual mode) linear model for data showing a substantial nonlinearity; or on the contrary, if we use quadratic model for linear data, we make a mistake that manifests by the measurement error increase, at least. If such a problem is detected by the statistical test implemented in the software, Protocol contains the word "*not feasible*"; otherwise it contains the word "*feasible*".
Weighted regression used?	Indicates, whether the weighted regression was used (*Yes* or *No*), that is whether a heteroscedastic errors model was invoked effectively.
Calibration model parameters	Information about calibration model parameters. If the model is linear, intercept "Abs" and slope (linear trend coefficient) will be reported. If the model is quadratic, the quadratic term coefficient is reported as well. Linear term does not have the first derivative meaning in that case.
Parameter	Parameter name: "Abs" = intercept, X=linear term coefficient, X^2=quadratic term coefficient
Estimate	Parameter estimate.
Std. deviation	Standard deviation of the parameter estimate (its standard error).
Lower limit	Lower limit of the 1-α confidence interval.
Upper limit	Upper limit of the 1-α confidence interval.
Intercept significance	Test of the null hypothesis that intercept is equal to zero. Result of this test is interesting when validating a new method via comparison with an established method, among other situations.
Value	Intercept value, copied from the *Calibration model parameters* paragraph
Conclusion	Conclusion of the significance test. If the intercept is *not significant*, than we have no particular reason to believe that the calibration curve (or line) does not go through the coordinate origin (more precisely: we cannot reject the zero intercept hypothesis at the significance level α). Even if not significant, the intercept should remain in the model (the *Calibration* module, unlike the *Linear* regression module does not allow models without intercepts anyway). This is on purpose, since models without intercepts can lead to unwanted confidence interval distortion at x=0 and to the situation, where it is impossible to compute calibration limits. When the intercept test is *significant*, it is possible to reject the null hypothesis about its' population value being zero, accepting the alternative that the calibration curve/line does not go through origin, i.e. that $Y(X=0) \neq 0$.
Slope validation	This test is very useful when validating a new method (respectively, when comparing two methods).
Value	Linear term coefficient value (when the model is linear, this is the derivative of the calibration relationship). It is copied from the *Calibration model parameters* paragraph.
Linear term coefficient=1	Conclusion of the unit slope hypothesis test (*Yes* or *No*). If the model is linear, the tested coefficients have meaning of the calibration relationship derivative.
Method sensitivity	Sensitivity of a particular method is defined as the measuring device response (*Y)* change, when *X* is changed by one unit. When model is linear, the sensitivity is equal to the slope. When model is nonlinear (quadratic), the sensitivity is given by the derivative of the calibration relationship and changes with the X value. Therefore, the software gives the sensitivity at four important points: at $x = 0$, at the lowest data value **min**(x), in the middle of measured data range – i.e. at (**min**(x) − **max**(x))/2, and at the highest data value **max**(x).
Selected K	Selected K for use in the K*sigma method.

Blank signal standard deviation	The value entered as Sigma B in the Dialog panel (if that was entered). That is the user-inputted value of σ_{blank}.
Computed blank signal standard deviation	If the blank signal's standard deviation is not entered in the Dialog panel, residual standard deviation is used instead. This value is used for calibration limits calculation by the K*Sigma method. Residual standard deviation is usually higher than σ_{blank}, the K*sigma does not give reliable results then.
Calibration limits	Critical value, detection limit and limit of quantification for *Y* and *X*. They are computed by the following methods: K*Sigma, direct analyte method, direct signal method, combined Ebel-Kamm method and K*Sigma from regression method.
Yc, Yd, Yq, Xc, Xd, Xq, Yq(10sigma), Xq(10sigma)	Yc = Critical value for *Y*, Yd = detection limit for *Y*, Yq = limit of quantification for *Y*. Xc = Critical value for *X*, Xd = detection limit for *X*, Xq = limit of quantification for *X*. Yq(10sigma) and Xq(10sigma) are alternatives to the quantification limit. For *K*=3, they correspond to 10σ, while Xq and Yq correspond to 9σ, for *K*=3.
Calibration table	This paragraph collects the results, whose computation gives main motivation to the calibration procedure. *X* estimates for a new, unknown sample are obtained by the estimated calibration relationship inversion from the measuring device response reading *Y* . If the data spreadsheet does not contain any *Y* readings, or when *New samples reading* was not selected, this paragraph does not appear in the Protocol.
Sample number	Integer indicating the sample number.
Estimate of *X*	Estimate of *X* by inversion.
Lower limit	Lower limit of the 100(1−α)% confidence interval of *X,* computed by inversion.
Upper limit	Upper limit of the 100(1−α)% confidence interval of *X,* computed by inversion.
Indirect estimate	Indirect estimate from the so-called statistical calibration. This estimate is computed only when the *Indirect estimate* was selected and when the variability of Y is large enough. Otherwise, column of zeros is printed here.
New samples readings	New samples readings of the measuring device, if they were entered. The NA acronym (not available) denotes missing data.
Residual analysis	Analysis of the residuals after the model.
Residual sum of squares	Residual sum of squares.
Mean absolute residual	Average of absolute values of residuals.
Correlation coefficient	Correlation coefficient estimate. ***Remark:*** Correlation coefficient *cannot* be used to judge whether linear or quadratic model is appropriate for given data!
Measurement number	Integer denoting numbering the *X*,*Y* pairs.
Measured X	Value of the known standard, taken from the inputted data.
Measured Y	Value of the measurement device reading, taken from the inputted data.
Computed Y	Value of *Y* , computed for a given *X* from the calibration model.
Residual	Difference: (Measured Y − Computed Y)
Weight	Weight attached to a particular measurement. If the *Heteroscedastic errors* choice was not selected, this column contains ones only.

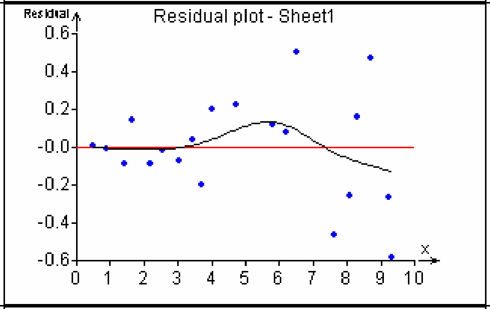
19.3. Graphical output



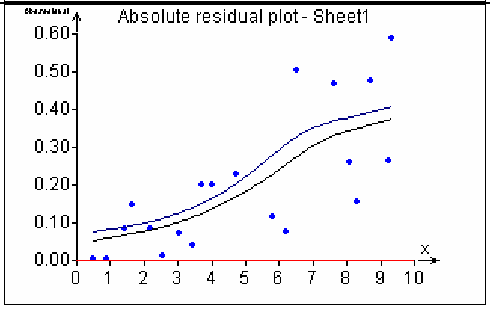


Calibration plot, showing estimated calibration relationship, together with the confidence band (solid red line). If there are measuring device readings *Y* for an unknown sample, and the *Plot inverse estimates* selection is marked, corresponding *X* estimates are plotted as well. Horizontal dashed lines correspond to Y values (in case of repeated readings, they correspond to average and the related confidence interval. Vertical lines correspond to *X* estimates and appropriate confidence intervals, obtained by inversion of the calibration relationship.

Upon double clicking on the plot, a new dynamic window is opened. In this window, further plot operations can be performed, see below.



Residual plot. Plot of residuals $e_{i,}$ fitted by a nonparametric, kernel estimate K(e_i) (black line). Substantial curvature can serve as a warning that the calibration relationship is nonlinear and is not described by the selected model satisfactorily. Observed curvature often relates to the presence of an outlier.



Absolute residual, |e_i| plot. Nonparametric, kernel regression estimate is superimposed (two dashed curves). The two curves depict estimates of standard deviation as a function of *X,* i.e. σ (x). The upper (blue) curve is given by the square root of the residual squares fit, √K(e_i^2). The lower (black) curve is given by the kernel fit of the absolute residuals, K(|e_i|). The upper curve tends to be a better estimate of $\sigma(x)$ (when the residuals behave normally). This is because $\sigma = \sqrt{\frac{1}{n}\sum \sigma_i^2}$, a quantity which is estimated rather directly in this case. The lower curve is more robust.

19.4. Interactive calibration plot

Double clicking on calibration plot invokes a new window with an interactive plot. This window can be used for inspection or even reading the X estimates, while keeping all other interactive plot features. While the mouse is moving *above* the calibration curve, *Y* coordinate is shown together with the corresponding *X* estimate and confidence interval obtained by inversion, as seen on the Fig.106. When the mouse is moving *below* the calibration curve, *x* coordinate is shown, together with the corresponding *Y* estimate, together with the confidence interval. When a particular detail is magnified (by zooming-in), plotted values can be read with a substantial precision. Nevertheless, when interested in *Y* estimates for a given *X,* or *X* estimates for a given *Y,* the "calibration calculator" should be used. The calculator can be invoked by clicking on the *Interactive estimates,* button.
The *Interactive estimates* window has 8 fields. *X* and *Y* cursor coordinates, relative to the interactive plot, appear on the uppermost line, originally. These values can be edited, however. Beneath the *X*

field, there is the corresponding *Y* estimate, accompanied by its confidence interval. On the other hand, beneath the *Y* field, there is the corresponding *X* estimate, accompanied by the appropriate confidence interval. By clicking on the X or Y field, *X* or *Y* value can be entered from keyboard. After pressing <Enter>, appropriate estimates are computed. For instance, if we want to compute *X* estimate for the measurement device reading $Y = 25.7$, we click on the *Y* field in the *Interactive estimates* window. Next, we erase the content of the *Y* field and enter the value 25.7. The request is submitted by pressing <Enter>. Subsequently, *X* estimate, lower (*X*–) and upper (*X*+) limits or the $100(1-\alpha)\%$ confidence interval obtained by the inversion of the previously estimated calibration relationship appear beneath the *Y* field subsequently.

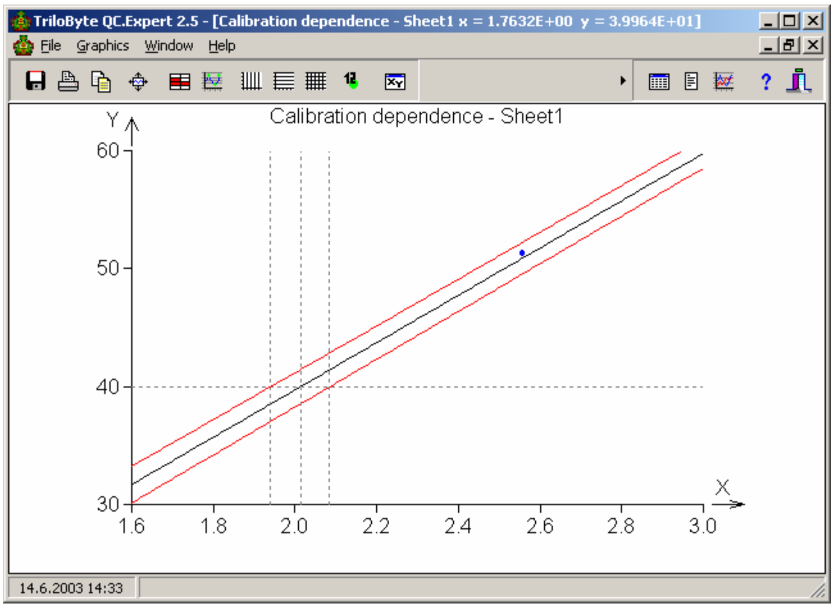
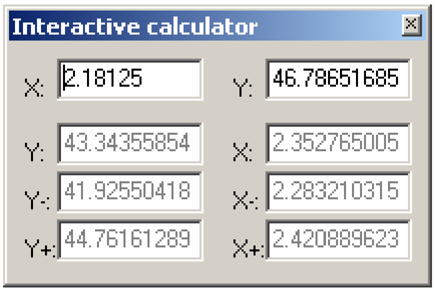

TriloByte QC.Expert 2.5 - [Calibration dependence - Sheet1 x = 1.7632E+00 y = 3.9964E+01]
File
Graphics
Window
Help
Calibration dependence - Sheet1
Y
60
50
40
30
X
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
14.6.2003 14:33

Fig. 108 Interactive estimates window with the calibration plot



X: 2.18125
Y: 46.78651685
Y: 43.34355854
X: 2.352765005
Y-: 41.92550418
X-: 2.283210315
Y+: 44.76161289
X+: 2.420889623

Fig.109 Interactive estimates

20. Shewhart control charts

Menu:	QC.Expert	Control charts

Control charts are constructed to decide whether a process is **under statistical control** and to monitor any departures from this state. This means that stability of some process properties over time is tested using certain statistical assumptions about the process (data it produces). Commonly considered properties are mean, variance (standard deviation), distribution shape or proportion of faulty items. The Shewhart control charts were invented in 1932. They are based on monitoring events, which are very unlikely when the controlled process is stable. Any incidence of such an event is taken as an alarm signal suggesting that stability of the process was broken and the process changed. Upon receiving such signal, possible causes of the change should be investigated and some correcting steps taken. One example of such an unlikely event is the situation when the control limits (UCL or LCL) are exceeded. They are constructed as ±3σ limits, so that when the process is under control, they are exceeded with relative frequency of 0.27%. In addition to the LCL and UCL limits, warning limits (LWL and UWL) are constructed at ±2σ as well as limits at ±σ. More complicated rules specifying event having low probability when the process is under control can be constructed. Some of these rules can be selected in the Shewhart control chart dialog panel, see Figure 30, by default all available rules are selecteed. They are:

1. One point exceeds LCL/UCL.
2. Nine points above/below the central line.
3. Six consecutive points show increasing/decreasing trend.
4. Difference of consecutive values alternates in sign for fourteen points.
5. Two out of three points exceed LWL or UWL limits.
6. Four out of five points are above/below the central line and exceed ±σ limit.
7. Fifteen points are within ±σ limits.
8. Eight consecutive values are beyond ±σ limits.

Shewhart charts are used in two steps:
1. Chart construction
2. Chart application

The goal of the construction step is to specify the central line (CL) and control limits so that they describe the real process correctly. When these values are not given in advance, they are set to mean and interval which contains 99.3% of available training data. For normal data, the interval is constructed ±3σ interval around arithmetic average, with the statistical characteristics based on observations of the process under control, excluding outliers or otherwise suspect data points. The chart is then applied to control the process, using a specified set of rules.

The QC.Expert offers seven common types of Shewhart control charts:

X-bar and S, X-bar and R, X-individual for continuous variables, np, p, u, c for discrete quality attributes. Continuous data can be transformed, see 6.11 to improve normality. Control limits for backtransformed data can be asymmetrical.

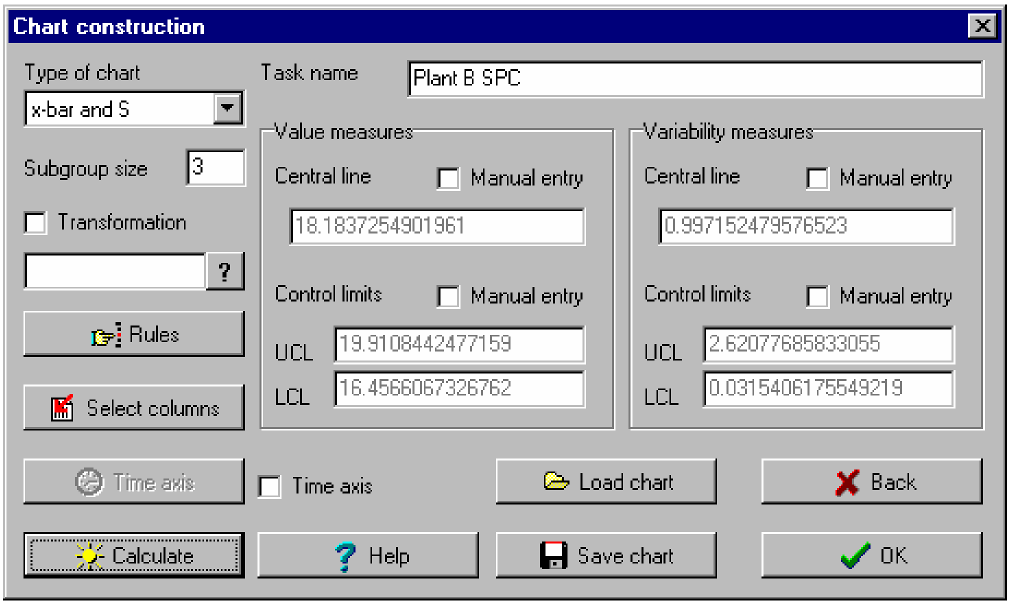


Chart construction

Type of chart: x-bar and S

Task name: Plant B SPC

Subgroup size: 3

□ Transformation

Value measures

Central line □ Manual entry

18.1837254901961

Control limits □ Manual entry

UCL 19.9108442477159

LCL 16.4566067326762

Variability measures

Central line □ Manual entry

0.997152479576523

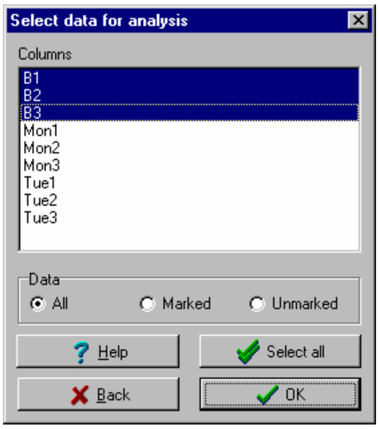
Control limits □ Manual entry

UCL 2.62077685833055

LCL 0.0315406175549219

□ Time axis

Fig. 110 Shewhart control chart construction dialog panel

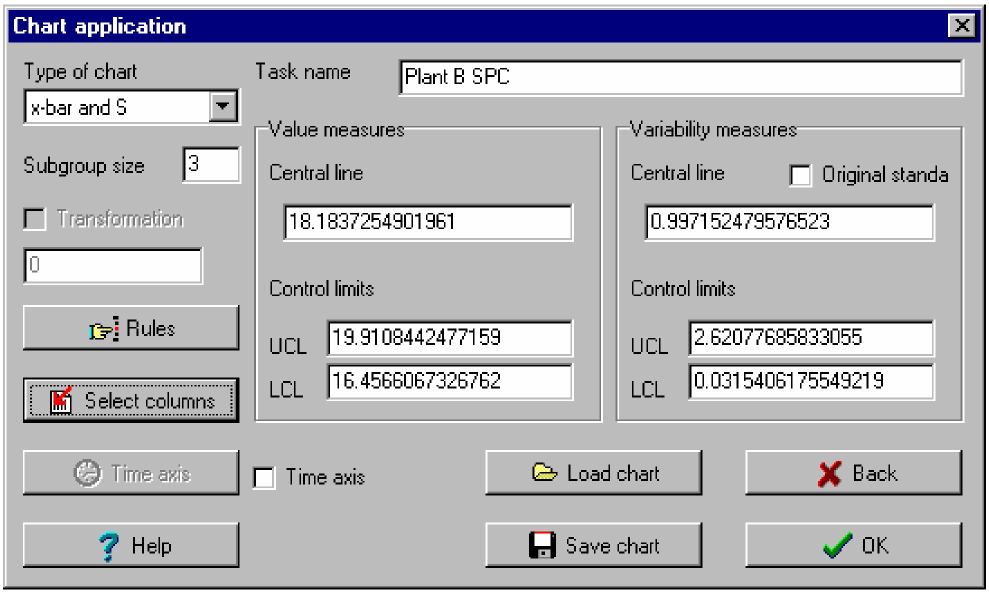


Columns

- B1 (selected)
- B2 (selected)
- B3 (selected)
- Mon1
- Mon2
- Mon3
- Tue1
- Tue2
- Tue3

Data: ● All ○ Marked ○ Unmarked

Fig. 111 Rows in selected columns represent subgroups for chart construction



Type of chart: x-bar and S

Task name: Plant B SPC

Subgroup size: 3

□ Transformation

0

Value measures

Central line: 18.1837254901961

Control limits

UCL: 19.9108442477159

LCL: 16.4566067326762

Variability measures

Central line □ Original standa

0.997152479576523

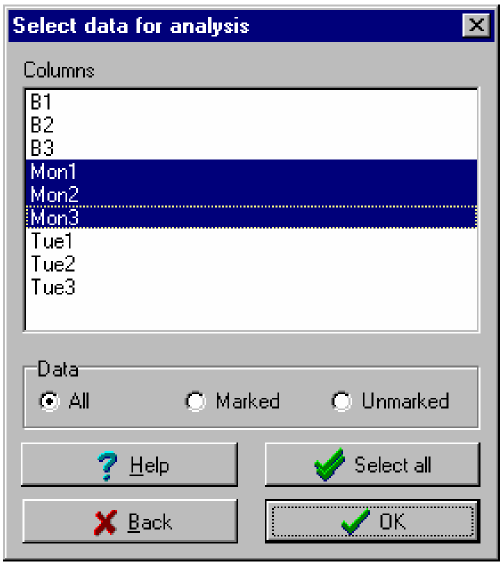
Control limits

UCL: 2.62077685833055

LCL: 0.0315406175549219

□ Time axis

Fig. 112 Shewhart control chart application dialog panel

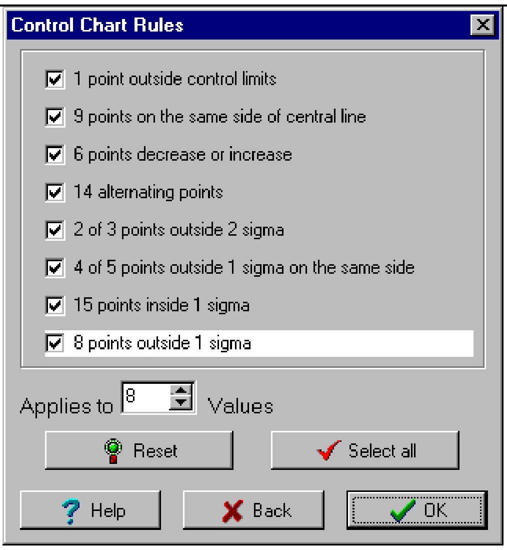


Columns

- B1
- B2
- B3
- Mon1 (selected)
- Mon2 (selected)
- Mon3 (selected)
- Tue1
- Tue2
- Tue3

Data: ● All ○ Marked ○ Unmarked

Fig. 113 Columns selected for application of the previously constructed chart



Control Chart Rules

☑ 1 point outside control limits

☑ 9 points on the same side of central line

☑ 6 points decrease or increase

☑ 14 alternating points

☑ 2 of 3 points outside 2 sigma

☑ 4 of 5 points outside 1 sigma on the same side

☑ 15 points inside 1 sigma

☑ 8 points outside 1 sigma

Applies to 8 Values

Fig. 114 Rules selection panel

20.1. Statistical assumptions

Simple Shewhart control charts X-bar and X-individual should not be used in the following cases:

1. Normality of chart data is rejected, see the Test for normality, 5.1.2.
2. Data are dependent or show a linear trend, see Autocorrelation and Test for linear trend, 5.1.2.
3. Data are heteroscedastic, i.e. their variance is not constant.
4. Several controlled variables are correlated, see Correlation, 5.3.2.

Possible remedies in such situations are:

1. When the departure from normality is caused by an outlier, the outlier should be excluded (only at the chart construction stage). When the departure is caused by systematic skewness or kurtosis of the data distibution, a subgroup size increase might help. Skewness problems can be removed by transformation. High kurtosis (see 5.1.2.) problems cannot be simply solved by the transformation technique implemented in the program, so that different quantile construction than the simple 3sigma approach should be used.
2. Detrended data (see data smoothing notes in 5.1.2) should be used. The EWMA dynamical chart (see 7.2) can be even better.
3. The EWMA dynamical chart should be used, see 7.2.
4. The Hotelling chart for multivariate data should be used, see 7.3.

20.2. Capability analysis, capability indices

Process capability indexes (PCI) are used to assess how successful the process control is. The simplest interpretation of a capability index is the following: when it is smaller than 1, process is not capable (does not satisfy given requirements), when the index is larger than 1, process is capable (satisfies given requirements). A more detailed process classification is sometimes used sometimes: capable (PCI<1), not capable (PCI ≥ 1) and highly capable (PCI ≥ 1.3). It is more appropriate to use not only the capability index estimates, but to accompany them by their confidence intervals, see *Capability indexes*, 6.4.2. A more strict classification then defines capable process as a process with **lower confidence limit for the capability index** larger than 1. The basic C_p index should be used only when the central line is given by the data mean. Confidence interval can be made narrower when more data points are used.

20.3. Transformations in control charts

The transformation technique can be helpful when dealing with asymmetrical data distribution. Skewed distributions can occur quite frequently, e.g. for low (trace) pollutant concentrations, product purity level, relative values close to 100%, physical variables like strength, time measurements, volume or surface related variables like mass, size of small particles etc. When the problem is not properly recognized and skewness is not taken into account when constructing a control chart, the chart information value might be drastically reduced. Neglecting skewness during chart construction causes that even if the process is under control, one control limit is frequently exceeded, while the other limit is not reached at all. When the subgroup size is large, effective skewness suppression can be expected, considering the central limit theorem. When significant asymmetry is found in the Transformation module, see 5.6 (it is often accompanied by normality rejection in the Basic data analysis module, see 5.1.2.), classical Shewhart chart should not be used. Transformation produces asymmetrical control limits for X. Variability chart is based on transformed data, so that the scale of S or R charts differs from the scale of the original data. Capability indexes C_p, C_{pk}, C_{pm}, C_{pmk} are computed from transformed data, which satisfy normality assumption. **Warning:** transformation parameter value has to be the same for both chart construction and chart application!

20.4. X-bar and S, X-bar and R charts

Menu:	QC.Expert	Control charts	Construction
			Application

20.4.1. Data and parameters

Data rows correspond to subgroups, see Table 6-1. Each subgroup has to have at least two values. Missing values are allowed.

Table10 X-bar chart data, number of subgroups = 10, subgroup size = 3

Time	NO$_x$1	NO$_x$2	NO$_x$3
1.11.96	110.9	101.6	114.6
2.11.96	113.3	120.8	116.3
3.11.96		106.1	107.6
4.11.96	236.2	230.8	238.5
5.11.96	110.3	105.6	104.9
6.11.96	115.6	113.6	122.2
7.11.96	128.3	138.6	127
8.11.96	139.9	133.9	
9.11.96	124.3	127.8	113.9
10.11.96	118.6	118	116.5

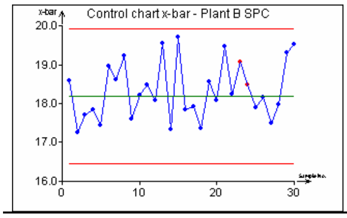
Chart construction. Parameters are inputted in the Shewhart control chart construction dialog panel, see Figure 30. *X-bar and S* or *X-bar and R* types can be selected in the *Chart type* window. Upon clicking the *Select columns* button, data columns are specified. At least two columns have to be selected. Data should come from a process under statistical control. Occasional outliers or otherwise suspect data as well as data violating specified rules should be omitted. The paragraph 6.1. describes actions to take when some other assumptions are violated. When a transformation is necessary, the transformation parameter can be specified. When the value is not known in advance, it can be computed upon clicking the question mark, ? button. When some of the chart parameters are known in advance, they can be inputted manually upon checking *Manual entry*. The *Rules* button invokes the Rules dialog panel, where some of the rules can be excluded/included by the >>, << buttons. Any change of the rules should be properly justified. Default rules are set by clicking the *Initialize* button. Computed chart parameters can be saved to a file by clicking the *Save parameters* button.

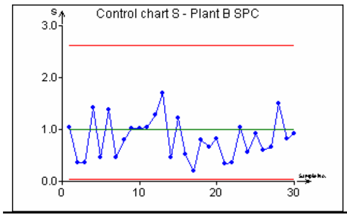
Chart application. As soon as the chart parameters are determined, the chart can be used to control data from the same process. Parameters for control chart application can be set in the Shewhart control chart application dialog panel, see Figure 31. The panel is similar to the construction panel, the UCL, LCL limits for x-bar and variability chart (S or R) are supposed to be known at this stage. Appropriate data columns are selected in the panel. Chart parameters are either entered manually when the *Manual entry* selection is checked, or they are read from a file upon clicking the *Read parameters* button. The file has to contain parameters for the same chart type and the same subgroup size. Rules can be modified by clicking the Rules button. The *Chart* button produces control chart and protocol output. Chart parameters can be saved by clicking the *Save parameters* button.

20.4.2. Protocol

No transformation/Transformation	Indicates whether any transformation was used to bring data closer to normality. When transformation was applied, control limits can be asymmetrical, depending on data distribution. S chart is based on the transformed data, so that the scale differs from the original data scale.
Chart type	X-bar.
Maximum subgroup size	Maximum subgroup size, i.e. number of columns, selected upon clicking the Select columns button.
Row number	Number of subgroups, i.e. points plotted in the chart.
Central line	Central line of the chart.
UCL	Upper control limit.
LCL	Lower control limit.
Variability	Specifies how the process variability is expressed.
Baseline	Baseline for standard deviation control chart.
Capability indexes	(see 6.10.)
Cp	Capability index $C_p = (UCL - LCL) / (6s)$ and its 95% confidence interval.
Cpk	Capability index C_{pk} = min(UCL–average ; average–LCL) / (3s) and its 95% confidence interval.
CI (Heavlin)	Heavlin 95% confidence interval.
CI (Kushler)	Kushler 95% confidence interval.
CI (Franklin)	Franklin 95% confidence interval.
Cpm	Capability index $C_{pm} = (UCL - LCL) / (6\sqrt{(s^2 + (\text{average} - ZL)^2)}$ and its 95% confidence interval.
Cpmk	Combined capability index C_{pmk}.
Rules violations	List of points, violating rules specified previously in the Rules panel. Points violating rules (corresponding to averages or standard deviations) are printed red bold, violated rule is specified.
Data	Printout of the input data.

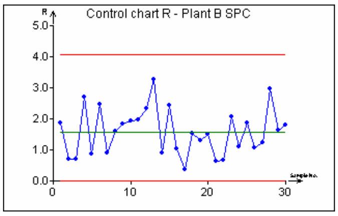
20.4.3. Graphical output

X-bar chart 	Standard X-bar control chart. Points violating specified rules are marked red. A variability chart (S or R) should be inspected before looking at the X-bar chart. The central line corresponds to the process baseline, the upper and lower lines correspond to the control limits.
S chart	Standard deviation control chart. Points exceeding limits for standard deviation are marked red. Variability charts (S or R) should be inspected before inspecting X-bar or X-individual charts. The green line corresponds



to the baseline, the upper and lower lines correspond to the control limits.

R chart



Range control chart. Range is calculated as the maximum-minimum difference within a subgroup. When the subgroup size is larger than 2, S chart is generally preferred. Points exceeding limits for range are marked red. Variability charts (S or R) should be inspected before inspecting X-bar or X-individual charts. The green line corresponds to the baseline, the upper and lower lines correspond to the control limits.

20.5. *X-individual and R charts*

Menu:	QC.Expert	Control charts	Construction
			Application

X-individual chart is used when no rational subgroups can be created. Individual data points are plotted in the chart, hence the chart is much more sensitive to departures from normality than the X-bar charts. X-bar charts (see 6.4.) should be used whenever possible. Absolute differences of consecutive values are plotted in the R chart.

20.5.1. Data and parameters

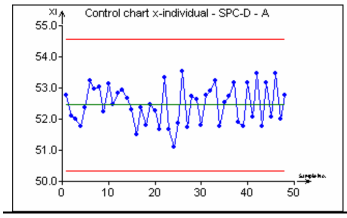
Data are expected in one column. No missing values are allowed. X-individual and R chart type is selected in Shewhart control chart application. Further steps are similar to those described in 6.4.1.

20.5.2. Protocol

X-individual chart protocol is analogous to the X-bar chart protocol, described in 6.4.2.

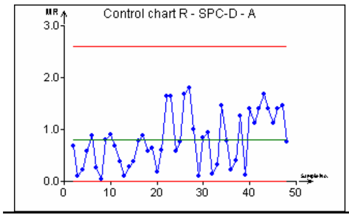
20.5.3. Graphical output

X-individual chart



Standard X-individual control chart. Points violating specified rules are marked red. R chart should be inspected before looking at the X-individual chart. The central line corresponds to the process baseline, the upper and lower lines correspond to the control limits.

R chart



Moving range chart. Since each points corresponds to a difference of two consecutive X-individual values, the chart starts from the second time point. The points exceeding limits for standard deviation are marked red R chart should be inspected before inspecting the X-individual charts. The green line corresponds to the baseline, the upper and lower lines correspond to the control limits.

20.6. *np-chart*

Menu:	QC.Expert	Control charts	Construction

np chart is useful when controlling number of faulty items within a batch. A binomial distribution is assumed for the faulty items number. It is a chart for integer valued variables, where batch corresponds to subgroup. Number of faulty items out of np items contained in a batch is plotted in the chart. Batch size might vary. For discrete quality attributes (like the number of faulty items here) no special variability chart is constructed.

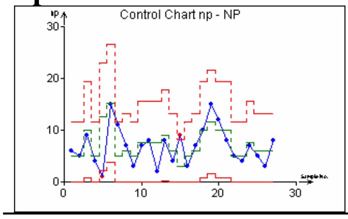
20.6.1. Data and parameters

Data are organized in two columns. The number of faulty items is in the first column, batch size (i.e. the total number of items in the batch) has to be in the second column. In the *Shewhart control chart construction* dialog panel, the np chart type is selected. Upon clicking the *Select columns* button, data columns are selected. Neither transformation nor additional rules are used for attributes charts (np, p, c, u). When baseline or limit values are known from previous analysis or given by a standard, they can be specified manually in the *Value* window within the Manual entry part of the panel. When the values are not known, they can be computed by clicking the *Calculate* button (*Manual entry* must not be checked). Baseline, LCL, UCL values can be saved into a file by the *Save parameters* button. Pressing the OK button plots the chart.

20.6.2. Protocol

NP	Number of faulty items by batches.
N0	Batch sizes.
BL	Baseline for a given batch size.
LCL	Lower control limit for a given batch size.
UCL	Upper control limit for a given batch size.
P	Average proportion of faulty items.
Out of limits	List of values out of control limits.

20.6.3. Graphical output

np chart 	Control chart for the number of faulty items. Points falling outside the control limits are marked red. The baseline is blue, the control limits red, their values increase with increasing batch size, hence they are not constant when the batch size changes.

20.7. p chart

Menu:	QC.Expert	Control chart	Construction

The p chart is useful when controlling proportion of faulty items in a batch. Because the number of faulty items follows a binomial distribution, it is a chart for attributes. Batch corresponds to subgroup. The proportion of faulty items within a batch is what is plotted in the chart. Therefore, the chart value is always between zero and one. The batch size might vary. No separate chart for variability is constructed.

20.7.1. Data and parameters

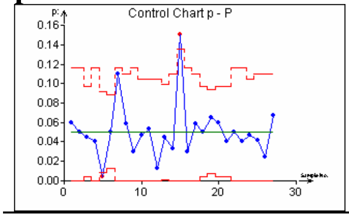
Data are organized in two columns. Proportion of faulty items is in the first column, batch size (i.e. the total number of items in the batch) has to be in the second column. In the *Shewhart control chart construction* dialog panel, the p chart type is selected. Upon clicking the *Select columns* button, data columns are selected. Neither transformation nor additional rules are used for attributes charts (np, p, c, u). When baseline or limit values are known from a previous analysis or given by a standard, they can be specified manually in the *Value* window within the Manual entry part of the panel. When the values are not known, they can be computed by clicking the *Calculate* button (*Manual entry* must not be checked). Baseline, LCL, UCL values can be saved into a file by *Save parameters*. Pressing the OK button plots the chart.

Protocol

P	Proportion of faulty items by batches.

N0	Batch sizes.
LCL	Lower control limit for a given batch size.
UCL	Upper control limit for a given batch size.
BL	Baseline.
Out of limits	List of points falling out of control limits.

20.7.2. Graphical output

p chart  	Control chart for proportion of faulty items. Points falling outside the control limits are marked red. The baseline is blue, the control limits are red, all of them depend on the batch size. When the batch size changes, their values are not constant but increase with the sample size.

20.8. c chart

Menu:	QC.Expert	Control charts	Construction

c chart is a less frequently used chart type. It is a chart for controlling the number of faults found in a defined amount of product. Such number can have only a nonnegative integer value and is assumed to follow a Poisson distribution. Because the controlled characteristic is discrete, the chart is a control chart for attributes. Unlike for np or p charts, the number of faults of the same type is recorded on each piece of product (or a small group of related pieces), where each individual fault might or might not cause whole piece to fail (e.g. number of knots counted within 10 meters of a fabric). The number of faults *c* is plotted in the chart and the size of a controlled piece of product has to stay constant. No special chart for variability is constructed.

20.8.1. Data and parameters

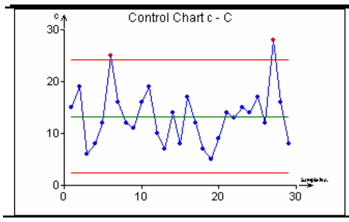
Data, i.e. the fault counts, counted on individual product pieces are in one column. The c chart is selected in the *Shewhart control chart construction* dialog panel. Data column is specified upon pressing the *Select columns* button. Neither transformation nor special rules are used for any of the attributes control charts (np, p, c, u). When the baseline or the control limits are known from previous data or given by a standard, they can be inputted upon checking Manual entry. When their values are not known, Manual entry must not be checked and their computation is requested by the Calculate button. The Save parameters button saves the baseline and the control limits to a file. Pressing the OK button produces the chart.

20.8.2. Protocol

Data	Number of faults counted on individual product pieces.
BL	Baseline.
LCL	Lower control limit.
UCL	Upper control limit.
Out of limits	List of values falling out of control limits.

20.8.3. Graphical output

c-chart	Control chart for controlling number of faults. Data exceeding control limits are marked red. The baseline is green. The control limits are red and constant.

20.9. *u chart*

Menu:	QC.Expert	Control charts	Construction

This chart type is utilized less frequently. It is used to control number of faults on a piece of product of variable size. Such number can have only nonnegative integer values and is assumed to follow a Poisson distribution. Because the controlled characteristic is discrete, the c chart is a control chart for attributes. Unlike for np or p charts, the number of faults of the same type is recorded on each piece of product (or a small group of related pieces), where each individual fault might or might not cause the whole piece to fail (e.g. number of knots counted within 10 meters of a fabric). The number of faults *u* is plotted in the chart. The size of a controlled piece of product is allowed to vary (this is the only difference from the c chart). No special chart for variability is constructed.

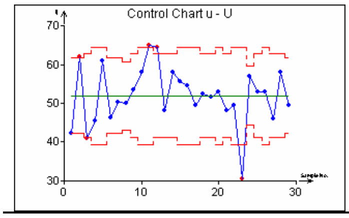
20.9.1. Data and parameters

Data, i.e. the fault numbers, counted for individual product pieces are in one column. Sizes of inspected product pieces have to be in the second column. The u chart is selected in the *Shewhart control chart construction* dialog panel. Data column is specified upon pressing the *Select columns* button. Neither transformation nor special rules are used for any of the attributes control charts (np, p, c, u). When the baseline or the control limits are known from previous data or given by a standard, they can be inputted upon checking Manual entry. When their values are not known, Manual entry must not be checked and their computation is requested by the Calculate button. The Save parameters button saves the baseline and the control limits to a file. Pressing the OK button produces the chart.

20.9.2. Protocol

U	Number of faults counted on inspected pieces of product.
N0	Sizes/amounts of product controlled.
U1	Relative number of faults, standardized to one unit of product, U1=U/N0. This is what is plotted in the chart.
LCL	Lower control limit.
UCL	Upper control limit.
BL	Baseline.
Out of control limits	List of points falling outside control limits.

20.9.3. Graphical output

u chart 	Control chart for fault numbers. Points falling out of control limits are marked red. The baseline is green, the control limits are red. The control limits depend on the amount /size of the product controlled. When the amount/size varies, the limits are not constant.

21. Other control charts

The control chart types listed below are recommended as alternative and additional tools to the Shewhart control charts. When compared with classical charts, they have some advantages and some disadvantages. More difficult construction and interpretation are on the minus side, while much larger sensitivity (up to an order of magnitude) to the shifts of the mean value (CUSUM), applicability even for data showing long cycles, trend or nonconstant variance (EWMA – dynamic modification) are on the plus side. When several correlated variables are controlled simultaneously, the Hotelling control chart is appropriate.

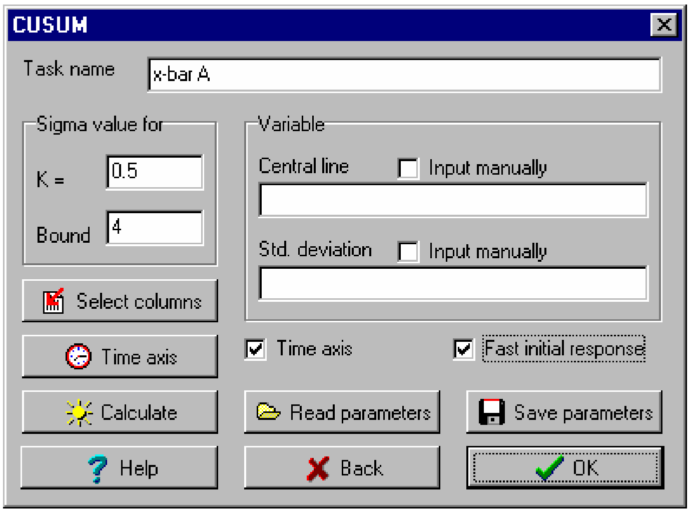
21.1. CUSUM

Menu:	QC.Expert	Other Control Charts	CUSUM

CUSUM control chart and its efficient modification by Lucas (V-mask is not needed) is recommended mainly when a process mean shift needs to be detected fast. The method is based on summing differences from the goal cumulatively (Cumulative SUMs). Imbalance of differences from the goal (more differences of one sign) is quickly detected. Sensitivity of this chart is given by the *k* parameter, which can be inputted in the CUSUM chart dialog panel, see Figure 33. When the data are normal and independent, the technique is much more efficient than the Shewhart X-bar or X-individual chart – number of data points necessary to detect 1 standard deviation departure from the baseline is about 10 times smaller. When a point falls outside the limits, it is taken as an alarm signal. When remedy steps are taken immediately after obtaining the signal, the FIR technique can be applied. The technique is useful as a quick check of whether the corrective action was successful. It is based on moving the point immediately following the remedy action just below/above the limit on the same side where the previous violation occurred. If the action was not successful, the limit is violated again.

21.1.1. Data and parameters

The CUSUM chart parameters are inputted in the CUSUM chart dialog panel, see Figure 33. *K* coefficient is entered in the *Sigma value* section. It specifies how large deviation in units of standard deviation from a specified baseline (goal or target) should be detected. Usual values are 1 or 0.5. When the value is too small, false alarms are frequent. On the other hand, when the coefficient value is too large, efficiency of the chart decreases. Limits are specified in the *Limit* field. 5 is recommended for a normal regimen and 4 for a strengthened regimen. Data columns are selected upon clicking the *Select columns* button. When several columns are selected, rows should correspond to different subgroups, since the row means are computed and used for further analysis. When the baseline and standard deviation are known, they can be entered manually upon checking *Manual entry*. FIR is enabled when checking *FIR*, see the previous paragraph.



CUSUM

Task name: x-bar A

Sigma value for

K = 0.5

Bound 4

Variable

Central line □ Input manually

Std. deviation □ Input manually

☑ Time axis

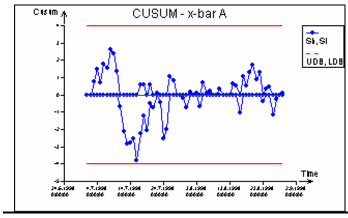
☑ Fast initial response

Fig. 115 CUSUM chart dialog panel

21.1.2. Protocol

Number of data points	Number of data rows used for chart construction.
Baseline	Target value, given or computed from data.
Standard deviation	Standard deviation value, given or computed from data.
Detectable shift	Minimum shift from the target to be detected.
FIR	Yes or No indicating whether the technique was used.
Limits +-	Inputted limits.
Violating (+)	Number of data points violating upper limit.
Violating (-)	Number of data points violating lower limit

21.1.3. Graphical output

CUSUM chart 	CUSUM chart for detecting systematic changes from the target (central line). Violation of the limits (red lines) is taken as an alarm signal.

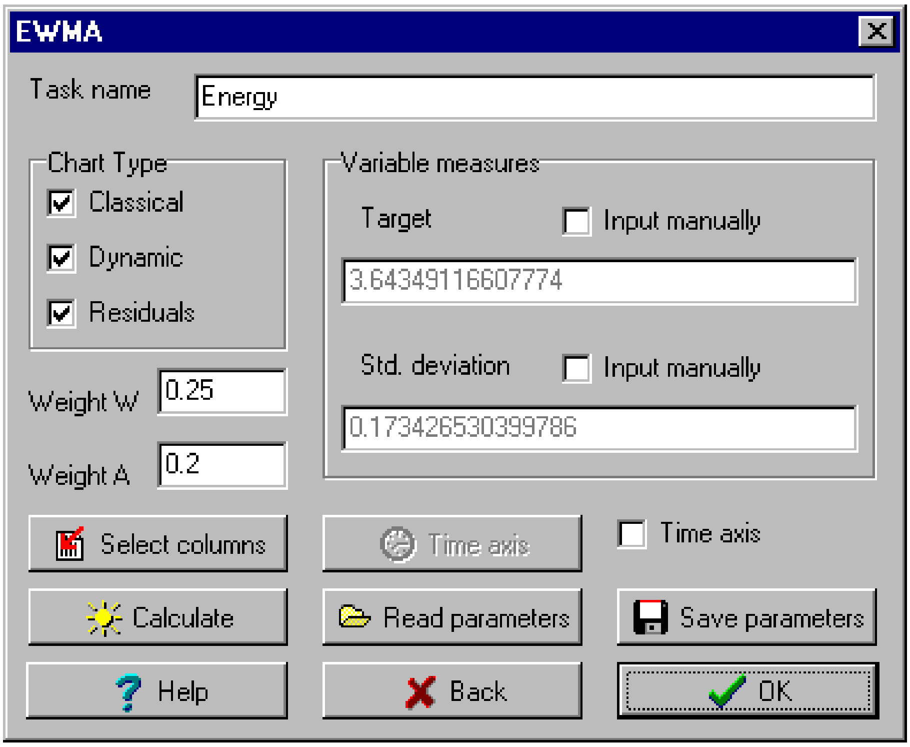
21.2. EWMA

Menu:	QC.Expert	Other Control Charts	EWMA

The Exponentially Weighted Moving Averages (EWMA) chart is based on the values $X_i = W.x_i + (1 - W).X_{i-1}$. W ($0 < W < 1$) is a weight specifying how fast is the EWMA response to change in the process mean. When W=1, the chart is equivalent to the X-bar or X-individual chart. Higher W values produce smoother chart which is less sensitive to sudden changes. This module can also produce a dynamical EWMA modification, useful when the controlled process shows long term oscillations or trend so that other chart types cannot be used. Such a situation can be indicated e.g. by significant autocorrelation or time trend , found in the Basic data analysis module. The dynamical EWMA chart tolerates long term process mean or variance changes, but identifies short term, sudden changes. The dynamical EWMA chart is controlled by the A parameter inputted in the dialog panel, see Fig. 119. The EWMA residual chart is used for the same purpose as the dynamical EWMA chart.

21.2.1. Data and parameters

Data are organized in one or more columns. When more than one column is selected, row means are computed and used in subsequent analyses, so that rows should correspond to subgroups. Baseline (target mean value) and standard deviation are entered in the EWMA control chart dialog panel., checking *Manual entry*, see Figure 34. Unknown values can be estimated from data. At least one chart type must be selected and W, A weights entered. When Manual entry is not checked, the program computes baseline and limits from data. The Save parameters button saves parameter values to a file. The Read parameters button reads parameters from a specified file. The OK button runs computations, producing protocol and graphical output.



EWMA

Task name: Energy

Chart Type

☑ Classical

☑ Dynamic

☑ Residuals

Weight W: 0.25

Weight A: 0.2

Variable measures

Target □ Input manually

3.64349116607774

Std. deviation □ Input manually

0.173426530399786

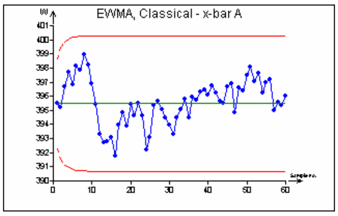
□ Time axis

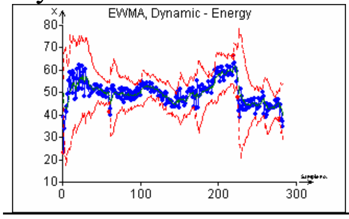
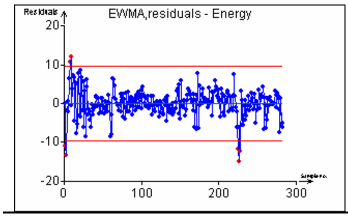
Fig. 116 EWMA chart dialog panel

21.2.2. Protocol

Number of data points	Number of data rows
Baseline	Baseline (target for mean) inputted in the dialog panel (see Figure 34)or computed.
Standard deviation	Standard deviation) inputted in the dialog panel (see Figure 34.
Weight W for EWMA	Weight for the EWMA construction.
Weight A for dynamic EWMA	Weight for dynamic the EWMA construction.
Number of data points out of classical control limits	Number of data points violating the EWMA control limits.
Number of data points out of dynamical control limits	Number of data points violating dynamical the EWMA control limits.
Number of residuals out of ±3σ limits	Number of data points violating residual the EWMA control limits.
Mean square	Mean squared difference from the baseline.

21.2.3. Graphical output

EWMA chart 	EWMA chart with nonconstant limits. Violation of the red limits is taken as an alarm signal.

Dynamic EWMA chart 	Dynamic EWMA chart tolerates slow changes in the mean. Violation of the control limits (red) is taken as an alarm signal.
Residual EWMA chart 	Residual EWMA chart has similar use as the dynamic EWMA chart. Violation of the control (red) limits is taken as an alarm signal

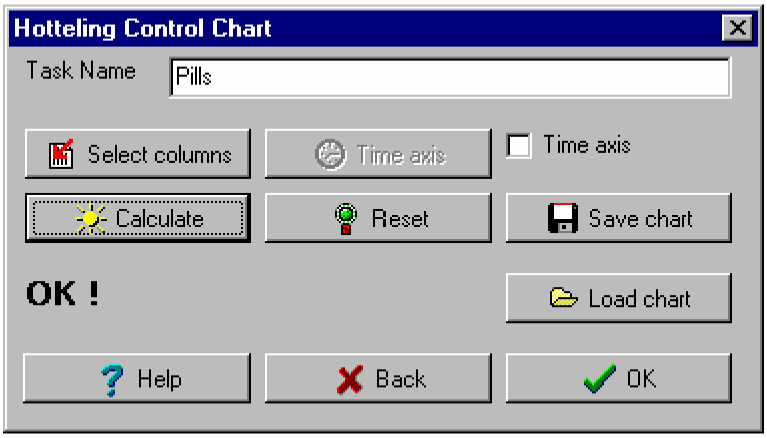
21.3. Hotelling T2 chart

Menu:	QC.Expert	Other Control Charts	Hotelling chart

Hotelling control chart is used when several correlated variables are controlled simultaneously. Within this module, the data are assumed to consist of individual (multivariate) data points, hence the chart is a multivariate analogue of the X-individual chart. Such data might be used in X-individual control charts for separate variables only if the variables are not correlated – their correlation coefficients should not be significant, see the Correlation module description, 5.3. Hotelling chart reduces the information about individual variables to one characteristic, namely distance from the (multivariate) mean computed with respect to the covariance matrix (Mahalanobis distance). The distance cannot be negative and respects interrelationships among different variables. The lower limit is always set to zero.

21.3.1. Data and parameters

Data are organized in columns, corresponding to different variables. Rows with missing data are ignored completely.



Hotteling Control Chart

Task Name Pills

□ Time axis

OK !

Fig. 117 Hotelling chart dialog panel

Setting chart parameters

Parameters are estimated from a training dataset, obtained when the process was under control. Upon clicking the *Select columns* button, data columns are specified. The *Reset* button resets all parameter values set previously. The *Calculate* button computes mean vector and covariance matrix. Parameters can be saved to a file by *Save parameters* button. The OK button starts the chart computations. The baseline is given by the mean vector estimated from the training dataset. Allowable amount of variability is given by complete covariance matrix, estimated from the training dataset. The mean vector is outputted to the protocol. The covariance matrix can be inspected by viewer called from the *QC.Expert-Viewer-Hotelling control chart* menu.

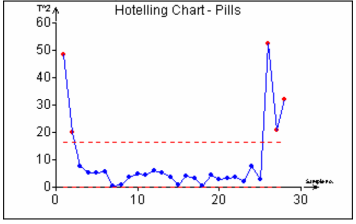
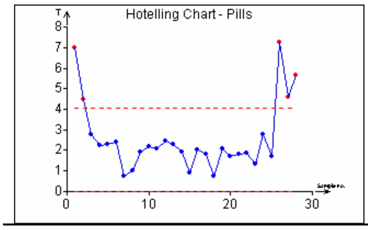
Control chart application

Chart parameters obtained previously and saved in a file are read by clicking the *Read parameters* button. The OK button runs computations.

21.3.2. Protocol

Column name	Data column names.
Target mean	Means computed from the training dataset.
Mean	Means for the process data for which the chart is applied.
Chart parameters	
LCL, UCL	Lower and upper control limits, lower limit is always set to zero.
Out of limit	List of data points exceeding the upper limit.
Time	Times at which UL was exceeded.
Number	Identification of data points exceeding UL.
Value	T^2 values for data points exceeding UL.

21.3.3. Graphical output

Hotelling chart   	Distances of individual (multivariate) data points from the target mean, expressed as Hotelling T^2 values. The other of the two charts shows T values, i.e. square roots of T^2. Both charts have exactly the same interpretation. This characteristic takes into account simultaneous behavior of all variables. The lower control limit is always set to zero, the upper control limit is given by a chi-square distribution quantile, depending on the number of variables used.

22. Capability analysis

Menu:	QCExpert	Capability

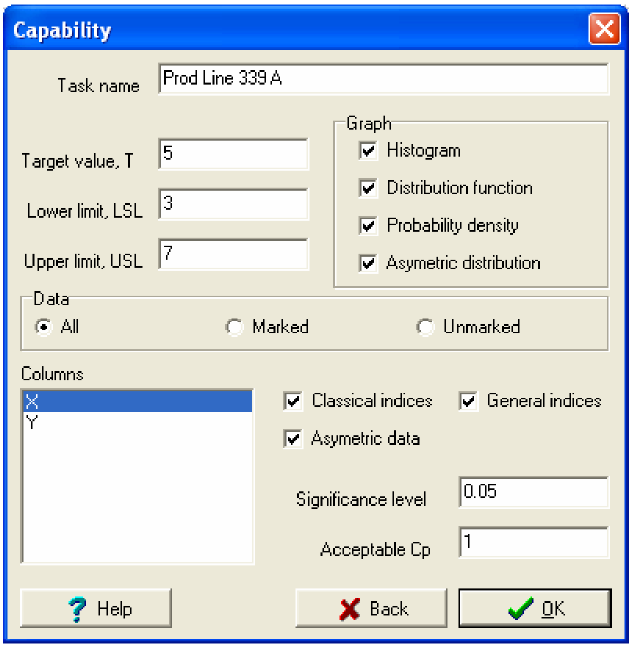
This module computes the capability index, c_p and the performance index, p_p, based on data and user-specified limits. Additional values, like ARL are given as well. The module allows for one-sided specifications as well as for asymmetric (non-normal) distribution.

22.1. Data and parameters

Measured values of the quality characteristics of interest are entered as data. The module expects that the data are in one column. Target value has to be specified in the Dialog panel. In addition, at least one of the specification limits has to be entered (LSL, *Lower Specification Limit* or USL, *Upper Specification Limit*). In the *Columns* field, data column is selected. One can specify, whether all data, marked data, or unmarked data will be used for computations in the *Data* field. In the *Plots* field, graphical output requests are to be specified. List of available plots appears in the paragraph 22.3 below.

When only one specification limit is defined for the process under control (no matter whether lower or upper), it is entered in the dialog panel, while the field for the other limit is left blank. A desired *Confidence level* can be specified as well. It is used subsequently for calculation of confidence interval, capability and performance indexes. *Cp limit* is the value, below which we consider the process as being not capable. In the Protocol output, all index values and their confidence interval limits smaller than *Cp* are marked in red. Usually, 1 is selected for *Cp*.

If the *Classical indexes* field is marked, classical capability and performance indexes: c_p, c_{pk}, c_{pm}, p_p, p_{pk}, p_{pm} are computed, together with additional characteristics. Definitions of these indexes are shown below. When only one specification limit is specified, classical indexes are not computed. Then, one has to mark the *General indexes* selection – and the c_{pk}^* index (based on probabilistic grounds) is computed. This generalized index can be used for one-sided specification limit or asymmetric data, violating the usual normality assumption. (Distributional normality test can be found in the *Elementary statistics* module. If the *Asymmetric data distribution* selection is checked, the software allows for a possibility that the data come form asymmetric (skewed) distribution. The c_{pk}^* calculation is then adjusted via preliminary application of the exponential transformation of the data. Quantile function F^{-1} value (needed during c_{pk}^* calculations) is computed after the transformation. Warning: if the *Asymmetric data distribution* selection is not checked, the software goes straight ahead and uses „forcefully“ normal model, even though the true data generating distribution is not normal. Hence, if one is not sure about distributional symmetry, it is a good strategy to leave the selection checked. Further detail about properties and motivation of the exponential data transformation can be found in the manual for the Transformation module. If the data are not normal, classical indexes commonly give unrealistically optimistic impressions. They often overestimate true values (although they can be underestimate as well). Hence, if the data are not approximately normal, the classical indexes should not be used.



Capability

Task name: Prod Line 339 A

Target value, T: 5

Lower limit, LSL: 3

Upper limit, USL: 7

Graph: ☑ Histogram ☑ Distribution function ☑ Probability density ☑ Asymetric distribution

Data: ● All ○ Marked ○ Unmarked

Columns: X (selected), Y

☑ Classical indices ☑ General indices ☑ Asymetric data

Significance level: 0.05

Acceptable Cp: 1

Fig. 118 Dialog panel for Capability

$$c_p = \frac{USL - LSL}{6\sigma_C}, \quad c_{pk} = \frac{\min(USL - \bar{x}, \bar{x} - LSL)}{3\sigma_C}, \quad c_{pm} = \frac{(USL - LSL)}{6\sqrt{\sigma_C{}^2 + (\bar{x} - T)^2}}$$

$$p_p = \frac{USL - LSL}{6\sigma_P}\ , \ p_{pk} = \frac{\min(USL - \bar{x}, \bar{x} - LSL)}{3\sigma_P}\ , \ p_{pm} = \frac{(USL - LSL)}{6\sqrt{\sigma_P{}^2 + (\bar{x} - T)^2}}$$

$$\sigma_C = \sqrt{\frac{1}{n-1}\sum_1^n [x_i - \bar{x}]^2}\ , \ \sigma_P = \frac{1}{d_2}\frac{\sum_2^n |x_i - x_{i-1}|}{n-1}, \quad kde\ d_2 = 1.128$$

$$p_{zm} = F_N\left(\frac{\bar{x} - LSL}{\sigma_C}\right) + 1 - F_N\left(\frac{USL - \bar{x}}{\sigma_C}\right)$$

$$ARL = 1/p_{zm}$$

$$c_{pk}^* = -\frac{1}{3}F^{-1}\{1/ARL\},$$

where F^1 is the inverse function to the distribution function (or the quantile function) for the normal distribution.

Remark: Because the estimate is generally not equal to the true value, it is more appropriate not to concentrate only on the point estimate, but to consider associated confidence interval as well. A particular, rater more conservative strategy is typically suggested: behave as if the true the lower confidence interval limit was equal to the true index value. Remember that, for instance if the index comes out as c_p=1.001 with associated confidence interval stretching from 0.8 to 1.2, that the process is very likely not capable (c_p<1)! If the index comes out as c_p=1.2, with the confidence interval ranging from 1.0 to 1.4, it is very unlikely that the process is not capable.

22.2. Protocol

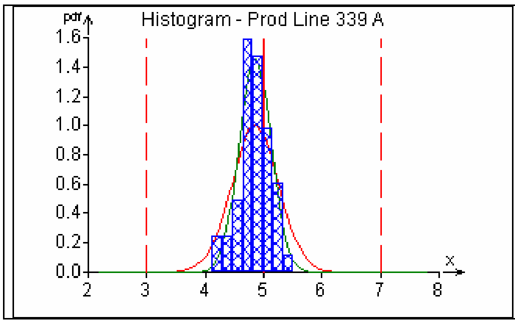
Capability and performance under the normality assumption	Normal distribution based calculations are performed only when both specification limits are given. When only one limit is given, report contains items from the „Cpk for asymmetrically distributed data“.
Project name	Name of the data spreadsheet
Target value	User-specified target parameter value.
Specification limits	
LSL	Lower specification limit (if specified by the user).
USL	Upper specification limit (if specified by the user).
CP limit	Lowest capability resp. performance index value that is acceptable. Values, smaller than the CP limit will be marked in red.
Capability indexes	
Arithmetic average	Arithmetic average computed from the data.
Standard deviation	Standard deviation, estimated from the data, σ_C
+/- 3sigma	Lower and upper limit of the ±3σ_C interval around the arithmetic mean
Z-score	Z-scores correspond to the lower and upper data part
Index	
Cp	Classical capability index value, c_p computed from σ_C

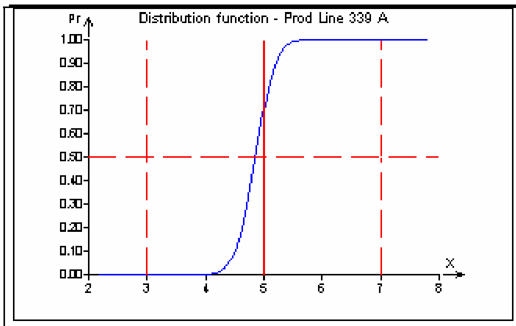
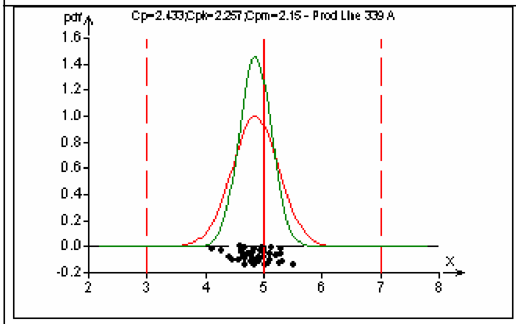
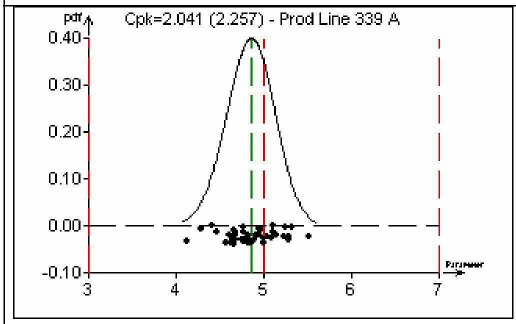
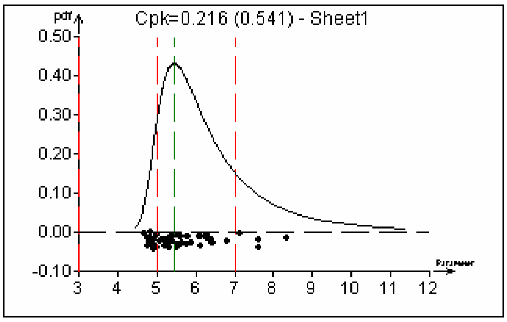
Cpk	Classical capability index value, c_{pk} computed from σ_C
Cpm	Classical capability index value, c_{pm} computed from σ_C
Lower limit	Lower limit of the confidence interval for a particular index.
Upper limit	Upper limit of the confidence interval for a particular index.
Performance limits	
Arithmetic average	Arithmetic mean computed from the data
Standard deviation	Standard deviation computed from the data, σ_C
+/- 3sigma	Lower and upper limit of $\pm 3\sigma_C$ interval around the arithmetic mean
Z-score	Z-scores correspond to the lower and upper data part
Index	
Pp	Classical performance index value, p_p computed from σ_P
Ppk	Classical performance index value, p_{pk} computed from σ_P
Ppm	Classical performance index value, p_{pm} computed from σ_P
Lower limit	Lower limit of the confidence interval for a particular index.
Upper limit	Upper limit of the confidence interval for a particular index.
Probability of exceedance	Probability that the upper, or lower specification limit, p_{zm} will be exceeded. This number can be understood as the probability that the next measurement will fall above upper or below lower specification limit.
Expected relative frequency of exceedance in %	It can be understood as the expected number of the measurements falling above the upper or below the lower specification limit in the next 100 measurements taken under the same circumstances.
Expected relative frequency of exceedance in PPM	It can be understood as the expected number of the measurements falling above the upper or below the lower specification limit in the next 1,000,000 measurements taken under the same circumstances.
Probability of being out of the SL	Probability that any of the specification limits will be exceeded. This number can be understood as the probability that the next measurement will fall beyond any of the specification limits.
Relative frequency of being out of the SL in %	It can be understood as the expected number of the measurements falling beyond any of the specification limits in the next 100 measurements taken under the same circumstances.
Relative frequency of being out of the SL in PPM	It can be understood as the expected number of the measurements falling beyond any of the specification limits in the next 1000000 measurements taken under the same circumstances.
ARL	*Average Run Length* is the expected number of measurements between two consecutive specification limit exceedances.
Cpk for asymmetrically distributed data	
Sample size	Number of the data points used for computations.
Corrected average	Expected value estimate, corrected for the data distribution skewness. When the data are symmetrically distributed, this characteristic is equal to the arithmetic average, see the *Transformation* module, paragraph 24.2.3, page 24-196.
Target value	User-defined target.
CP limit	Lowest acceptable c_p value. Values lower than this limit will be marked

	in red.
Specification limits	User-supplied specification limits.
Probability of exceedance	Probability that the upper, or lower specification limit, p_{zm} will be exceeded. This number can be understood as the probability that the next measurement will fall above upper or below lower specification limit.
Expected relative frequency of exceedance in %	It can be understood as the expected number of the measurements falling above the upper or below the lower specification limit in the next 100 measurements taken under the same circumstances.
Expected relative frequency of exceedance in PPM	It can be understood as the expected number of the measurements falling above the upper or below the lower specification limit in the next 1,000,000 measurements taken under the same circumstances.
Probability of being out of the SL	Probability that any of the specification limits will be exceeded. This number can be understood as the probability that the next measurement will fall beyond any of the specification limits.
Relative frequency of being out of the SL in %	It can be understood as the expected number of the measurements falling beyond any of the specification limits in the next 100 measurements taken under the same circumstances.
Relative frequency of being out of the SL in PPM	It can be understood as the expected number of the measurements falling beyond any of the specification limits in the next 1000000 measurements taken under the same circumstances.
ARL	*Average Run Length* is the expected number of measurements between two consecutive specification limit exceedances.
Cpk	Generalized capability index estimate, c_{pk}^*, defined for both symmetrically and asymmetrically distributed data, both two sided and one-sided specification limits. This characteristic should be used whenever the data are not symmetrically distributed.
Cpk limits	Lower and upper confidence interval limits for c_{pk}^*.

22.3. Graphical output

The capability module provides four plot types: three show density and one distribution function. The first three plots, that is: Histogram, Distribution function and a Density are plotted only when the *Classical indexes* selection is checked. The last plot: Density of the transformed data is plotted only when the *General index* selection is checked.

	A simple graphical tool, which can be used to compare data against the specification limits. Data are summarized by the histogram, kernel density estimate (red curve) and fitted normal density (Gauss' curve) plotted as a green curve. Vertical lines show the target, lower and upper specification limit. The location of the fitted Gauss' curve maximum corresponds to the arithmetic mean of the data. It should be as close to the target value as possible.

	Normal cumulative distribution function (integrated probability density function), based on the parameters estimated form the data (under the normality assumption). Vertical lines correspond to the target and specification limits. Horizontal line corresponds to the probability of 0.5. This line intersects the cumulative distribution curve at the point, whose x-coordinate corresponds to the arithmetic mean of the data. This plot can be used to read probabilities of the normal variable being lower than or equal to a value given by the x-coordinate. The reading can be done more precisely using the *Detail* function in the interactive mode of this plot after a double-click.
	Probability density curves. Red curve corresponds to the kernel estimate, Gauss' curve with parameters estimated from the data is plotted in green. If these two curves differ markedly, normality of the data is suspicious. Normality should be checked formally then, using an appropriate normality test. This test can be found in the *Elementary statistics*, module. Dashed lines correspond to the specification limits and to the target. Individual data points are plotted below the x-axis. For a better readability, the points are randomly scattered in the vertical direction (small amount of random jitter is added). Estimates of the classical indexes c_p, c_{pk} a c_{pm} are listed in the plot's header.
 	Probability density curve for the transformed data. The meaning of the plot is very similar to the previous one. The density is estimated via the exponential transformation. More details about the transformation can be found in the Transformation module. If the *Asymmetric data distribution* is not checked before the calculations, transformation is not used and normal density curve is plotted. Asymmetry of the data distribution can be checked graphically by inspection of the probability density curve on this plot. The c_{pk}^* index estimate is listed in the plot's header. When both specification limits are given, the classical c_{pk} index is listed in parenthesis as well. If the two values differ markedly, c_{pk}^* should be used. Illustrative examples on the left panel here show curve shapes for symmetric and asymmetric data distributions.

23. Pareto chart

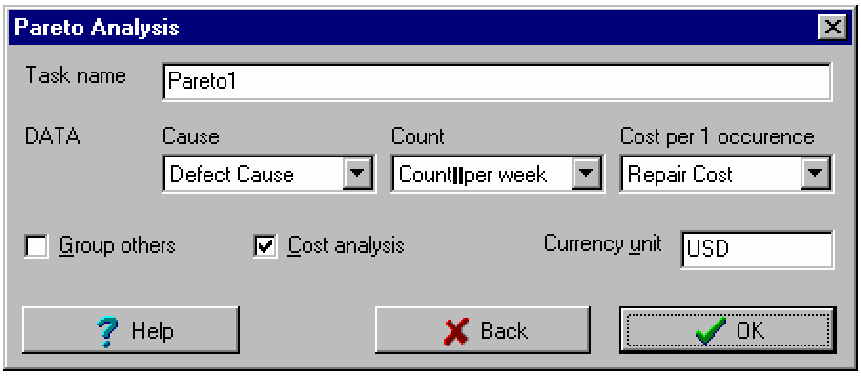
Menu:	QC.Expert	Pareto analysis

Pareto analysis is used to judge frequency and importance of various items, e.g. faults, errors etc. The Pareto chart is based on ordering various items by their amount. The Pareto 80/20 rule says that 80% of problems is caused by 20% of causes. The rule has been found approximately valid in many practical situations. When expenditures or financial losses are known, the Pareto analysis can be performed on them as well.

23.1. Data and parameters

At least two data columns are required. One of them contains names of items as character strings, the other contains their respective frequencies. When cost analysis is required, an additional column has to contain cost values (e.g. costs incurred by various type of damage as items). The cost analysis is performed when the *Cost analysis* option is checked in the Pareto dialog chart, and *Cost* column is specified. When the *Merge others* option is checked, items with small values are merged into the Other category while keeping the Other category smaller than any other individual category. This might be useful when there is many items. An example of the Pareto analysis data:

Defect Cause	Count per week	Repair Cost
Flange Packing	40	5
Corrosion A	3	60
Corrosion B	35	20
Turn-cock	14	130
Cover B	5	52
Nuts	62	13
Condenser	21	28
Ball bearing H	5	300
Ball bearing M	17	220
Hose	36	40


Pareto Analysis

Task name: Pareto1

DATA

Cause: Defect Cause

Count: Count‖per week

Cost per 1 occurence: Repair Cost

□ Group others

☑ Cost analysis

Currency unit: USD

Fig. 119 Pareto chart dialog panel

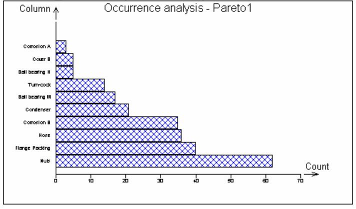
23.2. Protocol

Cost unit	Cost currency unit as a text inputted in the panel, see Figure 36.
Frequency table	Items sorted by their frequencies.
Item	Item identification (e.g. a damage type).
Number	Frequency of items.
Cost	Inputted costs for individual items. This *is applicable only when the Cost analysis option is checked*)
Item proportion	Relative frequency of items.
Cumulative item proportion	Cumulative item frequency.
Total cost	Total cost for each item.
Cost proportion	Relative cost for each item..
Cost table	This table contains the same columns as the frequency table, with cost information used in place of frequencies.

Merge	This table is created only if Merge was checked in the dialog panel, see Figure 36.
Item frequency	Frequency for each item.
Item proportion	Relative frequency for each item.
Cost analysis	Inputted costs for individual items. This *is applicable only when the Cost analysis option is checked*).
Relative cost analysis	Relative costs (in % of the total cost). This *is applicable only when the Cost analysis option is checked*).

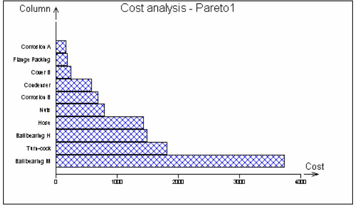
23.3. Graphical output

Frequency analysis



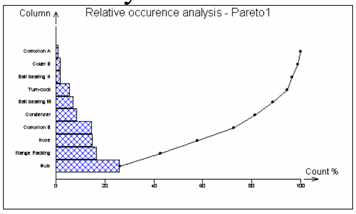
Items ordered by their frequencies.

Relative frequency analysis



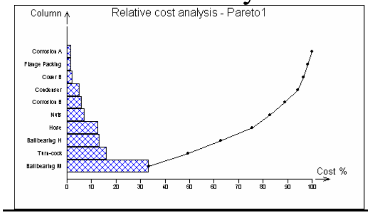
Items sorted by their relative frequencies (total number of all items is 100%). Ordering is the same as in the previous chart. The cumulative frequency curve is also plotted.

Cost analysis



Items sorted by their costs. The total item cost is the product of the item frequency and the cost of one item unit.

Relative cost analysis



Items sorted by their relative costs, ordering is the same as in the previous chart. The cumulative frequency curve is also plotted.

24. Acceptance Sampling

24.1. Acceptance sampling for attributes

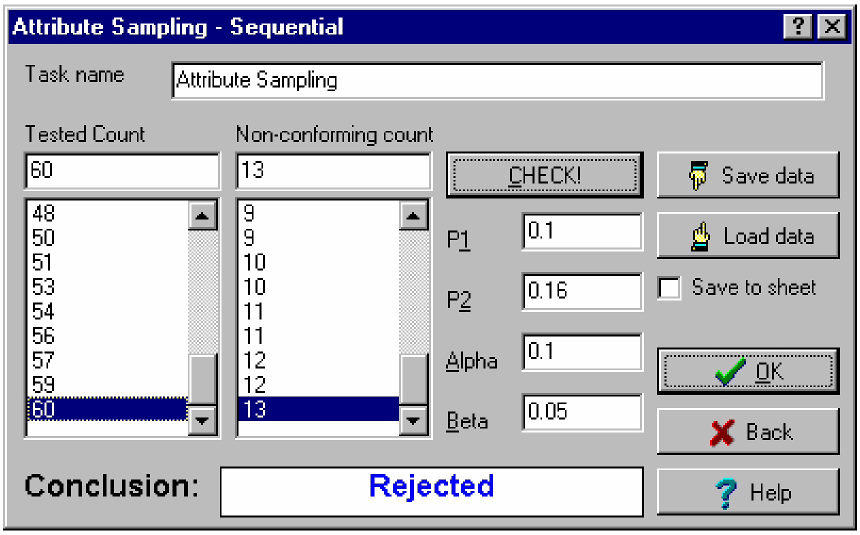
Menu:	QC.Expert	Acceptance sampling	Attributes

This module is designed to help with the decision whether to accept or reject a lot consisting of individual items, which can be classified as good or faulty (conforming or nonconforming). The decision is based on sequential random sampling. Each item drawn is inspected and information about its status is inputted into the program. After sufficient amount of information has been collected, the program suggests whether the lot should be accepted or rejected. The procedure is called *Sequential acceptance sampling* and usually requires to test only a relatively small number of items before a decision is made. Apart from appropriate parameter choice, the randomness of the sampling is the crucial part of the procedure. When numbers can assigned to items (e.g. order within the lot), tables of random numbers can be used. Alternatively, the random number generator from the *Simulation* module can be used instead, see 5.8.

24.1.1. Data and parameters

The acceptance sampling plan parameters are entered in *the Acceptance sampling for attributes* dialog panel. They are: producer risk (*Alpha*), consumer risk (*Beta*), acceptable proportion of faulty items corresponding to *Alpha* (P1) and unacceptable proportion of faulty items corresponding to *Beta* (P2). The results of individual item checks are inputted in the *Number of items tested* and *Number of faulty items* fields, immediately after each check is finished. Both *Number of items tested* and *Number of faulty items* are cumulative numbers since the beginning of the sampling procedure. The *Number of items tested* has to increase, the *Number of faulty items* must not decrease. The acceptance sampling procedure is finished when message "ACCEPTED" or "REJECTED" appears in the *Conclusion* window instead of the "UNDECIDED" message. A plot can be requested at any stage by pressing the OK button. The plot is a part of the acceptance sampling protocol. The individual test results can be placed to the current sheet by pressing the *Save* button. The table is then attached to the acceptance sampling protocol. When the individual test results are available in the current data sheet, they can be read by the *Read* button. The columns containing the number of items tested and the number of faulty items have to be selected. The first number must increase, the second number must not decrease.

Producer risk *Alpha* is defined as the probability (risk) of rejecting a good lot, i.e. a lot with acceptably small proportion of faulty items (smaller than P1). Consumer risk is defined as the probability (risk) of accepting a bad lot, i.e. a lot which has unacceptably large proportion of faulty items (larger than P2). When *Save to sheet?* selection is checked, the number of tests and the number of faulty items are saved to the current sheet whenever OK or *Save* button is pressed.



Task name: Attribute Sampling

Tested Count	Non-conforming count
60	13
48	9
50	9
51	10
53	10
54	11
56	11
57	12
59	12
60	13

P1: 0.1
P2: 0.16
Alpha: 0.1
Beta: 0.05
☐ Save to sheet

Conclusion: Rejected

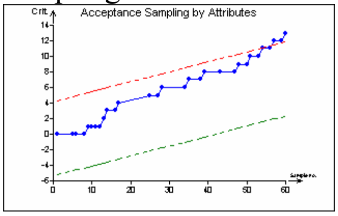
Fig. 120 Acceptance sampling for attributes dialog panel

24.1.2. Protocol

	Values entered in the panel shown in Fig. 20.
P1	Acceptable proportion of faulty items.
P2	Unacceptable proportion of faulty items.
Alfa	Producer risk..
Beta	Consumer risk.
Number of items tested,	Table, containing the number of items checked and

number of faulty items	the number of faulty items.
Conclusion	Conclusion related to an individual row.

24.1.3. Graphical output

Sequential acceptance sampling chart 	Graphical illustration of the acceptance sampling procedure. The cumulative number of tested items is plotted on x-axis, while cumulative number of faulty items is plotted on y-axis. The red (upper) line corresponds to the rejection region boundary. When the line is crossed, no further tests are necessary and the whole lot is rejected immediately. The green (lower) line corresponds to the acceptance region boundary. When the line is crossed, no further tests are necessary and whole lot is accepted immediately. When the number of faulty items is between these two lines, sampling has to continue. The probability of making a decision increases as the number of inspected items increases.

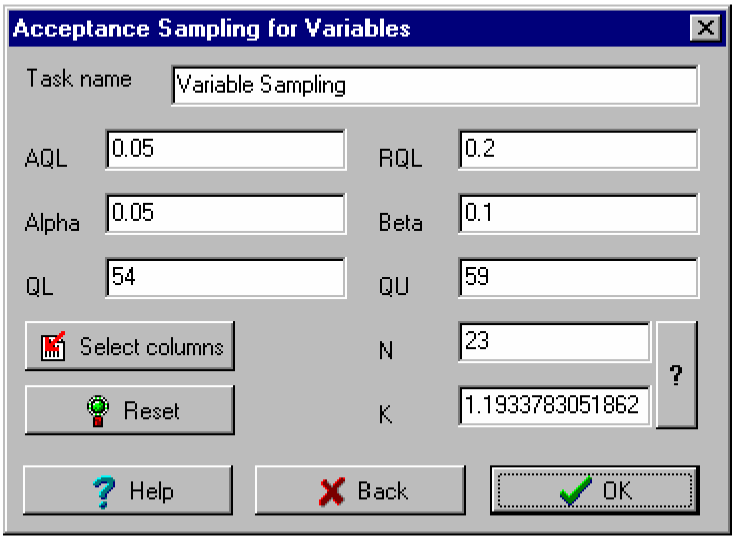
24.2. Acceptance sampling for variables

Menu:	QC.Expert	Acceptance sampling	Variables

The module helps to decide whether a quantitative characteristics of a material (e.g. size, purity, strength, mass, humidity) satisfies given requirements. Acceptance sampling for variables is useful even for materials, for which no individual pieces can be distinguished (e.g. fluid or gaseous materials). In such cases, a sample of appropriate size from the whole batch is taken. The sample is subsequently analyzed (measured). Size of an individual sample is related to the measurement method and its properties. Sampling has to be done in a random way, keeping in mind that each part of a sampled batch has to have the same probability of becoming a part of a sample.

24.2.1. Data and parameters

The values of producer risk *Alpha* and consumer risk *Beta*, together with corresponding acceptable probability of unacceptable quality *AQL* and unacceptable probability of unacceptable quality *RQL* are entered in the Acceptance sampling for variables dialog panel, see Figure 21. Meaning of these four parameters is analogous to the meaning of parameters of the attributes acceptance sampling, see 5.4.1. QL stands for the minimum value allowed for the measured variable (i.e. lower limit), similarly QU stands for the maximum value allowed. When only one of these two is given (e.g. minimum metal content in an ore, maximum heavy metals content in milk), the other field is left blank. Common values for *Alpha*, *Beta, AQL, RQL* are set by the *Initialize* button. After the parameter values have been entered, the required sample size N and coefficient K are computed upon pressing "?". These two values specify **acceptance sampling plan**. Upon pressing the *Select columns* button, the user is prompted to select data columns. The *Compute* button runs computations necessary to analyze the data. Each column should contain N data points corresponding to one batch. When computations are finished, the program produces a plot and gives a conclusion in words. Producer risk *Alpha* is the probability (risk) of rejecting a good batch (i.e. batch of acceptable quality), while consumer risk Beta is the probability (risk) of accepting a batch of unacceptably poor quality.



Acceptance Sampling for Variables

Task name: Variable Sampling

AQL: 0.05

RQL: 0.2

Alpha: 0.05

Beta: 0.1

QL: 54

QU: 59

N: 23

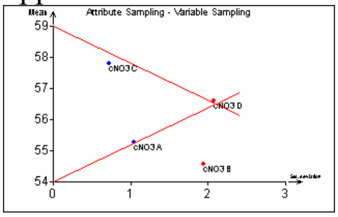
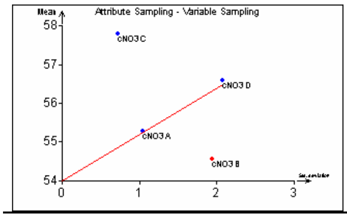
K: 1.1933783051862

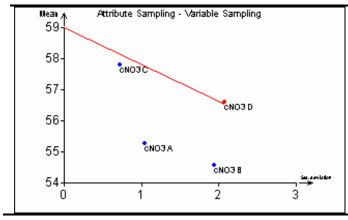
Fig. 121 Acceptance sampling for variables dialog panel

24.2.2. Protocol

	Values entered in the Acceptance sampling for variables dialog panel, Figure 21.
	Minimum acceptable level.
QL	Maximum acceptable level.
QU	Acceptable probability of unacceptable quality.
AQL	Unacceptable probability of unacceptable quality. Producer risk.
RQL	Consumer risk.
Alpha	
Beta	
Mean	Column average.
Sigma	Column standard deviation.
Conclusion	Conclusion in words (acceptable/unacceptable).

24.2.3. Graphical output

Acceptance sampling with upper and lower limits  	When both limits are specified, a batch is accepted if the corresponding point is within the limits given by the red lines. When a point falls above or below the triangle, the corresponding batch is not accepted because the measured values are too high or too small. When a point falls right from the triangle, no conclusion can be drawn because the variance of the measurements is too large. Measured values should then be checked in the Basic data analysis module for outliers. One might need to take more precise measurements.
Acceptance sampling with lower limit  	When only the lower limit is specified, a batch is accepted if the corresponding point falls above the line. It is clear that acceptance is influenced both by more desirable mean and by smaller variance. Unacceptable batches are marked red.
Acceptance sampling with upper limit	When only the upper limit is specified, a batch is accepted if the corresponding point falls below the line. It is clear that acceptance is influenced both by more desirable mean and by smaller variance.



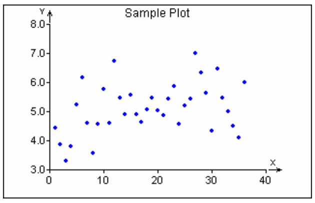
Unacceptable batches are marked red.

25. Plotting

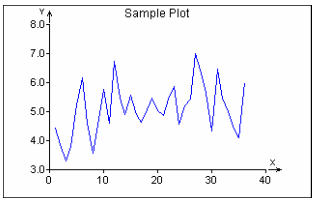
Menu:	QCExpert	Plotting

Module graphs offers various tools for visualization of uni- and multivariate data. Settings and options in different types of graphs allow for modifications and customiztions of the graphs. In the following table is a brief graphical overview of the possible graphs. Details of each grapf type will be delt with in the following paragraphs.

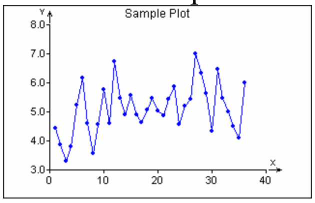
Points



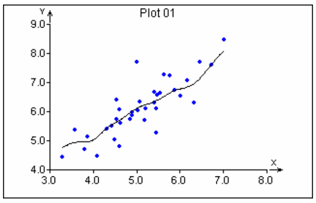
Lines



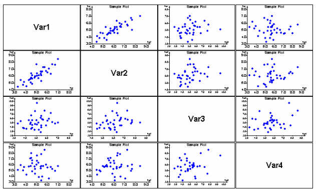
Connected points



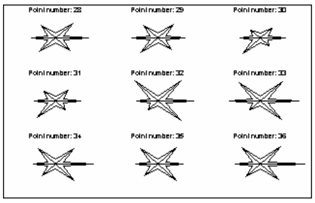
X-Y Scatter



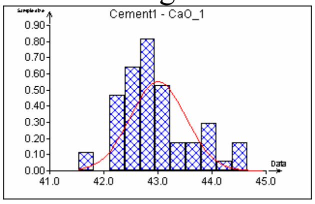
X-Y Matrix



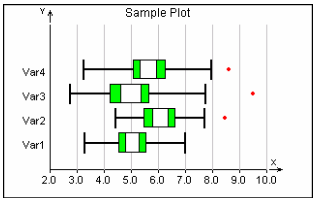
Star Plot



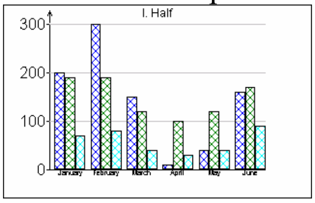
Histogram



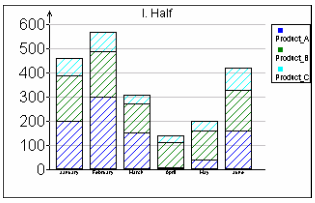
Box Plot



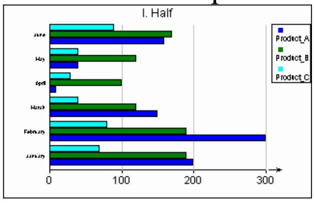
Bar Group



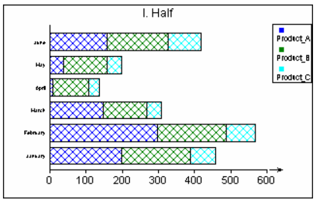
Bar Stacked



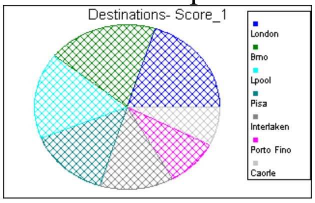
H-Bar Grouped



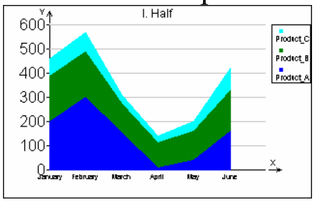
H-Bar Stacked



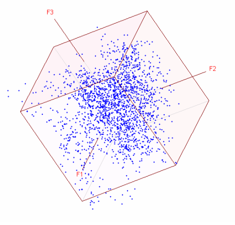
Pie Graph



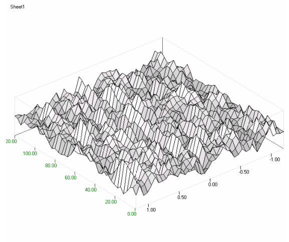
Area Graph



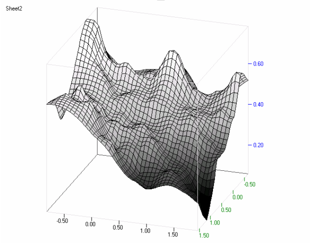
3D-Point Plot



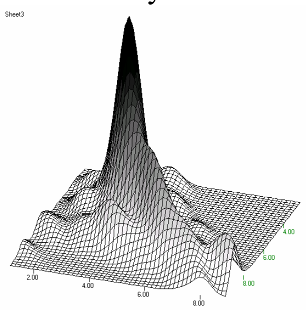
3D-Surface Plot



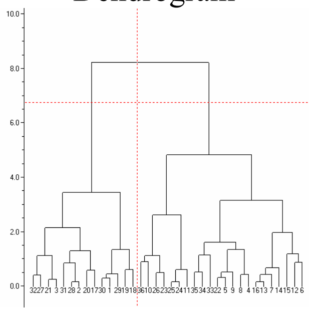
3D-Spline



3D-Density for 2 vars

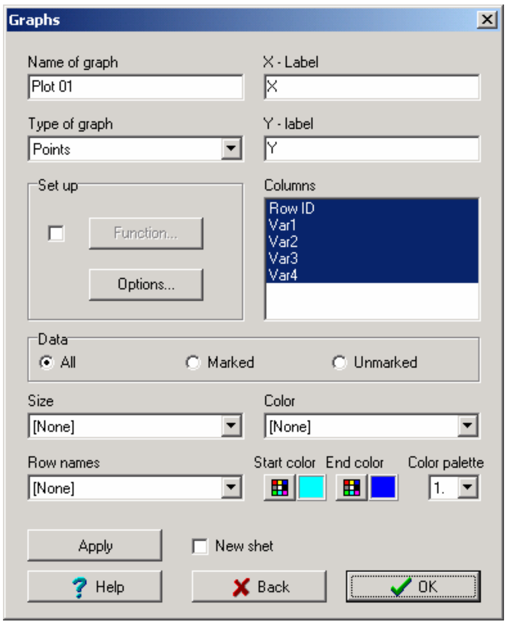


Dendrogram



25.1. Data and Parameters

Input data are expected in one or more columns in current data sheet. By selecting the Graphs menu item, the Graphs dialog window will be opened. The header may be written in the field *Name of graph*. Default header is taken from the name of the respective data sheet. Description of axes may be specified in *X-Label* and *Y-Label* fields. The reqired type of graph is selected in the drop-down list *Type of graph*. Options and setings are specific to every graph type and will be described in detail for each graph in the respective paragraphs below. The field *Columns* specifies which columns of the data sheet will be used, the group *Data* in the dialog window allows to specify rows to be used for the graph using marking the data, for marking data see paragraphs 4.1.1.1, page 4-14, or 4.3.3, page 4-28. If the checkbox New sheet is checked, every new graph will plot on new graph sheet, otherwise last graph will be overwritten each time a new graph is generated, unless the dialog window Graphs is closed. The pushbutton *Apply* is used to create a graph and leave the dialogbox open for further graphs, while tho *OK* button will create graph and close the dialogbox.



Graphs

Name of graph: Plot 01

X - Label: X

Type of graph: Points

Y - label: Y

Set up
□ Function...
Options...

Columns
Row ID
Var1
Var2
Var3
Var4

Data
● All ○ Marked ○ Unmarked

Size: [None]

Color: [None]

Row names: [None]

Start color End color Color palette: 1.

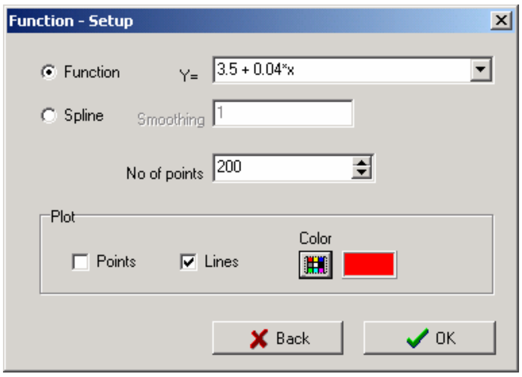
Fig. 122 Dialog window for the module Graphs

The module *Graphs* does not generate any output to *Protocol* with the exception of Dendrogram.

25.2. Graph types

25.2.1. Points plot

Creates a plot of data in the form of individual points, on the X-axis is the ordinal index and on the Y-axis is the value, see illustration A. The *Function* button enables to add a curve of a specified function. The function may be specified in the dialog box *Function* (Fig. 123) in field *Y*= in the form *f*(*x*). using common functions and mathematical notation, e.g.: **`3.5+0.05*x`**. Alternatively, a spline may be selected to fit the data. Spline will produce a non-parametric kernel smoothing using specified smoothing parameter. Function will always be plotter only once per plot, while spline will be calculated and plotted for each data column. *No of points* specifies how many intervals will be used to plot the curve of function/spline. The style of plotting the function/spline is affected by selection in the group *Plot*: Points will plot the function value just in the data points, lines will plot the curve. *Color* in the *Function-Setup* dialog window will specify the color of the line. The button *Options* opens *Options* dialog window, Checking the checkbox Legend will add legend in the plot for identifying columns, see illustration C.



Function - Setup

● Function Y= 3.5 + 0.04*x

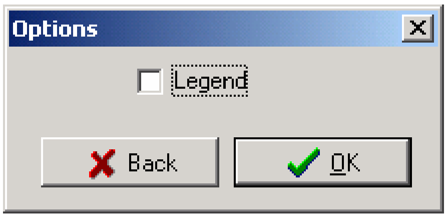
○ Spline Smoothing 1

No of points 200

Plot

□ Points ☑ Lines Color 🟥

Fig. 123 Function setup in module Graphs



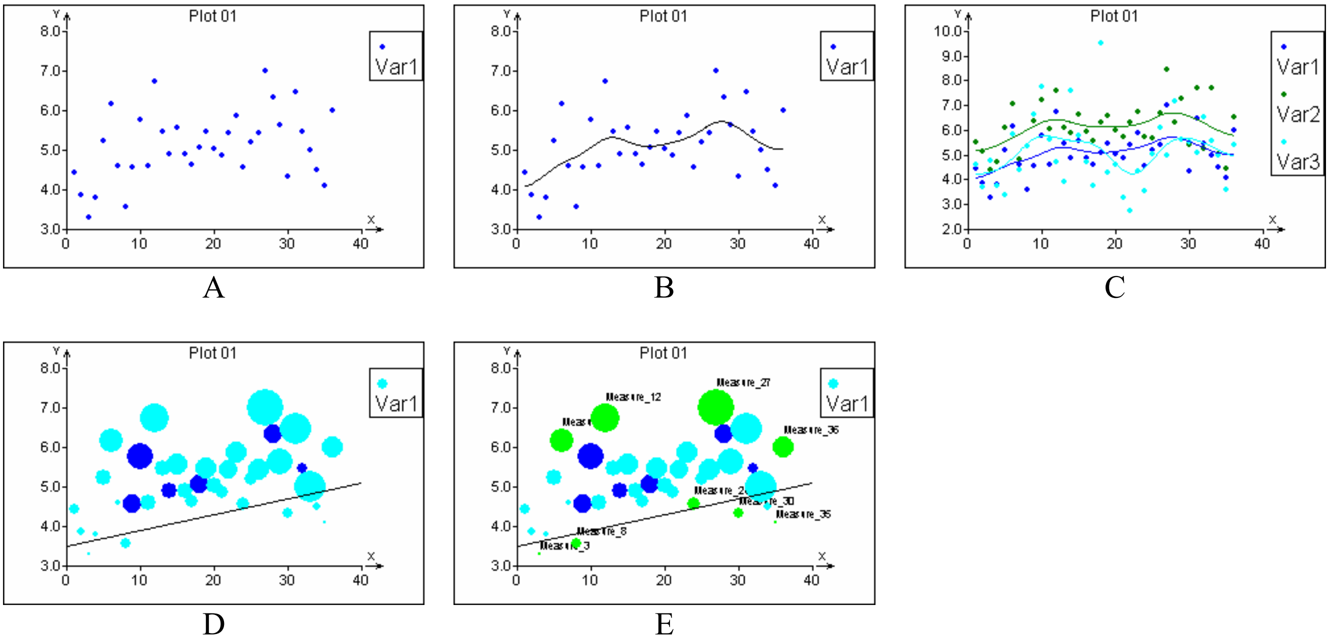
Options

□ Legend

Fig. 124 Legend setup in Points plot

In the Points plot it is possible to select further two columns in *Size* and *Color* drop-down fields to specify size and color of the plotted points. If a column is specified in *Size*, then diameter of each point is determined linearly by the values in the specified column within the range of this column. Similarly, if a column is specified in *Color*, then the color of each point is determined linearly between selected colors in the color space by the values in the specified column within the range of this column. Using color and size allows to visualize up to four dimensions on one 2D plot, specially in the XY-Scatter plot, see below. Number of data in all selected columns must be the same for the plot to work properly. Points plot may be combined with function (illustration D). If a column with row names is selected, these names may be used to describe the points by clicking on them with mouse in interactive plot (after double-clicking on the plot), see illustration E, for details on interacive plots, see paragraph 4.3.3 on page 4-28.

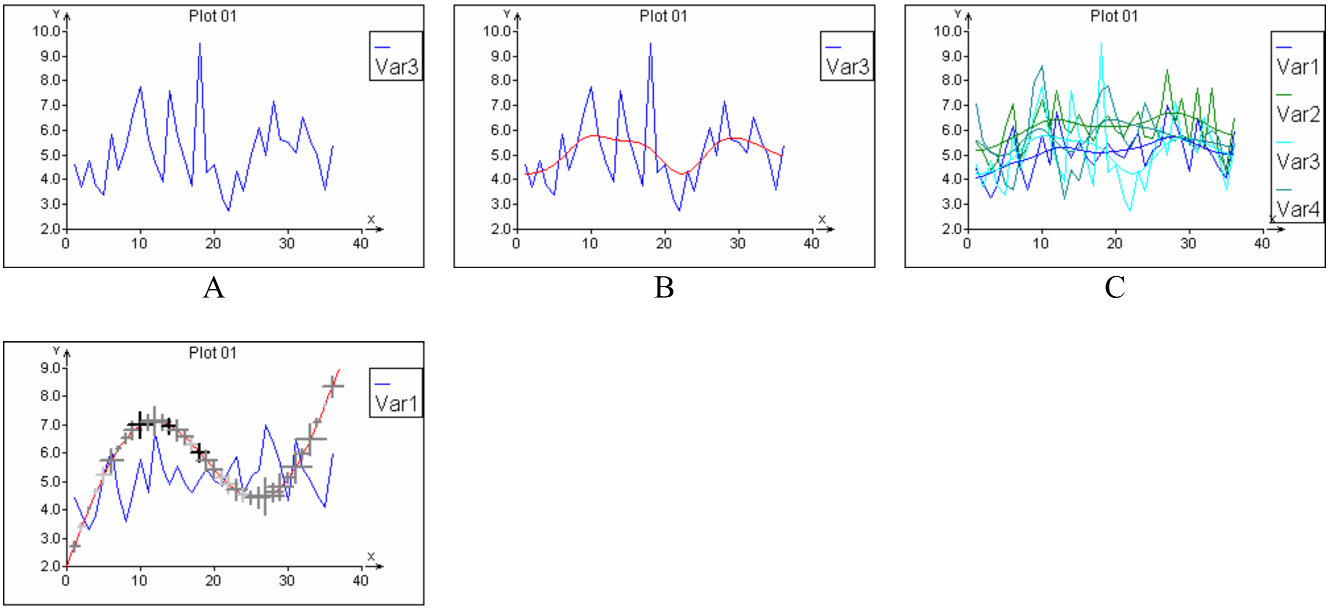
Illustrations



25.2.2. Lines plot

Creates a plot of data in the form of connected line segments, on the X-axis is the ordinal index and on the Y-axis is the value, see illustration A. The *Function* button enables to add a curve of a specified function. The function may be specified in the dialog box *Function* (Fig. 123) in field *Y*= in the form *f*(*x*). using common functions and mathematical notation, e.g.: **`0.5+3*sin(x/6)+0.05*x`**. Alternatively, a spline may be selected to fit the data, illustration B. Spline will produce a non-parametric kernel smoothing using specified smoothing parameter. Function will always be plotter only once per plot, while spline will be calculated and plotted for each data column. *No of points* specifies how many intervals will be used to plot the curve of function/spline. The style of plotting the function/spline is affected by selection in the group *Plot*: Points will plot the function value just in the data points, lines will plot the curve. *Color* in the *Function-Setup* dialog window will specify the color of the line. The button *Options* opens *Options* dialog window, Checking the checkbox *Legend* will add legend in the plot for identifying columns, see illustration C. Columns in Color ans Size drop-down menus will afect the points (crosses) of points plotted on selected function or spline, illustration D.

Illustration



A B C

D

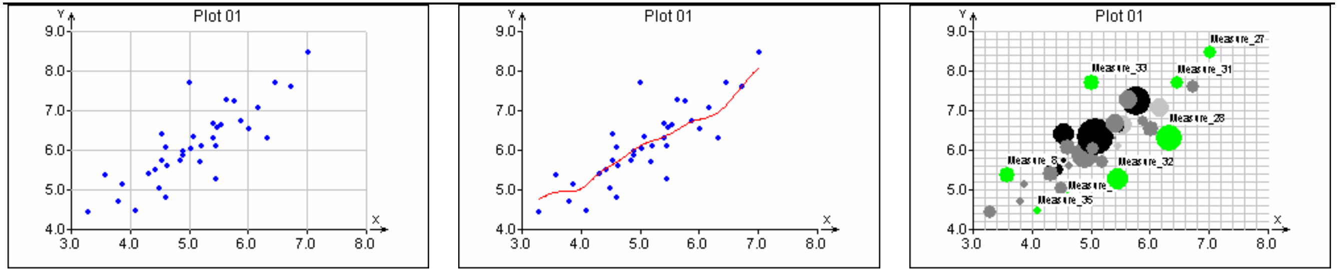
25.2.3. Connected points

Connected points plot is a combination of the two preceeding plots and its use is the same.

25.2.4. X-Y Scatter plot

X-Y Scatter plot displays two variables in one plot. Two columns must be selected in the field *Columns*. Data in the first column wil be plotted on the X axis, second column on the Y axis, see illustration A. Controls of this plot is similar to the Points plot as described above in paragraph 25.2.1. When combined with other two columns to control color and size of the plotted points, it is possible to visualize four dimensions in one plot, see illustration C, where some correlation between color and size can be spotted. As in the Points plot, the plotted data may be fitted with kernel smoothing curve (illustration B) or a user function can be added. Double clicking on the plot will allow to zoom in or to select and label separate points with row names specified in *Row names* drop-down field.

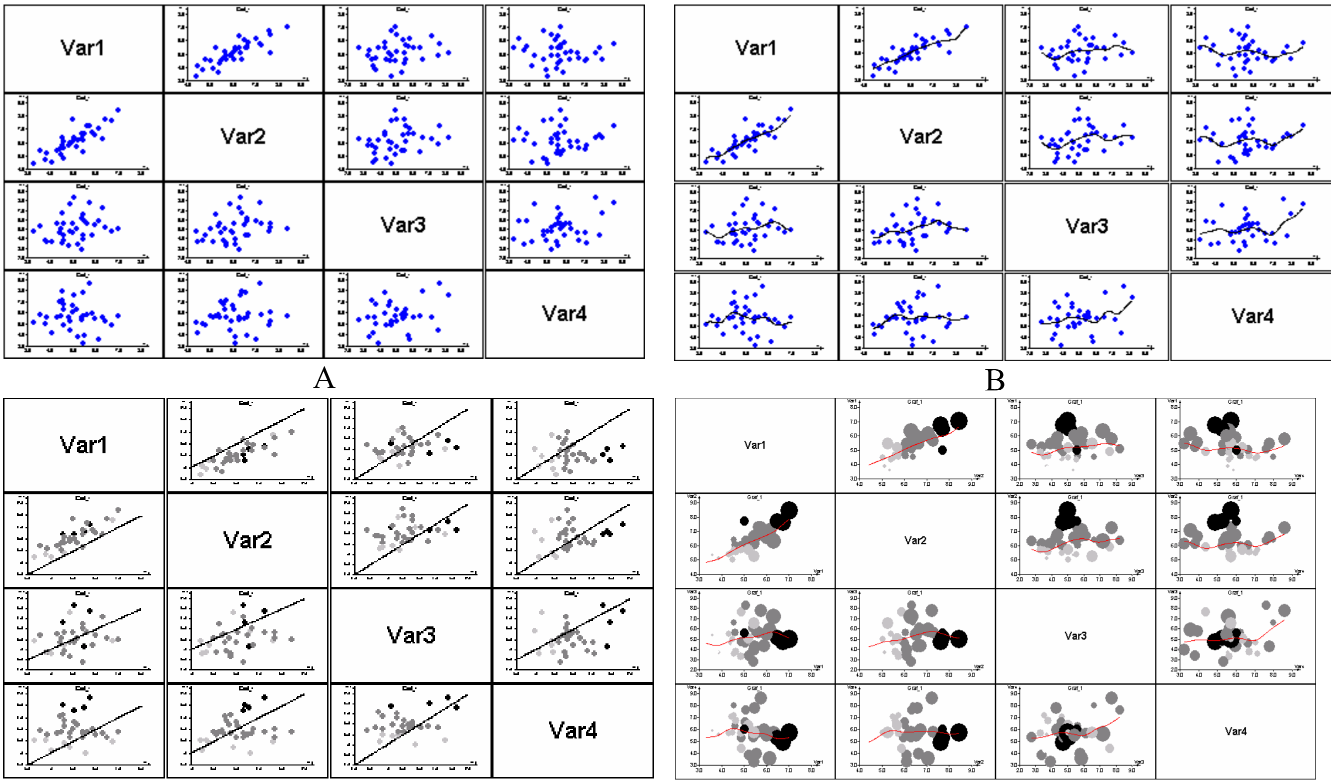
Illustrations



A B C

25.2.5. X-Y Matrix Plot

This plot is in fact a generalization of the previous X-Y Scatter plot and an alternative of the graphical output of the Correlation module, see paragraph 16.1, page 16-104. It plots scatter plots for all pairs of the selected columns, illustration A. Input data must have at least two columns. Like in the case of X-Y Scatter plot, smoothing curve or a user function may be displayed in each graph using *Function* button, illustration B, C. Other two columns may be used to control size and color on the plots as shown on illustrations C and D.

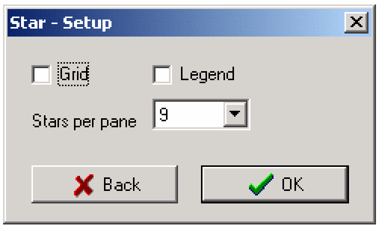


A B

C D

25.2.6. Stars plot

Stars plot is a standard tool to visualize multidimensional data. Input data are two or more columns selected in the *Columns* drop-down field. In this plot every data point represented by one row is displayed as one star. Each tip of a star represents one value in the corresponding row. Long tip means big value. This plot is used to review multidimensional data and to find similar or dissimilar values. Different behavior of data can be seen on the last two stars in illustration E. Number of plotted stars correspnds to the number of rows in selected columns. The number of stars to be displayed in one graph (1 to 64) can be set in *Options* window, see illustrations. Checking *Legend* will display the meaning of the pins, illustration B. Checking Grid in *Options* will add a radius in each tip, compare illustrations A and D.

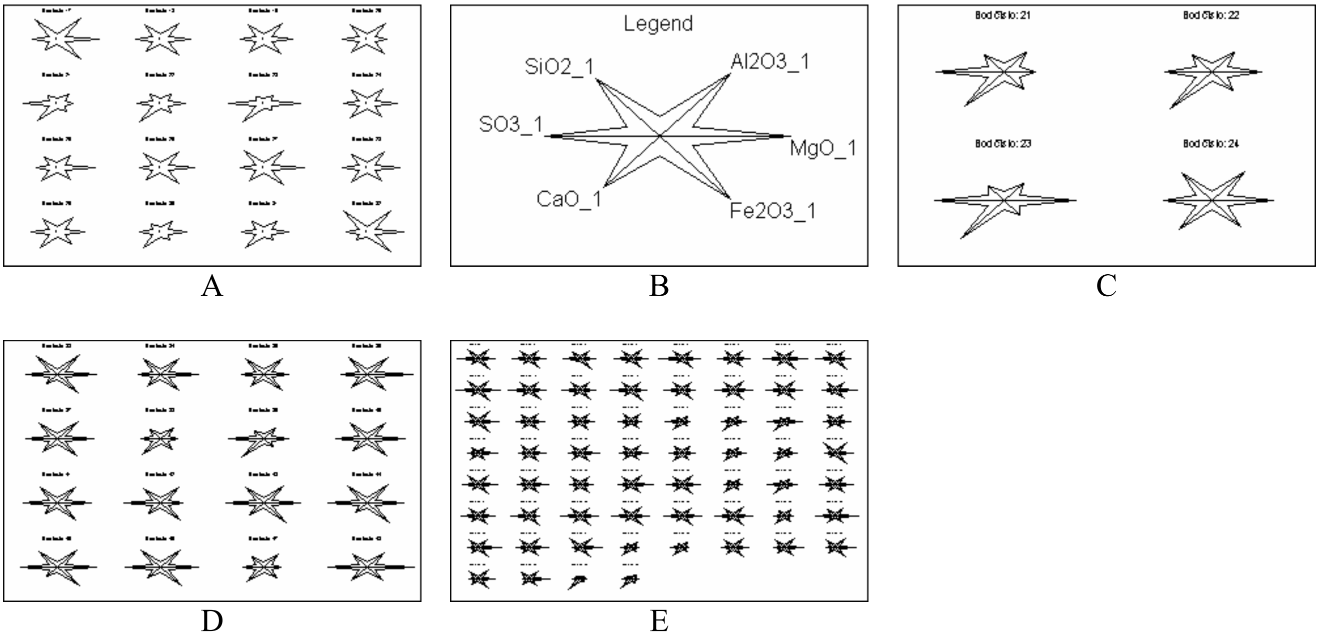


Star - Setup

□ Grid □ Legend

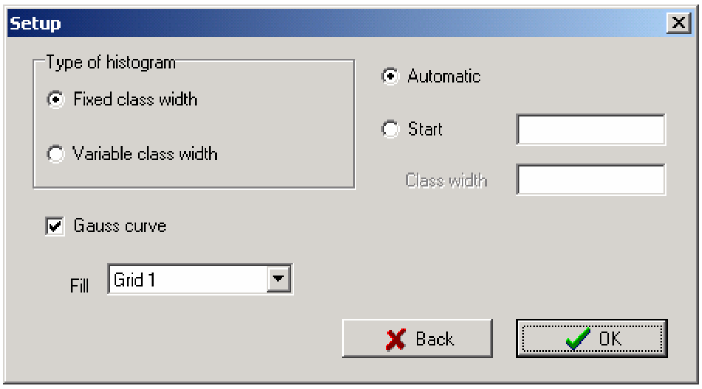
Stars per pane: 9

Fig. 125 Options for Star plot



25.2.7. Histogram

Draws histograms for all selected columns. By pressing *Options* button it is possible to set properties common to all histograms to be plotted. *Fixed Class width* will plot classical histogram with all bars of the same width, the height of the bars is the count of data in the class, illustrations A, B, C, D. Variable class width will generate variable class histogram, where there is constant number of data in each class (bar). This type of histogram has unit area and has usually higher information value. On illustration A-E histograms are constructed for the same data, only the variable class histogram (E) clearly suggests possible bi-modal distribution. For fixed class, user can specify start and class width, these two values are used for all selected columns and can be used to compare several data samples. Checking *Gauss curve* will generate a normal distribution probability density curve over the histogram, histogram is transformed to the scale of the probability density curve, so that the height of the bars is rather probability density than count. In the Fill field one of three filling grids may be selected.



Setup

Type of histogram
● Fixed class width
○ Variable class width

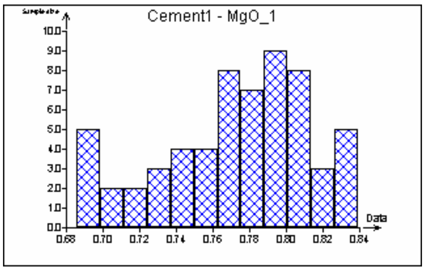
● Automatic
○ Start
Class width

☑ Gauss curve

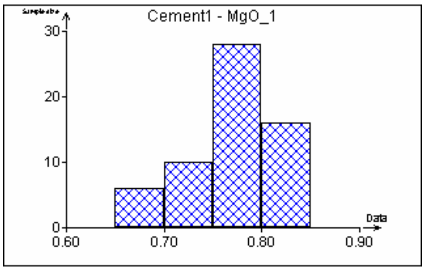
Fill: Grid 1

Fig. 126 Setup dialog window for Histogram

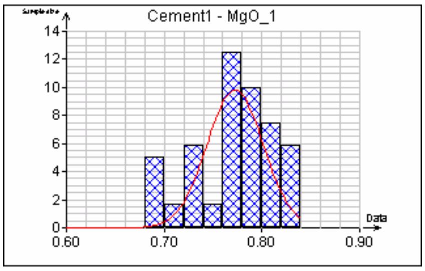
On illustrations A-E there are histograms for the same data with various settings: Automatic class width (A), too wide class (B) and too narrow class (D). Variable class width (E) reveals apparent (yet not statistically confirmed) bimodality of data at 0.77 and 0.79, which is not obvious from any of the previous histograms.



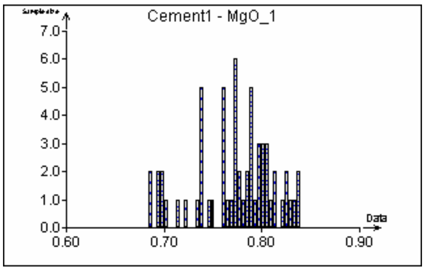
A



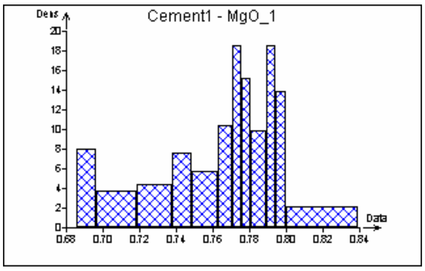
B



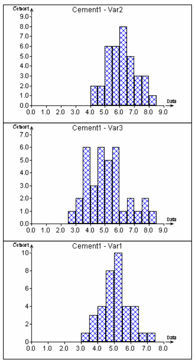
C



D



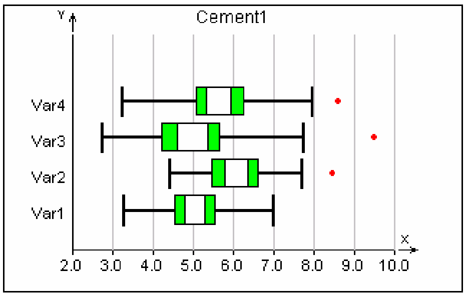
E



F

25.2.8. Box plot

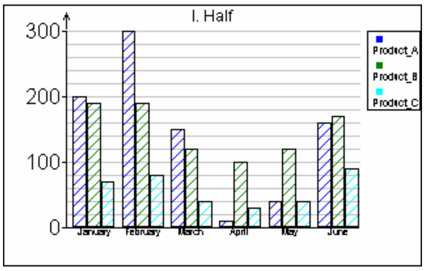
This plot draws box plots for all selected columns, see illustration. Box plot is a standard diagnostic tool used to assess symmetry of data and presence of outliers. The large box contains 50% of the data, its upper edge corresponds to 75th percentile, its lower edge to the 25th percentile. Median is located in the middle of the white rectangle inside the green box. Width of the white rectagle inside the green box corresponds to the width of the confidence interval for the median. Two black lines correspond to the inner fence. The data points outside the inner fence are marked red. They might be considered as outliers. This plot is also generated by ANOVA, see paragraph 12, p. 12-82.



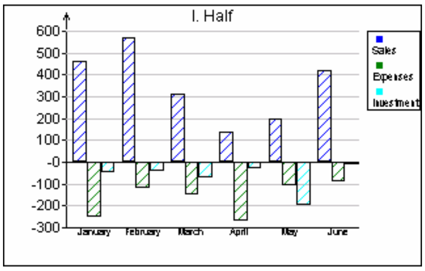
A

25.2.9. Vertical and Horizontal Bar plots

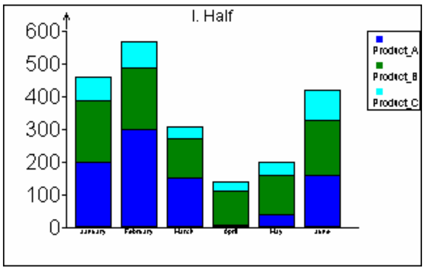
These plots are to visualize data for different classes. Data are expected in one or more columns, one column may contain row names, which should be specified in the *Row names* field. Button *Options* will allow to display a legent on the graph and to select fill type for the bars. Grouped bars will construct a separate bar for each row and each column, illustration A, B, E, F. Bars start from zero and values for bars may contain negative numbers, which plot on the negative part of axis, illustrations B, F, H. Columns are distinguished by color. On Stacked bars, the values form all columns are added into one bar for each row, illustrations C, D, G, H.



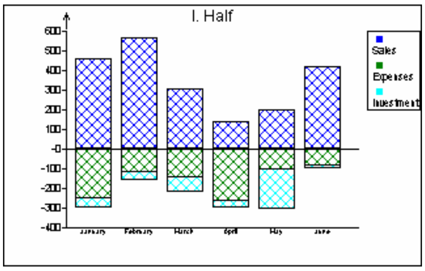
A



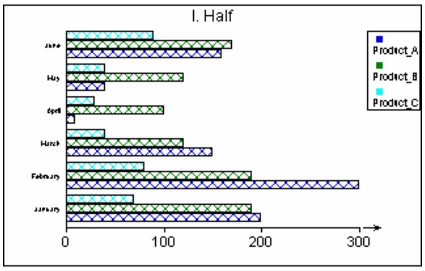
B



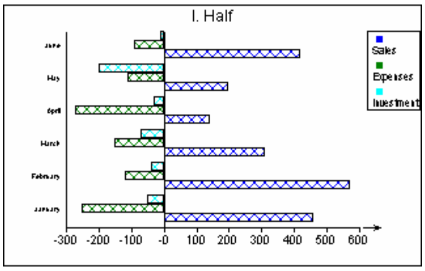
C



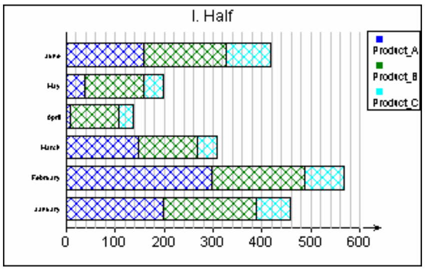
D



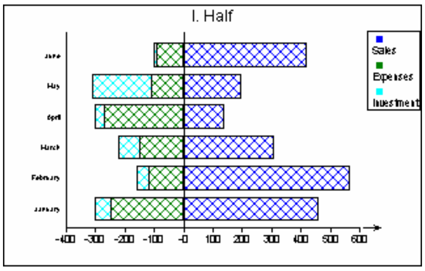
E



F



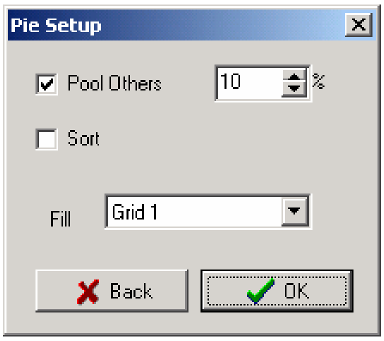
G



H

25.2.10. Pie chart

Pie chart is used to visualize parts of a defined unity as fractions of a circle. Data may be in one or more columns and must be non-negative (negative data are taken as zeros). One pie chart is constructed for each selected column. Row names if specified are used in legend on the plot, see illustration A, C. If no row names are specified, the legend shows the absolute data values, illustration B. The checkbox *Pool others* in *Options* dialog box will cause the specified fraction of values in columns to form one section on the chart named „Others", illustration E. If the checkbox *Sort* is checked, the data is sorted ascendingly, illustration D. Fill type for the bars selected.



Pie Setup

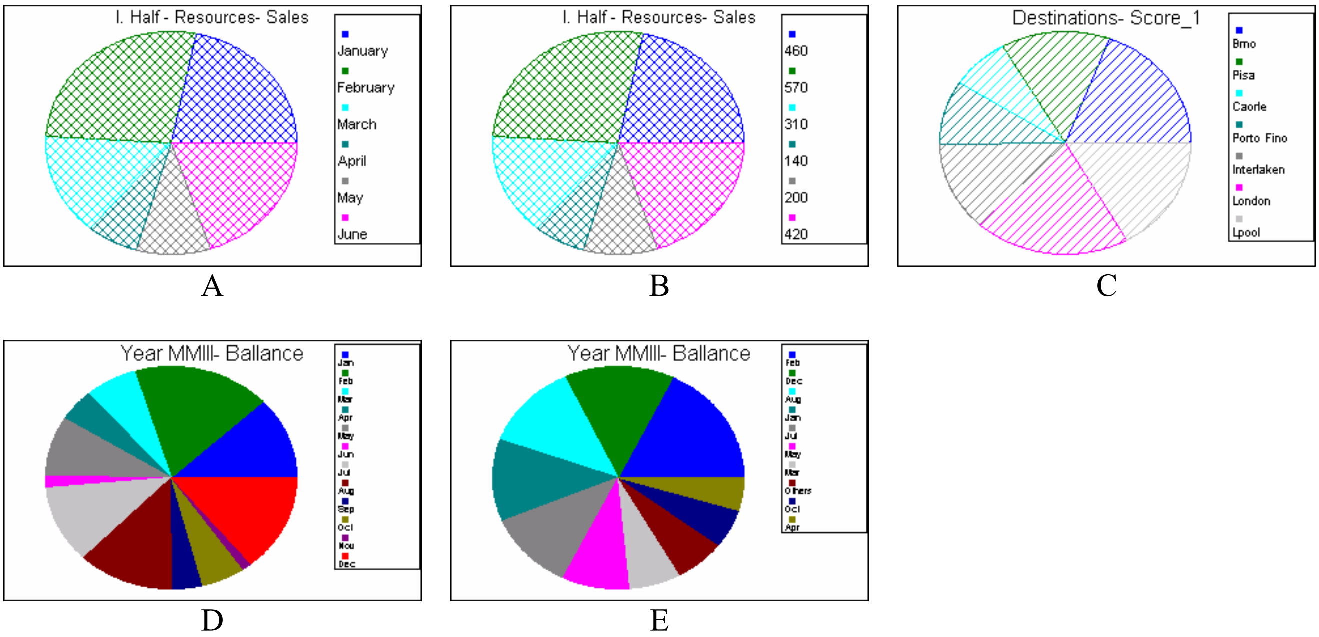
☑ Pool Others 10 %

☐ Sort

Fill Grid 1

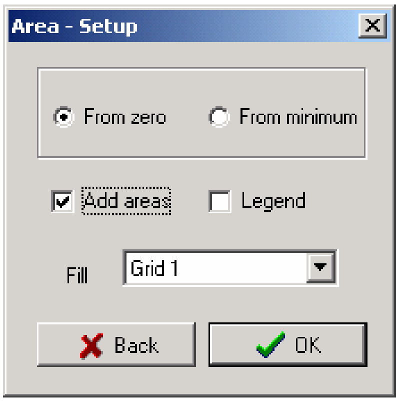
Fig. 127 Options for the Pie chart plot

Illustrations



25.2.11. Area plot

Area plot has similar use as the bar plot. Values in the selected columns are ploted on the Y axis. Basic form of the area plot is shown on illustration A. In the case of high values with low variability it may be more suitable to use *From minimum* option instead of the standard *From zero*, see illustration B. If more columns are selected, the areas may be added together by checking *Add* checkbox (illustraiton C). Negative data can be plotted only on the non-added plot, illustrations E, F. In the added plot, negative values are taken as zeros. Fill type for the bars selected in the options dialog window.

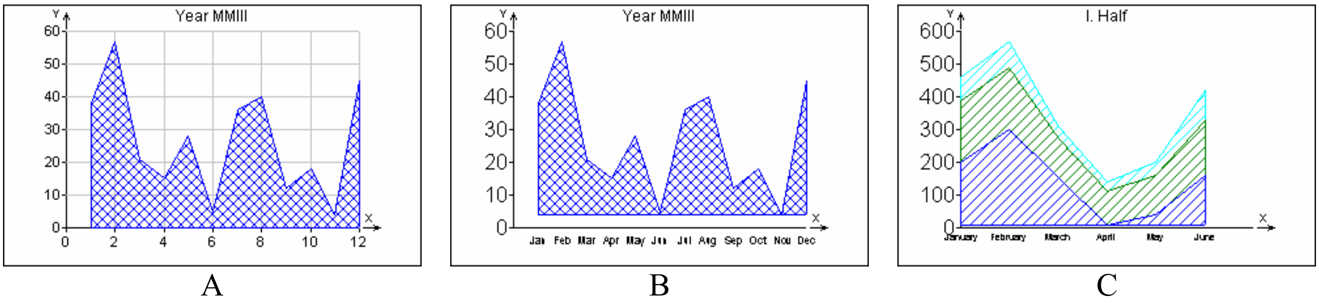

Area - Setup

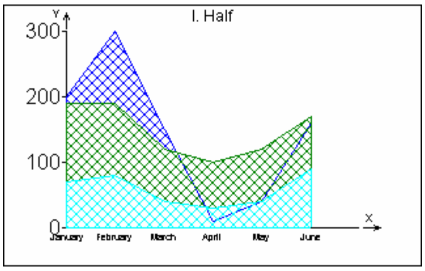
● From zero ○ From minimum

☑ Add areas □ Legend

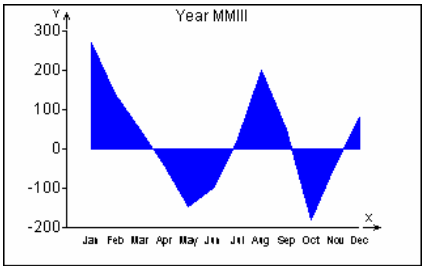
Fill Grid 1

Fig. 128 Options for the Area plot

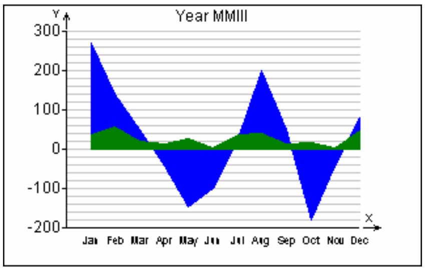




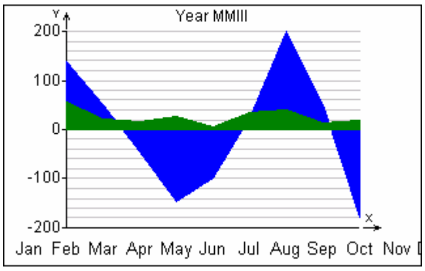
D



E



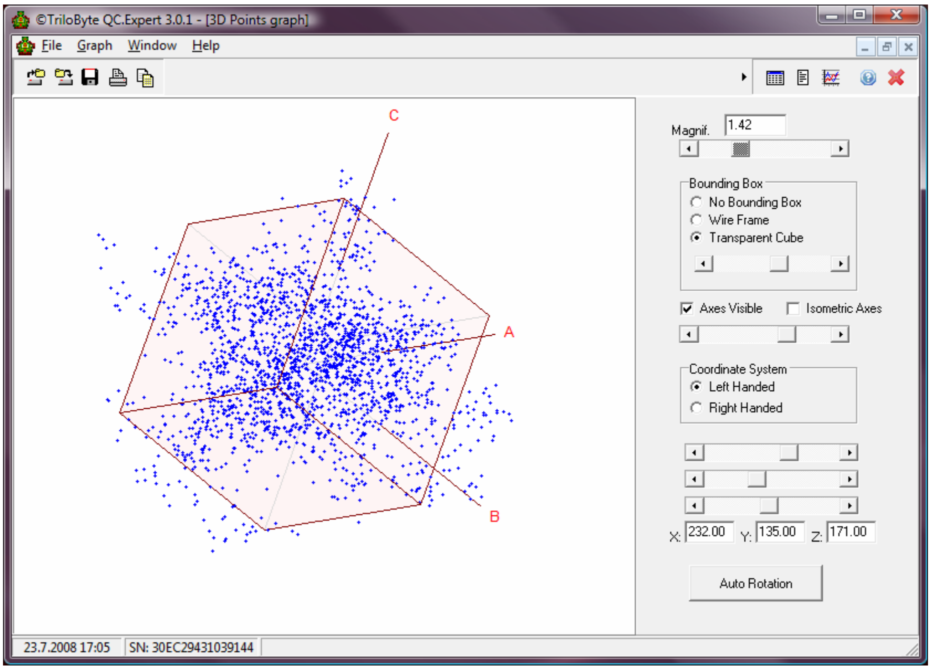
F



G

25.2.12. 3D-point plot

Plots 3 selected columns in a 3d-point plot. This type of a plot is especially suitable for assessing structure, properties relationships or homogeneity of a 3-dimensional data. The 3d-plot can reveal relationships which cannot be observed at 2d scatter plots, pair correlations, etc. 3d-point plot is an extention of the 2-scatter plot. The plot can be rotated, continuously scaled and moved in the graphical window. An individual point can be labeled with a mouse click either by its row number or by a column selected in the *Graphs* dialog window. The scale of the three axes is set so that the values appear normalised. Actual non-normalised scale can be set by the checkbox *Isometric Axes*.



Magnif. 1.42

Bounding Box
○ No Bounding Box
○ Wire Frame
● Transparent Cube

☑ Axes Visible ☐ Isometric Axes

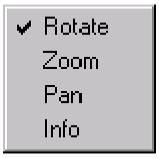
Coordinate System
● Left Handed
○ Right Handed

X: 232.00 Y: 135.00 Z: 171.00

23.7.2008 17:05 SN: 30EC29431039144

Fig. 129 3D-Point Plot dialog window

To display the 3d-point Plot, first select the graph type and three data columns (see Fig. 122 on page 25-198), then press OK or Apply. Non-contiguous columns are selectes with Ctrl-mouse click. Primarily, the plot is controlled by a mouse. It can be rotated, scaled and moved. Use right mouse-click to select the mouse function:

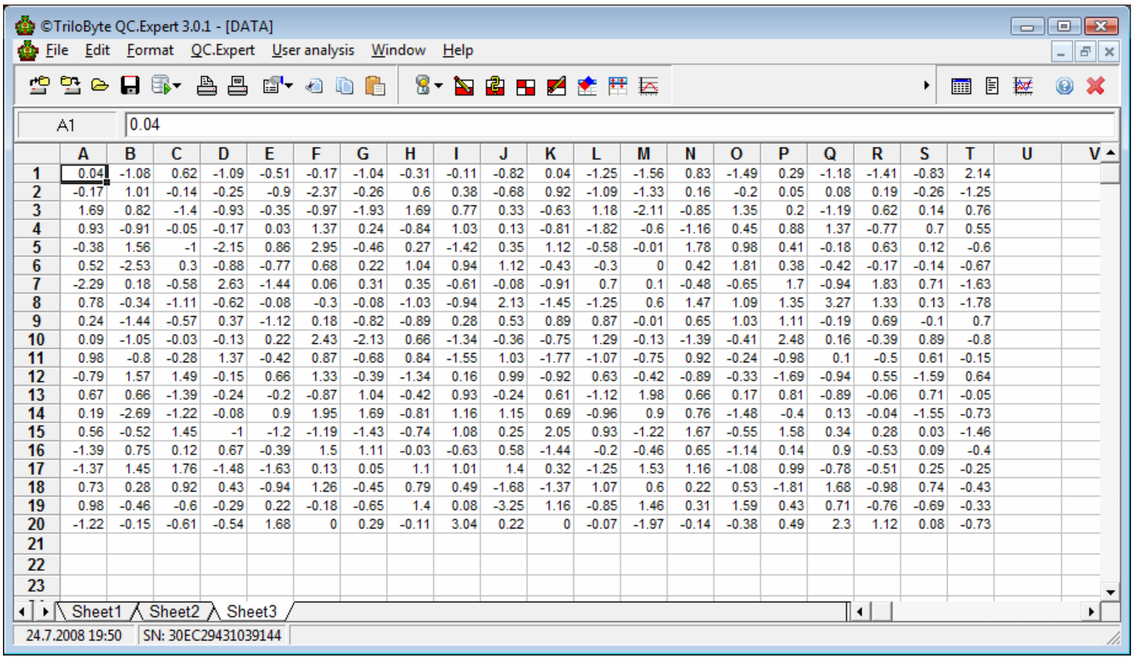


Rotate – rotate the plot with mouse in any direction to explore the data and observe shapes and unusual fetures in data, like outliers, clusters, non-linearity, correlations.
Zoom – move the mouse up or down to move the plot closer or farther.
Pan – move the plot with a mouse.
Info – Click the mouse near a point in the plot to display its row number or (if selected previously in the *Graphs* dialog window as *Row names*, see Fig. 122 on page 25-198) its description. This function is applicable only to one point at a time.

The right part of the plot window offers controls for the plot including setting the scale, boxing style, axes visibility, isometry (real scale of the data), left- or right-handed orientation. The Auto Rotation button will rotate the plot automatically in random directions. From the main menu you can copy, save or print the plot.

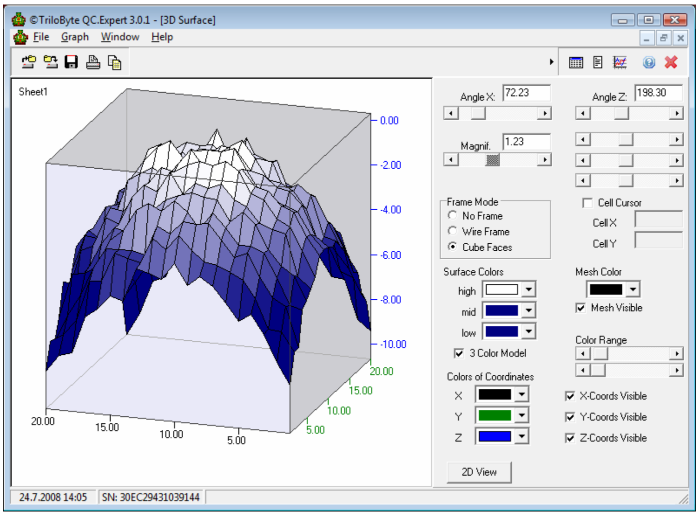
25.2.13. 3D-surface plot

This plot will display an *n* x *m* data matrix as a 3d-plot. Rows and columns are used as X and Y coordinates, the values in the matrix are plotted on the Z axis. Following examples illustrate the use of the plot.


©TriloByte QC.Expert 3.0.1 - [DATA]
File Edit Format QC.Expert User analysis Window Help
A1
0.04
Sheet1 Sheet2 Sheet3
24.7.2008 19:50
SN: 30EC29431039144

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	0.04	-1.08	0.62	-1.09	-0.51	-0.17	-1.04	-0.31	-0.11	-0.82	0.04	-1.25	-1.56	0.83	-1.49	0.29	-1.18	-1.41	-0.83	2.14
2	-0.17	1.01	-0.14	-0.25	-0.9	-2.37	-0.26	0.6	0.38	-0.68	0.92	-1.09	-1.33	0.16	-0.2	0.05	0.08	0.19	-0.26	-1.25
3	1.69	0.82	-1.4	-0.93	-0.35	-0.97	-1.93	1.69	0.77	0.33	-0.63	1.18	-2.11	-0.85	1.35	0.2	-1.19	0.62	0.14	0.76
4	0.93	-0.91	-0.05	-0.17	0.03	1.37	0.24	-0.84	1.03	0.13	-0.81	-1.82	-0.6	-1.16	0.45	0.88	1.37	-0.77	0.7	0.55
5	-0.38	1.56	-1	-2.15	0.86	2.95	-0.46	0.27	-1.42	0.35	1.12	-0.58	-0.01	1.78	0.98	0.41	-0.18	0.63	0.12	-0.6
6	0.52	-2.53	0.3	-0.88	-0.77	0.68	0.22	1.04	0.94	1.12	-0.43	-0.3	0	0.42	1.81	0.38	-0.42	-0.17	-0.14	-0.67
7	-2.29	0.18	-0.58	2.63	-1.44	0.06	0.31	0.35	-0.61	-0.08	-0.91	0.7	0.1	-0.48	-0.65	1.7	-0.94	1.83	0.71	-1.63
8	0.78	-0.34	-1.11	-0.62	-0.08	-0.3	-0.08	-1.03	-0.94	2.13	-1.45	-1.25	0.6	1.47	1.09	1.35	3.27	1.33	0.13	-1.78
9	0.24	-1.44	-0.57	0.37	-1.12	0.18	-0.82	-0.89	0.28	0.53	0.89	0.87	-0.01	0.65	1.03	1.11	-0.19	0.69	-0.1	0.7
10	0.09	-1.05	-0.03	-0.13	0.22	2.43	-2.13	0.66	-1.34	-0.36	-0.75	1.29	-0.13	-1.39	-0.41	2.48	0.16	-0.39	0.89	-0.8
11	0.98	-0.8	-0.28	1.37	-0.42	0.87	-0.68	0.84	-1.55	1.03	-1.77	-1.07	-0.75	0.92	-0.24	-0.98	0.1	-0.5	0.61	-0.15
12	-0.79	1.57	1.49	-0.15	0.66	1.33	-0.39	-1.34	0.16	0.99	-0.92	0.63	-0.42	-0.89	-0.33	-1.69	-0.94	0.55	-1.59	0.64
13	0.67	0.66	-1.39	-0.24	-0.2	-0.87	1.04	-0.42	0.93	-0.24	0.61	-1.12	1.98	0.66	0.17	0.81	-0.89	-0.06	0.71	-0.05
14	0.19	-2.69	-1.22	-0.08	0.9	1.95	1.69	-0.81	1.16	1.15	0.69	-0.96	0.9	0.76	-1.48	-0.4	0.13	-0.04	-1.55	-0.73
15	0.56	-0.52	1.45	-1	-1.2	-1.19	-1.43	-0.74	1.08	0.25	2.05	0.93	-1.22	1.67	-0.55	1.58	0.34	0.28	0.03	-1.46
16	-1.39	0.75	0.12	0.67	-0.39	1.5	1.11	-0.03	-0.63	0.58	-1.44	-0.2	-0.46	0.65	-1.14	0.14	0.9	-0.53	0.09	-0.4
17	-1.37	1.45	1.76	-1.48	-1.63	0.13	0.05	1.1	1.01	1.4	0.32	-1.25	1.53	1.16	-1.08	0.99	-0.78	-0.51	0.25	-0.25
18	0.73	0.28	0.92	0.43	-0.94	1.26	-0.45	0.79	0.49	-1.68	-1.37	1.07	0.6	0.22	0.53	-1.81	1.68	-0.98	0.74	-0.43
19	0.98	-0.46	-0.6	-0.29	0.22	-0.18	-0.65	1.4	0.08	-3.25	1.16	-0.85	1.46	0.31	1.59	0.43	0.71	-0.76	-0.69	-0.33
20	-1.22	-0.15	-0.61	-0.54	1.68	0	0.29	-0.11	3.04	0.22	0	-0.07	-1.97	-0.14	-0.38	0.49	2.3	1.12	0.08	-0.73

Fig. 130 20 x 20 data matrix for the 3D surface plot



©TriloByte QC.Expert 3.0.1 - [3D Surface]

File Graph Window Help

Sheet1

Angle X: 72.23

Angle Z: 198.30

Magnif. 1.23

Frame Mode
○ No Frame
○ Wire Frame
● Cube Faces

□ Cell Cursor
Cell X
Cell Y

Surface Colors
high
mid
low
☑ 3 Color Model

Mesh Color
☑ Mesh Visible

Color Range

Colors of Coordinates
X
Y
Z

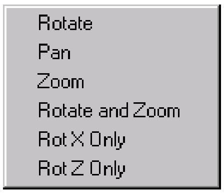
☑ X-Coords Visible
☑ Y-Coords Visible
☑ Z-Coords Visible

2D View

24.7.2008 14:05 SN: 30EC29431039144

Fig. 131 3D-surface plot for spatial data, 20 x 20 data matrix

The plot menu is activated by the right mouse click. The menu offers the following control functions. The 3D-plot can be rotated, zoomed and moved interactively with mouse. The Z-level of the surface can be coloured with 2 or 3 colors. There are more controls on the 3D-spline plot control panel which are used to modify the look of the plot. The button *2D View* is used to display an orthogonal projection into XY plane. For this view it is recommended to uncheck the *Mesh Visible* checkbox and select contrast colors. Right mouse click will display the control menu:



Rotate – Mouse movement will rotate the plot.
Pan –Move the plot with mouse.
Zoom – Mouse movement will zoom the plot.
Rotate and Zoom – Rotate with left mouse button, zoom with right mouse button.
Rotate X – Rotate only around X-axis
Rotate Z – Rotate only around Z-axis

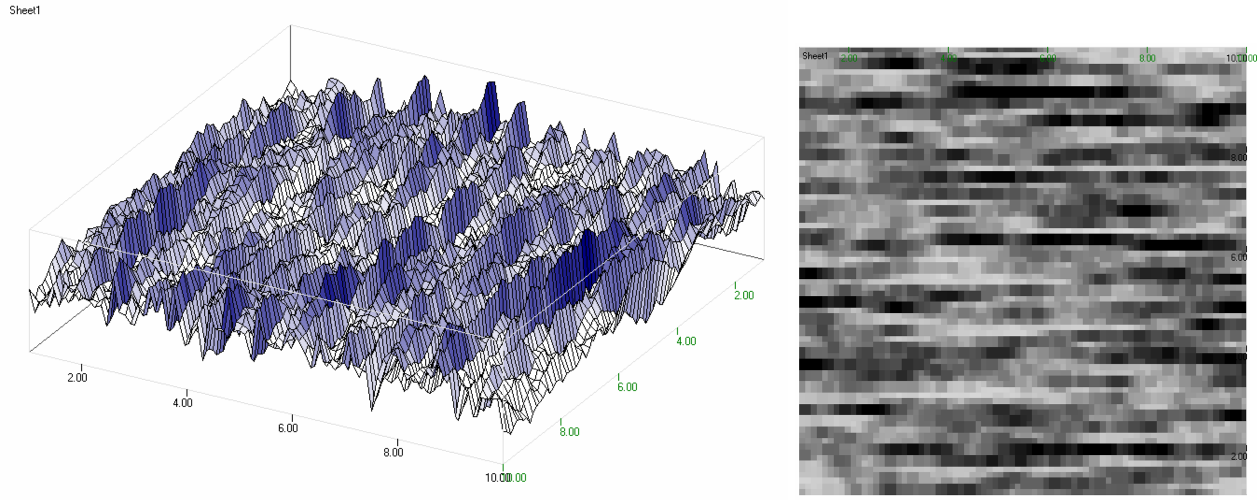


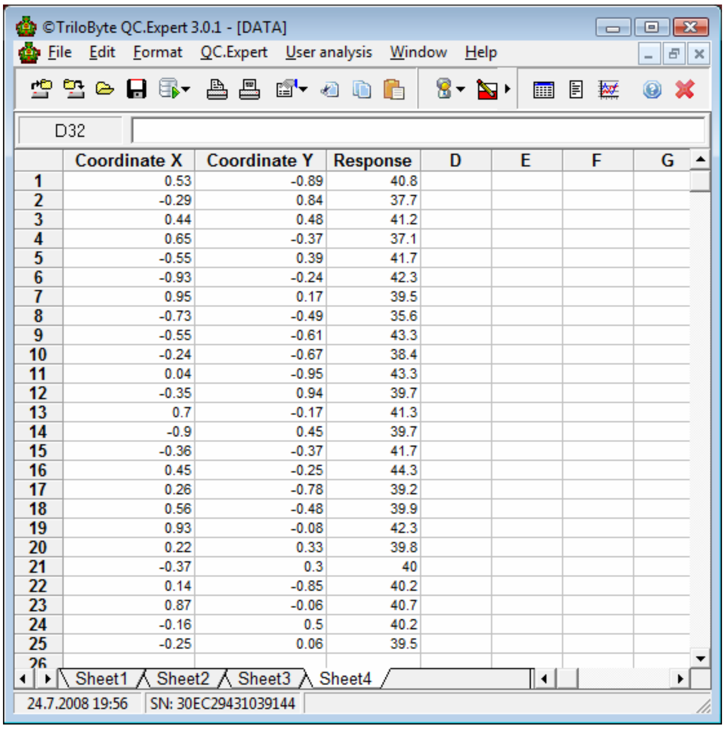
Fig. 132 a, b Multivariate time series view in general angle and in its orthogonal projection.

25.2.14. 3D-spline

This plot smoothes data in three columns with a Gaussian local kernel smoother

$$z_s(x,y) = \frac{\sum_{i=1}^{n} z_i \exp - \left(\frac{r_{xy}}{u} \sqrt{(x_i - x)^2 + (y_i - y)^2} \right)}{\sum_{i=1}^{n} \exp - \left(\frac{r_{xy}}{u} \sqrt{(x_i - x)^2 + (y_i - y)^2} \right)},$$

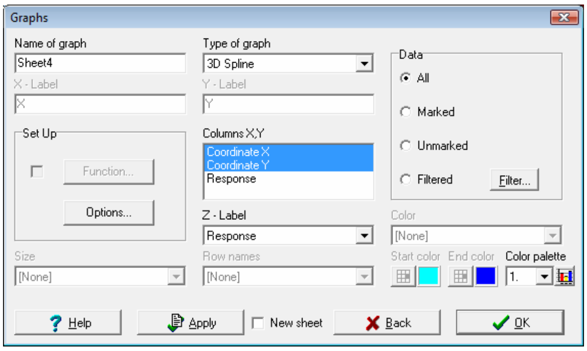
where u is the smoothing parameter, the coefficient r_{xy} is calculated from range of x and y. Thus, the smoothing parameter is independent of the ranges of variables.



	Coordinate X	Coordinate Y	Response	D	E	F	G
1	0.53	-0.89	40.8				
2	-0.29	0.84	37.7				
3	0.44	0.48	41.2				
4	0.65	-0.37	37.1				
5	-0.55	0.39	41.7				
6	-0.93	-0.24	42.3				
7	0.95	0.17	39.5				
8	-0.73	-0.49	35.6				
9	-0.55	-0.61	43.3				
10	-0.24	-0.67	38.4				
11	0.04	-0.95	43.3				
12	-0.35	0.94	39.7				
13	0.7	-0.17	41.3				
14	-0.9	0.45	39.7				
15	-0.36	-0.37	41.7				
16	0.45	-0.25	44.3				
17	0.26	-0.78	39.2				
18	0.56	-0.48	39.9				
19	0.93	-0.08	42.3				
20	0.22	0.33	39.8				
21	-0.37	0.3	40				
22	0.14	-0.85	40.2				
23	0.87	-0.06	40.7				
24	-0.16	0.5	40.2				
25	-0.25	0.06	39.5				
26							

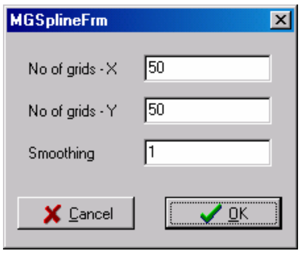
Fig. 133 Three columns for construction of the 3D spline

Data are expected in three columns, In the dialog box, two columns x, y must be selected in the field *Columns X,Y* and one response column must be selected in the *Z-Label* field. In Options form, the user may choose the density of the spline framework, or grid resolution and the value of the smoothing parameters. For most cases the default values can be left unchanged.



Name of graph: Sheet4
Type of graph: 3D Spline
X - Label: X
Y - Label: Y
□
Columns X,Y: Coordinate X, Coordinate Y (selected); Response
Z - Label: Response
Data: ● All ○ Marked ○ Unmarked ○ Filtered
Color: [None]
Size: [None]
Row names: [None]
Start color: 🟦 End color: 🟦 Color palette: 1.
□ New sheet

Fig. 134 3D-spline dialog box

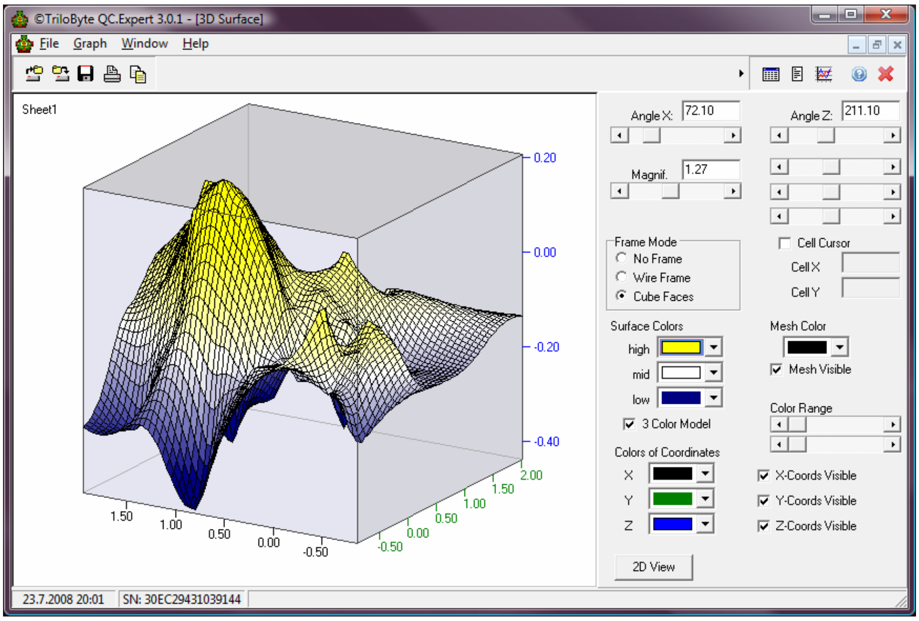


MGSplineFrm

No of grids - X: 50

No of grids - Y: 50

Smoothing: 1



©TriloByte QC.Expert 3.0.1 - [3D Surface]

Angle X: 72.10

Angle Z: 211.10

Magnif.: 1.27

Frame Mode: ○ No Frame ○ Wire Frame ● Cube Faces

□ Cell Cursor

Surface Colors: high 🟨, mid ⬜, low 🟦

☑ 3 Color Model

Mesh Color: ⬛

☑ Mesh Visible

Colors of Coordinates: X ⬛, Y 🟩, Z 🟦

☑ X-Coords Visible

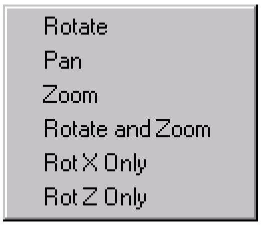
☑ Y-Coords Visible

☑ Z-Coords Visible

23.7.2008 20:01 | SN: 30EC29431039144

Fig. 135 3D-spline for data in 3 columns (X, Y, Z) can characterize response surfaces

Similarily like the previous plot, the 3D-spline can be rotated, zoomed and moved interactively with mouse. The Z-level of the surface can be coloured with 2 or 3 colors. There are more controls on the 3D-spline plot control panel which are used to modify the look of the plot. The button *2D View* is used to display an orthogonal projection into XY plane. For this view it is recommended to uncheck the *Mesh Visible* checkbox and select contrast colors. Right mouse click will display the control menu:



Rotate – Mouse movement will rotate the plot.
Pan – Move the plot with mouse.
Zoom – Mouse movement will zoom the plot.
Rotate and Zoom – Rotate with left mouse button, zoom with right mouse button.
Rotate X – Rotate only around X-axis
Rotate Z – Rotate only around Z-axis

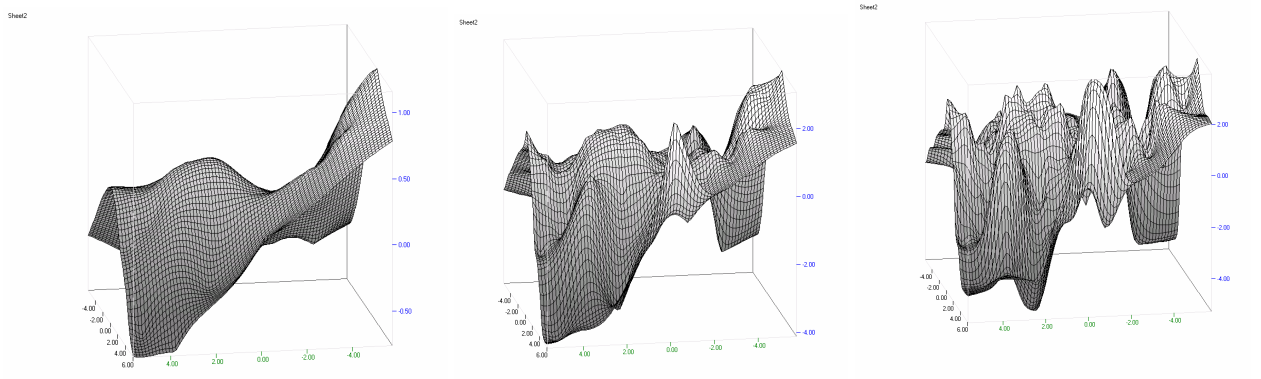


Fig. 136 a, b, c Non-parametric kernel smoothing for the same measured data with smoothing parameter 3, 1 a 0.5

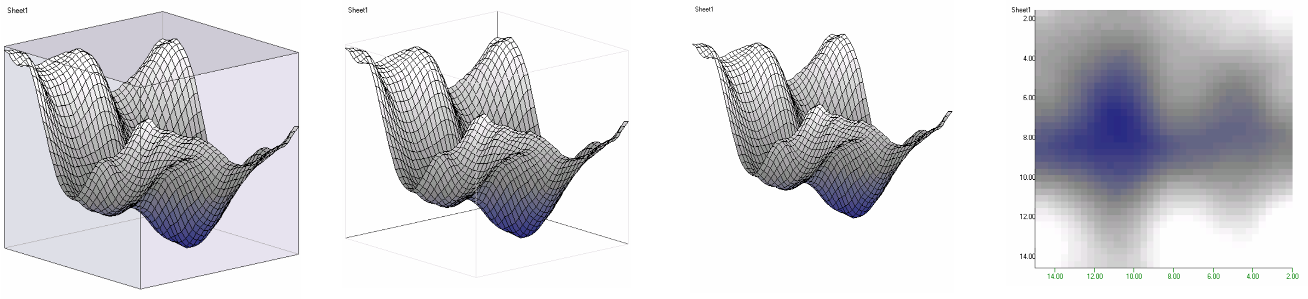


Fig. 137 a,b,c,d Various framing for 3D-plot and a 2D-view

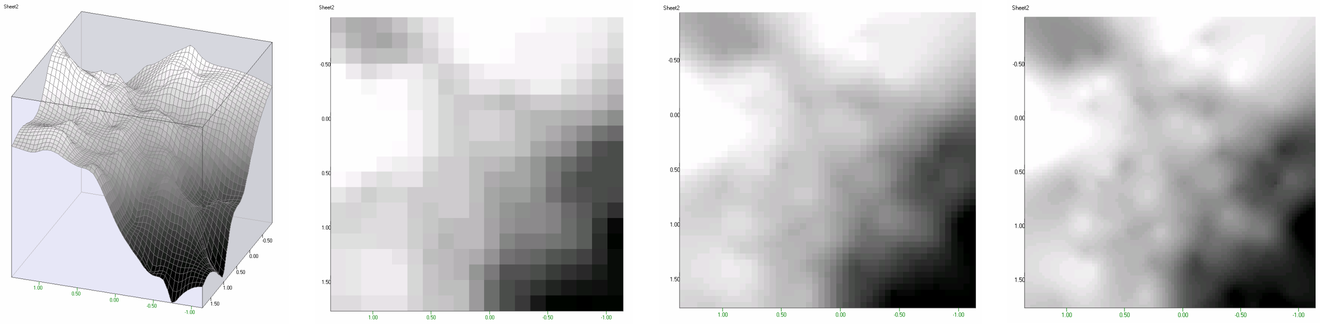
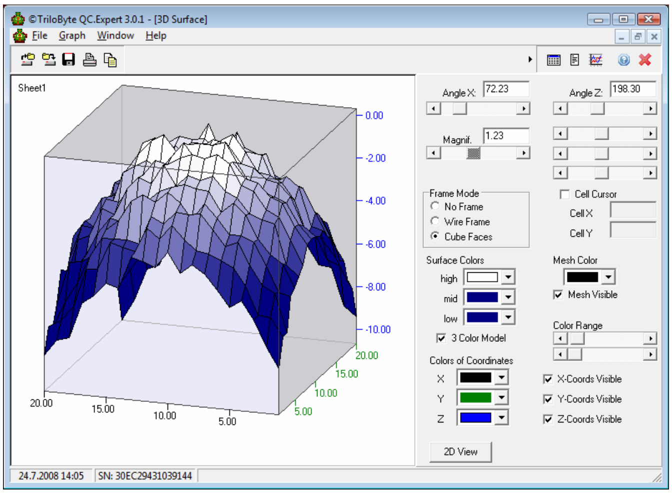


Fig. 138 a,b,c,d 2D-Same data, different grid resolution



Angle X: 72.23

Angle Z: 198.30

Magnif. 1.23

Frame Mode
○ No Frame
○ Wire Frame
● Cube Faces

□ Cell Cursor

Cell X

Cell Y

Surface Colors

high

mid

low

☑ 3 Color Model

Mesh Color

☑ Mesh Visible

Color Range

Colors of Coordinates

X

Y

Z

☑ X-Coords Visible

☑ Y-Coords Visible

☑ Z-Coords Visible

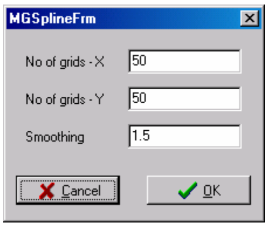
Fig. 139 Spline with smoothing parameter u = 0.01, spline with small u is an interpolation alternative of the 3D-Surface plot.

25.2.15. 3D-kernel probability density for 2 variables

This plot is used to estimate and visualize a non-parametric estimate of probability density for 2 variables. The density function is constructed as a Gaussian kernel density estimate with a user-defined kernel width (or smoothing parameter) and grid resolution.

$$f_s(x,y)=\frac{1}{K}\sum_{i=1}^{n}\exp-\left(\frac{1}{u}\sqrt{\frac{(x_i-x)^2}{s_x}+\frac{(y_i-y)^2}{s_y}}\right),$$

where u is the relative kernel width and s_x and s_y are standard deviation estimates for x and y respectively. Smoothing parameter and grid resolution are defined in the *Options* form. Plot controls are the same as in the previous plot, see paragraph 25.2.14. Examples of use are given below.


MGSplineFrm

No of grids - X: 50

No of grids - Y: 50

Smoothing: 1.5

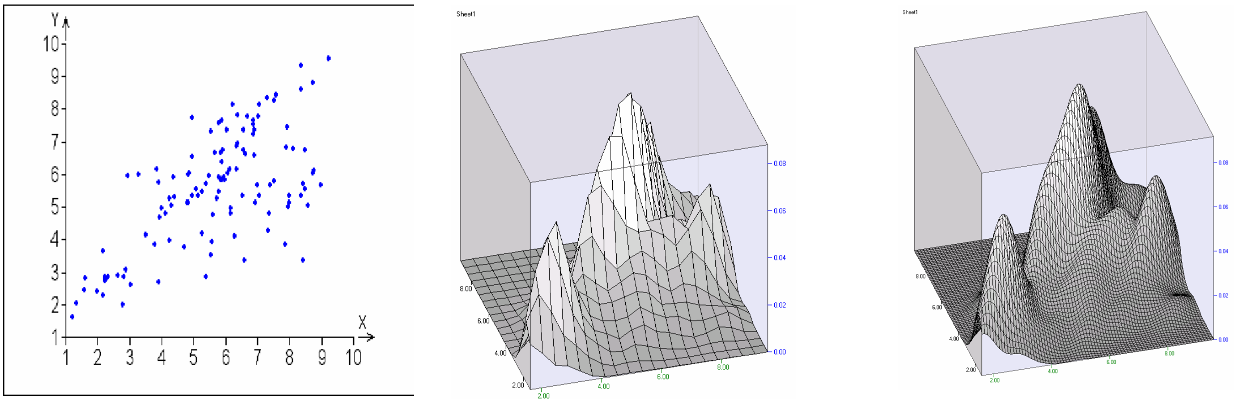


Fig. 140 Data plot and corresponding probability density estimates with 2 different grid resolutions

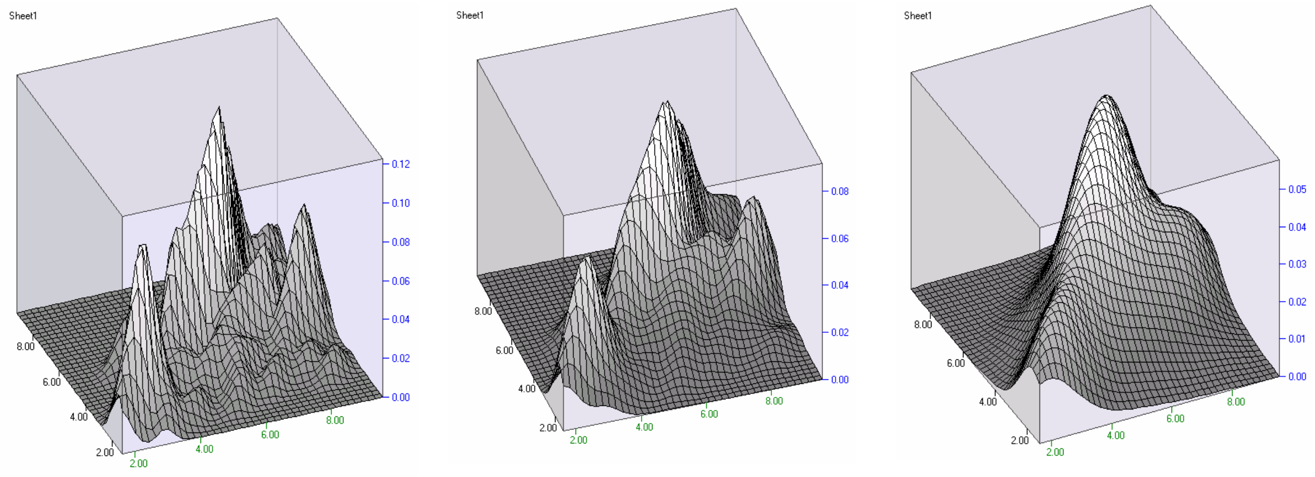


Fig. 141 a, b, c Kernel probability density estimates for the same data with 3 different smoothing parameters

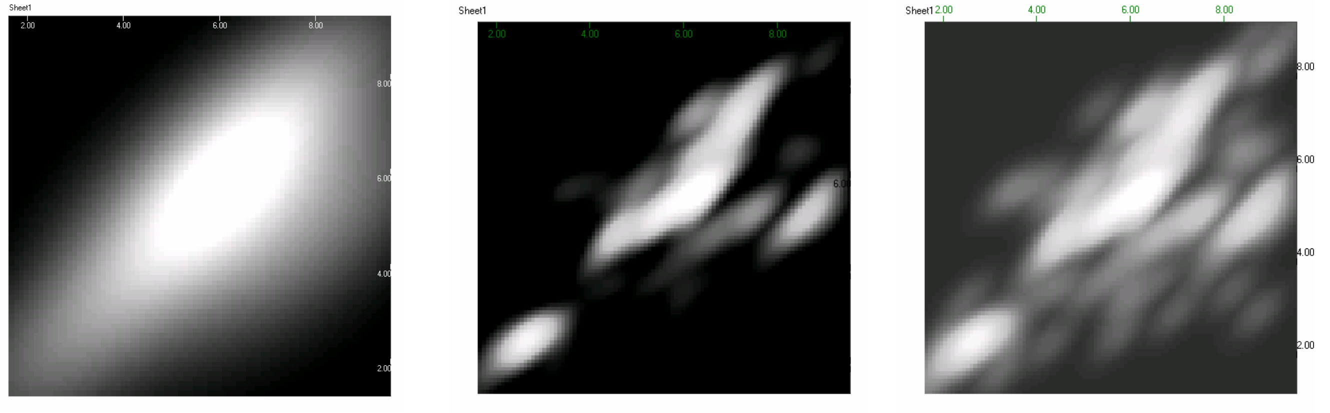
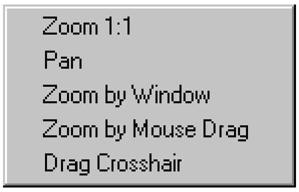
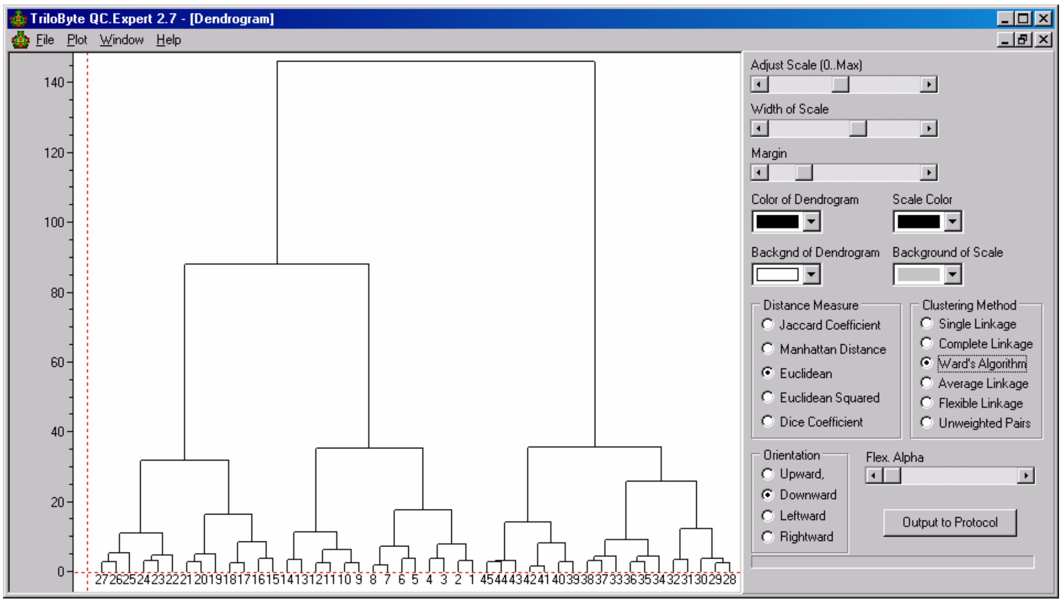


Fig. 142 a,b,c Projection of the previous plots into 2D contour plots with smoothing parameter 2 (a) and 0.7 (b, c)

25.2.16. Cluster analysis - Dendrogram

Dendrogram is a useful tool in cluster analysis. It is a tree-like plot where the multidimensional data represented by rows are connected by lines. The length of path connecting two points roughly defines the dissimilarity or distance between the two points. Thus, we can see if there are any distinct groups of data, we can recognize data that are close to each other. The shape of the dendrogram may be strongly dependent on the selected distance measure and clustering method. This module offers 5 distance measures and 6 clustering methods, one of which (Flexible linkage) is parametrized and can be adjusted continuously. Variables are expected in *m* columns and *n* rows, the data matrix may have more columns than rows. Individual rows are taken as points in *m*-dimensional space. After defining columns and constructing the dendrogram we can use controls in the right part of the plot window to modify the dendrogram. On the right mouse click we get interactive menu:





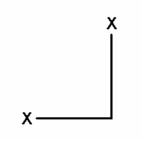
Distance Measure
○ Jaccard Coefficient
○ Manhattan Distance
● Euclidean
○ Euclidean Squared
○ Dice Coefficient

Clustering Method
○ Single Linkage
○ Complete Linkage
● Ward's Algorithm
○ Average Linkage
○ Flexible Linkage
○ Unweighted Pairs

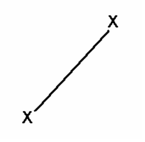
Orientation
○ Upward,
● Downward
○ Leftward
○ Rightward

Fig. 143 Dendrogram control window with parameters and dendrogram tree

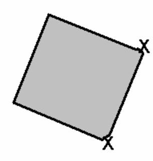
Distance measures definition:



Manhattan distance $$d_{ij}=\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right|$$



Eucleidian distance $$d_{ij}=\sqrt{\sum_{k=1}^{m}\left(x_{ik}-x_{jk}\right)^2}$$

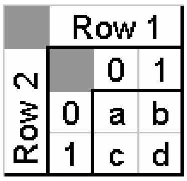


Squared Eucleidian distance $$d_{ij}=\sum_{k=1}^{m}\left(x_{ik}-x_{jk}\right)^2$$

Continuous Jaccard coefficient $$d_{ij}=\frac{\sum_{k=1}^{m}x_{ik}x_{jk}}{\left\|x_i\right\|\cdot\left\|x_j\right\|-\sum_{k=1}^{m}x_{ik}x_{jk}},$$

where $||x||$ is the quadratic norm of x.

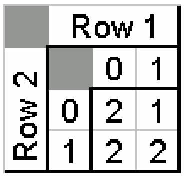
For binary data (eg. ones and zeroes) four similarity parameters *a*, *b*, *c*, *d* are computed: *a* = number of cases where in both rows to compare are zeroes in the same column, *b* = Number of cases with 1 in the first row and 0 in the second row, *c* = Number of cases with 0 in the first row and 1 in the second row, *d* = number of cases where in both rows to compare are ones. Binary data example is given below.



		Row 1	
Row 2		0	1
	0	a	b
	1	c	d

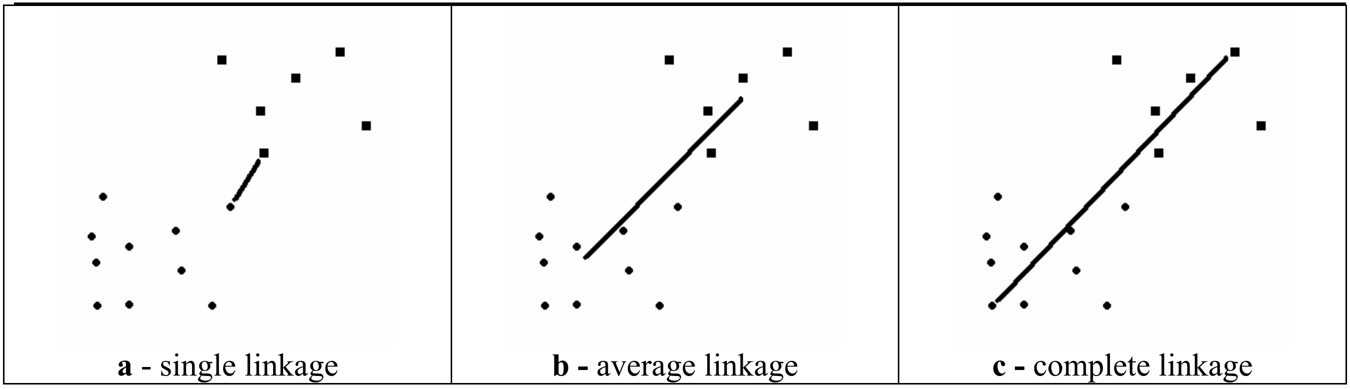
0	0	0	1	0	1	1
0	**0**	**1**	**1**	**0**	**1**	**1**
0	1	1	1	1	1	1
0	0	0	1	0	1	0
0	**1**	**1**	**0**	**0**	**1**	**0**
0	1	0	0	0	0	1
1	1	0	0	1	0	0

For the given example, when we compare the second and fifth row, we get: *a*=2, *b*=1, *c*=2, *d*=2. Dissimilarity coefficient d_{25} is then computed using Jaccard or dice coefficient:



		Row 1	
Row 2		0	1
	0	2	1
	1	2	2

Jaccard coeficient $d_{ij}=\dfrac{a}{a+b+c}$ and dice-coeficient $d_{ij}=\dfrac{2a}{2a+b+c}$.



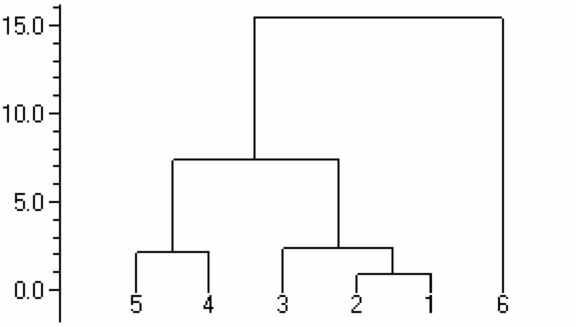
a - single linkage

b - average linkage

c - complete linkage

Fig. 144 a, b, c Three methods to build clusters

3-dimensional data example				
Object no.	p1	p2	p3	
1	2	2	2	
2	2	3	2	
3	4	1	3	
4	7	7	7	
5	8	7	5	
6	12	16	-7	



With the button *Export to Protocol* we get the following table in the Protocol window which is read as follows: Objects 1 and 2 with distance 1 form cluster 7; objects 4 and 5 have distance 2.23 and form cluster 8. Cluster 7 and object 3 form cluster 9, distance between 3 and 7 is 2.45, etc. The distances are plotted on the vertical axis.

Obj.1	Obj.2	New cluster	Distance
1	2	7	1
4	5	8	2.2361
7	3	9	2.4495
9	8	10	7.4833
10	6	11	15.5242

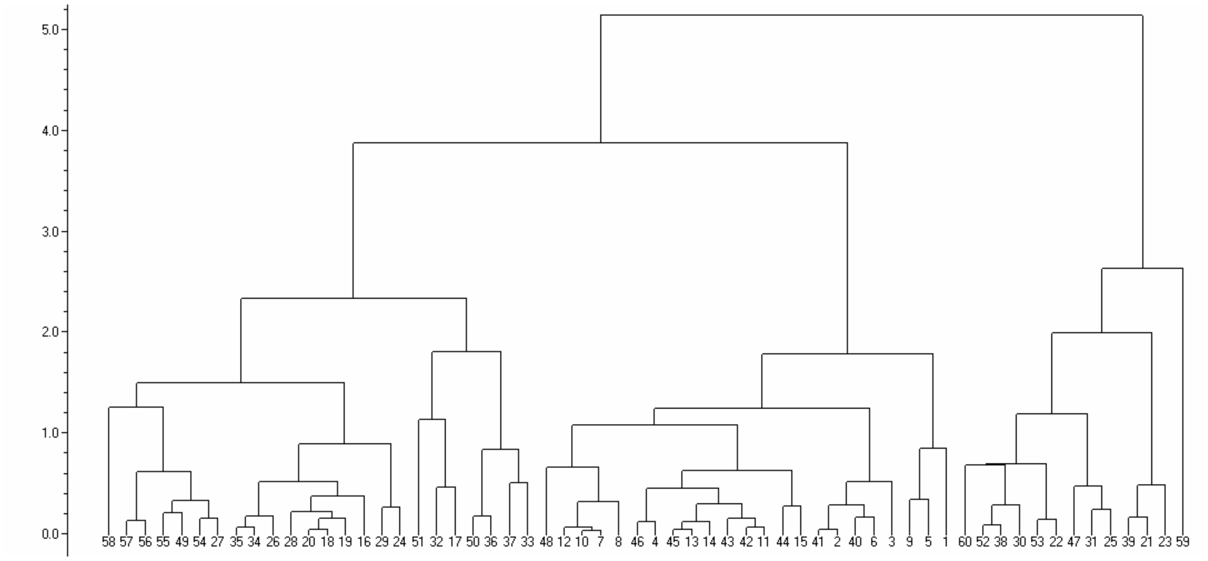
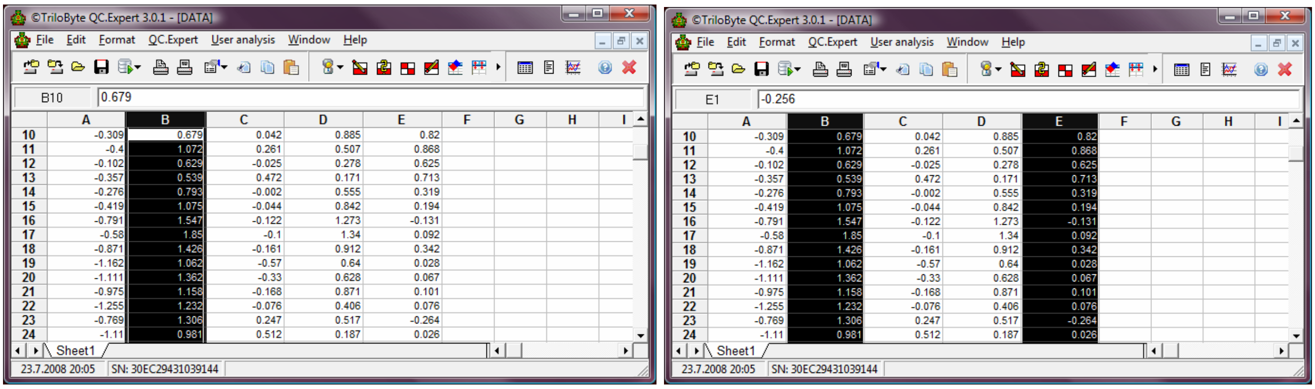


Fig. 145 Example dendrogram for 6 variables on 60 samples of cement, 3 dominant clusters are detected, containing 24, 23 a 13 points(or cement samples)

26. Quick data plot

Menu:	QCExpert	Data plot

Data plot function is used for quick look at univariate (one column) or bivariate (two columns) data. Unlike in othe modules, here the data to be plotted must be selected. To select two non-adjacent columns use Ctrl-mouse click. To select whole column click the mouse on the column header. Having selected the data to be plotted use *Data plot* from menu or click the *Data plot* button in the main toolbar. In case of one selected column, the data will be plotted against index, in case of two selected columns the data will be plotted as the X-Y scatter plot. In the plot, data can be marked by mouse as in any other plot.



B10 0.679

	A	B	C	D	E	F	G	H	I
10	-0.309	0.679	0.042	0.885	0.82				
11	-0.4	1.072	0.261	0.507	0.868				
12	-0.102	0.629	-0.025	0.278	0.625				
13	-0.357	0.539	0.472	0.171	0.713				
14	-0.276	0.793	-0.002	0.555	0.319				
15	-0.419	1.075	-0.044	0.842	0.194				
16	-0.791	1.547	-0.122	1.273	-0.131				
17	-0.58	1.85	-0.1	1.34	0.092				
18	-0.871	1.426	-0.161	0.912	0.342				
19	-1.162	1.062	-0.57	0.64	0.028				
20	-1.111	1.362	-0.33	0.628	0.067				
21	-0.975	1.158	-0.168	0.871	0.101				
22	-1.255	1.232	-0.076	0.406	0.076				
23	-0.769	1.306	0.247	0.517	-0.264				
24	-1.11	0.981	0.512	0.187	0.026				

E1 -0.256

	A	B	C	D	E	F	G	H	I
10	-0.309	0.679	0.042	0.885	0.82				
11	-0.4	1.072	0.261	0.507	0.868				
12	-0.102	0.629	-0.025	0.278	0.625				
13	-0.357	0.539	0.472	0.171	0.713				
14	-0.276	0.793	-0.002	0.555	0.319				
15	-0.419	1.075	-0.044	0.842	0.194				
16	-0.791	1.547	-0.122	1.273	-0.131				
17	-0.58	1.85	-0.1	1.34	0.092				
18	-0.871	1.426	-0.161	0.912	0.342				
19	-1.162	1.062	-0.57	0.64	0.028				
20	-1.111	1.362	-0.33	0.628	0.067				
21	-0.975	1.158	-0.168	0.871	0.101				
22	-1.255	1.232	-0.076	0.406	0.076				
23	-0.769	1.306	0.247	0.517	-0.264				
24	-1.11	0.981	0.512	0.187	0.026				

Fig. 146 One selected column (a) and two columns (b) selected by Ctrl-mouse click.

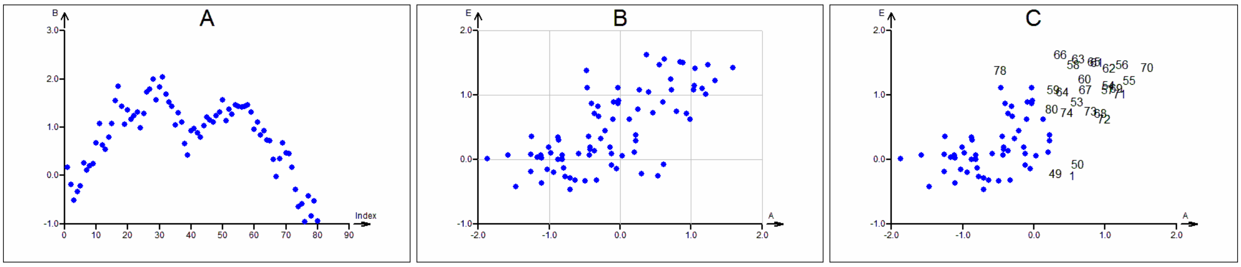


Fig. 147 Plotted data can be identified by mouse (fig C)

27. Appendices

27.1. List of figures

27.2. Suggested reading

(listed in alphabetical order)

1. Draper, R.D., Smith, H.: Applied Regression Analysis, John Wiley&Sons, 1998
2. Grant, E. L., Leavenworth, R.S.: Statistical Quality Control, McGraw Hill, 1996
3. Kotz, S.: Process Capability Indices, Chapman&Hall, 1993
4. Krzanowski, W.J.: Principles of Multivariate Analysis, Oxford University Press, 1993
5. Mittag, H. J.; Rinne, H.: Statistical Methods of Quality Assurance, Chapnam&Hall, 1993
6. Montgomery, D. C.: Introduction to Statistical Quality Control, Chapnam&Hall, 1990
7. Myers, R. H.; Montgomery, D. C.: Response Surface Methodology, Wiley, 1995
8. Ryan, T. P.: Statistical Methods for Quality Improvement, Wiley, 1986
9. Shewhart, W. A.: Economic Control of Quality of Manufactured Product. D. Van Nostrand, New York, 1931
10. Thompson, J. R.; Koronacki, J.: Statistical Process Control for Quality Improvement, Chapman&Hall, 1993

28. Index