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 is appropriate in such cases.

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.

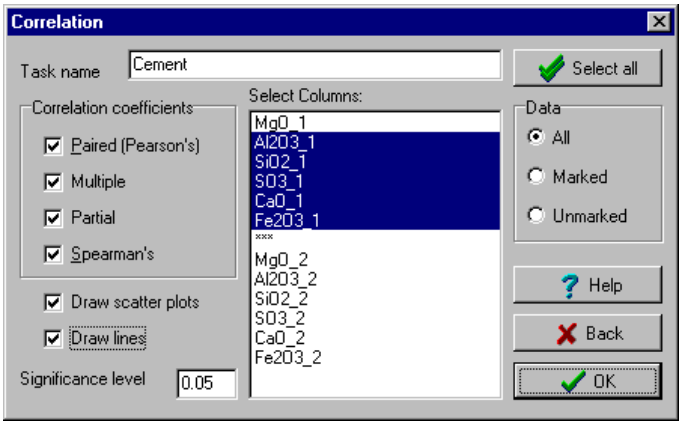
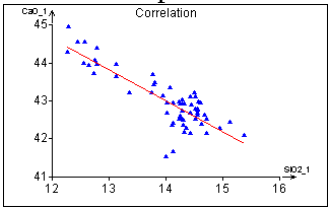


Fig. 1 Correlation dialog panel, columns A,B,C,D selected

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 pairwise 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.

Graphs

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.