Electronic supplementary materials *Mendeleev Commun.*, 2023, **33**, 419–421

An efficient method of searching for correlations between unlimited datasets to provide forecasting models

Alexander Yu. Tolbin

* * *

The numerical algorithm described in the article is implemented in the form of the CORRELATO program [1], which is available to readers. This is a demo version, which has significant limitations compared to the Full version (see Table S1).

The input file for the calculation is an ASCII table, the columns in which are separated by a tab character. The first line is headings, which are recommended to use Latin letters. At the intersection of rows and columns are the properties of substances (activity, equilibrium constant, wavelength, etc.), i.e. the array of considered substances is associated with rows, and the columns with properties.

Table S1. Information about versions of CORRELATO program.

	Demo	**Full**
Distribution	Binary executable for 64-bit Windows (ZIP archive)	PHP source code for Linux[1]
Suggestion	Free DOWNLOAD	Request[2]
Maximum number of columns	2	Unlimited
Maximum number of rows	10	Unlimited
Maximum number of iterations	50	Unlimited
An array of numbers for raising components to a power[3]	–3 to +3 in steps of 1, including 0	Any
Option to exclude a set of elements from the analysis	No	Yes
Special analysis of "outliers"	Simple	Statistical
Open MPI interface support	No	Yes[4]

[1] Requires CLI PHP interpreter to be installed; the source code can be also embedded in web applications;

[2] Contact the developer - Prof. Alexander Yu. Tolbin: tolbin@ipac.ac.ru;

[3] See Eqn. 2;

[4] It is possible to run one or multiple processes per node; MPI interface is required.

The Demo version of CORRELATO is downloaded as a ZIP archive which contains a 64-bit Windows executable correlatto_win.exe, configuration file correlatto_win.ini, and example input file input.txt. The input file for calculations is an ASCII table, with the columns separated by a TAB character. The first row is the column headings. Before starting the program, the user should configure the settings (see Table S2).

Table S2. Configuration parameters of CORRELATO program (correlatto_win.ini file).

Parameter	Sample values	Description
Y_cols	1,3	Column numbers (counting from 1) to form X or Y assets. In the demo version, user can use no more than two columns for preparing both X and Y assets (comma delimiter without spaces).
X_cols	2,4	
pearson_threshold	0.6	The threshold for Pearson coefficient. Below this value, results are skipped.
iter_count	50	The maximum number of iterations. For the demo version, no more than 50 iterations are allowed.

Now run correlatto_win.exe. A command window is opening to proceed with calculations. File plot.txt will appear to contain the results of combinations of the selected columns (a series of XY ASCII tables). Below is an example of the calculation (plot.txt):

```
Iteration # 10, Pearson: 0.9233, bad component: 2

A^(-2)*B^3 vs. log(G^3)
0.41016170146922    1.0813148788927E-9
0.40061672511065    2.08333333333333E-10
0.41963725920371    5.1939058171745E-10
0.42904440076229    8.4526008002462E-5
0.45686503314917    0.000489795918367 35

Iteration # 11, Pearson: 0.9752, bad component: 4

A^2*B vs. F*G^(-2)
1.1455058873675     5780
1.048875432526      3888
0.95568157950011    1559.52
0.93162879768128    2111.85
0.90259869073597    264.6
```

When logarithmic scale is considered, a decimal logarithm is used. The chart for Iteration #11 would look like presented in Fig. S1.

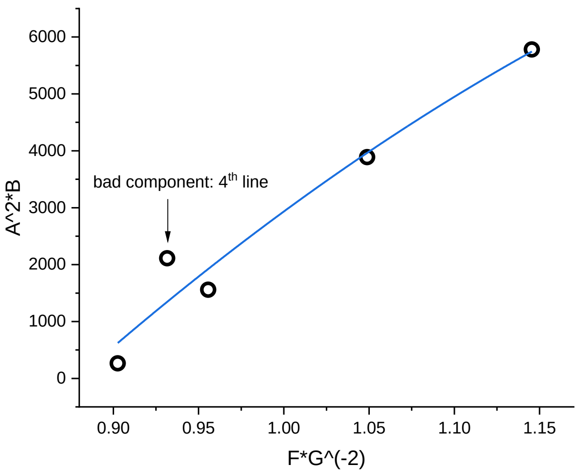

Figure S1. An example of a correlation when looking for relationships between the data presented in the input file (input.txt). Captions of X and Y axes show the analytical form of random combinations between the properties of substances.

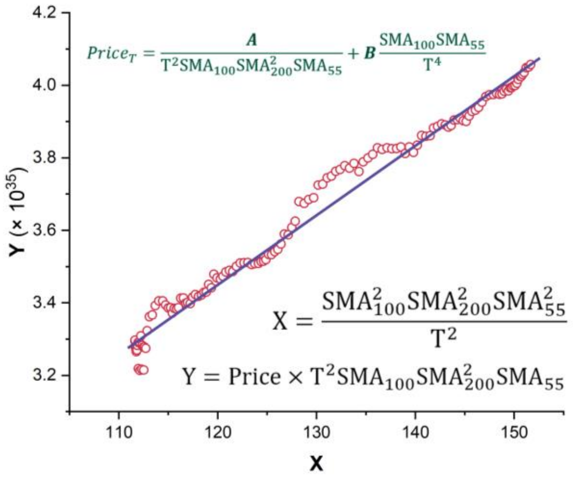

Figure S2. An example of the correlation of SMA values with the stock price; *A* and *B* are the fitting coefficients.

Bad components are the rows that can be excluded to improve the relationship. The full version of the CORRELATO program is available for Linux OS (Table S1), has no restrictions, and is periodically improved.

Another example demonstrates the ability to predict stock prices after several *T* periods based on the simple moving averages (SMA), which can be predicted with high accuracy in the short term (Fig. S2).

A very important detail should be noted. Thus, correlations provided for functional analysis (high *r*-Pearson) relevant only for the data for which they were established. Using other arrays may give incorrect results. However, if we consider zonal parametric analysis (low *r*-Pearson), there is the possibility to select elements according to their characteristics, including new data that did not participate in the correlation analysis.

REFERENCE

1. A.Y. Tolbin, *Establishing correlations between unlimited datasets - Correlato, Certificate of state registration of computer program No 2022613888 (RU).* 2022.