CONVERTING BIG DATA TO BIG ANALYSIS AND CONTROL – THE SECRET IS TO PRE-SCREEN

Pre-Screen is a user-friendly software package from researchers at the Centre for Process Analytics and Control Technology (CPACT) which has been specifically developed to make the analysis of large data sets as fast and visual as possible. Many commercial data analysis packages do not fully address the initial data cleaning and data conditioning tasks which can consume up to 80% of the time required to develop useful models. The unique software toolkit has been developed specifically to improve model quality and reduce the time spent pre-screening large industrial data sets but also has many other applications in data analysis. It is made freely available to all members of the Centre for Process Analytics and Control Technology (CPACT).

Introduction

In a previous article we discussed how the time was ripe for the petroleum-based industries to embrace the concept of "Big Analysis" using the data they already possess and continue to generate in ever larger quantities (1). A large amount of data currently exists inside the chemical process-based industries and more data is being created daily. Harnessing that data and providing new ways to utilize it for ever-increasing complex analytics could provide companies with unprecedented capabilities. Petro-chemical companies are increasingly understanding that their data is an extremely valuable asset since the ability to control and optimize processes, cut costs, drive new innovations, virtualise experiments, create and answer complex business questions relies on the efficient and timely analysis of that data.

However, in practice many in industry struggle with the practicalities of this challenge as commercial data analysis packages are often complex to use and do not fully address the initial data cleaning and data conditioning tasks which can consume up to 80% of the modelling time (2). This was widely recognised by the members at CPACT, the leading industry/academic network for research on all areas of process performance monitoring and control. This realisation led to a work programme to develop a new toolbox to make the analysis of large data sets as fast and visual as possible while also accessible for process and control engineers, analytical scientists and academic researchers. In this short article we will cover some of the main features of the toolbox, but a detailed description of the wider capabilities of Pre-screen has been published elsewhere (3).

Pre-Screen has been compiled in MATLAB 2012a and MATLAB 2015a and has been tested on both Windows 7 and Windows 10 systems and makes use of the appropriate Mathworks MCR package which is free to download. Pre-screen is routinely used on data sets with typically 150 variables and 7,000 samples and has recently been applied to a data set with 125,000 samples. The software has been rigorously tested and evaluated by BP Chemicals Ltd., Saltend Lane, Hull and other industrial and academic members of CPACT.

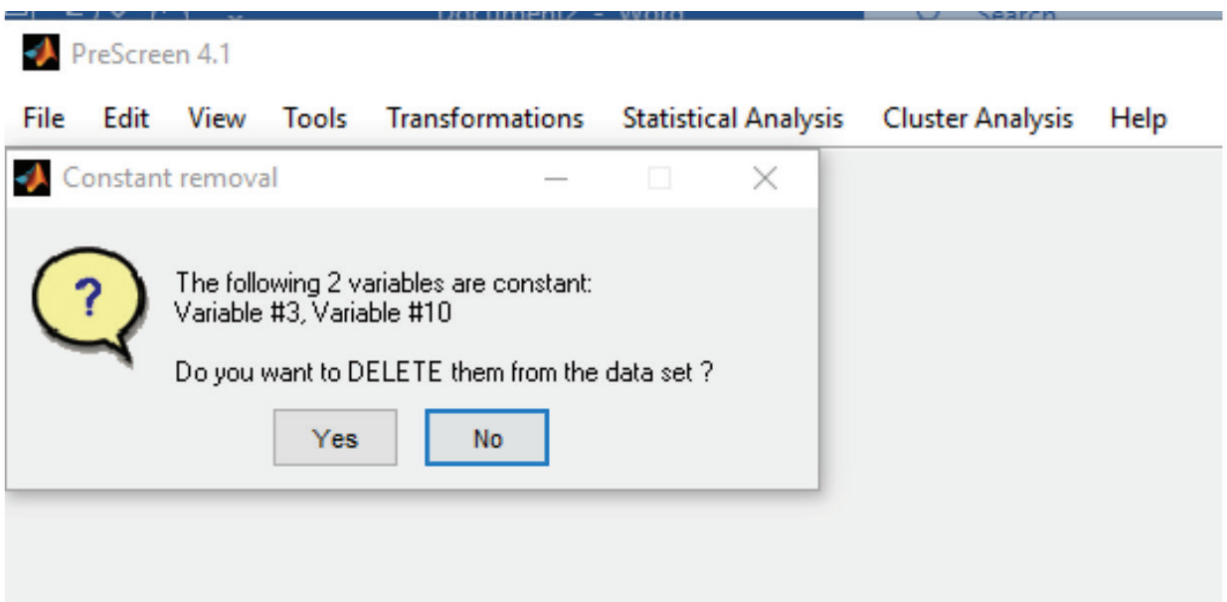


Fig 1. Screenshot of constant variable detection flagged on loading the data file.

Pre-Screen:- the Highlights

The interface is highly visual, interactive and easy to use and the main features can be summarised as follows: -

Ease of Loading Data sets:

Pre-screen can be applied to a wide range of data sets and can

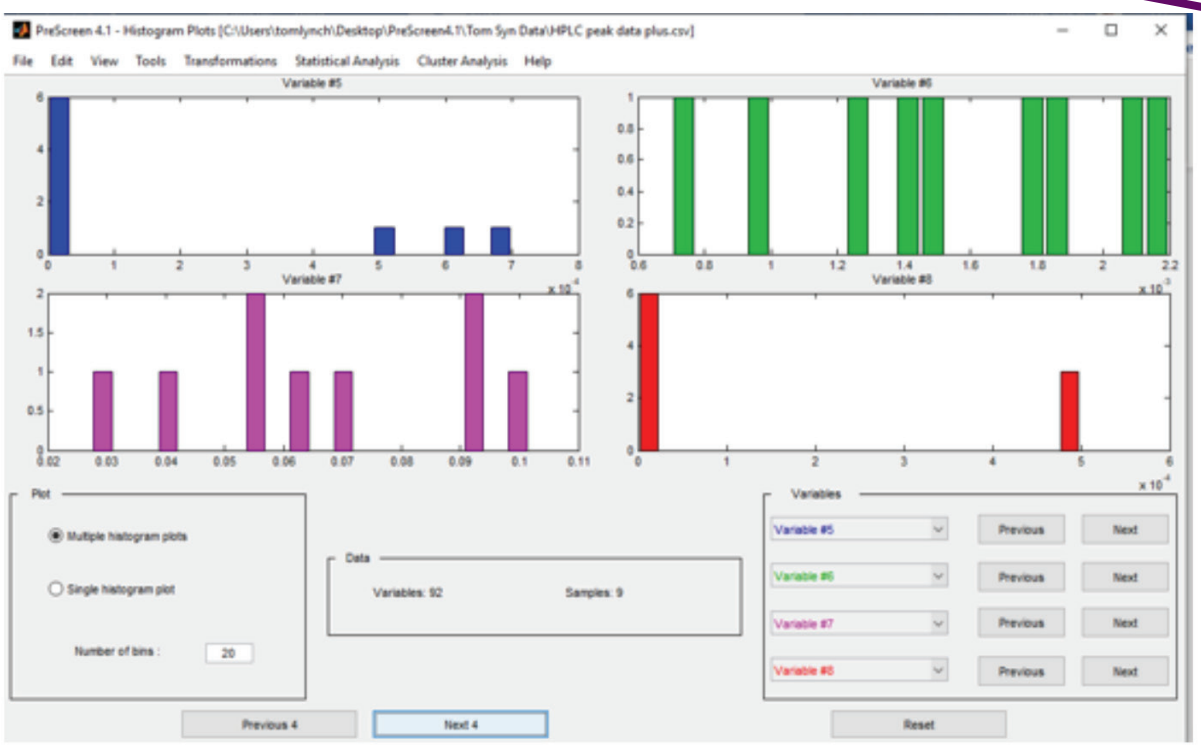


Figure 2.Histogram plots allowing visual screening of the variable data and its values and spread across the sample set.

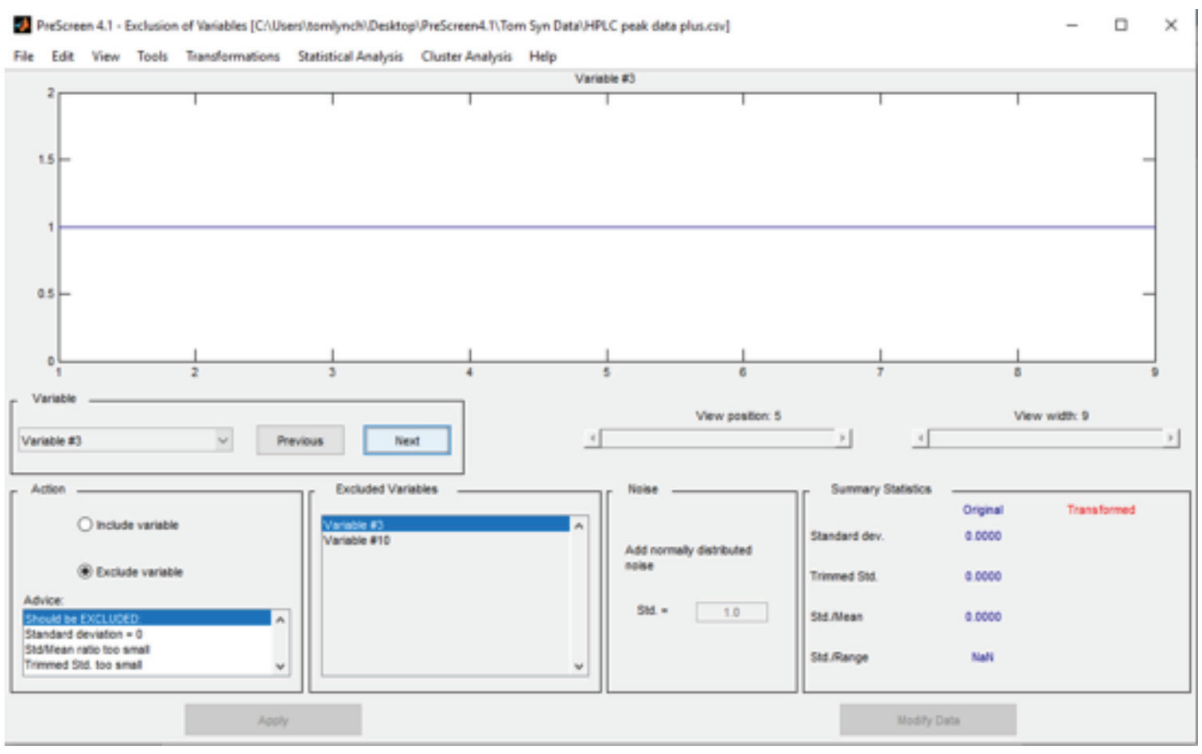


Figure 3 Screenshot of the Exclusion of Data tool applied to variable 3 showing that the added constant data should be excluded from the data set.

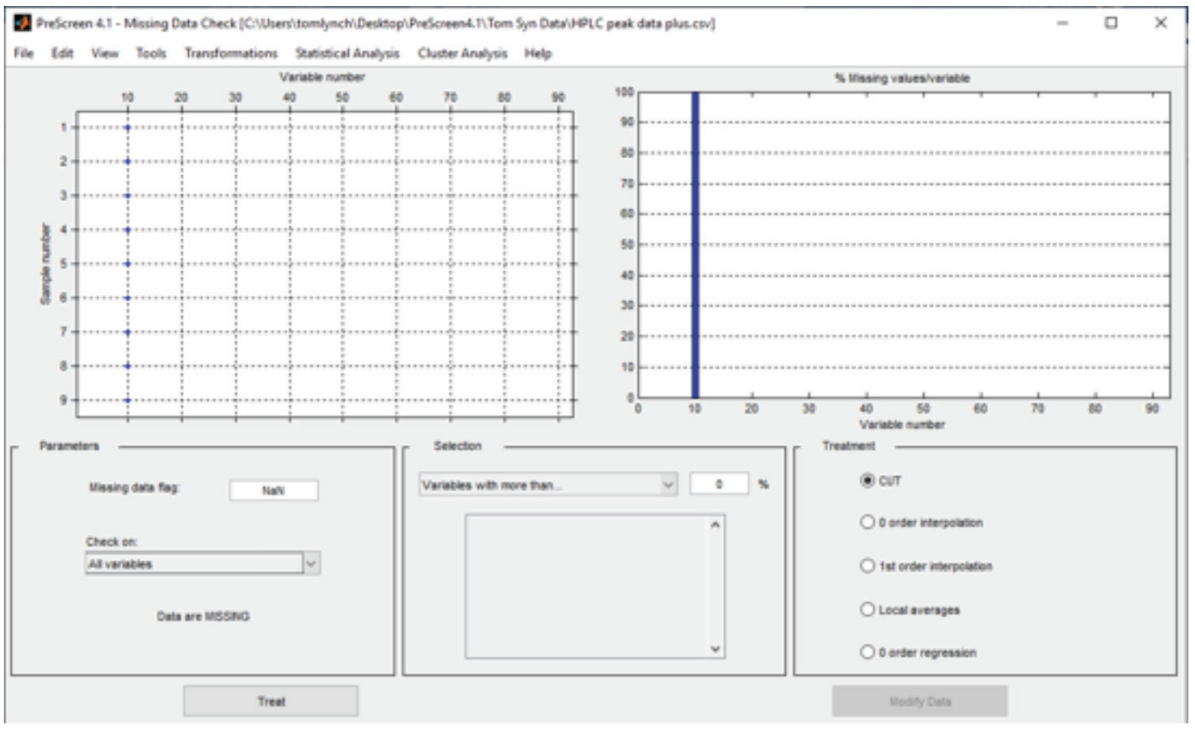


Figure 4: Missing Data Tool screenshot identifying variable 10 had no numeric data.

take data in a number of formats such as a MAT file (using the .MAT extension), a CSV file (using .CSV extension), an ASCII file (using a .txt or any other extension) or directly from commercial process information systems such as OSI PI with data imported from OSI PI using the PI Datalink Excel add-on. Data tags are imported as text files. All data operations and variable history are automatically recorded and saved in chronological order.

Missing Data:

The automatic detection of missing values in a data set requires the definition of format standards in the provision of the raw data. Usually a numerical flag is selected where: (i) the value of the flag must be outside the process measurement range, to ensure that normal measurements are not identified as missing observations (for example, the number 0 is not an appropriate choice for the flag since it is often an acceptable value for a number of measurements); (ii) attention needs to be paid to the data type used in the data pre-processing software. If data files are in binary format and the data set is read as a matrix of floating-point numbers, an integer flag could be misinterpreted, or the value could be changed because of rounding errors. In Pre-Screen, an automatic check is carried out on the full data set to assess in which variables missing data has occurred.

Data visualisation:

One of the major drivers for Pre-Screen was to have a highly visual user-friendly graphical user interface (GUI) which can be used in an intuitive way and can be used without the need for MATLAB. The toolbox is very user-friendly having been developed with industrial colleagues and is highly interactive and straightforward to use. It provides a highly visual approach to data pre-screening (messy data cleaning) and pre-processing and exploits the potential of Principal Component Analysis (PCA) based multivariate process performance monitoring / multivariate statistical process control (MSPC). Pre-Screen focuses on using PCA rather than Partial Least Squares (PLS) since PCA allows the user to select those process measurements and information from software (virtual) sensors and spreadsheet computations as additional process performance information. PLS requires the selection of the relevant variables related to the prediction of process outputs and in this respect is more restrictive than a PCA approach. The software includes unique data cleaning operations; data plotting in terms of time series plots; normality plots (univariate and multivariate); summary statistics (mean, standard deviations covariance, correlations, skewness and kurtosis); semi-automatic missing data analysis and rectification; outlier data identification and removal; data transformations, data filtering; cross correlation analysis; data transformations (mathematical and time shifting); scatter plots to observe possible relationships; loadings and contribution plots; histogram plots; normal probability plots, parallel coordinate plots and plot copying to word files. Active multiscreen visualisation and working allow simultaneous multivariate analysis plots, time series plots, scatter plots, normal probability plots and correlation plots – all observable simultaneously for enhanced process understanding and fault diagnostics.

Exclusion of Variables:

Tools for the selection of significant variables are the statistical properties, such as mean, median, standard deviation, range and trimmed standard deviation, and the inspection of the time series plots. These enable the user to distinguish between variability caused by process noise and variability resulting from real process changes. Exclusion of variables focuses on those variables with low 'trimmed standard deviation'. In Pre-Screen, the trimmed standard deviation is calculated on 95% of the data. The test criteria can be changed in the Exclusion Tests frame, according to the user's knowledge of the process. A series of tests are carried out: Test (i) is compulsory if the variable needs to be scaled prior to modelling and investigates whether the standard deviation is small when compared with the range detectable by the measuring device; Test (ii) checks if the standard deviation is small compared with a predetermined range; Test (iii) checks if the trimmed standard deviation is non-zero; Test (iv) checks for outlying observations in the variable. All the tests are optional, apart from Test (i) which is compulsory if the variables must be scaled prior to modelling. The number of tests to consider and the decision whether to accept the suggestion to remove the variable or not is ultimately based on user knowledge of the process. Variables can be included or excluded. If it is desired to include a variable with a standard deviation which did not pass the tests; random normal noise can be added to the variable by setting its standard deviation and modifying the data by the addition of noise. An example is a control valve position, a measurement that can provide important process understanding and knowledge.

A Simple Example using HPLC Data

Pre-Screen has such a range of capabilities it can be applied to a wide range of industrial data analysis and we could not cover it all in an article such as this. So here is a simple example of applying it to HPLC analysis of 9 samples that were claimed to be from the same multi

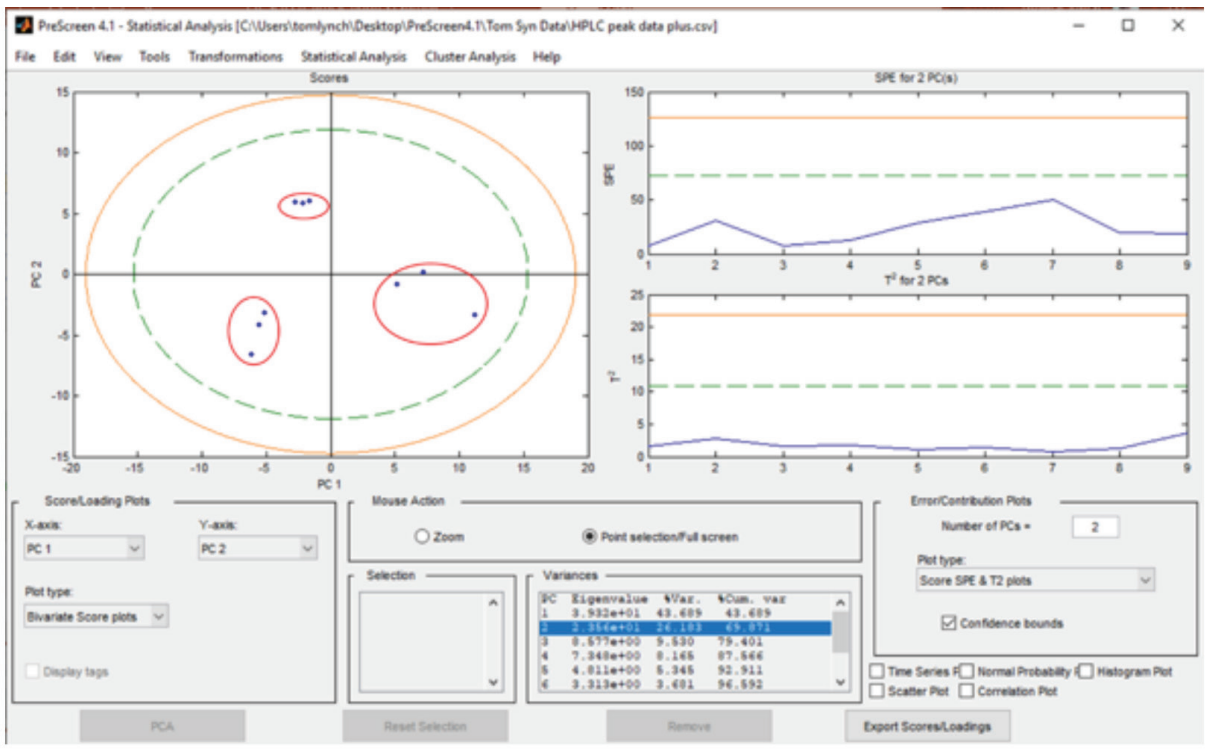


Figure 5. Multivariate PCA analysis results screenshot for HPLC data

component chemical product. However, these products exhibited a range of different performance characteristics in use and the question was why? The Pre-Screen analysis was carried out by one of the authors who had never used Pre-Screen or had any training. The actual data analysis including formatting the data file from Excel took under an hour and provided a unique insight into the samples origin and quality.

The HPLC analysis detected 90 unique peak retention times across the 9 samples with no clear patterns emerging by visual comparison of the chromatograms. A .csv file of the peak area data for all samples was edited to contain additional data including a variable with constant value 1 (as variable 3) and a variable with random text (as variable 10) to test some of the features of Pre-Screen. The data file was loaded and the additional data was immediately identified by Pre-Screen (Figure 1) as being constant with an option to remove those variables.

To illustrate further features of the program the suspect data was not deleted at this stage. The data was then reviewed visually using histogram plots selected from the View menu. A screenshot of histogram plots for variables 5,6,7 and 8 is shown in Figure 2. The x-axis of each histogram shows a variable (peak area) value with the y-axis showing how many of the 9 samples that had that value.

So in Figure 2 variable 8 had a value of zero (no peak detected) in 6 samples and a peak area of 5 $x10^{-4}$ in the remaining 3 samples whereas peak 6 is present in every sample but has a different peak area in each sample.

The data was then examined using the Exclusion of Data and Check on Missing Data Tools. The screenshot in Figure 3 shows the check on Variable 3 (added constant value data) and clearly shows the value was constant at 1 and should be excluded. It also identifies Variable 10 should be excluded.

The screenshot of the Missing Data Tool is shown in Figure 4 and it clearly identifies that Variable 10 (the added text data) has no numerical data and could be cut from the data set.

The data was then processed using a Multivariate PCA from the Statistical Analysis tools which produced the output shown in Figure 5.

The plot of PC1 versus PC2 shows the data appears to split into 3 main groups with 3 samples in each which have been highlighted in red in Fig 5. On further discussion with the sample originators it was discovered that these 3 groups corresponded to different manufacturers and also correlated with performance differences in use. Furthermore, in one of the groups the points are tightly clustered which indicates they are very similar in terms of composition whereas the other 2 groups seem to have a greater variation in their composition between samples. This could be very important information in terms of batch to batch product quality and the tool could be used to monitor the quality of future batches based on adding new analysis data to the historical data set.

So, by a very quick analysis of the HPLC data set, valuable information could be obtained in understanding the compositional variation in these supposedly identical products from different manufacturers which could be correlated with their performance. It would have been very difficult, if not impossible to achieve this by conventional comparison of chromatograms and although this is a simple example compared to many of the other applications for Pre-Screen it does serve to illustrate the power and ease of use of the application for data screening and analysis.

Application to MSPC:

Statistical Process Control (SPC) concepts and methods and in particular Multivariate Statistical Process Control (MSPC), or sometimes termed Multivariate Statistical Process Performance Monitoring, have become very important in the process industries. The objective is to monitor the performance of a process over time to verify that the process is remaining in a "state of statistical control". The concepts and methods of MSPC are complementary to those of feedback and advanced process control. MSPC monitoring methods are applied above the process and its automatic control system in order to detect process behaviour that indicates the occurrence of a special event or 'fault'. By diagnosing causes of the special events and removing them (rather than simply continuing to compensate for them), the process is improved. Multivariate Statistical Process Control (MSPC) schemes focus on monitoring the stability of the process mean and are based on the statistical metrics of the Squared Prediction Error (SPE) and Hotelling's T2 statistic based on the process variables.

To Conclude

Pre-Screen is a unique, user-friendly software package from research partners at CPACT which has been specifically developed to make the analysis of large data sets as fast and visual as possible. It has been developed through collaboration between experts from academia and industry users to overcome many of the issues encountered in real industrial applications. The unique software toolkit has been developed specifically with the aim to improve model quality and reduce the time spent pre-screening large industrial data sets but it can also be applied across a large range of numerical data analysis scenarios.

Pre-Screen is also one a suite of related data analysis and model building software packages and toolboxes developed by the CPACT consortium to meet the needs of its industrial members; the most important packages are MultiDAT, DoEMan and Spectral Shooter and further details of these can be found on the CPACT website (www.CPACT.com).

References

(1) Tom Lynch, Eric Little, Big Data, Smart Data and Big Analysis, Petro Industry News Sept 2018

(2) Villalba Pedro, Javier Sanchis, Alberto Ferrer, A graphical user interface for PCA Based MSPC: A benchmark software for multivariate statistical process control in MATLAB, Chemometrics and Intelligent Laboratory Systems, 2019, 185, 135–152.

(3) Gang Yi, Craig Herdsman and Julian Morris, A MATLAB toolbox for data pre-processing and multivariate statistical process control, Chemometrics Intell. Lab. Syst., 2019, 194.

Author Contact Details

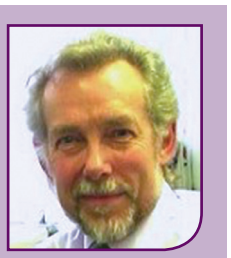
Tom Lynch[1], Tom Lynch Analytical Consultancy • Cricket House, High St, Compton, Newbury, RG20 6NY

• Email: **tomlynch.lynch@ btinternet.com**

Julian Morris[2], • 2 Centre for Process Analytics and Control Technology (CPACT), c/o Department Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL

[1]To whom all correspondence should be addressed


Tom Lynch


Julian Morris