MS2 Data Mining & Visualisation Project Update

AJM Consulting

Issue 2

October 2006

Number of MS2 Process Analytics Users Continues to Grow

We now have 12 companies on board as development partners for our MS2 process analytics system. Discussions are well advanced for at least six more to join us over the next few weeks.

Representing a range of company sizes and markets, these companies are involved in both organic synthesis and inorganic processes, pharmaceuticals, surface coatings, petrochemicals, oil and agrochemicals. Many are multinational.

Our concept of inviting these companies to become development partners, rather than simply supply the software once fully developed, has been remarkably successful. Without doubt, the MS2 system is more advanced, more powerful and more flexible because of this approach. We have learned much about the differing needs of different users, particularly with regard to the import of data prior to its analysis.

Another aspect of MS2's development which has been heavily influenced by the diversity of needs is the Data Pre-Screening functionality (see page 5). Although still at an early stage in its development and with only a few functions in place, it is already making the practical application of the system far simpler and speedier.

For some of the larger data sets in particular, we have also found the new Projection to Standard Model facility (page 3) to be very powerful, identifying true multi-variate covariance with a clarity not previously possible.

As the earlier projects mature, we are delighted that most companies are taking maintenance contracts, which enable them to take advantage of the new features as they are released.

Professor Julian Morris accepts Associate Director Role with AJM Consulting

We are pleased to announce that Professor Julian Morris FREng FIChemE FInstMC CEng has agreed to become a part-time Associate Director of the company.

Prof. Morris's role will be to guide further development of the MS2 multivariate analysis and process data mining system and to develop and participate in its industrial application.

He is currently Professor of Process Control in the School of Chemical Engineering and Advanced Materials at Newcastle University and co-Director of the Centre for Process Analytics and Control Technology (CPACT). His involvement with these organizations will continue. He is an acknowledged leader in the development of multivariate process performance monitoring technologies.

Inside this issue:

CPACT

We are particularly pleased to be working closely with the Centre for Process Analytics and Control Technology at Newcastle University and are grateful for the assistance we receive from this renowned centre of multivariate research and development.

We live in exciting times!

Demand for our innovative data mining technology continues to grow

Welcome to the second issue of our data mining project newsletter. It is certainly a very exciting time for us as our clients see the first results from the complex process investigations for which the MS2 Process Analysis System is designed and as new clients continue to join us.

The level of interest being shown in the MS2 Process Analysis System, and in the benefits it can provide, has taken us by surprise. Some of the most prominent process companies in the world are now adopting our system, and the range of problems to which it can be applied continues to widen.

We have some very exciting new features now available, such as Projection to Standard Model and Principal Component Time Animation. In addition to being very powerful tools for analysis of historic data, these are also vital components of the on-line event detection version which is now in design. Some functions of the Data Pre-Screening Module, together with powerful data import scripting facilities, are also now available.

I am also personally delighted to welcome Professor Julian Morris FREng as an Associate Director as described on page 1. We have certainly come a long way since he and I had the first tentative discussions about multivariate developments within MS2 last November.

Alan Mason

Managing Director

Case Study - Aesica Pharmaceuticals

Aesica Pharmaceuticals Ltd, the Active Pharmaceutical Ingredients company based in Cramlington, Northumberland, supplies customers on every continent. Since the 1970's it has been the supplier of choice for customers ranging from small start-ups to the global giants of the pharmaceutical industry, a position achieved through advanced technology and manufacturing excellence.

The Company's wide range of complex manufacturing processes is backed up by advanced analytical techniques. State of the art computer controlled production processes enables it to be responsive to customers' needs and highly flexible. cGMP excellence is a universal target and FDA audits have demonstrated that its standards are amongst the highest in the world.

Aesica is committed to maintaining its excellent quality and hence to a process of continual striving for improvement. To assist in developing their process knowledge to new levels of sophistication, the Company has now invested in the MS2 Process Analysis System from AJM Consulting to assist in gaining an ever-deeper understanding of the complexity of its processes. Developed by AJM Consulting, this system uses innovative algorithms and visualisation tools to identify the consequences of interaction of process variables and to pinpoint areas in which manufacturing performance can be improved still further.

"The ability of the MS2 process analysis system to co-display a multitude of variables simultaneously then quickly identify interactions has already proved to be very useful.

Whilst confirming some previously held beliefs there have been some genuine surprises in some of the causes behind process variability. It is this ability to handle and analyse diverse data sets in a user-friendly format which I believe will yield significant benefit to our many processes"

John Budge,

Aesica Pharmaceuticals

Projection to Standard Model

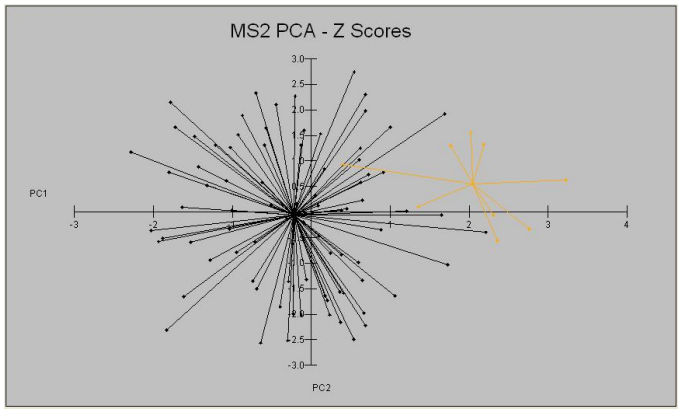
With the normal principal component analysis calculation, the entire data set is used. If a sub-set of samples is then selected, the differences between that sub-set and the data set as a whole can be plotted. This in itself is a very powerful technique and is proven to identify many covariate relationships and causes of poor performance. Whilst this is useful in many applications, we have found that, in some cases, better results are obtained if a technique known as Projection to Standard Model (PSM) is used.

With PSM, a data sub-set is first selected, typically one which contains known good samples. The PCA calculation is performed for this, which results in certain factors being created (the eigenvectors and eigenvalues). These factors are then retained, and used to calculate the principal components for the entire data set. Hence, a selected group of samples, such as those exhibiting a particular problem, is compared not with the entire data set but with the known good samples.

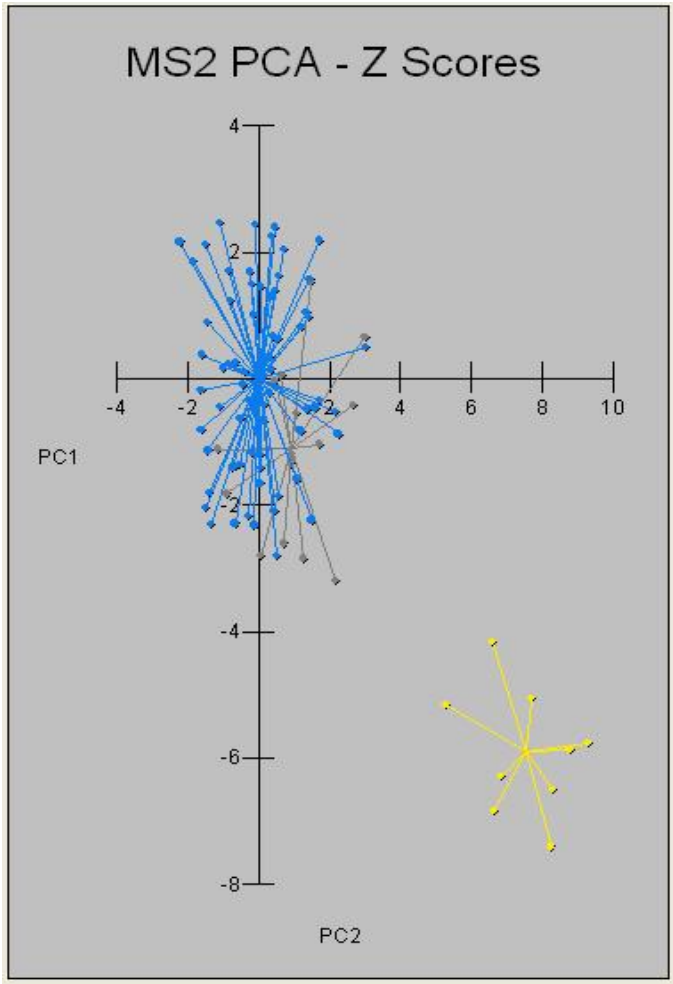
There are several applications in which this technique has proved to be highly beneficial. For example, in several projects MS2 is identifying changing operating conditions over time. A period of known good operation can be used as the initial calculation (often known as the “training set”) and the factors for these used to calculate the scores for particular periods of poor operation. Not only can the difference between these be shown clearly, but also the patterns generated by plotting the various principal components against each other can indicate the type of process problem which has caused the poor quality.

In another example, many batches are produced on a wide variety of process items. A complex process involving many variables, there are several key types of problem resulting in poor quality. PSM has shown itself to be very good at not only indicating the presence of a problem but in identifying the type of problem involved.

Not only is PSM now available in MS2, but we have also added a function to simplify the task of defining the standard model data, which may be done by selecting one or many variables and applying filters to them.



In this example, principal components for the entire data set are calculated. Bad samples are then selected and highlighted in this score plot. Whilst there is obviously a difference, it is imprecise.



In this example (which uses PSM), the good samples are shown in blue, and the bad samples (yellow) are calculated against the good samples. The resultant difference is far more pronounced.

Parallel Coordinates versus Principal Component Analysis

We have frequently been asked why it is that MS2 uses both parallel coordinate visualization (PCV) and principal component analysis (PCA).

The reason is simple - parallel coordinate visualization is a uni-variate method and principal components are multi-variate.

Parallel co-ordinates are very good at identifying situations in which primary variables affect, or are affected by, other primary variables. For instance, in figure 1 the PCV shows, for samples selected as having a high value of variable 3 (Free SO_3), a combination of low drier hours (variable 6) and shift 3 being involved (variable 2) are linked. In instances such as this the information lies directly with the process variables .

But in many cases this is not so, and the information lies in the way that process variables co-vary.

Figure 2 shows a parallel co-ordinate visualization for a data set in which covariance is important. Variable 5 (quality) has had a filter applied to show bad batches. Looking at the other axes, there is no apparent linkage and virtually every axis has a wide spread of samples exhibiting poor quality.

In figure 3, PCA has been calculated, and the principal components plotted to the right of the process variables. There is clearly a message starting to develop, since there are now clear linkages visible between the poor quality samples and their principal components.

The tools in MS2 can then be used to drill down further through the data to arrive at the co-plot of primary variables shown in figure 4. Again, bad batches are highlighted in yellow.

It can clearly be seen that there is a relationship between temperature and pressure which affects quality. Neither of these process variables affects quality directly - good product can be made throughout the entire temperature range and throughout the entire pressure range - but the covariance between temperature and pressure is critical.

In an example such as this, the principal component scores give a very clear indication of process abnormality. They can identify this with much smaller deviations than would be detected by a process control system with an alarm on the individual process variables.

This technique will be particularly useful in the on-line version of MS2, now in development, which will be able to give an early indication of the onset of such problems.

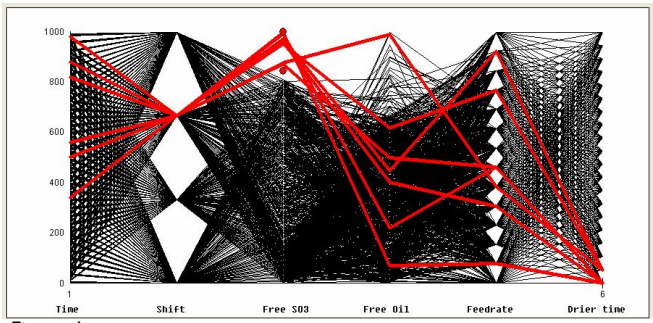


Figure 1

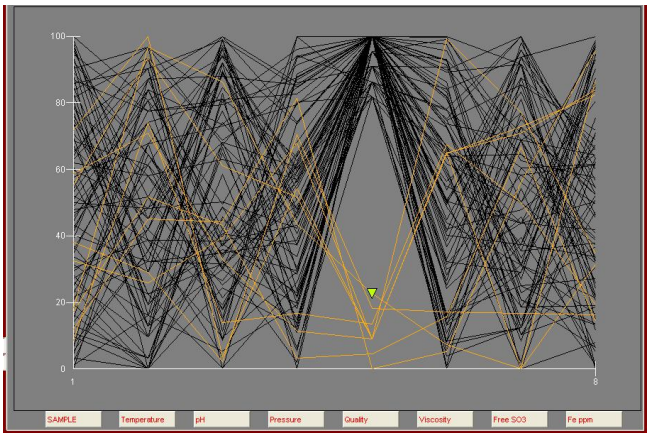


Figure 2

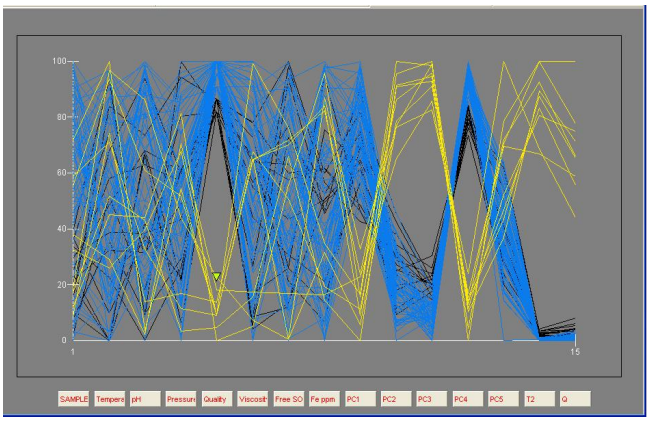


Figure 3

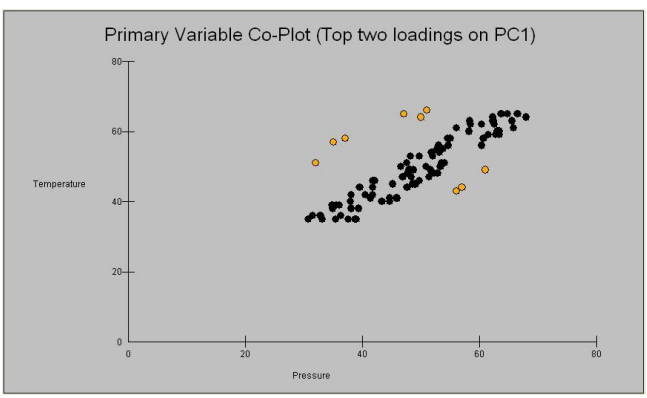


Figure 4

Data Pre-Screening

In many of our investigations we have found that the data is not suitable for analysis in the form it is received. Since we are accepting many different types of data this is not surprising.

Accordingly we have developed a screening facility which is applied to the raw process data imported into MS2, prior to any analysis being carried out. Figure 1 shows a variety of screening rules being applied to a simple data set.

Figure 1—rules being applied to raw data

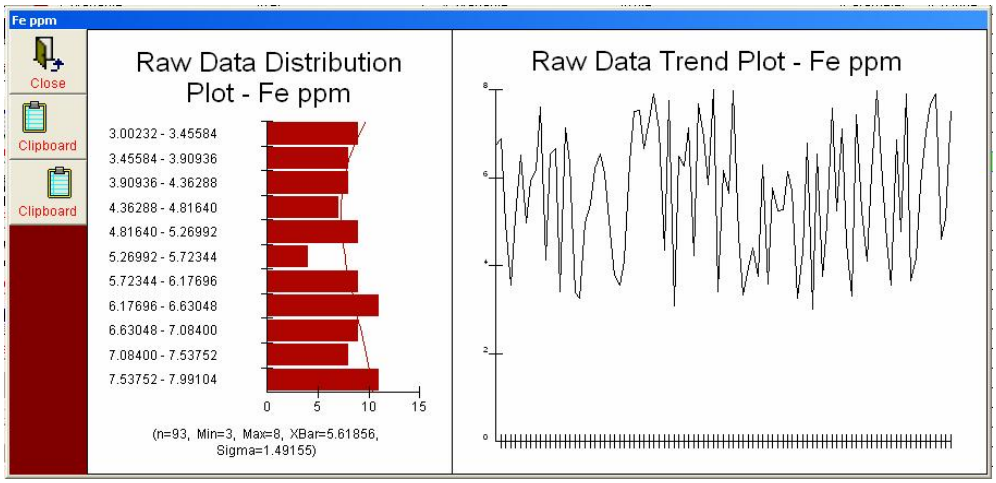


Figure 2– distribution and trend characteristics of raw data being examined to help identify screening requirements

Missing data is a major issue, particularly where data has been entered manually into a spreadsheet or Access database. A rule can be applied which removes any samples with the specified variable having a value of zero. Where the data for a variable is generally very poor, that variable can be excluded from analysis altogether. Alternatively it is possible to retain the last good value if a variable has a missing value, or to extrapolate between good values. The latter is particularly valuable when data is imported from several spreadsheets, representing data collected at different times and from different parts of the process.

Spurious values are another issue, such as a pH of 50! Here, rules can be applied to exclude any samples with the specified variable being above or below permitted values.

Non-numeric values are common. Since the range of problems which MS2 can address is very large, it is inevitable that some variables are not numeric, such as plant item identity or operator name. A group of rules can classify these variables, giving them a numeric value which can then be plotted.

Moving averages are sometimes necessary, particularly in continuous operations, where the variable has high inherent noise.

Variable names can be changed, such as removing part of a lengthy descriptive text to make visibility on screen easier.

As the system develops we anticipate that many more rules will be added to the system.

For large data sets, particularly those with many variables, the operation of identifying which rules need to be applied is quite time-consuming. We noticed this when setting up a large analysis ourselves! Accordingly, MS2 now has a facility which easily displays the distribution characteristics and trend data for each variable; an example is shown in Figure 2.

AJM Consulting

Europarc Innovation Centre
Europarc
Grimsby
North East Lincolnshire
DN37 9TT

Phone: 01472 500306
Fax: 01472 500307
E-mail: cathy.oswin@ajm.co.uk
www.ajm.co.uk

Creating Manufacturing Advantage

Formed in 1990, AJM Consulting was the first company to relocate to Europarc when it opened in 1998. We specialise in the application of IT to the process industries and have worked for over 100 clients (a partial list is shown to the right).

Our Clients:

- ABB
- Arthur D Little
- AEA Technology
- Aesica Pharmaceuticals
- A H Marks
- Bemis
- BP
- British Sugar
- Bridon
- British Paper & Board
- Camaxys
- Cambridge Consultants
- Carbon Trust
- ConocoPhillips
- DEFRA
- DICIDA
- DTI
- Eli Lilly
- Fibres Worldwide
- GlaxoSmithKline
- Glinojeck Sugar Refinery
- Grotech Production
- Hickson
- Holliday Pigments
- Humber Chemical Focus
- Huntsman
- ICI
- Impress/HCCTA
- Imerys
- J R Crompton
- Lakeland Laboratories
- Luxus
- Manro
- Norsk Hydro
- PICME
- PSE
- Roche
- Rotork
- Scott Bader
- Servelec
- Severn Trent Water
- SIRA
- Spiritus Consulting
- Strategem Consulting
- Stepan
- STG
- Sulzer
- Syngenta
- Synthomer
- Technical Absorbents
- Tensachem
- Total
- University of Newcastle
- University of Sheffield
- University of Strathclyde

Frequently Asked Questions

What is the range of function options available?

Some users require all MS2's functionality whilst others only require parallel co-ordinate visualization of a small number of samples. Consequently we provide MS2 at a number of levels, reflecting both the range of functions provided and cost.

- Level 1 provides parallel coordinates for a limited number of variables and samples
- Level 2 provides a more powerful parallel coordinate visualization suitable for larger data sets
- Level 3 adds principal component analysis
- Level 4 adds complex batch analysis including the ability to include process variable trends (known as multi-way PCA)

Is training available?

Yes. We plan to offer our first training course in late November and this will be announced on our website and to all users shortly.

What is the timescale for on-line development?

We are currently planning to develop the on-line predictive system during winter 2006-7 with the first demonstration system being available by Christmas.

Are you looking for development partners for on-line?

Very much so. This methodology has proven itself very useful indeed so far; we have learned much valuable knowledge about the needs of our users which we have translated into functionality within the system.

Will there be any user group meetings?

Yes. Since we now have 12 users, with several people from each company involved, and since we anticipate many more companies getting involved in the near future, we are planning to hold a meeting at which we will invite open discussion and, in particular, to develop the "wish list" of functions people most want to see added in future releases.

What hardware and operating system requirements are there?

For hardware, virtually any modern PC is sufficient. There is an absolute minimum of 256MB of RAM, but we recommend 512MB at least. For very large data sets more may be advantageous. Our recommended operating system is Windows XP Professional, and our tests have shown no problems so far using the new Vista operating system.

Do you provide consultancy or are you just selling the system?

We find that, typically, all users require some assistance in defining and using the system. We are also offering to provide specific consultancy to clients who have an issue they need to resolve but do not want to buy the system.

How much longer will you be offering Development Partner relationships?

For the forthcoming on-line development we have just started. For the off-line system now available, this is coming to an end. The system is now available for purchase in the normal way.