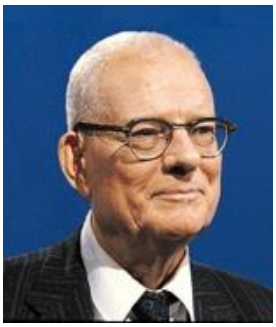
MS2 Data Mining & Visualisation Project Update

AJM Consulting

Issue 1

July 2006

Understand the Process to Improve Profitability



"If you are responsible for the management or improvement of a process then you need to understand process variability."

Dr W. Edwards Deming

"Information lies not in any individual process variable but in how they co-vary"

Prof. Julian Morris FREng

In order to make high quality product efficiently, consistently and safely, with minimal waste and maximized competitiveness, it is necessary to minimize variability. But process factories are complex, and understanding this variability to identify the causes can be a very difficult task.

Technologies which enable this exist and are well proven in large factories with very significant improvements in profitability being generated. But until now, such multivariate analysis has been impractical for many smaller sites, or indeed for most individual applications at large facilities. Many companies are unaware of the existence of solutions, or do not understand how they can be used. Cost, particularly for small companies, has been an issue, as has the amount of time needed from scarce on-site experts. Availability of data is also a common concern.

Our aim for this project, therefore, was to develop a suite of multivariate tools which are designed for small to medium process manufacturing factories. Ease of use, low cost and flexibility were key goals in the design process.

By involving process manufacturers in the development of the system at an early stage we ensured that the design process was efficient and addressed their requirements from the start. Using actual plant data, rather than simulations, enabled us to identify rapidly what analyses and resultant visualizations are needed and how they are used.

As the number of applications continues to grow we are identifying further functionality which is needed and are incorporating this into the system design, thereby developing an increasingly powerful manufacturing improvement tool whilst retaining the original concepts of ease of use and availability.

However, our original concept of designing the system specifically for small sites has been overtaken by events! Large sites such as major oil refineries are identifying applications and today the MS2 system is being applied to factories both large and small.

Inside this issue:

Funding from Yorkshire Forward

AJM Consulting bid for and won financial support for part of this project from DTI's competitive Research and Development Fund, awarded through Yorkshire Forward. We are grateful for this support which has enabled the project to become reality.

PHASE 2 - Take-off!

Our innovative data mining functionality is helping industry beat the competition

Welcome to the first issue of our data mining project newsletter. We'll be releasing these regularly over the coming months as this exciting project develops. Its purpose is to provide a background to the project, together with an understanding of the business benefits which are achievable and the technologies which enable them. It informs existing partners of progress and describes how others can get involved.

Since the pace of this project is such that any brochures would be outdated before they left the printers, we hope that these newsletters will inform you of the technical content of the developing system and the progress of the project as a whole.

The big news is the massive interest shown in our project from manufacturers. Friday, July 14 will be forever etched in our memory as the day when no less than three new partners agreed to come on board. For us, as a consultancy accustomed to perhaps four or five new clients in a year, it is proof that the flexible approach we are taking meets the needs of the wide range of process industries which we serve.

Many industries, such as speciality chemicals or surface coatings, face intense competition from overseas. Only by ensuring that process manufacturing operations are as efficient as possible can these industries survive and prosper. The aim of this project is to enable sites of any size to benefit from advanced technologies which were previously only available to large sites and at high cost.

In Phase 1 of the project we developed the basic principles of the data mining and visualisation techniques which we are integrating into the MS2 system.

With the launch of Phase 2 in July, we take these concepts further to increase the power and flexibility of the system to improve ease of use and the range of problems it can address.

Alan Mason

Managing Director

CPACT

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

We have been involved with CPACT for years, and know the organisation well.

Last November we exhibited at the Chemical Engineering North 2005 exhibition at Harrogate, where CPACT also had a stand presence. Having seen our system, which included a rudimentary parallel coordinate visualisation, several visitors to the stand commented that, if we had the technology that CPACT were exhibiting, combined with ours, we would be "onto a winner".

That very evening, Prof. Julian Morris and Alan Mason had dinner together. The rest, as they say, is history.

Case Study—Holliday Pigments

Holliday Pigments is the world's largest manufacturer of ultramarine blue pigments, exporting to over 80 countries.

Holliday Pigments' site at Hull is highly complex; part batch, part continuous. Consistency is the key to maximized quality. This can only be achieved by in-depth understanding of the causes of variability. One of the earliest companies to join phase 1 of our development, Holliday Pigments uses the MS2 Process Analysis System to mine the considerable amount of their data which is available and to identify operating parameters which enable consistently high quality production.

"Principal component analysis has been of particular interest and has given us some valuable insights into options to explore in process control strategies and planned maintenance activities"

Glyn Jagger,

Operations Director

Industrial Partners

From the start of system development in March this year to the present time, we have been delighted by the number of companies which have signed up as development partners. In addition to those listed here, many other companies are currently in discussions with us and we expect many new contracts shortly.

A H MARKS

Holliday Pigments

synthomer

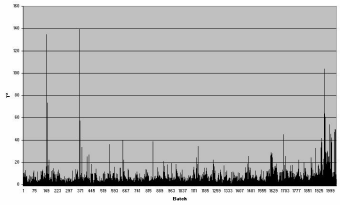
ConocoPhillips

Fibres Worldwide

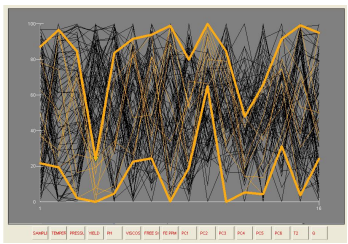
Technical Absorbents
The Absorbent Solutions Provider

JOTUN

Aesica
PHARMACEUTICALS LIMITED



Hotelling's T2 is one of several new multivariate statistics being developed for Phase 2.



Multivariate statistical results (principal components, T2 and Q) can be co-plotted with process values and quality results in the parallel coordinate visualisation.

What's in Phase 2?

Phase 1 saw the development of the basic visualization functionality, including parallel coordinates, correlation matrices and principal component analysis. The basic batch structure for the underlying database was also developed.

In Phase 2, these will be enhanced, for instance by the addition of a more powerful animation facility in parallel coordinates.

Further multivariate capability includes Hotelling's T2 and the Q statistic. Together, these substantially improve the amount and quality of information which can be derived from multivariate analysis. It is also possible to display principal component values, together with T2 and Q, as part of the parallel coordinate graph. Confidence bounds are also being developed.

Function blocks, for instance to calculate yield or time to peak exotherm, will be added.

More comprehensive drill-down is now a reality; from any data analysis graph a simple mouse click reveals the next layer of detail, right back to primary batch records if they are within the model.

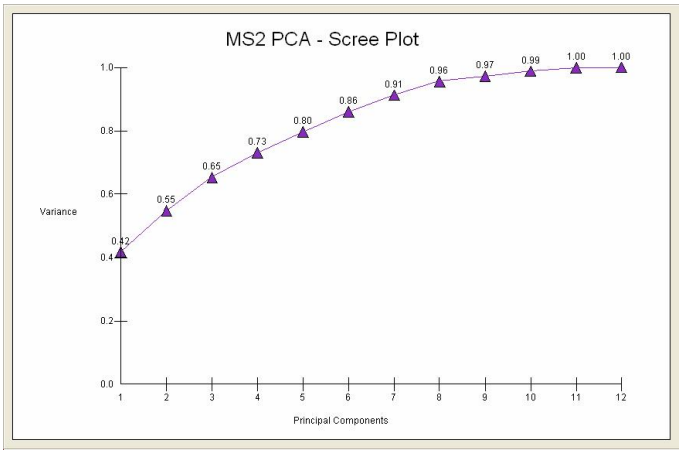
For continuous processes in particular, time dynamics are critical. For instance a process taking two days from start to finish cannot be analysed by simple time-based readings. It is necessary to offset these so that data from an early part of the process relates to the same material at a later stage. This is a significant part of our Phase 2 development.

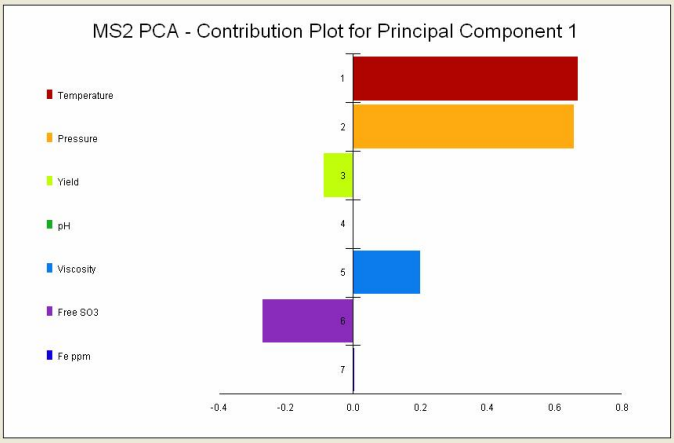
Data pre-screening and associated validation functions will increase the ease of use and reliability of observations and conclusions.

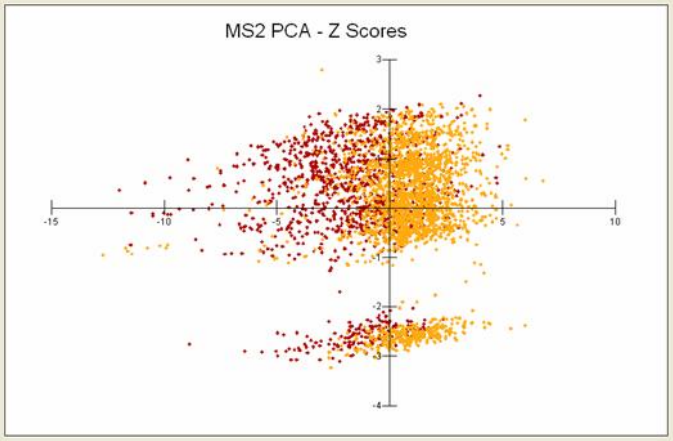
On-line help, training examples and documentation are all part of Phase 2.

Some of the many new features in Phase 2 are currently the subject of patent application.

Principal Component Analysis







Principal component analysis (PCA) is a powerful multivariate statistical analysis technique that is designed to extract major features within a data set. It can be applied in many fields such as face recognition, image compression and data mining. Patterns in data sets of high dimensionality can be hard to find using traditional SPC techniques. PCA can quickly identify patterns and express the data in such a way as to highlight its similarities and differences. Several statistics can then be identified that will provide important information regarding process plant information.

Principal component analysis is extremely useful for finding structure in data sets. PCA rotates the data into a new set of axes called principal components. The value of each point, when rotated to a given axis, is called the principal component value (or Z-Score). By plotting the data on these axes, we can immediately spot major underlying structures. Another main advantage of PCA is that most of the variations within the data are reflected in the first few Principal Components. This means that once the patterns in the data have been found, the data can be compressed. This reduces the number of dimensions without much loss of information.

After the Z-Score plots are used to identify patterns, contribution plots can be used to trace this back to primary process and quality variables. When monitoring a process plant using PCA, the statistics that tend to be of most use are Hotelling's T^2 statistic and the Q statistic.

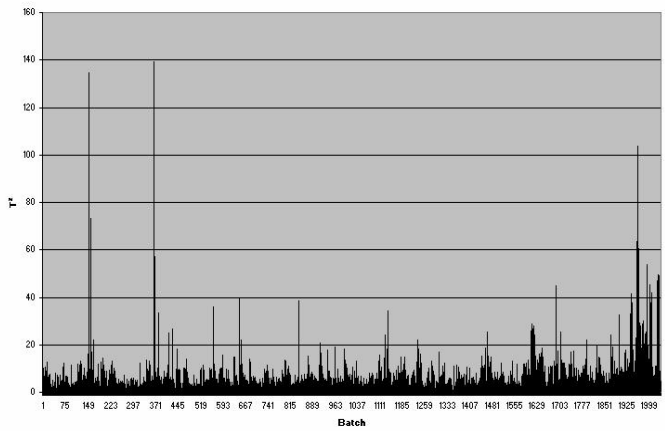
In traditional univariate statistical analysis, each variable has its own control limit. In multivariate statistical analysis, this is not a sufficient check to verify that the process is in control. A batch can be within the control limits for its individual variables using univariate methods but can still end up a bad batch. This is because in multivariate situations, the control limits for each variable will depend on the current values of the other variables.

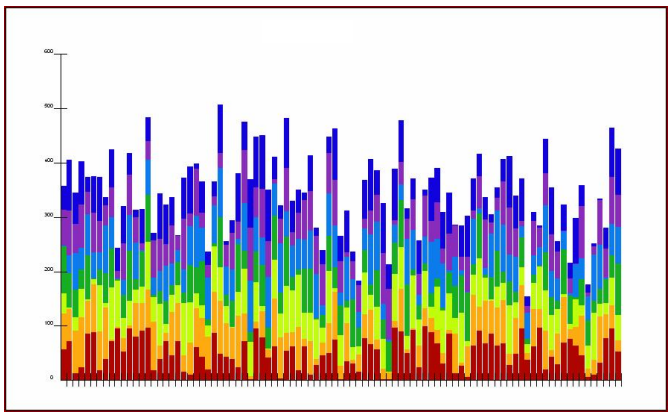
Hotelling's T^2 statistic is the multivariate equivalent of a univariate control limit. This differs from univariate control limits in the way that each batch will only have one value for the T^2 statistic, instead of a separate control limit for each

variable. A T^2 chart offers an immediate visual display of any batch that is out of control. In this context, out of control means when the magnitude of the variables is not within expected limits. If a batch has a high T^2 value, contribution plots can be displayed for the T^2 statistic for that particular batch. The contribution plot will indicate which principal component values are causing the high T^2 value. As described above, this can be traced back to primary variables by looking at the contribution plots for the principal components identified from the T^2 contribution plot.

Another statistic that tends to be of most use when monitoring a process plant using PCA is termed the squared prediction error (SPE). This is often called the Q-statistic. The Q-statistic provides a measure of how the relationships between the process variables compare with those identified under normal operating conditions. This is extremely useful when a batch is bad but the control limits and T^2 statistic indicate that the process is within control. Although none of the variables have gone out of control, the correlation between the variables is not the same as expected under normal operating conditions. Therefore the process is out of control. The Q-statistic measures this change in the correlation between the variables.

By combining these methods and statistics, principal component analysis is extremely powerful and useful. This is why it is now the most common research technique for finding patterns in data sets of high dimension.





A typical process investigation study

For most of the sites where the MS2 data mining system is being implemented, we have adopted a standard approach which uses very little staff time, and it is working very well.

At a preliminary meeting (usually when the client is being presented with MS2 and its concepts and is deciding to purchase) a real problem is identified; one in which performance would be improved if the causes could be identified. A "champion" is then identified who will be our primary point of contact with the client.

A start-up meeting is then arranged at which the nature of the problem is described in sufficient detail to allow us to understand it. The data which is required to populate the MS2 data mining system is provided at this point.

At our offices, the data is loaded onto MS2 after any relevant data translations have been configured and preliminary investigations into the problem and linkages to possible causes are identified. During this process we typically require a few phone calls or email discussions with the client's point of contact.

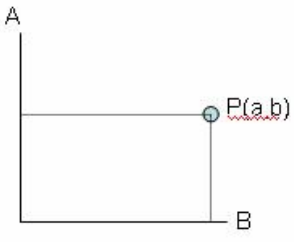
Once any modifications needed to interpret the data have been made and some apparent linkages between problem and possible causes defined, a workshop is held at the client site to present our conclusions. At this meeting, which can have several site staff present, other queries can be run, to allow "pet theories" to be examined.

Finally, the client is offered a copy of the MS2 data mining system, pre-configured with the site data, and shown how to use it. This system also includes a facility to load data representing other problems using a standard spreadsheet structure.

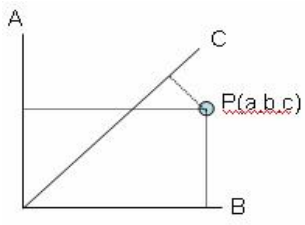
Parallel Coordinate Visualisation

Where many variables exist it is useful to be able to compare them, to discover whether any hidden relationships exist. But normal graph techniques only permit two or three variables to be plotted. Analysing a complex situation in this way can result in hundreds of graphs, and is not only tedious but inflexible.

What is needed is a way of plotting the relationships, if any, between many variables. Parallel coordinates are simply a way of plotting many variables (one current project requires 57) in a way which enables any relationships to be discovered. In the typical XY graph below, the values for variables A and B are shown, for an individual sample, as a point P:

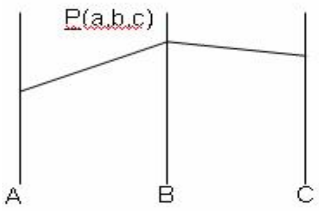


Similarly, it is possible to plot three values:

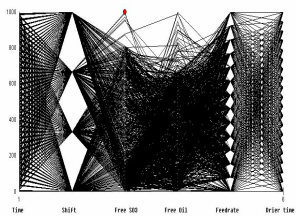


However, plotting more than three is not possible. But if the axes are drawn differently, the possibilities are limitless.

If we take the three axes for variables A, B and C in the previous graph, and plot them in parallel to each other, we can show the values for the variables on each axis:

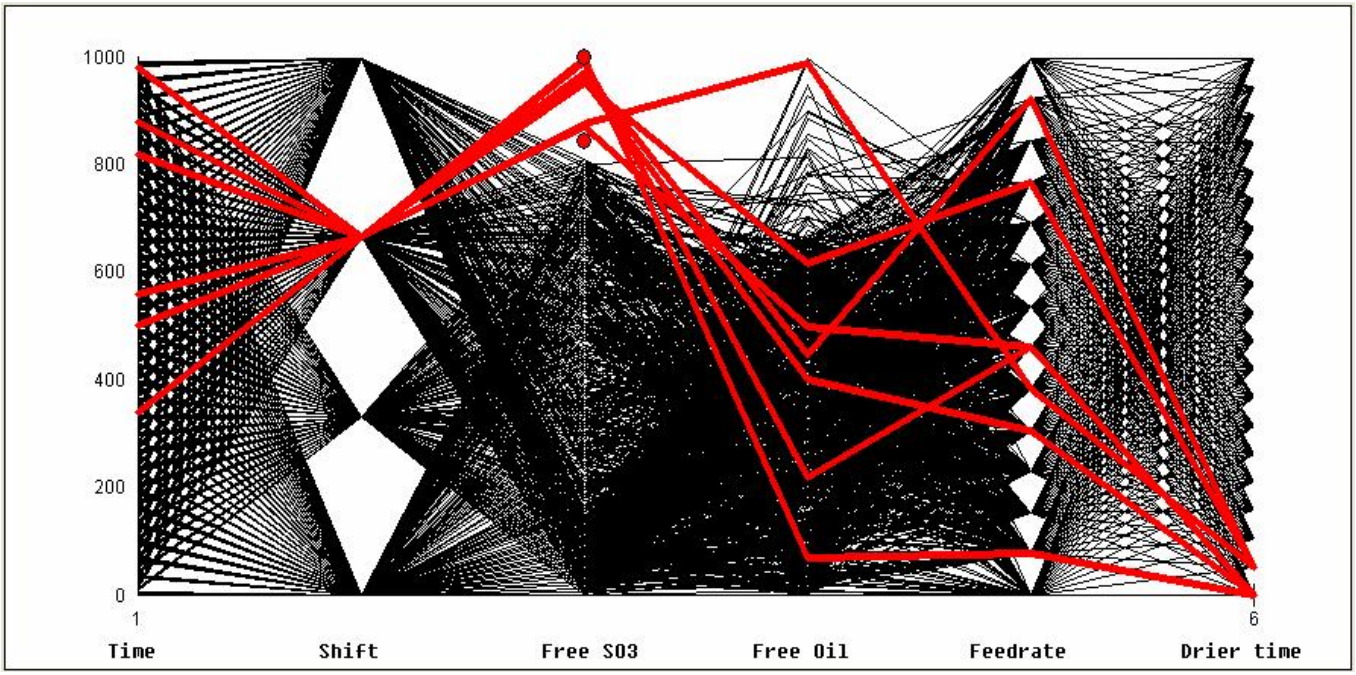


The sample which was shown as the point P now becomes the line P. Once the axes are parallel, many variables can be shown. Each vertical axis can describe different processing variables, such as temperatures, flows or pressure; quality values such as pH or viscosity; shift or operator identities, and so on. In each case, the system scales the axis to fit the lowest and highest value for the variable within the data set:



If filters are applied to an axis, and any sample with a value for that axis which is between the upper and lower filter limits is highlighted, relationships can be clearly seen. In the example at the bottom of the page, a filter has been applied to the Free SO_3 variable, and a clear relationship between this being high, a low drier time and an individual shift can be identified.

Other features relating to MS2's parallel coordinate visualization include distribution curves and counts, highlighting of any associated trend graphs, integration with principal component displays and trans-axial animation.



Getting the data

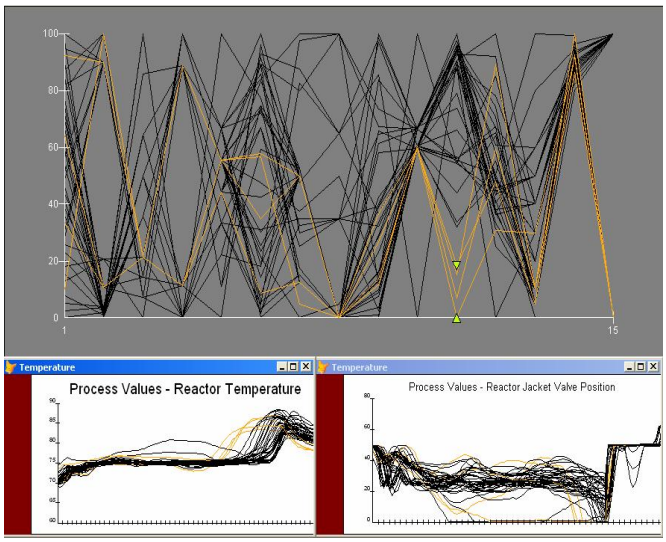
We have overcome one of the major obstacles to data mining - getting the data - by developing a range of modules which can receive incoming data and translate it into the standard structures used in the existing MS2 manufacturing system.

We recognize that, depending on the application, this data can have many sources, for instance it might be available as a spreadsheet or Access database; it might be batch data from a distributed control system (DCS) or trend data from a process historian.

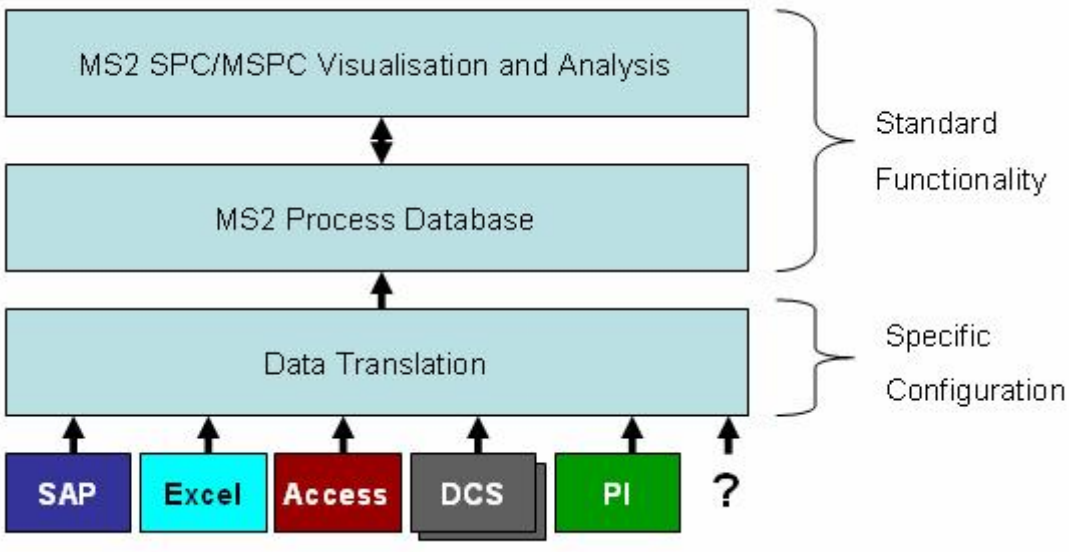
We can also analyse different types of data; quality data, for example, is usually one data point per batch, whilst a temperature trend could have hundreds of data points.

It is also important to screen the data - to take a very simple example, if a data point is missing the calculations must allow for this, or any results based on range calculations will be meaningless.

For many applications, data from several sources can be integrated to give very powerful visualisation. For example, a batch reaction could have attributes such as start and end times held on a DCS, with the corresponding trend data in a historian. With MS2, combining the two into a single visualization is simple (see right).



In this example, the parallel coordinate plot shows one-per-batch data points for both quality and process values, and batches with a specific quality parameter range have been selected. Trend data has been extracted from a data historian for the relevant time periods, and the trend information displays show the selected batches in highlighted colour.



Overall architecture of the MS2 Data Mining System

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 Business Advantage

Formed in 1990, AJM Consulting was the first company to relocate to Europarc when it opened in 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
- AEA Technology
- Aesica Pharmaceuticals
- A H Marks
- Arthur D Little
- BP
- Bemis
- 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
- ICI
- Impress/HCCTA
- Imerys
- Jotun Paints
- 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
- Scheduling Technology
- Sulzer
- Synthomer
- Technical Absorbents
- Tensachem
- University of Newcastle
- University of Sheffield
- University of Strathclyde
- UMIST
- Yorkshire Forward

Frequently Asked Questions

Will there be a Phase 3?

Oh yes. This is planned to start in late 2006 and end in 2007.

Will the system ever be on-line?

Yes. This is one of the main issues we will address in Phase 3. But first we need to complete all aspects of off-line analysis and extend our range of multivariate algorithms.

What else is in Phase 3?

Wait and see! We have some great ideas for unique functions and we want to keep them that way.

Should we contract as development partners now or wait for full product release?

Obviously you can do either, but we recommend joining now as a partner. Not only will it be considerably cheaper than licence and consultancy purchase but also you will get the chance to influence development to address your specific issues. And you will get the system earlier.

Will there be a maintenance contract / upgrade path?

Yes, and the costs will be announced during Phase 2.

Is MS2 Data Mining designed for batch or continuous processes?

It is fully applicable to both. It has some specific functionality to further assist analysis of S88-structured batch processes.

Is it single or multi-user?

It can be either. Licence control uses a hardware USB key which can either be standalone or networked.

Is training provided?

It will be. Once we are sufficiently far into Phase 2 to confirm the details of algorithms and interfaces we will develop training materials and courses.

Can we use it for other things?

Yes. In addition to any configuration which is specific to your own application, MS2 also permits the import of a spreadsheet which you can create. All MS2 univariate and multivariate analysis functions are available to the data held in this spreadsheet.

What operating system does it require?

Windows XP or Vista.

What is the anticipated time we need to invest in Phase 2?

Obviously dependent on the individual project but typically around two man days. Most clients are finding it useful to involve more staff for training purposes.

Can we buy some functionality only, at a reduced cost?

Yes. There will be a range of licence options.

Is there any cost penalty if we subsequently upgrade to a more comprehensive licence?

No.

Cost?

Talk to us.