# Chemometrics - A holistic approach to troubleshooting.

Michael J. McCann, McCann Science, Chadds Ford, PA,
Charles A. Bishop, Bishop Consultancy Ltd., England.

**Key Words:** Thermal modeling, Applications, Statistical process control, Operation

## ABSTRACT

It is common when troubleshooting to isolate the problem area and to look at that area in great depth. This is particularly true for complex processes.

However, this can mean that interactions that are distanced from each other within the whole process are missed. The data used can include information on parts of the process that might normally not be included. If there is more than one supplier of raw materials, this information can be incorporated, as too could be data from an earlier wet coating process. Thus, this technique can be regarded as truly holistic in nature.

The use of chemometrics or multivariate analysis enables the whole process to be analysed at once. Using tools such as hierarchical cluster analysis and principal components analysis enable interactions to be identified and, if necessary, tested. This technique can be used on production processes, it is not exclusive to research, and it is not intrusive so that it is not disruptive to production.

## INTRODUCTION:

Finding out what ails a process can be difficult if the problems do not show any direct link to immediately observable and related circumstances. The usual assumption is that trouble is caused by something proximate. Often it is hoped that statistical analysis will reveal the answers from current plant data. There are problems with relying on statistical methods if you are not an expert (and sometimes even if you are). [See Reference 1]. We see two key issues:
(a) analysing data and not finding a cause/effect relationship when one exists
(b) analysing data and finding a spurious cause/effect relationship and not knowing that it is spurious.

We wanted to try the use of a package of data analysis software that would allow anyone to get all the real relationships out of the available data without making any prior assumptions as to what was the cause of the trouble.

We had experience of some results (not published) which showed that Pirouette [2] software had been instrumental in identifying some parameters, which nobody had considered before the work was done as being relevant, to be the source of a film processing problem. For this investigation we couldn't get enough real data for which we knew with certainty all the relationships, true and false so we split the investigation between authors. One (MJMcC) was to generate the data and the other (CAB) was to investigate, from zero experience base, the use of the software. We used a series of tests, starting with very simple data streams, moving on to the main example shown here.

The data files were generated using MATHCAD [3] which includes facilities for generating streams of random numbers with known and controllable probability distributions, as well as all the necessary computations to run dynamic models [4] of web coating processes.

## MODELLED SYSTEM: TIME SERIES

After a simple sample had been run to get make sure we had the file formats right, the first tests were made with time series data. Using a model of a vacuum web coating process akin to that in Ref 4, the available records showed plant operating data for the processing of a whole roll of polymer film. The data was available on a second by second basis and with web speed around 0.1 m/s, correspondingly at about 10cm intervals along the web. There was a deliberate cyclic speed variation introduced into the web speed, which then showed up slightly in coating thickness and a control system which tended to overshoot on the power, but no "random" variables. The user of the Pirouette software was not told of the anomalies. The available variables were time, heater power, speed, length run, controlled temperature and resulting thickness.

The Pirouette analysis showed 3 distinct regions as shown in Figure 1. Since Pirouette uses colour coded displays, the colour versions of the figures in this paper are available in [5].

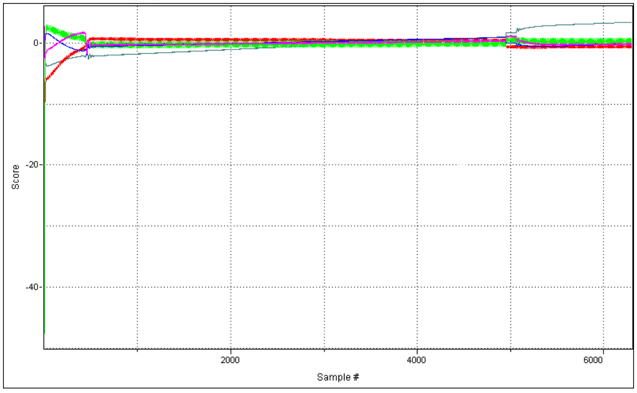

Figure 1. Scores against sample number show distinct regions.

This plot shows the influence of the data on the processing. The first region shows large spikes due to the pumpdown and switching on of the sources. The second region is fairly uniform and then the final region has a disturbance from switching off the sources. The analysis needed to be done in two ways one was to analyse the centre region only to determine the fine detail of the deposition interactions. The other required that multiple complete processing data sets be used to be able to include the pumpdown and switching to determine if those parts of the process were also contributing to the product variation. It was decided to work only on the deposition region and a new data set Data03 was generated.

With the same sort of model coater, a file set was created which looked only at the part of the time when coating was being done, so that the transients for warming up and stopping were eliminated. Furthermore, although the speed wobble had been removed, there was a long period wave in speed, rising slightly on average, to sweep over some dynamic performance range. The data file for the whole run was sampled at 1 minute intervals to reduce the file to manageable size. The odd behaviour was a mechanical disturbance, generated from the take-off roll so that its periodicity became shorter and shorter as the run progressed. It showed up as a variation in thickness, cycling in step with the rotation of off-take reel. With a core diameter of 100mm, and a web speed of between 0.05 and 0.1 m/s, this was detectable in data sampled at 1 second intervals, but was very confusingly aliased into a wave of apparently widely varying frequency by sampling at 1 minute intervals. See Figure 2.

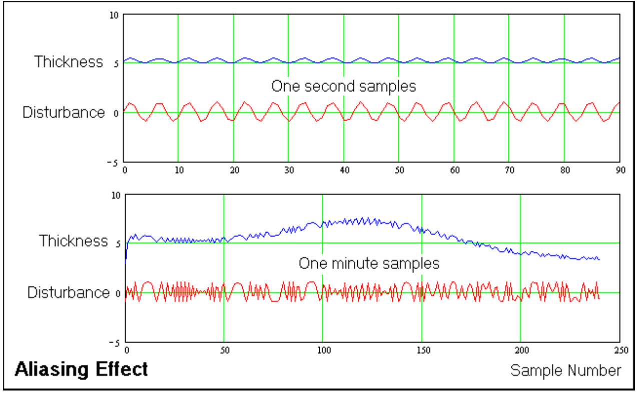

Figure 2. Data sampling interacts with system frequencies to give strange patterns.

To give a fair chance to the analyst, a set of short time records, with 1 second sampling, were produced, each representing a few minutes running, but separated by half hour intervals. In principle, it would be possible to relate the thickness to distance run and hence deduce the underlying pattern. Nothing was "random" in these data.

The Data03 set of files were analysed in two ways, each file individually and then in two groups. The individual analyses showed that factor 2 was strongly influenced by a variable or variables that had a changing frequency. There was not a clear indication of which of the variables was the root cause. The grouped data showed the initial set to be significantly different from the rest. All the sets were well defined and this would have enabled a model to be generated to classify future data into the same classes.

Hindsight has shown that I (CAB) over-analysed the individual data sets which created confusion. The 'scores' plot for the data set 'a', Figure 3, does look to show the correct disturbance waveform but without knowing the answer I had not at this time determined the source.

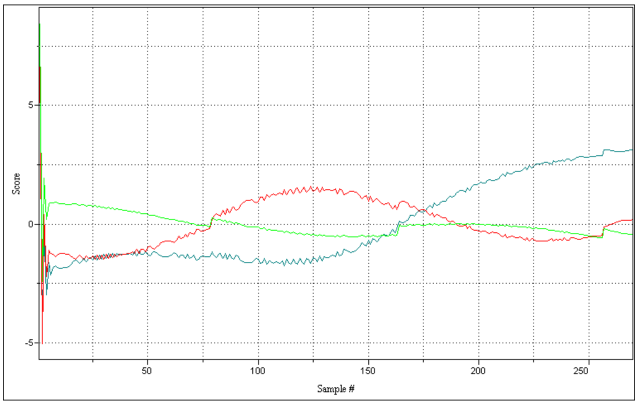

Figure 3. Scores plot for first part of Data03.

## MODELLED SYSTEM: ROLL SERIES

To test the method on random data, a set of models was constructed which progressed, without the analyst being aware of it, from totally random (no internal relationships in the data) to having a link between history and results, but all masked by process variations.

The data as presented showed the useful product yield from each of 500 rolls. The 500 rolls represented about 250 days work. The data for each roll consisted of date of fabrication, date of processing in coater, relative humidity and ambient temperature at both processing and at original fabrication (6 history variables), set point temperature, speed and vacuum pressure during coating (3 process variables), pinhole count, adhesion score, average coating thickness and yield of finished product (4 quality variables).

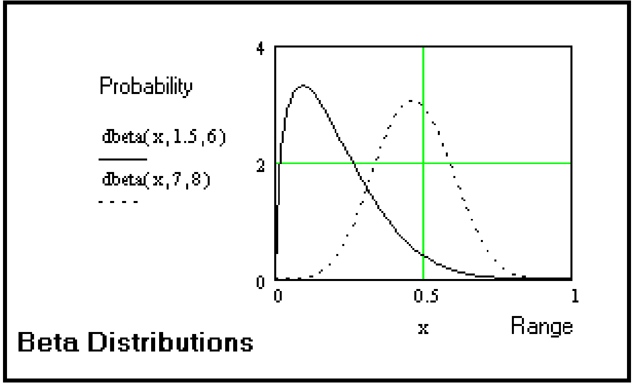

Figure 4. Most of the distributions were skewed, and bounded.

None of the random number generators was Gaussian, all were bounded, and had generally very skewed probability distributions. See Figure 4. The relationships built in, when used, were likewise typically non-linear.

The weather data (hot sticky summers, cool dry winters) was the same for both fabrication site and processing site. See Figure 5, "Weather".

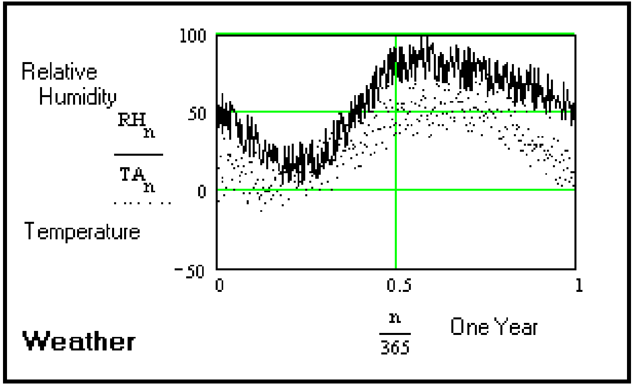

Figure 5. Weather data for both fabrication and processing.

There was a random, predominantly about 45 days, delay between fabrication and processing (coating). See Figure 6.

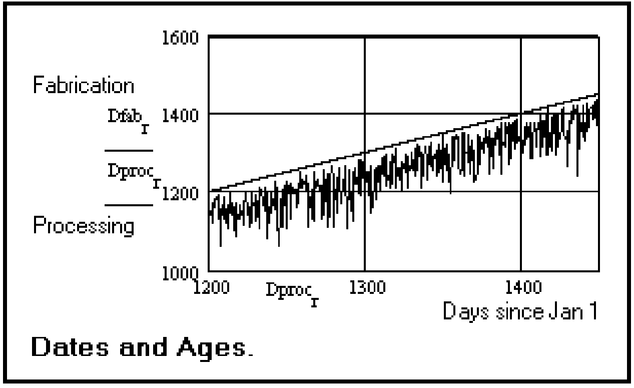

Figure 6. Relationship between date of fabrication and date of coating process.

The first of these tests (Data04) simply used random variables for all the data. The only correlation was that caused by the weather model which superimposed separate random variables on annual cycles of mean temperature and mean humidity.

The first thing the processing showed was that there were three points that were completely different from everything else. This is highlighted in Figure 7 where the different factors are plotted against each other and it can clearly be seen that there are three points that stand out. These are from three consecutive rolls and going back and looking at the data shows that there was a missing decimal point. The three rolls were eliminated from the spreadsheet and the processing re-done. Figure 8 shows a similar plot of factors once these three points are eliminated. Here it can be seen that there is a large scatter of points, which indicates there is little correlation within the data except for the weather effects.

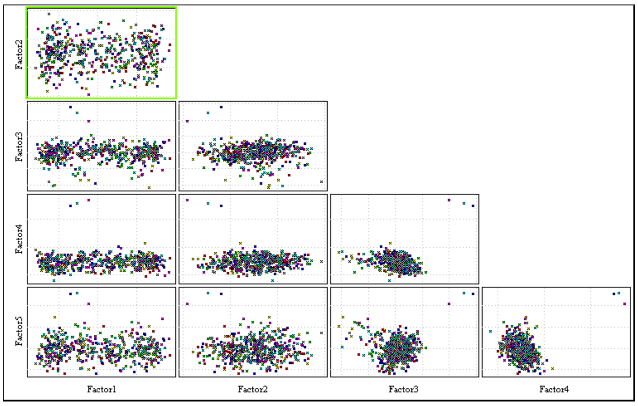

Figure 7. Cross plots of factors show outliers

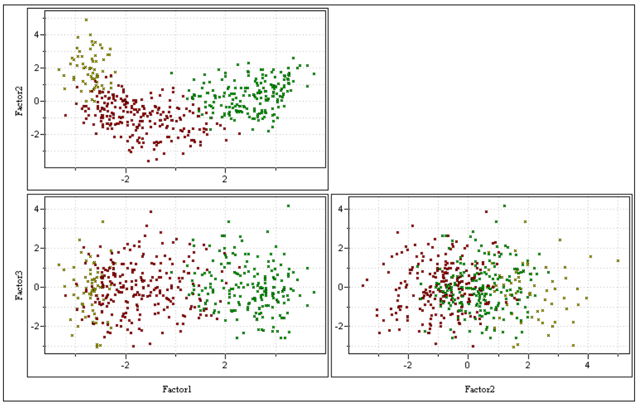

Figure 8 Outliers removed

In the next test (Data05), it was admitted that the yield depended on the thickness, the pinhole count and the adhesion score, since it would make sense that anyone working on such a process would know that. (See Figure 9, showing the connections between variables.)

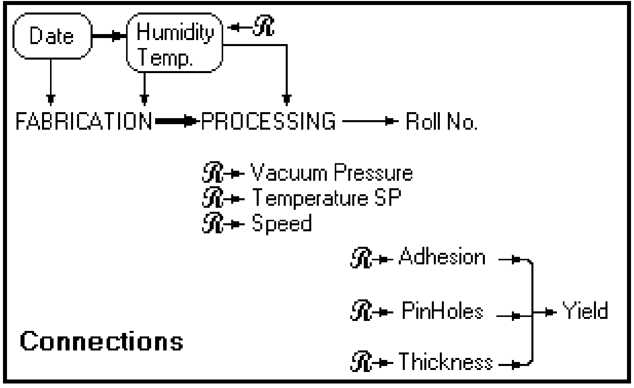

Figure 9. Connections between variables.

The specific relationship was not stated. Also not stated was that there was still no other cause-effect link in the model between either the six history variables or the three process variables and the resulting four measured quality values.

The apparent results of Data05 were very much the same as for Data04. At this stage the independent and dependant variables had not been separated.

In the final test (Data06), there was a (cryptic) relationship associated with the history. The buried rule was that if the average of processing ambient temperature and fabrication ambient temperature exceeded 40 degrees and the delay exceeded 15 days, then the adhesion went down by an amount proportional to the product of average temperature excess and time delay. See Figure 10.

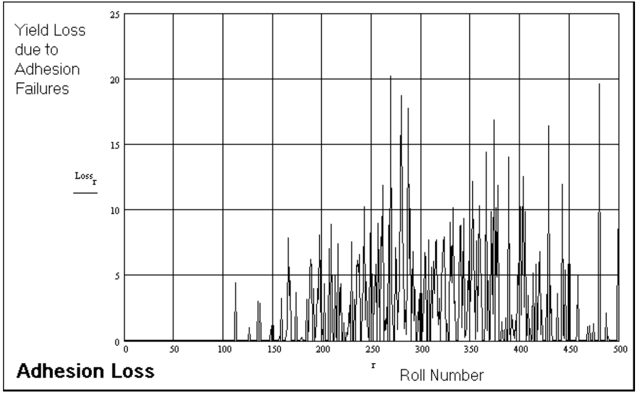

Figure 10. Adhesion loss for each roll.

As in the previous test, this affected yield as well, but there was a lot of other noise in the system. See Figure 11.

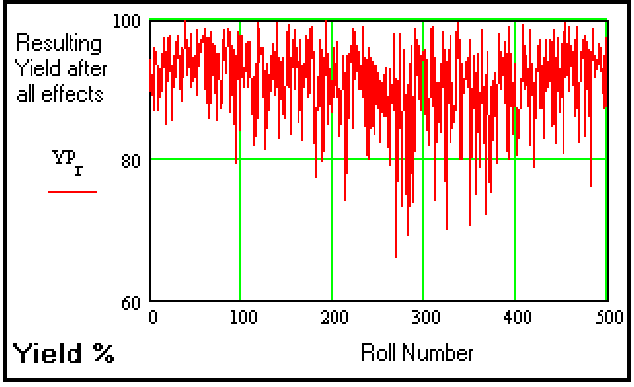

Figure 11. Yield allowing for all variations.

Data06 was processed several times. The information that the Vapour pressure, Thickness, Yield, Adhesion and Pinholes count were dependant variables was added into the processing.

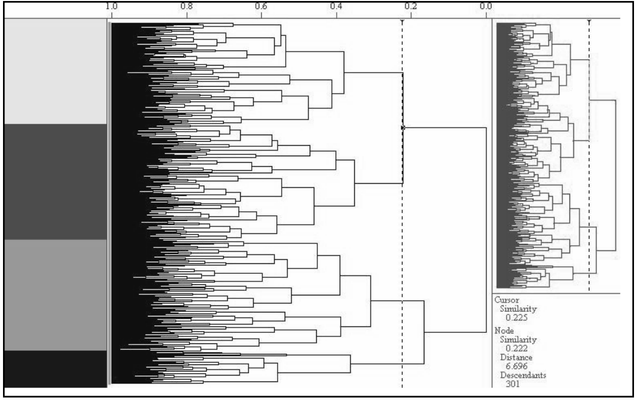

Figure 12. Hierarchical Cluster Analysis of Data06

The processing showed the data could best be described by three principal components. The 3D plot is shown in Figure 13.

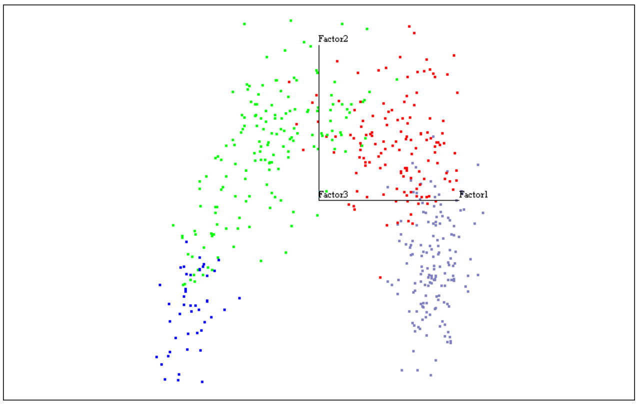

Figure 13. 3D plots can be rotated.

I then looked at the contributions that the different independent variables made to these three principal components. These are shown in the two loadings plots shown in figures 14 and 15.

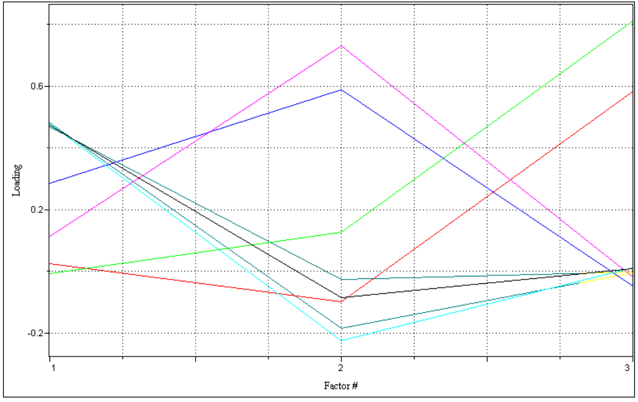

Figure 14. Loading plots. Variables contribute to factors.

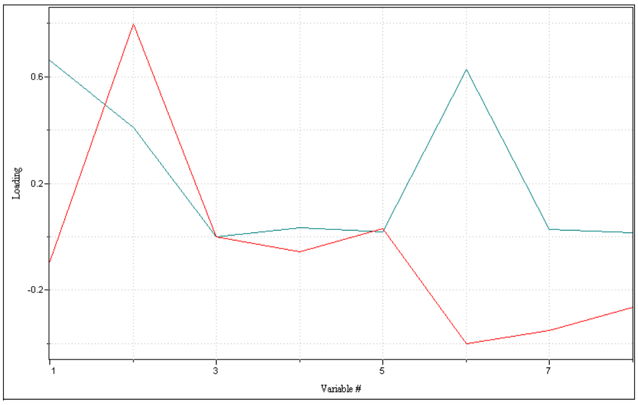

Figure 15. Loading Plot. Variables contribute to principal components.

The correlations that these highlight are that the dates of fabrication and processing and the temperature and humidity at fabrication are the largest contributors to principal component 1. The temperature and humidity at processing are slightly less significant contributors.

The temperature and humidity at processing are the main influences with the dates of fabrication and processing being the minor influences on principal component 2.

The speed and temperature of the source are the main influences on principal component 3. They will both affect the thickness of a deposited coating. However, in this model they did not (MJMcC). As a sanity check the variables are plotted against each other, these are shown in figure 16.

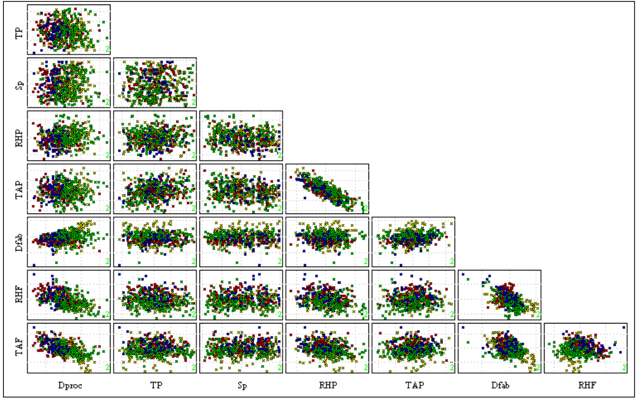

Figure 16. Cross plots of original data.

The tightest correlation can be seen between the temperature and humidity at the time of processing. The other correlations between dates of fabrication and processing and temperature and humidity at fabrication can also be seen. What the precise relationship is between the dates and the yield, adhesion, pinhole count and thickness was not determined. There are a number of other options that can be utilised which I had started to investigate. These looked in more detail at the contributions to each of the dependant variables. Figures 17,18,19 and 20 show some of these plots.

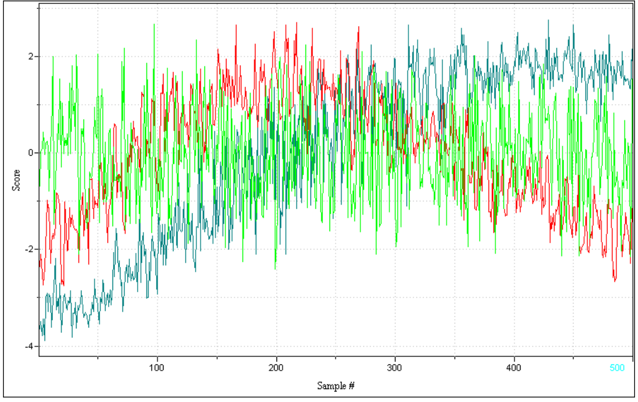

Figure 17 Data contribution to 3 principal components.

Figure 17 shows how the data contributes towards the three principal components. There appear to be two trends showing in this plot, one showing on the principal component 1, moving from negative to positive and the second one showing on principal component 2, which moves from negative to positive to negative.

The processing was taken on a stage further to try to produce a model that can be used to describe each of the dependant variables. A sample of the output is shown in Figure 18.

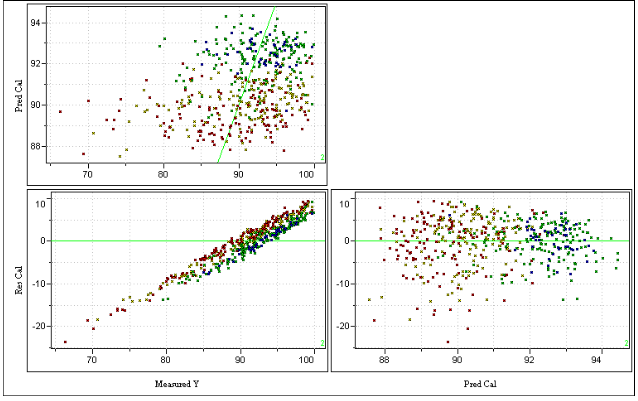

Figure 18. Using components to model the data.

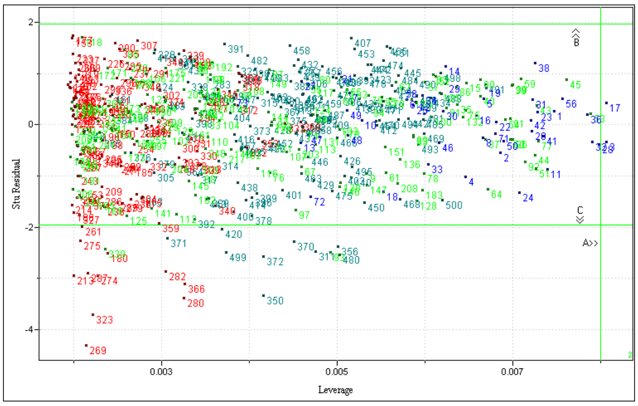

Figure 19. Investigating effects of individual samples

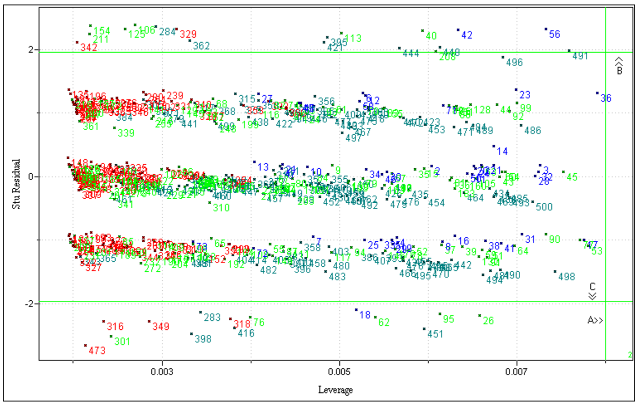

Figure 20. Some of the data was quantised.

Figures 19 and 20 have all the individual samples identified on the plots. Any of the points that fall outside the central box (boundaries A, B, C) are from samples that possibly had an undue influence on the outcome of the processing. The reason for identifying the samples is to allow checking of the input data to confirm that the entries are 'typical' and within the normal operating range. Any that are suspect can be highlighted and excluded allowing the data to be reprocessed to refine the model. Figure 19 suggests there is a skewed distribution to the data.

The temptation, now that I know the relationships, is to re-evaluate the data and imply they give information that I did not extract before having that knowledge. Without knowing the mathematics that has been used to manipulate the data the reality is the apparent skewing of the output could be an artifact of the processing and not of the input data.

Figure 20 shows the data having groupings in a series of levels. This is an output that is typical of a coarsely quantised variable. If the variable is thought to be one of the critical ones then this would indicate that either the measurement used on the variable or the control of the variable or both need to be upgraded.

## OBSERVATIONS.

The Pirouette system generally identified the important drivers for the underlying relationships. It didn't respond well to some input streams, notably time series data where process dynamics linked successive samples. There were some false positives.

There were limits on how much data could be handled, and that limit could be below the level where the causal links in a very noisy system could be elucidated.

Some of this difficulty is related to how the software is set up and used. In general using a dedicated machine with a high clock speed and as much RAM as possible made the processing much more robust. Some of the teething troubles we had were finding out what created difficulties. It was always good policy to "eyeball" as much of the data as practical to make sure there were no blanks or mistakes. The software has been improved to help in this process over recent upgrades. I have seen data sets that have had many tens of variables and tens of thousands of data points, which have run easily. It can be done, although this also had several man-months of effort dedicated to making this happen. There is no doubt that given the appropriate training the quality of the analysis of the data sets would have been improved.

Typically, each data set needed to be processed several times modifying the process each time to refine the output. The same is also true of the whole process in real life. Initially it is good policy to log data from as many process parameters as can be thought of, even the ones that everybody is convinced are irrelevant . Once sufficient data has been collected and processed decisions can be made as to which parameters are truly irrelevant and where data collection can be stopped. There are also likely to be others where it is indicated that there is insufficient control or measurement and there needs to be some process upgrade. Once this upgrade is done, the whole process starts again to further refine the process and the model. Once this point has been reached, the software can then be used to develop the process control algorithms.

This whole process of chemometrics is time consuming and intellectually taxing but the end result can be a full understanding of the whole of the manufacturing process. The tools are out there for us to be able to fully understand any manufacturing process from raw materials through to end product no matter how complex it is. The question is now is one of do we have the will to try?

What we found very encouraging was that with so little starting experience and without application of real statistical skills, the software pointed in the right direction each time. It is hoped that this will similarly encourage others to take up the challenge of chemometrics.

## REFERENCES

[1] Donald, J., Editorial in IEE Computing and Control Engineering Journal, February 1999, pp2-3:

[2] Pirouette 2.03. Infometrix Inc., Seattle, Washington.

[3] MATHCAD is a product of MATHSOFT, Cambridge MA. Version 8P was used for these trials.

[4] McCann, M.J. and Jones, D.P., "Web Coating Dynamic and Thermal Wrinkling Model", Proc. 41st Annual Technical Conference, Society of Vacuum Coaters, Boston, MA, April 1998.

[5] Internet website: http://www.mccannscience.com under applications, web coating (until 2000.04.30)