The value of pictures

A.M.C. Davies

Norwich NIR Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK.

I know that I have said it before, but it is something that is so important to the application of chemometrics that I will continue to repeat it. **You must look at the data.** In journalism it is said that a picture is worth a thousand words; in chemometrics a graph is worth a hundred statistics. Just to emphasise the general point, look at the graphs in Figure 1 each produced from 10 pairs of values for x and y.

The plots look very different but the statistics for the four sets A–D for: mean *x*, mean *y*, standard deviation *x*, standard deviation *y*, correlation, regression equation are **all the same!** They are a set of data invented by Frank Anscombe[1] to demonstrate the importance of putting data into graphs. At our recent Chemometrics for Beginners training course we were discussing how to define chemometrics and I was saying that nowadays I include Paul Geladi's "application of computer science"[2] in my original definition of "the use of mathematics and statistical techniques to aid the comprehension and utilisation of chemical and physical information". However, Ian Cowe, who was one of the instructors, proposed a quite different way of looking at it. He put chemometrics between data and output statistics (e.g. regression equations) and said "what we (chemometricians) do is mainly to look at pictures". Thank you Ian for that perception!

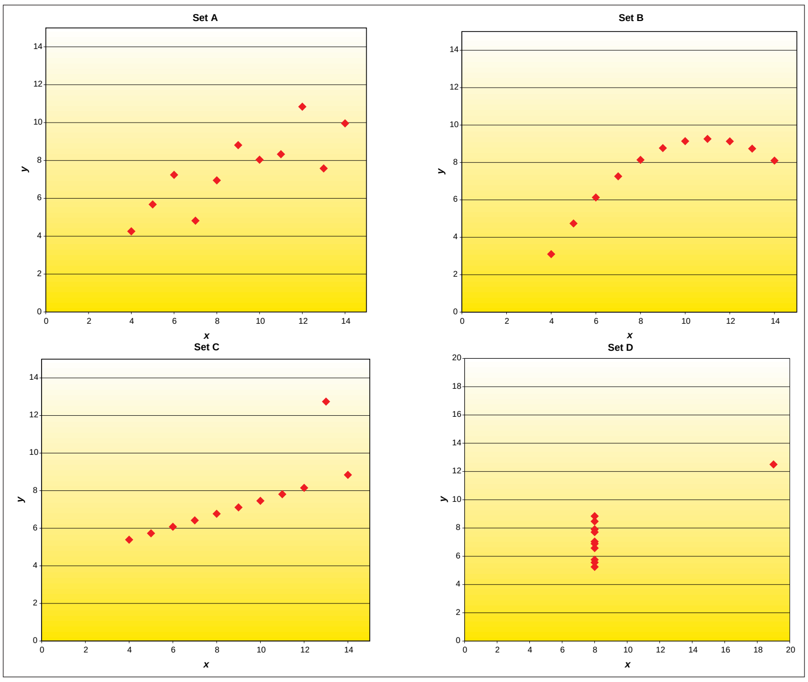


Figure 1. Four plots of *x*–*y* data sets A–D.

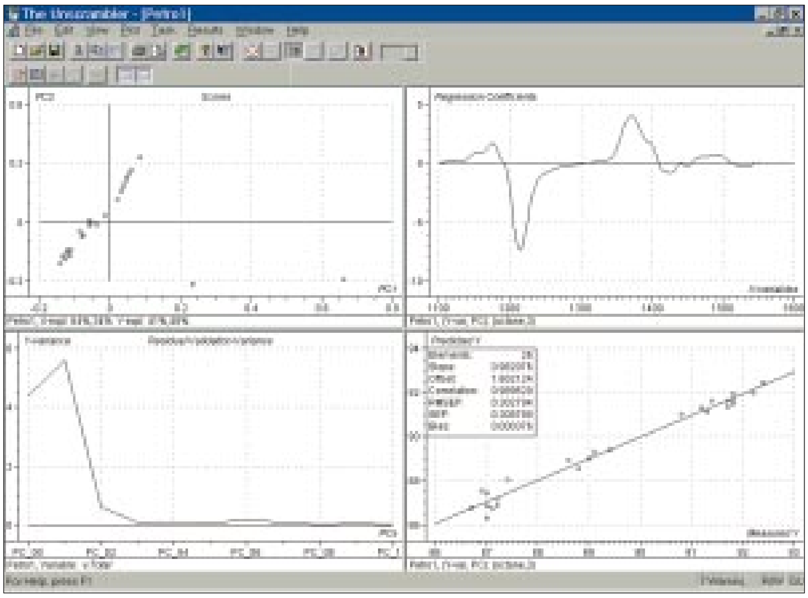


Figure 2. The standard Unscrambler regression overview.

Graphics for regression analysis

The specific topic for this column was suggested by a delegate on the training course when we started to look at calibration. "I thought you just looked at r^2 and *SEP*; why do we have all these graphs"?

The answer is to make sure we are not making any mistakes. PLS regression analysis is a complex process so we need a set of tools (graphs) to help us look at many different aspects. I am currently reviewing the new release (7.01) of the "Unscrambler" program (Camo AS, Oslo, Norway) and so these examples come from that package. I am using a small data set provided with the software, which consists of spectra of petrols and the goal is to produce a good calibration for octane number.[3]

When Unscrambler completes the calculation of a PLS or PCR you are

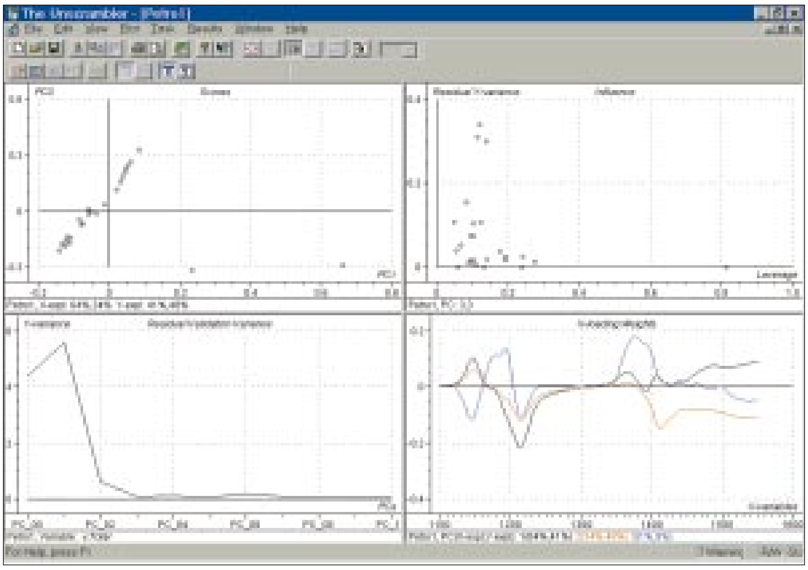


Figure 3. My preferred regression overview.

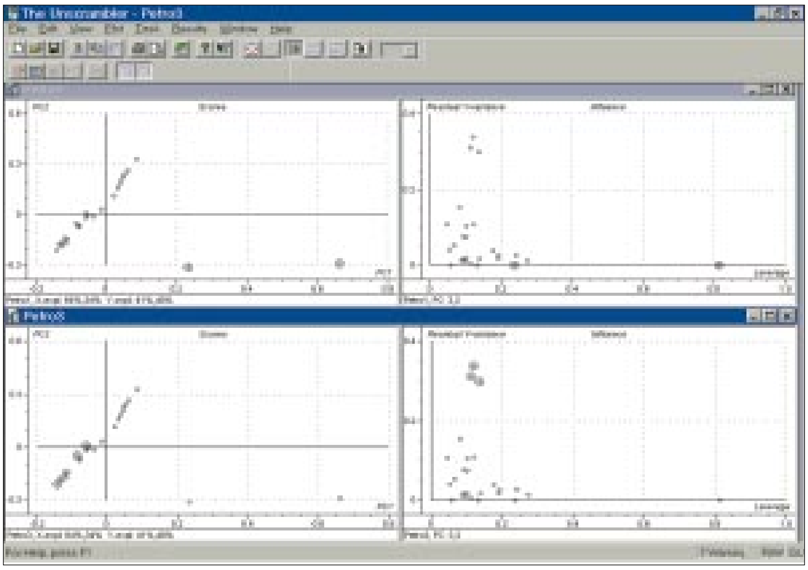


Figure 4. Use of the influence plot to indicate samples in the factor space which have high leverage (upper) or large residuals (lower).

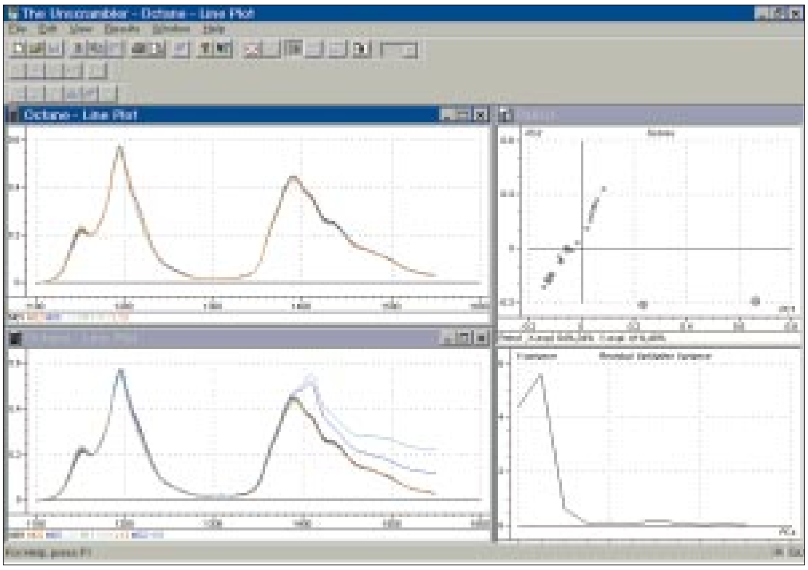

The Unscrambler - Octane - Line Plot
Octane - Line Plot
Scores
Residual Validation Variance

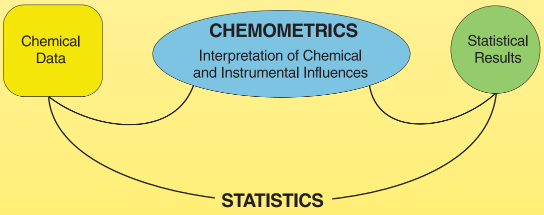
Figure 5. Some of the spectra plotted with and without the outliers in the scores plot identified as the probable cause of the anomalous increase in *y*-variance.

presented with four graphs as shown in Figure 2. They are a scores plot (factor 1 or PC1 versus factor 2 or PC2), a regression coefficient versus wavelength plot, an actual versus predicted plot (for calibration and validation data) and a residual standard deviation versus factors (or PCs) in the calibration. These are all useful plots but there are more, some of which are more important. I would much prefer not to have the actual versus predicted plot as one of the first four. It is impossible not to look at it first and get the r^2 and *SEP* values! Figure 3 shows what my preferred first four would be. The new ones are plots of loading weights for each factor (or PC) used in the optimum regression equation and an influence plot. An influence plot shows what contribution each sample is making to the calibration. It is a plot of residual *y* variance versus leverage, which is a statistic measuring the relative importance of each sample to the calibration. So let's look at the information they offer us.

We begin by a "global" view of all four. Three of the four are showing abnormal or undesirable distributions. The scores plot looks very one sided, which is caused by two outliers; the influence plot shows one sample with very high leverage and several with high variance and the residual *y* variance plot actually shows an **increase** in *y* variance with the first factor. There is something clearly unhealthy about this calibration. Yet if we had caught a sneak view of the predicted versus actual plot, the statistics looked quite acceptable. I am not going to suggest that there is any "correct" order for the detailed look at these graphs but this is what I did. I started with the scores and influence plots. Are the same samples outliers in both plots? The outliers in the scores plot could result in samples with high variance or high leverage. Unscrambler has a very convenient feature whereby you can mark a sample in one plot and it is marked in all other plots of samples (by a circle around the point). So we can mark the outliers in the scores plot and see where they are in the influence plot and as shown in the upper part of figure 4. The most extreme sample in the scores is the one with the highest leverage in the influence plot. The second outlier also has quite high leverage so where are the samples with high *y* residual? Move to the influence plot and mark the three samples with the largest residual *y* and we see in the lower part of Figure 4 that they are (reasonably) randomly distributed in the bulk of the samples. At present we have been looking at the influence plot for the third factor. If we

Ian Cowe's explanation for the difference between statistics and chemometrics

The aim of a Statistician is to reduce the information in a set of data to a small number of statistical variables, which summarise the relationship between various sets of data. He may have no knowledge of chemistry or instrumentation.



A Chemometrician brings knowledge of the chemical and sometimes the instrumental influences, which affect the data. The aim here is often to display the data in ways that allow chemical interpretation of the system. This may involve transforming the data in ways which "bring out" features which were not evident from the raw data or deriving new variables which are functions of the original data. Chemometricians make their living by "adding value" to the process of statistical analysis by bringing in skills which Statisticians lack.

Ian Cowe is the Chemometric Projects Coordinator at Foss Electric Development (UK) Ltd. Ian is particularly known for his work in introducing Principal Components Analysis into NIR spectroscopy.

look at the first and second factor we find that the same sample always has high leverage and this suggests an explanation for the fact that our residual standard deviation plot shows an increase after the first factor. This may be due to the fact that this sample has such an effect on the calibration model that the prediction becomes worse than assuming all samples are the average value (this is the value for no factors in the model). If we had not done it earlier then it is time to look at a plot of the raw data. This can be displayed in a separate window and then tiled to compare them. Figure 5 shows a plot of some of the samples and a plot of the same samples with the two outliers. Clearly these samples are very different and they must be excluded from the calibration.

This analysis will be continued in my next column and we will see how the other plots (and some more) are used before we decide on a final calibration.

References

1. F.J. Anscombe, "Graphs in statistical analysis", *American Statistician* **27,** 17 (1973).

2. P. Geladi, "CHEMO in chemometrics", *Spectroscopy Europe* **9(4),** 31 (1997).

3. You can get a demonstration CD-ROM from Camo using this data, by visiting their web site: http://www.camo.no.