

More pictures from PLS regression analysis

A.M.C. Davies

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

In my previous column,¹ I looked at four plots of data, which are used to assess the success of a PLS calibration for the octane number of a set of petrol samples.² In this article we will see some more and hope that with this armoury of weapons we will be able to make some reasonable decisions.

In the previous article we had identified two outliers as shown in Figure 1. So now I can do the PLS calibration, using full cross-validation³ on the data without the outliers. Four views of the calibration are shown in Figure 2.

The scores plot (top left) shows a reasonable, if some what structured, distribution of samples. The residual validation variance plot (bottom left) shows a rapid drop for two factors (called PCs in the plot) and then a baseline until the 11th factor after which it begins to rise. The program suggests a conservative use of only two factors and with a rather small sample set this seems reasonable. The predicted vs. measured scatter plot (bottom right) suggests that we have a reasonable calibration with no obvious outliers. There is only a small bias so the *RMSEP* (root mean square error in the prediction) and the *SEP* (standard error of prediction) are about equal and a satisfactory correlation, r ($r^2 = 0.976$, i.e. 97% variance explained). The x loading weights plot (top right) shows how the calibration works. In both factors there is a negative peak at 1218 nm, which does not correspond to a peak in the spectrum and this ought to be explained.

In Figure 3, I have plotted spectra from samples with low values of octane number (left) and spectra from samples with high or low octane number (right) for just the first peak in the spectrum, 101 data points. There is a clear separation between high and low samples around 1218 and this is being exploited by the calibration. This finding suggests that the data from the first peak might be sufficient for a calibration and so I have re-run the calibration just using the first 101 data points. The results are shown in Figure 4.

The left-hand side graphs show how the loading weights from the individual

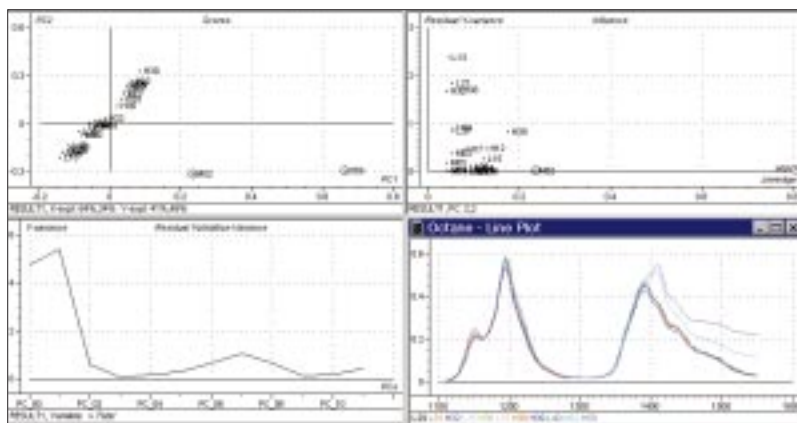


Figure 1. Discovery and identification of two outliers in the octane data set.

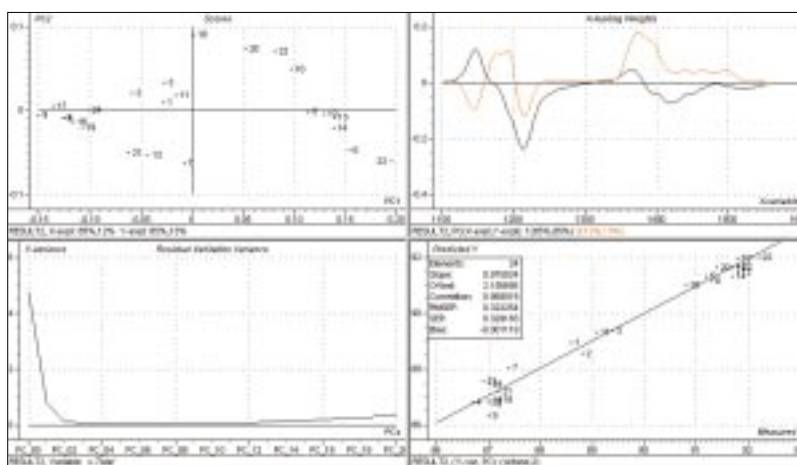


Figure 2. PLS calibration of octane data with two outliers removed.

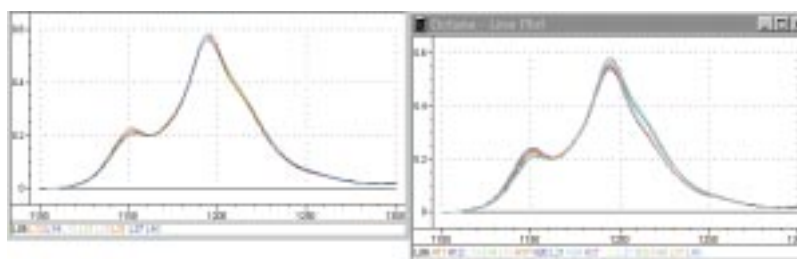


Figure 3. Spectra of low octane value samples (left) and high and low octane value samples (right).

factors are translated into the regression coefficients that are applied to the spectral data. The right-hand side graphs indicate that this calibration is equivalent to the previous one. An alternative, and useful, way of looking at predicted results is to plot the error against the predicted. I have done this for both calibrations in Figure 5.

Note that there is a scale difference between the plots (outside my control). Both plots appear to indicate an improvement in prediction with increased octane value but there is little to choose between the two. Sometimes these plots will show very interesting shapes, which indicate that additional, or different processing of the data is required. As the short range calibration is using 101 instead of 226 variables it would be preferred if we have no additional information. We can obtain some additional information because we have a validation set of samples. WE SHOULD ALWAYS HAVE A VALIDATION SET OF SAMPLES! The prediction results for these samples, produced by the two calibrations, are shown in Figure 6.

The results from Figure 6 indicate that the short-range calibration has bias, RMSEP and r^2 that are considerably worse than the full range calibration, which demonstrates the value of having an independent validation set. It is interesting to note that this sample set contains two samples that look similar to the outliers that were rejected from the calibration. However, neither calibration produces large errors for these samples.

Please note: This is a demonstration set of samples; "in real life" we would want about ten times as many samples.

References

1. A.M.C. Davies, *Spectroscopy Europe* **10(4)**, 28–31 (1998).
2. You can get a demonstration CD-ROM from Camo using this data, by visiting their web site: <http://www.camo.no>.
3. See A.M.C. Davies, *Spectroscopy Europe* **10(2)**, 24–25 (1998) for a discussion of cross-validation.

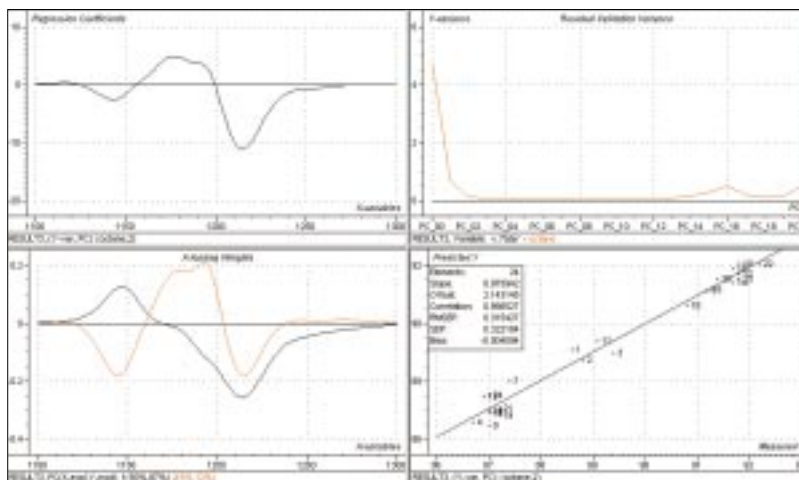


Figure 4. Calibration for octane number based on 101 data points.

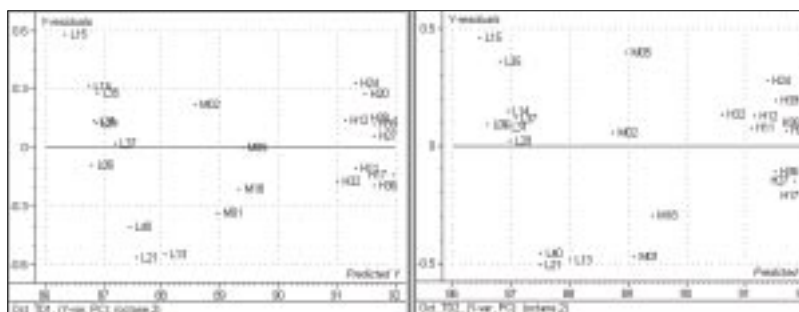


Figure 5. Distribution of errors for the full range (left) and short range (right) calibrations.

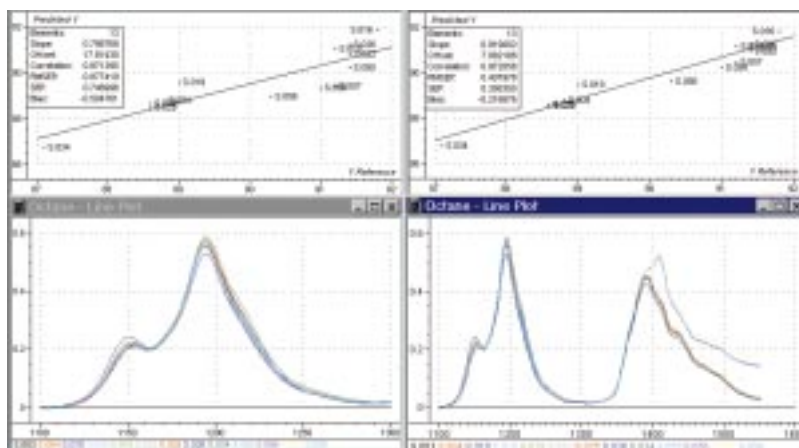


Figure 6. Prediction results for the two calibrations.