CAMO
Delivering Tomorrow's Innovations today

The Unscrambler®

Determination of caffeine in decaffeinated coffee by NIR spectroscopy

In the production of decaffeinated coffee the manufacturer needs to know how much caffeine is still in the coffee to ensure product quality. For this purpose they must be able to tell whether the caffeine concentration is below or above 0.1 %.

The usual way to determine the caffeine concentration is to use HPLC. This method is very accurate but has some disadvantages, eg it's very costly.

The coffee company wants to replace this method with a method that is faster, easier and cheaper to use. A method that fulfills these requirements is Near Infrared Reflectance (NIR) spectroscopy.

For these reasons the coffee company is interested in making a model, which can give the caffeine concentration of a given sample from the NIR measurements.

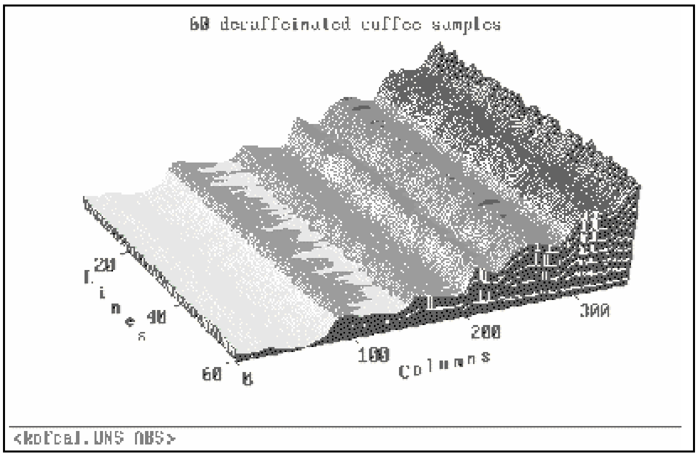


Fig 1 Raw data for 60 coffee samples

Input data

To make such a model several different coffee samples were collected and the caffeine content was measured with both the HPLC and the NIR method. Most of the samples are made from mixtures of different coffee sorts but some samples are made from only one sort. For the NIR measurement a Bran + Luebbe Infralyzer 500 instrument was used. The spectrum from 1100 nm to 2500 nm in steps of 4 nm giving a total of 351 wavelengths was recorded for each sample. The measurements are arranged in matrices, the x-matrix contains the spectra and the y-matrix contains the HPLC measurements that are regarded as the true concentration. Each line in the matrices corresponds to one sample, and the columns in x correspond to the wavelengths.

To be able to ensure the models' ability to predict future samples two sets of data were collected: one set with 60 coffee samples and one with 40 samples, a total of 100 samples. The samples in the two sets were collected so the expected variations in caffeine concentration and the variations in coffee sorts will be spanned.

Analysis

The model was originally developed by TNO-Nutrition and Food Research, in cooperation with Bran+Luebbe and the coffee company Marvelo Food Combany BV, the Netherlands, using **The Unscrambler** package from CAMO AS. To investigate the correlation in the data, two multivariate techniques were used: Principal Component Analysis (PCA) and PLS-regression.

To get a first impression of the data it is possible to use The Unscrambler®'s immense graphic functions. It can be used to visualize the measurements. The matrix plot in figure 1 shows the spectra for the 60 calibration samples.

We can see the caffeine content of the different samples using the Data - Edit menu. Below is shown a part of the figures.

Sample id	#	CAFFEINE
COLOM-P	1	0.150
COLOM-Q	2	0.200
COLOM-R	3	0.060
COLOM10	4	0.150
COLOM25	5	0.230
COLOM15	6	0.170
GEVA-33	7	0.040
GEVA-37	8	0.050
GEVA-42	9	0.060
...		
...		
BRASIL-2	59	0.035
GEVA-M 60	0.120	

Scatter Correction

Since the samples consist of ground coffee beans, we expect some scattering effects in the measurements due to different particle size and packing. To investigate if there are any of these effects we plot the individual spectra versus the average spectrum. As seen in figure 2 the lines do not follow the diagonal, i.e., there are some scattering effects. To handle the problems that arise from this we have two possibilities. We can either use Multiplicative Scatter Correction or take the second derivative of the spectra. MSC is a method designed to remove both additive and multiplicative noise effects in reflectance spectroscopy.

Because of the scatter problems the further analysis is done on the corrected data. The correction is done very easily in The Unscrambler from the Data - Transform - Multiplicative S. C. menu.

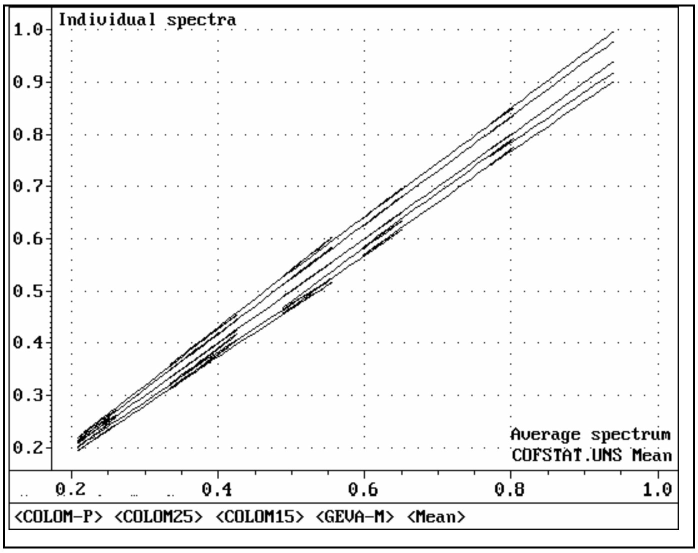


Fig 2 Individual spectra plotted against the average spectrum to detect any scatter.

Classification

First we run a PCA model on all the samples, both the calibration and the validation set, to see if they span the same variation. Furthermore we want to see if it is possible to recognize the different coffee sorts, especially to see whether the pure coffee samples are different from the others. For validation of the PCA model the quick leverage correction method is used, since we are only interested to find the similarities of the samples.

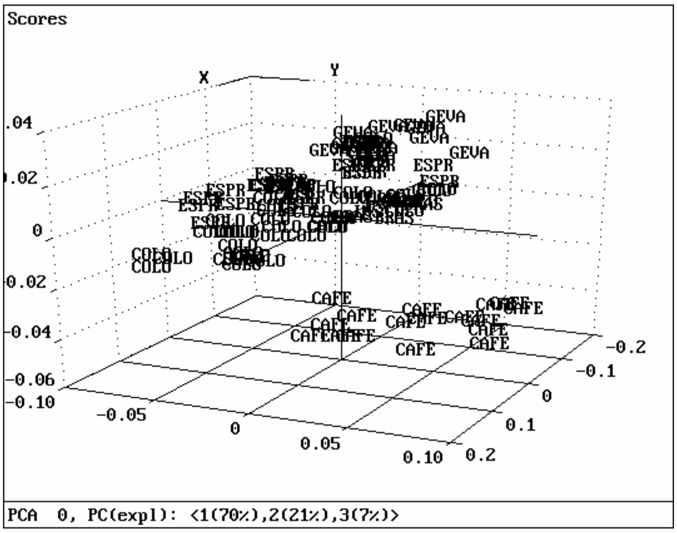


Fig 3 Scores for three first principal components

Scores

The score plot is a "map" over the characteristics of the samples. Samples placed in the same area of the plot are similar. Samples placed far from each other are dissimilar.

In figure 3 the scores from the first three components are plotted and there is a group of points that separates from the others. Since each point represents a sample this means that the samples are different from the other samples. If we examine the sample names, which identify the objects, we see that it is the samples made from the pure coffee that separates out. This means that we have to build one model for all samples that are blends of coffee and one model for the samples made from pure coffee. To keep these samples outside the calibration we go to the Model menu and use Remove objects.

Calibration

Now we are ready to make the PLS model from the spectra and the HLPC data. For this regression we use the test set validation. After a few seconds the model is ready and we use the ready-to-use model plots to examine the model named PLS 1.

Variance

The variance plot in figure 4 shows very clearly that we need eight principal components (the curve has a minimum), to model the caffeine content in the samples. We also see that the variance increases slightly at seven PCs. This is an indication of outliers, samples that are in some way dissimilar to the others or perhaps even erroneous. We have to investigate the model further to find out what causes the variance to increase.

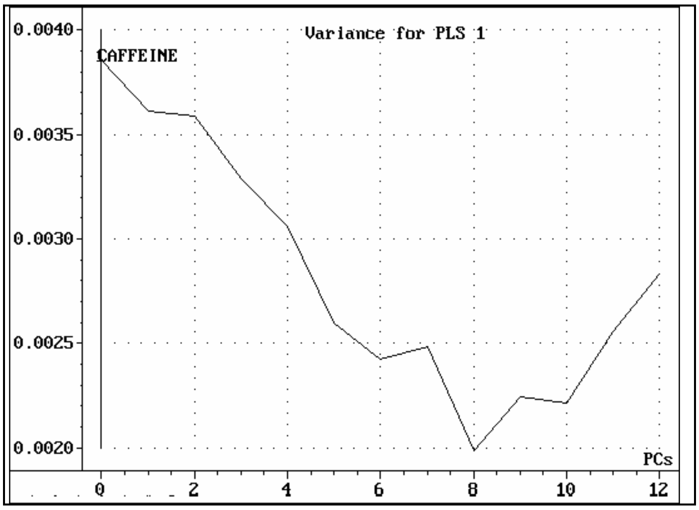


Fig 4 Residual variance. Eight pricinpal components is significant.

Outliers

The most likely reason for an increase in the variance is the presence of outliers. There are different plots we can use to find and identify these.

TU-plot

One is the TU-plot where the scores for the X-matrix T are plotted versus the scores for the Y matrix U. To get this plot we run the TvsU macro that comes with The Unscrambler. By studying this plot for all significant principal components, we identify two outliers; number 28 and 73 in PC no 4.

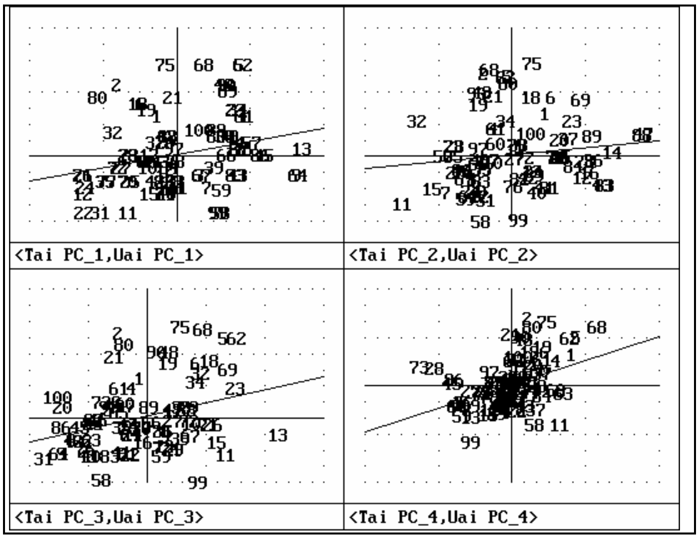


Fig 5 Scores from X-matrix plotted versus scores from Y-matrix in the TU-plot.

Also the score plot can be used to find outliers. In figure 6 the scores from PC number one and four are plotted, we see that there is a large group placed in the middle of the plot, the normal samples. Then there are some samples placed at a distance from the others. These are dissimilar to the "normal" samples, - they are either extreme values or erroneous in some way. We note their names but do not yet consider them as erroneous. First we check their caffeine content. This shows us that some of them, e.g., 68, just are extreme values - they are at the upper limit of the measured values. We see that 28 and 73 also show up here.

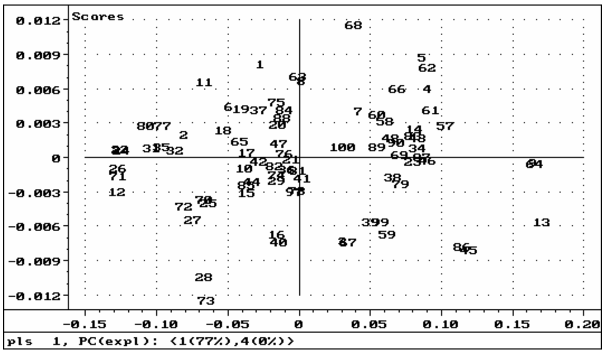


Fig 6 Scores for PC no 1 and 4.

Predicted versus measured

Yet another way to find outliers is to plot the predicted values versus the measured. The plot for a model with eight PCs in figure 7 shows us another outlier, number 80.

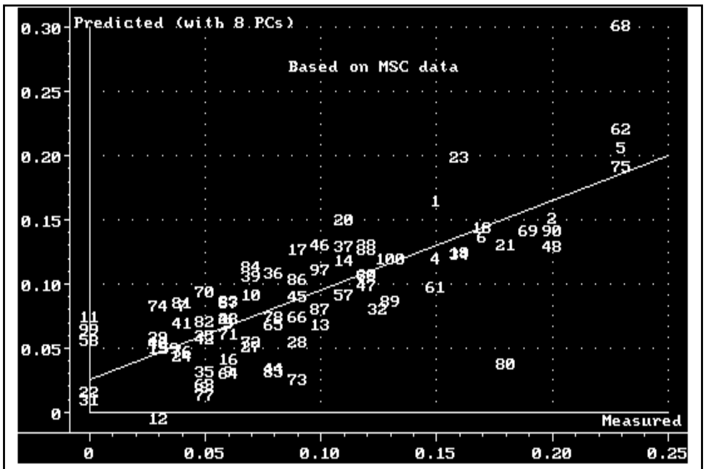


Fig 7 Predicted vs. measured caffeine content with an eight PCs model.

Prediction ability

After these three outliers are removed, we recalibrate and now the increase in the variance has disappeared and simultaneously the prediction ability of the model has increased. To be able to measure the prediction ability we use the value of RMSEP (also called the SEE), root mean square of prediction. For the model with the outliers the program calculated it to be 0.037 and for the model without it was calculated to 0.035. This is within the accepted range of 0.03 - 0.04.

Since many spectrophotometers are not able to preprocess data with MSC we try to use the second derivative of the spectra. This model has the same outliers, but gives a worse prediction compared to the MSC pretreatment - the RMSEP was calculated to 0.038. This is still within the accepted range and might therefore be used.

Reduction of variables

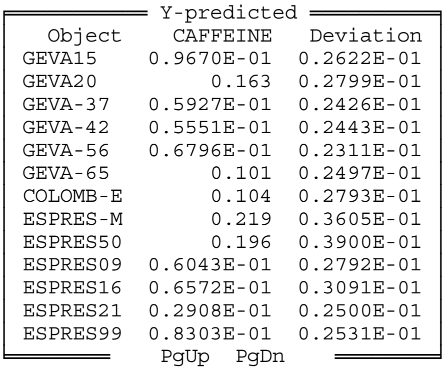
Traditional regression coefficients are also calculated so the relationships between the caffeine content and the spectra can be expressed as a regression equation on the form Y = b0 + b1*X1 + b2*X2 + ... +b351*X351. These coefficients may be exported back to the instrument software, and thus down-loaded to the instrument for on-line prediction.

The ones with the highest positive or negative values are the most important for the model, and could be used as a guideline to pick out a few of the wavelengths, eg. if we want to use a filter instrument instead.

Therefore let us try to make a model for the second derivative spectra with only the reflectance measurements at five wavelengths, number 67, 140, 202, 290 and 336. To keep the other wavelengths outside the calibration we go to the Model menu and weight them to zero. A model with two principal components gave a RMSEP of 0.037, better than the full spectrum model. This is because we have removed noisy wavelengths.

Routine prediction

To use the model to predict new samples we first have to read in a new data set. Then we go to the Prediction menu. In a few seconds the prediction is done and we can see the predicted values of the caffeine content:



```
═══════════ Y-predicted ═══════════
  Object     CAFFEINE    Deviation
GEVA15     0.9670E-01  0.2622E-01
GEVA20          0.163  0.2799E-01
GEVA-37    0.5927E-01  0.2426E-01
GEVA-42    0.5551E-01  0.2443E-01
GEVA-56    0.6796E-01  0.2311E-01
GEVA-65         0.101  0.2497E-01
COLOMB-E        0.104  0.2793E-01
ESPRES-M        0.219  0.3605E-01
ESPRES50        0.196  0.3900E-01
ESPRES09   0.6043E-01  0.2792E-01
ESPRES16   0.6572E-01  0.3091E-01
ESPRES21   0.2908E-01  0.2500E-01
ESPRES99   0.8303E-01  0.2531E-01
═══════════   PgUp   PgDn   ═══════
```

It is of course possible to see the prediction as a plot in the Plot - Predicted menu. In figure 8 the predicted values are plotted with the uncertainty shown as bars.

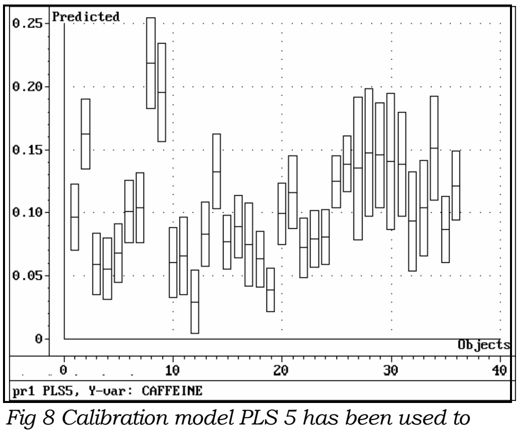


Fig 8 Calibration model PLS 5 has been used to predict the caffeine content in 34 new samples.

This uncertainty is calculated from the similarity of the samples to the calibration samples: the more similar they are, the smaller is the uncertainty. In this plot some samples have larger bars than the others, these are samples number 28-34. This is because these samples are coffee of the Cafei type and the model was made from a calibration set that didn't include samples of this coffee type (we removed them).

Conclusions

- By using the Unscrambler it was possible to first classify samples and then make a model that made it possible to predict the concentration of caffeine with very good accuracy.
- The model has been used with the instrument for routine quality control in more than a year, with good results.
- Outliers were detected using the Unscrambler graphics that allow easy interpretation of data. We may find which samples and wavelengths are the most important for the model.
- It was very easy to find the number of significant principal components, and the validation feature helps to avoid overfitting and to estimate the prediction error expected in future predictions.
- From these results it is easy to optimize the model to get the best prediction ability. Furthermore it is simple to select fewer wavelengths that can be used with a filter instrument.
- The best model was based on the MSC pretreated data but since the instrument used in this application cannot do MSC, we tried to use the 2nd derivative of the spectra, which gave almost as good a model.
- It is possible to download the model to the spectrophotometer for routine measurements, by expressing the model as a traditional regression equation.

The Unscrambler® is developed and marketed by

Camo AS
Olav Tryggvasons gt. 24, N-7011 Trondheim, Norway
Phone: (+47) 7351 4966, Fax: (+47) 7351 4257

US Office:
Camo Smart
1480 Route 9 North
Woodbridge, NJ 07095
Phone: 732-602-8886 x36