Determination of octane number from spectroscopic data

In many applications there is a need to determine non-chemical parameters from spectroscopy data. This is often considered a problem, since the spectroscopic methods were developed to reflect chemical variations in samples, where one looks for unique peak information to separate one constituent from another. The multivariate *PLS-regression* method is very effective at extracting variance information from complex, seemingly diffuse data, which we can utilize to relate octane number in gasoline samples to light absorbance in the Near Infrared wavelength range.

The following application demonstrates how UOP/Guided Wave Inc has successfully applied multivariate calibration techniques to estimate the octane number of refinery products from Near Infrared spectral data. This application high-lights the techniques made possible with The Unscrambler multivariate software package to develop calibration models for the lab and on-line. The data below were used in a feasibility study for one of their petro-chemical industry customers, where the control of octane number is required. The application is also of value for regulatory agencies responsible for verifying octane number in commercial establishments. Use of NIR analysis combined with multivariate calibration and prediction provides a large time savings over the traditional methodology for this analysis.

1. Problem

Make a model that predicts octane number from spectroscopic data. There are no selective wavelengths in the spectra, so univariate regression is not possible (see section 4).

2. Input data

To *make a model* we have prepared a training data set (calibration set): For each of 26 representative gasoline samples (objects), that are considered to span the important variations, we have recorded NIR absorbance spectra at 226 wavelengths (X-variables no. 1 - 226) and octane number (Y-variable). They are stored in the matrices XTrain and YTrain.

To *predict* (estimate, determine) octane number in new samples, we have 14 gasoline samples with absorbance readings (226 wavelengths), but with unknown octane number. They are stored in the matrix XNew. We will use the model to predict the octane numbers of these samples.

3. Plot raw data

We use The Unscrambler software package for multi-variate analysis and graphical presentation. By using the **Matrix plot** facility, plotting XTrain, we can study the spectra for all the samples. Scaling shows that the clearest peak is 1194 nm.

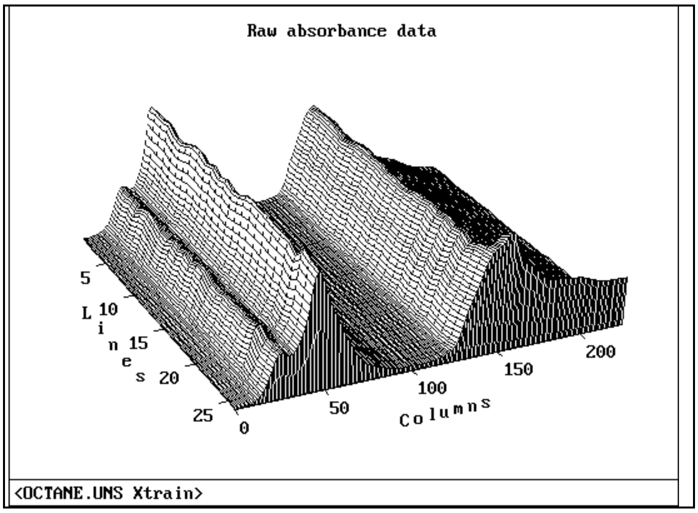


Fig 1 Raw data absorbance spectrum for 26 gasoline samples

4. Univariate regression

The **General 2-vector plot** lets us try a univariate regres-sion by plotting the clearest absorbance peak (X48 = 1194 nm) versus measured octane number (Y). The regression is not at all suitable for prediction.

5. Multivariate regression

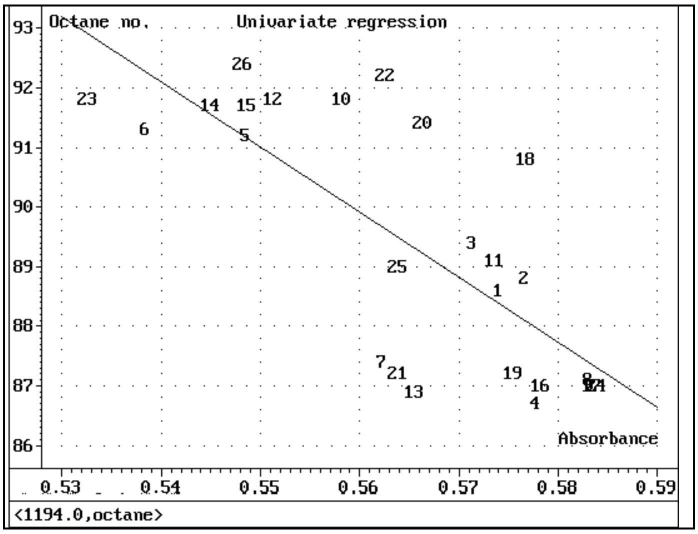
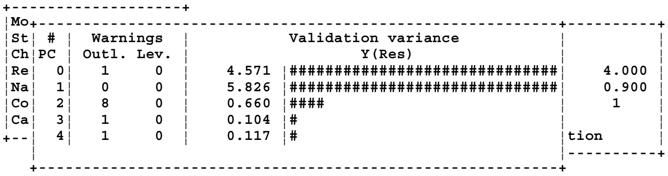


Fig 2 Univariate regression gives a bad prediction, although we use the clearest peak.

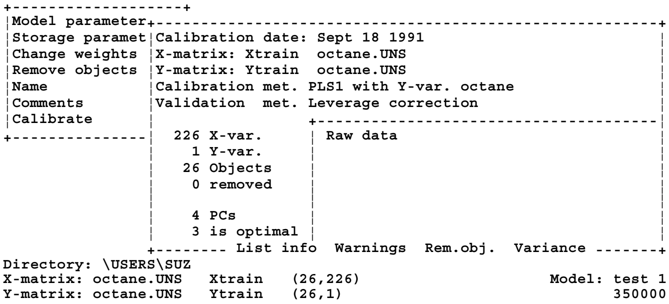
We read the training data set into the program. In the **Model menu** we choose regression method and model parameters. We choose PLS (Partial Least Squares, since the information in Y-variables is important for the decomposition of the X-matrix) and a quick validation method – Leverage correction – to make the first model.

The calibration output screen (below) gives an overview of the generated model; outliers and prediction error (residual variance) after each PLS component (PC) (also see section 7):



```
+-------------------+
|Mo+---------------------------------------------------------+----------+
|St| #  |  Warnings   |               Validation variance      |          |
|Ch|PC  | Outl. Lev.  |                        Y(Res)          |          |
|Re|   0|   1      0  |     4.571 |##############################|    4.000 |
|Na|   1|   0      0  |     5.826 |##############################|    0.900 |
|Co|   2|   8      0  |     0.660 |####                         |      1   |
|Ca|   3|   1      0  |     0.104 |#                            |          |
+--|   4|   1      0  |     0.117 |#                            |tion      |
   |                                                           |----------+
   +-----------------------------------------------------------+
```

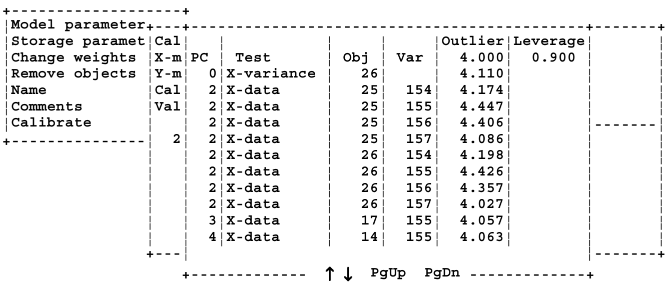
We also get an overview of the model; names, comments, data sets and model parameters used, optimal number of PCs, etc. When scanning the directory for models, this information is available to help us keep track of all models and data files.



```
+-------------------+
|Model parameter+----------------------------------------------------+
|Storage paramet|Calibration date: Sept 18 1991                      |
|Change weights |X-matrix: Xtrain  octane.UNS                        |
|Remove objects |Y-matrix: Ytrain  octane.UNS                        |
|Name           |Calibration met. PLS1 with Y-var. octane            |
|Comments       |Validation  met. Leverage correction                |
|Calibrate      |                 +----------------------------------|
+---------------|  226 X-var.     | Raw data                         |
                |    1 Y-var.     |                                  |
                |   26 Objects    |                                  |
                |    0 removed    |                                  |
                |                 |                                  |
                |    4 PCs        |                                  |
                |    3 is optimal |                                  |
                +-------- List info  Warnings  Rem.obj.  Variance -------+
Directory: \USERS\SUZ
X-matrix: octane.UNS   Xtrain   (26,226)                    Model: test 1
Y-matrix: octane.UNS   Ytrain   (26,1)                             350000
```

Model overview

The calibration output screen above indicated *outlier warnings* in several of the computed PCs, ie warnings for samples (objects) and/or variables that migth be abnormal. Via the menu in the model overview (above) we get a detailed list of the warnings:



```
+-------------------+
|Model parameter+---+-----------------------------------------+-------+
|Storage paramet|Cal|    |            |      |     |Outlier|Leverage|       |
|Change weights |X-m|PC  | Test       | Obj  | Var |  4.000|  0.900 |       |
|Remove objects |Y-m|   0|X-variance  |   26 |     |  4.110|        |       |
|Name           |Cal|   2|X-data      |   25 |  154|  4.174|        |       |
|Comments       |Val|   2|X-data      |   25 |  155|  4.447|        |       |
|Calibrate      |   |   2|X-data      |   25 |  156|  4.406|        |-------|
+---------------|  2|   2|X-data      |   25 |  157|  4.086|        |       |
                |   |   2|X-data      |   26 |  154|  4.198|        |       |
                |   |   2|X-data      |   26 |  155|  4.426|        |       |
                |   |   2|X-data      |   26 |  156|  4.357|        |       |
                |   |   2|X-data      |   26 |  157|  4.027|        |       |
                |   |   3|X-data      |   17 |  155|  4.057|        |       |
                |   |   4|X-data      |   14 |  155|  4.063|        |       |
                +---|                                             |-------+
                    +------------- ↑ ↓  PgUp  PgDn -------------+
```

Warnings from calibration model OCT ver 0

The list shows that objects number 25 and 26 are constantly pointed out as outliers (and also no 17, 14 in later PCs).

6. Graphical examination of outliers

Let's take a closer look at the objects! We leave the **Model** menu, open the **Plot**-menu and choose **Scores** plot which shows the projected locations of the objects onto the principal components, ie *which samples affect the model and how?*

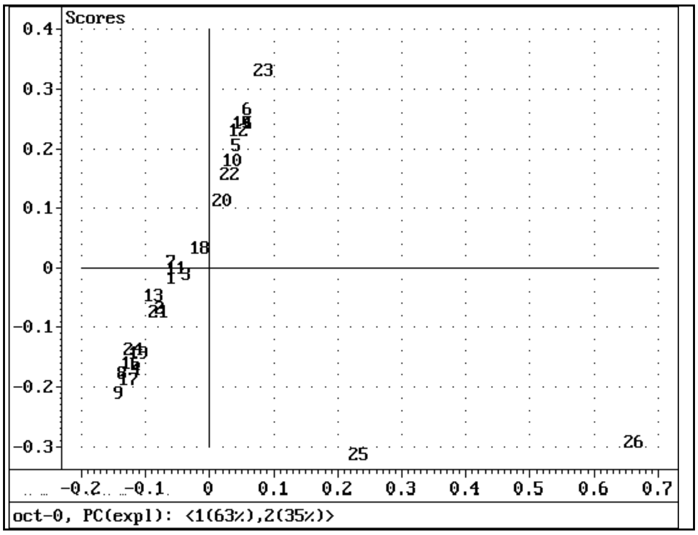


Fig 3 Scores for PC 1 versus PC 2.

The objects placed far from the origin have the most influence on the model. When plotting the scores for PC 1 versus PC 2 we can see that objects no 25 and 26 are placed in a group separate from the others. They also affect the model a lot (since they have high score values in PC 1 (which models most of the total variations). As indicated at the bottom of the plot, the two first PCs describe 63% + 35% = 98% of the total X-variance.

It seems reasonable to believe that objects 25 and 26 really are outliers - abnormal samples that give bias to the model and makes it useless for prediction. It is quite obvious that The Unscrambler has found errors in those two samples, and if you look closely at the matrix plot (Fig. 1), you may discover that the spectra for these samples deviate a bit from the others. The outlying samples in fact contain alcohol.

7. Recalibration with outliers removed

We will now perform a new calibration with objects no 25 and 26 removed from the calibration set. We go back to the **Model** menu, where we can mark these objects to be kept outside the calibration. We also change the validation method to the more conservative Cross validation. Then we start a new calibration run.

The calibration screen output (below) now shows no outlier warnings. The bars of #-signs indicate the prediction error (residual variance) after each PC. Their numerical values are given too. We also get fewer PCs this time; the two first PCs describe most of the total variations in Y. The variances are more easily studied by using the ready-to-use Variance plots.

```
+-------------------+
|Mo+----------------------------------------------------+-----------+
|St| # |  Warnings  |         Square error of prediction |           |
|Ch|PC | Outl. Lev. |                    Y               |           |
|Re|  0|  0     0   |     4.747 |##############################|     4.000 |
|Na|  1|  0     0   |     0.781 |#####               |     0.900 |
|Co|  2|  0     0   |     0.104 |#                   |    1      |
|Ca|  3|  0     0   | 0.81E-01 |#                    |           |
+--|  4|  0     0   | 0.66E-01 |                     |n          |
   |                                                  |-----------+
   +--------------------------------------------------+
```

It is however easier to study the results graphically, so let's go to the **Plot** menu again.

8. Interpreting the calibration model

Variance

Let's first look at the **Variance plot**, that shows how well the model describes the variations in the data. We can study the variance as Explained variance or Residual variance, for X-variables or Y-variables. Here is the Explained variance for the Y-variable (Octane number).

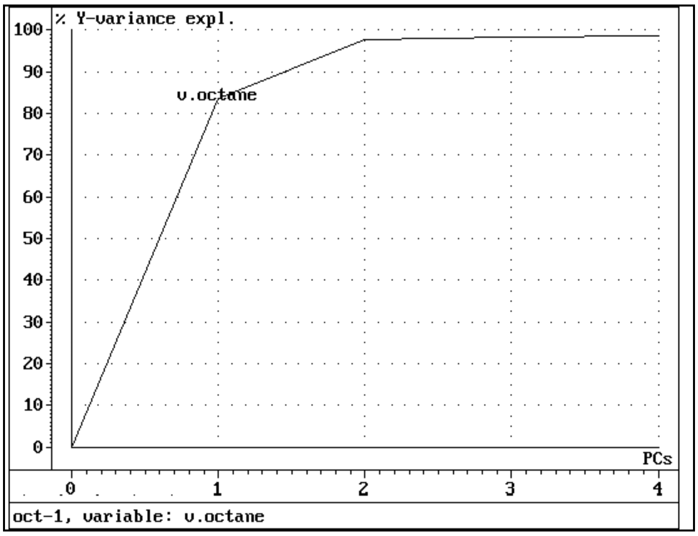


Fig 4Explained variance. Two PCs describe 98 % of the total variations in Y.

Based on this plot we choose how many PCs to include in the model. Generally we look for the number of PCs that minimizes the residual variance (maximizes the complementary explained variance), but without taking more than absolutely necessary, to ensure that we don't overfit (model noise). Two PCs explain 98 % while three PCs explain 99%.

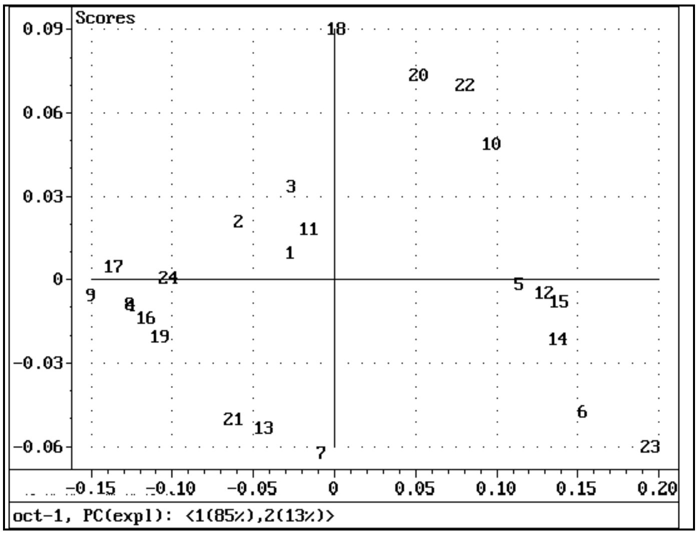


Fig 5 The combination of PC 1 and PC 2 describes a variation in the samples, seen as subgroups (encircled by hand).

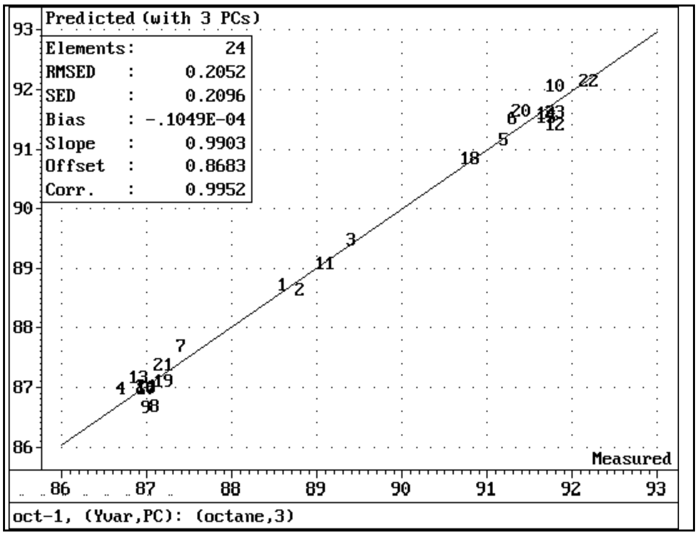
Scores

When plotting scores for the two first PCs we see no obvious outliers. However we see *subgroups*. By looking into the samples and their characteristics we may be able to interprete the meaning of the principal components. It seems for example that objects 1-2-3-11 have something in common. We can identify these groups according to their type of gasoline. (By naming the objects in a smart way, eg reflecting their composition or origin, we can sometimes see patterns more easily, since the program allows us to plot names instead of numbers if desired.)

Prediction ability

The **Pred/meas plot** (here with 3 PCs) shows the correspondence between the known octane numbers and octane number as predicted by the model.

We see here the same subgroups as in the Score plot! The groups represent samples with the same octane number. We can also plot Predicted vs measured using a two PC model. This gives a lower correlation, 0.995, why that model has a somewhat worse prediction ability.

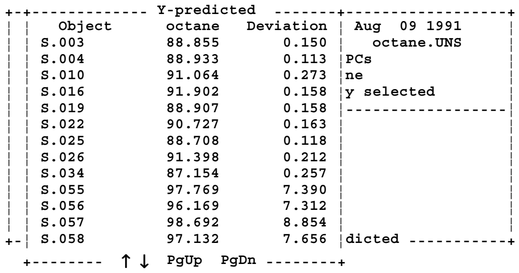

Predicted (with 3 PCs)

```
Elements:        24
RMSED    :   0.2052
SED      :   0.2096
Bias     : -.1049E-04
Slope    :   0.9903
Offset   :   0.8683
Corr.    :   0.9952
```

Fig 6 Predicted vs measured octane numbers with a three-PC model. Good correlation: 0.995.

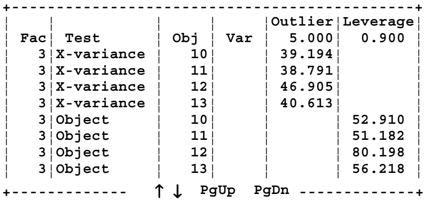
9. How to predict *new* samples?

We can now read a new data set, XNew, containing only absorbance readings for 14 new gasoline samples. We then open the **Predict** menu, enter the name of the model to use; OCT version 1, and how many PCs to use; 3. The prediction takes place immediately and information about the prediction run and numerical values of the predicted octane number show up in a window.



```
+-+------------- Y-predicted  -------+-------------------+
| |   Object      octane  Deviation | Aug  09 1991      |
| | S.003         88.855      0.150 |   octane.UNS      |
| | S.004         88.933      0.113 |PCs                |
| | S.010         91.064      0.273 |ne                 |
| | S.016         91.902      0.158 |y selected         |
| | S.019         88.907      0.158 |------------------ |
| | S.022         90.727      0.163 |                   |
| | S.025         88.708      0.118 |                   |
| | S.026         91.398      0.212 |                   |
| | S.034         87.154      0.257 |                   |
| | S.055         97.769      7.390 |                   |
| | S.056         96.169      7.312 |                   |
| | S.057         98.692      8.854 |                   |
+-| S.058         97.132      7.656 |dicted ------------+
  +--------  ↑ ↓  PgUp  PgDn --------+
```

We also get outlier warnings for objects 11, 12, 13, and 14.



```
+-----------------------------------------+
|   |           |     |     |Outlier|Leverage|
| Fac| Test     | Obj | Var |  5.000|  0.900 |
|   3|X-variance|   10|     | 39.194|        |
|   3|X-variance|   11|     | 38.791|        |
|   3|X-variance|   12|     | 46.905|        |
|   3|X-variance|   13|     | 40.613|        |
|   3|Object    |   10|     |       | 52.910 |
|   3|Object    |   11|     |       | 51.182 |
|   3|Object    |   12|     |       | 80.198 |
|   3|Object    |   13|     |       | 56.218 |
+-------------   ↑ ↓  PgUp  PgDn ------------+
```

We'll however plot them, since it is easier to evaluate graphs:
The **Predicted plot** shows the predicted octane number with *uncertainty limits*.

The model fails at predicting objects number 11 - 14. The program has once again detected erroneous samples. (This was also verified by the person who prepared the samples. They contained alcohol.)

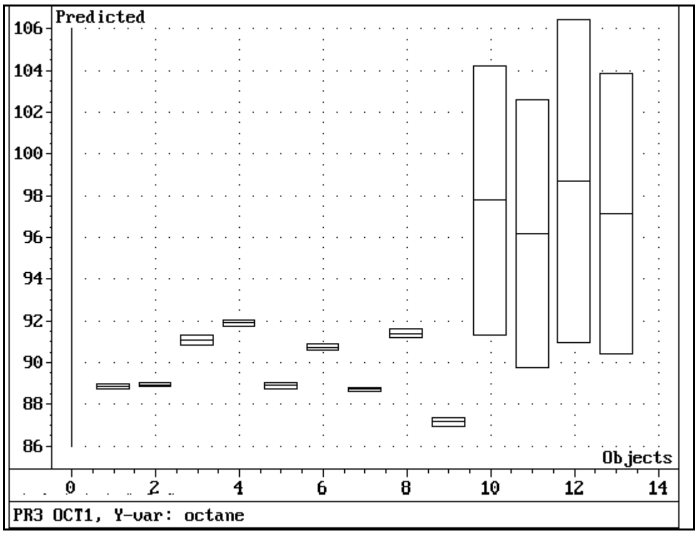


Fig 7 Calibration model OCT ver 1 has been used to predict octane numbers in 14 new samples.

10. The PLS regression model

The PLS regression model relates a set of *X-variables* (here spectra) to a set of *Y-variables* (here octane numbers). This is accomplished through a set of abstract latent variables called PCs or principal components. Each *PC* represents one systematic variation in the data.

The value of each PC for each sample is called a *score*. The *loadings* are the regression coefficients from each variable to each PC. The matrix equations used to relate these terms are

Y = TP + E
X = TQ + F

where T = PC scores, P = X-variable loadings, Q = Y-variable loadings, E = X-residuals (error), and F = Y-residuals.

Once the model is complete, it may be used to determine the Y-variables only based on the X-variables. The regression model is best interpreted and examined by using *graphical* presentation of the terms, as we have seen in this application note.

11. Traditional regression model

However, The Unscrambler program also calculates the B-coefficients which can be used to express the relations between X and Y as the more commonly known regression equation;

Y = B0 + B1*X1 + B2*X2 + ... + BN*XN

This equation is often implemented in for example on-line prediction models with spectroscopy instruments or other measurement instruments.

The B-coefficients can be read from a plot or a table:

12. Conclusions

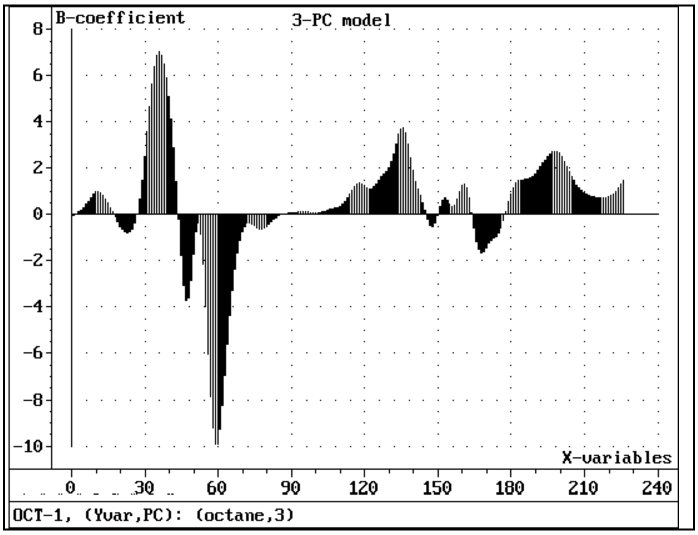


Fig 8B-coefficients using 3 PCs

By using PLS with The Unscrambler software package we were able to make a calibration model that gave *very accurate predictions* of the non-chemical parameter Octane number from spectroscopic data. The program automatically *detected erroneous samples.* The ready-to-use plots enabled a *visualization* of the calibration model, making interpretation easier.

The Unscrambler® is developed and marketed by CAMO Process AS,
Nedre Vollgate 8, N-0158 Oslo, Norway. Tel: (+47) 2239 6300, Fax: (+47) 2239 6322, E-mail: camo@camo.no web page: www.camo.no