periodica polytechnica

*Chemical Engineering*
*57/1–2 (2013) 15–20*
*doi: 10.3311/PPch.2165*
*http://periodicapolytechnica.org/ch*
*Creative Commons Attribution*

RESEARCH ARTICLE

# Partial Least Squares Model based Process Monitoring using Near Infrared Spectroscopy

*Tibor* Kulcsár / *Gábor* Sárossy / *Gábor* Bereznai / *Róbert* Auer / *János* Abonyi

Received 2012-09-26, accepted 2013-02-27

***Abstract***

*On-line analyzers are widely used in chemical and oil-industry to estimate product properties and monitor production process. Partial Least Squares regression (PLS) is known as bilinear factor model as it projects input (***X***) and output (***Y***) data into low dimensional spaces. We present how this projection can be utilised in process monitoring and validation of on-line analysers. We apply the proposed methodology in a diesel fuel mixer where main product properties are estimated from near infrared spectra. Results show that the developed 2 Dimensional Partial Least Squares (2DPLS) model not only gives better property estimation performance than the currently applied Topological Near Infrared modelling tool (TOPNIR), but it is also able to provide informative map of operating regimes of the process.*

***Keywords***

*on-line analyser · PLS · multi-dimensional scaling (MDS) · near infrared spectrum*

***Acknowledgement***

*We acknowledge the financial support of the Hungarian State and the European Union under the TAMOP-4.2.2.A-11/1/KONV-2012-0071.*

*This work was presented at the Conference of Chemical Engineering, Veszprém, 2012.*

**Tibor Kulcsár**
University of Pannonia, Department of Process Engineering, Veszprém, H-8200, Hungary

**Gábor Sárossy**
**Gábor Bereznai**
**Róbert Auer**
MOL Ltd. Department of DS Development Analytics, MOL Hungarian Oil and Gas Plc. R&M Division, DS Development, POB. 1., Százhalombatta, H-2443, Hungary

**János Abonyi**
University of Pannonia, Department of Process Engineering, Veszprém, H-8200, Hungary
e-mail: janos@abonyilab.com

## 1 Introduction

Control of measured process values (e.g. temperature, pressure, flow rate) does not always ensure that product properties (e.g. density, cloud point, flash point) will be in desired ranges. Some of these properties in chemical and oil-industry are not measured online (e.g. cetane index, aromatic field, sulfur content) or not at the frequency necessary for real time control (e.g. flash point, density, cold filter plugging point). The objective of the development of software sensors and online analysers is to support the control of product properties which cannot be measured online or offline measurements would be expensive.

Interaction of signals like temperatures, pressures, flow rates or absorption intensities can be used for calculating unmeasured product properties (flash point, density etc.). Soft sensors are especially useful in data fusion, where measurements of different characteristics and dynamics are combined.

Near infrared spectroscopy is a widely used on-line measurement technique. There are several multivariate models and methods to support the prediction of product properties based on Near-Infrared (NIR) spectra. These methods can be separated into parametric models (e.g. linear regression, multi-linear regression, Partial Least Squares regression (PLS) ) and nonparametric methods (e.g. k-NN[1], False Nearest Neighbors (FNN), Neural Networks, Topological Near-Infrared Modeling [2, 3] - TOPNIR). The main difference between these two classes is that the nonparametric techniques cannot extrapolate.

The key idea of the paper is the utilisation of the multivariate signal of NIR analysers not only for building models to estimate product quality but also to use it in process monitoring and validation of models used in on-line analysers.

A PLS based prediction model has been developed to support both prediction and visalisation (monitoring)[4]. Datasets taken from the Dune Refinery of MOL Ltd were analysed. The PLS model is applied to estimate cold filter plugging point, density and one property of distillation. For monitoring the latent space of the PLS model is used. A special orthogonalisation algorithm was applied. The presented mapping is able to visualise the data and give information about the distribution of operating regimes and the quality of the model.

## 2 Spectroscopic Modeling

The main task of the spectroscopic modeling is to find relation between recorded spectra and relevant material properties, $\mathbf{y}_k = f(\mathbf{x}_k)$, where $k$ represents the index of the samples [5, 6]. Data driven identification of models require spectral databases. The first part of the database contains the recorded and preprocessed spectra, $\mathbf{X} = \left[\mathbf{x}_1^T \ldots, \mathbf{x}_N^T\right]$, where $N$ represents the number of samples available for model building. In our case the on-line ABB spectrometer records spectra in range 4000 - 4800$cm^{-1}$. The recorded spectra contains 195 equally distributed absorbance values in the recorded range, $\mathbf{x}_k = [x_{k,1}, \ldots, x_{x,n}]$, where $n = 195$.

The second part of the training set represents property values ($\mathbf{y}_k = [y_{k,1}, y_{k,2}, \ldots, y_{k,m}]$) as output variables of the prediction model. For model identification the set of $N$ samples of these properties are also arranged in a matrix form, $\mathbf{Y} = \left[\mathbf{y}_1^T \ldots, \mathbf{y}_N^T\right]$.

Figure 1 shows a spectral database which contains 651 samples.§

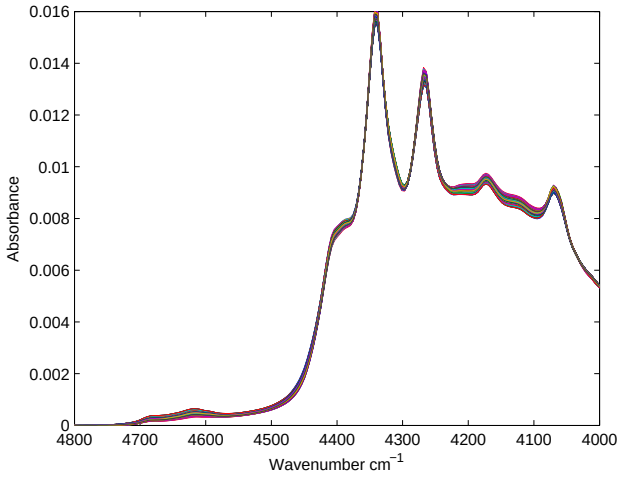



**Fig. 1.** $DS_1$ spectral database containing 651 spectra used for property estimation in a diesel fuel mixing process.

Since prediction model should provide good performance in the whole range of the operational regime of the process the development of an appropriate model requires properly distributed training set. Unfortunately Figure 1 does not give any useful information about the distribution of the data. To get more insight into the structure of the high dimensional spectral database visualisation techniques should be applied that are able to map the original $n = 195$ dimensional space into an easily visualisable two-dimensional map. In the following section such PLS based method will be presented.

### 2.1 PLS Concept

Partial least squares (PLS) is a perfect method for constructing predictive models from large number and correlated input variables [7].

PLS was developed in the 1960s by Herman Wold as an econometric technique, but soon it become widely applied tool of in chemical engineering [8]. In addition to spectrometric calibration, PLS is often applied to monitoring and controlling industrial processes; since compley process can easily have hundreds of process variables[4].

PLS tries to find the multidimensional direction in the $\mathbf{X}$ space the input variables that explains the maximum multidimensional variance direction in the $\mathbf{Y}$ space of the output variables. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among $\mathbf{X}$ values. By contrast, standard regression fails in these cases.

The general underlying model of multivariate PLS is

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2}$$

where $\mathbf{X}$ is an $n \times m$ matrix of predictors, $\mathbf{Y}$ is an $n \times p$ matrix of responses; $\mathbf{T}$ and $\mathbf{U}$ are $n \times l$ matrices that are, respectively, projections of $\mathbf{X}$ (the X score, component or factor matrix) and projections of $\mathbf{Y}$ (the Y scores); $\mathbf{P}$ and $\mathbf{Q}$ are, respectively, $m \times l$ and $p \times l$ orthogonal loading matrices; and matrices $\mathbf{E}$ and $\mathbf{F}$ are the error terms, assumed to be i.i.d. normal. The decompositions of $\mathbf{X}$ and $\mathbf{Y}$ are made so as to maximize the covariance of $\mathbf{T}$ and $\mathbf{U}$.

### 2.2 2DPLS based Visualization

For the two-dimensional visualization of the PLS model the algorithm developed in [4] was applied. In this subsection the most important details of this technique are summarized based on [4].

Two components that are informative for visualization may be obtained in several ways. One example is principal components of predictions (PCP), where in the scalar response case $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ normalization is used as one component, while residuals of $\mathbf{X}$ not contributing to $\mathbf{y}$ are suggested for use as the second component.

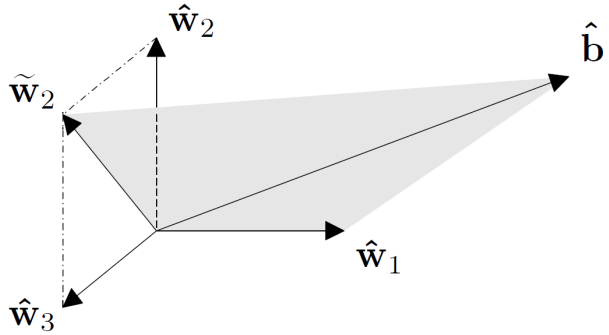



**Fig. 2.** 2D PLS mapping

The basic idea behind the applied mapping is illustrated in Figure 2. The estimator $\hat{\mathbf{b}}$ is found in the space spanned by loading weight vectors in $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_A]$ i.e. it is a linear combination of these vectors. It is, however, also found in the plane defined by $\hat{\mathbf{w}}_1$ and a vector $\tilde{\mathbf{w}}_2$ orthogonal to $\hat{\mathbf{w}}_1$, which is a linear combination of the vectors $\hat{\mathbf{w}}_2, \hat{\mathbf{w}}_3, \ldots, \hat{\mathbf{w}}_A$.

The matrix $\tilde{\mathbf{W}} = [\hat{\mathbf{w}}_1, \tilde{\mathbf{w}}_2]$ is thus the loading weight matrix in a two-component PLS solution (2PLS) giving exactly the same estimator $\hat{\mathbf{b}}$ as the original solution using any number of components. What matters in the original PLS model is not the matrix $\hat{\mathbf{W}}$ as such, but the space spanned by $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_A$. In the 2PLS model this represents the plane spanned by $\hat{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$ that is essential. Note that all samples in $\mathbf{X}$ (row vectors) in the original PLS model are projected onto the space spanned by $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_A$.

Samples may thus be further projected onto the plane spanned by $\hat{w}_1$ and $\tilde{w}_1$, and form a single score plot containing all $\mathbf{y}$-relevant information. When for some reason e.g. $\hat{\mathbf{w}}_2$ is more informative than $\hat{\mathbf{w}}_1$, a plane through $\hat{\mathbf{w}}_2$ and $\hat{\mathbf{b}}$ may be a better alternative. It will in any case result in a 2PLS model that gives the estimator $\hat{\mathbf{b}}$, as will in fact all planes through $\hat{\mathbf{b}}$ that are at the same time subspaces of the column space of $\hat{\mathbf{W}}$.

## 3 Application example

Presented research focuses to two tasks. The first task is the development of a prediction model that can estimate product properties based on spectra taken by online NIR analysers. The second task is the development a monitoring tool based on the visualisation of the same spectra [9].

Datasets collected at the Dune Refinery of MOL Ltd (Hungary) are analyzed. The first dataset ( "$DS_1''$) contains 651 samples collected from a diesel fuel mixing process. Approximately twenty material properties are estimated. The second data set ( "$DS_2''$) consists of 67 samples collected from a different process.

### 3.1 Prediction of product properties

The prediction performance of the models is measured by the correlation coefficient defined as:

$$\mathbf{R}(i,j) = \frac{\mathbf{C}(i,j)}{\sqrt{\mathbf{C}(i,i)\mathbf{C}(j,j)}} \tag{3}$$

where $\mathbf{C}$ is the covariance matrix and it's calculated as $\mathbf{C} = \mathrm{cov}(\mathbf{y}, \hat{\mathbf{y}})$.

All the presented algorithms including the $k$-nn algorithm that TOPNIR utilises have been implemented in MATLAB. Similarly to the global statistics feature of TOPNIR we calculated the basic measures of for the $k = 3$ case. As it can be seen results are a bit better than the global statistics of TOPNIR. Exact numerical reproduction of results was not possible since the documentation of software and related patent do not contain every details and tricks related to the calculation of distances. Table 1 shows that the $N$ number of the available samples differs for each properties. Among the 651 spectra only 560 were different and in most of cases only a fragment of the properties were measured.

Firstly the effect of dimensionality of latent space of the PLS model has been analysed (from 2 to 48 dimensions). To perform an adequate comparison leave-one-out and 10-fold cross validation technique was applied. On Figure 3 the performances (correlation coefficients)[10] of the PLS models are shown.

**Tab. 1.** Global statistics of data used for estimation

| Property | k-NN | $N$ | TOPNIR |
|---|---|---|---|
| Density | 0.983 | 441 | 0.971 |
| CI | 0.698 | 384 | 0.411 |
| CFPP0 | 0.972 | 229 | 0.964 |
| CFPP | 0.859 | 380 | 0.810 |
| CloudPt | 0.967 | 378 | 0.941 |
| FlashPt | 0.885 | 379 | 0.832 |
| T10 | 0.978 | 383 | 0.966 |
| T50 | 0.952 | 328 | 0.916 |
| T90 | 0.896 | 383 | 0.814 |
| E250° | 0.928 | 365 | 0.995 |
| E350° | 0.651 | 361 | 0.459 |
| E360° | 0.741 | 342 | 0.588 |
| PolyCycl | 0.741 | 331 | 0.559 |
| TotAro | 0.952 | 327 | 0.910 |
| VISC | 0.988 | 67 | 0.951 |

As it is shown in this figure, the accuracy of the model increases rapidly by increasing the dimensionality of the latent space from 2 to 6 dimensions, however, it reaches a maximum since when the complexity of the model is higher than the complexity of the modelled system.

**Tab. 2.** Effect of the number of latent variables to the performance of the model (correlation between the estimated and measured variables are shown).

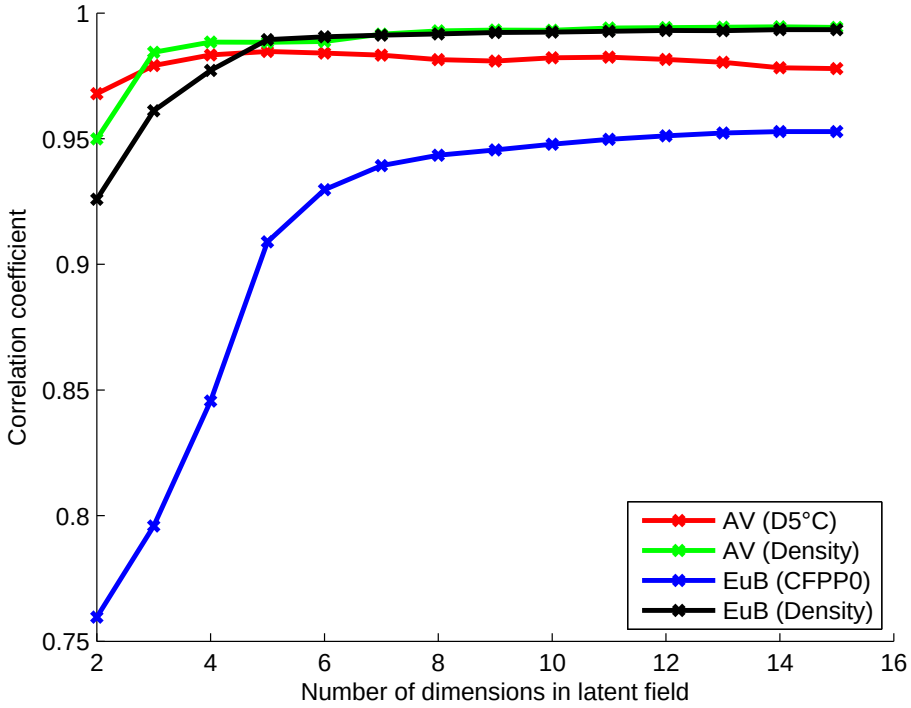
| | Latent dimensions | | | | | |
|---|---|---|---|---|---|---|
| Property | 2 | 6 | 12 | 18 | 24 | 48 |
| Density | 0.776 | 0.988 | 0.993 | 0.993 | 0.993 | 0.989 |
| CI | 0.130 | 0.204 | 0.190 | 0.420 | 0.344 | 0.272 |
| CFPP0 | 0.657 | 0.942 | 0.947 | 0.953 | 0.921 | 0.888 |
| CFPP | 0.516 | 0.755 | 0.769 | 0.728 | 0.703 | 0.610 |
| CloudPt | 0.668 | 0.924 | 0.950 | 0.958 | 0.955 | 0.943 |
| FlashPt | 0.408 | 0.596 | 0.878 | 0.901 | 0.895 | 0.854 |
| T10 | 0.428 | 0.732 | 0.908 | 0.946 | 0.941 | 0.938 |
| T50 | 0.694 | 0.922 | 0.970 | 0.971 | 0.957 | 0.910 |
| T90 | 0.432 | 0.654 | 0.849 | 0.895 | 0.868 | 0.796 |
| E250° | 0.660 | 0.879 | 0.955 | 0.954 | 0.927 | 0.904 |
| E350° | 0.044 | 0.077 | 0.308 | 0.259 | 0.174 | 0.006 |
| E360° | 0.115 | 0.374 | 0.431 | 0.397 | 0.341 | 0.190 |
| PolyCycl | 0.169 | 0.377 | 0.429 | 0.441 | 0.434 | 0.381 |
| TotAro | 0.765 | 0.885 | 0.905 | 0.880 | 0.862 | 0.771 |
| VISC | 0.898 | 0.991 | 0.999 | 0.999 | 0.999 | 0.999 |

### 3.2 Visualization of operating regimes

In section 2.2 a special method was presented that can map the PLS latent space into two dimensional space by orthogonal signal correction. This method has been compared with Principal Component Analysis [11] (PCA) and Topological Near-Infrared Modeling [2, 3] (TOPNIR) developed specifically to visualize NIR spectra and building topological prediction models with the help of resulted maps [12, 13].

TOPNIR uses nonlinear equation pairs (referred as aggregates) based on a small set of absorption values. Usually 4-6 characteristic wavelengths are selected to formalise a given aggregate that somehow reflects material property. To maximise

**Fig. 3.** Effect of PLS latent field's dimensionalit



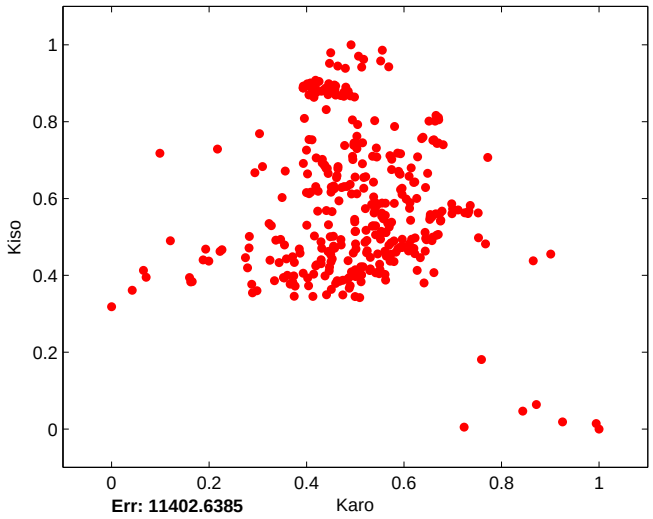




**Fig. 4.** Visualization of $DS_1$ using aggregates

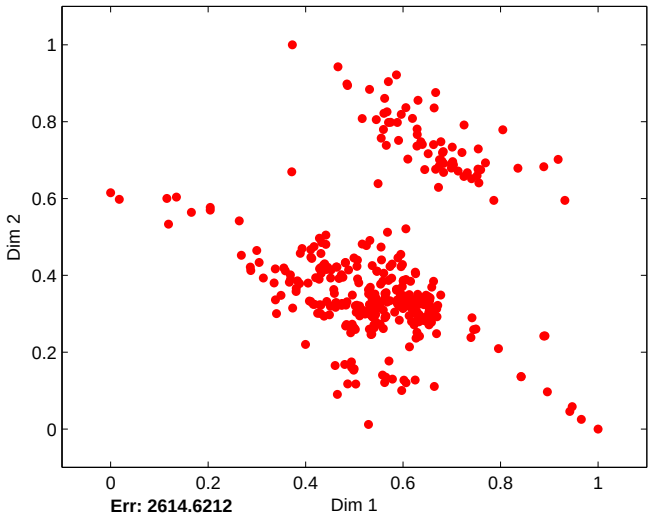



**Fig. 5.** Visualization of $DS_1$ using PCA

the information content of the mapping among 14 predefined aggregates the less correllated pairs were selected (Figure 4 ).

Figure 5 shows the mapping of PCA with the first two principal component [11]. This map is more informative. As it can be seen, the database contains samples from two different operating modes (summer and winter diesel) and some this mapping is able to separate these operating regimes.

Results of 2D PLS can be seen on Figure 6 and 7. The PLS model is more informative since it also utilizes output variables for the mapping. Figure 6 shows the mapping using the *Density* as output property. Comparing this mapping with the mapping of obtained using *CFPP* (see Figure 7) one can easily see that operating regimes have much more impact to the CFPP than density.

As it can be seen PLS correctly reflects the operating regions and much more able to detect outliers than aggregate based mappings.

In the second part of the case study we demonstrate how outlier samples can be identified in the mapped space. As it can be seen on the Figure 9 the $DS_2$ contains two samples which are really far from the normal operational range (top right corner). The aggregate based mapping can not identify these samples exactly, it finds only one outlier of two.

As it can be seen on Figures 10 and 11 the 2D PLS gives detailed information for outlier detection. Comparing these plots, TOPNIR based mapping ( Figure 8 ) and PCA ( Figure 9 ) it can be concluded that the 2PLS technique is the most efficient to detect outliers in the spectral or in the property space.

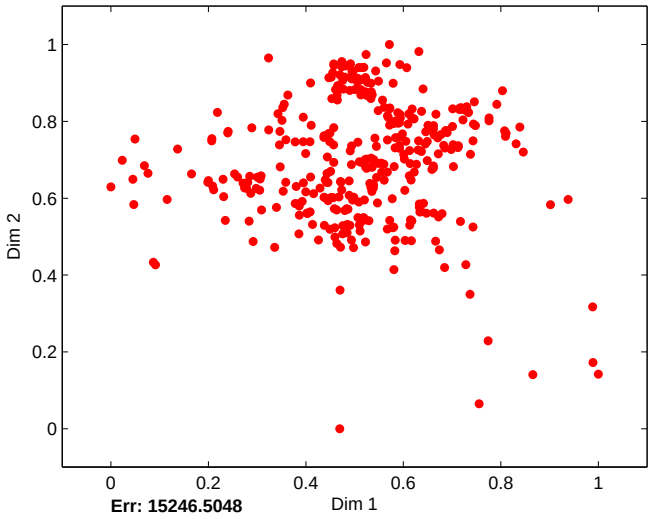



**Fig. 6.** Visualization of $DS_1$ using PLS (Density)

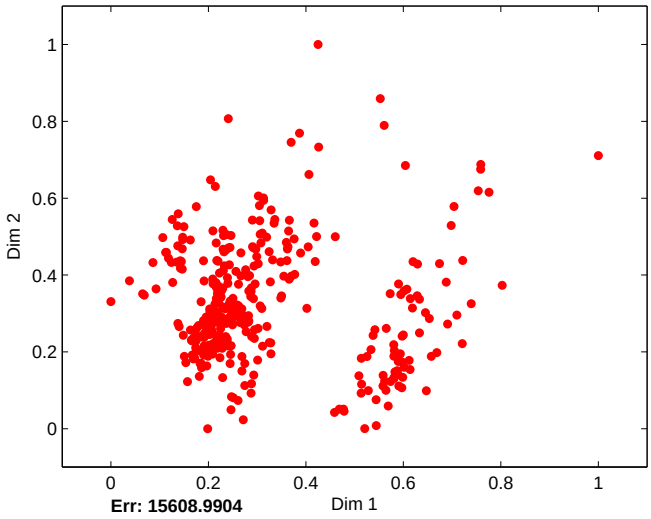



**Fig. 7.** Visualization of $DS_1$ using PLS (CFPP0)

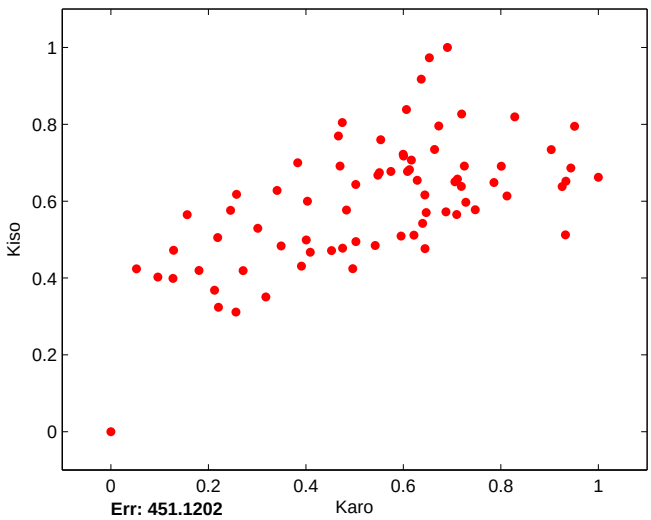



**Fig. 8.** Visualization of $DS_2$ using aggregates

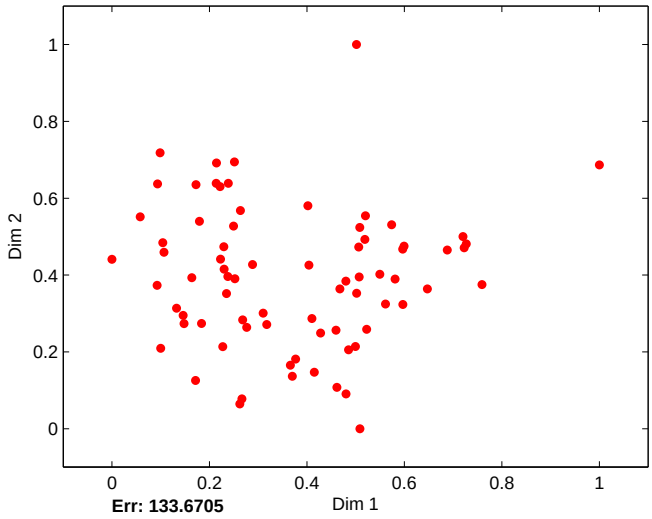



**Fig. 9.** Visualization of $DS_2$ using PCA

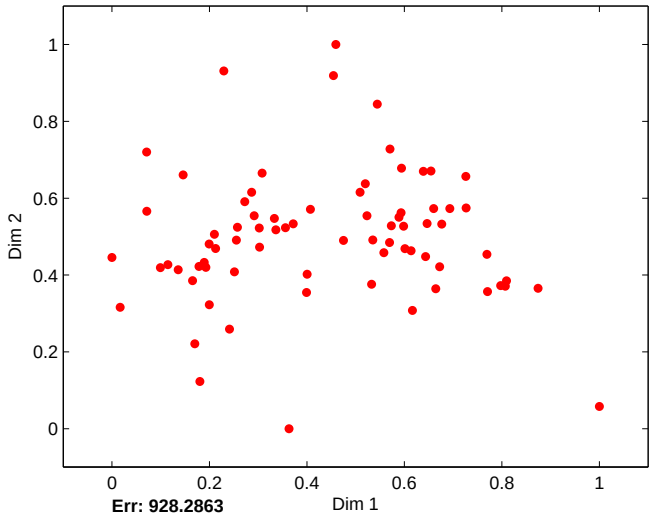



**Fig. 10.** Visualization of $DS_2$ using PLS (Density)

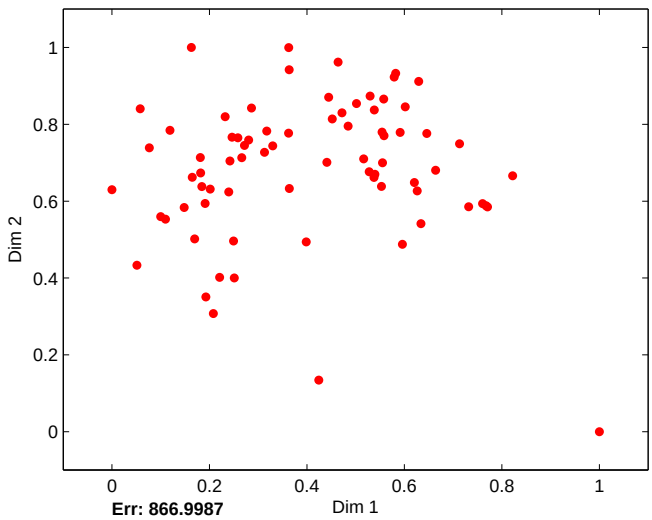



**Fig. 11.** Visualization of $DS_2$ using PLS (D5°C)

Such analysis gives the user information not only about operating regimes but also about quality of the models, so the presented mapping is able to give hints the modeller how to enhance model performance by the proper selection of the training data.

## 4 Conclusion

On-line analysers use indirect measurement combined with prediction model to support process control and monitoring. Near infrared spectroscopy is a widely used on-line measurement technique. There are several multivariate models and

methods to support the prediction of product properties based on NIR spectra. Model development cannot be a fully automatised process, human supervision and intervention is always needed. To support model development it is very informative to visualise the hidden structure of complex spectral database in a low-dimensional space. Industrial applications require easily implementable, interpretable and accurate projection. TOPNIR utilises heuristic nonlinear functions (aggregates) for the mapping of spectra as high dimensional object. We proposed a much more sophisticated approach that can be used simultaneous prediction and visualisation. We adapted a technique that allows the application of PLS also for visualisation of spectral database.

Datasets taken from the Dune Refinery of MOL Ltd were analysed. The PLS model is applied to estimate cold filter plugging point, density and one property of distillation. The main benefit of this technique is that it allows the extention of the operating region of the model by extrapolation. The proposed PLS based model is able to simultaneously predict unmeasured material properties and monitor the state of the process. Process monitoring is realized in orthogonal two dimensional plots. These plots can also be used for the effective identification of outliers.

## References


1 **Wu Y, Ianakiev K, Govindaraju V**, *Improved k-nearest neighbor classificationn*, Pattern Recognition, **35-1**, (2002), 2311-2318.

2 **Descales B, Lambert D, Llinas JR, Martens A, Osta S, Sanchez M, Bages S**, *Method for determining properties using near infra-red (NIR) spectroscopy*, Eutech Engineering Solutions, (2000), DOI US6.070.128.

3 **Sonbul YR**, *Topological near infrared analysis modeling of petroleum refinery products*, Saudi Arabien Oil Company, 2005, DOI US6.897.071 B2.

4 **Ergon R**, *Informative PLS score-loading plots for process understanding and monitoring*, Journal of Process Control, **14**, (2004), 889–897.

5 **MacGregor JF, Kourtl T**, *Statistical process control of multivariate processes*, Control Fag. Practice, **3**(3), (1995), 403–414.

6 **Browett WR, Stillman MJ**, *Computer-aided chemistry II. A spectral database management program for use with microcomputers*, Computers and Chemistry, **11**, (1987), 73–82.

7 **Kramer N, Boulesteix AL, Tutz G**, *Penalized Partial Least Squares with applications to B-spline transformations and functional data*, Chemometrics and Intelligent Laboratory Systems, **94**, (2008), 60–69.

8 **Wold H**, *Partial least squares*, In: **Kotz S, Johnson NL** (eds.), Encyclopedia of Statistical Sciences, Vol. 6, Wiley, New York, 1985.

9 **Kopanakis I, Theodoulidis B**, *Visual data mining modeling techniques for the visualization of mining outcomes*, Journal of Visual Languages and Computing, **14**, (2003), 543-589.

10 **Yamamotoa H, Yamaji H, Fukusaki E, Fukuda H**, *Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting*, Biochemical Engineering Journal, **40**, (2008), 199–204.

11 **Esbensen KH, Geladi P**, *Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice*, Comprehensive Chemometrics, (2009), 211–226.

12 **Blasco X, Herrero JM, Sanchis J, Martiez M**, *A new graphical visualization of n-dimensional Pareto front for decision-making in multiobjective optimization*, J. Information Sciences, **178**, (2008), 3908–3924.

13 **Greenacrea M, Hastieb T, Sanchis J**, *Dynamic visualization of statistical learning in the context of high-dimensional textual data*, Web Semantics: Science, Services and Agents on the World Wide Web, **8**, (2010), 163–168.