UNIVERSITEIT GENT

FACULTY OF ENGINEERING AND ARCHITECTURE

APPLICATION OF PARTIAL LEAST SQUARES REGRESSION FOR UNDERSTANDING AND PREDICTION OF FOULING IN THE TRANSFER LINE HEAT EXCHANGER OF A STEAM CRACKER

Nenad D. Ristic, Victor Abrahamsson, Kristina Franz, Kevin M. Van Geem, Guy B. Marin

Laboratory for Chemical Technology, Ghent University

http://www.lct.UGent.be

Budapest, Hungary, 26th of May, 2017

Outline

- **Introduction**
- **Experimental study**
- **Data structure**
- **Variable selection PLSR modelling**
- **Conclusions**

Steam cracking

Heart of a petrochemical plant

The main source of ethylene, propylene and other valuable hydrocarbons

Ethane

LPG

Naphtha

Gas Condensate

Atmospheric Gas Oil

Steam cracker

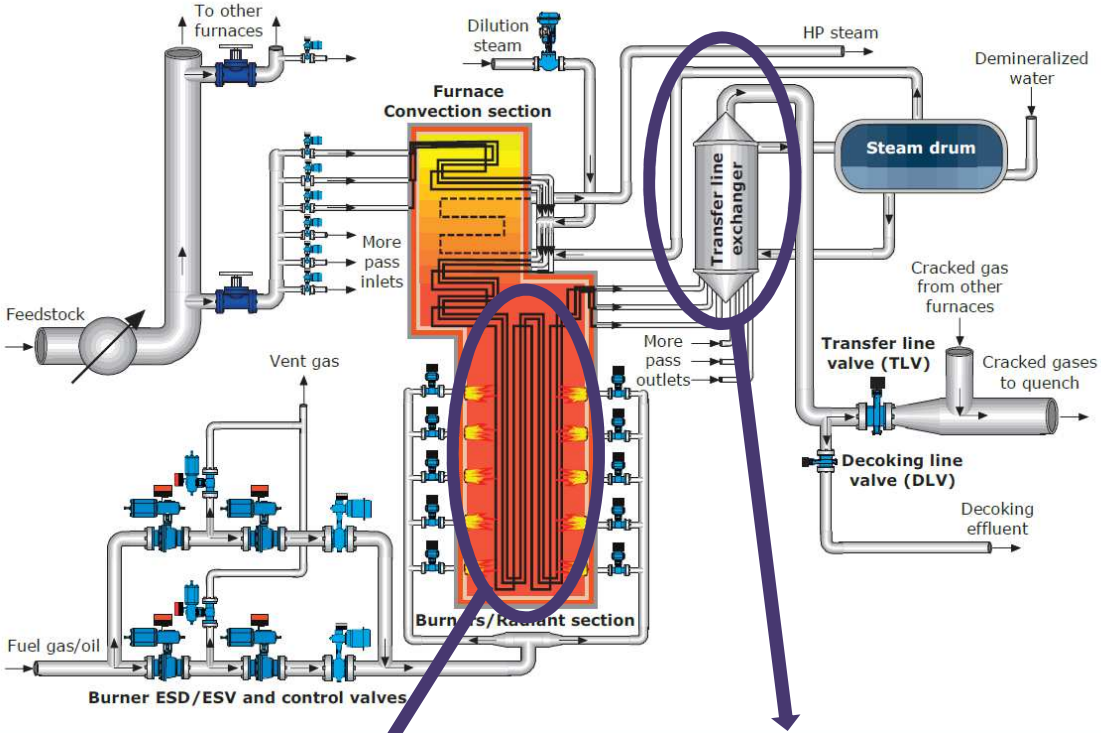
Ethylene

Propylene

1,3-Butadiene

BTX

Fouling



Fouling in the reactor coil and the ***TLE***

1. Thermal resistance
2. Pressure buildup
3. Decrease in selectivity towards olefins

Parameters influencing ***fouling***:

- Operating conditions, i.e. T, P, δ
- Material properties
- ***Feedstock composition***

DREAM

Prediction of fouling based on feedstock composition?

Outline

- **Introduction**
- **Experimental study**
- **Data structure**
- **Variable selection PLSR modelling**
- **Conclusions**

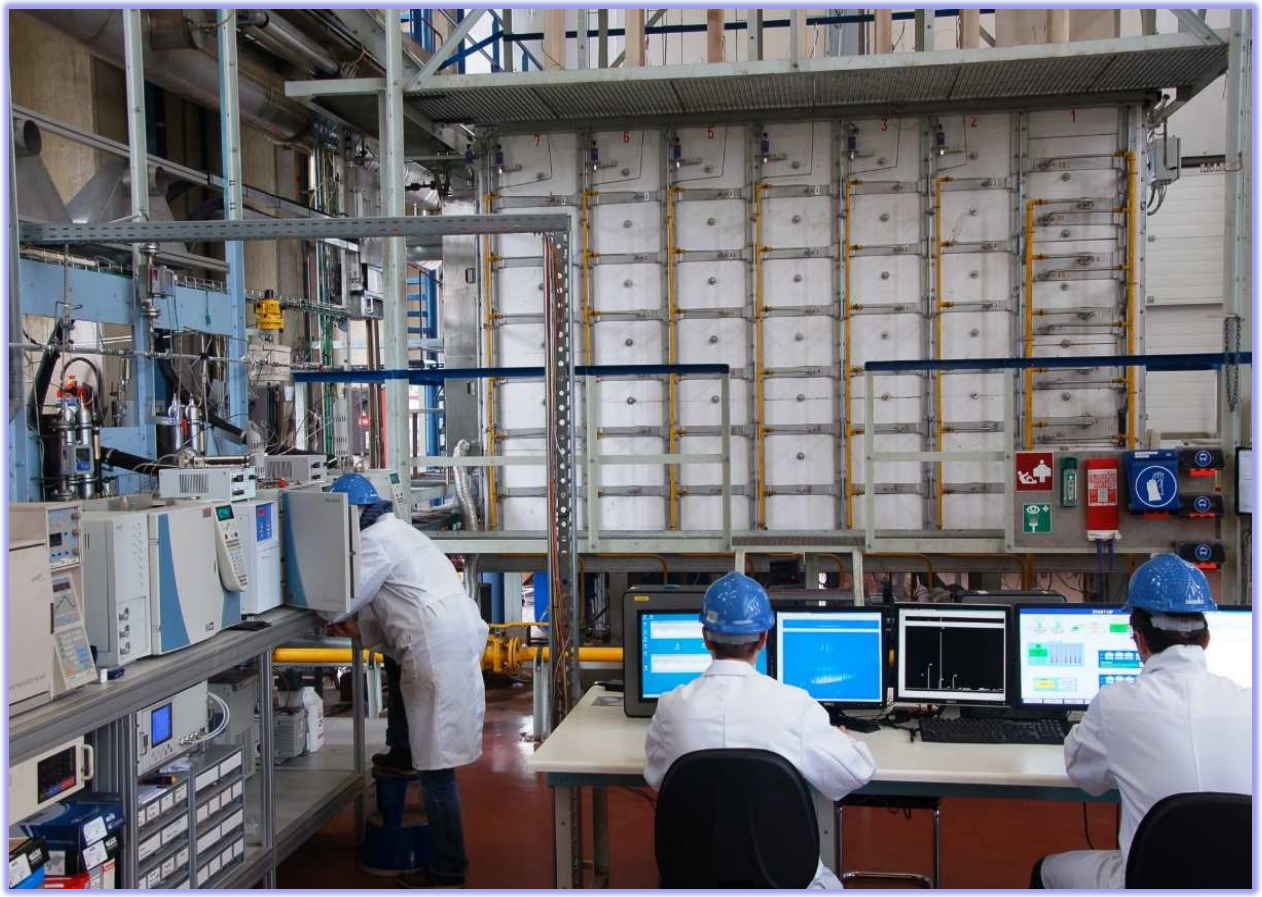
Pilot scale steam cracking experiments

Pilot scale steam cracking

- Duration - 6 hour
- Identical process conditions
- ***7 gas condensate*** feedstocks

Coke deposited in TLE is burned off (800 to 900 °C) after each experiment

CO and **CO2** concentrations are monitored (0.1 Hz) in the outlet stream (app. 1.5 hours)



$$Cokes\ (g) = MMc \sum_{i=0}^{N} \frac{\dot{Q}}{360}\frac{\left(Y_{co_i} + Y_{co_{2i}}\right)\cdot p_i}{RT_i}$$

MMc – Molecular Mass of carbon

$\dot{Q}$ – Volumetric flow rate

p – Pressure

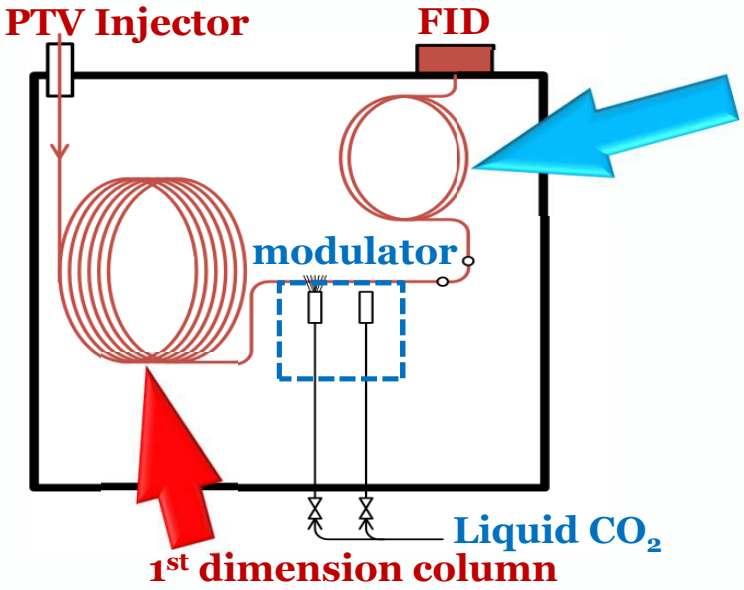
R – Gas Constant

Y_{co_i}- Molar concentration of CO

$Y_{co_{2i}}$- Molar concentration of CO2

T – Temperature

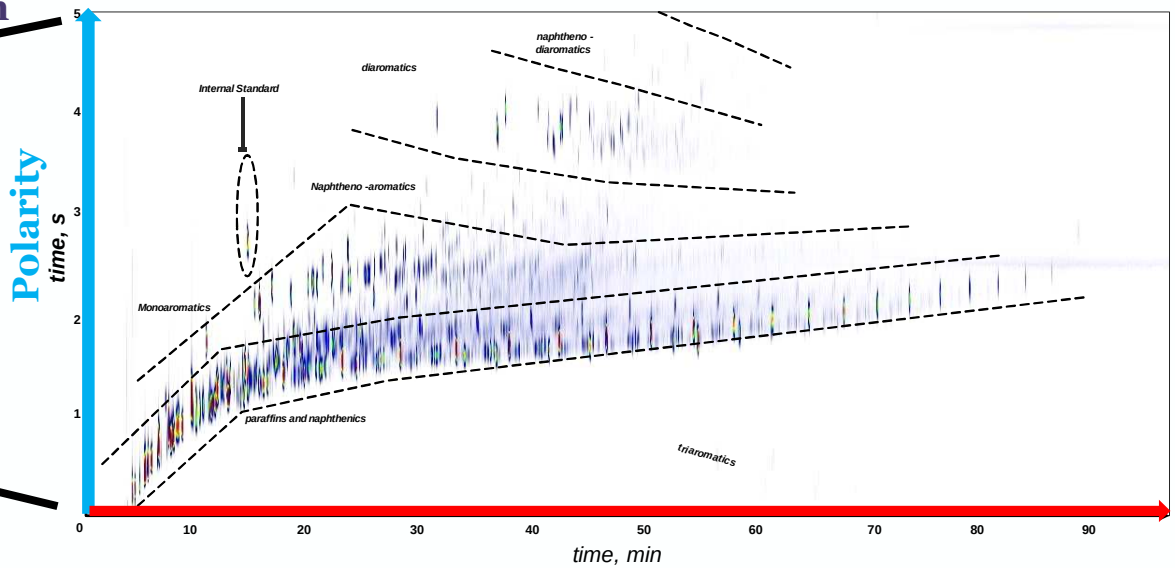
Compositional Characterization



2nd dimension column
BPX-50, SGE, 2 m, 0.15 mm, 0.15 µm

High resolution separation of **GC × GC** enables unrevealing the **complex** nature of petroleum streams

Due to the large number of detected molecules the number of **samples (7)** is unfavorable for exploratory analysis



Outline

- **Introduction**
- **Experimental study**
- **Data structure**
- **Variable selection PLSR modelling**
- **Conclusions**

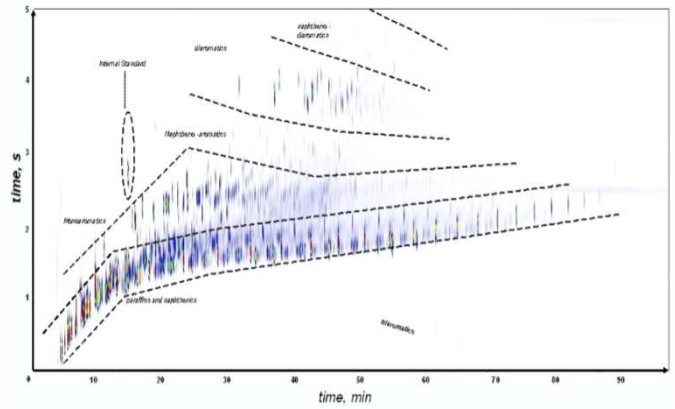
Data structure

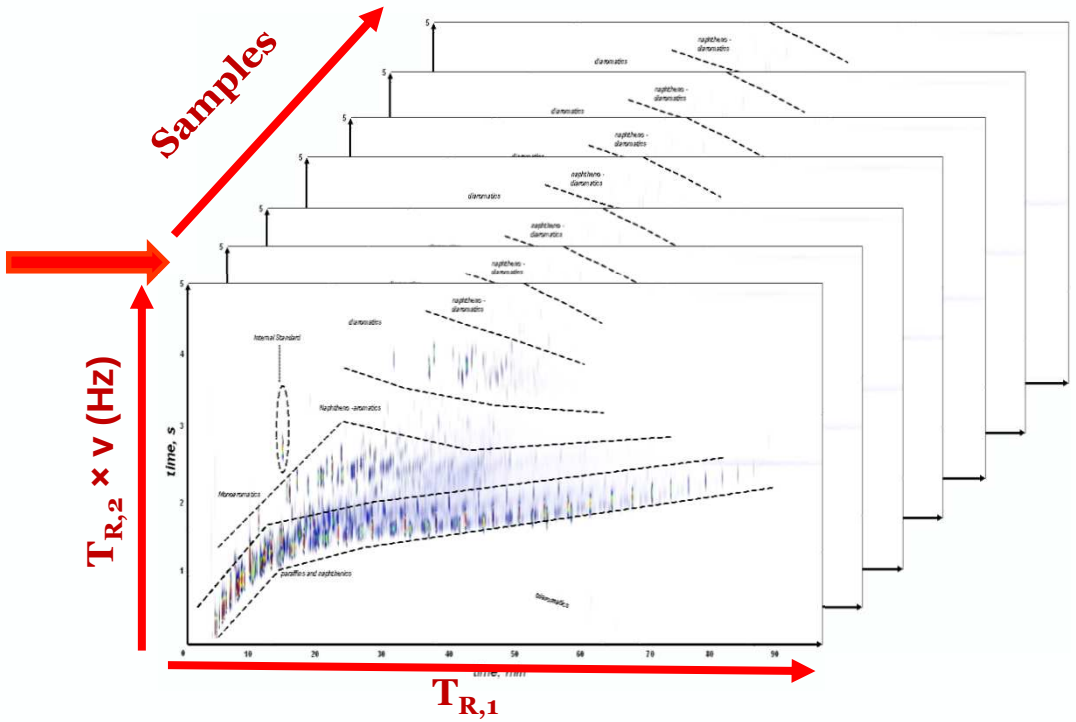
RAW X DATA **1D** → $T_{R,1} \times v$ (Hz)

Samples

Gas condensate	1 (0.01s)	2 (0.02s)	…	…	N (96.6 min)
1	X_{11}	X_{12}	…	…	X_{1N}
2	X_{21}	X_{22}	…	…	X_{2N}
3	X_{31}	X_{32}	…	…	X_{3N}
4	X_{41}	X_{42}	…	…	X_{4N}
5	X_{51}	X_{52}	…	…	X_{5N}
6	X_{61}	X_{62}	…	…	X_{6N}
7	X_{71}	X_{72}	…	…	X_{7N}

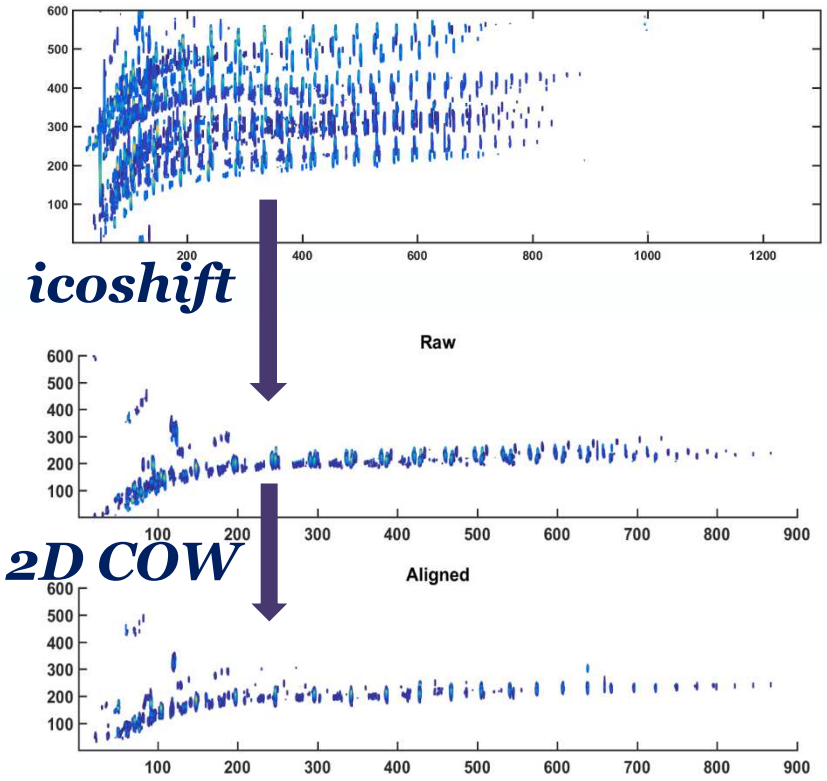
2D → $T_{R,1} \times T_{R,2} \times v$ (Hz)





Exploratory analysis requires ***data alignment*** solving retention time shift issues

Data alignment



Icoshift **-** applied on complete structure 1D data and subsequently on each 5 modulation segment

$$C_y^x(u) = \int_{-\infty}^{+\infty} Y(t+u)X(t)dt$$

$$x(f) = F\big(X(t)\big) = \int_{-\infty}^{+\infty} X(t)e^{2\pi ift}dt$$

$$\mathrm{X}(t) = F^{-1}\big(x(f)\big) = \int_{-\infty}^{+\infty} x(f)e^{-2\pi ift}df$$

2D COW

1D – segment of 20 data points with a allowable movement of 2 data points

2D – segment of 80 data points with a allowable movement of 30 data points

Partitioned chromatogram

(1,1)	(1,2)	(1,3)
(2,1)	(2,2)	(2,3)
(3,1)	(3,2)	(3,3)

Warped chromatogram

(1,1)	(1,2)	(1,3)
(2,1)	(2,2)	(2,3)
(3,1)	(3,2)	(3,3)

$$\sum_{j=0}^{n-1} \widetilde{corr}\left(\tilde{X}_{(l_j+u_j+1):(l_{j+1}+u_{j+1})}, Y_{(l_j+u_j+1):(l_{j+1}+u_{j+1})}\right)$$

Tomasi, G., et al. (2011). "Icoshift: An effective tool for the alignment of chromatographic data." Journal of Chromatography A **1218**(43): 7832-7840.

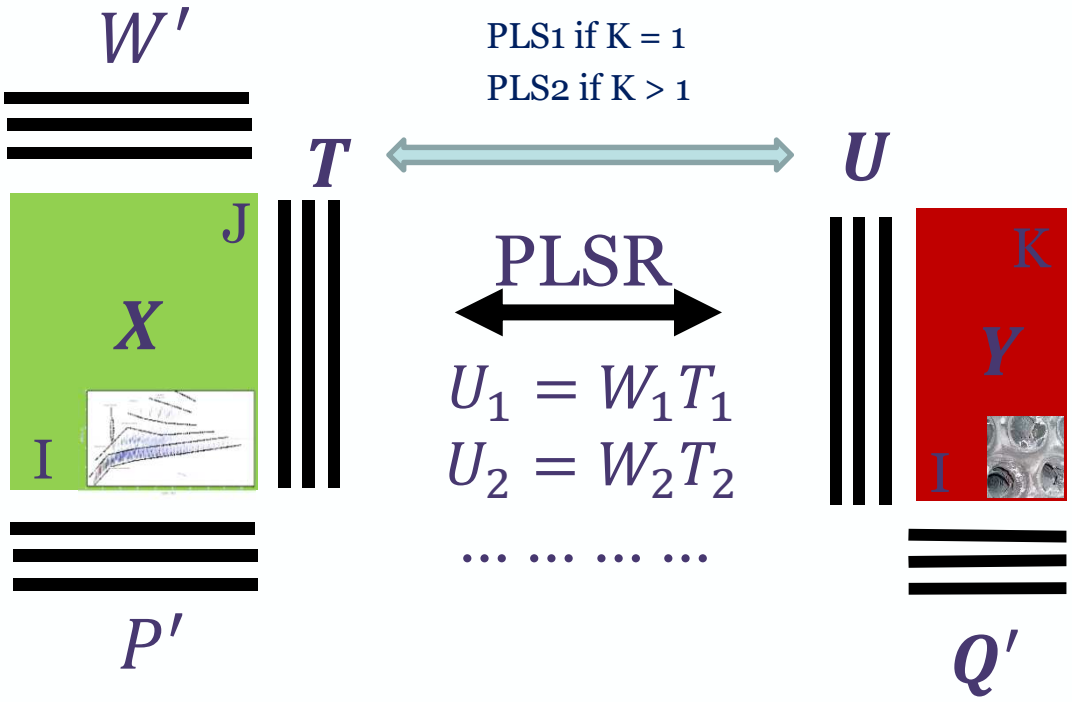
Zhang, D., et al. (2008). "Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data." Analytical Chemistry **80**(8): 2664-2671.

Outline

- **Introduction**
- **Experimental study**
- **Data structure**
- **Variable selection PLSR modelling**
- **Conclusions**

Partial least squares regression (PLSR)

Maximize covariance of X and Y
(NIPALS or SIMPLS algorithms)



X=TP′
Y=UQ′

T=XW
Y=TQ′
Y=XWQ′

B =WQ′
(regression coefficients are not independent)

Y=XB

$$X = T\,P' + E_X \qquad Y = U\,Q' + E_Y$$

Direct multiple linear regression (MLR) cannot handle collinear and large rank matrixes

Wold, S., M. Sjöström and L. Eriksson (2001). "PLS-regression: A basic tool of chemometrics." Chemometrics and Intelligent Laboratory Systems 58(2): 109-130.

Variable selection - Methods that require initially valid PLSR model

Variable importance in projection (VIP)

$$VIP_j = \sqrt{\frac{\sum_f^F w_{jf}^2 \cdot SSY_f \cdot J}{SSY_{total} \cdot F}}$$

$$SSY_f = \boldsymbol{b}_f^2 \boldsymbol{t}_f' \boldsymbol{t}_f \qquad SSY_{total} = \boldsymbol{b}^2 \boldsymbol{T}' \boldsymbol{T}$$

w_{jf} - weight value for variable j of component f

SSY_f - the sum of squares of $\boldsymbol{y}$ for the f^{th} component

J - the number of variables

SSY_{total} - the sum of squares of the total explained variance of $\boldsymbol{y}$

F - the total number of components, *i.e.* LVs

Selectivity ratio (SR)

$$\boldsymbol{X} = \hat{\boldsymbol{X}}_{TP} + \boldsymbol{E}_{TP} = \boldsymbol{t}_{TP}\boldsymbol{p}_{TP}' + \boldsymbol{E}_{TP}$$

$$\boldsymbol{t}_{TP} = \frac{\boldsymbol{X}\boldsymbol{b}}{\|\boldsymbol{b}\|} \quad \boldsymbol{p}_{TP} = \frac{\boldsymbol{X}'\boldsymbol{t}_{TP}}{\boldsymbol{t}_{TP}'\boldsymbol{t}_{TP}} \quad SR_j = \frac{SS_{explained,j}}{SS_{residual,j}}$$

$$SS_{explained,j} = \left\|\boldsymbol{t}_{TP}\boldsymbol{p}_{TP,j}'\right\|^2 \quad SS_{residual,j} = \left\|\boldsymbol{E}_{TP,j}\right\|^2$$

t_{TP} – *scores vector*

p_{TP} - *loadings vector*

SS – sum of squares for each variable j

Regression Vector

Regression coefficients which are a single measure of association between each variable and the response can be utilized based on the assumption that variables with larger coefficient are more influential and therefore more important

The regression vector, VIP and SR were acquired straight from the PLS_Toolbox

Mehmood, T., K. H. Liland, L. Snipen and S. Sæbø (2012). "A review of variable selection methods in Partial Least Squares Regression." Chemometrics and Intelligent Laboratory Systems 118: 62-69.

Rajalahti, T., R. Arneberg, F. S. Berven, K. M. Myhr, R. J. Ulvik and O. M. Kvalheim (2009). "Biomarker discovery in mass spectral profiles by means of selectivity ratio plot." Chemometrics and Intelligent Laboratory Systems 95(1): 35-48

Farrés, M., S. Platikanov, S. Tsakovski and R. Tauler (2015). "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation." Journal of Chemometrics 29(10): 528-536.

Variable selection – Screening methods

RReliefF algorithm

$$W_j = \frac{P_{diff\ C\ |\ diff\ F_j}\, P_{diff\ F_j}}{P_{diff\ C}} - \frac{\left(1 - P_{diff\ C\ |\ diff\ F_j}\right) P_{diff\ F_j}}{1 - P_{diff\ C}}$$

$P_{diff\ C}$ - *the probability that the two nearest variables have different predictions*

$P_{diff\ F}$ - the probability that the two nearest variables have different values for the independent variable

$P_{diff\ C\ |\ diff\ F}$ -the probability that the two nearest variables have both different values for the independent and dependent variable

F-test statistic or Fisher ratio

$$F = \frac{MS_{regression}}{MS_{residuals}}$$

MS - mean squares calculated for each variable

MATLAB Statistics and Machine Learning Toolbox

Patchava, K. C., M. Benaissa and H. Behairy (2015). Improving the prediction performance of PLSR using RReliefF and FSD for the quantitative analysis of glucose in Near Infrared spectra. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.

Model validation

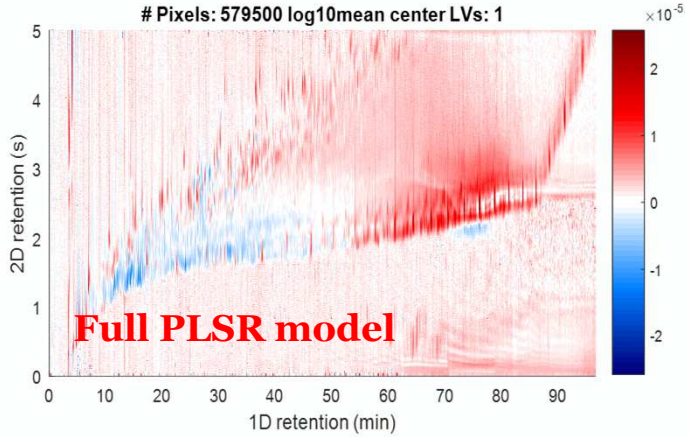
1. Both ***X*** and ***Y*** data blocks are Mean Centered (***MC***)
2. ***Logarithmical*** transformation of ***both blocks***
3. Evaluation via calculation of ***RMSECV*** established using ***leave-one out*** methodology
4. With increasing the number of ***LV*** the threat of overfitting increases
5. Models are build by ***iterative*** increase of the ***relevance cut-off value,*** with a constrain that the model must not have less than ***10 000 pixels***

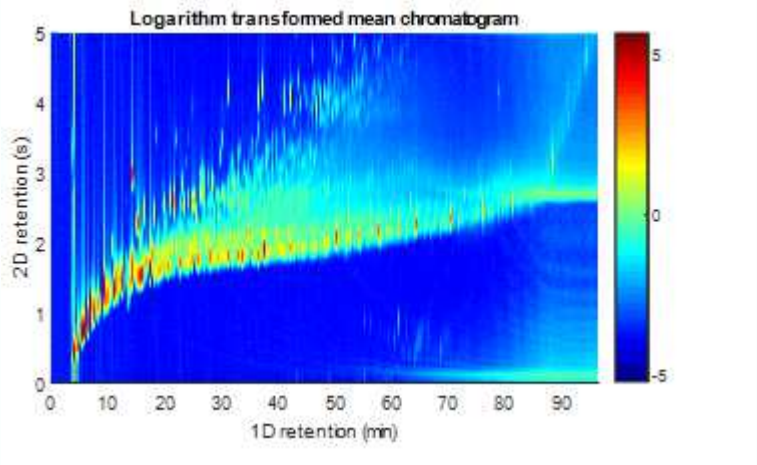
Results	**Preprocessing**		**PLSR - RMSECV (g deposit/6 h)**	
Method	**X-block**	**Y-block**	**1LV**	**2LV**
Full	log10 MC	log10 MC	0.95	0.91
Regression vector	log10 MC	log10 MC	**0.35**	0.44
VIP	log10 MC	log10 MC	**0.38**	0.43
SR	log10 MC	log10 MC	0.47	0.45
RReliefF	log10 MC	log10 MC	0.42	0.62
Fisher ratio	log10 MC	log10 MC	**0.38**	0.51

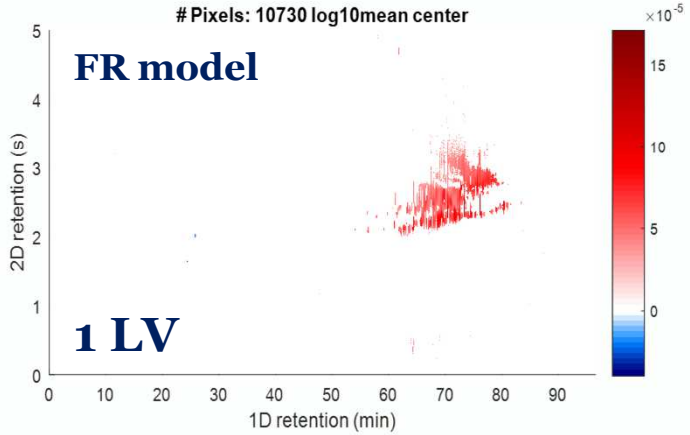
MC – mean centered **log10 – logarithm base 10**

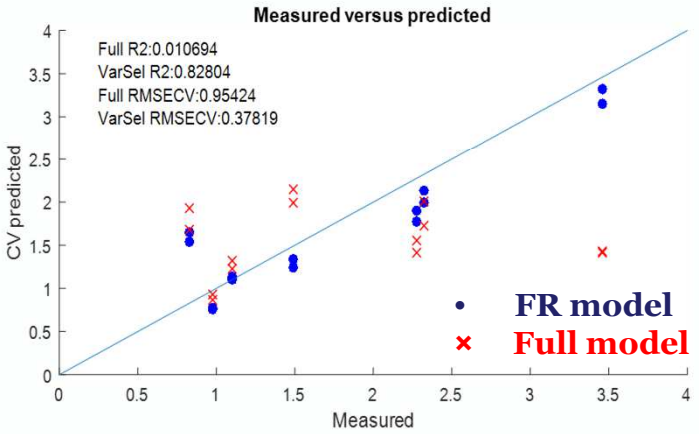
Fisher ratio

Pixels corresponding to aromatics and naphthenics with long aliphatic side chains and high boiling point paraffins serve as the best TLE fouling predictors



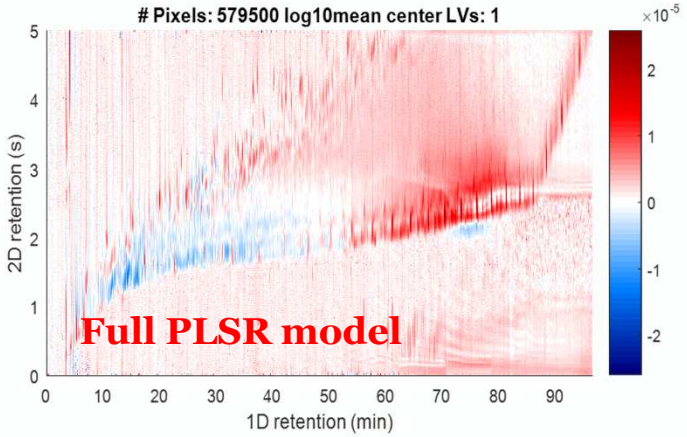


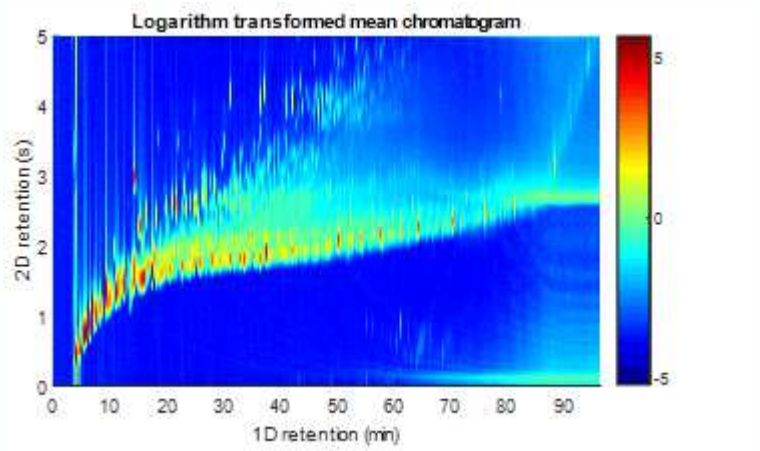


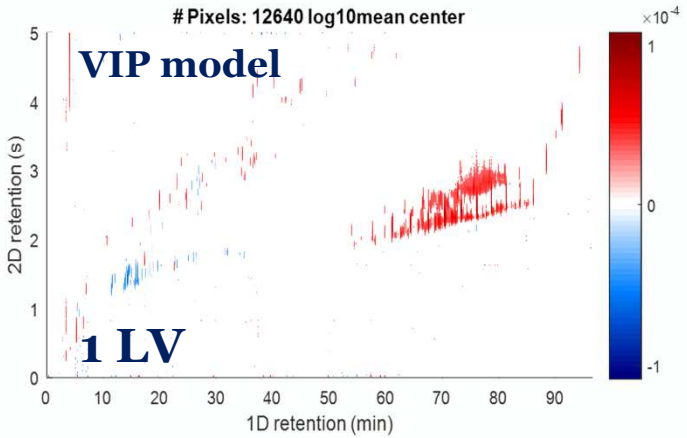


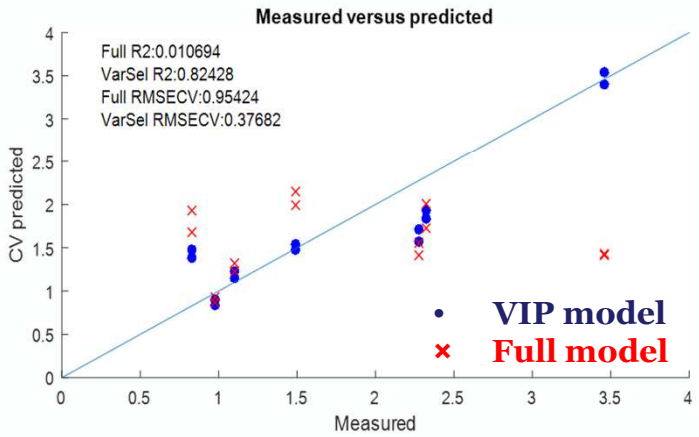
Variable importance in projection

Confirms that aromatics and naphthenics with long aliphatic side chains along with poly-aromatics and high boiling point paraffins serve as good TLE fouling predictors



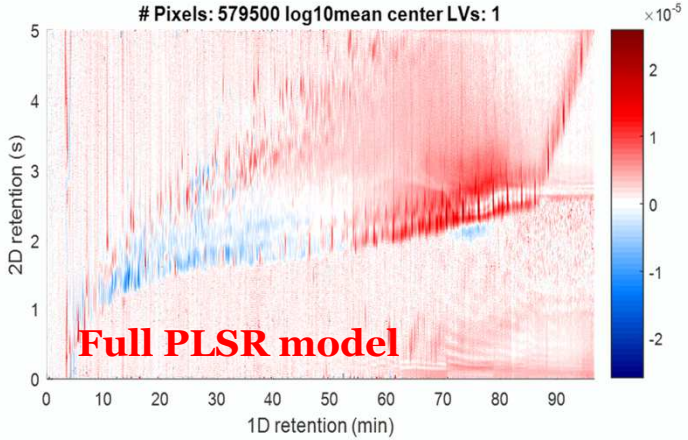


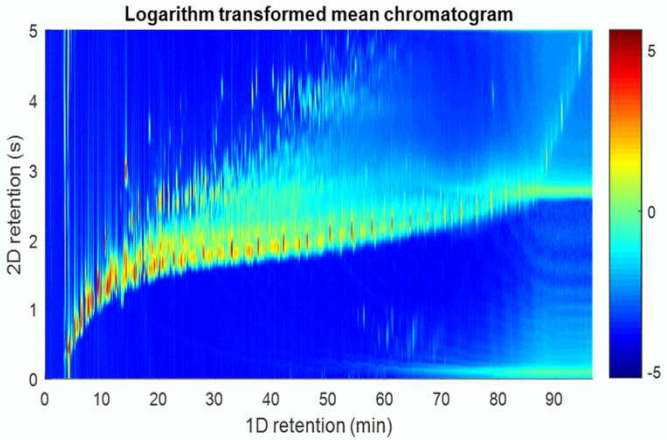


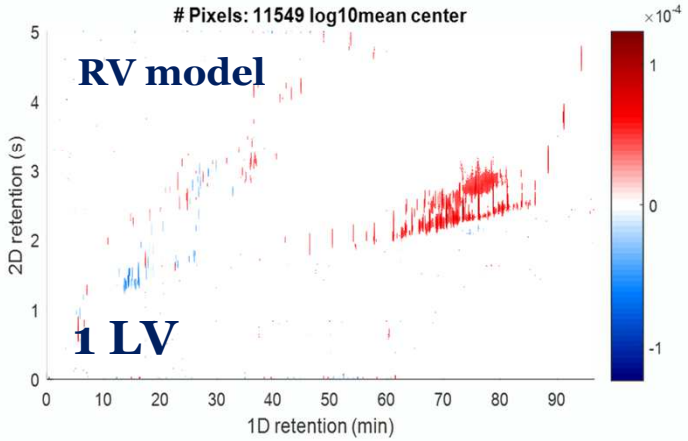


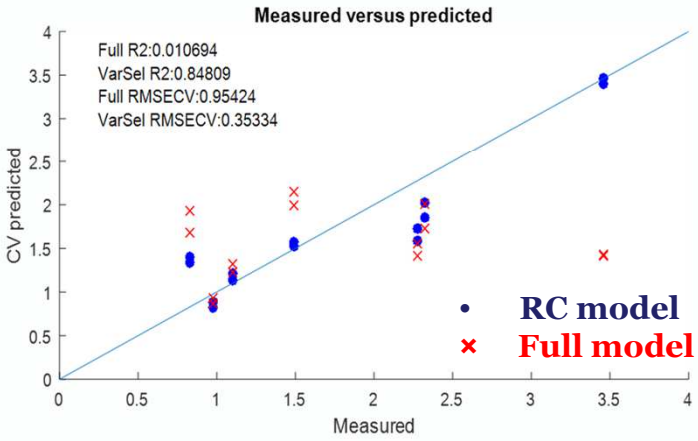
Regression vector

Same pixels, corresponding to aromatics and naphthenics with long aliphatic side chains, poly-aromatics and to some extent to high boiling point paraffins serve as good TLE fouling predictors









The method enabled establishing of the most accurate TLE fouling prediction model with an average relative error of 20 %

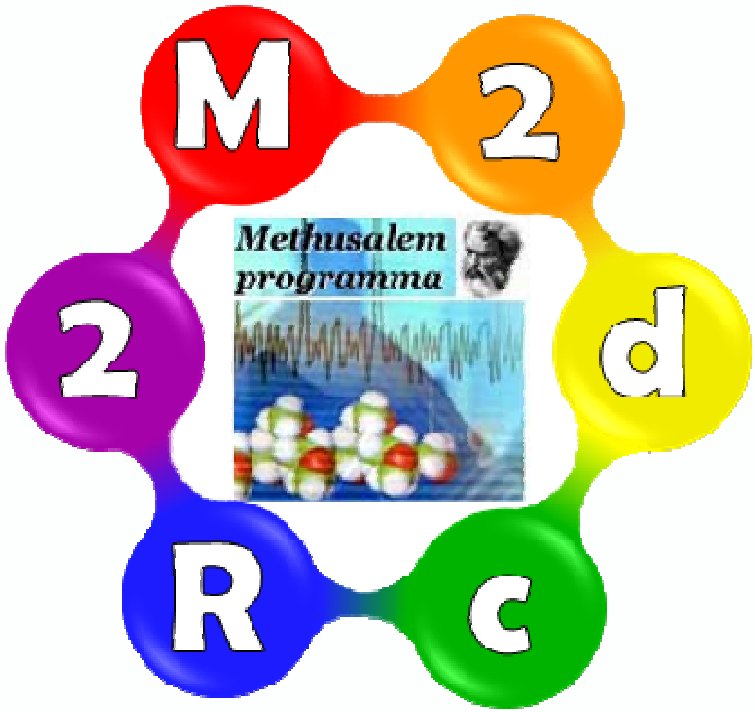
Outline

- **Introduction**
- **Experimental study**
- **Data structure and PCA**
- **Variable selection PLSR modelling**
- **Conclusions**

Conclusions

- Pixel-based PLSR models can be applied within the calibration sample range
- Identical analytical method must be used for characterization of each sample, slight shifts in retention can be aligned
- Variable selection techniques are able to detect the chromatographic regions, i.e. chemical compounds, with the strongest correlation with the studied phenomenon
- Monoaromatics and naphthenics with a high boiling point are the best TLE fouling predictors
- Composition of the feeds is successfully correlated with the TLE fouling, providing a predictive model with an average relative error of 20 %

Acknowledgments



Long Term Structural Methusalem Funding by the Flemish Government

Q & A

Thank you for your attention!

Journal of Chromatography A, 1501 (2017) 89–98

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Comprehensive two-dimensional gas chromatography in combination with pixel-based analysis for fouling tendency prediction

Victor Abrahamsson[a], Nenad Ristic[b], Kristina Franz[b], Kevin Van Geem[b,*]

[a] Lund University, Department of Chemistry, Centre for Analysis and Synthesis, Lund, Sweden
[b] Ghent University, Laboratory for Chemical Technology, Ghent, Belgium

ARTICLE INFO

Article history:
Received 18 October 2016
Received in revised form 19 February 2017
Accepted 11 April 2017
Available online 14 April 2017

Keywords:
Comprehensive two-dimensional chromatography
Pixel-based analysis
Feature selection
Steam cracking
Fouling

ABSTRACT

Fouling tendencies of a series of gas condensates were evaluated using comprehensive two-dimensional gas chromatography with flame ionization detection and sulfur chemiluminescence detection. A pixel-based approach was applied in order to identify parts of the chromatograms which were associated with the reactor coil fouling. Particular emphasis is given in this work to evaluate several feature selection methodologies along with various data preprocessing procedures. It was found that both aspects were crucial for studying the fouling tendencies and, as part of the subsequent partial least squares model development, predominantly the feature selection. Based on the flame ionization detector chromatograms and using the RReliefF algorithm for feature selection, a partial least squares regression model with one latent variable resulted in a root mean square error of the cross-validation of 0.65 g deposit/6 h (17%). Based on the sulfur chemiluminescence detector chromatograms, the F-statistics feature selection generated a slightly better partial least squares regression model compared to using RReliefF, thus generating a model using one latent variable with a root mean square error of the cross-validation of 0.81 g deposit/6 h (21%). Heavy aromatic compounds and heavy sulfur containing compounds were negatively associated with the fouling rate. Both were crucial in developing a partial least squares model with good prediction power, however, worked independently as predictors.

Glossary

BTX – Benzene, Toluene, Xylene
LPG – Liquefied petroleum gas
T - Temperature
P - Pressure
δ – Dilution
TLE – Transfer Line Heat Exchanger
GC × GC - comprehensive two-dimensional gas chromatography
PCA – Principle Component Analysis
PC – Principle component
PLSR - Partial Least Squares Regression
VIP - Variable importance in projection
SR - Selectivity ratio
RMSECV - Root Mean Square Error of Cross Validation
LV - latent variable
MC - Mean Centered
FR - Fisher Ratio
RC - Regression Coefficients
PTV - Programmed Temperature Vaporising injector
FID - Flame Ionization Detector