FACULTY OF **ENGINEERING AND ARCHITECTURE**

Evaluation of variable selection approaches for pixel-based analysis of GC×GC data

***Victor Abrahamsson*[1], *Nenad Ristic*[2], *Kristina Franz*[2] *and Kevin Van Geem*[2]**

[1]Lund University, Department of Chemistry, Centre for Analysis and Synthesis

[2]Ghent University, Laboratory for Chemical Technology

E-mail: Victor.Abrahamsson@chem.lu.se

Aim

- Study the effect of gas condensate chemical composition on **reactor coke formation during steam cracking.**
- Evaluate **variable (feature) selection** methodologies prior to partial least squares regression (PLSR).

Introduction

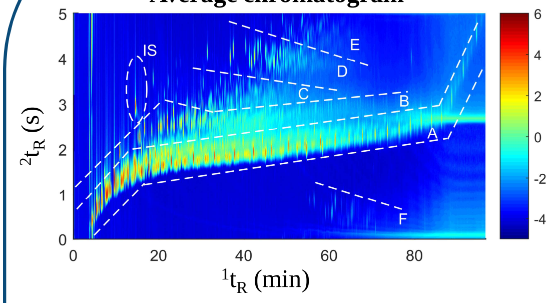
- **Pixel-based analysis** enables integration free interpretation of GC×GC data. Therefore all information is retained and analysis is swift.[1]
- **PLSR** can be used to correlate information in chromatograms with *e.g.* petroleum properties.[2]
- **Variable selection** is needed to remove redundant information and to avoid focusing on large peaks in chromatograms which might not be important.[3]

GC×GC

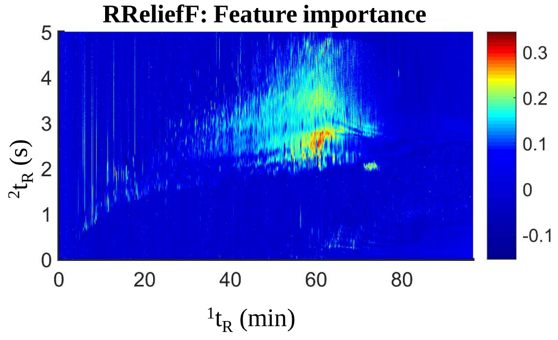
- 50 m dimethyl polysiloxane column (RTX-1 PONA, 0.25 mm I.D., 0.5 µm film thickness)
- 2 m phenyl polysilphenylene-siloxane (BPX50, 0.15 mm I.D., 0.15 µm film thickness)
- Dual-stage cryogenic (liquid CO_2) modulator
- Flame ionization detection
- 8 gas condensates with duplicate analysis and QC samples

Results and discussion

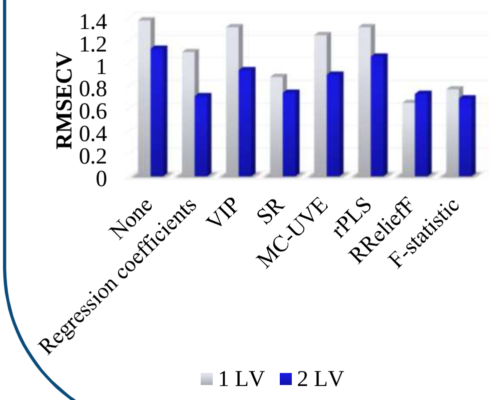
Average chromatogram



RReliefF: Feature importance



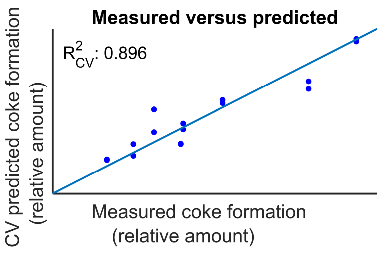
Comparison of feature selection methods



Representation of the mean of all chromatograms with a logarithmic transformation to visualize the large span between intensities of various analyte groups.

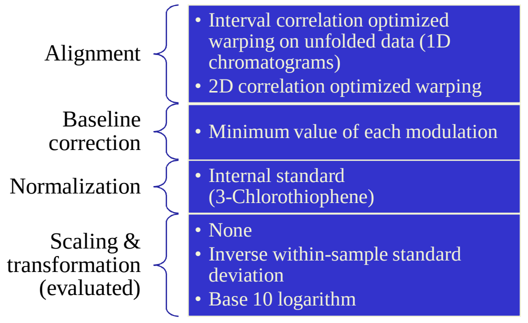
The group-type separation is categorized by: (A) paraffins and napthenes, (B) monoaromatics, (C) napthenoaromatics, (D) diaromatics, (E) napthenodiaromatics, (F) triaromatics. Added 3-Chlorothiophene was used as internal standard (IS).

Several feature selection methodologies were evaluated. The RReliefF algorithm resulted in the lowest root mean square error of the cross-validation (RMSECV).



*Cross-validation (CV) was performed using leave-one-sample-out, *i.e.* both replicates.

Signal processing



Alignment
- Interval correlation optimized warping on unfolded data (1D chromatograms)
- 2D correlation optimized warping

Baseline correction
- Minimum value of each modulation

Normalization
- Internal standard (3-Chlorothiophene)

Scaling & transformation (evaluated)
- None
- Inverse within-sample standard deviation
- Base 10 logarithm

Conclusions

- Feature selection is a crucial part of the development of PLSR models
- Feature selection indicates important regions in chromatograms
- RReliefF was the most efficient method
- RReliefF offers independent evaluation and does not require an initial PLSR model.
- Reactor coke formation was associated with heavy aromatic compounds and could be used as a predictor

References

(1) Furbo, S., et al. (2014). Anal. Chem. **86**(15): 7160-7170.

(2) Pierce, K. M., et al. (2015). Data Handl. Sci. Techn. **29:** 427-463.

(3) Andersen, C. M. and R. Bro (2010). J. Chemometr. **24**(11-12): 728-737.

Acknowledgements

•Swedish Research Council VR

•Royal Physiographic Society in Lund

Analysdagarna, Umeå, 14/06/2016