

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

Victor Abrahamsson, Nenad Ristic\*, Kristina Franz\* and Kevin Van Geem\*

Lund University, Department of Chemistry, Centre for Analysis and Synthesis (CAS), The Green Technology Group (GTG), P.O. Box 124, SE-221 00 Lund, Sweden, \*Ghent University, Laboratory for Chemical Technology, BE-9052 Zwijnaarde, Belgium.

Comprehensive two-dimensional gas chromatography (GC×GC) offers unsurpassed peak capacities ranging up to ten times as high as compared to one-dimensional GC [1]. Yet, petroleum analysis presents a difficult challenge for the analytical chemist. The number of compounds in a petroleum fraction increases exponentially with the boiling point and can quickly exceed the peak capacity of both one-dimensional GC and as well as GC×GC [2]. This naturally makes traditional peak integration difficult and a time-consuming task. Pixel-based analysis of GC×GC generated data offers a direct approach for chemical evaluation of samples. Intermediate steps of integration are avoided, thus subjectivity in either manual integration or setting the peak selection parameters is omitted [3]. In pixel-based analysis each detector scan is treated as a variable in the two-dimensional chromatographic space of  $^1t_R$  and  $^2t_R$ . Chemical information in areas where peaks are too small or the resolution is too poor for integration, is still retained and no information is discarded [3]. For established routines the analysis is quick and completely automated. However, the retained data becomes both a blessing and curse, as it is not easily analyzed in its raw format. Multivariate regression techniques like partial least squares regression (PLSR) enables the prediction of sample properties or measurements where the number of variables exceed the number of observations. However, PLSR is not insensitive to measurement noise and it may be difficult to establish a model where the variables with the highest intensities do not actually function as predictors [4]. Therefore, variable selection becomes a very important aspect of pixel-based analysis of GC×GC generated data.

Here, we present a comparison of commonly used variable selection techniques for PLSR such as filtering by regression coefficients, variable importance in projection and the selectivity ratio as well as more recently introduced techniques such as recursive weighted PLS, Fisher ratio analysis and RReliefF. The comparison is demonstrated based on an example where the chemical information within GC×GC-FID chromatograms is used for exploring the underlying chemistry and to make prediction models of coke formation in a steam cracking process. It is shown that variable selection techniques based on already established PLSR models may not be sufficient or even feasible and that independent algorithms may be beneficial when no prior knowledge about the needed pre-processing exists.

## References

1. Oh, C., et al., *Journal of Chromatography A*, 2008. **1179**(2): p. 205-215.
2. Nizio, K.D., et al., Harynuk, *Journal of Chromatography A*, 2012. **1255**: p. 12-23.
3. Furbo, S., et al., *Analytical Chemistry*, 2014. **86**(15): p. 7160-7170.
4. Andersen, C.M. et al., *Journal of Chemometrics*, 2010. **24**(11-12): p. 728-737.