Conferentia Chemometrica 2019

Karcag, Hungary
Hotel Nimród
September 08-12, 2019

Research Centre for Natural Sciences
Hungarian Academy of Sciences
Chemometrics and Chemoinformatics Working Group of the Hungarian Academy of Sciences

International Organizing Committee

K. Héberger (Hungary) chair

R. G. Brereton (UK)

J. H. Kalivas (USA)

O. M. Kvalheim (Norway)

Kunal Roy (India)

R. Todeschini (Italy)

K. Varmuza (Austria)

A. Voelkel (Poland)

I. G. Zenkevich (Russia)

Local Organizers

Head:

Károly Héberger, Research Centre for Natural Sciences, Hungarian Academy of Sciences

Administrative Affairs:

Dávid Bajusz and Anita Rácz, Research Centre for Natural Sciences, Hungarian Academy of Sciences

Financial Affairs:

János Elek SciencePort Ltd.

ISBN 978-963-7067-38-9

Edited by Károly Héberger

Scientific Program of the Conferentia Chemometrica 2019: an overview

Sept. 08, Sunday: 16:00–18:00 Registration 18:00–20:00 Get-together party;

Sept. 09, Monday: 8:00-8:50 Registration, 8:50-9:00 Opening

September 09 Monday		September 10 Tuesday		September 11 Wednesday		September 12 Thursday	
09:00–09:30	**L01 O.M. Kvalheim**	**09:00–09:30**	**L08 T. Bocklitz**	**09:00–09:30**	**L16 C. Ruckebush**	**09:00–09:30**	
09:30–10:00	**L02 G. Tóth**	**09:30–10:00**	**L09 K. Baumann**	**09:30–10:00**	**L17 M. Daszykowski**	**09:30–10:00**	**L24 T. Lundstedt**
10:00–10:30	**L03 K. Roy**	**10:00–10:30**	**L10 A. Smilde**	**10:00–10:30**	**L18 L. Pieszczek**	**10:00–10:30**	**L25 J. Elek**
10:30–11:00	Break	**10:30–11:00**	Break	**10:30–11:00**	Break	**10:30–11:00**	Break
11:00–11:30	**L04 R. Todeschini**	**11:00–11:30**	**L11 R. Brereton**	**11:00–11:30**	**L19 I. Stanimirova**	**11:00–11:30**	**L26 D. Bajusz**
11:30–12:00	**L05 Y. Monakhova**	**11:30–12:00**	**L12 J. Abonyi**	**11:30–12:00**	**L20 A. Surkova**	**11:30–12:00**	**L27 K. Heberger**
12:00–13:30	Lunch	**12:30–13:30**	Lunch	**12:30–13:30**	Lunch	**12:30–13:30**	Lunch
13:30–14:00	**L06 A. Voelkel**	**13:30–14:00**	**L13 M. Novič**	**13:30–14:00**	**L21 L. Guillou**		
14:00–14:30	**L07 T. Baczek**	**14:00–14:30**	**L14 V. Fonseca Diaz**	**14:00–14:30**	**L22 S. Kovács**	**14:00–**	**Departure**
14:30–15:30	**POSTER SESSION**	**14:30–15:00**	**L15 B. Hemmateenejad**	**14:30–15:00**	**L23 M-h. Kim**		
		15:00–15:30	Break	**15:00–15:30**	Break		
15:30–17:00	**Visit to Cumania Museum**	**15:30–16:30**	**POSTER SESSION**	**15:30–16:30**	**POSTER SESSION**		
17:00–18:00	Dinner	**17:00–**	**Excursion Morgó Tcharda Dinner &**	**18:00–19:00**	Break		
18:00–21:00	**Free time (Thermal bath)**		**City mayor's speech**	**19:00–**	**Banquet (Best Poster Award) City souvenirs**		

Sunday evening, Sept. 08, 2019

16:00–18:00 **Registration**
18:00–20:00 **Get-together party**

Monday morning, Sept. 09, 2019

08:00–08:50 **Registration**
08:50–09:00 **Opening, technical information**

Variable selection, allocation, classification, QSAR, cross-validation

09:00–09:30 **L01 Olav M. Kvalheim:**
What is (are) the best measure(s) of variable importance for interpretation and variable selection in multivariate regression models?

09:30–10:00 **L02 Gergely Tóth:**
Optimalization of R^2 and related quantities by allocation

10:00–10:30 **L03** Pravin Ambure, Agnieszka Gajewicz, M Natalia DS Cordiero and **Kunal Roy:**
QSAR model development from small data sets: A new workflow with integration of data curation, double cross-validation and consensus prediction tools

10:30–11:00 Coffee Break

Ranking, multivariate calibration, pattern identification in spectra

11:00–11:30 **L04 Roberto Todeschini,** Francesca Grisoni and Davide Ballabio:
Deep Ranking Analysis by Power Eigenvectors (DRAPE): a wizard for ranking and multi-criteria decision making

11:30–12:00 **L05 Yulia Monakhova,** Bernd Diehl:
Transfer of multivariate regression models based on 1D and 2D spectra between high-resolution NMR instruments: application to lecithin and heparin

12:00–13:30 Lunch Break

Monday afternoon, Sept. 09, 2019

Data evaluation in chromatography, applications, metabolomics

13:30–14:00 **L06** **Adam Voelkel,** B. Strzemiecka, M. Sandomierski:
Chemometric support in physicochemical evaluation of industrial materials

14:00–14:30 **L07** Lucyna Konieczna, Anna Krawczyńska, **Tomasz Bączek:**
Assessment of amino acids' variability in plasma, cerebrospinal fluid and exhaled breath condensates in pediatric leukemia patients

14:30–15:30 **POSTER SESSION**

15:00–15:30 Coffee Break

15:30–17:00 **Visit to Cumania Museum**

17:00–18:00 Dinner

18:00–21:00 **Thermal bath**

Tuesday morning, Sept. 10, 2019

Machine learning, prediction, QSAR, metabolomics

09:00–09:30 **L08** **Thomas Bocklitz:**
Chemometrics and machine learning for analysis of Raman-related data

09:30–10:00 **L09** **Knut Baumann**, V. Schlenker, A. ter Laak, N. Heinrich:
Incorporating left-censored bioactivity data in predictive regression models

10:00–10:30 **L10** **Age K. Smilde** and Thomas Hankemeier:
Numerical Representations of Metabolic Systems

10:30–11:00 Coffee Break

Probability theory, preference maps

11:00–11:30 **L11** **Richard G. Brereton**:
The use and misuse of p values and related concepts

11:30–12:00 **L12** **János Abonyi:**
Mixture of QSAR Models–Learning Gating Functions to Combine of pK_a Predictions

12:00–13:30 Lunch Break

Tuesday afternoon, Sept. 10, 2019

Toxicity prediction, bilinear modeling, multiblock data

13:30–14:00 **L13** Liadys Mora Lagares, Nikola Minovski, Viktor Drgan, Marjan Tušar **Marjana Novič:**
PgP transport activity - in connection to the efflux of toxicants or drugs

14:00–14:30 **L14** **V. Fonseca Diaz**, W. Saeys:
Robustness control in bilinear modeling based on maximum correntropy

14:30–15:00 **L15** E. T. Bayat, K. Baumann, **Bahram Hemmateenejad:**
Improved rank estimation using randomization and dependency

15:00–15:30 Coffee Break

15:30–16:30 **POSTER SESSION**

17:00 **Excursion, Morgó Tcharda, Dinner &**

Wednesday morning, Sept. 11, 2019

Hyperspectral chemical images, classification

09:00–09:30 **L16** **Cyril Ruckebusch:**
Towards easier and more reliable multivariate curve resolution of hyperspectral images

09:30–10:00 **L17** **M. Daszykowski,** L. Pieszczek:
Spatial advantage of hyperspectral imaging and construction of rigorous classifiers

10:00–10:30 **L18** **L. Pieszczek,** M. Daszykowski:
Determination of plastic particle size using discreet information hidden in NIR-HSI images

10:30–11:00 Coffee Break

11:00–11:30 **L19** **Ivana Stanimirova**:
Uncertainty SIMCA – a classification method that includes measurement uncertainty information

11:30–12:00 **L20** **Anastasiia Surkova**, A. Bogomolov, A. Legin, D. Kirsanov:
Calibration model transfer between optical multisensor systems and full-scale spectrometers

12:00–13:30 Lunch

Wendesday afternoon, Sept. 11, 2019

Data processing, rank estimation, classification, drug design

13:30–14:00 **L21** **C. Guillou and L. Guillou:**
Data processing of spectra of RAMAN portable handheld devices

14:00–14:30 **L22** L. D. Koren, L. Lőrincz, **Sándor Kovács,** G. Kun-Farkas, B. Vecseriné Hegyes, L. Sipos:
Comparison of classifiers for commercial beers and identifying patterns

14:30–15:00 **L23** **Mi-hyun Kim**, H. Kim, S. Lee, S. Ahn, S. Kumar:
3D-Chemocentric target deconvolution of unprecedented drug scaffolds

15:00–15:30 Coffee Break

15:30–16:30 **POSTER SESSION**

19:00– **Conference Dinner (Distribution of the Best Poster Award)**

Thursday morning, Sept. 12, 2019

Spectra evaluation, consensus modeling, classification

09:00–10:00

09:30–10:00 **L24** **Torbjörn Lundstedt**, K. Lundstedt-Enkel, K. Bennett, C. Russell, R. Martín-Jiménez, M. Campanella, S. Mole, J. Petschnigg and J. Trygg:
Endogenous Metabolic Profiling as a Fundament in Personalized Theranostics

10:00–10:30 **L25** **János Elek**, E. Markovics, Zs. Komka M. Szász:
Evaluation and understanding of near infrared spectra in sports diagnostic examinations

10:30–11:00 Coffee Break

11:00–11:30 **L26** **Dávid Bajusz**, A. Rácz, K. Héberger:
Data Fusion Methods as Consensus Scores for Ensemble Docking

11:30–12:00 **L27** A. Rácz, D. Bajusz, **Károly Héberger**:
Comparison of performance parameters for machine learning classifiers

12:30–13:30 Lunch

14:00– **Departure**

Poster sessions

Monday afternoon: 15:30–16:30

Tuesday and Wendesday afternoon: 15:30–16:30

P01 Dávid Bajusz, Anita Rácz, Károly Héberger: Fingerprint similarity metrics in cheminformatics, metabolomics and other fields

P02 Barbara Biró, A. M. Sipos, A. Kovács, K. Badak-Kerti, K. Pásztor-Huszár, A. Gere: Sensory evaluation of cricket-enriched oat biscuits using check-all-that-apply analysis

P03 Zsuzsanna Guld, D. Nyitrainé Sárdy, A. Gere, A. Rácz: Multi-level comparison of Hungarian wines using advanced chemometric methods

P04 Loránd Románszki, Szilvia Klébert, and Károly Héberger: How can surface roughness be estimated at best?

P05 Sara Mostafapour, Bahram Hemmateenejad: Net analyte signal-based supervised preprocessing: application in pattern recognition

P06 Kabiruddin Ikramuddin Khan, K. Roy: Consensus QSAR modeling for the toxicity of organic chemicals against *pseudokirchneriella subcapitata* using 2D descriptors

P07 S. Lee, S. Ahn, Mihyun Kim: How to measure distribution for a pair of target classes?

P08 Péter Király, Dániel Kovács, Gergely Tóth: Remarks on some validation parameters

P09 Daniel Kovács, Péter Király, Gergely Tóth: Sample-size dependence of validation parameters

P10 Máté Mihalovits, S. Kemény: Application of regression control chart in pharmaceutical on-going stability study to detect out-of-trend results

P11 Zsolt I. Németh, I. Mészáros, Cs. Millei-Raffai, A. Vágvölgyi, R. Rákosa: Discrimination of aqueous solutions by PCA-DA assessment based on FT-IR spectrometry

P12 R. Rákosa, M. Vargovics, J. Jakab, Zsolt I. Németh: Applicability of FT-ATR-IR spectrometry in identification of *mycellium* cultures

P13 Adrián Pesti, E. Kontsek, G. Smuk, S. Gergely, A. Kiss: Mid-infrared imaging based lung cancer subtype determination

P14 Éva Pusztai, Sándor Kemény: Process capability indices when the usual assumptions fail, a tolerance interval approach

P15 Anita Rácz, Dávid Bajusz, Károly Héberger: Unsupervised data reduction: How to set the intercorrelation limits optimally.

P16 A. Rybińska-Fryca, A. Mikołajczyk, J. Łuczak, M. Paszkiewicz-Gawron, A. Zaleska-Medynska, M. Paszkiewicz and T. Puzyn: Thermal stability of ionic liquids under the conditions of synthesis of TiO2-based photocatalysts: chemometric studies

P17 Mariusz Sandomierski, Zuzanna Buchwald, Monika Zielińska, Adam Voelkel: Chemometric methods in characteristic of zeolites for specific applications

P18 L. Sipos, I. F. Boros, K. Madara, L. Csambalik, Attila Gere: Multi-criteria decision making–Comparing lettuce types by their phytonutrient content

P19 Daniel Szabó, A. Ács, F. Auer, B. Rojkovich, Gy. Nagy, P. Géher, G. Sármay, L. Drahos, K. Vékey: Analysis of protein glycosylation in *rheumatoid arthritis*

P20 Anastasia Surkova, A. Bogomolov, A. Legin, D. Kirsanov: Optical multisensor system for fat and protein determination in milk

P21 Kurt Varmuza P. Filzmoser, N. Fray, H. Cottin, S. Merouane, O. Stenzel, J. Kissel, C. Briois, D. Baklouti, A. Bardyn, S. Siljeström, J. Silén, M. Hilchenbach: Composition of cometary particles versus distance to sun during sample collection - based on multivariate evaluation of mass spectral data (Rosetta/COSIMA)

P22 Tsvetomil Voyslavov, E. Mladenova, R. Balkanska: Self-organizing maps as an approach for monofloral bee honeys botanical origin determination

P23 Máté Csontos, J. Elek, I. Bácskai, P. Arany, Infrared analysis of chemically modified 3D printed PLA scaffolds

P24 Zoran Stamenković, Ivan Pavkov, Milivoj Radojčin, Krstan Kešelj, Siniša Bikić, Attila Gere: Multicriteria optimization of raspberry convective drying processes

Lectures

What is (are) the best measure(s) of variable importance for interpretation and variable selection in multivariate regression models?

Olav M. Kvalheim[1]
[1] Department of Chemistry, University of Bergen, Norway
E-mail: olav.kvalheim@uib.no

Measures for variable importance are crucial for interpretation and variable selection using partial least squares (PLS) or other methods for latent variable regression modelling. Many measures have been proposed and assessed for their appropriateness and usefulness. The most used approach in chemometrics is without doubt variable influence in projections (VIP) [1], but other methods such as selectivity ratio (SR) [2-3] and the related method significance Multivariate Correlation (sMC) [4] as well as modifications of VIP adapted to Orthogonal PLS [5] have been proposed and compared [6-8]. The results of comparisons, however, are not conclusive. One reason for confusion is that the results partly depends on the purpose of the modelling, but also erroneous implementation of methods have occurred.

In this work, we discuss the assumptions and limitations of VIP, SR and sMC for interpretation and variable selection using PLS and compare results for three data sets from applications within mixture analysis, process analysis and metabolomics.

References

[1] S. Wold S, M. Sjöström and L. Eriksson L., *Chemometr. Intell. Lab. Syst*. **58** (2001) 109-130.
[2] T. Rajalahti, R. Arneberg, F.S. Berven, K.-M. Myhr, R.J. Ulvik, O.M. Kvalheim, *Chemometrics & Intell. Lab. Syst.*, **95** (2009) 35-48.
[3] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.-M. Myhr and O.M. Kvalheim, *Analytical Chemistry* **81** (2009) 2581-90.
[4] T. N. Tran, N. L. Afanador, L. M.C. Buydens and L. Blanchet, *Chemom. Intell. Lab. Syst.* **138** (2014) 153-160.
[5] B. Galindo-Prietoa, L. Eriksson and J. Trygg, *J. Chemometrics* **28** (2014) 623–632.
[6] C.M. Andersen and R. Bro R. *J. Chemometrics* **24** (2010) 728–737.
[7] M. Farrésa, S. Platikanova, S. Tsakovskib and R. Tauler, *J. Chemometrics* **29** (2015) 528–536.
[8] J. P.M. Andries, Y. Vander Heyden and L.M.C. Buydens, *Analyt. Chim. Acta* **760** (2013) 34-45.

Optimization of R^2 and related quantities by allocation

G. Tóth[1], *P. Király and D. Kovács*
Institute of Chemistry, Loránd Eötvös University
Pázmány s. 1/a 1117 Budapest, Hungary
[1] E-mail: toth@chem.elte.hu

Design of experiments is the field where we would like to get maximum necessary information from the minimal number of experiments. The basic idea is to find the most feasible independent variable set for the modelling. It can be performed *a priori* to the experiments, after some exploratory experiments or to select a best subset of large number of existing experiments. There are several ways to do it, we can use fix designs, e.g. factorial designs at different levels. We can apply an optimality design, where a mathematical extremum of a given optimality function calculated on the predictor and/or response data is used. There are several subset selection methods, from random methods to maximal orthogonal ones including different stratified schemes as well. These are sometimes connected to the task of splitting the sample into a training set and a test one.

The aim of our study is to check the effect of the allocation of the predictor variables on some performance and optimality parameters. We show which typical allocations provide good or weak performance parameters, how we can tune these parameters. We calculated also the correlation among the performance parameters and the optimality parameters to have some insight into the feasibility of using optimality design.

We show mostly our simulated data on sets containing two predictor variables and one response variable that can be adequately modelled by linear regression. The predictor variables were generated along different trends, but some fix designs were applied as well. The main generated trend was to pull the variables from the middle of the variable space to the corners of the accessible domain.

R^2, CCC and RMSE belonging to quantification of goodness-fit-internally seem to be improved if our data are close to the corners. In the case of RMSE a close to corner like arrangement was optimal. Surprisingly, in the case of 10 data points a three level composite fix design provided even better results. The internal robustness parameters like Q^2_{loo} and CCC_{loo} behave similarly, a close to corner like allocation is the most feasible, but the fix design is superior. The same is valid for the external predictivity parameters, Q^2_{F1}, Q^2_{F2} and Q^2_{F3}.

A widely accepted scheme is in design of experiments that the variance of the model parameters should be minimized. This value is also one that is minimal at a close to corner allocation, and again the fix design is superior. Our study is based on simulated data, so we know the 'exact' model. It means we can calculate an error integral for every model. This error integral depends on the allocation similarly to the model parameter variance.

Cross optimization can be performed, if a performance or optimality parameter highly correlates with a desired quantity, like the total parameter variance or integrated error. We found in the case of the total model parameter variance, that many of the performance and optimality parameters (e.g. R^2, CCC, Q^2_{loo}, A-, D-optimality) have rank correlations around 0.9. In the case of the integrated error the best rank correlations were around 0.7. Unfortunately, this does not mean, that the best sets of the cross-optimized features reasonably overlap. For example, the overlap of the best 1% is only 0.6 for e.g. R^2 and the model parameter variance, while in the case of integrated error they are very small, only 0.04. It means, cross optimization may perform weak, questioning the efficiency of optimal designs, especially that of fix design schemes.

QSAR model development from small data sets: A new workflow with integration of data curation, double cross-validation and consensus prediction tools

Pravin Ambure[1], *Agnieszka Gajewicz*[2], *M Natalia DS Cordiero*[1] *and Kunal Roy*[3]
[1]REQUIMTE/Department of Chemistry and Biochemistry,
University of Porto, 4169-007 Porto, Portugal
[2]Laboratory of Environmental Chemometrics, University of Gdansk, Gdansk, Poland
[3]Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
E-mails: kunalroy_in@yahoo.com , kunal.roy@jadavpuruniversity.in
URL: https://sites.google.com/site/kunalroyindia/

Quantitative structure-activity/property relationship (QSAR/QSPR) models [1] are used mainly for two purposes, prediction of the endpoint values for untested chemicals for data gap filling, and physicochemical and mechanistic interpretations of the structure-response relationships. We frequently come across small data sets (with number of data points 25 -50) for some specialized endpoints. Due to shortage of experimental data for such endpoints, it is desirable to develop (QSAR/QSPR) models in order to fill data gaps. However, it is difficult to develop a properly validated and robust QSAR/QSPR model from a small data set due to several reasons, including not using a part of the available data for model development for test set validation, bias in descriptor selection due to a fixed composition of the small training set, presence of outliers regarding both chemical and biological domains in the training data. To address these problems, we suggest here a workflow involving modeling of the whole small data set (*i.e.*, without data set division) integrating three major steps: data curation, double cross-validation and consensus predictions. In the data curation step, we try to identify structural and response range outliers (compounds which are sufficiently different from the rest with respect to chemical features and/or response values) [2] and also activity cliffs (compounds which are similar in chemical features, but much different in response values). For double cross-validation, we carry out leave-many-out cross-validation in different iterations [3] and select the best model based on the lowest error of the respective validation sets [4]. Finally, the best selected models are applied for consensus predictions for the query compounds using simple average and weighted average (weighting based on the mean absolute error computed from leave-one-out prediction errors of all training compounds) of predictions [5]. Finally, to demonstrate the applicability of the workflow, case studies are performed using a few small data sets.

The suggested workflow is available for free public use in the form of a software tool in the site https://dtclab.webs.com/software-tools

References

[1] Roy K (editor), Advances in QSAR Modeling. Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences. Springer, 2017, http://www.springer.com/in/book/9783319568492
[2] Roy K, Kar S, Ambure P, *Chemometr. Intell. Lab. Syst.*, **145**, 2015, 22-29, http://dx.doi.org/10.1016/j.chemolab.2015.04.013
[3] Roy K, Ambure P, *Chemometr. Intell. Lab. Syst.*, **159**, 2016, 108-126, http://dx.doi.org/10.1016/j.chemolab.2016.10.009
[4] Roy K, Das RN, Ambure P, Aher RB, *Chemometr. Intell. Lab. Syst.*, **152**, 2016, 18-33, http://dx.doi.org/10.1016/j.chemolab.2016.01.008
[5] Roy K, Ambure P, Kar S, Ojha PK, *J. Chemometr*, **32**, 2018, e2992, http://dx.doi.org/10.1002/cem.2992

Deep Ranking Analysis by Power Eigenvectors (DRAPE): a wizard for ranking and multi-criteria decision making.

Roberto Todeschini, Francesca Grisoni and Davide Ballabio
Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca. P.zza della Scienza 1, 20126, Milano, Italy.
E-mail: roberto.todeschini@unimib.it

Ranking and multi-criteria decision-making approaches are useful tools to analyse multivariate data and obtain useful insights into data structure and the relationships between samples and variables. In this study [1], we present a new ranking approach, named Deep Ranking Analysis by Power Eigenvectors (DRAPE), which is based on the Power-Weakness Ratio analysis and provides a set of sequential rankings. Such a sequential ranking procedure allows to gather deeper insights into the analysed dataset. Moreover, by a "retro"-regression procedure, the relevance of each variable in determining the final rankings can be assessed, while a consensus ranking can be obtained by a Principal Component Analysis (PCA). In this lecture, we present the theory of the novel method, and show two applications to real datasets.

Reference

[1] Roberto Todeschini, Francesca Grisoni, Davide Ballabio, Deep Ranking Analysis by Power Eigenvectors (DRAPE): A wizard for ranking and multi-criteria decision making *Chemometrics and Intelligent Laboratory Systems* **191** (2019) 129–137. https://doi.org/10.1016/j.chemolab.2019.06.005

Transfer of multivariate regression models based on 1D and 2D spectra between high-resolution NMR instruments: application to lecithin and heparin

Yulia Monakhova[1,2], *Bernd Diehl*[1]
[1] Spectral Service AG, Emil-Hoffmann-Straße 33, 50996 Cologne, Germany;
[2] Institute of Chemistry, Saratov State University, Astrakhanskaya Street 83, 410012 Saratov, Russia; E-mail: yulia.monakhova@spectralservice.de

Chemometric techniques have become essential to model NMR profiles of complex mixtures. However, the routine implementation of quantitative multivariate models on different high-resolution NMR devices is not trivial and requires standardization.

In this study different calibration transfer methods were applied and compared for authenticity control of lecithin vegetable origin and molecular weight of heparin. Three NMR spectrometers with magnetic field strength of 500 MHz (BBFO Plus probe and BBFO Prodigy cryo probe) and of 600 MHz (BBP cryoprobe) were utilized in this study. For multivariate modelling piecewise direct standardization (PDS) and direct standardization (DS) were employed.

In case of lecithin the application of PDS resulted in almost twofold enhancement in prediction of partial least squares (PLS) model of 1D NMR spectra in comparison with the modelling of non-standardized data. PDS also showed the best performance for estimating heparin molecular weight using 2D DOSY measurements resulting in a significant decrease in root mean square error of prediction (RMSEP) from 647 Da and 393 Da without standardization to 513 Da and 135 Da when PDS was applied for heparin and low molecular weight heparin, respectively.

The study showed that standardization methods are useful when quantitative multivariate model has to be applied for the spectra recorded on a secondary NMR spectrometer even with another magnetic field strength.

Acknowledgement
The work was supported by the Russian Science Foundation (project 18-73-10009).

Chemometric support in physicochemical evaluation of industrial materials

A. Voelkel[1], *B. Strzemiecka*[1], *M. Sandomierski*[1]
[1]Institute of Chemical Technology and Engineering,
Poznan University of Technology, Poland,
E-mail: Adam.Voelkel@put.poznan.pl

In all applications of nanomaterials control over their surface properties are crucial for the interface aspects. Indeed, the strong interfaces are critical regions that ensure not only the long term use of the polymer composites but contribute also to the performances of the end products. For this reason, they must be controlled at the molecular level with appropriate chemistry strategies in order to ensure their long term stability.

In this paper the usefulness of Inverse Gas Chromatography technique in the nanomaterials characterization is presented. It is shown that this technique is sensitive and accurate to notice changes in monolayer on the surface after e.g. its modification, influence of storage, synthesis conditions. Moreover, it enables to study nanomaterials in the real conditions which means no special pretreatment is needed before measurements *e.g*. degassing, heating. The measurements are performed for materials just as received.

Dynamic Vapor Sorption enables to test adsorption properties by using different adsorbates (such as water, alcohols, alkanes) the same as Inverse Gas Chromatography without special pretreatment of the investigated materials under real conditions.

Both techniques can be easily applied for quick and accurate surface characterization of wide range of nanomaterials such as fillers for polymers, biomaterials.

The properties of the aluminosilicates with the different Si/Al ratio are closely related to the amount of aluminum and silicon in the tested materials. The surface properties of the synthesized aluminosilicate does not depend on the source of alumina and silica used as substrate. The particle size of the material agglomerates visibly increases with the increase of the amount of aluminum. The water sorption increased with the increase of the silica amount.

Acknowledgement
This work was supported by PUT grant 03/32/SBAD/0900.

Assessment of amino acids' variability in plasma, cerebrospinal fluid and exhaled breath condensates in pediatric leukemia patients

Lucyna Konieczna, Anna Krawczyńska, Tomasz Bączek
Department of Pharmaceutical Chemistry, Medical University of Gdańsk, Gdańsk, Poland,
E-mail: tomasz.baczek@gumed.edu.pl

The main goal of the study was searching for new metabolic signatures of selected amino acids (AA) in plasma (PL) and cerebrospinal fluid (CSF) samples as well as in exhaled breath condensates (EBC). It was confirmed with the use of new bioanalytical methodology that composition of AA in PL, CSF and EBC alters during acute lymphoblastic leukemia (ALL) in pediatric patients. Levels of the following AA were measured: alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, homoarginine, hydroxyproline, isoleucine, leucine, lysine, methionine, norvaline, phenylalanine, proline, serine, threonine, tryptophan, tyrosine and valine. Profiling of AA in PL and CSF in children with ALL was established in the first step of the research.

Statistically significant differences between three analyzed groups: children with ALL at the moment of diagnosis, patients during chemotherapy – 15 days or 33 days from the diagnosis and healthy controls were noticed. Results confirmed relationships between AA levels and diagnosis of leukemia and the moment of therapy. These results were then combined with the analysis of AA in completely non-invasive matrix, namely EBC. AA profiles before and after the therapy were observed as different ones for the tested EBC samples. And, what is even more significant and interesting, profiles of AA observed for patients after the therapy resembled the profiles obtained for healthy volunteers.

References

[1] L. Konieczna, M. Pyszka, M. Okońska, M. Niedźwiecki and T. Bączek, *J. Chromatogr. A*, **1542** (2018) 72-81.

Chemometrics and machine learning for analysis of Raman-related data

Thomas Bocklitz[1,2]
[1] Institute of Physical Chemistry and Abbe Center of Photonics (IPC), University of Jena
[2] Leibniz Institute of Photonic Technology (IPHT)
E-mail: thomas.bocklitz@uni-jena.de

Untargeted imaging modalities, like Raman spectroscopy based imaging or multimodal imaging, *e.g.* the combination of coherent-anti-Stokes Raman scattering (CARS), second-harmonic generation (SHG) and two-photon-excited fluorescence (TPEF), could already demonstrate their unique potential for disease diagnostics and treatment monitoring [1,2]. For example, both imaging modalities were utilized for cancer diagnostics and for the quantification of chronic inflammatory bowel diseases. These studies indicated that the utilization of the huge potential of multimodal imaging and Raman spectroscopic imaging for diagnostic tasks require sophisticated machine learning methods and chemometrics. These computational methods aim at the translation of the optical imaging data into medical information. For multimodal images this translation process is composed of image standardization, correction procedures, image enhancement methods, feature extraction procedures and/or deep learning methods. For Raman spectroscopic imaging correction and standardization procedures for the spectral data are necessary [2]. After this so-called pre-processing was carried out to clean the data, regression and/or classification methods can be applied to extract medical relevant information. This contribution shows how the whole data pipeline needs to be tailored to the measurement modalities and how the feature extraction or the deep learning methods have to be adapted to the samples and the specific diagnostic task to be solved.

Acknowledgment
Financial support of the EU, the 'Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft', the 'Thüringer Aufbaubank', the Federal Ministry of Education and Research, Germany (BMBF), the German Science Foundation, the Fonds der Chemischen Industrie, the Carl-Zeiss Foundation and Leibniz association via the ScienceCampus 'InfectoOptics' are greatly acknowledged.

References

[2] T. Bocklitz; S. Guo; O. Ryabchykov; N. Vogler and J. Popp, *Anal. Chem.*, **88** (2016) 133-151
[1] N. Vogler; S. Heuke; T. W. Bocklitz; M. Schmitt and J. Popp, *Annual Review of Analytical Chemistry*, **8** (2015) 359-387

Incorporating Left-Censored Bioactivity Data in Predictive Regression Models

K. Baumann[1], *V. Schlenker*[1], *A. ter Laak*[2], *N. Heinrich*[2]
[1] Institute of Medicinal and Pharmaceutical Chemistry, Technische Universität Braunschweig, Beethoven Straße 55, 38106 Braunschweig, Germany;
E-Mail: k.baumann@tu-braunschweig.de
[2] Bayer AG, Drug Discovery, Pharmaceuticals, 13342 Berlin, Germany

In industrial drug discovery, bioactivity data for weakly active compounds are often incompletely measured so that no exact pIC_{50} or pK_i value is known. In these cases, it is only known that the pIC_{50} or pK_i value is smaller than a certain cut-off value. Data of this type are called left-censored. Such data frequently occur in econometrics and environmental chemistry (values below the determination or detection limit) and efficient regression algorithms to include those data into calibration models are well known [1]. However, in the latter two application areas only few predictors are typically processed. Regression algorithms for handling hundredths or thousands of predictors are not available. For right-censored survival data, high-dimensional regression algorithms have been published. Yet, the censoring mechanism in these cases is very different from the one for left-censored data.

Here, we present semi-parametric and nonparametric extensions of linear latent variable regression models (*i.e*. Principal Component Regression and Partial Least Squares Regression) as well as nonlinear decision forests to left-censored data. Robustness and predictive capability for the different techniques will be studied. Critical issues with respect to estimating intercepts and the predictive ability will be discussed. The specifically tailored algorithms will also be compared to the naïve case where the censored data are handled as if they were uncensored. The competition between tailored and naïve techniques boils down to a typical bias-variance tradeoff, which will be illustrated with data from pharmaceutical industry.

Reference

[1] D. R. Helsel, *Statistics for censored environmental data using Minitab® and R*. John Wiley & Sons, Hoboken, NJ, USA, 2012, 2nd Ed.

Numerical Representations of Metabolic Systems

Age K. Smilde[1] *and Thomas Hankemeier*[2]
[1] Biosystems Data Analysis, University of Amsterdam, Netherlands (a.k.smilde@uva.nl)
[2] LACDR, Leiden University, Netherlands

In metabolomics we perform measurements. These measurements produce numbers, which is not the same as data: data are numbers including their meaning. Data can have different properties depending on how the numbers are measured. One property is measurement scale, which ranges from ratio-scaled data to nominal-scaled data. Another property is comparability across rows and columns of our data table. These different properties will be explained by simple examples from metabolomics data analysis practice. It will also be shown what the repercussions are of those properties for the type of statistical analysis to employ.

The use and misuse of *p* values and related concepts

Richard G. Brereton
School of Chemistry, Cantocks Close, Bristol BS8 1TS, United Kingdom,
E-mail: r.g.brereton@bris.ac.uk

The origins of chemometrics were primarily in analytical and physical chemistry, where answers were often known to higher precision than could be predicted by multivariate measurement techniques such as spectroscopy. Hence hypothesis / significance testing was of little interest in the first few decades. As the application of chemometrics moves into areas such as metabolomics and heritage science, the aim of experiments often is to prove a hypothesis or determine the significance of a variable, requiring different objectives to classical chemometrics.

The presentation discusses the origins of the idea of *p* values and significance tests from the work of Fisher, and the related but distinct concepts of type 1 and 2 errors and hypothesis tests from the work of Neyman and Pearson, both in the early 20th century. These ideas became merged in the approach of Null Hypothesis Significance Tests in the 1940s. Various related definitions including the power of a test, the prevalence of true positives, and false positive rates will be described, and how many confuse these ideas.

The relationship between *p* values and false positive rates will be described, how a *p* value of 0.05 could, for example, be the result of a false positive rate of 0.5 or 50%. The reproducibility of *p* values is quite low, and they show wide variability as will be demonstrated. The current controversy over the use of *p* values will be discussed.

In multivariate studies, *p* values have limited use unless variables are orthogonal, and the dependability of statistical indicators such as *F* and *t* values on orthogonality is described. There are problems interpreting *p* values and related statistics if variables (such as potential markers) or factors (such as squared terms in models) are not orthogonal.

With the widespread expectation especially of biomedical scientists to cite *p* values, there is often unrealistic pressure on chemometricians, in order to justify papers and grants. Many are not aware of the problems using these methods, or the distinction between false positive rates and *p* values.

Mixture of QSAR Models–Learning Gating Functions to Combine of pK_a Predictions

János Abonyi
MTA – PE Lendület Complex Systems Monitoring Research Group
Department of Process Engineering, University of Pannonia
POB. 158, Veszprém, H-8201, Hungary
E-mail: janos@abonyilab.com

Quantitative structure-activity relationships (QSARs) models predict physical properties or biological effects based on chemical structure descriptors. Some studies investigate methods for combining multiple QSAR tools to gain better predictive performance for various toxic endpoints [1]. The mixture of experts (ME) is one of the most popular methods developed to combine data-driven models. ME is established based on the divide-and-conquer principle in which the problem space is divided between a few neural network experts, supervised by a gating network. Applications of MEs have been seen in various areas, but the potential of gating function-based exploration and control of the application domain of QSAR models has not been studied yet.

This work incorporates QSAR models into the ME model structures. As the combined experts (QSAR tools) are not updated, our work focuses to the model structures, the types of gating functions, the gradient descent and Expectation-Maximization (EM) methods that minimize the goal-oriented ME error function, the selection of the input variables and the exploration of the applicability domains of the QSAR models.

The applicability of the concept is demonstrated based on combining four acidic pK_a predictions of the OECD QSAR Toolbox. Predicting pK_a values (logarithm of the acid dissociation constant) of pharmaceutical substances is both challenging and essential as the knowledge of the possible ionization states is essential in drug development. The analysis of the pK_a OASIS database (1930 chemicals) illustrates that mixture of Chemaxon, OASIS Consensus, Electric, and Regression pK_a models improves the accuracy of the individual models even when the gating function is not based on additional molecular descriptors, so the ME utilizes only the predictions of the QSAR models.

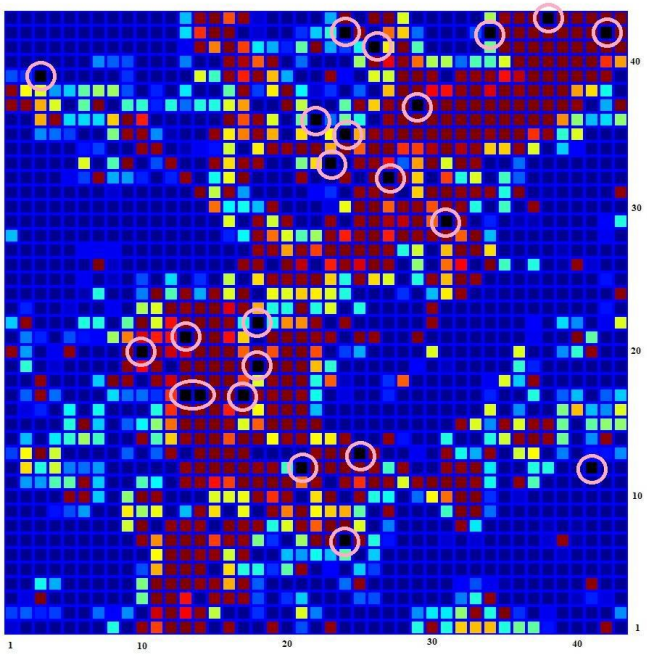
References

[1] Prachi Pradeep, Richard J. Povinelli, Shannon White & Stephen J. Merrill, *Journal of Cheminformatics*, **8** (2016) Article number 48.

P-gp transport activity–in connection to the efflux of toxicants or drugs

Liadys Mora Lagares[1,2], *Nikola Minovski*[1], *Viktor Drgan*[1], *Marjan Tušar*[1], *Marjana Novič*[1]
[1] National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia
E-mail: marjana.novic@ki.si
[2] Jožef Stefan International Postgraduate School,
Jamova 39, SI-1000 Ljubljana, Slovenia

P-glycoprotein (P-gp) is a transmembrane protein, playing significant roles in the process of drug discovery. P-gp affects absorption, distribution, and elimination of different compounds and it is mainly expressed in intestines, liver, kidneys, heart, colon, and placenta. P-gp is responsible for resistance of cells to xenobiotics, particularly the anticancer drugs, giving rise to the multidrug resistance (MDR) phenomenon by mediating the active transport of these drugs from the intracellular to the extracellular compartment [1]. Moreover, studies showed that P-gp contributes to resistance to pesticides in certain pest species, and it also contributes to decrease toxicity by removing compounds from cells in mammals [2]. We have developed an *in silico* multiclass classification model capable to predict the probability of a compound to interact with P-gp (inhibitors, substrates, and non-interacting compounds) [3]. In this study the model was applied to several compounds, either drug or drug-like candidates, in order to pay attention to the likelihood of a compound being transported by P-gp, since this contributes to whether a compound actually reaches its intended target or it is removed from the cell before exerting its action. Distribution of the positive (blue), negative (red) and 24 double-negative compounds (P-gp non-inhibitor, P-gp non-substrate) in the response map of the non-active class is shown below. Pink circles represent the only absolutely certain non-active compounds in the dataset, double-tested as inhibitors and substrates.



References

[1] E.M. Leslie, R.G. Deeley, S.P. Cole, *Toxicol. Appl. Pharmacol*. **204**, (2005), 216–237.
[2] K. Sreeramulu, R. Liu, F.J. Sharom, *Biochimica et Biophysica Acta* **1768**, (2007) 1750–1757.
[3] L.M. Lagares, N. Minovski and M. Novič, *Molecules*, **24**, (2019) 2006; doi:10.3390/molecules24102006.

Robustness control in bilinear modeling based on maximum correntropy

V. Fonseca Diaz[1], *W. Saeys*[2]
[1] KU Leuven Department of Biosystems, MeBioS, Kasteelpark Arenberg 30, 3001 Leuven, Belgium. E-mail: valeria.fonsecadiaz@kuleuven.be;
[2] KU Leuven Department of Biosystems, MeBioS, Kasteelpark Arenberg 30, 3001 Leuven, Belgium. E-mail:wouter.saeys@kuleuven.be

In this work, we present the development of bilinear regression models for multivariate calibration based on maximum correntropy criteria (MCE) whose robustness can be easily controlled. MCE regression methods have proven to be effective in the presence of non-gaussian noise or outliers in the data and are particularly competitive when the degree of robustness against them should be controlled. Within the context of linear regression models with data compression, MCE versions of PCR and SIMPLS are implemented based on literature. In addition, a modification of MCESIMPLS is proposed, which is named Weighted MCESIMPLS. Thanks to the controllable robustness of MCE models, observations are up-weighted or down-weighted during the calibration process rendering robust models with soft discrimination of samples. Such a weighting represents an important advantage, especially for cases when datasets contain borderline outliers or distributions with heavier tails. Simulations and applications to real cases are presented to demonstrate and discuss the performance of the different methods in comparison to classical and robust PCR and SIMPLS (Fig. 1). The presented methods based on MCE have been implemented in Python using the machine learning sci-kit learn classes and were found to have a low computational cost and a straightforward implementation.

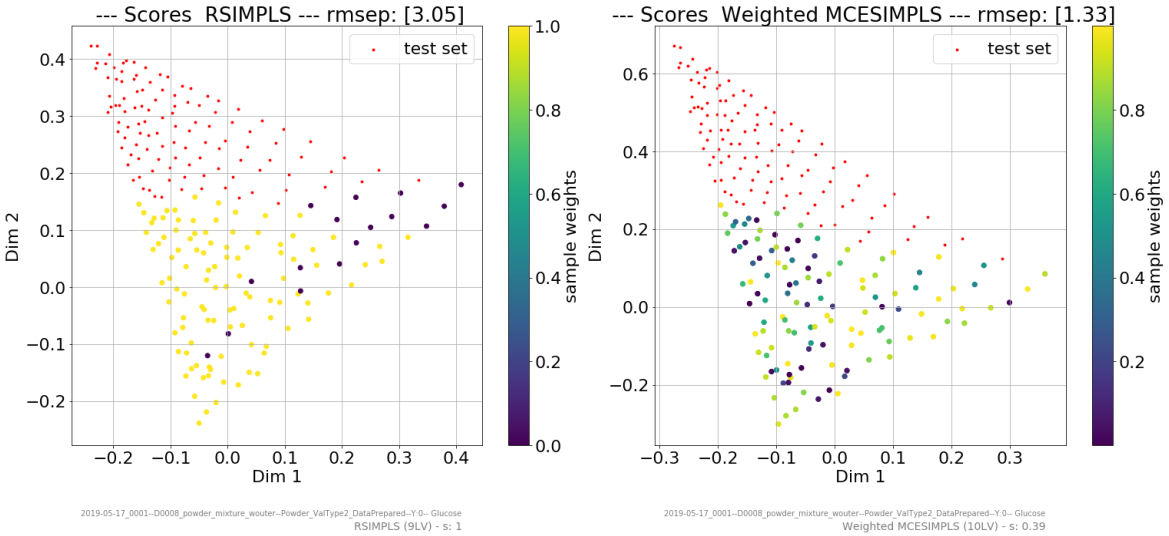


Figure 1. *Visualization of the sample weighting by RSIMPLS and Weighted MCESIMPLS for the prediction of the glucose content in powder mixtures of Glucose, Casein and Lactate.* The corners of the triangle correspond to the pure powders. *Because of the strategic split between calibration and test data, RSIMPLS completely discards the corner of pure Glucose, challenging the prediction on the test set. Weighted MCESIMPLS manages to up-weight samples at this corner, contributing to improve the predictions on test samples as shown by the drastic reduction in the RMSEP value.*

References

[1] M. Hubert, P. J. Rousseeuw and S. Van Aelst, *Statistical Science*, *23*(1) (2008), 92–119.

[2] R. He, B. Hu, W. Zheng and X. Kong, *IEEE Transactions on Image Processing*, *20*(6) (2011), 1485–1494.

[3] Y. Feng, X. Huang, L. Shi, Y. Yang and J. Suykens, *Journal of Machine Learning Research*, *16 (2015)*, 993–1034.

[4] J. Peng, L. Guo, Y. Hu, K. Rao and Q. Xie, *Chemometrics and Intelligent Laboratory Systems*, *161* (2016), 27–33.

Improved rank estimation using randomization and dependency

Elaheh Talebanpour Bayat,[1] *Knut Baumann,*[2] *Bahram Hemmateenejad*[1]
[1] Chemistry Department, Shiraz University, Shiraz, Iran
E-mail: hemmatb@shirazu. ac.ir
[2] Institute of Medicinal and Pharmaceutical Chemistry,
University of Technology Braunschweig, Germany

An important issue in principal component analysis or factor analysis is estimating the correct chemical rank of a data matrix prepared from chemical instruments. This allows chemists to know the number of chemical species/chemical equilibria in the chemical system under study. Because of this, rank estimation has been the subject of many different research studies and hence different indices have been suggested to determine the chemical rank. Chemical data are always associated with artefacts like noise (especially heteroscedastic noise) and variable background. These make rank estimation a hard task. Very recently, we used two non-parametric measures of dependency including maximum information coefficient (MIC) and distance correlation (DC) and proposed new indices for rank estimation [1]. While these indices showed good sensitivity and reliability for rank estimation in several simulated and real data sets, they failed in the presence of sever peak overlapping or highly variable backgrounds.

Here, we combined our dependency approach with the randomization method [2], which has been used previously by different researchers for rank estimation too, and developed improved indices for estimating matrix rank. In this approach, the dependency indices (dependency between principal components) are calculated two times iteratively: for the native data matrix and for the randomly shuffled data matrix. During iteration, the contribution of principal components is removed from original data matrix consecutively. The algorithm converges when the no significant differences are observed between the native and randomized data.

Interestingly, it was found that the dependency measures calculated for the randomized data are almost the same for all PCs. This means that dependency measure of the randomized data matrix is not required to be calculated for all PCs. It is just enough to calculate it for the randomized original data. Since randomization is repeated many times, the previous rank estimation methods based on randomization required long computation time. However, in our approach the computation time decreases dramatically.

Our new approach was done on different simulated data sets in the presence of severe perturbations e.g. high collinearities between variables, low signal to noise (SpN) ratio and etc. The results revealed that our approach has superiority performance in comparison with the other strategies in term of computational time and accurate rank determination. However, multiple procedure could also assess true dimensionality with consuming time (e.g. around 64%) more than single protocol in MIC and DC for simulated data sets.

References

[1] E. T. Bayat, B. Hemmateenejad, M. Akhond, M. M. Bordbar, K. Baumann, *J. Chemom*., **33** (2019) e3102.
[4] S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold, K. Faber, *J. Chemom*., **21** (2007) 427-439.

Towards easier and more reliable multivariate curve resolution of hyperspectral images

Cyril Ruckebusch
University Lille, LASIR CNRS, Lille, France
E-mail: cyril.ruckebusch@univ-lille.fr

Hyperspectral imaging is an analytical technique by which the spatial information available in regular images and the spectral information provided by multivariate instrumental platforms are combined at the pixel level. Since the seminal paper by Geladi [1], hyperspectral image analysis has been, in chemometrics, mainly focused on the spectroscopic signals, which was implicitly assumed to carry some selective information, and the full data set (all the spectral pixels) was almost systematically exploited. On the other hand, as already pointed by Craig [2] in 94, pure spectrum pixels are sufficient to describe the whole variability of the measured pixels for linear unmixing and can be taken as a reduced data set.

With these ideas in mind, we resume in this presentation our recent efforts towards easier and more reliable multivariate curve resolution of hyperspectral images. We will start by introducing image processing constraints which can be imposed on the solutions of MCR-ALS at each least-squares iteration. These constraints provide an easy and flexible way to achieve both the aforementioned complementary goals as well as better unmixing of highly spectrally-mixed and spatially-structured hyperspectral imaging data. We will highlight the effect of these spatial constraints on the rotational ambiguity of the solutions provided by MCR-ALS of hyperspectral images.[3,4] We will then point that even when dealing with highly mixed hyperspectral imaging data, not all the measured spectral pixels are equally important for the resolution of the data at hand. Some of those pixels, those on the outer envelope of the data, form a representative subset of the data that can be taken as a reduced (compressed) data set for multivariate curve resolution of hyperspectral images. This reduced data, sometimes consisting of a few spectral pixels only, carries the same amount of information as all the spectral pixels. In practice, these pixels can be found by taking the support (vertices) of the convex hull of the actual data distribution in the PCA subspace. Our results [5] show that multivariate curve resolution on reduced data yield essentially the same solution as for full data. We also point that even for complex strongly mixed hyperspectral imaging data, the compression ratio can be huge. In addition, removing noisy, redundant and highly mixed pixels simplifies the MCR problem a lot: rank determination is easier, initial estimations are more robust and calculation is of course faster.

References

[1] P. Geladi, *et al.*, *Trends in Analytical Chemistry*, 1992;
[2] M.D. Craig *IEEE Transactions on geoscience and remote sensing*, 1994;
[3] S. Hugelier *et al. Journal of Chemometrics*, 2015;
[4] M.Ghaffari, *et al. Analytica Chimica Acta*, 2019;
[5] M.Ghaffari, N. Omidikia, C. Ruckebusch, *submitted to Analytical Chemistry.*

Spatial advantage of hyperspectral imaging and construction of rigorous classifiers

M. Daszykowski, L. Pieszczek
[1] Institute of Chemistry, University of Silesia in Katowice, 9 Szkolna Street, 40-006 Katowice, Poland,
E-mail: michal.daszykowski@us.edu.pl

Hyperspectral imaging is recognized as a well-suited technique for revealing the distribution of chemical components on a sample surface. This feature is particularly useful when sample homogeneity is of great concern. Taking also into account that the so-called bush-brum hyperspectral cameras are very rapid and non-destructive, these imaging systems often support online quality assessment of products at diverse production environments. Combining the hyperspectral imaging with advanced chemometric processing of multivariate spectral data offers the possibility to extract chemical information from images and perform the qualitative and quantitative analysis.

In our study, we have explored the spatial advantage of hyperspectral imaging [1] in order to enhance the performance of one-class classification models and advance their validation. The proposed framework involves the Monte Carlo procedure [2,3] introduced to facilitate the selection of the optimal number of components and enable the comprehensive validation of a classification model. In particular, the Monte Carlo procedure offers the possibility to calculate several different figures of merit in a function of the number of components at every iteration. Therefore, with little effort, one can also extend figures of merit with the corresponding uncertainties of their estimation.

The prediction power of any classification model depends on the representativeness of training samples used to construct logic rules. In this study, we also examine the performance of rigorous classification models. It assumes that logic rules are optimized, taking into account the variability explained by training samples and samples from the remaining groups.

Acknowledgment
This research was supported by the National Science Centre, Poland (research grant no. UMO-2018/29/N/ST4/01547)

References

[1] L. Pieszczek, M. Daszykowski, *Chemometr. Intell. Lab. Syst*., **187** (2019) 28-40.
[2] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, *Analyst*, **141** (2016) 1060-1070.
[3] L. Pieszczek, H. Czarnik-Matusewicz, M. Daszykowski, *Meat Science*, **139**, (2018) 15–24.

Determination of plastic particle size using discreet information hidden in NIR-HSI images

L. Pieszczek, M. Daszykowski
Institute of Chemistry, University of Silesia in Katowice, 9
Szkolna Street, 40-006 Katowice, Poland,
E-mail: pieszczek.lukasz@gmail.com

Usually, the diffuse near-infrared spectroscopy is used to study the chemical composition of solid samples. Non-chemical sources of variation are frequently considered as uninformative part of the instrumental signal. This is why in typical applications of the diffuse near-infrared spectroscopy the scattering effect is considered as a disturbing factor. To date many pre-processing methods have been developed to account for scattering. Among them, the multiplicative scatter correction is probably most popular.

The first attempts to use scattering as a valuable source of analytical information has been described in the literature more than 30 years ago [1]. At the same time, successful and comprehensive implementations of this idea for solving practical analytical problems were published just in the last few years [2,3]. Recent publications confirm and further extend the use of scattering for modeling [3], but they are mainly dealing with analytical problems encountered in the field of pharmaceutical technology.

A large part of production involves the production of polymer materials and their further processing. Effective control of the size of polymer granules, used for instance, during the injection compression molding process, is a step of process control aiming to prevent unexpected changes during plastic production.

The aim of this study was to evaluate the particle size distribution of polymer samples using the NIR hyperspectral camera. Two types of polymers, PMMA and HDPE, were grounded, separated by mechanical sieves, and split into 8 different particle size fractions (size range between 0.075 and 1.5 mm). Three replicates for each polymer fraction were prepared. Obtained polymer samples were measured with the FX17e Specim camera (spectral range between 900 and 1700 nm). The scattering regression models were built using the principal component regression and the partial least squares regression. To test the performance of these models three independent fractions of polymers were used. Obtained predictions indicated relatively small differences between the observed and modeled size of polymer particles. Therefore, one can conclude that size of plastic particles can be estimated using the scattering effect characterized by the NIR-HSI technique.

Acknowledgment
This research was supported by the National Science Centre, Poland (research grant no. UMO-2018/29/N/ST4/01547)

References

[1] J. L. Ilari, H. Martens and T. Isaksson, *Applied Spectroscopy*, **42** (1988) 722–728.
[2] W. Kessler, D. Oelkrug and R. Kessler, *Analytica Chimica Acta*, **642** (2009) 127–134.
[3] V. Pauli, Y. Roggo, P. Kleinebudde and M. Krumme, *European Journal of Pharmaceutics and Biopharmaceutics*, **141** (2019) 90–99.

Uncertainty SIMCA–a classification method that includes measurement uncertainty information

I. Stanimirova[1]
[1] Institute of Chemistry, University of Silesia in Katowice, 9 Szkolna Street, 40-006 Katowice, Poland
E-mail: istanimi@us.edu.pl

Soft independent modeling of class analogy, SIMCA [1], is a classification technique, the classification rules of which are created by constructing a classic principal component analysis, PCA, model of definite complexity for each group separately. In general, the PCA solution is only optimal when the errors for all measurements that were collected are independent and identically distributed (*i.i.d*.) normal. However, this assumption is often not met when analyzing data obtained from complex natural samples for which the measurement uncertainty varies systematically with the magnitude of the signal. A straightforward way to improve the classification solution for the collected data is to use methods that incorporate *a priori* knowledge about possible sources of variation in the modeling process instead of classic approaches. Thus, in order to create an uncertainty SIMCA approach either the maximum likelihood principal components analysis, MLPCA [2], or maximum likelihood common factor analysis, MLCFA method [3] can be used instead of the classic PCA. Both, MLPCA and MLCFA, methods have been proposed to solve the bilinear decomposition problem for data in which variables are subject to errors in measurement.

In this work, we compare the properties of these two candidate methods and illustrate them for simulated and hyperspectral imaging data. The evaluation of the complexity of such an uncertainty SIMCA model and the prediction for the independent test set are also important issues that will be addressed.

References

[1] S. Wold, *Pattern Recogn*., **8** (1976) 127–139.
[2] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, *J. Chemometrics*, **11** (1997) 339-366.
[3] D.N. Lawley, A.E. Maxwell, *Biometrika*, **60** (1973) 331-338.

Calibration model transfer between optical multisensor systems and full-scale spectrometers

A. Surkova[1,2], *A. Bogomolov*[1,2], *A. Legin*[1], *D. Kirsanov*[1]
[1] Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia;
E-mail: melenteva-anastasija@rambler.ru
[2] Samara State Technical University, Samara, Russia;

The advances of modern optical engineering allow creating simplified analytical devices based on inexpensive commercially available light sources, optical fibers, and detectors. The typical example of such devices is optical multisensor system consisting of several light-emitting diodes (LEDs) and a photodetector.

A built-in multivariate calibration model in optical multisensor system is required for compensating the lack of sensitivity and selectivity of individual sensor channels. Since each optical single- or multisensor system has a specific response pattern (absorption maxima, shape of the analytical signal, width of emission bands, etc.), it is nearly impossible to standardize such devices and their components at the production stage. Therefore, an individual calibration model is required for each optical multisensor system and this can be time-consuming and expensive. The re-calibration of each new device can be avoided using model transfer techniques, i.e. the adaptation of the calibration model developed for one instrument to another by mathematical transformations.

Various methods of model transfer were suggested in literature. The calibration transfer procedure can be performed either correcting the regression model parameters (for example, slope and bias correction) or by transforming spectral data of the second instrument into the format of the first instrument prior in order to use the same calibration model. In the present work, direct standardization and slope and bias correction methods were tested to transfer calibration model between the laboratory full-scale spectrometer and 4-channel LED-based optical sensors. The set of 25 samples consisting of cobalt and copper aqueous solutions prepared according to the diagonal design was measured with UV/Vis/NIR spectrometer and three LED-based optical multisensor systems. The prediction performance of different model transfer methods was compared and discussed.

The proposed model transfer approach can be widely applied in practice as it allows using the laboratory calibration models in production process control, and *vice versa*.

Acknowledgement

This study was supported by the RFBR-NSFC project #18-53-53016 GFEN_a,

New Psychoactive Substances: Data processing of spectra of RAMAN portable handheld devices

Claude Guillou[1], *Lyoma Guillou*[2],
[1] European Commission - Directorate General Joint Research Centre,
Directorate F – Health, Consumers and Reference Materials,
via E. Fermi, 2749 21020 Ispra (VA) - Italy;
[2] Free lance Data processing, Le Kremlin-Bicètre France

The Joint Research Centre is providing scientific support to the Customs Laboratory European Network for the identification of unknown substances and in particular for New Psychoactive Substances (NPS) [1]. NPS, also known as "legal highs" or "designer drugs", are recently introduced recreational drugs, whose chemical nature or effects may be similar to well-known substances of abuse such as cannabis, cocaine, MDMA, morphine, heroine, etc... Among them the class of synthetic opioids and fentanyl analogues are of very high concern in reason of their very high potency and the quick and strong addiction that they induce in the users of such substances. They also constitute a high risk of occupational health exposure for Customs or Police officers in charge of control of seizures and inspection of postal packages.

Raman spectroscopy allows the recording of spectra through transparent packaging reducing thus the risk of exposure to harmful substances. This makes it very attractive for controls by Customs which are now increasingly equipped with Raman handheld portable devices.

We will present a data processing workflow, developed in R, of heterogeneous Raman data (various brands, models, spectral resolution, laser wavelengths *etc*…) gathered from Custom user libraries collected in several EU member States.

We will show how this process can contribute to cross-checking and cross-validation of data contributing to the quality check providing thus more robust reference data for controls.

Besides the processing of spectra, a first key step is also the unambiguous naming or coding of substances allowing their clear and unambiguous identification by chemists and industry, trade operators and Customs officers. We will also discuss how this can be supported by the systematic use of InChi and InChiKey codes developed by IUPAC and NIST and designed for communication of chemical information through the use of computers and information technology [2].

The continuation of this work towards the establishment of a possible cross-instrument open-format CLEN spectral data library (not depending on manufacturers proprietary data format) will be discussed. Finally, the possibility of establishing models of classification of substances based on their spectral characteristics instead the classical method of finding a match in libraries will also be discussed [3].

References

[1] C. Guillou, F. Reniero, J. L. Vicente, M. Holland, K. Kolar, H. Chassaigne, S. Tirendi, H. Schepers, *Current Pharmaceutical Biotechnology* , **19** (2018) 91–98.

[2] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, *Journal of Cheminformatics*, **7** (2015) 23–57.

[3] J. Omar, B. Slowikowski, C. Guillou, F. Reniero, M. Holland, A. Boix, *Journal of Raman Spectroscopy* , **50** (2019) 41–51.

Comparison of classifiers for commercial beers and identifying patterns

D. Koren[1], *L. Lőrincz*[2], *S. Kovács*[3], *G. Kun-Farkas*[1], *B. Vecseriné Hegyes*[1], *L. Sipos*[4]
[1] Department of Brewing and Distilling, Faculty of Food Science, Szent István University, 45 Ménesi út, H-1118, Budapest, Hungary
[2] Egis Pharmaceuticals PLC, 116 Bökényföldi út, H-1165 Budapest, Hungary
[3] Department of Research Methodology and Statistics, Institute of Sectorial Economics and Methodology, Faculty of Economics and Business, University of Debrecen, 138 Böszörményi út, H-4032 Debrecen, Hungary
E-mail: kovacs.sandor@econ.unideb.hu
[4] Department of Postharvest Sciences and Sensory Evaluation, Faculty of Food Science, Szent István University, 29-43 Villányi út, H-1118 Budapest, Hungary
E-mail: sipos.laszlo@etk.szie.hu

In this study 13 properties (alcohol-, real extract-, flavonoid-, anthocyanin, glucose, fructose, maltose, sucrose content, EBC and L*a*b* color, bitterness) of 21 beers (alcohol-free pale lagers, alcohol-free beer-based mixed drinks, beer-based mixed drinks, international lagers, wheat beers, stouts, fruit beers) were determined. Multiple Factor Analysis was performed for the whole data and 5 clusters (target classes) were determined, then a bootstrapping was applied to establish a balanced data so as every cluster should contain 100 samples and the total sample size is 500. Next 13 supervised learning algorithms were applied to classify each brand into the target classes. Furthermore 5 error rates were recorded: resubstitution error rate (RER), Cross-validated error rate (CV), bootstrap error (BOOT), leave-one-out (LOO) and train-test error rate (TRAIN). The MFA could discriminate 5 groups which can be characterized by some analytical parameters. The other multivariate methods performed similarly. The methods can be discriminated best based on the bootstrap, CV and LOO. The best estimation methods are the C4.5, CSMC4 and CSCRT, these performed best along the flavonoid content and EBC color. It has been identified that the methods most sensitive to the properties are the MLR and NBC. The classification ability fluctuated greatly in case of three properties (glucose, maltose, sucrose). A remarkable fluctuation has been experienced in case of L*a*b* color parameters, flavonoid content, EBC color and bitterness by NBC and MLR methods.

Acknowledgement
This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The Project is supported by the European Union and co-financed by the European Social Fund (grant agreement no. EFOP-3.6.3-VEKOP-16-2017-00005 and EFOP-3.6.1-16-2016-00016). This research project was supported by the Doctoral School of Food Sciences (Szent István University).

3D-Chemocentric target deconvolution of unprecedented drug scaffolds

M. Kim,[1]* *H. Kim*,[2] *S. Lee*,[1] *S. Ahn*,[3] *S. Kumar*[1]
[1] Department of Pharmacy & Gachon Institute of Pharmaceutical Science, College of Pharmacy, Yeonsu-gu, Incheon, Republic of Korea;
Gachon University, Republic of Korea;
E-mail: kmh0515@gachon.ac.kr
[2] Department of Data Management, KEIS, Republic of Korea; [3]Department of Financial Engineering, College of Business, Ajou University, Republic of Korea;

In drug discovery, how much diverse structures of drugs are applicable for both selective and effective targeting? Even though it can be a very curious question in academia, current approaches combining cheminformatics with medicinal chemistry commonly use 'targets' as their queries so that 'structural uniqueness or diversity of drugs' cannot be considered with either the highest priority or the uncoupling with SAR . Therefore, rational drug design or virtual screening using a built molecular database cannot consider 'structural diversity of a drug' enough *e.g*. [1]. However, a part of 'not existing compounds' can be unfound treasure island [2]. In this presentation, I introduce (1) the definition of problems on my research, (2) divided redesigned sub-researches [3], (3) how to control 3D-similarity as an interactive variable. In particular, I intensively focus on a comparison method for characterizing the structures consisting of 3D similarity vector between every ligand (as a query) within a target.

References

[1] Jang, C. *et al*. Identification of novel acetylcholinesterase inhibitors designed by pharmacophore-based virtual screening, molecular docking and bioassay. *Sci. Rep.* **8** (2018) 14921*; Chosen Top100 in Chemistry at 2018.*
[2] Venkanna, A. *et al*. Pharmacological use of a novel scaffold, anomeric N,N-diarylamino tetrahydropyran: molecular similarity search, chemocentric target profiling, and experimental evidence, *Sci. Rep.* **7** (2017) 12535-12552.
[3] Kim, H.* *et al*. The comparison of automated clustering algorithms for resampling representative conformer ensembles with RMSD matrix, *J. Cheminformatics,* **9** (2017) 21-47.

Endogenous Metabolic Profiling as a Fundament in Personalized Theranostics

Torbjörn Lundstedt[1,2,7], *Katrin Lundstedt-Enkel*[1,3], *Kate Bennett*[1], *Claire Russell*[4], *Rebeca Martín-Jiménez*[4], *Michelangelo Campanella*[4], *Sara Mole*[5], *Julia Petschnigg*[5] and *Johan Trygg*[1,7]

[1] AcureOmics AB, Umeå, E-mail: torbjorn.lundstedt@acureomics.com
[2] Department of Pharmaceutical Chemistry, BMC, Uppsala University, Sweden
[3] Department of Environmental Toxicology, EBC, Uppsala University, Sweden.
[4] Department of Comparative Biomedical Sciences, Royal Veterinary College, UK
[5] MRC Laboratory for Molecular Biology, University College London, UK
[6] Umeå Plant Science Centre, Swedish University of Agricultural Science, Umeå, Sweden.
[7] Umeå Research group for Chemometrics, Institute of Chemistry, Umeå University, Sweden.

Metabolomics has grown into an established tool in research for; i) Diagnosis, *i.e*. classification; ii) Identification of biomarkers in relation to *e.g*. diseases; iii) Dynamic studies when identifying effects from, for instance, medical treatment, changes in life style, environmental or genetic changes.

In this presentation the use of metabolomics as a tool in drug discovery and diagnostics will be highlighted. In the first part the differences in biochemical profiles between healthy volunteers and persons with the diagnosis *rheumatoid arthritis* (RA) are discussed and identification of novel biochemical pathways for understanding the underlying factors of the disease are presented. In addition to this, two novel methods to restore a disturbed metabolic profile caused by a mutation in Cln3 will be discussed.

In the next part a comparison to different animal models is made, in order to identify the most relevant animal model for describing the disease in humans. The animal models are used for evaluation of novel treatments.

In the last part, an example from the BATCure project will be presented for a CLN3 disease yeast model, comparing the btn-1 mutant vs. wild type. In addition, from the zebrafish (*Danio rerio*) CLN2 disease model, we compared tpp1-/- with the metabolic profile of wild type. Results will be presented and discussed in relation to metabolic profiles and biochemical pathways and how these findings can help us to identify novel methods of treatments.

Evaluation and understanding of near infrared spectra in sports diagnostic examinations

J. Elek[1], *E. Markovics*[2], *Zs. Komka*[3] *M. Szász*[4]
E-Mail: elek@scienceport.hu
[1] Science Port Ltd, [2] University of Debrecen
[3] University of Physical Education, [4] Scitec Institute

The daily water intake is a very important generally, but has a pronounced importance in case of professional sportsmen. The hydration of the nervous system plays the more important role in the concentration efforts, thus has many times more impact on the sport performance under extreme conditions than the actual state of the muscles.

In a previous study near infrared spectra from different points of the body of male and female volunteers in different ages were recorded and evaluated. The volunteers were asked not to drink for a certain time, then they drunk water or isotonic drinks, and the NIR spectra were recorded and evaluated: visualization of the hidden patterns was carried out by principal component analysis, while the conclusions of these visual observations were checked by ANOVA. Isotonic drinks were also tested and the hydration effect was compared to the use of pure water.

As mentioned, for the sportsmen the fatigue does not only appear due to the exhaustion of the muscles, but the dehydration of the nervous system which is essential for a focused concentration not only the absorption, but the retention of the water is particularly important. We have examined the effect of using isotonic drinks instead of water on the ability of the body to retain the water under physical/thermal stress.

The extreme stress tests were carried out in the framework of a cooperation with the Hungarian Olympic Canoe Team and Scitec Institute where the traditional sport diagnostic parameters (such as cardiovascular and respiratory parameters, blood chemistry, etc.) and the NIR infrared spectra were recorded during the experimental phases.

A very typical figure of PC1 extracted from the NIR data vs. time is shown below. The maximal strain was put on the volunteers at 25 min, where their vascular system was practically collapsed. The collapse and the restitution can be excellently followed by analysing the NIR spectra.

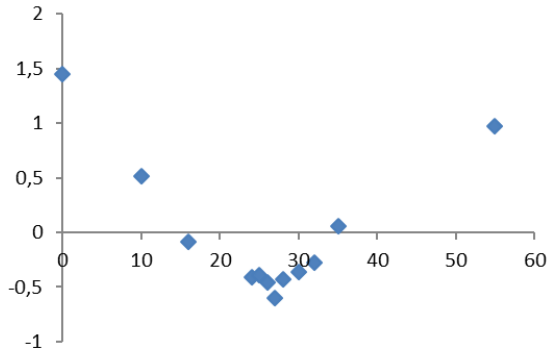


Figure 1
PC1 *vs*. sampling time

Our recent efforts target the mapping of the relationship between the principal components extracted and the medical parameters recorded during the study. The presentation gives an insight of the recent state of these researches.

Data Fusion Methods as Consensus Scores for Ensemble Docking

Dávid Bajusz[1], *Anita Rácz*[2], *Károly Héberger*[2]
[1] Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary, e-mail: bajusz.david@ttk.mta.hu
[2] Plasma Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

Molecular docking has proven to be an invaluable tool of structure-based drug discovery in the past three decades, powering many virtual screening campaigns [1]. Docking embodies a delicate balance between speed and precision: despite being based mostly on principles of molecular mechanics (and often involving a thorough evaluation of various intermolecular interaction terms), it is still capable of examining large ligand sets (up to the order of 10^6 ligands) in a reasonable time. Hence, ligand docking lies in-between the fields of molecular modelling and cheminformatics. Ensemble docking is a widely applied concept in structure-based virtual screening—to at least partly account for protein flexibility—usually granting a significant performance gain at a modest cost of speed: it involves docking the ligands into more, conformationally diverse structures of the target protein.

From a molecular modelling perspective, many important methodological questions regarding ensemble docking have been studied, such as ensemble size, selection algorithm and other factors [2]. However, if we look at ensemble docking from a cheminformatics point of view, it evidently presents a task of data fusion, i.e., how do we rank or select from the screened compounds when we have multiple docking score values for each of them? (More technically: how do we define the consensus docking score for a compound, given *n* docking scores on the individual protein structures?) While several data fusion methods exist – and have been thoroughly studied for ligand-based virtual screening [3] –, an extensive study to explore their application for ensemble docking has still been lacking.

Our recent work has addressed this question by a thorough statistical comparison of several data fusion options on multiple datasets, coming from ensemble docking-based virtual screenings on kinases, a G-protein coupled receptor, an oxidoreductase enzyme, and a nuclear receptor [4]. A combination of performance metrics is applied for the evaluation, including area under the receiver operating characteristic (ROC) curve (AUC), average precision (AP), Boltzmann-enhanced discrimination of ROC (BEDROC) and sum of ranking differences (SRD) values. Ultimately, we point the attention of the cheminformatics and drug discovery communities to the possible applications of data fusion rules (especially the geometric and harmonic means) in the context of ensemble docking.

References

[1] C. Sotriffer, *Virtual Screening: Principles, Challenges, and Practical Guidelines*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2011.
[2] O. Korb, T. S. G. Olsson, S. J. Bowden, R. J. Hall, M. L. Verdonk, J. W. Liebeschuetz, J. C. Cole, *J. Chem. Inf. Model*., **52** (2012) 1262–74.
[3] P. Willett, *J. Chem. Inf. Model*., **53** (2013) 1–10.
[4] D. Bajusz, A. Rácz, K. Héberger, *Molecules*, **24** (2019) 2690.

Comparison of performance parameters for machine learning classifiers

Anita Rácz[1], *Dávid Bajusz*[2], *Károly Héberger*[1]
[1] Plasma Chemistry Research Group, [2] Medicinal Chemistry Research Group,
Research Centre for Natural Sciences Hungarian Academy of Sciences,
Magyar tudósok krt. 2, 1117 Budapest, Hungary
E-mail: heberger.karoly@ttk.mta.hu

The goodness of the (binary or systematic) classification problems could be determined with several performance parameters, which can frequently give deviating results for the user. The same problem occurs with the use of machine learning (and deep learning) algorithms. The optimization of these methods is very important, because the effectiveness of the methods strongly depends on the optimization protocol.

In this study our aim was to build up a well-determined protocol for each type of the used machine learning algorithms. Moreover, we performed a multi-level comparison with the use of the different performance parameters and machine learning methods. The robust but sensitive sum of ranking differences [1] and other chemometric tools (COVAT heatmaps [2]) were applied for the comparison. The workflow was carried out for more ligand sets and protein targets (in the case of activity classification) with *in-vivo* toxicity/activity data.

Finally, the most consistent performance parameters and the best classification methods were selected using appropriate optimization protocols. Monte-Carlo cross-validation and other cross-validation options [3] were implemented in the comparison process to validate our results.

Acknowledgement
This work was supported by the National Research, Development and Innovation Office of Hungary (NKFIH) under grant numbers OTKA K 119269 and KH_17 125608.

References

[1] K. Héberger Sum of ranking differences compares methods or models fairly. *TrAC Trends Anal. Chem*. **29** (2010) 101-109. https://doi.org/10.1016/j.trac.2009.09.009
[2] F. Andrić, D. Bajusz, A. Rácz, S. Šegan, K. Héberger, Multivariate assessment of lipophilicity scales–computational and reversed phase thin-layer chromatographic indices, *J. Pharm. Biomed. Anal*., **127** (2016) 81-93. http://dx.doi.org/10.1016/j.jpba.2016.04.001
[3] K. Héberger and K. Kollár-Hunek, Comparison of validation variants by sum of ranking differences and ANOVA, *J. Chemometr*. **33** (2019) pp. 1-14 Article number: e3104 https://doi.org/10.1002/cem.3104

Posters

Fingerprint similarity metrics in cheminformatics, metabolomics and other fields

Dávid Bajusz[1], *Anita Rácz*[2], *Károly Héberger*[2]
[1] Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary, E-mail: bajusz.david@ttk.mta.hu
[2] Plasma Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

Fingerprints (binary data structures) are ubiquitous in many, often unrelated scientific areas as a compact and (ideally) unique representation of e.g. molecular structure or sample composition. In molecular fingerprints (the former), bit positions are associated to the presence or absence of certain substructures [1], while in metabolomics fingerprints (the latter), bit positions are associated to the presence or absence of certain components/metabolites. The similarity of such binary fingerprints is calculated for various purposes: hierarchical clustering of samples, virtual screening for drug candidates, *etc*. A large number of similarity metrics were proposed by researchers of diverse fields over the past century: they were collected in a recent work [2], however, most have remained virtually unknown for the broader scientific community.

We have recently conducted several large-scale comparative studies to explore an exhaustive set of binary similarity metrics for various applications, using the robust statistical comparison method, sum of ranking differences (SRD) [3]. We have found that from a small pool of the most commonly known similarity metrics, the well-known and generally favored Tanimoto coefficient is indeed a valid preference for molecular fingerprints [4]. In metabolomics, we have demonstrated that with a suitable similarity metric, binary metabolomic fingerprints can be used for the clustering of samples with minimal misclassification (as compared to the use of quantitative chemical composition data) [5]. Our most recent work in the overlapping fields of molecular modelling and cheminformatics has highlighted six similarity measures as better or comparable alternatives for the popular Tanimoto similarity coefficient for protein-ligand interaction fingerprints [6]. A compact, open-source Python package, implementing the 51 similarity metrics collected by Todeschini *et al*. [2] for binary data formats was published at:

https://github.com/davidbajusz/fpkit.

References

[1] D. Bajusz, A. Rácz, K. Héberger, in *Comprehensive. Medicinal Chemistry*. III (Eds.: S. Chackalamannil, D.P. Rotella, S.E. Ward), Elsevier, Oxford, 2017, pp. 329–378.
[2] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, *J. Chem. Inf. Model.*, **52** (2012) 2884–2901.
[3] K. Héberger, *TrAC Trends Anal. Chem.*, **29** (2010) 101–109.
[4] D. Bajusz, A. Rácz, K. Héberger, *J. Cheminform.*, **7** (2015) 20.
[5] A. Rácz, F. Andrić, D. Bajusz, K. Héberger, *Metabolomics*, **14** (2018) 29.
[6] A. Rácz, D. Bajusz, K. Héberger, *J. Cheminform.*, **10** (2018) 48.

Sensory evaluation of cricket-enriched oat biscuits using check-all-that-apply analysis

B. Biró[1*], *A. M. Sipos*[1], *A. Kovács*[2], *K. Badak-Kerti*[2], *K. Pásztor-Huszár*[3], *A. Gere*[1]
[1] Szent István University, Faculty of Food Sciences, Department of Postharvest Science and Sensory Evaluation, Villányi út 29-43. Budapest, H-1118, Hungary
*Corresponding author e-mail: barbarabirophd@gmail.com
[2] Szent István University, Faculty of Food Sciences, Department of Grain and Industrial Plant Processing, Villányi út 29-43. Budapest, H-1118, Hungary
[3] Szent István University, Faculty of Food Sciences, Department of Refrigeration and Livestocks' Products Technology, Ménesi út 43-45. Budapest, H-1118, Hungary

Entomophagy, or "eating insects" was typical in the prehistoric era and is still an integral part of the gastronomy of more than a hundred countries. The topic gained scientific interest in the past few years and since then numerous papers have been published. Today, more than 2000 edible species are known: bugs, ants, wasps, bees among others, in nearly every developmental stage [1].

According to several researches, edible insects are favourable from nutritional perspective: they can also be used as food and feed [1]. However, there is a lack of appropriate scientific literature on product development and the sensory evaluation of insect-based food products. Only a few publications deal with such developments, (*e.g.*: bread, meat patty, protein bar and cereal snacks), and these publications do not include the large-scale consumer sensory acceptance tests [2].

In our study, four oat based biscuit samples were made using oat flour, buckwheat flour and *Acheta domesticus* (house cricket) powder. The control sample contained 0 % cricket powder, while the enriched ones contained 5 %, 10 % and 15%, respectively. A total of 67 consumers evaluated the samples in a one-week session. Participants rated multiple liking attributes and filled out a check-all-that-apply (CATA) questionnaire for several properties related to appearance, odour, texture and flavour.

CATA experiment is a simple task that does not necessitate a specific training of the accessors. The panellists are given a set of products; each sample is presented with a list of attributes [3]. The task is to check all those attributes that considered to be appropriate for the given product. Even though CATA questions yield only binary outcomes (1 – checked data, 0 - unchecked data), the method has been found to demonstrate high discriminative capability among the product samples [4]. There are several sensometric methods available to analyse CATA data, there is no agreement on their use. The aims was i) to compare different CATA data evaluation methods in order to define the most discriminating one and ii) to describe the sensory attributes of the tested cricket enriched oat biscuit samples.

References

[1] L. Kouřimská, and A. Adámková, *NFS Journal,* **4** (2016) 22-26.
[2] A. Gere, D. Radványi, and K. Héberger, *Innov. Food Sci. Emerg.,* **52** (2019) 358-367.
[3] G. Ares, and S. R. Jaeger, *Rapid sensory profiling techniques,* (2015) 227-245.
[4] D. Valentin, S. Chollet, M. Lelièvre, and H. Abdi, *Int. J. Food Sci. Tech.,* **47** (2012) 1563-1578.

Multi-level comparison of Hungarian wines using advanced chemometric methods

Zs. Guld[1], *D. Nyitrainé Sárdy*[1], *A. Gere*[2,*], *A. Rácz*[3]
[1] Szent István University, Department of Oenology, H-1118 Budapest, Ménesi út 45.
[2] Szent István University, Faculty of Food Science, Sensory Laboratory, Villányi út 29-43, H-1118 Budapest, Hungary
E-mail: gereattilaphd@gmail.com
[3] 3Plasma Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1117 Magyar tudósok krt. 2 Budapest, Hungary

The aim of this study was to compare different Hungarian Kadarka, Kékfrankos and Cabernet Franc wines produced and aged by the same methods and to compare the two types of sensory analysis methods as well: the 100-point system of the International Organisation of Vine and Wine (OIV), and quantitative descriptive analysis (QDA). Both tests were conducted by 12 assessors of the University of Pécs, Institute for Regional Development, Faculty of Horticulture and Oenology. This study provides conclusions about the use of sensory analysis methods, highlighting the advantages and disadvantages of QDA and the OIV system. Principal component analysis (PCA), analysis of variance (ANOVA) and multiple factor analysis (MFA) were used for the evaluation of the data.

Our results showed that the sensory panel was able to discriminate the samples by both sensory methods, however the information provided by them was significantly different. ANOVA clearly showed that the two methods have different sensitivity when comparing wines (produced and reference wine samples) and QDA proved to be the more sensitive, as well as more detailed, method.

In general, OIV is able to show the general quality of the wines, while QDA coupled with proper chemometric methods is able to describe why the given samples received good or bad OIV scores.

How can surface roughness be estimated at best?

Loránd Románszki[1] *Szilvia Klébert*[2], and *Károly Héberger*[2,*]
[1] Functional Interfaces Research Group; [2] Plasma Chemistry Research Group
Institute of Materials and Environmental Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2, 1117 Budapest, Hungary
E-mail: heberger.karoly@ttk.mta.hu

Plasma treatments of textiles have been extensively studied in the last decade since textile industry is continuously trying to fulfill the ascendant society demand in terms of enhanced hydrophobic or hydrophilic properties, better printability, intelligent filtration properties, flame retardation, biocompatibility, self-cleaning finishes, *etc*.

Untreated and treated PET textile samples were investigated by an EVO 40 scanning electron microscope (SEM, Carl Zeiss AG, Oberkochen, Germany) at 20 kV acceleration voltages.

The surface roughness of individual PET fibers before and after plasma treatment was measured by a Dimension 3100 atomic force microscope (AFM) equipped with a Nanoscope IIIa controller (Digital Instruments/Veeco, Santa Barbara, California, USA) on 2×2, 5×5, 6.7×6.7 and 8×8 μm^2 areas using silicon cantilevers in contact mode with 512×512 points resolution.

Nine surface roughness measure have been defined.

This research was focused to answer the following questions: Which is the best roughness measure from the nine usually defined one? How they are grouping? Which ones are equivalent with others? Are there significant differences from among them? How large is the optimal window? Does the roughness increases during plasma treatment?

The roughness measures were evaluated by using sum of ranking differences (SRD) [1,2] and generalized pair correlation method (GPCM) [3]. The two techniques allowed the selection of the best roughness measures: R_{max}, -the vertical distance of the highest point from the lowest point, R_z, the mean vertical distance of the five highest points and five lowest points from the mean plane and R_{pm}, the mean vertical distance of the five highest points from the mean plane.

It is not worth to calculate all nine measures; the larger the determination window - the better. Plasma treatment increases the roughness (independently from the roughness measure used).

Acknowledgement

The authors thank the support of the National Research, Development, and Innovation Office of Hungary (OTKA, contract No K 119269). This work was partially funded also by the National Competitiveness and Excellence Program, Hungary (NVKP_16-1-2016-0007).

References

[1] K. Héberger, Sum of ranking differences compares methods or models fairly. *TRAC - Trends Anal. Chem.* **29** (2010) 101-109.

[2] K. Kollár-Hunek and K. Héberger, Method and Model Comparison by Sum of Ranking differences in Cases of Repeated Observations (Ties). *Chemometr. Int. Lab. Syst.*, **127** (2013) 139-146.

[3] Károly Héberger and Róbert Rajkó, Generalization of Pair-Correlation Method (PCM) for Nonparametric Variable Selection. *J. Chemometr.* **16** (2002) 436-443.

Net analyte signal-based supervised preprocessing: application in pattern recognition

Sara Mostafapour *Bahram Hemmateenejad*
Department of Chemistry, Shiraz University, Shiraz, Iran
E-mail: hemmatb@shirazu.ac.ir

Pattern recognition (PR) is a kind of exploratory data analysis method that automatically recognizes patterns and regularities in data. Generally, PR can be used for either clustering (unsupervised learning) or classification (supervised learning) purposes. Multivariate data analysis methods often need to be preprocessed since there are systematic variations in the response data (**X**-variables) that are unrelated to the class labels (y-variable). The unrelated variations in **X** may disturb the multivariate modelling, cause imprecise prediction for new samples [1]. For removal of unrelated variations, differentiation and signal correction methods have been used. Almost all of these methods are unsupervised, i.e., they correct **X** variables without using information in y-variable.

Projection has been used as a supervised preprocessing method. In its simplest form, the projection of X to Y can be considered. As stated by Wold et al., this calculation may work for training set but since no y-values are available for new samples, the x-vectors of these new samples cannot be orthogonalized in the same manner as the training set. Orthogonal signal correction (OSC), to remove the variations in **X** that are unrelated to y, was then suggested by Wold et al. [1]. However, OSC is associated with some drawbacks like having an internal time consuming iteration to find orthogonal components and does not giving a unique solution.

Here, we used the net analyte signal (NAS) [2] concept to suggest a supervised preprocessing method for using in pattern recognition and classification and it is called Net Analyte Preprocessing (NAP). This method, which is based on the simple projection method, instead of projecting **X** on **Y**, works by projecting the **X**-variable of one class on the **X**-variable of another class and vice versa. In this manner, the **x**-vector of each sample is projected twice (on the space of the samples of the same class and on the space of the samples of another class). This method can be easily applied for new compounds by projecting the x-vector of the new samples on the matrix spaces of the training data matrices. The preprocessed data can be simply used as input of both supervised and unsupervised pattern recognition methods.

This method was used for classification of different proteomics datasets [3], e.g., to identify pancreatic cancer. Improved classification results was obtained using NAP, e.g., accuracy was increased from 0.6 (without NAP) to 0.81 (using NAP).

References

[1] J. Trygg, S. Wold, *J. Chemometrics*,**16**, (2002) 119-128.
[2] A. Lorber, *Anal. Chem.*, **58**, (1986) 1167-1172
[3] https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

Consensus QSAR modeling for the toxicity of organic chemicals against *pseudokirchneriella subcapitata* using 2D descriptors

K. Khan[1], *K. Roy*[1,*]
[1]*Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, 188 Raja S C Mullick Road, 700032, Kolkata, India*
E-mail: kunalroy_in@yahoo.com

The two most commonly used measurements in ecological risk assessment include no-observed-effective concentration (NOEC) and x% effective concentration (ECx) where x can be 5-100. The existing methods of determination of NOEC and ECx via laboratory testing involve considerable cost and time; additionally these experiments cannot be performed for all possible endpoints. However, computational tools like quantitative structure-activity/toxicity relationship (QSAR/ QSTR) modeling can help filling the data gap. The QSAR technique for the risk assessment of chemical compounds is recommended by various regulatory agencies like European Centre for the Validation of Alternative Methods (ECVAM), European Union Commission's Scientific Committee on Toxicity, Ecotoxicity, and Environment (CSTEE) and United States Environmental Protection Agency (US EPA). The current report proposes robust, externally validated consensus quantitative structure-activity relationship (QSAR) models developed from 334 organic chemicals for the prediction of effective concentrations of chemicals for 50% and 10% inhibition of algal growth. 2D descriptors having definite physicochemical meaning were calculated from Dragon and PaDEL-descriptor software tools. Model development, validation and interpretation were performed following the strict guidelines of Organization for Economic Co-operation and Development (OECD).

For feature selection, genetic algorithm along with stepwise selection was used, while the final models were developed from partial least squares regression technique in order to obviate any chance of intercorrelation among descriptors. The variables like MLOGP, MR and LogKow (experimental lipophilicity) exert highest positive contributions in controlling the aquatic toxicity, whereas polar groups such as oxygens in the form of SO_2OH (*n*SO2OH descriptor) and alpha hydrogen (H-051 descriptor) showed an inverse correlation with the algal toxicity. The applicability domain analysis was carried out using the DModX technique available in SIMCA-P software in order to set the predefined chemical space to obtain reliable predictions for unknown organic chemicals. The obtained models against pEC_{50} endpoints were then used to predict toxicity of 64 organic chemicals not having definite observed response. Finally, the prediction reliability indicator tool was used to assess the confidence with which unknown compounds were predicted. The obtained results also emphasize on the use of consensus modeling and its application in reducing prediction errors. The obtained QSAR models can act as helpful tools for identification and prioritization for chemicals of highest concern, production of safer alternatives within the scope of REACH regulations for hazardous chemicals.

How to measure distribution for a pair of target classes?

S. Lee,[1] *S. Ahn*,[2] *M. Kim*,[1]*
[1] Department of Pharmacy & Gachon Institute of Pharmaceutical Science, College of Pharmacy, Yeonsu-gu, Incheon, Republic of Korea;
Gachon University, Republic of Korea;
E-mail: kmh0515@gachon.ac.kr
[2] Department of Financial Engineering, College of Business,
Ajou University, Republic of Korea;

When we compare a pair of compounds, the similarity of the pair depends on alignment method, chosen descriptors, and metric. In order to predict promising targets of unprecedented drug scaffold, the comparison requires between groups (classes). In this presentation, we investigate an informatical comparison method for characterizing the structures consisting of 3D similarity scores between every ligand (as a query) within a target. For the representative similarity distribution of the structures, Gaussian Mixture Model (GMM) applies to the probabilistic distributions of Tanimoto coefficient (as a similarity metric) from the structures corresponding to the classes. Hyperparameters of GMM components can be acquired using expectation-maximization (EM) algorithm for each distribution. We introduce Kullback-Leibler (K-L) divergence to measure the discrimination ability. K-L divergence between a representative target class distribution and the distributions from each structure corresponding to each query could quantify the probability that the query belongs to the specific pharmacological class.

References

[1] Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **22** (1951) 79-86.
[2] Nalewajski, R.F.; Parr, R.G. Information Theory, Atoms in Molecules, and Molecular Similarity. *Proc. Natl. Acad. Sci. U. S. A.* **97** (2000) 8879-8882.
[3] Venkanna, A. *et al*. Pharmacological use of a novel scaffold, anomeric N,N-diarylamino tetrahydropyran: molecular similarity search, chemocentric target profiling, and experimental evidence, *Sci. Rep.* **7** (2017) 12535-12552.

Remarks on some validation parameters

P. Király[1], *D. Kovács and G. Tóth*
Institute of Chemistry, Loránd Eötvös University
Pázmány s. 1/a, 1117 Budapest, Hungary
[1] e-mail: peter0kiraly@gmail.com

Despite the frequent use of validation parameters there are several basic features of the most common ones that are not clear for the average user. In our parallel study on the sample size dependence of these metrics we faced several misconceptions. A large part of these uncertainties is caused by neglecting conditions when the parameters exhibit a special behavior, and the lack of comprehensive comparison on mathematical statistical basis. Also, the existing QSAR validation guidelines [1] have to be taken into account, according to which the internal parameters can be divided into two groups in terms of assessing the goodness-of-fit or robustness, while the external ones are applied to check predictivity. The results shown here are based on the datasets detailed in our study presented on the sample size dependence of validation parameters. The same R code was used and extended.

We calculated the correlation and the rank correlation between different performance parameters. These data were plotted against the sample size. We found that the internal-internal, external-external and internal-external pairs behave differently. The largest correlations are between the pairs of those internal parameters which are used for estimating the goodness-of-fit. If one internal parameter for goodness-of-fit and one internal parameter for robustness is tested, their rank correlation is weaker at small sample sizes, but it becomes similar to that of the goodness-of-fit - goodness-of-fit pairs above a medium sample size. It means, up to this sample size the two qualities of indicated by internal parameters, the goodness-of-fit and the robustness are distinguishable, but above a certain sample size the parameters are redundant. In the case of the external-external pairs the correlations are still significant. The external-internal pairs are weakly correlated. To summarize: our data show that all the three aspects of validation can be accessed at small sample sizes, but the internal check of robustness is not informative above a given sample size.

We plotted the $\bar{r}^2_m$ metric as a function of Δr^2_m, both of which have been developed and suggested by Roy, Ojha et al. [2]. Furthermore, we have compared the results and the respective values of R^2_{adj}, Q^2_{loo} and Q^2_{F2}. We believe that if the size and quality of the sample allows the creation of several test sets, this kind of graphs might be a useful visual asset throughout the course of model selection. The division of the graph provides an easy schematic classification of the models.

We have minor comments, too: i) We collected some misunderstandings of R^2 and R^2_{adj} used in unconstrained regression and in constrained cases. ii) In the case of CCC our data and the literature search showed that many of the proposed benefits of the metric are similar to the ones already characteristic of R^2. Usually the benefits of CCC were stated without comparing its performance to R^2. iii) We explained the difference between Q^2_{F1} and Q^2_{F2} with respect to the test/training ratio. Consequently, in addition to the purpose of application training-test division may determine which of them is to be used. We propose some modification in the calculation of Q^2_{F3} [3] to take into account the correct number of degrees of freedom.

References

[1] http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf
[2] P. K. Ojha, I. Mitra, R. N. Das and K. Roy, *Chemometrics and Intelligent Laboratory Systems*, **107** (2011) 194-205.
[3] V. Consonni, D. Ballabio and R. Todeschini, *Journal of Chemical Information and Modeling* **49** (2009) 1669-1678.

Sample size dependence of validation parameters

D. Kovács[1], *P. Király* and *G. Tóth*
Institute of Chemistry, Loránd Eötvös University
Pázmány s. 1/a 1117 Budapest, Hungary
[1] e-mail: kovacsdaniel995@gmail.com

The feasibility of mathematical models based on experimental results is usually checked by validation both in industry and in science. In our research we investigated a so far barely discussed area, namely the dependence of statistical validation parameters on the size of the sample taken.

A code in R programming language has been written and applied in the study to handle the dataset chosen, executes fitting of multivariate linear curves and calculates all validation metrics at hand. The metrics were chosen from ref. [1]. The first dataset comprised observations of four dependent variables and a single independent variable related to the electric performance of a combined cycle power plant [2]. The second dataset contained a single dependent and 8 independent variables with the former describing a mechanical property of concrete and the latter providing compositional data and age [3]. The first dataset is one that can be efficiently modelled by multivariate linear regression (termed as good), while the second one can only weakly be modelled by such a function (weak).

Through the course of our simulations, samples of different sizes were taken from the above-mentioned large datasets. Sample sizes were chosen as follows: 30, 40, 50, 75, 100, 200, 500 and 1000. Following linear modelling internal and external validation parameters were calculated. The internal ones can be divided into two groups according to assessing the goodness-of-fit or robustness, while the external ones are usually used to check predictivity.

We have observed that the unadjusted R^2, RMSE, *s*, MAE and CCC parameters are misleading as they overestimate the goodness-of-fit of models when the sample size is small. The scaling of the R^2 to R^2_{adj} corrects the trend only for the good dataset. The Q^2_{loo} and CCC_{loo} parameters shows the correct trend that models on large sample sizes are better, and the SEP_{rdCV} parameter behaved correctly as well.

We calculated several leave many out Q^2s. All of them supports the models on large data sizes. Surprisingly, it is possible to scale them close to identical by correcting the degrees of freedom of the models. It means, the simple-to-calculate Q^2_{loo} gives the same results as the complicated leave many out versions.

We calculated the *y*-scrambled, *y*-randomized and *x*-randomized R^2 and Q^2_{loo} parameters. The three different methods provided close to identical results in average, therefore we suggest using the simplest *y*-scrambling method.

The sample size dependence of the external Q^2_{F1}, Q^2_{F2} and Q^2_{F3} parameters, the Roy-Ojha $\bar{r}^2_m$ metric and CCC_{test} were calculated as well with a test:training ratio of 1:4. These parameters provided the correct trends with respect to the sample size, but their sensitivity to sample size differed. Q^2_F measures exhibit remarkably different behavior even among themselves in the case of minor sample sizes. The large sample limit of the Q^2_{F2} parameter was significantly different from those of the other two similar metrics in the weak regression case.

References

[1] A. Rácz, D. Bajusz and K. Héberger, *SAR and QSAR in Environmental Research*, **26**, (2015) 683-700.
[2] P. Tüfecki, *Electrical and Energy Systems*, **60**, (2014) 126-140.
[3] I-C. Yeh, *Cement and Concrete Research*, **28**, (1998) 1797-1808.

Application of regression control chart in pharmaceutical on-going stability study to detect out-of-trend results

M. Mihalovits[1], *S. Kemény*[2]
[1] Department of Chemical and Environmental Process Engineering,
Budapest University of Technology and Economics,
E-mail: mihalovits@mail.bme.hu
[2] Department of Chemical and Environmental Process Engineering,
Budapest University of Technology and Economics

In pharmaceutical stability studies the important chemical, biological and physical attributes of the drug are being monitored over time. Regression curve is fitted to the data obtained in the study which is then used to estimate the shelf-life for the investigated batch. Two type of stability studies can be distinguished: pre-approval and on-going stability studies. Pre-approval stability studies are applied to obtain shelf-life for the registration of the drug product. The purpose of on-going stability studies is two-fold: they are used in on-going production to test whether the regulatory requirement regarding the shelf-life is met and to test whether the production is in-control that is the quality of the drugs are predictable.

ICH Q1E guide is intended to give support in the statistical methods to be used in evaluation of stability studies data. The guide is focused around preapproval stability studies while on-going stability studies and their quality assurance aspect are not considered there. Out-of-trend (OOT) data detection is one of the most important topic that is not discussed in the guide. One of the type of OOT nature is when the data is outlier in the space of the independent variable within a stability study. That is, the expected value of the OOT data within the investigated batch differs from the true value of the measured attribute at a given time point. The source of these points are usually an error in the analytical process. OOT data points distort the estimation of shelf-life may resulting in a false decision regarding the quality of the drug being produced. In this work regression control chart method is adapted for the use in on-going stability studies to detect OOT data points within the batch under investigation. The speciality of stability studies is that usually 8-10 data points are obtained throughout the study and each point is potentially OOT, which means that the control chart should be used as early in the study as possible.

The downside of the application of short Phase I is that the uncertainty of the parameters of the chart (centre line and acceptance limits) is great and the effectiveness of the chart is questionable. To overcome this problem, different considerations are suggested in this work. The effectiveness of the control chart method in stability studies was investigated and characterized with statistical power.

Discrimination of aqueous solutions by PCA-DA assessment based on FT-IR spectrometry

Zs. I. Németh[1], *I. Mészáros*[2], *Cs. Milic Raffai*[2], *A. Vágvölgyi*[3], *R. Rákosa*[1]
[1] Institute for Chemistry, University of Sopron, Sopron, Hungary, H-9400 P O Box 132;
E-mail: nemeth.zsolt@uni-sopron.hu
[2] Sopron Waterworks, Sopron, Hungary, H-9400 P O Box 41
[3] Institute of Forest- and Environmental Techniques, University of Sopron, Sopron, Hungary, H-9400 P O Box 132

In the 2000s years, appearing the sampling technique ATR (attenuated total reflectance) revolutionized the measurability of liquid samples in FT-IR spectrometry. The IR spectra of aqueous solutions have become recordable with its application. The FT-ATR-IR spectrum of aqueous solutions is dominated by own absorbance of the water that overshadows the absorbance contribution of the solutes in low concentrations. Multivariate data analysis without some data pre-processing is not able to bypass the determinant role of the water in the formation of the resultant absorbance of aqueous solutions. To comparison or discrimination of the aqueous solutions on the basis of FT-IR spectrometry, the IR absorbance of the solutes determining the quality difference is to be uncovered from the resultant spectrum.

To achieve this goal, theoretically supported spectrum pre-treatment strategy is to be conceived and applied. Utilized the spectrum of distilled water as background, such difference spectra of aqueous samples can be obtained in the wavenumber range from 1000 to 1500 cm^{-1} that are robustic against the spectral shift of the water spectrum since this spectral range has a quasi-horizontal line of water. The stochastic alterations of the offside and drift of the water plateau hidden in the primary spectra can be eliminated by linear detrending of difference spectra. To lowering the noise, some smoothing methods (moving window, Savitzky-Golay, etc.) can be applied. Ultimately, the concentration-dependent absorbance deviations of the solutes can be compensated by the application of some normalisation method (e.g. SNV). After PCA decomposition of the spectra transformed by spectrum pre-treatment strategy proposed by us, the quality of classification into the own group is characterized by the Mahalanobis distances from DA method. In fact, the discrimination of aqueous solutions with our assessment strategy is just guided back to the absorbance characters of the various solutes.

The applicability of outlined assessment strategy is exemplified for four aqueous solutions (tap water, urine, input sewage and output waste water samples from a wastewater treatment plant). As a reference, the results of PCA-DA without spectrum pre-treatment are also introduced.

Acknowledgement
The research was funded by the EFOP-3.6.2-16-2017-00010 “RING 2017” project.

Applicability of FT-ATR-IR spectrometry in identification of mycelium cultures

R. Rákosa[1], *M. Vargovics*[1], *J. Jakab*[2], *Zs. I. Németh*[1]
[1]Institute for Chemistry, Faculty of Forestry, University of Sopron,
H-9400, 4 Bajcsy-Zs, Sopron, Hungary;
E-mail: nemeth.zsolt@uni-sopron.hu
[2]Institute of Silviculture and Forest Protection, Faculty of Forestry,
University of Sopron, H-9400, 4 Bajcsy-Zs

Phytopathogenic fungi genera cause serious economic and ecological damage all over the world. To identify and characterize these pathogens and to develop an adequate protection strategy requires rapid diagnostic procedures. The classic microbiological, immunological, and molecular methods are expensive, time consuming and suitable only for a limited number of fungi. Fourier transform infrared spectroscopy (FT-IR) has been applied widely for the examination of biological samples such fungi species and their strains as, for example, those of *Phytophthora*. They are responsible for severe diseases in various forests in Hungary.

Phytophthora cultures (*P. cambivora, P. taxon raspberry, P. cactorum, P. plurivora, P. lacustris, P. gonapodyides*) grown on standardized potato dextrose agar (PDA) were investigated. The method based on reflection technique (ATR) was developed for the spectral discrimination of mycelium cultures. To reduce the stochastic effects on the spectra, which can increase the extents of the groups, different data pre-processing methods were applied. After having executed pre-treatments, the spectra were assessed by principal component analysis (PCA) and linear discriminant analysis (LDA). Performing the PCA decomposition on the spectra enabled to separate the various fungal strains from each other. Although groups of mycelium cultures can be clearly identified in the score plot of the first two PCs, an overlapping appears between the groups of *P. plurivora* and *P. lacustris*. Despite of experienced overlapping, the first two PCs discrimination accuracy of 94% was obtained. Because of increasing the classification quality, LDA was extended to the multidimensional space defined by the deterministic principal components.

Ackowlegement
The research was funded by the EFOP-3.6.2-16-2017-00010 “RING 2017” project

Mid-infrared imaging based lung cancer subtype determination

A. Pesti[1], *E. Kontsek*[1], *G. Smuk*[2], *S. Gergely*[3], *A. Kiss*[1]
[1] 2nd Department of Pathology, Semmelweis University Budapest;
pesti.adrian@med.semmelweis-univ.hu
[2] Department of Pathology, University of Pécs
[3] Department of Applied Biotechnology and Food Science,
Budapest University of Technology and Economics

Introduction: Lung cancer is the most common tumorous disease. The two main types are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Squamous cell carcinoma (SCC) and adenocarcinoma (AC) are the two most frequent subtype in NSCLC. Life science associated mid- and near-infrared based microscopic techniques are developing exponentially, especially in the past decade. This is a potential non-destructive approach to investigate malignancies. Our long term project is to develop a spectroscopy based diagnostic tool, revolutionizing medical diagnostics, especially cancer identification. Vibrational spectroscopy could be a promising approach to reach this.

Aims: Our goal was to differentiate lung cancer subtypes by their label-free mid-infrared spectra using supervised multivariate data analyses.

Material and methods: The following settings based on our preliminary research from the last years. Formalin-fixed paraffin-embedded (FFPE) samples were selected from the archive of the 2nd Department of Pathology Semmelweis University and the archive of Department of Pathology University of Pécs. Three subtypes were selected each group 10-10 cases SCLC, SCC and AC. 2 μm thick sections were cut and laid on aluminium coated glass slides. Transflection optical setup was applied with the Perkin-Elmer Spotlight 400 infrared microscope. 250 μm × 600 μm areas were imaged and the so-called mid-infrared fingerprint region (1800-648 cm^{-1}) was investigated with cluster analysis (CA), principal component analysis (PCA), linear discriminant analysis (LDA) and support vector machine (SVM) methods. On the other hand we tried convolutional neural networks (machine learning).

Results: In multivariate data analysing methods the CA and PCA helped us to identify outliers, and bias components. The LDA model resulted a 60-80% accuracy (depending on the settings). The SVM performed better than the LDA (70-85%). The confusion matrices showed that the SCC and AC are overlapping each other, the SCLC spectra are better separated. The machine learning method gave worse results (50-67 %).

Conclusions: Mid-Infrared imaging might be used to differentiate FFPE lung cancer subtypes, but the spectra pretreatment has a major effect on the results. The supervised methods gave controversial results on the non-filtered spectra. Overall SCLC and SSC worked much better than AC except the machine learning tool. Some hidden errors in the training set must be revealed to gain better accuracy. Altogether, our results support our long term project, the feasibility of infrared imaging to identify different tissues under *in vivo* conditions. Therefore, the technique might be applied to judge the resection margins of malignancies by multivariate data analysis methods.

Acknowledgement: This work was supported by ÚNKP 19-3 from the National Scientific Research Foundation.

Process capability indices when the usual assumptions fail, a tolerance interval approach

Éva Pusztai[1], *Sándor Kemény*[1]
[1] Budapest University of Technology and Economics, Department of Chemical and Environmental Process Engineering,
E-mail: pusztaie@mail.bme.hu

Statistical indices- like process capability (C_p) or process performance (P_p) index- are widely used in the field of quality management. These clever indicators make illustrative the relationship between the width of the specification interval and the width of the process variability. The latter is characterized by the tolerance interval, which contains major part of the population with high confidence. In the original concept this tolerance interval is calculated using simple models. The ultimate objective of calculating these indices is to give information about the proportion of non-conforming parts in the process (in the population).

However, the use of C_p, P_p is based on certain statistical assumptions. If at least one of them does not fulfill, the calculated value of C_p (P_p) is not able to give information about the process. Our work is dealing with the case that the quality characteristic of interest is a normally distributed random variable, the process is in control, but three sources of variability are present. The model is a two-way nested ANOVA

In order to establish the proportion of non-conforming parts the calculation of the ratio of the population beyond the specification limits is needed. According to this, the quantile of the distribution shall be determined that are equal to the specification limits. Thus, the task is to calculate tolerance interval for the $\mathcal{N}(\mu, \sigma_A^2 + \sigma_{B(A)}^2 + \sigma_e^2)$ distribution. In practical cases the variance components are unknown and to be estimated. To estimate the ratio of non-conforming parts two approximate calculating methods which are coherent with the definition of C_p are investigated, as well.

The aim of this work is to compare the results of the two approximate methods with the tolerance interval based (correct) calculation method. In addition, the effect of the ratio of the variance components to the differences is discussed.

Unsupervised data reduction: How to set the intercorrelation limits optimally?

Anita Rácz[1], *Dávid Bajusz*[2], *Károly Héberger*[1]
[1] Plasma Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
E-mail: Anita.racz@ttk.mta.hu
[2] Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

Introduction: Descriptor (pre-)selection is an integral part of QSAR workflows: It usually involves the removal of descriptors with missing values, constant values across the whole dataset, or collinear (inter-correlated) descriptors.

Methodology: Sum of ranking differences (SRD) is a basic and simple tool for ranking methods or models in every field of science [1]. SRD realizes a multicriteria decision making, MDCM or post-Pareto optimality analysis [2,3]. First, a data matrix is formed, the columns of which contain the information to be compared (*e.g*. methods, models, *etc*.) and the rows contain the samples, objects, *etc*. Then, reference column is defined, *i.e*. an "exact golden reference" or the average, minimum, maximum, or other data fusion possibility. The data values in each column are ranked in increasing magnitude and these rankings are compared to the ranks of the reference column. The absolute differences between the rank-variables and the rank-reference columns in each case are calculated and summed. These values (SRD values) give us the ordering of the variables. The smaller the SRD value, the better (or the more consistent) the variable. Finally, the rankings are validated using a randomization test and an n-fold cross-validation. SRD was coupled with ANOVA and applied four, carefully selected datasets

Conclusions: The choice of intercorrelation limits during molecular descriptor preselection has a significant effect on the outcome.

Overall the lower (around 0.80) limits deteriorate the resulting models (by removing valuable descriptors). The region between 0.95 and 0.9999 is applicable (recommended) for variable reduction, keeping in mind that it is worth to check more than one limit before finalizing the selection, as the specific choice is inherently dataset dependent.A seemingly insignificant change (like a setting of 0.9999 instead of 0.999) can remove a significant number of descriptors.

We would strongly suggest to authors of future QSAR studies to disclose the specific intercorrelation limits they applied for reproducibility of the whole modeling workflow.

Acknowledgement
This work was supported by the National Research, Development and Innovation Office of Hungary (NKFIH) under grant numbers OTKA K 119269 and KH_17 125608.

References

[1] Károly Héberger, Sum of ranking differences compares methods or models fairly, *TRAC - Trends in Analytical Chemistry*, **29** (2010) 101-109
[2] Anita Rácz, Dávid Bajusz and Károly Héberger, Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters, *SAR and QSAR in Environmental Research*, **26**, (2015) 683-700.
[3] Joao Lourenco and Luiz Lebensztajn Post-Pareto Optimality Analysis with Sum of Ranking Differences, *IEEE Transactions on Magnetics*, **54**, Issue: 8 (Aug. 2018) pp. 1-10. https://doi.org/10.1109/TMAG.2018.2836327

Thermal stability of ionic liquids under the conditions of synthesis of TiO_2-based photocatalysts: chemometric studies

A. Rybińska-Fryca[1*], *A. Mikołajczyk*[1], *J. Łuczak*[4], *M. Paszkiewicz-Gawron*[2], *A. Zaleska-Medynska*[2], *M. Paszkiewicz*[3] *and T. Puzyn*[1]

[1]Laboratory of Environmental Chemometrics, Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland, anna.rybinska@phdstud.ug.edu.pl

[2]Department of Environmental Technology, Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland

[3]Department of Environmental Analysis, Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland

[4]Department of Process Engineering and Chemical Technology, Chemical Faculty, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland

Ionic liquids (IL) are salts composed of large organic cation and inorganic/organic anion, which by definition are liquid below 100 °C. They have generated interest due to the specific properties such as negligible volatility, high polarity, and ionic conductivity. However, the real strength of ILs is related with the possibility to modify the ions structure and as a consequence to "tune" their physicochemical properties [1]. Moreover, ILs can be used as reaction environment for nano- and microstructures preparation, replacing the volatile organic solvents. It was shown that structure, thus properties of the final product (IL-TiO_2) are determined by the ability of IL to interact with growing TiO_2 particles and adsorption of the salt at the TiO_2 surface or incorporation in the crystalline structure.

ILs were found to have ability to activate the TiO_2 photocatalyst in the visible irradiation (Vis, $\lambda > 420$ nm). The mechanism of photoexactation under Vis for the samples modified with ionic liquids probably occurred in two ways, depending on the degree of ILs decomposition: (i) by creating a surface complex with energy transfer or (ii) by doping with nitrogen atoms originated from the ionic liquid [2,3].

In order to improve process of design and synthesis of new IL-TiO2 photocatalyst, we decided to investigate how the structure of the ionic liquid affects the thermal stability under the conditions of synthesis. The developed approach consists three steeps: (1) creating virtual library of 836 combinatorically created and theoretically characterized ILs structures (candidates); (2) selection of most promising ILs structures for efficient TiO_2-based photocatalyst synthesis; (3) chemoinformatic analysis of the relationship between the structure of the ionic liquid and its thermal stability under the conditions of synthesis [4].

References:

[1] H. Weingartner, *Angew Chem Int Ed Engl*., **47**, (2008) 654–670.

[2] M. Paszkiewicz-Gawron et al., *ACS Sustain. Chem. Eng*., **6**, (2018) 3927–3937.

[3] J. Łuczak *et al*., *ChemCatChem*, **9**, (2017) 4377–4388.

[4] A. Rybińska-Fryca *et al*., Thermal stability of ionic liquids under the conditions of synthesis of TiO_2-based photocatalysts: experimental and theoretical approach (in preparation)

Chemometric methods in characteristic of zeolites for specific applications

Mariusz Sandomierski[*], *Zuzanna Buchwald, Monika Zielińska, Adam Voelkel*
Poznan University of Technology, Institute of Chemical Technology and Engineering, Poznań, Poland
E-mail: mariuszsandomierski@wp.pl

Zeolites are crystalline aluminosilicates with a porous structure which is crucial for many industrial applications. Zeolites are promising for detoxication, controlled drug delivery and tissue engineering. They are also used as fillers for many polymers. Composites based on methacrylate resins and zeolites show high mechanical strength. There are several different types of zeolites in the environment which are non-toxic to living beings what affects their potential application in the drug release process. One of the zeolite types used in this process is type A. Type A zeolites are flavorless, odorless, harmless and have high cation exchange capacity. Ability of zeolites to ion exchange allows for incorporation of some amount of the calcium cations in their structure. Ca^{2+} ions are considered to be a confirmed anticaries agents. Calcium ions delivered from the external sources to the human mouth environment are able to rebuild the hydroxyapatite.

The purpose of this work was to receive type A and X zeolites with high ion exchange capacity and to apply these materials as active fillers in the dental composites with the remineralizing potential and potential material for drug release. All zeolites were subjected to the ion exchange process. As a result, a calcium form of these materials were prepared. The effectiveness of synthesis was confirmed by infrared spectroscopy, X-ray Diffractometry, scanning electron microscopy, energy dispersive spectroscopy and nitrogen adsorption/desorption measurements.

The remineralizing potential was specified as an ability to release calcium ions during the incubation in saline with the use of inductively coupled plasma-mass spectrometry. Composites containing calcium form of zeolites proved to have the ability to release calcium ions. The ability to release calcium ions and good mechanical properties indicates potential of prepared composites in dental applications. The second type of application was the release of drugs. Zeolite properties as a carrier of drugs and the possibility of their release were tested using UV-VIS spectroscopy. The obtained results were evaluated using chemometric methods.

Acknowledgments
This work was supported by the Polish Ministry of Science and Higher Education (Grant No. 03/32/SBAD/0900)

Multi-criteria decision making–Comparing lettuce types by their phytonutrient content

L. Sipos[1], *I. F. Boros*[1], *K. Madaras*[2], *L. Csambalik*[2], *A. Gere*[1]
[1] Department of Postharvest Sciences and Sensory Evaluation, Faculty of Food Science, Szent István University, 29-43 Villányi út, H-1118 Budapest, Hungary,
E-mail: sipos.laszlo@etk.szie.hu;
[2] Department of Ecological and Sustainable Production Systems, Faculty of Horticultural Science, Szent István University, 29-43 Villányi út, H-1118 Budapest, Hungary

Consumption of fruits and vegetables rich in valuable phytonutrients is proven to have several health benefits[1]. Despite the fact, that lettuce (*Lactuca sativa* L.) contains 95% of water, it is a valuable source of dietary fibers, folate, vitamins and phenolic compounds with antioxidant properties[2]. International scientific literature evaluates the phytonutrient content of varieties and types mostly from a single point of view, concentrating on independent parameters. Our aim was to provide a clear message to consumers with regards to the phytonutrient content of these leafy vegetables.

Sum of ranking differences[3] (SRD) as a powerful statistical method was used and the theoretical best type (formed from the collected dataset[4,5] as that got the higher value from analyzed attributes of every types) was the reference for the comparison of lettuce types in this study. So theoretical best type was formed as that got the higher value (Max) from analyzed attributes of every types. Pairwise comparison of the significant difference of the uncertainty of SRD values defined with leave-one-out cross-validation (LOO). According to the results of the Bartlett's sphericity test and Kaiser-Meyer-Olkin test, the principal component analysis (PCA) was acceptable. Cluster analysis of the factor score data of the PCA (Agglomerative hierarchical cluster, Euclidean distance, Ward's method) was also applied. Different statistical methodologies – validated SRD method, PCA, CA – gave consensual results, which shows the stability of the sequence of lettuce types. Based on the phytonutrient values of different lettuce types, Leaf (red) type was the closest to theoretical best type, followed by Leaf (green), Romaine (green), while those with the lowest performance were Crisphead and Butterhead types. Both Sign and Wilcoxon tests confirmed, that each lettuce types significantly divide from each other (α=0.05), except the Crisphead and Butterhead types, which belong to the same group. As a solution, validated SRD can support decision-making, being a quick and effective tool for sample ranking.

Acknowledgement
This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The Project is supported by the European Union and co-financed by the European Social Fund (grant agreement no. EFOP-3.6.3-VEKOP-16-2017-00005 and EFOP-3.6.1-16-2016-00016). AG thanks the support of Premium Postdoctoral Researcher Program of Hungarian Academy of Sciences and the National Research, Development and Innovation Office of Hungary (OTKA, contract No K119269).

References

[1] D. F. Birt, S. Hendrich and W. Wang, *Pharmacol. Ther.*, **90** (2001) 157–177.
[2] C. Nicolle, A. Carnat, D. Fraisse, J. Lamaison, E. Rock, H. Michel, P. Amouroux and C. Remesy, *J. Sci. Food Agric.*, **84** (2004) 2061–2069.
[3] X. Liu, S. Ardo, M. Bunning, J. Parry, K. Zhou, C. Stushnoff, F. Stoniker, L. Yu and P. Kendall, *LWT- Food Science and Technology*, **40** (2007) 552–557.
[4] M. J. Kim, Y. Moon, J. C. Tou, B. Mou and N. L. Waterland, *J. Food Compos. Anal.*, **49** (2016) 19–34.
[5] K. Héberger and K. Kollár-Hunek, *J. Chemom.*, **25** (2011) 151–158.

Analysis of protein glycosylation in *rheumatoid arthritis*

D. Szabó[1,2], *A. Ács*[1,3], *F. Auer*[4], *B. Rojkovich*[5], *Gy. Nagy*[5], *P. Géher*[5], *G. Sármay*[4], *L. Drahos*[1], *K. Vékey*[1]
[1]Hungarian Academy of Sciences, MS Proteomics Research Group
[2]Eötvös Loránd University, Hevesy György PhD School of Chemistry
[3]Semmelweis University, Károly Rácz School of PhD Studies
[4]Eötvös Loránd University, Department of Immunology
5. Polyclinic of the Hospitaller Brothers of St. John of God, Rheumatology Department III., Budapest, Hungary
Email: szabo.daniel@ttk.mta.hu

Rheumatoid arthritis (RA) is a chronic, progressive, systemic autoimmune disease, estimated to affect 0.5-1% of the adult population. RA exerts influence on, among other processes, the glycosylation of the proteins. Glycosylation consists of various oligosaccharide structures covalently linked to the proteins. It is most often characterized by the glycosylation pattern, which comprises the relative amount of these structures. Uncovering changes induced by RA in the glycosylation of proteins could lead to greater understanding of its pathology as well as earlier or more accurate diagnosis of the disease.

The glycoproteomic analysis of the N-glycosylation of immunoglobulin G (IgG) extracted from sera samples of both healthy and diseased individuals was performed. IgG was chosen as the compound of interest because it is the most abundant immunoprotein in human serum. Relative concentration of 23 different oligosaccharide structures was measured by HPLC-MS experiments. Based on the individual glycan abundances further parameters corresponding to certain structural features were determined, namely: the degrees of fucosylation (F), galactosylation (G), sialysation (S), and the ratio of structures featuring a glucosaminoglycan unit at the bisecting position (B). Differences in these derived parameters between healthy and diseased individuals can be related to the alteration of the biological processes governing the presence of such structural features.

To uncover the differences between the glycosylation pattern of the healthy and diseased samples multivariate statistical methods were used. Principal Component Analysis was performed on the relative concentration data of various glycoforms, as well as on the structure-related parameters (S, F, G, B). Linear Discriminant Analysis was employed to make an efficient differentiation between healthy individuals and RA patients.

Optical multisensor system for fat and protein determination in milk

A. Surkova [1, 2], *A. Bogomolov* [1, 2], *A. Legin* [1], *D. Kirsanov* [1]
[1] Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia;
[2] Samara State Technical University, Samara, Russia;
E-mail: melenteva-anastasija@rambler.ru

Optical multisensor system is a device composed of several one-channel sensors optimized for specific analytical tasks. Modern methods of multivariate data analysis allow compensating non-selective response of individual sensors and ensuring the accurate qualitative or quantitative analysis of object under study. Optical multisensor systems can be successfully applied in various fields, such as food quality control, cancer diagnostics and in-line process monitoring in biotechnology. Development of optical multisensor for a particular application includes several steps: finding an optimal configuration (i.e. the light-emitting diodes (LEDs) number and their spectral characteristics) of a sensor, building the calibration model based on a representative set of samples, solving the problem of model transfer and finally, providing an independent model update system (for instance, using the cloud software).

In the present work, an approach to the construction of an optical multisensor system for milk analysis has been developed. Visible and short-wave near infrared (Vis/SW-NIR) region is the most promising for the development of optical multisensor systems, since the details, such as light sources, optical fibers and detectors are low costs and commercially available for a wide range of producers. Recently proposed scatter-based method for fat and protein content determination in milk in the region 400–1100 nm [1–2] can be used as a basis for such low-cost portable milk analyzers. Preliminary full-spectral investigation should be carried out on representative dataset to find the most relevant spectral intervals that could be used to select appropriate LEDs for an optical multisensor system. To achieve this goal quantitative models for fat and protein determination in milk have been built in the region 400–1100 nm using partial least-squares (PLS) regression. Different ways to construct global calibration models for raw and normalized (homogenized milk with standardized nutrient content) milk have been developed. A particular attention was paid to calibration transfer between different types of the instruments.

The reported results can be used to construct a portable LED-based sensor system for different type of milk.

Acknowledgement
This study was supported by the RFBR-NSFC project #18-53-53016 GFEN_a.

References

[1] A. Bogomolov, S. Dietrich, B. Boldrini, R.W. Kessler, *Food Chem.*, **134** (2012) 412-418.
[2] A. Bogomolov, A. Melenteva, *Chemometr. Intell. Lab. Syst.*, **126** (2013) 129-139.

Composition of cometary particles *versus* distance to sun during sample collection - based on multivariate evaluation of mass spectral data (Rosetta/COSIMA)

*Varmuza K.**[1], *Filzmoser P.*[1], *Fray N.*[2], *Cottin H.*[2], *Merouane S.*[3], *Stenzel O.*[3], *Kissel J.*[3], *Briois C.*[4], *Baklouti D.*[5], *Bardyn A.*[6], *Siljeström S.*[7], *Silén J.*[8], *Hilchenbach M.*[3]

[1] Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Research Unit Computational Statistics, Vienna, Austria
E-mail: kurt.varmuza@tuwien.ac.at

[2] Laboratoire Interuniversitaire des Systèmes Atmosphériques, Université Paris Est Créteil et Université Paris Diderot, Créteil, France

[3] Max Planck Institute for Solar System Research, Göttingen, Germany

[4] Laboratoire de Physique et Chimie de l'Environnement et de l'Espace, Université d'Orléans et du CNES, Orléans, France

[5] Institut d'Astrophysique Spatiale, Université Paris Sud, Orsay, France

[6] DTM, Carnegie Institution of Washington, Washington, DC, USA

[7] Bioscience and Materials, Research Institute of Sweden, Stockholm, Sweden

[8] Finnish Meteorological Institute, Helsinki, Finland

The instrument COSIMA [1] onboard of the ESA spacecraft Rosetta collected dust particles in the neighborhood of comet Churyumov-Gerasimenko [2]. The distance to the sun during the sample collections varied between 1.2 and 3.8 AU (AU for astronomical unit, 150 000 000 km). The chemical composition of the particle surfaces was characterized by COSIMA using TOF-SIMS (time-of-flight secondary ion mass spectrometry). A set of about 3000 spectra has been selected, and relative abundances for CH-containing positive ions (from organic compounds [3-6]) as well as elemental ions (from minerals [7]) define a set of multivariate data. Evaluation by chemometric techniques indicates different compositions of samples collected at different distances to the sun. The applied methods comprise (1) considering the compositional nature of the used mass spectral data and centered log-ratio transformation [8]; (2) robust PCA [9]; (3) KNN-classification with repeated double cross validation [10].

Acknowledgement
Supported by the Austrian Science Fund (FWF), project P 26871 - N20.

[1] Kissel J., *et al.*: *Space Sci. Rev.*, **128**, 823 (2007)
[2] Schulz R., *et al.*: *Nature*, **518**, 216 (2015)
[3] Fray N., *et al.*: *Nature*, **528**, 72 (2016)
[4] Bardyn A., *et al.*: *MNRAS*, **469,** Suppl_2**,** S712 (2017)
[5] Fray N., *et al.*: *MNRAS*, **469,** S506 (2017)
[6] Varmuza K., *et al.*: *J. Chemometrics*, **32**, e3001 (2018)
[7] Stenzel O. *et al.*: *MNRAS*, **469,** Suppl_2, S492 (2017)
[8] Filzmoser P., Hron K., Templ M.: *Applied compositional data analysis*, Springer Nature, Cham, Switzerland (2018)
[9] Hubert M. *et al.*: *Technometrics*, **47**, 64 (2005)
[10] Varmuza K., Filzmoser P.: In Khanmohammadi M. (ed.), *Current applications of chemometrics*, p. 15-31; Nova Science Publishers, New York, USA (2015)

Self-organizing maps as an approach for monofloral bee honeys botanical origin determination

Ts. Voyslavov[1], *E. Mladenova*[1], *R. Balkanska*[2]
[1] Faculty of Chemistry and Pharmacy, University of Sofia,
1 J. Bourchier Blvd., 1164 Sofia, Bulgaria,
E-mail: voyslavov@abv.bg
[2] Department "Special branches –bees", Institute of Animal Science,
2232 Kostinbrod, Bulgaria

From early times, bee honey has been used in traditional medicinal practices in a lot of different cultures. Different kinds of honey (acacia, lime, sunflower, rapeseed, etc.) have different medicinal uses, as well as taste and flavor. The botanical origin of the product is very important for the customers.

Pollen analysis is the only certified approach for botanical origin determination of different monofloral bee honeys all over the world. The method is very expensive, time consuming and requires highly qualified specialists in order to obtain accurate and useful results.

The aim of the present study is to find a reliable approach, combining the determination of well-known physicochemical parameters and selected chemical element contents with intelligent multivariate statistical technique, instead of classical pollen analysis. The neuron nets method is useful for multivariate data processing for different types of samples, but according to authors' paper research, the method of Self-organizing maps is not applied in honey science. A good separation was obtained in accordance with the botanical origin of the honey samples. According to the U-matrix all honey samples formatted few groups for different monofloral honeys respectively.

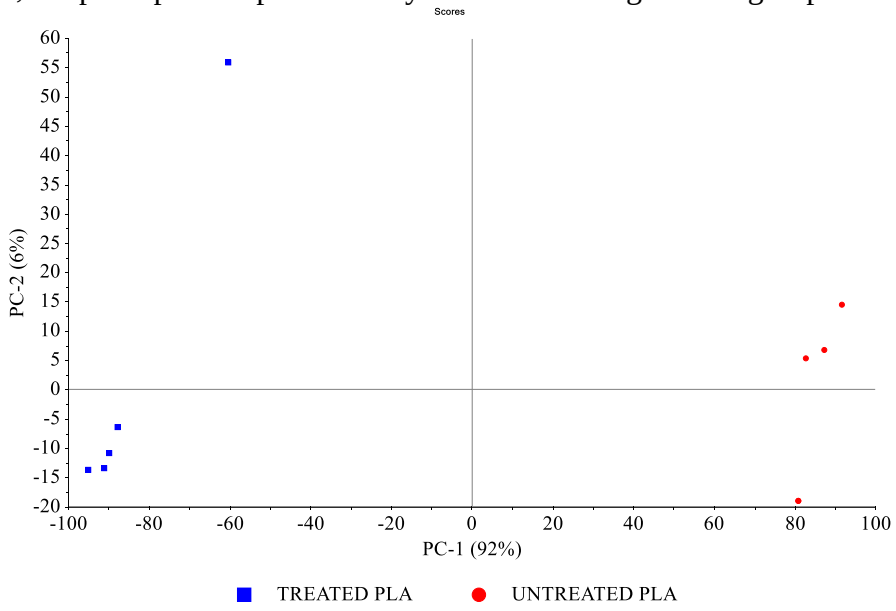
Infrared analysis of chemically modified 3D printed PLA scaffolds

M. Csontos[2], *J. Elek*[1], *I. Bácskai*[2], *P. Arany*[2]
[1] Science Port Kft, E-mail: elek@scienceport.hu
[2] University of Debrecen, E-mail: csontos.mate@science.unideb.hu

The widespread use of 3D printing made the computer engineered implants more easily available for the medical sciences. For the proper fine-tune of the product to certain applications, chemical functionalization might be necessary, which allows the optimization not only the physical but the biological properties as well.

In this case, amino side chain compounds were used to modify the biocompatibility of poly lactic acid “artificial bone” implants. The scaffold was placed in aqueous solution of the modifiers for 24 hours then rinsed with distilled water. The analytical task was to investigate if the modification of the poly-lactic acid happens through simple surface absorption or covalent chemical bonds are formed during the process.

MID-IR spectroscopy was chosen for the analysis with ATR sampling. While practically no visible difference was observed between the spectra of the treated and untreated PLA samples, the principle component analysis resulted in significant group formation.



Looking into the loading values it can be uniequivocally showed that the origin of the spectral differences can be attributed to the formation and transformation of amide C-N and C=O bonds, respectively. Therefore, it can be stated that the classification between the treated and untreated samples is based on the presence of newly formed amide bonds and the parallel transformed acyl groups [1].

Principal component analysis proved to be an excellent tool to visualize the slight changes in the FT-IR spectra, and by understanding the loading plots chemical evidence was presented to support the hypothesis of covalent bonding between the PLA skeleton and the modifier substance.

Reference

[1] A.J. Domb, L. Turovsky, R. Nudelman, Chemical interaction between drugs containing reactive amines with hydrolysable insoluble biopolymers in aqueous solutions, Pharmaceutical Research, Vol. 11, No. 6, pp. 865-868, 1994

Multicriteria optimization of raspberry convective drying processes

Zoran Stamenković[1], *Ivan Pavkov*[1]*, *Milivoj Radojčin*[1],
Krstan Kešelj[1], *Siniša Bikić*[2], *Attila Gere*[3]
[1]Faculty of Agriculture, University of Novi Sad, TrgDositejaObradovića 8,
21000 Novi Sad, Serbia, *Correspondence: ivan.pavkov@polj.uns.ac.rs
[2]Faculty of Technical Science, University of Novi Sad,
Bulevar cara Lazara 6, 21000 Novi Sad, Serbia,
[3]Faculty of Food Science, Szent István University, Villányi út, 29-43,
H-1118, Budapest, Hungary

Serbia is one of the leading countries in the world for the production and trade of raspberries. In regard to raspberry shelf life, only 10 % is immediately used for processing or sold on open markets and 85 – 90% of produced Serbian raspberries are frozen [1].

Processing of raspberries often leads to the loss of their natural properties, which is even more pronounced by using heat treatment and oxygen presents. This is particularly present in the most commonly used convective drying technology. Therefore, freeze-drying of raspberry is established as a standard in the industry due to the preservation of natural properties. On the downside, researches have shown that freeze-drying is one of the most energy-demanding technologies leading to a higher production cost [2,3].

Stamenković *et. al*., 2019., shows that the dried raspberry variety Polana, which is intended as an additive to some confectionery products, biscuits, cookies, dairy product *etc.*, may be considered as a proper substitution for freeze-dried raspberry. The experiment involves twelve different convective drying regimens applied on raspberry and the product quality was compared with freeze-dried samples by physicochemical criteria.

Convective drying regimens were performed by combining three influenced convective drying factors: air temperature at 60, 70 and 80 °C, air velocity at 0.5 and 1.5 $m \cdot s^{-1}$ and raspberry condition before drying (fresh and frozen).

Used quality criteria are chemical properties (L-ascorbic acid, total phenolic content, flavonoid content, anthocyanin content and anti-oxidative activity), physical properties (volume shrinkage, shape change and rehydration capacity) and mechanical properties (hardness and crispiness).

Sum of ranking differences method (SRD) [4] was used to compare the convective drying processes based on the measured properties. Quality parameters obtained with freeze-dried technology are used as benchmark criteria in order to define the most suitable method to replace freeze drying. From technological perspectives, convective drying of fresh raspberry at T=60 °C with air velocity of 1.5 $m \cdot s^{-1}$ proved to be the most suitable drying method. It was obvious that fresh raspberries had less physical changes than frozen ones. Chemical criteria show that convective drying reduced L–ascorbic acid content 80-99.99%, but less than 50% for other biologically active compounds as compared to freeze-dried raspberries. However, SRD analysis provides further insights to the differences and similarities among the used drying methods.

References

[1] Serbia, S.Y.o.t.R.o. Published and printed by: Statistical Office of the Republic of Serbia, Belgrade. (2018).
[2] W.Yueyue, D. Xu, R.Guangyue, L. Yunhong, *Drying Technology. Comparative study on the flavonoids extraction rate and antioxidant activity of onions treated by three different drying methods*, **37** (2018)245-252.
[3] R. Stanisław, *TEKA Kom. Mot. Energ. Roln. – OL PAN.Energy consumption in the freeze - and convection-drying of garlic*, **9** (2009) 259–266.
[4] K. Héberger,*TrAC - Trends in Analytical Chemistry*, **29** (2010), 101–109.

Author index

L.M. Lagares	L13
A. Legin	L20, P20
S. Lee	L25, P07
L. Lőrincz	L22
J. Łuczak	P16
T. Lundstedt	L24
K. Lundstedt-Enkel	L24
K. Madara	P18
E. Markovics	L25
R. Martín-Jiménez	L24
S. Merouane	P22
I. Mészáros	P11
M. Mihalovits	P10
A. Mikołajczyk	P16
Cs. Millei-Raffai	P11
N. Minovski	L13
E. Mladenova	P22
S. Mole	L24
Y. Monakhova	L05
S. Mostafapour	P05
Gy. Nagy	P19
Zs. I. Németh	P11, P12
M. Novič	L13
D. Nyitrainé Sárdy	P03
K. Pásztor-Huszár	P02
J. Petschnigg	L24
A. Pesti	P13
L. Pieszczek	L17, L18
M. Paszkiewicz-Gawron	P16
M. Paszkiewicz	P16
I. Pavkov	P24
É. Pusztai	P14
T. Puzyn	P16
A. Rácz	L26, L27 P01, P03, P15
M. Radojčin	P24
R. Rákosa	P11, P12
B. Rojkovich	P19

L. Románszki	P04
K. Roy	L03, P06
C. Ruckebusch	L16
C. Russel	L24
W. Saeys	L14
A. Rybińska-Fryca	P16
M. Sandomierski	L06, P17
G. Sármay	P19
V. Schlenker	L09
J. Silén	P22
S. Siljeström	P22
A.M. Sipos	P02
L. Sipos	L22, P18
A.K. Smilde	L10
G. Smuk	P13
I. Stanimirova	L19
O. Stenzel	P22
B. Strzemiecka	L06
A. Surkova	L20, P20
D. Szabó	P19
M. Szász	L25
A. ter Laak	L09
R. Todeschini	L04
G. Tóth	L02, P08, P09
J. Trygg	L24
M. Tušar	L13
A. Vágvölgyi	P11
M. Vargovics	P12
K. Varmuza	P21
K. Vékey	P19
A. Voelkel	L06, P17
T. Voyslavov	P22
A. Zaleska-Medynska	P16
Z. Stamenković	P24
M. Zielińska	P17