**Drushbametrics Project**

Ninth Winter Symposium on Chemometrics

# Modern Methods of Data Analysis

Russia, Tomsk, February 17 – 21, 2014

**Drushbametrics Project**

Ninth Winter Symposium on Chemometrics

# Modern Methods of Data Analysis

Russia, Tomsk, February 17 – 21, 2014

Russian Chemometrics Society

Scientific Consul for Analytical Chemistry RAS

Semenov Institute of Chemical Physics RAS

Vernadsky Institute of Geochemistry and Analytical Chemistry RAS

National Research Tomsk Polytechnic University

With support of

Russian Foundation for Basic Research

Umetrics

Drushbametrics

WSC 9

project

Ninth Winter Symposium on Chemometrics

# Modern Methods of Data Analysis

## The organizing committee

**Co-chair person**

Lev Gribov

Sergey Romanenko

**Secretary**

Ekaterina Cherednik

**Members**

Sergey Kucheryavskiy

Oxana Rodionova

Alexey Pomerantsev

Federico Marini

Sergey Zhilin

Tomsk Polytechnic University, Lenina av. 30, 634034 Tomsk, Russia

http://wsc.chemometrics.ru/wsc9 email: wsc9@chemometrics.ru

## Thanks

The WSC-9 organizers and participants wish to express greatest appreciation to the following conference sponsors for their valuable economic and friendly help:

Russian Foundation for Basic Research (RFBR)

Umetrics

CAC 2014 Organizing Committee

Finally, we are grateful to all the WSC-9 attendees, lecturers, accompanying persons, and visitors for their interest to the conference.

**See you again at the next WSC-10 conference!**

# Useful information

## Conference and activities

Symposium sessions will be held at the conference room at the administrative building, excepting February 20 when session will be organized in building 3. Skis, sauna, etc. are available for rent at the administrative building.

## Meals

All meals (buffet) will be served in administrative building. The banquet will take place in the administrative building.

## Scores & Loadings

The «Scores and Loadings» meeting will be held in administrative building with drinks at reasonable prices.

## Communication

The three Russian cellular networks, Beeline, MTS and Megafon, have a proper coverage around the hotel. Internet WiFi is also available.

## Connection with Tomsk

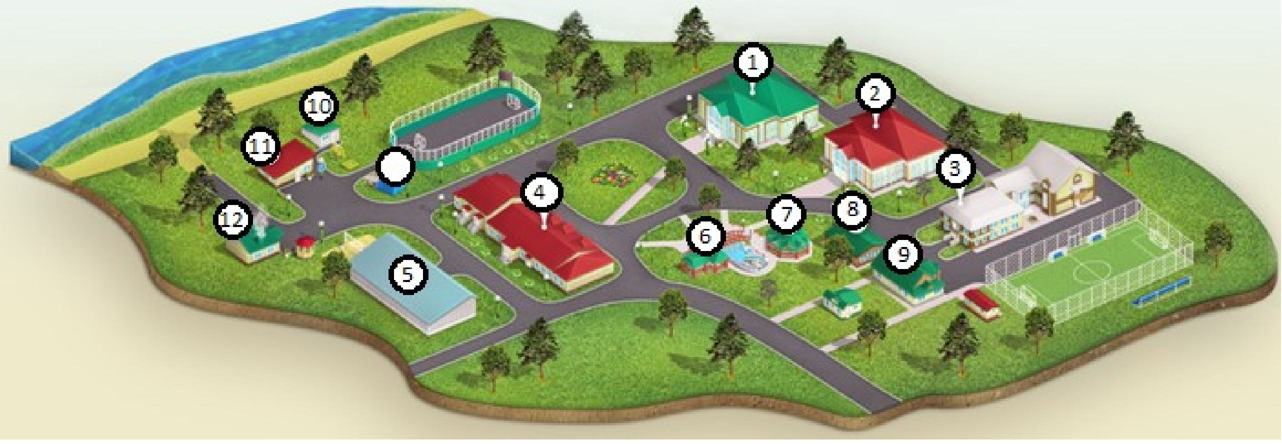
The hotel is located about 30 km from Tomsk. Tomsk is available by taxi (300 rub), or by the public buses.

## Miscellaneous

The conference official language is English.

Everyone is encouraged to have his/her badge attached, both during the symposium and social activities.

## Map of Hotel Tom



1 Building №1

2 Building №2

3 Building №3

4 Administrative building

5 Rent a ski, etc

6 Porch №1

7 Porch №2

8 Porch №3

9 Staff room

10 Shooting Area

11 Spa-complex

12 Sauna

## Useful Phone Numbers

| | |
|---|---|
| Ekaterina Cherednik, symposium secretary | +7(923)404 1684 (mobile) |
| Tom reception | +7(3822) 96-71-64 (local) |

**Monday, 17 February, 2014**

| | |
|---|---|
| 09:00 – 12:30 | **Registration + Coffee-break (11:00-11:30)** |
| 13:00 – 14:00 | **Lunch** |
| 14:00 – 14:30 | **Opening** |
| **Session 1** | **Chair: Sergey Romanenko** |
| 14:30 – 15:00 | **T1** *Pentti Minkkinen* Weighting error – a potential source of systematic measurement errors in process analysis |
| 15:00 – 15:30 | **T2** *Oxana Rodionova* NIR measurements with a fiber-optic probe |
| 15:30 – 16:00 | **T3** *Dmitry Kirsanov* Novel calibration design for multiple components |
| 16:00 – 16:30 | **Coffee-break** |
| **Session 2** | **Chair: Alexey Pomerantsev** |
| 16:30 – 17:30 | **L1** *Andrey Bogomolov* Designing a multi-component calibration experiment |
| 17:30 – 18:00 | **T4** *Irina Yaroshenko* Potentiometric multisensor system for clinical diagnostics of urolithiasis |
| 18:00 – 18:30 | **T5** *Maxim Savinkov* Presentation of CSort company |
| 18:30 – 19:00 | **Free time** |
| 19:00 – 20:00 | **Dinner** |
| 20:00 – 22:00 | **Scores & Loadings** |

**Tuesday, 18 February, 2014**

| | |
|---|---|
| 08:30 – 09:30 | **Breakfast** |
| **Session 3** | **Chair: Oxana Rodionova** |
| 09:30 – 10:30 | **L2** *Alexey Pomerantsev* Dual data driven SIMCA as a one-class classifier |
| 10:30 – 11:00 | **T6** *Vladislav Galyanin* Selecting optimal spectral regions for sensor analysis |
| 11:00 – 11:30 | **Coffee-break** |
| 11:30 – 12:00 | **T7** *Evgeny Karpushkin* Morphology assessment of polymer hydrogels using multivariate analysis of viscoelastic and swelling properties |
| 12:00 – 13:00 | **Free time** |
| 13:00 – 14:00 | **Lunch** |
| **Session 4** | **Chair: Pentti Minkkinen** |
| 14:30 – 15:30 | **L3** *Sergey Shary* Maximum consistency method for data fitting under interval uncertainty |
| 15:30 – 16:00 | **T8** *Sergey Kumkov* Estimation of corrosion parameters for Metals in Oxygen Containing Molten Salts by Methods of Interval Analysis |
| 16:00 – 16:30 | **Coffee-break** |
| **Session 5** | **Chair: Federico Marini** |
| 16:30 – 17:30 | **L4** *Yizeng Liang* A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems |
| 17:30 – 18:00 | **T9** *Anastasiia Melenteva* Modeling of fat and protein content in raw milk based on historical spectroscopic data |
| 18:00 – 18:30 | **T10** *Aleksey Skvortsov* Possible improvements to multivariate curve resolution with particle swarm optimization: Estimation of Rotation Ambiguity |
| 18:30 – 19:00 | **Free time** |
| 19:00 – 20:00 | **Dinner** |
| 20:00 – 22:00 | **Scores & Loadings** |

**Wednesday, 19 February, 2014**

| | |
|---|---|
| 08:30 – 09:30 | **Breakfast** |
| **Session 6** | **Chair: Yizeng Liang** |
| 09:30 – 10:30 | **L5** *Yuri Kamambet* Confidence intervals, noise filtering and outliers |
| 10:30 – 11:00 | **T11** *Qing-Song Xu* Identifying Bioactive Signature in Natural Products through Chromatographic Fingerprint |
| 11:00 – 13:00 | **Cultural program** |
| 13:00 – 14:00 | **Lunch** |
| **Session 7** | **Chair: Isneri Talavera** |
| 14:00 – 15:00 | **L6** *Federico Marini* A novel multiway methods for regression on tensors with shifts along one mode |
| 15:00 – 15:30 | **T12** *Sergey Zhilin* Macrolevel Study of Russian Science Citation Index using PCA for Interval-Valued Data |
| 15:30 -16:00 | **T13** *Eketerina Delakova* The results of solving tasks of medical diagnosis for dental patients on the basis of chemometrics methods |
| 16:00 – 16:30 | **Coffee-break** |
| 16:30 – 18:30 | **Poster Session** |
| 18:30 – 20:00 | **Free time** |
| 20:00 – 21:00 | **Dinner** |
| 21:00 – 22:00 | **Scores&Loadings** |

**Thursday, 20 February, 2014**

| | |
|---|---|
| 08:30 – 09:30 | **Breakfast** |
| **Session 8** | **Chair: Sergey Shary** |
| 09:30 – 10:30 | **L7** *Isneri Talavera* Classification of continuous multi way data via dissimilarity representation |
| 10:30 – 11:00 | **T14** *Ekaterina Cherednik* Resolution of overlapped peaks in stripping voltammetry with stepwise mathematical resolution method |
| 11:00 – 12:00 | **Free time** |
| 12:00 – 13:00 | **Lunch** |
| 13:00 – 19:00 | **Excursion** |
| 20:00 – 00:00 | **Banquet** |

**Friday, 21 February, 2014**

| | |
|---|---|
| 08:30 – 09:30 | **Breakfast** |
| 10:30 | **Departure** |

# Abstracts

## L1 Designing a multi-component calibration experiment

*Andrey Bogomolov*
*Samara State Technical University, Samara, Russia*
*J&M Analytik AG, Essingen, Germany*

The necessity of multivariate design of experiment (DoE) for the efficiency of analysis and subsequent data modeling is commonly recognized [1]. However, in spite of its well-established theory and practice [2], the modern DoE is mainly focused at the optimization problem, i.e. finding a set of experimental parameters producing maximal or sufficiently high value of a chosen merit. The problem of designing an optimal calibration experiment stays almost beyond consideration by the theory of DoE.

One of the practically most important needs of quantitative mixture analysis is simultaneous calibration and prediction of several constituents from the same multivariate measurement, e.g. spectrum, using possibly few designed samples. Examples of this kind are numerous and include: two active ingredients in a tablet, fat and protein in milk, ethanol and glucose in the fermentation medium, to name a few. For robust regression modeling, the analyte concentrations (i.e. factors) in calibration samples should be spread on many levels. This distinction makes the classical experimental designs, typically operating in two to five levels, disadvantageous. Attempts to construct an economical DoE effective in the case of a few factors at many levels are rare [3, 4].

The present lecture reviews existing approaches to the calibration DoE and presents a new one. Suggested diagonal designs for a multi-component calibration experiment provide uncorrelated factor variations in as many levels as the number of samples. An independent validation set is provided by the very design scheme. Practicability of the diagonal designs is illustrated by selected applications.

1. R. Leardi, *Analytica Chimica Acta*, **652**, 161–172 (2009).

2. Eriksson L., Johansson E., Kettaneh-Wold N., Wikström C., & Wold S. (2008). Design of experiments: Principles and applications (3rd ed.). Umeå: Umetrics AB.

3. A. Bogomolov, S. Dietrich, B. Boldrini, R.W. Kessler, *Food Chemistry*, **134**, 412–418 (2012).

4. R.G. Brereton, *Analyst*, **122**, 1521–1529 (1997).

# L2 DUAL DATA DRIVEN SIMCA AS A ONE-CLASS CLASSIFIER

*A.L. Pomerantsev*[1,2]
[1]*Semenov Institute of Chemical Physics RAS, Moscow, Russia*
[2]*Institute of Natural and Technical Systems RAS, Russia*

SIMCA, the method of soft independent modeling of class analogy, was proposed about 40 years ago [1]. This approach is a natural extension of a well-known method of principal component analysis (PCA). The main SIMCA function is qualitative data analysis that can be formulated as following. Let us consider a class of objects **X** (substance, materials, products). A set of analytical measurements (spectra, chromatograms, etc.) represents the class of objects accounting for a possible variability. Suppose we have a new object **y**, for which we have to decide whether **y** belongs to class **X,** or not. An example of such a problem is the recognition of counterfeit drugs [2]. SIMCA has been revised repeatedly [3-6]. Today this method is very popular in analytical chemistry (chemometrics), but almost unknown outside. SIMCA provides a unique opportunity to make classification accounting both for the Type-I error α (false rejection) and the Type-II error β (false acceptance), however this is used extremely rare. The SIMCA theoretical base is thoroughly developed, but most of the analytical studies contain gross errors, which are repeated from publication to publication.

The presentation is going to bridge the gap and to provide a general SIMCA concept as a data driven method. The following items will be considered

- How PCA relates to SIMCA
- The distance measures in use: score distance, orthogonal distance, total distance
- What statistics are used in SIMCA and how this statistics are distributed
- How to make the decision at a given the Type I error α
- How to calculate the Type II error β

Presentation is illustrated with simple examples.

1. Wold S., *Pattern Recognition*, **8**, 127–139 (1976).
2. Rodionova O.Y., Pomerantsev A.L., *Trends Anal. Chem*., **29**, 781–938 (2010).
3. Nomikos P., MacGregor J.F., *Technometrics*, **37**, 41–59 (1995).
4. Pomerantsev A.L., *J. Chemometrics*, **22**, 601–609 (2008).
5. Hubert M., Rousseeuw P.J., Vanden Branden K., *Technometrics*, **47**, 64–79 (2005).
6. Pomerantsev A.L., Rodionova O.Y., *J. Chemometrics*, 2013 (DOI: 10.1002/cem.2506).

## L3 Maximum Consistency Method for Data Fitting under Interval Uncertainty

*Sergey P. Shary*
*Institute of Computational Technologies, Novosibirsk, Russia*

For the linear regression model $b = a_1x_1 + a_2x_2 + \cdots + a_nx_n$, we consider the problem of data fitting under interval uncertainty. Let an interval $m\times n$-matrix $\boldsymbol{A} = (\boldsymbol{a}_{ij})$ and an interval $m$-vector $\boldsymbol{b} = (\boldsymbol{b}_i)$ represent the input data and output responses of the model respectively, such that $a_1 \in \boldsymbol{a}_{i1}, a_2 \in \boldsymbol{a}_{i2}, \dots , a_n \in \boldsymbol{a}_{in}, b \in \boldsymbol{b}_i$ in the $i$-th experiment, $i = 1, 2, \dots , m$. It is necessary to find the coefficients $x_1, x_2, \dots, x_n$ that best fit the above linear relation for the data given.

A family of values of the parameters $x_1, x_2, \dots, x_n$ is called *consistent* with the interval data $(\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \dots , \boldsymbol{a}_{in}), \boldsymbol{b}_i$, $i = 1, 2, \dots , m$, if, for every index $i$, there exist such point representatives $a_{i1} \in \boldsymbol{a}_{i1}, a_{i2} \in \boldsymbol{a}_{i2}, \dots , a_{in} \in \boldsymbol{a}_{in}, b_i \in \boldsymbol{b}_i$ that $a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i$ . The set of all the parameter values consistent with the data given form a *parameter uncertainty set*. As an estimate of the parameters, it makes sense to take a point from the parameter uncertainty set providing that it is nonempty. Otherwise, if the parameter uncertainty set is empty, then the estimate should be a point where maximal “consistency” (in a prescribed sense) with the data is achieved.

The parameter uncertainty set is nothing but the solution set $\Xi(\boldsymbol{A}, \boldsymbol{b})$ to the interval system of linear equations $\boldsymbol{A}x = \boldsymbol{b}$ defined in interval analysis: $\Xi(\boldsymbol{A}, \boldsymbol{b}) = \{ x \mid Ax = b$ for some $A$ from $\boldsymbol{A}$ and $b$ from $\boldsymbol{b} \}$. For the above data fitting problem, we propose, as the consistency measure, the values of the *recognizing functional* of the solution set $\Xi(\boldsymbol{A}, \boldsymbol{b})$, which is defined as

$$\mathrm{Uss}\,(x, \boldsymbol{A}, \boldsymbol{b}) = \min_{1\le i\le m} \left\{ \mathrm{rad}\,\boldsymbol{b}_i + \sum_{j=1}^{n} (\mathrm{rad}\,\boldsymbol{a}_{ij})\,|x_j| - \left| \mathrm{mid}\,\boldsymbol{b}_i - \sum_{j=1}^{n} (\mathrm{mid}\,\boldsymbol{a}_{ij})\,x_j \right| \right\},$$

where “mid” and “rad” mean the midpoint and radius of an interval. The functional Uss “recognizes” the points of $\Xi(\boldsymbol{A}, \boldsymbol{b})$ by the sign of its values: $x \in \Xi(\boldsymbol{A}, \boldsymbol{b})$ if and only if Uss $(x, \boldsymbol{A}, \boldsymbol{b}) \ge 0$. Additionally, Uss has reasonably good properties as a function of $x$ and $\boldsymbol{A}, \boldsymbol{b}$.

As an estimate of the parameters in the data fitting problem, we take the value of $x = (x_1, x_2, \dots, x_n)$ that provides maximum of the recognizing functional Uss (Maximum Consistency Method). Then,

- if the parameter uncertainty set is nonempty, we get a point from it,

- if the parameter uncertainty set is empty, we get a point that still has maximum possible consistency (determined by the functional Uss) with the data given.

In our work, we discuss properties of the recognizing functional Uss, interpretation and features of the estimates obtained by the Maximum Consistency Method as well as correlation with the other approaches to data fitting under interval uncertainty.

1. S.P. Shary, *Automation and Remote Control*, **73** (2), 310–322 (2012).
2. S.P. Shary, I.A. Sharaya, *Computational Technologies*, **18** (3), 80–109 (2013), (in Russian).

## L4 A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems

*Yi-Zeng Liang, Yong-Huan Yun,and Qing-Song Xu*
*College of Chemistry and Chemical Engineering, Central South University, P. R. China*
*School of Mathematics and Statistics, Central South University, P. R. China*

Two chemical modeling spaces, say component spectral space and measured variable space, are firstly defined, respectively. From this point of view, classical calibration and inverse calibration can be two kinds of multivariate calibration in chemical modeling. It is worth noting that the intrinsic difference between these two calibration models is not fully investigated. The net analyte signal (NAS) proposed by Lorber[1] based on orthogonal projections can be regarded as the theoretic summary for clssic calibration. Also, the tensor calibration for high dimensionally linear data is its natural extension. However, in the case of complex analytical systems, NAS cannot be well defined in inverse calibration due to the existence of uninformative and/or interfering variables. Therefore, application of the NAS cannot improve the predictive performance for this kind of calibration, since it is essentially a technique based on the full-spectrum. From our perspective, variable selection can significantly improve the predictive performance through removing uninformative and/or interfering variables. Although the need for variable selection in the inverse calibration model has already been experimentally demonstrated, it has not aroused so much attention. In this study, we first clarify the intrinsic difference between these two calibration models and then use a new perspective to intrinsically prove the importance of variable selection in the inverse calibration model for complex analytical systems. In addition, we have experimentally validated our viewpoint through the use of one UV dataset and two generated near infrared (NIR) datasets.

## L5 Confidence intervals, noise filtering and outliers

*Yuri P.Kozmin*[2]*, Sergey Matsev*[1]*, Yuri,* *A.Kalambet*[1]
[1]*Ampersand Ltd., Moscow, Russia*
[2]*Institute of Bioorganic Chemistry RAS*

Recently developed optimal method of noise filtering is enhanced by outlier elimination module. Outlier elimination improves confidence intervals of the prediction and gives better results for noise filtering of time series. Principles and procedures of outlier detection and elimination are discussed. Practical examples of chromatographic signal filtering are presented.

## L6 SCREAM: A novel multiway methods for regression on tensors with shifts along one mode

*F. Marini*[1]*, R. Bro*[2]
[1]*Dept. of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro, Rome, Italy*
[2]*University of Copenhagen, Rolighedsvej 30, Frederiksberg C, Denmark*

Analytical instrumentation has developed to a point where many techniques provide outcomes that, for each sample, take the form of a landscape or a higher order array. Accordingly, several methods have been proposed during the years to directly process multi-way data, both for exploratory purposes (e.g. PARAFAC [1]) and for regression (multilinear PLS [2]) or classification (NPLS-DA, NSIMCA [3]). However, these same methods become less adequate, when the underlying profiles change shape from sample to sample, for instance in chromatography when there are retention time shifts in the elution profiles or in process analysis, when the batches have different lengths and/or are not synchronized. In such cases, if only a decomposition of the array is sought, reliable models can still be calculated using a suitable modification of the PARAFAC algorithm, called PARAFAC2 [4]. On the other hand, in the case of calibration problems, no alternatives to N-PLS have been proposed so far to cope with these limitations. To overcome this problem, in this communication a new regression method called SCREAM (Shifted Covariates REgression Analysis for Multi-way data) is proposed for calculating calibration models on multi-way arrays which present shifts (or shape changes) along one of the modes. The algorithm combines a PARAFAC-2 decomposition of the X array and a Principal Covariate Regression-like least squares criterion for the computation of the regression coefficients in a way which is analogous to what already described by

Smilde and Kiers in the case of other multi-way PCovR algorithms [5]. The method is tested on real and simulated datasets providing good results and performing as well or better than other available regression approaches for multi-way data.

1. Bro R., *Chemometr Intell Lab Syst*, **38**, 149–171 (1997).

2. Bro R., *J Chemometr*, **10**, 47–61 (1996).

3. Kiers HAL, ten Berge JMF, Bro R., *J Chemometr*, **13**, 275–294 (1999).

4. Salvatore E., Bevilacqua M., Bro R., Marini F., Cocchi M. In: De La Guardia M, Gonzalvez Illueca A (Eds.) Food protected designation of origin: methodologies & applications. Elsevier, Oxford, 2013, 339–382.

5. Smilde AK, Kiers HAL, *J Chemometr*, **13**, 31–48 (1999).

## L7 Classification of continuous multi way data via dissimilarity representation

*Diana Porro Muñoz[1], Isneri Talavera[1], Robert P.W Duin[2]*
*[1]Advanced Technology Applications Center Havana Cuba*
*[2]Delft University of Technology, The Netherlands*

Representation of objects by multidimensional data arrays has become very common for many research areas, as image analysis, signal processing and chemometrics.

Multiway data analysis is the extension of multivariate analysis when data is arranged in the multiway structure. Nevertheless, this type of data will not be analyzed optimally by the traditional multivariate methods, which do not take into account the multiway structure.

A number of methods for multiway analysis have been proposed. Most of these methods are for exploratory and regressions purposes. Classification has been less studied.

This work introduces the dissimilarity representation approach as a new tool for the classification of multiway data and also in this context a new 2D measure (2D Shape) for classification tasks in three ways spectral data is presented. This measure is based on the combination of ID measures taking into account the information of both measurements directions.

Even the good performance obtained comparing it with the 1D traditional approach and the 2D existing measures, this measure still has some limitations because does not take into account the simultaneous shape changes in both directions of the surfaces of 2D continuous data. To solve this problem a continuous multiway shape measurement (CMS) is presented, which is based on the differences between the gradients of objects. The new measure allows taking into account the complex multidimensional structure, such that the shape information of the surfaces (objects) can be considered in the dissimilarity representation of the objects. The way the measure has been defined, allows to use different gradient convolution kernels, according to the problem at hand. .

Results have corroborated the presented argument that considering the continuous multiway nature of these type of data in their analysis can lead to better results. Moreover, it is shown that by taking into account the information in more directions, results can be improved. This measure has also the advantage that is easily extended to high dimensions multiway data.

## T1 Weighting error – a potential source of systematic measurement errors in process analysis

*Pentti Minkkinen*
*Lappeenranta University of Technology, Lappeenranta, Finland*

The Weighting Error, *WE,* is a systematic error which is generated under certain conditions if the mean value of a lot is estimated as simple arithmetic mean. Cases where this can happen are:

- Lots consisting of several packages (sub-lots) of different sizes, each having different average concentrations. In this case it is obvious that the mean concentration of the analyte in the whole lot should be calculated as a *weighted average* by using the sizes of the sub-lots as statistical weights.
- Samples of constant size (volume or weight) are drawn from the target lot. If the density of the material varies in the lot and is *correlated* with the concentration of the analyte, the simple average estimate of the lot mean is *biased*. A correct average value is obtained if the individual results are weighted by using the densities or the masses of the samples as statistical weights in calculating the mean.
- The flow-rate of the moving stream of matter correlates with the concentration of the analyte and constant-volume samples are drawn and analysed, or combined into a composite sample. This procedure ignores the effect of the flow-rate variation and, consequently, the simple mean of the measurements is biased due to weighting error. As compositing of sub-samples is equivalent of averaging also the composite sample is biased. If the samples are cut with a constant speed proportional cross-stream cutter the sample masses can be used

> as weights in calculating the average because they are proportional to the flow-rate. Similarly, if these individual increments are combined into a composite sample before analysis, the composite sample is unbiased. If it is not possible to use proportional sampling, the flow-rates at the time of sampling can be used as weights, provided that reliable measurements are available. In this case, if composite samples are made, the increments should be drawn when a predetermined volume has passed the sampling device (volume-proportional sampling) to guarantee an unbiased composite sample.

Gy defined the *WE* in [1], but its properties were not further studied. In later publications Gy did not discuss this error further. His justification was that it can be eliminated if a correct cross-stream proportional sampling devices are used for cutting the samples. The mean of the sampling target (lot) should then be estimated as the weighted mean by using the sample masses as the weights in estimating the mean.

There are, however, many industrial sampling targets where proportional sampling is difficult and expensive to realize in practice, if not impossible. Consider, *e.g.* sampling dust emissions into the atmosphere from a stack of a large industrial soda recovery boiler, which might have a diameter up to eight meters. From this kind of sampling targets it is impossible to cut correct cross-sections from the stream. Only point intake samplers are in use today. A critique of traditional approaches to this type of challenging sampling situation can be found in Wagner & Esbensen [2,3].

As mentioned above, in order to eliminate the weighting error composite samples are sometimes collected so that the sampling device is coupled with a flow-meter, and an increment is taken when a predetermined total volume has passed the sampling point. In this approach the sampling interval is varied proportionally to the flow-rate. Some doubts have been presented concerning the effectiveness of this approach and difficulties have been experienced in convincing industrial partners to adopt this approach, especially because it is more difficult to implement than the systematic sampling at fixed time intervals. It is also often claimed that increasing the number of samples decreases the weighting error in estimating the lot average. It is also claimed sometimes that increasing the number of samples will decrease the weighting error. The study presented here shows, however, that this is not the case.

1. Gy P.M., Sampling of Particulate Materials, Theory and Practice, Elsevier, Amsterdam, 1982.
2. C. Wagner, K.H. Esbensen, *Chemical Engineering Research and Design*, **89** (9), 1572–1586 (2011), doi:10.1016/j.cherd.2011.02.028

3. C. Wagner, K.H. Esbensen, *Renewable and Sustainable Energy Reviews*, **18** (1), 504–517 (2012).

## T2 NIR measurements with a fiber-optic probe

*O.Ye. Rodionova*
*Semenov Institute of Chemical Physics RAS, Moscow, Russia*

When several near-infrared instruments are used in a network and a common chemometric model is applied for spectra processing, the instruments comparison is indispensable. Direct transferability often claimed by the producers should be treated with caution. It has been found experimentally that when measurements are performed with the help of a fiber-optic probe, the main source of spectra discrepancy is probes' sensitivity to contactless measurements. Here the influence of the probe-to-object distance on the acquired spectra is analyzed in detail. Special experimental set-ups are proposed to isolate various strongly influencing factors and to maintain stable measurement conditions. The application of an artificial standard instead of real-world objects helps to focus on the instrument/accessory characteristics.

It has been shown that a substantial impact on spectra quality is caused by the gap between the probe tip and the object. Spectra distortions include signal attenuation, low frequency effect, such as nonlinear baseline shift, and high frequency noise, mostly noticeable in the range where absorption is low. In some cases even a loss of peaks is observed (when object and reference have low intensity). Such spectra discrepancy is hardly compensated by common spectroscopic preprocessing methods. The performance of just two FT-NIR spectrometers equipped with fiber probes was analyzed in detail. However the experiments conducted with six other instruments from NIR network have revealed similar problems. Hence, the detected FPs differences are not attributable to the specific properties of the two tested probes. The problem is of a general nature.

As a rule, the producers of modern FT-NIR instruments claim high compatibility of the instruments of the same product line. However, the general instrument consistency does not guarantee the accessories' compatibility in various experimental set-ups. This issue should be taken into account when several FT-NIR spectrometers are consolidated in one network.

1. O.Ye. Rodionova, K.S. Balyklova, A.V. Titova, A.L. Pomerantsev, *Appl. Spectrosc.*, **67** (12), 1401–1407 (2013).

## T3 Novel calibration design for multiple components

*Dmitry Kirsanov, Vitaly Panchuk, Andrey Legin*
*Chemistry Department, St. Petersburg State University, St. Petersburg, Russia*

The multivariate calibration is a well-established technique in numerous fields of analytical chemistry. It is especially useful when dealing with multicomponent mixtures where classical least squares approach with a single variable fails because of e.g. a complex shape of analytical signal, lack of signal selectivity, etc. Usually to establish regression model one will require a set of reference data from another method/instrument, e.g. to calibrate NIR spectrometer for prediction of ash content in wheat grain one has to analyze all calibration samples with standard technique (burning in ash oven) first. This is the most straight-forward way of multivariate calibration and it allows for taking into account the influence of all the components in the real complex multicomponent mixtures since dealing with real samples. There are however certain cases when this direct approach cannot be used, e.g. when real samples are hardly available or very expensive. In this situation an obvious way for calibration is design of model mixtures simulating real samples. In case of only one component of interest an approach for design is quite obvious: the concentrations of the component must be evenly distributed along the concentration scale. When numerous components are of interest the situation is getting more tricky: e.g. to study all possible combinations of seven components with five particular concentration levels of each will require $5^7$=78125 different mixtures to be studied. Definitely this is far from being doable in a common laboratory practice. The literature on experimental designs for calibration with multiple components is quite sparse. The works of Brereton [1,2] with cyclic permutation design are worth of mentioning, however the number of samples in these designs is strictly fixed to provide for the orthogonality of components. We suggest an approach to calibration design which is valid with any number of components and any number of mixtures. Thus, the experimentator can adopt the design to his/her particular needs taking into account the number of mixtures affordable from “time and money” considerations. This approach is based on the algorithm of even distribution of points in a hypercube. The details on the algorithm and comparison of its performance with cyclic permutation design will be provided in the presentation.

1. R. Brereton, *Analyst*, **122**, 1521–1529 (1997).
2. J.A. Munoz, R. Brereton, *Chemometrics and Intelligent Laboratory Systems*, **43**, 89–105 (1998).

## T4 Potentiometric multisensor sytem for clinical diagnostics of urolithiasis

*Irina Yaroshenko*[1,2]*, Dmitry Kirsanov*[1,3]*, Luidmila Kartsova*[1]*, Andrey Legin*[1,3]
[1]*Chemistry Department, St. Petersburg State University, St. Petersburg, Russia*
[2]*Bioanalytical Laboratory CSU "Analytical Spectrometry", St. Petersburg, Russia*
[3]*Sensor Systems LLC, St. Petersburg, Russia*

The urolithiasis is a common disease affecting 20 % of the world population. At early stages the disease proceeds asymptomatically. The patient is aware of his illness when the stone is already formed and it obstructs normal organism activities. Early diagnostics is a half of the successful treatment, so biochemical analysis of urine is the main part to fight against the urolithiasis.

Certainly the modern analytical tools such as gas chromatography (GC), high performance liquid chromatography (HPLC), capillary electrophoresis (CE) are effective, sensitive and accurate. However, these methods have significant drawbacks – they are very expensive and require highly qualified staff, thus not every lab can afford them. The multisensor systems can be considered as a potential alternative. They can become a simple and effective tool for the clinical analysis of urine. The main idea of this approach is in employment of chemical sensor array with high cross-sensitivity in a complex liquid media and subsequent processing of the obtained analytical signal by means of multivariate data processing techniques.

In this experiment over one hundred urine samples from patients of Urolithiasis Laboratory and healthy people were collected. All samples were analyzed by CE as referent method [1]. The content of ions (calcium, magnesium, sodium, potassium, ammonium, chloride, sulfate, phosphate, oxalate, citrate, urate, creatinine), pH-level and density of urine were evaluated. These parameters can indicate a possible stone formation. All urine samples were also analyzed with potentiometric multisensor system. Various chemometric approaches were implemented to relate the data from electrochemical measurements with the reference data from CE and available medical diagnoses for patients.

Principal component analysis (PCA) and decision tree analysis (DTA) were applied for classification of samples (ill or healthy). Canonical correlation analysis (CCA) was used for observation of common information shared between two data sets on the same set of samples. The results are presented as similarity maps allowing for a visual assessment of common variance structure shared between the data sets. The quantification of particular urine components with multisensor system was the main objective of this work. PLS regression models were constructed with reference data from biochemical laboratory. It was found, that potentiometric multisensor system can determine concentrations of certain urine components with relative errors in the range 5-10%. These results are quite promising for further development of a simple and inexpensive diagnostic tool for urolithiasis. The experimental details along with the most interesting results of this study will be given in the presentation.

1. Sidorova A.A., Grigoriev A.V., *Journal of Analytical Chemistry*, **67** (5), 478–485 (2012).

## T6 Selecting optimal spectral regions for sensor analysis

*V. Galyanin[1], A. Bogomolov[1,2]*
[1]*Samara State Technical University, Samara, Russia*
[2]*J&M Analytik AG, Essingen, Germany*

Replacement of the wide-range general-purpose spectroscopy by inexpensive sensor systems customized for a specific application is a distinct trend in industrial process analysis today. One of the modern approaches applies light emitting diodes (LEDs) for the sample illumination with subsequent monochromatic detection of diffusely reflected or transmitted light.

The accuracy of such sensor systems greatly depends on the number of LEDs and their arrangement along the spectral region involved into the analysis. The system optimization can be done using full spectral data, and resulting multivariate model can be taken as a benchmark. Specific character of the problem makes existing variable selection techniques (including interval methods) inefficient.

New algorithmic approaches suggested in this work have been specifically intended for the development of sensor analyzers. The optimization has been tested on a previously published dataset [1] including 96 spectra of raw milk in visible and short-wave near infrared region used to predict fat and protein content by means of PLS regression. Optimization of the number and wavelength ranges of the light sources in the developed milk sensor was performed using two different approaches: exhaustive search for a reduced spectral resolution with a predefined grid and genetic variable selection algorithm specially adjusted for the problem at hand. It has been shown that the sensor system can be realized with five LEDs without any loss in accuracy (compared with full-spectral PLS prediction). Increasing LED number to seven noticeably improves the accuracy of fat content prediction and, moreover, reduces the model complexity compared to the spectral data.

Suggested approach can be adapted to a wide range of practical applications of optical spectral analysis.

1. A. Bogomolov, A. Melenteva, *Chemometrics and Intelligent Laboratory Systems*, **126**, 129–139 (2013), http://dx.doi.org/10.1016/j.chemolab.2013.02.006

## T7 Morphology assessment of polymer hydrogels using multivariate analysis of viscoelastic and swelling properties

*Evgeny Karpushkin*[1,2], *Andrey Bogomolov*[3,4]
[1]*Institute of Macromolecular Chemistry AS CR, v. v. i. Heyrovského nám. 2 162 06 Prague 6 Czech Republic*
[2]*Lomonosov Moscow State University, Moscow, Russia*
[3]*J&M Analytik AG, Essingen, Germany*
[4]*Samara State Technical University, Samara, Russia*

Among the various swollen hydrogels intensively used in biomedical applications, porous ones are extremely important. The hydrogel morphology and mechanical properties can be governed by the preparation conditions. In practice, changing the polymer nature is limited by requirements of stability and biocompatibility; external conditions (swelling medium, temperature, and pressure) cannot be chosen arbitrarily as well: a biomaterial should operate at physiological conditions. Therefore, changing concentrations of the solvent or the crosslinker at preparation are in fact the two major ways to alter the polymer hydrogel properties.

In the case of poly(2-hydroxyethyl methacrylate) (polyHEMA), both polymer and monomer are biocompatible. Porous polyHEMA hydrogels are easily prepared by phase separation occurring during polymerization when the amount of diluent exceeds the critical concentration. The mechanical and swelling properties of the resulting material are strongly affected by its porosity.

Direct methods to characterize utilitarian properties of the the polyHEMA support such as cell viability, growth, and proliferation are usually time-consuming and hardly suitable for accurate testing of large specimens series. Therefore, morphology and mechanical properties of the hydrogel materials are typically probed for the preliminary screening. Microscopy of swollen samples is typically used to characterize the porosity; however, light microscopy does not provide sufficient spatial resolution, while scanning electron microscopy of swollen matter is subjected to artifacts. Thus, indirect methods to deduce the morphological features from other data are of particular interest. As mechanical properties should often be determined for other reasons, it is natural to use them for morphology assessment as well.

In this work, we have developed an efficient approach to distinguish macroporous hydrogels from their other morphological forms, based on the PCA of data from conventional analytical techniques, without need for microscopic study.

The numerous variables that can be potentially used in PCA, including both experimental conditions and measured product characteristics, have been screened to exclude redundant, noisy, or low-informative sources. Furthermore, PCA has delivered additional knowledge of both product characterization techniques and the synthesis process.

### T8 Estimation of Corrosion Parameters for Metals in Oxygen Containing Molten Salts by Methods of Interval Analysis

*Kumkov S.I.*[1,3], *Yolshina L.A.*[2], *Yolshina V.A.*[2,3]
[1]*Institute of Mathematics and Mechanics, Ural Branch of RAS, Ekaterinburg, Russia*
[2]*Institute of High-Temperature Electrochemistry, Ural Branch of RAS, Ekaterinburg, Russia*
[3]*Ural Federal University, Ekaterinburg, Russia*

Earlier, corrosion process of aluminum was investigated [1]. The process implemented in a molten eutectic mixture of cesium and sodium chlorides that contained up to 30 wt.% of the sodium nitrate under the argon atmosphere in the temperature interval 793–903 K. Here, the corrosion is stipulated by the free oxygen and sodium nitrite appeared as a result of the thermal decomposition of the sodium nitrate. The corrosion process is followed by creation of the aluminum oxide that partially deposits as the oxide protective passive film on the surface of the metallic aluminum electrode and partially leaves the process in the form of the powder sediment. The coefficients of activity of creation and transformations of the reagents are of the most interest. But estimation of these parameters is strongly hampered by complexity of the process description, short length of measurement samples, and complete absence of any probabilistic characteristics of errors of the measured values. This practically excludes application of the standard statistical methods [2] for estimation of parameters of the corrosion process.

Under such conditions, methods of the interval analysis work efficiently on the basis of elaborated applied procedures and algorithms of estimation [3]. The goal of this work is development of a new approach for processing the mentioned experimental data. The essence of the suggesting method is in the following. The corrosion process and transformations of its components (Al, $O_2$, $Al_2O_3$, $NaNO_3$) are described by a kinetic system of differential equations of the first approximation. The vector of parameter, *i.e.*, the activity coefficients, comprises: $K_{\mathrm{NITRA,Al}}$ of transformation of nitrate in the direct reaction with aluminum, $K_{\mathrm{NTRA,T}}$ of the own nitrate thermal decomposition, and $K_{\mathrm{Al,O}}$ of aluminum transformation into the aluminum oxide by free oxygen. Equation for the oxygen kinetics is built using the principle of mass conservation and coefficients $K_{\mathrm{NTRA,T}}$ and $K_{\mathrm{Al,O}}$. Measuring is implemented at the beginning and at the end of the process. The following values of the process are measured: the mass of the metallic aluminum specimen before and after the corrosion test, its surface area, time of the corrosion interaction, anode potential, density of the anode current, the aluminum ions concentration, general mass of the aluminum (including its quantity in the form of the powder aluminum oxide), concentrations of the nitrate ions converted to the chloride melt according to chemical–analytical data determination. Estimation of the set of the admissible values of parameters is performed in the following way. On the basis of the measurements, the uncertainty intervals of the direct or indirect measurements (both at the beginning and the end of the process) of the components are built using the bounds on the errors of measuring. The sought-for set is a totality only of such parameters values, for which the integral curves of the process components pass through the mentioned intervals at the beginning and end instants of the process.

By variation of the activity coefficients, investigation results allow one to analyze redistribution of the aluminum oxide between the oxide film and the powder sediment.

*The work was supported by the RFBR Grants, nos. 12-01-00537 and 13-01-96055.*

1. Yolshina L.A. Mechanism of Formation of Oxide Nanopowders by Anodic Oxidation of Metals in Molten Salts – Nanomaterials: Properties, Preparation and Processes, ISBN: 978-1-60876-627-7, **NOVA** Publishers, New York, USA, 2010, pp. 255–293.

2. P 40.2.028-2003. The state system for providing unification of measuring. Recommendations of building the reference characteristics. Estimation of errors (uncertainties) of linear reference characteristics by using the least squares method. Official edition.

3. Kumkov S.I., *Rasplavy*, **3**, 86–96 (2010).

## T9 Modeling of fat and protein content in raw milk based on historical spectroscopic data

*Anastasiia Melenteva*[1]*, Andrey Bogomolov*[1,2]
[1]*Samara State Technical University, Samara, Russia*
[2]*J&M Analytik AG, Essingen, Germany*

Visible and adjacent short-wavelength near infrared (VIS/SW-NIR) spectroscopy can be successfully used for the quantification of raw milk [1]. Analysis in this area is economically very attractive due to the use of simple and accessible elements of optical instruments and also enabling in-line monitoring of production processes and products. The spectral analysis in the VIS-region is not widely used for milk, because the measurements are strongly complicated by the multiple light scattering by colloidal fat and protein particles. However, the scattered light contains quantitative information about the milk composition and multivariate analysis is a straightforward technique to extract it. The recently proposed new method is based on multiple light scatter and requires the calibration to be performed on raw spectral data, without a corrective preprocessing [2].

Modeling on samples of natural raw milk is considerably complicated due to the high variability of the milk composition. A precise and robust model should be based on a sufficient number of representative milk samples, covering all possible variability affecting the spectra, and hence, multivariate calibration. The main objective of the presented work was to investigate feasibility of constructing a global model based on historical spectroscopic data collected during a long period.

In this study, the problem of modeling based on historical data has been reviewed and approaches to the creation of global models were proposed. Availability of global models significantly facilitates application of the scatter-based method to laboratory and in-line industrial analysis of fat and protein content in the raw milk.

1. A. Bogomolov, *Food Chem.*, **134**, 412–418 (2012).
2. A. Bogomolov, A. Melenteva, *Chemom. Intell. Lab. Syst.*, **126**, 129–139 (2013).

## T10 Possible Improvements to Multivariate Curve Resolution with Particle Swarm Optimisation: Estimation of Rotation Ambiguity

*Skvortsov A.N.*

*St-Petersburg State Polytechnical University, Biophysics Department, St-Petersburg, Russia*

Multivariate curve resolution (MCR) employs a set of various techniques widely used in modern chemometrics. They include target factor analysis (TFA), resolving factor analysis (RFA), alternating least squares (ALS), and many other. These techniques aim to resolve data matrices into spectra and concentrations of individual chemical components, given a set of physicochemical constraints. Well known issue of these methods is the rotation ambiguity (RA). If the model is not sufficiently constrained, e.g. in the case of non-negativity constrains only; or the data matrix is rank-deficient, the solution may be not unique (ambiguity set). Except for the simple case of 2 components, RA is hard to predict *a priori*. However, in case of RA, a single MCR solution has very little practical value. So good MCR techniques are to numerically evaluate the extent of the ambiguity and report it to the researcher. There are many works devoted to RA in the literature, and several numeric approaches to its solution do exist (MCR-BANDS, MCR-FMIN etc). Unfortunately these approaches lose the elegance and simplicity of the initial methods: they base on complicated nonlinear algorithms, or produce huge amount of data. The latter data are hard to visualize, and they themselves need non-trivial analysis.

One of the possible options, which gains popularity, is to combine MCR with global search algorithms or evolution algorithms. The power of combining MCR with particle swarm optimization (PSO) has been recently shown (MCR-PSO; *e.g.* Parastar et al., Anal Chim Acta. 2013. 772:16). PSO builds a set of solutions (particles) which are updated by some simple search algorithm, typically including stochastic terms. This inner-level algorithm does not need to have good convergence *per se*. The convergence of PSO is provided by interaction of the particles (swarm intelligence), which is typically some simple potential function. In common PSO, interaction is optimized to find the minimum of generic target function (which is sum of squares of residues in case of MCR). However, we believe that the inner-level algorithm and interaction in PSO may be tuned specifically for solving MCR problems.

The present work focused on the ability of PSO-like algorithms to find ambiguity set for MCR. PSO was modified by adding small penalty function for similarity of solutions (repulsion potential, as it corresponds to repulsion between the particles). When several particles are in the ambiguity set and have the same values of target function, the repulsion potential is the sole force that controls the particles. It repels the particles from each other towards the borders of the ambiguity set. When the procedure finishes, it should in theory produce the set of solutions, which are the least similar to each other (in terms of the selected repulsion potential). The solutions are biased, but the bias can be made small by carefully selecting the repulsion potential.

Various types of repulsion potentials and weight coefficients were tested on common optimization tasks, simulated multivariate data, and real MCR problems (fluorescence spectra). Different inner-level algorithms were tested (stochastic search, ALS, RFA). The algorithm had to account for permutation and scale ambiguities of MCR, which incorrectly reduced penalty for similarity. For two-component data matrices the results were compared to the analytical boundaries of the ambiguity set. The results of the work have shown a fairly good ability of PSO with repulsion potential to span the ambiguity set of MCR solutions. The significant improvement of convergence of PSO over the respective parent method was confirmed. The developed approach is compatible with all types of constrains found in MCR problems including model-based MCR. It also retains the simplicity of the parent algorithms. The only cost for it is the computational expense for optimizing multiple particles, which is partially ameliorated by easy parallelization of PSO.

## T11 Identifying Bioactive Signature in Natural Products through Chromatographic Fingerprint

*Qing-Song Xu*
*School of Mathematics and Statistics, Central South University, Changsha, P.R China*

Bioactive component identification is a crucial issue in search for new drug leads. We provide a new strategy based on Sure Independence Screening (SIS) and Backward Variable Selection for PLS (BVSPLS), which is termed as the SIS-BVS-PLS method. SIS-BVS-PLS is not only able to find out the chief bioactive components, but also able to judge how many components should be there responsible for the total bioactivity. The method is totally "data-driven" with no need for prior knowledge about the unknown mixture analyzed, therefore especially suitable for effect-directed work like bioassay-guided fractionation. Two data sets, a synthetic mixture system of twelve components and a suite of *Radix Puerariae Lobatae* extracts samples, are used to test the identification ability of the SIS-BVS-PLS method.

*This work is financially supported by the National Nature Science Foundation Committee of P.R. China (Grants No11271374).*

1. Chau FT, Chan HY, Cheung CY, Xu CJ, Liang Y, Kvalheim OM, *Analytical chemistry*, **81** (17), 7217–7225 (2009).

2. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr KM, Kvalheim OM, *Analytical chemistry*, **81** (7), 2581–2590 (2009).

3. Fan J, Lv J, *Journal of the Royal Statistical Society*: *Series B*, **70** (5), 849–911 (2008).

4. Fan J, Song R, *The Annals of Statistics*, **38** (6), 3567–3604 (2010).

5. Xu Q-S, Liang Y-Z, Shen H-L, *Journal of Chemometrics*, **15** (3), 135–148 (2001).

6. Li H-D, Liang Y-Z, Xu Q-S, Cao D-S, *Journal of Chemometrics*, **24** (7–8), 418–423 (2010)

7. FernándezJA, Abbas O, Baeten V, Dardenne P, *Analytica Chimica Acta*, **642** (1–2), 89–93 (2009).

8. Chan H-y, PhD dissertation. Hong Kong Polytechnic University, 2010.

9. Jiang R-W, Lau K-M, Lam H-M, Yam W-S, Leung L-K, Choi K-L, Waye MMY, Mak TCW, Woo K-S, Fung K-P, *Journal of Ethnopharmacology*, **96** (1–2), 133–138 (2005).

10. Chen S-B, Liu H-P, Tian R-T, Yang D-J, Chen S-L, Xu H-X, Chan ASC, Xie P-S, *Journal of Chromatography A*, **1121** (1), 114–119 (2006).

## T12 Macrolevel Study of Russian Science Citation Index using PCA for Interval-Valued Data

*Petr A. Ledomskiy, Sergei I. Zhilin*
*Altai State University, Barnaul, Russia*

Russian Science Citation Index (RSCI) [1] is a bibliographic database of scientific publications in Russian founded in 2005. It accumulates more than 4.7 million publications of Russian authors, as well as information about citing these publications from more than 4000 Russian journals.

Performing an exploratory analysis of such a large dataset is potentially problematic especially when using visual methods. Scatter plots or PCA scores plots of original dataset allows one to learn a little of the innate structure of the data because the plots become visually poor when the number of observation grows up to hundreds and thousands.

One of fruitful ideas is to aggregate the data in some meaningful way to reduce them to a more manageable size for analysis. How data might be aggregated depends on the questions driving the analysis. New categorical observations have internal variations along with the familiar (between observations) variation of classical data. That is why classical methods cannot be directly applied to objects obtained by aggregation and need for adaptation. Apparently, the analysis of categorical observations instead of individual objects hides some data details and peculiarities but enables one to discover some important patterns and trends on the macro level, i.e., level of objects categories.

Interval-valued observations naturally arise as the simplest form of aggregating a larger data set. An interval box representing a category of data points can be easily constructed as a Cartesian product of variables' ranges for data points in the category. To make the analysis robust interquartile intervals can be used to represent a category on each variable.

Several variants of PCA for interval-valued data are known in the framework of the mentioned approach: Centers PCA (CPCA), Vertices PCA (VPCA) and Complete-information PCA (CIPCA) [2]. CPCA, VPCA and CIPCA share common PCA algorithm and differ only in the manner of introducing the covariance matrix for analyzed interval data matrix. After computation of the covariance matrix the algorithm assumes orthogonal projection of original interval observations to the subspace spanned by eigenvectors of the covariance matrix. Each of the interval-valued categorical observations is featured by interval-valued PCs and can be displayed as a rectangle in the 2D score plot. The center location of each rectangle represents the averaged value of the concerning interval-valued PC score, while the size gives a measure of dispersion associated with PC scores of data points from the concerned category.

We employ CPCA, VPCA and CIPCA to discover patterns and trends in the bibliometric data for journals indexed by RSCI. Journals are described by such variables as number of published papers, number of references per article, impact factor, total citation, etc. We studied journals grouped into categories by a scientific discipline on different levels of aggregation, by the "quality" of journals and some other criteria. The results of the analysis approve that in some cases it may be more beneficial to use interval-valued data instead of the original numerical data, and to achieve a clearer and more meaningful understanding of huge-volume and complex-structured data on a macro level.

1. Russian Science Citation Index — http://elibrary.ru/ project_risc.asp.
2. Wang H., Guang R., Wu J., *Neurocomputing*, **86**, 158–169 (2012).

## T13 The results of solving tasks of medical diagnosis for dental patients on the basis of chemometrics methods

*Delakova Ekaterina*
*The First Saint Petersburg State Medical University, Saint Petersburg, Russia*

Strategically important tasks for modern science are improving the quality of medical diagnosis, using of new information technologies in the work with the results of biomedical and clinical research, choosing optimal methods for the analysis of multivariate and various medical data.

The data of biomedical researches are the multidimensional heterogeneous attributes, which have different qualitative and quantitative characteristics, varying in different ranges. These include laboratory patient tests, various indices of somatic health status or data device-computer systems. The results of solving tasks of medical diagnosis (classification and clustering tasks) based on chemometrics methods and comparative analysis of their effectiveness in relation to other information technologies are offered in this report.

The input data are the results of dental examination of patients of different age on morphofunctional state of the oral mucosa and lips (the status of each patient is characterized by about 60 parameters). These data were obtained on the basis of Pavlov First Saint Petersburg State Medical University in which researches in the field of evaluation of the effectiveness of methods of multivariate medical information data analysis are being conducted. The central chemometrics method - the principal component analysis (PCA) is used for studying the data structures. PCA - modeling revealed scorecard with effective dimensionality containing 7 principal components. Also we have identified the parameter which have the greatest influence on the separation of data to classes.

The Classification problem –is referring the patient to one of the pre-formed classes for baseline survey data. This task is solved with the help of formal analogies independent modeling classes method. The teaching model was build and this model proved its efficiency in testing: when using three principal components, the key measure of model quality (residual dispersion) was 2%. The Recognition procedure is carried out, the condition of the patients is rightly assigned to the Class " presence of the disease " modeling error  was 4%. The comparative analysis of the effectiveness of classification models is carried out. As an alternative models the neural network technology, the algorithm KNN and the Bayes classifier were taken.

The method of cross-validation and methodology ROC-curves designing were used to assess the quality of each classification model. According to the results of cross-validation conventional indicators of statistical evidence [1] were calculated. There were sensitivity, specificity, likelihood ratio and predictive values for the presence or absence of disease. After comparing the obtained results SIMCA method showed the highest efficiency: sensitivity value (Se) was 90%, specificity value  (Sp) was 92%, predictive value (PPV, NPV) of 97% and 94%, respectively. Index AUC [2] was almost 90%, which also corresponds to a high quality model.

The obtained results showed high efficiency of chemometrics methods for solving problems of analysis, classification and clustering for the work with the data of biomedical researches.

1. Olive Jean Dunn, Virginia A. Clark: BASIC STATISTICS A Primer for the Biomedical Sciences // 2000 A JOHN WILEY &SONS, INC., PUBLICATION.

2. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers // 2004 Kluwer Academic Publishers.

## T14 Resolution of Overlapped Peaks in Stripping Voltammetry with Stepwise Mathematical Resolution Method

*Romanenko S.V., Shekhovtsova N.S., Cherednik E.A.*
*Tomsk polytechnic university, Department of ecology and human safety, Tomsk, Russia*

Overlapping of analytical signals is a common problem in many methods of analytical chemistry, and voltammetry in particular.

The curve fitting method is the one commonly used in resolution of overlapped peaks. But it also has a major drawback in case of high number of overlapped peaks, or in case of its complex shape. The curve fitting in the given conditions includes optimization of great number of parameters. Therefore it requires way too much of calculation and leads increase of uncertainty of results.

A new stepwise mathematical resolution method (SMRM) can be used to overcome the limitations mentioned above.

The basic idea of the SMRM approach is a one-by-one mathematical removal of a signal of a particular analyzed component from a complex (mixed) signal. The key to successful implementation of this method is the choice of an optimization criterion. Such a special criterion estimates the distortion of signal remainder in a part of the overlapping area. To obtain correct results by SMRM it is necessary for resolving signal to coincide with shape, height, position of signal removed from complex curve.

Effectiveness of stepwise mathematical resolution method was verified by resolution of simulated overlapping signals and signals of model chemical systems, obtained in stripping voltammetry.

The application of SMRM was shown by resolution of four-component voltammetric curves of Pt-Bi binary precipitates in stripping voltammetry. Besides proposed approach has been applied to X-ray diffraction overlapping signals.

## P1 The study of voltammetric behavior of the electrode/engine oil/marker system using PCA and PLS

*Bikmeev D.M., Sidelnikov A.V., Kudasheva F.Kh., Maistrenko V.N.*
*Bashkir State University, Ufa, Russia*

The areas of study of voltammetric and potentiometric electronic tongues are restricted mainly by electrochemically active materials, whereas non-electroactive components usually studied using the methods of impedance spectroscopy. At the same time, the multicomponent solutions consisting of electroactive and non-electroactive components (technical liquids, food industry, pharmaceuticals, environmental objects, etc.) aren't less relevant for voltammetric study. The voltammetric measurements in such systems can be carried out by use of the electrodes which are include the test substances. In this case, the test substances will influence to the magnitude of the detected signal and the shape of a voltammogram. However, physico-chemical characteristics and properties of these electrodes have been poorly studied.

Voltammetric study and applying of chemometric methods PCA and PLS in electronic tongue on the basis of carbon-paste electrodes containing electroactive and non-electroactive components for multicomponent solutions identification have been presented. The values of the peaks current constants on voltammograms of nitrocompounds reduction and the dependence of maximum of peaks current from scan rate, from accumulation time, from the polarizing voltage shape have been calculated. Using PLS the relationship between residual and Faraday currents on voltammograms of nitrocompounds reduction, physico-chemical characteristics of motor oils and IR spectroscopy data has been established.

*The reported study was partially supported by RFBR, research project No. 13-03-00471 A.*

## P2 Efficiency of «reactor-regenerator» system joint work in synthetic detergents manufacturing enhancing use of mathematical modeling methods

*I.O. Dolganova, E.N. Ivashkina, E.D. Ivanchina*
*Tomsk polytechnic university, Institute of natural resources, Russia*

The work carried out to enhance the efficiency of «reactor-regenerator» system joint work in linear alkylbenzenes manufacturing. Dependence of HF-catalyst activity on reactor and column regimes was studied. The ways of benzene alkylation rector and catalyst regenerator column joint work optimization were determined. They can be in

HF-catalyst optimal activity achievement with further system to the steady state transfer, as well as in date of probable failure predicting with possibility of column drainage to prevent it.

Over the last years there has been a huge success in production of synthetic detergents on the base of linear alkylbenzenes (LAB) as environmentally friendly technologies.

One of the ways of process efficiency enhancing without reconstruction of existing equipment can be a organization of technology-related blocks smooth operation, in this case, joint work of benzene with olefins alkylation reactor and HF- catalyst regeneration column.

Therefore the aim of this research is to find out the ways of LAB yield enhancing and «reactor-regenerator» system joint work stabilization.

It was found out, that

- value of optimum HF-catalyst activity, at which equilibrium of heavy aromatics (HAR) formation reaction and maximum selectivity for the desired product is reached, depends on alkylation reactor feedstock composition. It is increased from 0.44 to 0.6 rel. u while ratio of hydrocarbons ($C_{10}$ + $C_{11}$) / ($C_{12}$ + $C_{13}$) in alkylation reactor feed is increased from 0.676 to 1.033;
- Optimum flow quantity of HF from alkylation reactor to regenerator varies in the range from 3.6 to 4.7 m3/hr while flow rate of diolefines into the alkylation reactor increasing from 56.25 to 73.9 kg / hr.
- The effect of optimal mode of «reactor-regenerator» system joint work maintaining consists in increasing the revenue of the 36 million rubles. per year. It is provided by LAB yield and period of regenerator column stable operation increasing and also by ability to forecast a possible abnormality with use of developed computer modeling system.

*The study was supported by the Ministry of Education and Science of Russian Federation, the agreement 14.B37.21.0825 «Environmental properties of petrochemical products produced by hydrocarbons $C_{12}$-$C_{27}$ in industrial reactors processing forecasting with use of experimental and quantum chemical methods».*

## P3 Fiber optical ATR-IR spectroscopic analyzer of oil-contaminated soil

*V.V. Ermakov*[1]*, A.O. Guryanova*[1]*, A.Yu. Bogomolov*[1]*, D.E. Bykov*[1]*, T.V. Sakhorova*[2]*, V.G. Artyushenko*[2]

[1]*Samara State Technical University, Samara, Russia*

[2]*Prokhorov General Physics Institute, Russian Academy of Sciences*

Oil-contaminated soil and others heterophase wastes have a negative impact on the environment. Nowadays there is an increasing need in the environmental monitoring, rapid response to emergency situations and field control of contaminants assimilation.

The basis of the present analyzer is an infrared spectrometer. It has been equipped with a fiber optical probe that acquires IR absorption spectra of the soil through an attenuated total reflectance (ATR) measurement head. A portable infrared spectrometer is connected to the probe head with a flexible IR fiber cable. The probe crystal and fiber material were selected taking their transmission ranges into account.

Spectra of oil-contaminated soils have a group of characteristic peaks of the hydrocarbons between 2925 and 2860 cm$^{-1}$ that exhibits an evident correlation with oil content. At the same time, even the spectra of non-contaminated soils experience the contribution of organic substances of natural origin. A broad and intensive peak in the region 3500–3300 cm$^{-1}$ is related to the absorption of water significantly interfering with the signals of interest. The "fingerprint" region also comprises characteristic peaks of the hydrocarbon groups. Their intensity is lower and the interpretation difficult because of the signal overlay.

PLS regression model for the quantitative determination of the oil products in soils has been built based on both design laboratory and field samples. The experimental and modelling issues have been considered in the present work. The accuracyimprovement potentials are outlined.

The immersion probe can be used to control the composition of oil-contaminated soils in the field conditions without or with a minimal sample preparation. The method is also applicable under conditions of a bioremediation processes.

## P4 Calibration Transfer For Electronic Tongue

*M.M. Khaydukova*[1]*, D.O. Kirsanov*[1]*, V.V. Vietoris*[2]*, A. Legin*[1]

[1]*Laboratory of chemical sensors, Chemistry department, St. Petersburg State University, Mendeleev Center, St. Petersburg, Russia*

[2]*Faculty of biotechnology and food sciences, Slovak university of agriculture, Nitra, Slovakia*

There is a distinct trend in modern analytical chemistry towards development and application of fast, simple and inexpensive analytical methods. Despite the lack of sensitivity and selectivity in certain cases these methods are still able to produce reliable analytical results which are adoptable for industrial practice. One of the promising instruments for application in industry is potentiometric multisensor system– "electronic tongue". Such system consists of an array of specially designed non-specific chemical sensors. Sensor response is determined by the qualitative and quantitative composition of a sample and is recorded by an appropriate data acquisition system. Being non-selective in nature the response of this array however contains information on multiple species in samples and this information can be effectively extracted by multivariate data analysis techniques. Such measurement protocol allows for solving the problems of sample classification and numerical prediction of particular parameters.

As any other analytical method multisensor system requires calibration before measuring samples with unknown reference characteristics. Calibration is a two-stage procedure: 1) measurement of sample set with known reference parameters; 2) building a mathematic model. Once being established this model allows for prediction of the parameters of interest in new unknown samples. The problems with further application of such models are associated with necessity of periodical sensor changing in the array, sensor drift and untypical samples presented for prediction. All these issues require recalibration of the multisensor system which is a long and tedious process. This circumstance slows down the integration of these systems in industrial cycles. An obvious way to circumvent this limitation is to establish the mathematical methods for calibration transfer from one sensor array to another. While in the field of spectroscopic methods this task is already solved and several protocols such as e.g. PDS are suggested [1], in the field of multisensor systems this important issue was not intensively addressed.

This research is devoted to an attempt of transfer PLS regression model between two physically different sensor arrays having the same set of sensors. Details on samples, transfer protocols and results will be provided in the presentation.

1. Y.D. Wang, D.J. Veltkamp, B.R. Kowalski, *Anal. Chem.*, **63**, 2750–2756 (1991).

## P5 Usage of a cubic polynomial Bernstein-Bezier for describing a curvilinear baseline in inversion voltammetry

*Kuznetsov V.V., Romanenko S.V.*
*Tomsk Polytechnic University, Russia*

In inversion voltammetry, the approximation of the baseline under the peak on the voltammogram is an important task. Since the calculation error of concentration depends on correctness of the baseline position.

For some chemical elements (for example, Cd, Pb), a linear approximation of the baseline will be suffice. But in most cases, curve fitting is required (eg Zn, Cu, Se), analytical peak is usually located in concave or convex areas of voltammogram.

Main purpose of the presented work is to develop auniversal algorithm to describe the curved baseline under the analytical signal with form close to the peak.

The authors proposed a cubic polynomial Bernstein-Bezier baseline approximation. The polynomial consist of four points: first and last points of the polynomial determine the position of on the plane (nodes), and the second and third points define the curvature of a polynomial (support).

Nodal points of the polynomial lie on a voltammogram and define the boundaries of the analytical peak (the markup is provided by qualified laboratorian, or special mathematical algorithms).

Support points bend the line under the peak of the polynomial. The farther the support point from the nodal point is located, the stronger the whole curve will be distorted. Support points lie on a tangents of the voltammogram with formation of the nodal points of the polynomial, Determining the behavior of the tangent voltammograms outside of peak.

The abscissa for the support points is equal to the abscissa of the inflection points of the peak (the most stable point). Position of the support points can be adjusted depending on the distance between the nodal points and the center of the peak (the farther from the center the nodal point is located, the less influence of behavior voltammograms, outside of peak is).

Algorithm takes into account the behavior of nodes and support points of the polynomial with exponential and logarithmic dependence of the baseline under the peak. With the exponential dependence, the nodal points visually are shifted inward peak, while the logarithmic dependence, support points underreport the signal height (due to the rapid growth of the signal on the rising branch of the peak).

The algorithm was tested on such elements as Zn, Cu, Se, obtained by the method of "introduced-found" on model solutions using voltammetric analyzer TA-07, produced by Ltd. "Tehnoanalit" (a total of more than 200 experiments). The average error of measurement of the amount of substance amounted up to 9%, but not more than 20%. Marking of the peak was carried out in automatic mode.

## P6 Recognition of the emission spectra of laser-induced plasma by their correlation with modeled one

*T.A Labutin, S.M Zaytsev, A.M Popov, N.B Zorov*
*Lomonosov Moscow State University, Department of Chemistry, Moscow, Russia*

Qualitative analysis with the use of atomic emission spectroscopy is essentially a task of recognition of analyzed sample spectra. Spectra recognition is usually performed with step-by-step identification of all elements, comparing their strongest lines (fingerprint) with the experimental spectrum. The identification of emission lines in the Laser-Induced Breakdown Spectrometry (LIBS) is a very difficult and time-consuming task due to huge heterogeneity of laser plasma and strong matrix effect. Therefore, excitation conditions in mainly depend on a sample matrix, wavelength of the laser beam, external pressure, and others. For example, we cannot clearly assign the weak emission line in a spectrum of steels with the strong line of minor component or with the weak line of matrix elements (such as Fe, Cr, Ni). The elemental qualitative analysis of environmental objects, such as soils or rocks is even more difficult procedure. The problem is to assign each of the peaks with specific emission line of a certain element due to the strong superposition of emission lines in UV-VIS spectral range.

We have applied another algorithm to identify emission lines in LIBS. The algorithm implemented by three parts: simulation of the set of spectra corresponding to different temperature (T) and electron density (Ne), searching the best correlated pair of model spectrum and experimental one and attributing the peaks with certain lines. In order to construct the model spectra we used the parameters of atomic and ionic lines, levels, the mechanisms of the broadening of spectral lines and the selected parameters of the spectrograph. We tried different weighting procedures for attribution of the peak with the certain line, involving the values of transition probabilities and closeness to the center of observed peak. Suggested approaches provide a fast and accurate recognition of the LIBS spectra of soils and steels.

*This work was financially supported by Russian Foundation for Basic Research (grants No 13-03-13104 ofi_m_RZhD).*

## P7 Comparison of various chemometric techniques for quantitative analysis in complex industrial solutions

*Ekaterina Oleneva*[1], *Dmitry Kirsanov*[1], *Marina Agafonova-Moroz*[2], *Alexander Lumpov*[2], *Vasily Babain*[2], *Andrey Legin*[1]

[1]*Chemistry Department, St. Petersburg State University, St. Petersburg, Russia*

[2]*Khlopin Radium Institute, St. Petersburg, Russia*

Optical spectroscopy is widely known to be a very attractive tool for on-line process control in industry. Typically it is employed in combination with multivariate regression techniques such as PLS for quantification of various components in complex industrial mixtures. There is a growing interest for development of on-line control procedures for spent nuclear fuel reprocessing, since at the moment this industrial field lacks express and simple techniques allowing for rapid quantification of several key elements in process streams. There are several important components to analyze in such solutions, uranium, plutonium, neptunium and nitric acid are most important. While U, Pu and Np can be easily quantified with PLS regression from UV-Vis spectra [1], nitric acid has no its own distinct band in this spectral region. HNO3 can be quantified from IR measurements; however this will require addition of an extra spectroscopic probe to the monitoring system and additional costs.

It is known that nitric acid content has distinct influence on the shape of uranium bands in UV-Vis spectra [2]. It seems worth of trying to quantify nitric acid from these signals without involvement of additional IR methods. In this study we compare several different regression techniques (like PLS, SVM, etc) for quantification of nitric acid in complex mixtures with uranium and plutonium simulating typical composition of PUREX solutions from UV-Vis spectroscopic measurements.

1. Kirsanov D., Babain V., Agafonova-Moroz M., Lumpov A., Legin A., *Radiochim. Acta*, **100**, 185 (2012).

2. Warburton J., Smith N., Czerwinski K., *Sep. Sci. Techn*., **45**, 1763 (2010).

## P8 Optimum discretization of velocity scale in Mossbauer spectrum measurements

*V.V. Panchuk*[1,2]*, V.G.Semenov*[1,2]*, A.A. Goydenko*[1]*, S.M.Irkaev*[2]

[1]*Department of Chemistry, St. Petersburg State University, St. Petersburg, Russia*

[2]*Institute for Analytical Instrumentation, St. Petersburg, Russia*

Discrete spectral representation is widely used in modern analytical spectroscopy. Signal digitization and choice of discretization step Δ are important for this procedure. For small values of Δ the number of points in the observed spectrum will be high and therefore the accuracy of spectral function reproduction will be high. On the other hand a reduction of Δ will decrease spectra accumulation time to obtain necessary statistical accuracy. The accuracy is the value equal to the ratio of the signal amplitude to the average level of noise. For large values of the discretization step Δ number of channels is reduced, consequently, accuracy of reproduction of the spectral function is reduced due to the strong averaging of this function within the interval Δ. Thus, the optimal discretization provides a representation of the desired spectral function with the required precision and minimal number of channels. In this case, all points of the spectrum are significant to restore the original spectral function and the time required for their accumulation is minimal. Optimal sampling step is estimated by restoration error of the spectral function.

The immediate aim of the Mossbauer experiment is the registration of gamma rays intensity connected with the processes of resonance absorption or scattering in a certain energy range. Doppler modulation is used for energy scanning of gamma rays emitted by the source. Modulation is performed using a mechanical motion of the source relative to investigated sample in the required velocity range.

This work describes the general approaches for selection of the optimum channel widths at discretization of the velocity scale. The first approach is based on the analysis of the Kotel'nikov-Shannon's frequency criterion, the second one - on analysis of the errors which arise in the case of a differential representation of the spectrum. In this case, the analysis is conducted in two phases: the first phase is considering the ideal shape distortion of the spectrum due to the sampling velocity of the scale and the second takes into account the influence of the statistical spread in the spectrum. At the second stage, the quality of the spectrum (S/N) is estimated. As the result of these procedures one can find optimum spectral channel width providing for short accumulation time and statistical accuracy.

*The reported study was partially supported by RFBR, research project No. 13-08-00623-a.*

## P9 Software package for processing and modeling of the nuclear gamma resonance spectra

*V.V. Panchuk*[1,2]*, V.G. Semenov*[1,2]*, S.M. Irkaev*[1]
[1]*Department of Chemistry, St. Petersburg State University, St. Petersburg, Russia*
[2]*Institute for Analytical Instrumentation, St. Petersburg, Russia*

Nuclear gamma resonance spectroscopy (Mossbauer spectroscopy) is widely used for analytical purposes. The existing software packages for experimental spectra processing are usually focused on extracting qualitative information, leaving aside the problem of quantitative analysis, as well as preprocessing procedures (smoothing, filtering, deconvolution, etc.). The proposed software package includes a wide range of options and allows for both qualitative and quantitative analysis of samples.

The package structure is designed so that the user at the first stage can carry out pre-processing of the experimental data in order to improve the signal / noise ratio as well as to remove the outlier points. Application of the deconvolution removes the impact of apparatus function. Mathematical processing of experimental spectra is based on a linearized least-squares method. The form of spectral lines is assumed to be Lorentzian, however, it can be in the form described by the Voigt function, which better reflects the real situation. The program provides an option when the spectral lines form a continuous sequence (for example, the objects in the glassy state), leading to incorrect problem. In this case, the package allows you to find the distribution function of required parameters. Two-dimensional arrays can be analyzed to compare the experimental spectra with the spectra from the database.

The developed software package can be used not only in research laboratories and analytical centers, but in the educational laboratories using functions to create model spectra.

*The reported study was partially supported by RFBR, research project No. 13-08-00623-a.*

## P10 XRF with chemometric data processing for analysis of iron oxidation state

*V.V. Panchuk*[1,2]*, V.G. Semenov*[1,2]*, N.O. Rabdano*[1]*, A.A. Goydenko*[1]*, S.M. Irkaev*[2]
[1]*Department of Chemistry, St. Petersburg State University, St. Petersburg, Russia*
[2]*Institute for Analytical Instrumentation, St. Petersburg, Russia*

XRF (X-ray fluorescence) analysis is a widely employed spectroscopic method for elemental analysis which allows for determination of quantitative content for elements from beryllium to uranium. Iron is involved in many geological formations (hematite, magnetite, hydrogoethite etc.) and it is important for various construction materials. It is not only quantitative content which is important when analyzing such objects, but also the oxidation state of iron. To solve the last problem one can use the phase analysis methods. On the other hand the energy (or wavelength) of some X-ray fluorescence lines depends on the energy of the atom valence levels. X-ray fluorescence lines shift can be used to determine oxidation state and this was shown in numerous studies. However, the practical implementation of this approach is associated with certain difficulties mainly due to insufficient energy resolution of spectrometers and uncertain interpretation of results. The purpose of this work is to determine oxidation states of the iron atoms from X-ray fluorescence lines shifts using logistic regression, PCA and PLS-DA.

Solid samples containing iron with different oxidation states were analyzed: a-Fe, $Fe_2O_3$, $K_3[Fe(CN)_6]$, $K_4[Fe(CN)_6]\cdot 3H_2O$, $FeCl_2\cdot 4H_2O$, $FeCl_3\cdot 6H_2O$. The L - series lines of iron ($FeL_{\alpha 1\text{-}2}$, $FeL\beta_{1\text{-}3}$) were employed as analytical signals for determination of the iron atoms oxidation states. The reason for this choice is the energy of these lines which mostly depends on the energy of 3d electron levels involved in formation of chemical bonds.

Spectra of the samples were measured by wavelength dispersive X-ray fluorescence spectrometer with high resolution (Shimadzu XRF-1800). Measurement conditions were: TAP crystal with first order of reflection; X-ray tube with Rh anode, voltage 90 kV and current 45 mA. These conditions allowed for the largest changes in X-ray lines, good reproducibility and reasonable accumulation time.

The spectra were decomposed into single lines followed by logistic regression, PCA and PLS-DA. It was shown that all three methods can solve the task of oxidation state analysis.

## P11 Methodological aspects of application of peak evaluation techniques

*S.V. Romanenko, E.V. Larionova*
*Tomsk Polytechnic University, Russia*

Most frequently analytical signal has a peak and sigmoid shape. There are a few ways of the characteristic of the properties of signals. The method of the statistical moments of distribution is used for peak evaluation. Contour and tangent methods are applied for peak and sigmoid evaluation. Early systematic comparison of various approaches to peak evaluation has been carried out. In this work some methodological aspects of application of evaluation techniques for discrete signals will be considered. The ways of parameters calculation by these methods are described. Influence of a degree of digitization, noise level, a base line variation, and signals overlapping on stability of parameters calculation of all studied ways is investigated.

It is shown, that the error of a base line subtraction can essentially influence on values of the statistical moments and introduce the errors into calculation of them, especially, at the small analyte concentration. The main problem of application of the contour and tangent methods is connected with the necessity of derivatives application at the calculation of parameters. That is why the full suppression of noise is required. Main advantage of the tangent method is that the error of estimation of inflection points insignificantly affects on calculation of other parameters of triangular frame. Therefore, application of the tangent method at an estimation of size, for example, overlapped signals is especially justified.

## P12 The identification of motor oils using differential voltammetry and SIMCA

*Sidelnikov A.V., Bikmeev D.M., Kudasheva F.Kh., Maistrenko V.N.*
*Bashkir State University, Ufa, Russia*

In this work the results of research in the field of voltammetric electronic tongues for identification of the viscous organic solutions have been generalized. Gasoline engine oils of different brands have been studied. The variation of conditions of voltammograms registration (potential scan rate, potential range and polarizing voltage form) and the nature of the electroactive markers with applying the modern methods of chemometric data processing can extend the capabilities of analysis and the identification rapidness. The results of the research of voltammetric behavior of electroactive markers in the terms of differential voltammetry at the tubular and carbon-paste electrodes have been shown. The principal approaches for pattern recognition of engine oils using chemometric methods have been presented. The developed voltammetric electronic tongue has been allowed to expand the range of problems that can be solved by voltammetry – the identification of multicomponent mixtures as electroactive and non-electroactive substances without preliminary sample preparation and without using of high-cost and time-consuming analytical methods.

*The reported study was partially supported by RFBR, research project No. 13-03-00471 A.*

## P13 Soil classification for forensic purpose by using scanning electron microscopy with X-ray analyzer, color analysis and chemometrics

*Talavera I., Madrazo I.*
*Advanced Technology Applications Center Havana Cuba*

Soil forensic evidence samples are very difficult to process, due to the greater number of general and individuals characteristics presents at the same time and the low discriminative information that the surface layer has , and this part is the one that is collected in crime scene or in shoes and other objects belonging to a suspect.

The main purpose of this paper is demonstrating the feasibility and benefits of the Scanning Electron Microscopy with X-ray Analyzer coupled (SEM-EDS) for the analysis and chemical characterization of samples of soils of one municipality of Havana City, and the construction of an automatic classification model for soil samples discrimination, in correlation with the present-day genetic classification existing of this place, in order to predict the origin and soil type in unknown samples related with a case using Chemometrics tools for the multivariate processing of the data.

The data from the SEM-EDS analysis was submitted to an exploratory analysis using the Hierarchical Clusters Analysis, demonstrating the feasibility and differentiation of 6 types of soils of the 9 presents. Elements Mg, Si, K, Ca, Ti, Fe and Al were the most significant elements in the discrimination. From these results a model for the automatic classification for 6 types of soil samples was constructed using a Support Vector Machines (SVM) classifier. The model was validated with external samples not present in the training set with 97 % of efficacy. Another model built for himself from the data once SEM-EDS was gotten and the analysis of the color with the charts of Munsell and the parameters of the color of the Editor of images for Windows

Adobe Photoshop 7,0.( R, G, B, C, M, And, K, H, S, B1, L, a, b ), getting out a differentiation in 9 types of ground for PCA and for HCA, constructing a model of automatic classification for 9 classes remaining how best classifier the SVM. The model was validated with external samples with 100 % of efficacy.

## P14 Resolution of fully overlapping peaks in hyphenated chromatography

*Yu.P. Turov*
*Surgut State University, Surgut, Russia*

In hyphenated chromatography the concentration profiles of elution products are simultaneously analyzed by spectroscopy methods. Using a priori knowledge of the most general kind one may resolve hyphenated spectra of mixtures in terms of those of their components.

There are some closely related 'model free' methods under the names of evolving factor analysis, window factor analysis, and heuristic evolving latent projections.

R. Manne in his purely theoretical paper analyzed the limits of resolution methods for hyphenated chromatography and proved: «If for every interferent the concentration window of the analyte has a subwindow where the interferent is absent, then it is possible to calculate the spectrum of the analyte (Theorem 2)» [1].

Nevertheless the present paper will present some results of resolution of fully overlapping peaks in GC/MS data. Data matrix was artificially generated by multiplication of mass spectra model compounds (benzene and cyclohexane) by two component elution profiles: ***a*** – from R. Manne paper [1], and ***b*** – for model Gaussian and bimodal distribution profiles with coincident maximum (average of distribution).

Ad hoc normalizing- and boundary condition assisted iterative target transformation procedure of abstract factor analysis (analogues to non-negative iterative factor analysis) succeeded in resolving the problem in full – to extract both profiles and spectra from ***a-*** and ***b-*** hyphenated chromatography data sets.

Normalizing- and boundary conditions are the source of additional information in order to obtain unique resolution.

1. R. Manne, *Chemom. and Intell. Labor. Syst.*, **27**, 89–94 (1995).

## P15 Regional Grouping Method for Temperature Data

*Yu.V. Volkov*[1,2]

[1]*Tomsk Polytechnical University, Tomsk*

[2]*Institute of Monitoring of Climatic and Ecological Systems, Tomsk*

Sun is the main energy source for natural and climatic processes on the Earth. Energy is supplied in a carrier light of solar radiation and heat in different wavelengths. Depending on the latitude, the greatest amount of heat is produced at the lower atmosphere, directly adjacent to the earth's surface, they are heated to the highest temperatures. Terrestrial radiation determines the temperature regime and the corresponding circulation in the atmosphere. Temperature is the primary factor in the formation of weather and climate.

Series of monthly mean temperature have been obtained within the last 55 years at 333 meteorological stations located in the territory of Eurasia. Main purpose of this investigation is to determine regional characteristics of temperature changes.

The monthly mean temperature was changing during the period of record with forming of oscillatory process with quasi-period of one year. To characterize the weather of interest temperature changes that deviate from the annual cycle was used. However, is it difficult to select it.

The frequency spectrum of the temperature oscillation process is a narrow band with one mode in most cases, that allow to use the causality and introduce the phase of the oscillation, applying the theory of analytic signal [1-2]. Annual phase component is a linear function, it is removed from the phase by means of the least squares method in the interval of 55 years. The remaining phase fluctuations allow exploring consistently their coordinated behavior or synchronicity by means of Pearson correlation coefficient.

New algorithm of first pairwise correlation was used for calculation of all combinations of temperature series for each row of the group of temperature series that had close correlation coefficients above a predetermined level. After that for each group of rows arithmetic average value of the first level was calculated. These procedures are included in the iterative process in which input data are typical for the prior phase, while the output - phase following standard level for each of the temperature range and, consequently, for each weather station.

It was found that the iterative process has been designated for the usage of temperature convergent series. As a result, several groups of stations have been formed and for each group was calculated typical phase - the phase fluctuation model for the group. In this series in each group were highly correlated with the standard phase of their group and with a small typical phases of other groups.

It turned out that the selected group are located in the territory is compact enough that you can explain the presence of certain climates. Characteristically, the weather changes in average synchronous variant within these zones.

1. Vakman D. On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency // IEEE Trans. Signal processing. 1996. 44. № 4. P.791.

2. Cohen L. Loughlin P. Vakman D. On an ambiguity in the definition of the amplitude and phase of a signal // Signal Processing. 1999. 79. P.301.

## P16 Chemometric approach to identification of polyethylene by FT-IR/ATR spectroscopy

*A.G. Zarubin, O.N. Zarubina*
*National Research Tomsk Polytechnic University, Tomsk, Russia*

Recently, infrared ray spectroscopy has been used as a tool in analysis of variety of objects, such as petroleum, diesel, high-density polyethylene. Fourier transform infrared spectroscopy with attenuated total reflectance (FT-IR/ATR) has advantages in the small sample analysis and needs no special sample preparation . Materials based on polyethylene are increasingly used for the manufacture of pipes for transport of liquids. Therefore the quality of polymer materials is an important parameter in the overall evaluation of the material composition and structure. Using different mathematical approaches it became possible to calculate such parameters as kinematic viscosity, water content, and etc. However, sometimes transformation of initial data is needed.

In this work, FT-IR/ATR method in association with peak separation was applied for analysis of the high-density polyethylene pipes. Two different samples of the high-density polyethylene pipes were analyzed. We investigate the possibility to identify the polyethylene pipes samples using FT-IR/ATR spectroscopy. The parameter of identification,

$L_S = S_{2850}/ S_{730}$, is the ratio of the peak area of the valence band of the symmetric vibrations of –$CH_2$– group at 2850 $cm^{-1}$, $S_{2850}$, to the peak area of deformation pendular oscillations of group –$CH_2$– (crystalline phase) at 730 $cm^{-1}$, $S_{730}$.

IR spectra have a complicated structure that is an interference of individual signals of the functional groups to each other. To separate individual signals of functional groups the Peak Analyzer tool of software Origin was used. For each sample of the polyethylene pipe, five FT-IR/ATR spectra was obtained, and the mean values of $L_S$, and standard deviations of the mean were calculated. The average value for the first sample was $L_{S1} = 48.9$, and standard deviation of the mean was 0.8. The average value for the second sample, was $L_{S2} = 36.6$, and standard deviation of the mean was 1.3. The Euclidean distance between $L_S$ values can be a good criterion for identification of the polyethylene samples.

## P17 Selection of optimal regression algorithm for steel analysis by Laser-Induced Breakdown Spectrometry

*Sergey M. Zaytsev, Timur A. Labutin, Andrey M. Popov, Nikita B. Zorov*
*Lomonosov Moscow State University, Department of Chemistry, Moscow, Russia*

Laser-induced breakdown spectrometry (LIBS) is an emerging technique for materials analysis, based on laser sampling and simultaneous detection of the light emission from a laser-induced plasma. The possibility of a non-contact sampling by focused laser radiation makes possible real-time analysis of samples of different origin (construction and composite materials, coating layers, alloys, environmental and geological objects, pieces of the art, etc.) with minimal sample preparation and extremely short analysis time. These advantages of LIBS give an opportunity for rapid direct analysis of materials during operation process. Quantitative analysis of steels is focused both on impurities and doping components determination, but there are numerous spectral interferences due to extremely complex emission spectra of iron. In the field of control of both metallurgy process and assessment of constructions (e.g. railway rails) it is important to solve this problem for reliable and accurate analysis result.

Multivariate calibration techniques are commonly applied for analysis in the case of overlapping signals. In the present work we focused on searching an appropriate calibration strategy for LIBS determination either metal or non-metal components in various steels. We tried two multivariate regression methods (PCR and PLS) for solving the problem of spectral interferences for determination of silicon, chromium, nickel, manganese and carbon in steels. Data pre-processing includes the averaging of the spectra from several laser pulses and removing of the background as the minimal intensity in the spectrum. The stability of the model was verified by the one-leave-out cross-validation procedure. A special criterion for the determination an optimal number of principal components was used. We compared the results of multivariate regression with ordinary univariate linear regression. In cases of Mn and Cr we found experimental conditions to isolate an analytical lines and obtained a good univariate calibration with the use of baseline correction and an appropriate internal standard ($R^2$~0.996). For Si multivariate calibration provides worse results than univariate calibration despite the existing of spectral interferences of Si with the lines of Ni and Cr ($R^2$~0.86, $R^2$~0.94, respectively). Nevertheless, multivariate linear regression methods gives moderate prediction capability for all elements of interest in the case of overlapping signals ($R^2$~0.86 – 0.99).

*The work was financially supported by Russian Foundation for Basic Research (grant No. 13-03-13104-ofi_m_RZhD).*

# Participants

**Bogomolov Andrey**
Samara State Technical University
Chemometrician
Essingen, Germany
a.bogomolov@mail.ru

**Delakova Ekaterina**
The First Saint Petersburg State Medical University
Assistant
Saint Petersburg, Russia
ekaterina.delakova@gmail.com

**Ermakov Vasiliy**
Samara State Technical University
Head of laboratory
Samara, Russia
wassiliy@rambler.ru

**Galkin Evgeniy**
Csort
Head R&D depertment
Barnaul, Russia
info@csort.ru

**Inchakov Maksim**
JSC Gercules
NIR Coordinator
Klin, Russia
Maxim_inchakov@cargill.com

**Karpushkin Evgeny**
Moscow State University
Researcher
Moscow. Russia
eukarr@gmail.com

**Kirsanov Dmitry**
St. Petersburg State University

Associate Professor
Saint Petersburg, Russia
d.kirsanov@gmail.com

**Kuznetsov Vitaliy**
Tomsk Polytechnic University
PhD Student
Tomsk, Russia
kuvv@tpu.ru

**Cherednik Ekaterina**
Tomsk Polytechnic University
PhD Student
Tomsk, Russia
kinder_che@mail.ru

**Dolganova Irena**
Tomsk Polytechnic University
Engineer
Tomsk, Russia
dolganovaio@sibmail.com

**Marini Federico**
Sapienza University of Rome
Researcher
Roma, Italy
federico.marini@uniroma1.it

**Galyanin Vladislav**
Samara State Technical University
PhD Student
Samara, Russia
v.galyanin@gmail.com

**Kalambet Yuri**
Ampersand Ltd.
General Director
Moscow, Russia
kalambet@ampersand.ru

**Khaydukova Maria**
St. Petersburg State University
PhD Student
Saint Petersburg, Russia
khaydukova.m@gmail.com

**Kumkov Sergey**
Institute of Mathematics and Mechanics, Ural Branch, Russian Academy of Sciences
Senior Research Scientist
Ekaterinburg, Russia
kumkov@imm.uran.ru

**Labutin Timur**
Lomonosov Moskow State University
Researcher
Moskow, Russia
timurla@laser.chem.msu.ru

**Larionova Ekaterina**
Tomsk Polytechnic University
Associate Professor
Tomsk, Russia
evl@tpu.ru

**Oleneva Ekaterina**
St. Petersburg State University
Student
Saint Petersburg, Russia
ekaterina.oleneva@inbox.ru

**Pentti Minkkinen**
Lappeenranta University of Technology
Professor Emeritus
Lappeenranta, Finland
pentti.minkkinen@lut.fi

**Pomerantsev Alexey**
Semenov Institute of Chemical Physics RAS
Leading Researcher
Moscow, Russia
alexey.pomerantsev@gmail.com

**Romanenko Sergey**
Tomsk Polytechnic University
Professor
Tomsk, Russia
svr@tpu.ru

**Sergey Zaytsev**
Lomonosov Moskow State University
PhD Student
Moskow, Russia
sergz@laser.chem.msu.ru

**Shekhovtsova Nataliya**
Tomsk Polytechnic University
Associate professor
Tomsk, Russia
nssh@tpu.ru

**Skvortsov Aleksey**
Saint Petersburg State Polytechnic University
Professor
Saint Petersburg, Russia
colbug@mail.ru

**Turov Yury**
Surgut State University
Associated professor
Surgut, Russia
yuri_tom@rambler.ru

**Melenteva Anastasiia**
Samara State Technical University
PhD Student
Samara, Russia
melenteva-anastasija@rambler.ru

**Panchuk Vitaly**
St. Petersburg State University
Associate Professor
Saint Petersburg, Russia
vitpan@mail.ru

**Podchasov Anton**
Csort
Programming ingenier
Barnaul, Russia
info@csort.ru

**Rodionova Oxana**
Semenov Institute of Chemical Physics RAS
Leading Researcher
Moscow, Russia
oxana.rodionova@gmail.com

**Savinkov Maksim**
Csort
Director
Barnaul, Russia
info@csort.ru

**Shary Sergey**
Institute of Computational Technologies SB RAS
Senior Reseacher
Novosibirsk, Russia
shary@ict.nsc.ru

**Sidelnikov Artem**
Bashkir State University
Docent
Ufa, Russia
artsid2000@gmail.com

**Talavera Isneri**
CENATAV
Deputy Director
La Habana, Cuba
italavera@cenatav.co.cu

**Volkov Yuri**
Tomsk Polytechnic University
Associate professor
Tomsk, Russia
yvvolkov@tpu.ru

**Xu Qingsong**
Central South University
Professor
Changsha, Hunan, China
qsxu@csu.edu.cn

**Liang Yizeng**
Central South University
Professor
Changsha, Hunan, China
yizeng_liang@263.net

**Zhilin Sergey**
Altay State University
Associate Professor
Barnaul, Russia
sergei@asu.ru

**Yaroshenko Irina**
St. Petersburg State University
PhD Student
Saint Petersburg, Russia
papieva_irina@mail.ru

**Zarubin Alexey**
Tomsk Polytechnic University
Associate Professor
Tomsk, Russia
zagtpuru@gmail.com

## Notes