Thesis of Dissertation

Anita Rácz

Budapest
2016

SZENT ISTVÁN UNIVERSITY

CHEMOMETRICS AND FT-NIR SPECTROSCOPY IN FOOD ANALYSIS

Anita Rácz

Budapest
2016

PhD School

Name: PhD School of Food Science
Field: Food Science

Head: Prof. Gyula Vatai, DSc
Professor
SZIU, Faculty of Food Science,
Biotechnology and Process Design Institute
Department of Food Engineering

Supervisor(s): Marietta Fodor, PhD habil.
Associate professor
SZIU, Faculty of Food Science,
Food Quality and Safety and Nutritional Science Institute
Department of Applied Chemistry

Prof. Károly Héberger, DSc
Scientific advisor
Hungarian Academy of Sciences,
Research Centre for Natural Sciences
Institute of Materials and Environmental Chemistry
Department of Plasma Chemistry

The applicant met the requirement of the PhD regulations of the Szent István University and the thesis is accepted for the defense process.

....

Signature of Head of PhD School Signature of Supervisor

.....

Signature of Supervisor

INTRODUCTION

Fast, non-destructive and cost efficient analytical methods are frequently used in the area of food science and food industry. I used near infrared spectroscopy in my PhD work, which follows the above directive. Near infrared spectroscopy can be effectively used for the analysis of several food components. It can provide reliable quantitative and qualitative results as well. The quality control of food products is indeed important for both the producers and the consumers, because there is a huge competition on the market of food products, where the guarantee of quality is an essential condition. While the Hungarian standards and guidelines for food analysis are sometimes really outdated and recommend expensive procedures, the use of fast analytical techniques – such as NIR spectroscopy –opens a new way, which is economically more favorable and more efficient.

Modern analytical chemistry together with food analytics have gained increasing importance since the last decades of the twentieth century. The appearance of faster and more modern computers had a huge impact in the improvement of analytical methods, especially in the field of spectroscopy, where we work with huge amount of data. With the "data explosion" of the past few decades the time demand for calculation processes decreased with the appearance of high performance supercomputers. Chemometrics, as the application of statistics to multidimensional chemical datasets has become more and more popular after the mid-20th century, but the huge breakthrough was in the 80's – 90's with the appearance of desktop computers. Most of the best and valuable basic chemometric tutorials and research articles were written at that time.

We can find several applications of chemometric methods in food science as well. Dominant patterns, similarities and dissimilarities in the dataset can be revealed with the use of chemometric methods. Hidden connections come to light and help to make qualitative and quantitative analyses. Chemometrics is connected to spectroscopy, because a spectral dataset with the huge number of variables usually cannot be evaluated without special chemometric tools. Thus, chemometric analysis was essential for the evaluation of near infrared spectra, which were the most important and frequently used datasets in my doctoral work. Evaluation of the complex food analytical spectra can be carried out only with the help of chemometrics.

MAIN OBJECTIVES

My doctoral thesis can be decomposed to three larger parts according to the examined sample matrices: i) coenzyme Q10 dietary supplements ii) complex examination of energy drinks and iii) antioxidant capacities. Chemometric method developments are discussed separately in my thesis.
The aims of my work for each separate section were the following:

Chemometric method developments:

- Development of a novel variable selection method and application to the FT-NIR spectra of Q10 coenzyme dietary supplements. Increasing the goodness of models with variable selection (omitting thousands of variables).

- Development of *n*-class ROC curves, which can be used for the evaluation of multi-class problems. Testing and validation of the developed method with the use of a dataset of energy drinks. Comparison of pattern recognition techniques with *n*-class ROC curves.

- Optimization of the model building process for the random forest (RF) technique to increase the goodness of the models.

- Comparison and ranking of the performance parameters in regression analysis.

Coenzyme Q10 dietary supplements:

- Examination of the coenzyme Q10 content in dietary supplements with FT-NIR spectroscopy.

- Building appropriate calibration models, which can be validated with cross-validation and test samples as well.

- Replacement of time-consuming and expensive HPLC and other commonly used techniques.

- Comparison of the final PLS-R models with different variable selection techniques by sum of ranking differences (SRD).

Examination of energy drinks:

- Development of an easy high-performance liquid chromatographic (HPLC-UV) method using an international standard to provide a reference method for the determination of caffeine concentrations.

- Determination of the sugar content of energy drinks with the Schoorl method.

- Development of a novel, money- and time-saving method for the determination of caffeine and sugar concentration in energy drinks with FT-NIR spectroscopy. Internal and external validation of the final models.

- Classification of energy drinks based on their FT-NIR spectra. Differentiation of the energy drinks with taurine content, arginine content and without taurine or arginine content.

Examination of antioxidants:

- Comparison of antioxidant capacity assays using statistical methods (HCA, PCA, SRD and GPCM) and selecting the most representative method or methods for the available datasets based on time and cost efficiency.

III. MATERIALS AND METHODS

In my doctoral thesis chemometric analyses are usually based on FT-NIR spectra and the use of classical and modern analytical techniques. Chemometric method developments were also carried out, but I discuss these methods in the Results section.

3.1 FT-NIR spectroscopy

A Bruker MPA™ Multipurpose Fourier-transform near-infrared spectroscopy (FT-NIR) analyzer (Bruker Optik GmbH, Ettlingen, Germany) was used for FT-NIR measurements. The device is equipped with a quartz beam splitter, an integrated Rocksolid™ interferometer, a thermostated sample compartment equipped with a flow-through cuvette, and a Te-InGaAs detector working in the 800–2500 nm wavelength range (12,500–4000 cm^{-1} wavenumber). These parameters were used in transmission mode for the collection of absorption spectra for energy drinks. The solid coenzyme Q10 samples were measured by a rotatable sample wheel and a PbS detector. The spectral resolution was 8 cm^{-1}, the scanner speed was 10 kHz, and each spectrum was the average of 32 subsequent scans in both cases.

3.2 HPLC measurements

For the determination of the total coenzyme Q10 content, an Agilent 1200 HPLC (Agilent Technologies) system was used in isocratic mode on an Agilent Zorbax XDB C18 HPLC column (2.1 mm × 50 mm × 3.5 μm) followed by UV detection at 275 nm. The column temperature was set to 30 °C. The eluent consisted of the mixture of ACN:THF:water in 65:30:5 %v/v rate and the flow rate was 0.35 ml min^{-1}. The injection volume was 10 μl.

The international standard for the determination of caffeine content in coffee and coffee products (ISO 20481:2008) was adapted for the energy drink samples. Briefly, an Agilent 1200 HPLC (Agilent Technologies, Santa Clara, CA, USA) system was used for the HPLC-UV-based quantification of caffeine. An Agilent Zorbax XDB C18 HPLC column (4.6 mm × 150 mm × 5.0 μm) was used in isocratic mode at 40 °C. The flow rate was 1 ml min^{-1}, the injection volume was 20 μl, while the chromatographic run lasted for 18 min. UV detection was carried out at 273 nm, and additional peak purity measurements were executed at 260 nm in order to exclude samples containing impurities in the retention window of caffeine.

3.3 Sample preparation

For the HPLC measurements in the case of coenzyme Q10 the procedure was based on the AOAC Official Method 2008.07 and later optimized by Vass et al.

In the case of energy drinks for the HPLC-UV measurements, the samples were sonicated in an ultrasonic bath (type T2MODX; VWR) for 20 min; then, 50 µl of them was diluted to 1600 µl with ultra-pure water in vials. External calibration with peak area integration was used for the quantification of total caffeine concentration in the energy drink samples. The calibration points were the following: 2.5, 5.0, 10.0, and 20.0 ppm (because of the 32-times dilution). Here the only sample pretreatment step was pouring the samples into 10 ml vials for the FT-NIR analysis after the sonication.

In the case of coenzyme Q10 the only sample pretreatment for FT-NIRS analysis was the careful homogenization of tablets and the content of encapsulated products in a mortar.

3.4 Classical analytical method – Schoorl method

The Schoorl method was applied as the reference for the determination of sugar concentration in energy drink samples. This method is frequently used for the determination of sugar content in food analysis. In this method the reducing sugar components can be examined. The analysis is based on the OÉTI ÉLK 4.009 standard method.

The sugar content of energy drinks was calculated for invert sugar, because the samples were inverted with acidic hydrolysis. It was the most appropriate solution, because the exact sugar compound is indicated rarely on the energy drink cans.

3.5 Softwares and chemometric methods

FT-NIR spectra were evaluated with OPUS 6.5 (Bruker Optik GmbH, Ettlingen, Germany) and Unscrambler version 9.7 (CAMO Software, Oslo, Norway). In the pattern recognition section STATISTICA 12 (Statsoft Inc., Tulsa, OK, USA) with PCA (Principal component analysis), LDA (Linear discriminant analysis), RF (Random forest), BT (Boosted tree), PLS-DA (Partial least-squares discriminant analysis) methods were used. The SRD (Sum of ranking differences) and GPCM (General pair correlation method) techniques were connected to MS Excel as macros. A home-made Linux code was developed for calculating *n-class* ROC curves, which can be found in my doctoral thesis.

IV. RESULTS

4.1 Chemometric method developments

4.1.1 Interval selectivity ratio (iSR)

The basic idea of this method was developed by Rajalahti *et al*. The selectivity ratio can be calculated for each spectral variable with the following equation:

(1) $\mathrm{SR}_i = v_{exp,i}/v_{res,i}$, ahol $i = 1, 2, 3 \ldots m$.

Where v_{exp} (R^2) is the explained variance and v_{res} (squared error of calibration, SEC) is the residual variance. The original equation of selectivity ratio was slightly modified, introducing a square root in the denominator. A simple algebraic transformation (square root or division by the same number — the degree of freedom) does not change the tendencies observed. I modified the equation for spectral intervals:

(2) $\mathrm{SR}_i = R_i^2/\mathrm{RMSEC}_i$, $i = 1, 2, 3 \ldots m$.

where i=1, 2, 3 … m (m is the number of intervals). In this study, interval SR was used for the same reasons as iPLS. In the latter equation, degree of freedom is not used for R^2, but this equally influences all intervals. The higher the selectivity ratio, the higher the importance of the interval.

4.1.2 *n*-class ROC curves

My approach to the multi-class ROC (Receiver operating characteristic) analysis of the three-class classification problem presented in my thesis is based on the ‘one *versus* all’ method of Provost and Domingos. Thus, I calculated AUC values by taking each class as positives (and all the others as negatives) in turn. The weighted average of these AUC values gives the overall AUC of the given classifier method:

(3) $$\overline{\mathrm{AUC}} = \frac{\sum_{i=1}^{n} N_i \mathrm{AUC}_i}{\sum_{i=1}^{n} N_i}$$

The weights *Ni*-s are the number of samples of each class. We can visualize an overall ROC curve for a classifier method by plotting a “ROC-like” curve with the overall AUC value using the Hanley formula. The variance of the overall AUC value can be calculated with the law of error propagation:

(4) $$Var(\overline{\mathrm{AUC}}) = \frac{\sum_{i=1}^{n} N_i^2 Var(\mathrm{AUC}_i)}{\sum_{i=1}^{n} N_i^2}$$

When comparing more classifiers, ROC-like curves with AUCs corresponding to either confidence limits or the mean ± one SD of the overall AUC can also be plotted to decide whether the performances of the methods differ significantly. In my doctoral work, I plot curves corresponding to the mean ± one SD for the comparison of the methods.

Four classification methods were compared using ROC curves to identify the best one: two common ones (linear discriminant analysis, LDA and partial least squares, PLS) and another two (random forest, boosted tree) that are not applied as frequently as LDA or PLS yet. Two datasets were applied for the further statistical analysis: the first one contained the 90 energy drink samples' spectral data and the second was calculated from the spectral data with principal component analysis. Both of the matrices had a categorical variable with the classes of the sugar contents. In the final step of the method comparison a comparative plot based on PCA scores with the average ROC curves of each classification model was plotted, which can be seen in Fig. 1. All classification methods have very high AUC values, but the best one is without doubt the boosted tree method. It was also the best one in the case of the spectral dataset.

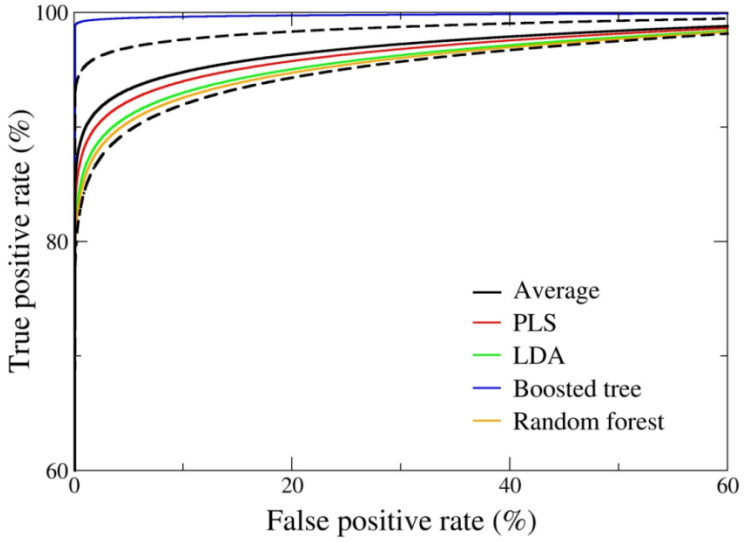

Figure 1: The final comparison of the four classification methods for the PCA scores dataset. The plot is the magnified version of the original one for better visualization. Dashed lines indicate ± 1 SD from the average.

4.2 Quantitative determination of coenzyme Q10 from dietary supplements

Initially, 52 dietary supplements were measured by FT-NIR spectroscopy in the range of 12,800–3,600 cm^{-1} (800– 2,700 nm). The average of each samples' spectra was used for the further statistical analysis. The part of the spectra between 12,800 cm^{-1} and 9,000 cm^{-1} (800–1,111 nm) was cut because it did not carry any systematic information. The first step before the building of the calibration model is the outlier detection. There were two spectral outliers, which were clearly different from the average spectrum. Principal component analysis was used as a verification method in spectral outlier detection. The first model was made for the original dataset, which contains fifty samples and the whole spectrum range was between 9,000 and 3,600 cm^{-1}. Because the calibration model cannot be considered perfect, I wanted to improve it with the application of fewer variables. Different variable selection techniques (for instance interval PLS, interval selectivity ratio, genetic algorithm) were used for the improvement of the original PLS regression model. Derivation was used as data preprocessing method in each case, which gave the best result in the model building phase. Five-fold cross-validation and test validation sets were also applied in the validation process for each regression model. Table 1 contains the performance parameters for the best three regression models. The first row in the table contains the original PLS regression model without any data preprocessing and variable selection.

Table 1: Performance parameters, scaling methods and number of latent variables of the best and original models
R^2: determination coefficient, Q^2: determination coefficient for the cross-validated model, RMSE: root mean squared error (C=for the calibration model, CV= for the cross-validated model, P= for the test validation) values. D is the shorter form of derivation data scaling method.

R^2	Q^2	RMSEC	RMSECV	RMSEP	Scal.	Comp. number	Var. select.
0,85	0,71	11,26	15,74	-	-	9	-
0,92	**0,87**	**8,16**	**10,58**	**8,82**	**D**	**7**	**iPLS**
0,90	**0,87**	**8,90**	**10,85**	**13,74**	**D**	**6**	**iSR**
0,91	**0,88**	**8,48**	**10,22**	**11,12**	**D**	**6**	**GA**

The final models were compared with sum of ranking differences, thus I could select the best and most consistent model. Average was used as the

reference and the predicted values were compared to this. According to this comparison the best model was clearly the iPLS variable selected one.

4.3 Quantitative determination and classification of energy drinks using near-infrared spectroscopy

In this part the results are discussed in two separate groups: i) the classification/pattern recognition models and ii) regression model building for the caffeine and sugar contents of energy drinks.

4.3.1 Classification of energy drinks

FT-NIR spectra of 108 energy drinks samples were evaluated with PCA and LDA. With the use of PCA as a data reduction technique, we could eliminate the limitation in the number of variables. The aim of the evaluation was to classify the energy drinks into three groups, based on whether (i) it contains arginine, (ii) it contains taurine, or (iii) no taurine and arginine are present in the samples. In the first step, the average spectra of the samples from 12,500 to 4000 cm^{-1} were used for principal component analysis. Standardization (standard normal variate) was applied as data preprocessing. After that, the first 20 PCA scores were used for further analysis with LDA. Forward stepwise model building and three-fold cross-validation were applied in the evaluation. Proper validation is very important; it should be tested, whether the results are artefacts or not. For this purpose, as another validation method for the model, X-scrambling randomization test was used three times. Figure 2 shows the final result with the comparison of a typical example for X-scrambling validation model. The earlier mentioned three groups can also be clearly classified based on LDA and PCA analysis (and only FT-NIR spectra) and the validation of the model returned good results as well.

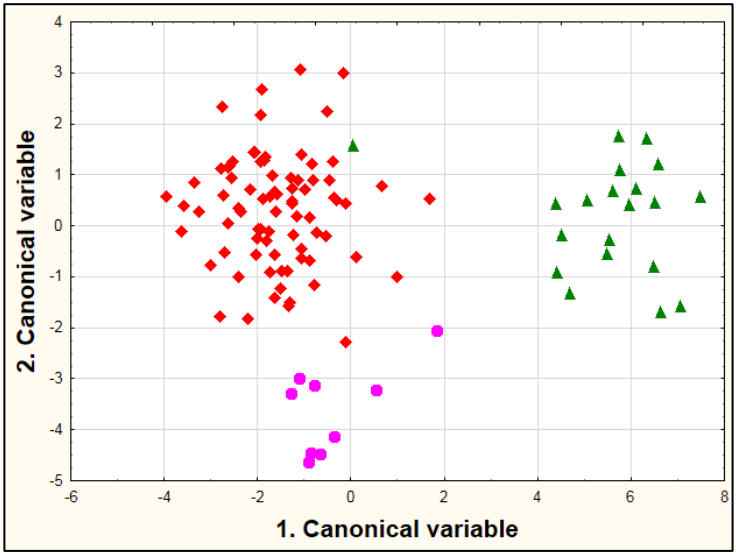

Figure 2: Classification model of the energy drinks: samples containing arginine are marked with pink circles, samples containing taurine are marked with green triangle and normal (without taurine and arginine) samples are marked with red squares.

4.3.2 Determination of caffeine and sugar content of energy drinks

Determination of caffeine content

The 42 original energy drink samples were measured first with the HPLC-UV method. The other 33 mixtures were prepared from the original ones. Every sample was measured three times with HPLC-UV, and then the average of the calculated caffeine concentrations was used for the FTNIR measurements as reference values. The running time of the HPLC-UV analysis was 18 min. The retention time for the caffeine peak was around 9.5 min. Every sample was examined three times from 10 ml vials with a quartz flow cuvette with the FT-NIR analyzer. Figure 2 shows an example of the measured spectra and its derivative form.

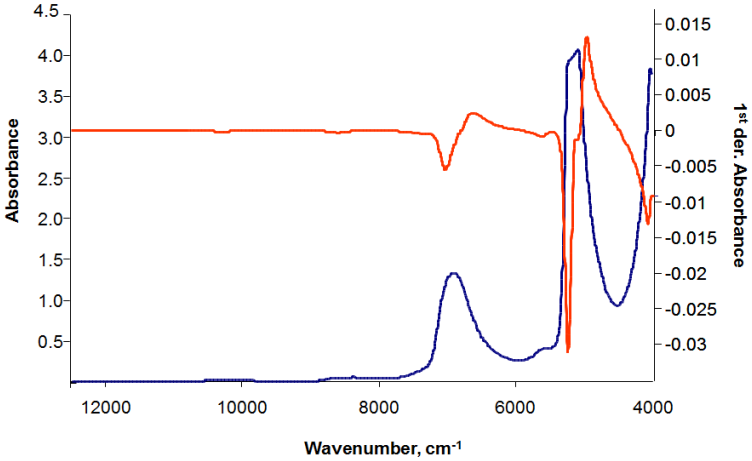

Figure 3: An example of the measured samples spectra and its derivative form. The original spectrum is marked with blue and the derivative is marked with red.

The concentration range of caffeine was between 118 and 338 ppm, based on HPLC-UV determination. Principal component analysis was used for spectral outlier detection. There was no spectral outlier in our dataset, thus the final number of samples was 75. The applied data preprocessing methods were derivation and standardization (standard normal variate). The number of latent variables was eight, which was chosen based on the global minimum of the root mean squared error of cross-validation (RMSECV). Seven-fold cross-validation was used as validation of the PLS regression model. Finally a test validation with 13 new, commercial samples was also done. The performance parameters of the final model can be seen in Table 2.

Determination of sugar content

Seventy-one original and 20 mixed samples (91 in all) were used for the determination of sugar content in the energy drinks. The mixture samples were made from the original ones with the use of different mixing ratios. The Schoorl method was applied as the reference for the determination of sugar concentration. This method is frequently used for the determination of sugar content in food analysis. Seventy-five of the 91 samples were chosen and measured in this way. However, the method has a large bias and relatively large standard deviation (namely 12.4 %), especially in the range of small amounts of sugar (1–2 g/100 ml). Thus, I decided to use and compare both of the original

(indicated on the can) and the measured values, because the nominal concentrations have less error (based on a simple weighing). In this case, every sample was analyzed three times from 10 ml vials in a quartz flow cuvette with an FT-NIR analyzer, as well. The average of the spectra was used for further chemometric analysis.

First, PCA was applied to detect spectral outliers. There were two spectral outliers in the dataset, which were omitted before the model building process. PLS regression was used for model building. The model optimization for the 89 samples was carried out with OPUS 6.5; first derivative and standardization were used for data preprocessing. The concentration range for sugar was between 0.0 and 14.9 g/100 ml. Six latent variables were sufficient for model building, based on the global minimum of the RMSECV curve. Model building was repeated with the reference dataset based on the sugar content measurements. The two spectral outliers (as in the previous case) were omitted from the dataset, thus the final number of samples was 73. In this case, the concentration range extends between 0.1 and 15.3 mg/100 ml. The model optimization processes were the same as in the previous case. Seven-fold cross-validation was used in both cases. The performance parameters of the two models can be seen in Table 2.

Table 2: Summary of the final regression models for caffeine and sugar content determination in energy drinks

	N	**C**	R^2	Q^2_{ext}	Q^2	***RMSECV***	***RMSEP***
Caffeine model	75	8	0,966	0,898	0,928	16,8	36,3
Sugar model (Schoorl)	73	6	0,943	0,935	0,919	1,13	1,23
Sugar model (nominal)	89	6	0,998	0,996	0,995	0,29	0,26

4.4 Comparison and ranking of antioxidant capacity assays with chemometric methods

Antioxidant capacity values for thirteen berry genotypes and twelve sour cherry cultivars were measured by seven antioxidant capacity assays in total (FRAP, TPC, TRSC, DPPH, ACL, and ACW for the berry samples and FRAP, TPC, TEAC, ACL, and ACW for the sour cherry samples). Every sample had two duplicates and each duplicate was measured three times. The average value of the measurements for each sample was used for the further statistical analysis. The first matrix (13 × 6) contained the berry sample antioxidant capacity values for six determination techniques (FRAP, TRSC, TPC, DPPH, ACL, and ACW). The second data matrix (12 × 5) contained the sour cherry sample antioxidant activity values for five techniques (TPC, FRAP, TEAC, ACL, and ACW). Both datasets were standardized before statistical evaluation.

First the comparisons and connections between the methods are shown in the results with PCA and HCA. Finally, the rankings produced by SRD and GPCM are presented. In the case of cluster analysis Euclidian distance and Ward's method were used as distance measure and linkage rule, respectively, for both datasets. In the evaluation of berry samples two groups were clearly separated, one contained the ACL and ACW techniques, and the other one contained the TRSC, DPPH and FRAP methods. In the case of sour cherry dataset also two clusters were observed. The ACW and ACL methods clearly form a distinct group and the other three are more closely connected to each other. The results were confirmed with PCA as well. With the use of the first two PCA loading vectors ACW, ACL and TPC were scattered, but DPPH and FRAP were in close proximity in the case of the berry dataset. In the case of the sour cherry dataset with the use of three PCA loading vectors the pattern could verify the results of cluster analysis (where ACW and ACL formed an individual cluster).

Before the SRD evaluation, the data matrix had to be preprocessed to convert all variables to the same scale. In this case, the average was chosen as reference for all of the datasets. In the diagram (Figure 4) the scaled results are used, which makes the methods comparable.

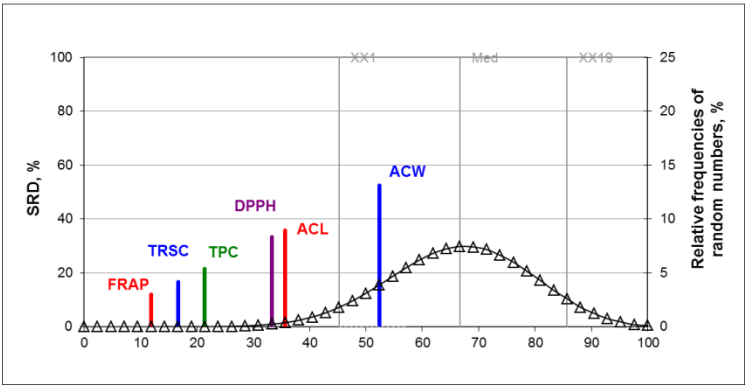

Figure 4: Evaluation of six antioxidant capacity methods using sum of ranking differences (berries dataset). The right y axis shows the relative frequencies for the black Gauss-like curve with triangles (exact theoretical distribution).

The lower the SRD value, the closer it is to the reference (to the average). Thus, FRAP can substitute all methods for antioxidant capacity with the smallest error. I have also plotted the random probability distribution curve (a Gauss like one), which helps to decide whether the applied method is better than or similar to the use of random numbers. All of the methods produce better results than random numbers, except for ACW. Validation of the ranking has been carried out using a randomization test and a seven-fold cross-validation. Nonparametric tests and *t*-test - based on the validation results - clearly indicated that SRDs for DPPH and ACL are derived from the same distribution.

For the sour cherry dataset, SRD analysis was applied in the same manner as for the previous case. The SRD result suggests that TPC has the smallest error out of the five applied methods and hence it can be used to replace all of the other methods. ACL and ACW methods are outside the acceptable region of the graph. FRAP and TEAC had the same SRD value, and their medians are indistinguishable according to the Sign and the Wilcoxon matched pair tests.

The general pair correlation method (GPCM) results were highly similar to the SRD results, but GPCM could even distinguish the methods DPPH and ACL in the first case, and FRAP and TEAC in the second case.

4.5 Comparison and ranking of the performance parameters in regression analysis

As an outlook in my PhD work I apply two QSAR (quantitative structure activity relationships) datasets as case studies for the comparison of model performance parameters. It was necessary, because with the use of these datasets, more general conclusions can be drawn and a lot of performance parameters could be calculated. The first dataset contained toxicity values of benzene derivatives as the Y variable, and the other contained biological activity values (IC_{50}) of N-substituted maleimides. Multilinear regression (MLR) was used for the evaluation, which is a commonly used regression technique in QSAR analysis. For SRD analysis both Case studies' dataset contained 20 performance parameters in the columns and in the first case there were 60 models in the rows. Row-average was used as the reference. Seven-fold cross-validation was used for the verification of the analysis. Figure 5 shows the validation results on a Box & Whisker plot for the first case study.

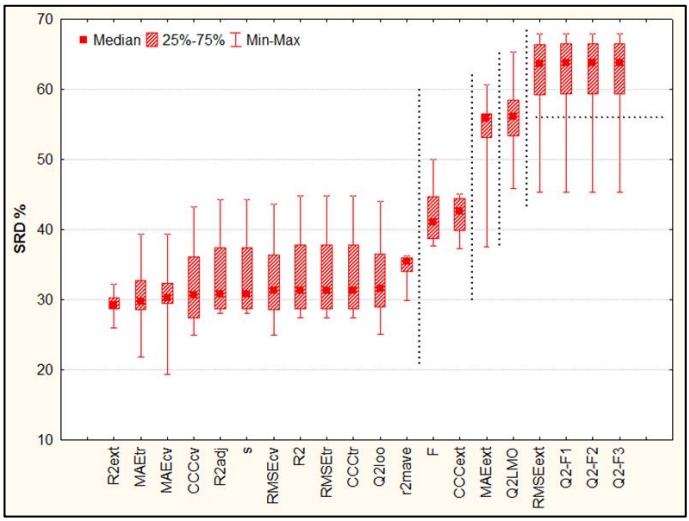

Figure 5: Comparison of performance parameters using seven-fold cross-validation of scaled SRD values. On the box and whisker plot the horizontal dotted line shows the 5 % error limit for random ranking. Vertical dotted lines show the 5% error limit for the Wilcoxon matched pair test.

There were a few performance parameters (Q2-F1, Q2-F2, Q2-F3 and RMSEext) that overlap with random ranking, but most of the parameters are located between zero and the 5 % limit for random ranking.

The same examination was carried out for Case study 2, where the number of the columns was 20 with the same performance parameters, but the number of the rows was 70 (since here the number of created models is 70).

In summary I can conclude that the performance parameters connected to training (calibration) and cross-validation, were the most representative and consistent group. The best and most consistent performance parameters for both datasets were RMSECV, CCC_{cv}, Q^2_{LOO} and MAE.

V. NEW SCIENTIFIC ACHIEVEMENTS

1. In my doctoral work I developed *n*-class ROC curves, which can be applied for the comparison and evaluation of models – including error estimation. Classification abilities of the models can be illustrated properly. Based on the FT-NIR spectra of energy drinks the different classification methods can be compared with *n*-class ROC curves. I also developed and used two parameter optimization processes for the random forest (RF) method. I verified that the boosted tree (BT) method has the best result for my datasets in classification. I used a home-made Linux code for the calculation of *n*-class ROC curves.

2. I developed three regression models, which can be used for the determination of coenzyme Q10 concentration in dietary supplements from their FT-NIR spectra in a fast, easy and environmentally friendly way. The models were validated with internal and external validation. With the use of these models, the frequently used, but slow and expensive HPLC methods can be replaced.

3. As a part of coenzyme Q10 concentration determination I also developed the interval selectivity ratio (iSR) method, which can be used for variable selection. I applied this method in an efficient way in the model building phase of the regression models.

4. I also developed PLS regression models for the determination of sugar and caffeine concentration in energy drink samples based on their FT-NIR spectra. The models are properly accurate and robust. I used internal and external validation for the verification of the models' goodness, thus these models can be used for sugar and caffeine concentration determination instead of the time consuming and expensive other techniques. On the other hand I can classify properly the taurine, arginine and normal (without the mentioned components) samples with my LDA classification model. This latter one can be good for quality control and identification of the energy drink samples' origin.

5. In the case of the comparison of different antioxidant capacity methods I found that the methods based on similar chemical backgrounds give more similar (unified) results in PCA and HCA analyses. ACW and ACL methods differed from the other techniques in the examination processes. The most consistent method based on chemometric evaluation were FRAP and TPC. They can replace the other ones, thus if we do not have enough time, these two method can give the best result with minimum error.

6. In the wide-spread comparison of performance parameters I found a big difference between the parameters based on internal and external validation. The performance parameters based on internal validation are usually better and more consistent ones. The most applicable ones can be: RMSECV, CCC_{cv} and Q^2_{LOO}. From the other point of view the external (test) validation performance parameters can carry interesting and useful information, because of their dissimilarity.

VI. CONCLUSIONS AND SUGGESTIONS

In my doctoral work I showed the useful opportunities of FT-NIR examinations and chemometric approaches with several examples. With my experiments and evaluations I have made gradual developments in the field of food science (food analytics) and chemometrics.

The market of coenzyme Q10 is nowadays still increasing, because lots of people believe in its health benefits. In the intense market competition, the quality control of these products plays and important role. In this segment my regression models can replace perfectly the commonly used, expenseive and relatively slow HPLC methods. On the other hand with the developed variable selection technique (iSR) I can give new opportunities for variable selection in chemometric analysis. The variable selection is a key step in the regression model building phase, which can be seen in the examination of coenzyme Q10 dietary supplements. The iSR variable selection method – together with the other two applied methods - was efficient in the variable selection phase of coenzyme Q10 concentration regression models, thus its application can be proposed for further regression analyses.

In the field of energy drinks I recognized that these non-alcoholic drinks have a huge impact in the Hungarian market because of the number of consumers. As ten years earlier there were only a few brands in the Hungarian market, now the numbers of brands are more than a hundred. Unfortunately the biggest consumer group is the youth, which includes children as well. Because of this special consumer group the quality control of these samples is highly important, in particular the determination of caffeine and sugar content, because they can be exposed to health risks. In my doctoral work I successfully built regression models to the caffeine and sugar concentration of the samples, which provide a fast and environmentally friendly way to examine the caffeine and sugar concentration. Thus these models can be used instead of the other commonly used analytical techniques. I examined around a hundred of samples in my work, which means that it is optimized to all of the energy drinks in the Hungarian market. In the field of classification I developed a model for the differentiation of taurine, arginine and normal energy drink samples (without taurine and arginine). This classification model can be good for quality control and identification of the energy drink samples' origin. With the use of the FT-NIR spectra of energy drinks I developed and tested a novel form of *n*-class ROC curves. The *n*-class ROC curves are used for the comparison of several classification methods with a very attractive visual interpretation. The less

frequently used classification techniques had as good or even better classification ability as the old and commonly used ones. In my case the boosted tree method gave the best classification result, thus it can be a good opportunity for further analyses.

In the world of antioxidant capacity measurements several determination techniques were developed; thus, a decision to choose the best one for the examination is sometimes complicated. Usually the use of more techniques is recommended. In this part of my doctoral work I raised the question, which method can give us the least error and the most consistent results. In my examinations (with two datasets) FRAP and TPC techniques proved to be the most consistent ones, thus these methods can replace the others, if we do not have sufficient time or it is too expensive to measure antioxidant capacity in several ways.

In regression model building, the use of appropriate performance parameters is always a problem not just in the case of FT-NIR spectra but in the case of pharmaceutical and other datasets as well. The regression models are judged in different ways based on the different performance parameters, thus the question of 'what should we use' is very important. In my case studies the external validation parameters can give extra information about the models, the internal validation parameters such as RMSECV or CCC_{CV} can give more consistent results. Based on my evaluations the results based on internal validation and cross-validation result should be emphasized in regression model building.

PUBLICATIONS

Journals with impact factor:

1. **A. RÁCZ**, K. HÉBERGER, M. FODOR (2016): Quantitative determination and classification of energy drinks using near-infrared spectroscopy.
Analytical and Bioanalytical Chemistry, Vol. 408. pp. 6403-6411.
(IF=3.125)

2. H. TIMA, **A. RÁCZ**, ZS. GULD, CS. MOHÁCSI-FARKAS, G. KISKÓ (2016): Deoxynivalenol, zearalenone and T-2 in grain based swine feed in Hungary.
Food Additives & Contaminants: Part B. Vol. 9. pp. 275-280.
(IF=1.467)

3. F. ANDRIC, D. BAJUSZ, **A. RÁCZ**, S. SEGAN and K. HÉBERGER (2016): Multivariate assessment of lipophilicity scales – computational and reversed phase thin-layer chromatographic indices.
Journal of Pharmaceutical and Biomedical Analysis, Vol. 127. pp. 81-93. (IF=2.867)

4. **A. RÁCZ**, D. BAJUSZ, M. FODOR and K. HÉBERGER (2016): Comparison of classification methods with "*n*-class" receiver operating characteristic curves: A case study of energy drinks.
Chemometrics and Intelligent Laboratory Systems, Vol. 151. pp. 34-43. (IF=2.321)

5. D. BAJUSZ, **A. RÁCZ** and K. HÉBERGER (2015): Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?
Journal of Cheminformatics, Vol. 7:(1) paper 20. 13 p.
(IF=4.547)
Citations: 36 (MTMT)

6. **A. RÁCZ**, N. PAPP, E. BALOGH, M. FODOR and K. HÉBERGER (2015): Comparison of antioxidant capacity assays with chemometric methods. *Analytical Methods: Advancing Methods and Applications*, Vol. 7. pp. 4216-4224.
(IF=1.821)
Citations: 3 (MTMT)

7. **A. RÁCZ**, A. VASS, K. HÉBERGER and M. FODOR (2015): Quantitative determination of coenzyme Q10 from dietary supplements by FT-NIR spectroscopy and statistical analysis.
Analytical and Bioanalytical Chemistry, Vol. 407:(10). pp. 2887-2898. (IF=3.436) Citations: 5 (MTMT)

8. **A. RÁCZ**, D. BAJUSZ, K. HÉBERGER (2015): Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters.
SAR and QSAR in Environmental Research, Vol. 26:(7-9). pp. 683-700. (IF=1.596)
Citations: 5 (MTMT)

9. O.H.J. CHRISTIE, **A. RÁCZ**, J ELEK, K. HÉBERGER (2014): Classification and unscrambling a class-inside-class situation by object target rotation: Hungarian silver coins of the Árpád Dynasty, ad 997–1301.
Journal of Chemometrics, Vol. 28:(4). pp. 287-292.
(IF=1.803)
Citations: 2 (MTMT)

10. **A. RÁCZ**, K. HÉBERGER, R. RAJKÓ, J. ELEK (2013): Classification of Hungarian medieval silver coins using X-ray fluorescent spectroscopy and multivariate data analysis.
Heritage Science, Vol. 1:(2) paper 1/1/2. 9 p.
(IF(Chemistry Central Journal)=1.663)
Citation: 3 (MTMT)

Journals without impact factor:

11. **RÁCZ, A.** (2015): Mire jó a kemometria?
Élet és Tudomány, Vol. 70:(4). pp. 118-120.

12. **RÁCZ, A.**, VASS, A., HÉBERGER, K., FODOR, M. (2015): Q10 tartalmú étrendkiegészítők hatóanyagtartalmának mennyiségi meghatározása FT-NIR spektroszkópiával.
Élelmiszer - Tudomány Technológia, Vol. 69:(2). pp. 25-32.

Book chapters:

1. **A. RÁCZ**, D. BAJUSZ, K. HÉBERGER:
Chapter 8: Cheminformatics/chemometrics in analytical chemistry
In **Chemoinformatics – a textbook**
(Editors: Johann Gasteiger, Thomas Engel)
WILEY, 2017 (in press)

2. D. BAJUSZ, **A. RÁCZ**, K. HÉBERGER**:**
Chapter 30010, title: Chemical data formats, fingerprints and other molecular descriptions for database analysis and searching
In **Comprehensive Medicinal Chemistry III.**
(Editors: Andy Davis, Colin Edge)
ELSEVIER, 2016 (in press)

3. D. BAJUSZ, **A. RÁCZ**, K. HÉBERGER:
Chapter title: Which performance parameters are best suited to assess the predictive ability of models?
In **Advances in QSAR modeling with applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences**
(Editor: Kunal Roy)
SPRINGER, 2017 (in press)

Conferences and Hungarian conference abstracts:

1. **RÁCZ, A.**, FILIP, A., BAJUSZ, D., HÉBERGER, K. (2016): Számításos és vékonyréteg-kromatográfiás lipofilicitási indexek összehasonlítása kemometriai módszerekkel.
KeMoMo–QSAR 2016 szimpózium, Szeged, 2016. május
http://www.chemicro.hu/QSAR/kovetkezo.html

2. **RÁCZ, A.**, HÉBERGER, K., FODOR, M. (2016): Energiaitalok minőségi és mennyiségi elemzése FT-NIR spektroszkópiával.
Aktualitások a táplálkozástudományi kutatásokban – VI. PhD Konferencia, Budapest, 2016. február (http://www.mttt.hu/index.php?content=57)

3. **RÁCZ, A.**, FODOR, M., HÉBERGER, K. (2015): Energiaitalok egy kemometrikus szemével.
KeMoMo–QSAR 2016 szimpózium, Szeged, 2015. május
http://www.chemicro.hu/QSAR/20150514.html

4. **RÁCZ, A.**, VASS, A., HÉBERGER, K., FODOR, M. (2015): Q10 tartalmú étrendkiegészítők minőségellenőrzése FT-NIR módszerrel.
Aktualitások a táplálkozástudományi kutatásokban – V. PhD Konferencia, Budapest, 2015. január. p. 32 (ISBN 978-963-88108-8-5)

5. **RÁCZ, A.**, FODOR, M., HÉBERGER, K. (2014): Változó kiválasztás – avagy legális út regressziós modellek javítására.
KeMoMo–QSAR 2014 szimpózium, Szeged, 2014. május
http://www.chemicro.hu/QSAR/kivonatok19/kivonat1902.html

6. **RÁCZ, A.**, PAPP, N., FODOR, M., HÉBERGER, K. (2014): Antioxidáns kapacitás meghatározási technikák összehasonlítása kemometriai módszerek segítségével. Aktualitások a táplálkozástudományi kutatásokban – IV. PhD Konferencia, Budapest, 2014. január. p. 18 (ISBN 978-963-88108-7-8)

7. **RÁCZ, A.**, ELEK, J., PAPP, G. (2013): Égetett szeszesitalok tömegspektrometriás és infravörös spektroszkópiás vizsgálatának többváltozós elemzése
KeMoMo–QSAR 2014 szimpózium, Szeged, 2013. április

8. **RÁCZ A.**, ELEK J., NEMES Z. (2012): Hogyan segíthet a modern műszeres analitika történelmi kérdések tisztázásában, avagy interdiszciplináris kutatások egy római kori sírkő körül.
XXXV. KEN, Szeged, 2012. október. p. 246 (ISBN: 978-963-315-099-3)

9. **RÁCZ, A.**, ELEK, J., NEMES, Z. (2013): Feltárhat-e rejtett összefüggéseket egy római kori sírkőről röntgenfluoreszcenciás elemanalízis adatok főkomponens-elemzése?
KeMoMo–QSAR 2014 szimpózium, Szeged, 2012. szeptember

10. **RÁCZ, A**., ELEK, J., HÉBERGER K., RAJKÓ, R., LENGYEL, A. (2011): Régészeti leletek XRF vizsgálata és értékelése, avagy mennyi adatra van szükségünk főkomponens elemzéshez
MKE 1. Nemzeti konferencia, Sopron, 2011. május
(http://www.mkenk2011.mke.org.hu/hu/tudomanyos-program.html)

Conferences – International conference abstracts:

1. D. BAJUSZ, **A. RÁCZ**, K. HÉBERGER (2015): Revival of an old debate: Cross- *vs*. External validation in QSAR modeling.
Conferentia Chemometrica 2015, Budapest (Hungary), 2015, September. p. L23 (ISBN: 978-963-7067-31-0)

2. **A. RÁCZ**, D. BAJUSZ, K. HÉBERGER (2015): Large scale statistical comparison of similarity metrics for fingerprint-based calculations.
Conferentia Chemometrica 2015, Budapest (Hungary), 2015, September. p. P04 (ISBN: 978-963-7067-31-0)

3. **A. RÁCZ**, D. BAJUSZ, K. HÉBERGER (2015)**:** *n*-class ROC curves as novel, intuitive tools for method comparison.
Conferentia Chemometrica 2015, Budapest (Hungary), 2015, September. p. P24 (ISBN: 978-963-7067-31-0)

4. K. HÉBERGER, D. BAJUSZ, **A. RÁCZ** (2015): Consistency of QSAR models: correct split of training and test sets, ranking of models and performance parameters. 8th International symposium on computational methods in toxicology and pharmacology integrating internet resources, Chios (Greece), 2015, June. p. 26

5. J. ELEK, A. MEZŐSI, **A. RÁCZ** (2013): Estimation of active ingredient content by multivariate calibration: NIR examination of tablets used in ED treatment.

Conferentia Chemometrica 2013, Sopron (Hungary), 2013, September. p. L26 (ISBN: 978-963-9970-38-0)

6. **A. RÁCZ**, J. ELEK, G. PAPP (2013): Multivariate data analysis of Hungarian spirit drinks' IR spectroscopic data. Conferentia Chemometrica 2013, Sopron (Hungary), 2013, September. p. P22 (ISBN: 978-963-9970-38-0)

7. O.H.J. CHRISTIE, **A. RÁCZ**, J. ELEK, K. HÉBERGER (2012): Object target rotation for principal components analysis: metal content data of hungarian silver coins from the Árpád dynasty.
XIII. Chemometrics in Analytical Chemistry, Budapest (Hungary), 2012, June. p. 72 (ISBN: 978-963-9970-24-3)

8. **A. RÁCZ,** J. ELEK, K. HÉBERGER, R. RAJKÓ, A. LENGYEL (2012): Principal component analysis of an XI-XV century silver coins' XRF dataset and verification by additional multilinear methods. WSC 8, Drakino, Russia, 2012, February.
(http://wsc.chemometrics.ru/wsc8/presentations/?page=2)

9. **A. RÁCZ**, J. ELEK, K. HÉBERGER, R. RAJKÓ, A. LENGYEL (2011): Classification of medieval silver coins using X-ray fluorescent spectroscopy and multivariate data analysis.
Conferentia Chemometrica 2011, Sümeg (Hungary), 2011, September. p. P33 (ISBN: 978-963-9970-15-1)

10. **A. RÁCZ**, J. ELEK, K. HÉBERGER, R. RAJKÓ, A. LENGYEL (2011): Principal Component Analysis of an XI-XV century silver coins' XRF dataset and verification by additional multilinear methods. Conferentia Chemometrica 2011, Sümeg (Hungary), 2011, September. p. L26 (ISBN: 978-963-9970-15-1)