NIR History: Progress in Processing and Evaluation of Spectral Data

K. Kaffka

Corvinus University of Budapest, Budapest, Hungary.
E-mail: karoly.kaffka@uni-corvinus.hu

My presentation is intended to review briefly the history of the role of chemometrics in the progress of NIR technology.

In the early 1960s a team of the Central Food Research Institute (CFRI), Budapest, Hungary) received the task to deal with measurement and control of Quality, in order to get good quality foods from the production lines consistently. This team developed a method, which was called "measurement alternation method" (the word "chemometrics" was not discovered yet). By this method, after measuring some independent physical parameters, the data were used to determine the composition.

Although this measurement alternation method that measured conductivity, density, refractivity, optical activity, viscosity, and other parameters was used successfully for determining the composition of different foodstuffs, such as wine, beer, pickling brine, and butter, usually a separate instrument was needed to measure as many physical parameters that had to be measured, which corresponded to the number of constituents needed, and a separate instrument had to be used for each parameter/constituent. The results were published in 1970 at the IMEKO Congress in Versailles, France. In several cases, due to the texture of foodstuffs, the measurements of physical parameters proved to be extremely difficult.

In the early seventies we heard about a new method, the NIR technology developed in Beltsville, USA at Karl H. Norris' Laboratory. Karl suggested the determination of composition by measuring the transmittance or reflectance parameters of the sample at different wavelengths, characteristic of the constituents.

Optical properties can be measured independently of the consistency of the sample, accurately and non-destructively, and many independent parameters can be measured with one and the same detector and instrument. Limiting the physical parameters to optical properties paradoxically facilitated the widening of the applicability of rapid physical methods in food analysis.

In 1971 at one of his trips in the USA, Dr. Karoly Vas, director of the Central Food Research Institute (CFRI) paid a visit to the Instrumentation Laboratory of the USDA Agricultural Research Centre in Beltsville, and he got acquainted with this new technology, that used the near infrared spectral region of electromagnetic radiation for multicomponent analysis. Recognising the importance and the perspectives of this technology on the one hand, and the similar mentality of the two teams in Beltsville and Budapest on the other, Dr. Vas paved the way for me to study this technology in Beltsville at Karl H. Norris Laboratory in 1973 to 1974. As a follow-up to this visit, the CFRI has sent several other researchers to study this technology for food composition analysis and quality determination.

In the early 1970s, a one-of-a-kind high intensity NIR spectrometer was developed at Beltsville (USA), interfaced with a computer, for food quality measurements. This served as a basic tool for collaborative studies with scientists from all over the world for a long time. In the late1970's the CFRI (Budapest, Hungary) purchased a NEOTEC Model 6350, scanning spectrometer and a TREBOR Model 90B NIR instrument. The Trebor 90 was developed by Trebor Industries, Gaithersburg, MD. It used NIR light-emitting diodes, worked in transmittance mode, and was the first instrument to work well with whole grain. Work began at both places (Beltsville and Budapest) on determining the relationship between optical properties (spectral data) of certain products, and their relationship to quality parameters (composition or the nature of the product).

In the near infrared spectral region the absorption bands are highly overlapping and weakly absorbing, and only improvements in instrumentation and advances in multivariate chemometric data analysis allowed meaningful results to be obtained from the complex spectrum. Thus, the advances in sophisticated chemometric techniques—with the help of which massive amounts of chemical information can be extracted from NIR spectra—greatly contributed to making NIR technology suitable for application in different fields of agriculture and industry.

In the 1970s the knowledge and the technique spread very rapidly. A team headed by W.F. McClure started research at North Carolina State University, a team headed by J.S. Shenk at Pennsylvania State University, another team headed by G.G. Dull in Athens Georgia at the USDA R.B. Russell Agricultural Research Center and a further one at the Kansas State University headed by D.L. Wetzel joined this activity and became schools of NIR technology in the USA. Other research groups entered into NIR research, including P.C. Williams and co-workers in Winnipeg, Canada; myself with a team in Budapest, Hungary; M. Iwamoto in Tsukuba, Japan; P.C. Flinn in Hamilton, Victoria, Australia, G.B. Cornish in Adelaide, Australia., and Wu Xiu Qin in Beijing, P.R. China.

In the early 1980s a world-wide boom in the application of the near infrared reflectance and transmittance technique was experienced, together with the widening of its fields of application. More and more research institutes and universities installed NIR instruments and started projects in the field of NIR spectroscopy. A.M.C. Davies in Norwich, U.K.; I. Murray in Aberdeen, U.K.; D. Bertrand in Nantes, France; R. Biston and P. Dardenne in Libramont, and M.J. Meurens in Louvain, Belgium; K.I. Hildrum, and T. Isaksson in Aas, Norway; W.B. Mroczyk in Poznan, Poland; R. Giangiacomo in Lodi, Italy; M. Moisio and A. Kinnunen in Helsinki, Finland; R. Frankhuizen in Wageningen, the Netherlands; all of these workers involved in food quality determination, joined - with their colleagues - the "family" dealing with research in the field of NIR spectroscopy, and each one became an NIR spectroscopy center for developing and spreading the technology.

The scanning type NIR spectrometers installed in the late 1970s and in the early 1980s were producing a fantastically large amount of spectral data, and the processing and particularly the evaluation of these data began. Fortunately in the early 1980s mathematicians and statisticians—seeing the large amount of meaningful data—found considerable satisfaction in dealing with these data. The science of chemometrics was born, and these new pioneers became the first chemometricians, giving the world inestimable help in processing and evaluating NIR spectral data.

Regarding **pre-treatments** of the NIR spectra—starting in the early 1970s—K.H. Norris and W.R. Hruschka introduced the running mean (the **box-car) smooth,** replacing the spectral value at each wavelength by the means of the values in a wavelength interval surrounding it.

For smoothing they also applied the **Savitzky–Golay principle,** fitting the NIR spectrum in a wavelength interval with a polynomial, using least squares. At the same time these workers also introduced the principle of the **derivative** (the first and second order finite-difference) method of resolving overlapping peaks, and removal of linear baseline shifts. In the early 1980s the **Fourier-transformation** method was introduced to pre-treat NIR spectra. W.F. McClure, A.M.C. Davies and D. Bertrand used this method for different purposes. For smoothing, the NIR spectrum must be transformed to the Fourier-domain. Then by eliminating the high frequency coefficients, and retransforming this modified Fourier spectrum into a NIR spectrum a smoothed spectrum can be achieved. During calibration (model) development, sample composition can also be correlated directly to the Fourier coefficients. Fourier transformation can correct for particle size effect, it minimises multicollinearity, and it is easy to generate derivatives. The next very useful step forward in spectral data pre-treatments was that of **multiplicative scatter correction (MSC)**, which involves regressing the spectral values onto the corresponding values of the average spectrum. H. Martens and co-workers developed the MSC concept in the early 1980s to eliminate the optical interference. D. Bertrand and his colleagues elaborated a method for correction of the spectral deformation caused by the effect of sample particle size variation. P.L. Geladi and his team confirmed the linearisation and scatter correction effects of MSC.

In the early 1990s improvements were made on MSC by H. Martens and E. Stark, who introduced **extended multiplicative signal correction (EMSC)** to eliminate additive and multiplicative effects, caused by varying particle size and optical pathlength. In 1992 T. Isaksson and B. Kowalsky introduced a further development of the MSC; the **piece-wise multiplicative scatter correction (PMSC)**. Other pre-treatments included different kinds of **normalisations**. Through normalisation, by subtraction, the spectral value of a spectrum at a single wavelength (at the reference wavelength) is subtracted from all the spectral values of the whole spectrum. Similarly at normalisation by division, the spectral value of a spectrum at a single wavelength is divided by all of the spectral values of the whole spectrum. Other mathematical transformations introduced during the mid 1980s included **standard normal variate (SNV)** and the **Norris pathlength correction (NPC)** for removing slope, base-line and pathlength variation. These concepts were introduced by I. A. Cowe, and K. H. Norris, and further developed by R. J. Barnes.

Regarding **evaluation** of the pre-treated NIR spectra, the task was to create a model (calibration) that would effectively describe the relationship between composition and spectral data of the investigated product, or determine (recognise) its nature. In the early days of the modern era of NIR technology (starting in the early 1970s) the **multiple linear regression (MLR)** method was preferred, using the spectral data measured at several specific wavelengths. The dependent variable (concentration of a component) was regressed against the independent variables (spectral data) by least squares fitting. Using more specific wavelengths, and more terms in the regression equation, the danger of overfitting, and the problems caused by the effect of multi-collinearity increased. In the mid 1970s K. H. Norris introduced a multiterm regression, in which each independent variable is a quotient of first or second derivatives, achieving better results.

In the early 1980s revolutionary progress was experienced in NIR technology, in the evaluation of the NIR spectra. I.A. Cowe and J.W. McNicol in Scotland and a Scandinavian chemometric school introduced the concept of **principal component regression (PCR)** and **partial least squares (PLS)** regression. These techniques used all of the spectral data (the full spectrum) and compressed these data into a small number of components as linear functions of the original spectral data. These compression methods solved the collinearity problem, and more stable regression equations and predictions were obtained. They were also useful for better understanding and interpretation of the data, and they could be used to detect outliers, overfittings, model errors, etc. The methods, their modified, extended, enhanced, combined versions and their various application possibilities were published in different ways (books, articles, users' manuals) and at different places (congresses, conferences, symposia, courses etc.) by the members of the Scandinavian school; by H. Martens, T. Naes, K.I. Hildrum, T. Isaksson, S. Wold, H. Wold, K.H. Esbensen, C. Borggaard and P.L.Geladi to mention just a few of them. The software for performing all these possible operations was already available, thus the method spread very rapidly. The collinearity problem was solved with PCR and PLS by projecting the spectral data from their original coordinate system (onto a subspace), and regressing the composition data onto the new coordinate values (scores) of this projection, thus concentrating the spectral values to their most dominant (at PCR) or to their most relevant (at PLS) dimensions. For **detection of outliers** in the chemical (reference) data H. Martens and T. Naes introduced in the late 1980s the "y-residuals" method, for detection of outliers in spectral data, R.D. Cook and S. Weisberg introduced in the early 1980s the "leverage" method (leverage is closely related to Mahalanobis distance). For deciding whether an outlier plays an important role in calibration or not the influence plot was introduced by T. Naes. T. Isaksson and T. Naes drew our attention to the problem of **selecting "good" samples** for calibration in 1990, and presented some strategies for solving the problem.

Parallel with the activity of the Scandinavian chemometric school, in the middle of the 1980s several teams in Europe and in the United States, mostly collaborating with the members of the Scandinavian school, also started to deal with chemometrics and its application in the NIR technology. A team around D. Bertrand, P. Robert and M.M.F. Devaux in France; experts around A.M.C. Davies, I.A. Cowe, B.G. Osborne, J.W. McNicol, T. Fearn, I. Murray, C.N.G. Scotter and G. Downey in UK and Ireland; experts around R. Biston, P. Dardenne, M. Meurens in Belgium; scientists around H.W. Siesler and Ch. Paul in Germany, a team in Beltsville (USA) K.H. Norris, W.R. Hruschka, D. Massie, S.R. Delwiche, J.B. Reeves, III and D. Slaughter; experts around W.F. McClure, E. Stark, H. Mark, P.R. Giffiths, D.E. Honigs, J.S. Shenk, M.O. Westerhaus, G. Dull, F.E. Barton, B. Kowalsky, J. Workman and D. Wetzel, all participated in development and applications.

Scientists, who just applied the new methods also contributed in the progress as they discovered the new difficulties and specified new tasks for the chemometricians. Using PCR or PLS all the most important problems arising in model building (in calibration) can be solved except the multivariate non-linearity. C.E. Miller revealed the spectroscopic basis for non-linearity in the early 1990s. For linearisation of the relationship between concentration and spectral data of transmission (T) or reflection (R), the log function (log1/*R* or log1/*T*) was applied, derived from Beer's Law, and sometimes also Kubelka–Munck transformation was used. Both of these transformations proved to be very useful, but as they are univariant transformations, they do not always improve multivariate non-linearity. Other useful linearisation methods were the **locally weighted regression (LWR)** and the **artificial neural network (ANN)**. LWR was introduced in 1990 by Naes and co-workers, based on the idea of local linearity and it is quite a straightforward extension of PCR. A general introduction to ANN was given by Y.H. Pao in 1989. The use of ANN in NIR technology was introduced by C. Borggaard and H.H. Thodberg in 1992 and T. Naes and co-workers in 1993.

As well as the tried and trusted role of chemometrics in quantitative NIR spectroscopy (quantitative analysis), chemometricians have also developed several useful methods for the purpose of classifying samples. All of these methods described the same general concept, which is to determine the actual nature of the sample, rather than determination of its composition. Thus, these methods are called, **qualitative methods**. With the growth of interest in qualitative analysis the principal component approach to classification has become very popular since the early 1980s. P. Robert, H. Mark, T. Naes and T. Isaksson introduced the **principal component analysis (PCA)** method for cluster analysis as a very simple visual technique, using the NIR spectra of the samples. The goal was to recognise and identify samples, or cluster them in groups and subgroups without using any prior information (unsupervised classification). Another very popular classification method was **discriminant analysis (DA)**. The goal here was to build classification rules (models) for a number of pre-specified groups. These rules can be later used for allocating unknown samples to the most probable group (supervised classification). Discriminant analysis is a calibration method, where the quantity to be calibrated for is a categorical group variable. In the early 1980s J. Rose reported an algorithm to identify pharmaceutical materials based on their NIR spectra. In 1985 H. Mark and D. Tunnell designed an algorithm for this purpose based on **Mahalanobis distances**. The Mahalanobis distance is similar to the Euclidean one, the difference is that the principle component directions are weighted according to the variability along them. T. Naes and co-workers discussed this method in detail in 1990. A **modified Mahalanobis distance** was introduced by T. Naes and T. Isaksson in 1992. The **soft independent modelling of class analogies (SIMCA)** were elaborated and introduced by S. Wold in 1983. This method also uses the calculation of principal components, where a different set of principal components is calculated for each different material for which the model is being created, thereby classifying unknown samples using the residual variance. In 1990 N.K. Shah and P.J. Gemperline further developed the SIMCA method by combining the residual variance and the Mahalanobis distance.

Another qualitative method was the **polar qualification system (PQS)** where a quality point was defined as the center of the spectrum of the investigated sample, represented in a polar coordinate system. PQS was introduced by K. Kaffka and L. Gyarmati in 1990. In 2004 Zs. Seregely and Sz. Velkei introduced a new **sample (spectrum) recognition tool (SRT).** The SRT is a new non-linear mathematical statistical method, with the help of which an unknown sample can be classified (recognised) extremely rapidly into the class to which it belongs. A model is created for each class by a training, and by a learning procedure. By this procedure the density distribution function of the spectral values of the training set is measured at each wavelength, and for each class to be determined, then these functions will be handled as probability density functions, and will serve as the model of the classes. The unknown sample (the spectrum of this sample) will be classified to the particular class in which the sum of the log probability values at each wavelength for this spectrum is the maximum. The result is given in a classification matrix where the values (the percentage of the correct recognitions) are shown by columns in vertical direction.

Coming to the end of my presentation I have to mention the names of T. Naes, T. Isaksson, T. Fearn and T. Davies who have been writing and editing chemometric columns in *NIR news* and in *Spectroscopy Europe* since the early 1990s. These articles have significantly helped the readers to understand, and to apply chemometric methods in near infrared spectroscopy. I add also the name of W.F. McClure, who has compiled a bibliography (CBIBL) of NIR literature references.

Summarising my talk; my intention was not to introduce the chemometric methods themselves, but to give you a short history of the influence of chemometric methods on near infrared spectroscopy. These methods have resulted in fantastic progress in the NIR technology, as a result of the interaction between users and chemometricians. The talk was not intended to be exhaustive, as I could not mention every single method, and I apologise to those scientists whose names were not mentioned in my presentation.