Chemometrics and Intelligent Laboratory Systems 45 (1999) 55–63

Chemometrics and
intelligent
laboratory systems

Making money with chemometrics

Mary Beth Seasholtz [*]

Dow Chemical, 1897 Building, Midland, MI 48667, USA

Received 23 October 1997; accepted 7 January 1998

Abstract

Chemometricians have been formally employed at Dow Chemical since 1988. In that time, chemometric methods have been applied in a number of analytical chemistry applications. These have resulted in making money for the company in a variety of ways, and several recent case studies are presented. These applications have been positive for the company in terms of (1) better process control, (2) faster verification of raw material identification and quality, and (3) faster analysis of wastewater. The analytical methods used are NIR and NMR spectroscopy. The chemometric methods include pattern recognition and multivariate calibration. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Chemometrics; Industrial applications; Multivariate calibration; Pattern recognition; NIR spectroscopy; NMR spectroscopy; Process analysis

Contents

[*] E-mail: mseasholtz@dow.com

PII: S0169-7439(98)00089-6

1. Introduction

The use of chemometrics in analytical chemistry has increased dramatically over the last 7 years at Dow Chemical. The formal effort to incorporate this technology began in 1988, with the hiring of Ken Beebe into the process analysis group within the Analytical Sciences Laboratory. The growth continued with the hiring of Randy Pell in 1990, and finally with the hiring of the author in 1992. In 1993, Ken opted for management rather than staying in the technical arena, leaving two formally trained chemometricians.

A team approach is taken for problem solving in order to take advantage of the expertise of various people. Depending on the venue of application, the team is obviously different. The approach taken in process analytical chemistry is depicted in Fig. 1. When a measurement need has been defined, the chemometrician works in conjunction with an analytical chemist and experts in both process engineering and process chemistry.

Other than the measurement itself, there are a multitude of additional issues which are considered. These include the hardware, software, communications, and the ease of cloning the technology. The analytical equipment must be located in a safe environment where it will not be damaged. Further, it must itself not be a danger to the surroundings (such as providing a spark source). The software must be developed so that the instrument can operate unattended, as well as have error checking to signal when the instrument is not performing correctly. Automatically communicating the results from the instrument to the process control instrument is also critical, and not always obvious to accomplish. Finally, Dow Chemical has manufacturing facilities located around the world. Therefore, when an application is successful at one plant, it is implemented at all locations which produce the same product. Therefore, the ease of duplicating the technology is considered in the development stage. The obvious issue here with respect to chemometrics is the ease of transferring models from one instrument to another.

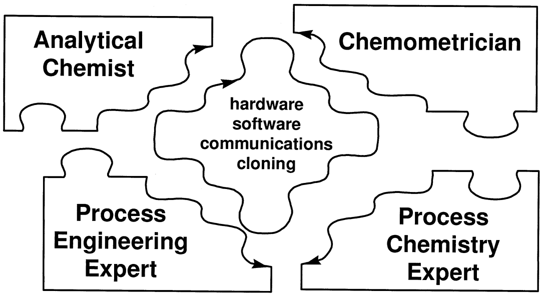

Fig. 1. Problem solving in process analytical chemistry in a team approach.

This paper documents four applications involving the use of chemometrics which have made money for Dow Chemical within the past 3 years. Two of the four are traditional on-line process analytical applications, while the other two are laboratory applications. Three of the four involve the use of FTNIR spectroscopy, and the fourth uses ^{1}H NMR spectroscopy. Three examples are calibration applications, while a fourth demonstrates the use of pattern recognition. The descriptions given below outline the application, discuss the chemometric aspect of the project, and summarize how this is making money for the company.

2. Example 1: identification of raw materials using NIR spectroscopy

When raw materials are delivered to a manufacturing site, there are two options upon delivery. Either the identification and quality of the material is believed to be acceptable, or analyses can be made to verify this before unloading begins. Traditional wet chemical analyses can be time consuming, and so the driver of the truck, train or boat may wait hours before unloading. For one particular plant, blind acceptance was not an option, but the wet chemical methods were laborious and too time consuming. Therefore, an effort was undertaken to develop a faster verification method.

The use of NIR spectroscopy has found a home in industrial settings for a variety of reasons. In this case, it was considered because of the ease of making a measurement of liquid samples. The hand held transmission probe is simply immersed into the liquid while the spectrum is acquired. Further, it was observed that the NIR spectra of the over 50 raw materials were visibly different to the eye. It was hypothesized that the identification could be automated with a classification routine.

One of the simplest classification methods is K-nearest neighbor (KNN) [1]. The identity of an unknown sample is assigned to be the identity of the samples from a library which are located close to it in measurement space. From an implementation point of view this is an attractive method. All that needs to be maintained is the library of samples with known identity. However, there is a serious (and well known) drawback of KNN, demonstrated in Fig. 2. Here is a simple example with two measurements and two classes. The identification of unknown X as an A is trivial. But, the identification of unknown Z as an A is not satisfactory, because it is located far from its nearest neighbors. With respect to the raw materials identification, this means if a low quality or a completely unknown material is delivered, it is always classified, with no warning that anything is wrong.

There are of course other classification methods such as SIMCA [2] which do not have this drawback. However, they are typically more difficult to implement in a manufacturing setting with personnel untrained in analytical chemistry, much less chemometics. For example, if samples are added to the library, SIMCA models must be reconstructed. Not only does this take time, but unsatisfactory results may be obtained due to the users unfamiliarity with the methodology. With KNN, no work needs to be done if the library is expanded. Therefore, the decision was made to implement the KNN algorithm, with an additional calculation to verify the reliability of the classification.

The 'goodness criterion', G, is calculated in the form of a t-test. First, the distance from the unknown sample to the nearest neighbor ('d' in Fig. 2 for unknown Z) is determined. Then, the nearest neighbor distances for the library samples in the identified class (class A in Fig. 2) are calculated. Finally, the average (m) and standard deviation (s) of the library values are determined. The goodness criterion, G, is calculated to be $G = (d - m)/(s)$. If the difference between d and m is large relative to s, the goodness criterion is large, and the classification of the unknown is questionable. Before deploying the system, an optimal cutoff for G is determined by R&D personnel using the initial installation of the library (typical values range from 4 to 10).

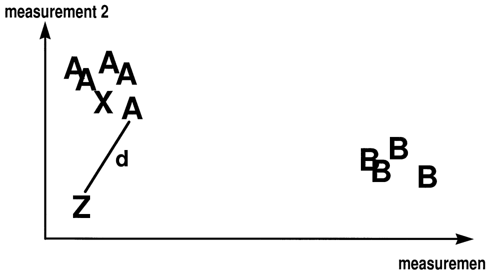

Fig. 2. KNN classification with two measurements. Classes A and B are in the library used to classify unknowns X and Z.

In practice in the plant, if a large G is observed, the probe is cleaned and the measurement is repeated. If it is still large, the wet chemical test is performed to confirm the identification and quality of the material. This information is then used to decide to either accept or decline the delivery. If G is smaller than the cutoff, the delivery is accepted with no delay.

How do we make money? For this application, money is made in several ways. First, the delivery process is much faster. The driver does not wait until a lengthy analysis is completed before transferring the material into the Dow facilities. This saves money in terms of time required to obtain the raw materials as well as the cost of a vehicle idling on Dow property (e.g., mooring costs for ships). Further, the implications of using wrong or low quality raw materials are huge. Low quality product might be produced which will be disposed. Or, customers will be dissatisfied if they receive the unacceptable product. And, unsatisfied customers are not typically repeat customers, so there will be a loss in sales.

3. Example 2: caustic stream analysis by NIR

Caustic/Salt systems are prevalent in the chemical industry. In incinerators, caustic (NaOH) and HCl are used to convert components such as Cl_2 and SO_2 to salts, NaCl and Na_2SO_4, respectively. Evaporators reduce the water in a caustic stream to obtain high concentrations of NaOH in the liquid phase and precipitate the NaCl which are sold as products. Finally, process scrubbers use caustic streams to con-

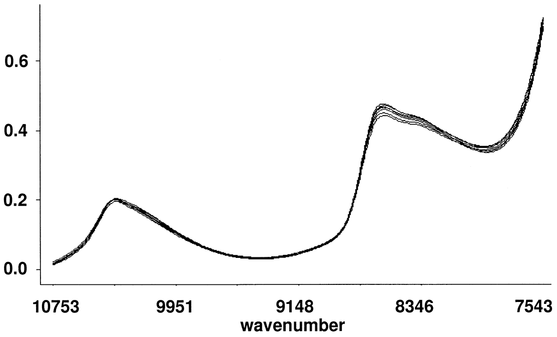

Fig. 3. NIR absorbance spectra of samples with varying amounts of NaOH, salt in water, 50–70°C.

vert H_2S in the presence of CO_2 to Na_2CO_3 and Na_2S.

For a particular application at Dow involving a caustic stream, there was a need for an on-line measurement of the NaOH and salt concentrations in water over a temperature range of 50–70°C. Various methods were considered, including physical measurements (refractive index, speed of sound, conductivity and density), titration, FIA, thermal neutron/ γ-ray capture, NMR and NIR. Under consideration was not only the applicability of the method for the analysis, but also the cost of the instrumentation, the calibration and the maintenance of the on-line analyzer over its many years of deployment.

NIR spectroscopy was chosen based on the demonstrations in the literature [3,4] and the availability of a NIR transmission probe such as the one shown in Ref. [5]. NIR absorbance spectra collected over 50–70°C, spanning the caustic and salt concentration ranges of interest are shown in Fig. 3. The spectra are dominated by the overtones and combinations of the OH bends and stretches of the water, with perturbations due to the varying temperature, and the presence of the caustic and salt. The PLS calibration method was used to develop a predictive model for the caustic and salt concentrations. A three factor model was constructed using first derivative data. The scores from this model are shown in Fig. 4, where each point represents one spectrum. The points form lines in the scores space. Close examination of the data set reveals that each line contains the spectra from an individual calibration sample, and the variation along the line is due to the varying temperature of the sample. This is a good demonstration of the ability of the PLS method to implicitly model a source of variation other than the component of interest.

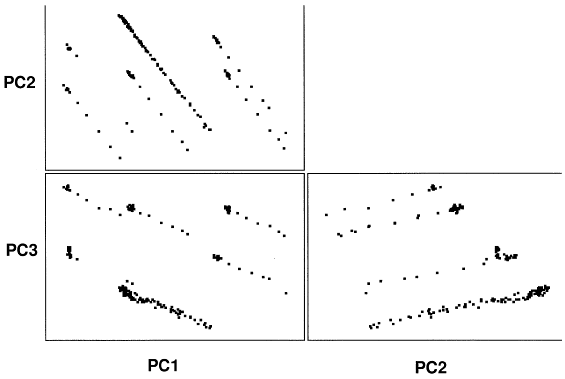

Fig. 4. Scores plot from the PLS calibration of the caustic/ salt/water system.

This analyzer was the first on-line analytical tool installed in this plant. As such, the chemical compositions in the plant had never been monitored in any way other than periodic grab sample analyses (1 per 8-h shift). Fig. 5 shows a portion of the data obtained demonstrating a periodic change in concentration. This decrease in concentration is counter-productive, and, if eliminated could improve the output from the plant significantly. However, until this analyzer was installed, the plant personnel did not know the periodic decrease took place.

How do we make money? The main advantage of the NIR on-line determination of the caustic and salt concentration is demonstrated in Fig. 6. The optimum set point is to have the concentration as high as possible. However, if the concentration is too high, there will be process problems. For example, salt may

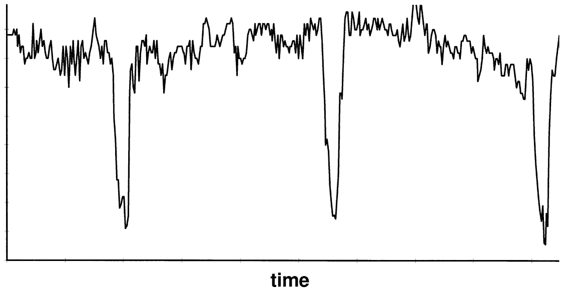

Fig. 5. On-line predicted concentrations from the NIR show periodic behavior.

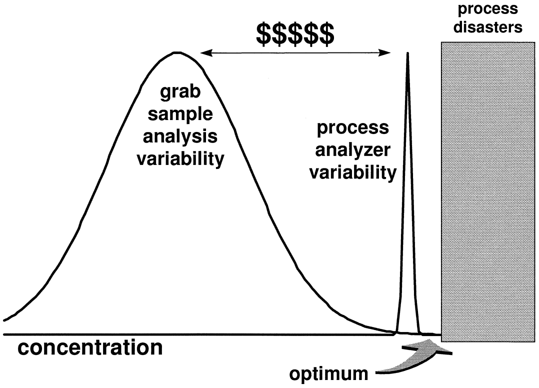
$$$$$
process disasters
grab sample analysis variability
process analyzer variability
concentration
optimum

Fig. 6. Comparison of process control via grab–sample analyses and on-line NIR analysis for the caustic/salt/water stream.

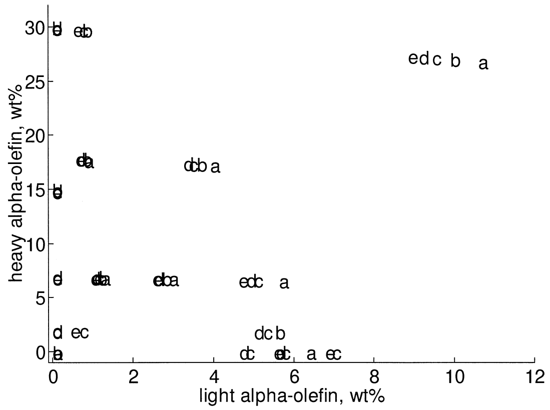

Fig. 7. Experimental points for the calibration of the NIR for polyethylene service. (a) 25°C, (b) 60°C, (c) 85°C, (d) 110°C, (e) 140°C.

come out of solution in a vessel not designed to handle solids, or an unwanted side reaction will begin to dominate. In the past the process was monitored based on the laboratory analysis of grab samples which has large variability. Therefore, the set point for the process was at a lower concentration to stay away from the potential disasters. The on-line analyzer has much improved precision (demonstrated by the narrower histogram in Fig. 6) and so it is possible to operate at a higher concentration while still confidently staying out of trouble. Therefore, money is made because of higher production rates, which lead to a higher profit.

4. Example 3: compensating for temperature effects in multivariate calibration: PLS fails, CLS shines for ONE application

The use of mid IR spectroscopy for process control in polyethylene processes via analysis of the feed stream is discussed in reference [6]. The use of NIR was investigated due to the advantage of being able to use fiber optic probes, which would minimize sampling difficulties as well as allow the analyzer to be located remotely from the process. The goal of the work discussed here was to calibrate the NIR to predict the concentrations of two α-olefins. One major consideration for the calibration experiments is the temperature of the process at the sampling point is known to vary from 25 to 140°C. It is known that the NIR spectra are sensitive to temperature variations, and so spectra of the calibration samples were collected over various temperatures. The experimental points for the calibration are shown in Fig. 7. The letters represent the five different temperatures of the samples. The spectra (baseline corrected) in the region of interest for the calibration spectra are shown in Fig. 8. Thorough examination of the data show that the temperature effect is quite significant. For example, the top five bands represent samples of similar concentrations with varying temperatures.

It is known that spectroscopy is fundamentally a w/v measurement because of the fixed pathlength. In these experiments the density of the solution is expected to vary significantly, and so it is not advisable to model in wt.% units. In the discussion below, the

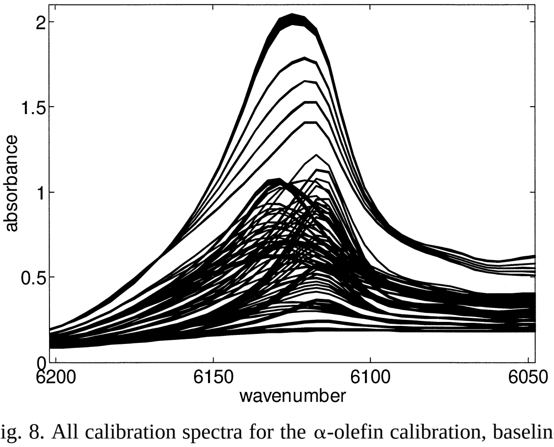

Fig. 8. All calibration spectra for the α-olefin calibration, baseline corrected at 7650 cm^{-1}.

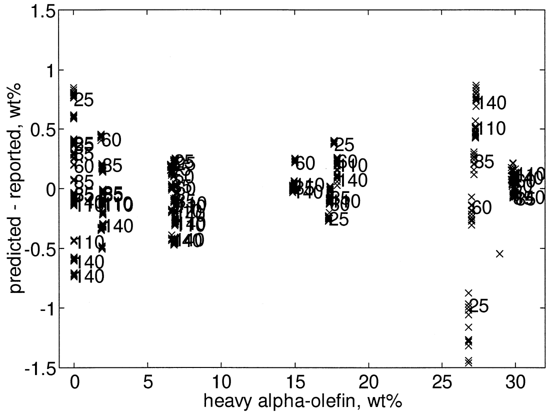

Fig. 9. PLS concentration residuals, heavy olefin, $r = 4$, 99.98% of the spectral variance described. For clarity, the temperatures for only a portion of the samples are labeled.

modeling is performed in w/v units and then converted to wt.% using the externally determined solution density.

The PLS concentration residuals for the light olefin are acceptable relative to the known errors in the reported concentrations (±0.5 wt.%). The concentration residuals for the heavy olefin shown in Fig. 9 are not satisfying for two reasons. First, the errors are significantly larger than the errors in the reported concentrations (±0.5 wt.%). Second, the residuals are structured as a function of temperature. For example, the $T = 140°C$ residuals are negative at low concentration, and positive at high concentration. Similarly, the $T = 25°C$ residuals are positive at low concentrations and negative at high concentration. Five individual PLS models from the data collected at the five discrete temperatures were then constructed. The concentration residuals are unstructured and are consistent with the errors in the reported concentrations. However, it is impossible to implement in practice, because the stream can be at any temperature in between 25 and 140°C. The comparison of these two approaches suggests that the traditional PLS model is not able to account for the variation of the spectra with respect to temperature. Therefore, the temperature was added as an additional measurement variable. It was scaled from 0 (25°C) to 1 (140°C) so as to match the scale of the spectra. However, this did not improve the model. Through examination of the raw data, the temperature is known to have a large influence on the spectra, and therefore adding it explicitly does not provide any additional information to the model. The conclusion is that the PLS is not able to model the heavy olefin while implicitly accounting for the light olefin and temperature.

Another approach for this calibration is classical least squares (CLS) modeling, [7] which assumes the Beer's Law model. Using this model, pure spectra were estimated at each of the five discrete temperatures monitored in the calibration (see Fig. 10). The light olefin has a higher absorptivity than the heavy olefin. Further, the light olefin has much more sensitivity to the temperature than the heavy one. As the temperature increases, the peak decreases in intensity and becomes broader, which is consistent with the theory of increased rotational transitions at higher temperatures for small molecules. The concentration residuals for the heavy olefin using the CLS model are shown in Fig. 11. They are within ±0.5 wt.%, indicating this model is adequate. To implement the CLS approach generally, the temperature of the stream must be known. Then, the pure spectra at a given temperature are estimated by interpolation of the pure spectra in Fig. 10.

This calibration problem is an interesting example where the PLS was not able to implicitly model the effect of one component (temperature). The CLS method on the other hand, explicitly removes the temperature variance before predicting the concentra-

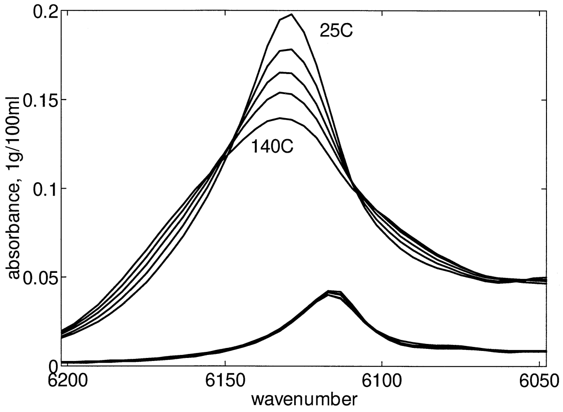

Fig. 10. Estimated pure spectra of the light olefin (top group) and the heavy olefin (bottom group).

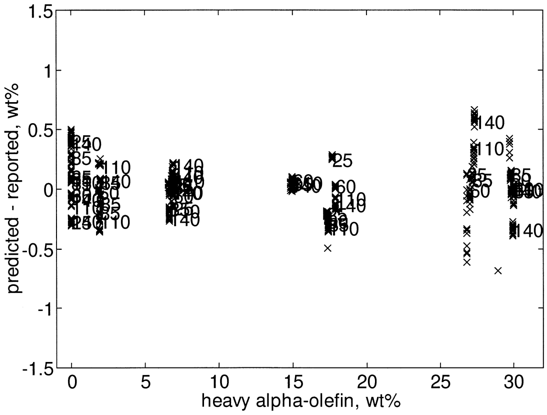

Fig. 11. CLS concentration residuals, heavy olefin. For clarity, the temperatures for only a portion of the samples are labeled.

tion. The reason this was a problem for only the heavy olefin is that it has a relatively small absorptivity and changes in this component are overshadowed by the temperature effect on the light olefin.

How do we make money? The NIR analyzer is used on-line for process control. The real time analysis and excellent precision of the instrument helps the plant make consistent product. The accuracy of the NIR helps the plant more quickly and accurately transition from one product to another. Both features reduce the amount of offgrade produced. Offgrade is either sold at a lower price or is disposed, both of which cost money as compared to selling prime product.

5. Example 4: analysis of organics in industrial wastewater using ^{1}H NMR

Industrial wastewater has significantly more variability than municipal wastewater due to the chemically diverse waste streams that must be treated. Further, relatively quick intervention must be taken before the biomass is destroyed from an undesirable component. If the biomass is damaged, it can take days to weeks for it to recover, requiring reduced production schedules throughout the site.

The 5-day biological oxygen demand, BOD$_5$, is one important measure of the condition of the wastewater. It is an empirical test to determine the relative oxygen requirements of wastewater. The test measures in part the molecular oxygen utilized during a specified period (5 days) for the biochemical degradation of organic material [8]. However, because it takes 5 days for a result to be obtained, the goal was to develop an analytical method which can predict the BOD$_5$ value in substantially less than 5 days.

^{1}H NMR spectroscopy can be used to analyze low level (ppm) organics in water. It has a 15-min response time for most samples and requires no sample preparation. The NMR spectrum of the low level organics is obtained using water suppression techniques to suppress the large water peak. The NMR fingerprints of various chemical plants are substantially different, demonstrating the sensitivity of this technique to varying components in the wastewater. Therefore, it is an ideal candidate for modeling the BOD$_5$.

A calibration set of 24 months of data was assembled for determining the model. Due to the seasonal nature of the manufacturing as well as the biomass behavior, it was important to have data over a long period of time. The calibration spectra are shown in Fig. 12. Cross validation was used to select the PLS model. The standard error of prediction (SEP) as a function of the number of factors is not ideal, with a minimum at one factor. Evaluating the magnitude of the SEP revealed that the one factor PLS model was not better than predicting all samples to be equal to the average BOD$_5$! Clearly, there was a problem with this model. Closer inspection of the data reveal that the NMR peaks vary in location. This variation is not due to instrumental instabilities, but rather from the chemical makeup of the sample. This is demonstrated by examining the propionate signal spiked into clean water and wastewater (see Fig. 13). The peaks at 2.4 ppm are significantly shifted.

There has been discussion in the literature about how the multivariate methods such as PLS and PCR are unequipped to model data with shifts such as this, and some solutions have been proposed [9]. One approach is to reduce the resolution of the analysis, so as to minimize the peak shift. An extreme case of that idea was employed here. The integrals of seven peaks were calculated and used for modeling. These seven peaks were thought to be important for BOD$_5$ based on the knowledge of the NMR spectra of wastewater. Using variable selection to choose from the linear,

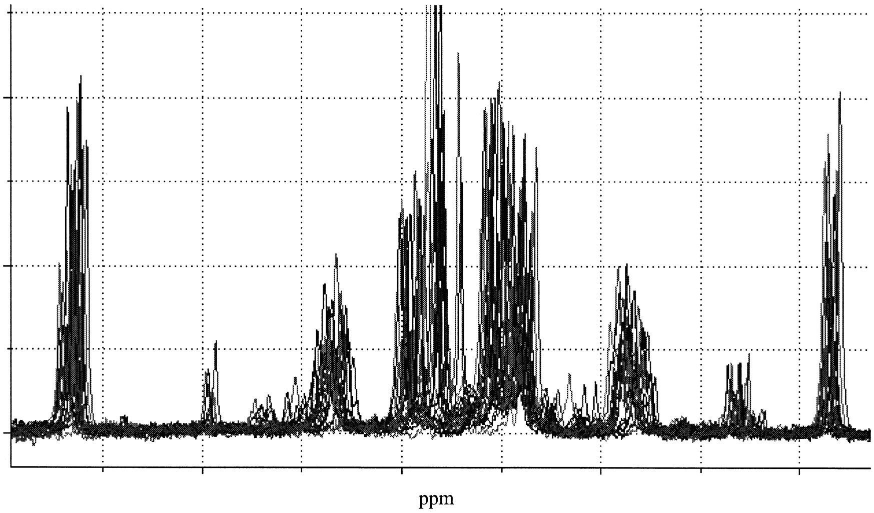

Fig. 12. NMR calibration data for BOD_5 modeling.

squared and 2-variable cross terms, a two variable model was developed. In contrast to the PLS model of the full spectra, this model is a good model for BOD_5. This model was used to predict subsequent samples over a 1-month period. The favorable comparison of the NMR predictions to the laboratory BOD_5 values is shown in Fig. 14.

How do we make money? Using the NMR for evaluating the wastewater, faster intervention can be taken if a problem arises (15 min vs. 5 days). And,

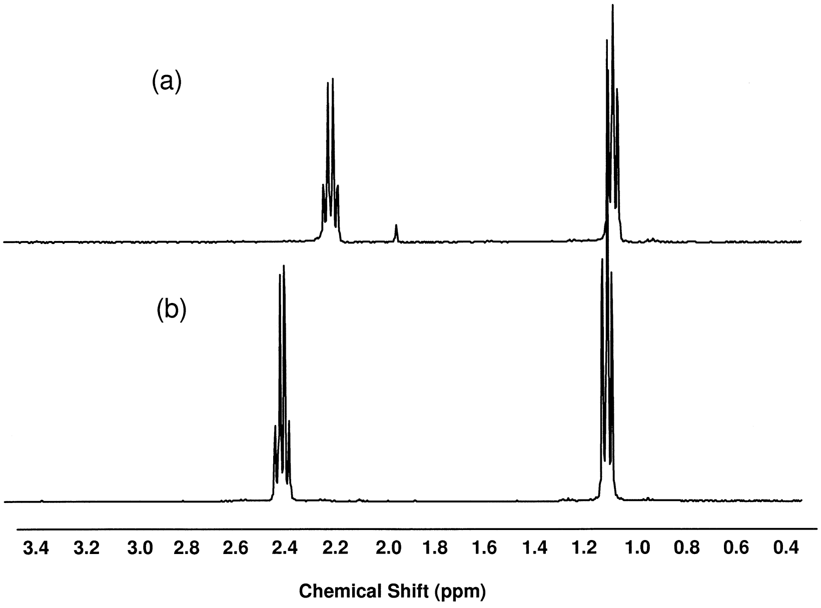

Fig. 13. ^1H NMR spectra. (a) Wastewater spiked with propionate and (b) clean water spiked with propionate.

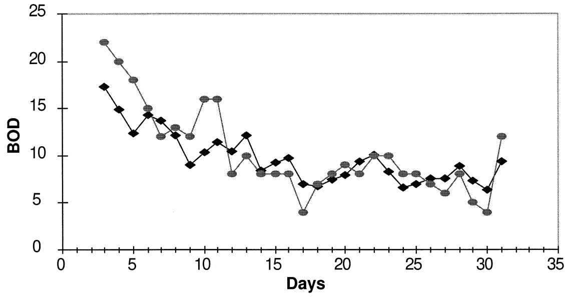

Fig. 14. Comparison of BOD5 (——•——) to the NMR model (——♦——).

making sure our wastewater is clean is the responsible thing to do.

6. Conclusions

In general, standard chemometric methods have been found to be extremely useful in industry. However, the selection of chemometric technology must match the need. And, the need is defined by the application as a whole, not just the data analysis challenges. One technical conclusion is that while PLS is a powerful method, it does not solve every calibration problem. Two of the three calibration examples discussed here required and alternate analysis method.

Acknowledgements

The author wants to acknowledge the colleagues involved in this work (in alphabetical order): Dave Albers, Ken Beebe, Bill Bender, Paul Chauvel, Claire Conboy, David Gay, Rich Harner, Barbara Kirsch, Del Lafevor, Bryant LaFreniere, Ray Mangold, Randy Pell, Steve Vien and Deniz Yalvac.

References

[1] B.R. Kowalski, C.F. Bender, J. Am. Chem. Soc. 94 (1972) 5632.
[2] S. Wold, Pattern Recognition 8 (1976) 127–139.
[3] E. Watson, E.H. Baughman, Spectroscopy 2 (1984) 44–48.
[4] K. Phelan, C. Barlow, J. Kelly, T. Jinguji, J. Callis, Anal. Chem. 61 (1989) 1419–1424.
[5] US Patent 5,988,155, R. Harner, C. Myers, 1991.
[6] US Patent 5,151,474, Lange, Denton, Weller, Chauvel, Farquharson, Ruhl, Winter, 1994.
[7] D. Haaland, E. Thomas, Anal. Chem. 60 (1988) 1193–1202.
[8] A. Eaton, L. Clesceri, A. Greenberg (Eds.), Standard Methods for the Examination of Water and Wastewater, 19th edn., prepared and published jointly by the American Public Health Association (Washington, DC), American Water Works Association, and the Water Environment Federation, 1995.
[9] J. Vogels, A. Tas, J. Venekamp, J. van der Greef, J. Chemom. 10 (1996) 425–438.