Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 80 (2006) 236 – 242

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

A NIR data set is the object of a chemometric contest at '*Chimiométrie 2004*'

Pierre Dardenne [a,*], Juan Antonio Fernández Pierna [b]

[a] *Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural Products Department, Chaussée de Namur no 24, 5030 Gembloux, Belgium*
[b] *Collaborateur Scientifique F.N.R.S. Unité de Statistique et Informatique, Faculté Universitaire des Sciences Agronomiques, Avenue de la Faculté 8, B-5030 Gembloux, Belgium*

Received 9 March 2005; received in revised form 13 May 2005; accepted 21 June 2005
Available online 15 September 2005

Abstract

The organisation committee of the symposium *'Chimiométrie 2004'* (http://www.chimiometrie.org/) held in Paris (30th November and 1st December) proposed on their web site a NIR data set. This data set contained 194 spectra of minced meat taken on an Infratec-Tecator instrument between 850 and 1050 nm. The participants were asked to present during the conference their own approaches to calibrate and predict two independent and blind test sets. The committee received nine answers and this paper summarizes the different ways the data were treated by the participants and the proposed approach of the authors. The conference session was quite interesting and considered valuable for a publication by the committee and the participants. The prediction performances expressed by the RMSEP vary within a ratio of 10 between the extremes values. Due to the non-linearity, methods based on classification gave the best prediction models. A simple ANN model gave the best results.

Keywords: NIRS; Near infrared; PLS; ANN; Classification; Meat

1. Introduction

During conferences concerning chemometrics there are generally many talks about theory, mathematics and sophisticated new methods. Practical applications can be covered only by their author(s) and the results always look promising but often optimistic. For this reason, the idea was to submit the same data set to different chemometricians and compare the approaches of the data treatments and calibrations. Four months before the conference the data were available on the web site. Three data sets were included: a calibration data set and two test sets. The calibration data included 194 spectra of minced meat taken on an Infratec-Tecator (nowadays Foss Analytical AB, DK) instrument between 850 and 1050 nm and the two test sets consisted of 15 spectra each. The participants knew the spectral region and the instruments but the product and the constituent to be predicted were not told. The instructions were simple: predict as accurately as possible the two test sets of 15 blind spectra using any kind of method and send to the jury a text file or a slide presentation with the proposed methodology and the predicted results. Among the nine answers, three proposed only linear methods (MLR and/or PLS), four used classification algorithms ("visually", CART, LWR and MI+kNN) and two applied ANN. The RMSEP of the fat content varied from 3.26 to 35.5 for the first test set and from 0.72 to 8.44 for the second test set. These results themselves remind us that a calibration modelling is not obvious.

2. Material and sample set selection

The supplied data set was chosen for different reasons: i) it is in the public domain and available on the following web site http://lib.stat.cmu.edu/datasets/tecator, ii) the file is

* Corresponding author.
E-mail addresses: dardenne@cra.wallonie.be (P. Dardenne), fernandez@cra.wallonie.be (J.A. Fernández Pierna).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.06.013

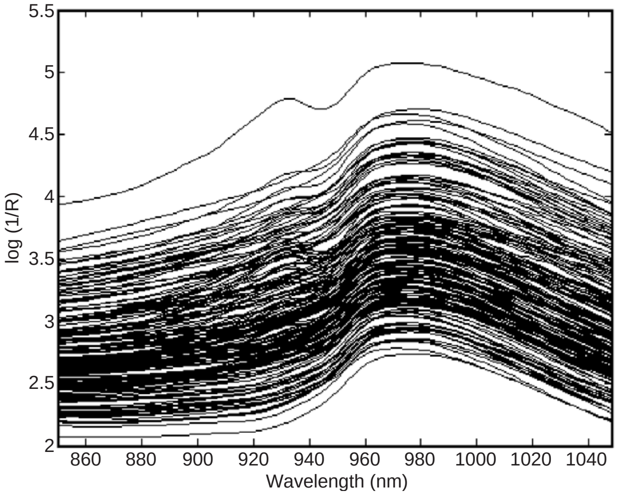

Fig. 1. Spectra of the calibration set.

quite small and easy to handle using any chemometric package and iii) the relationship between the analyte (fat content) and the spectra presented an obvious non-linearity which is more challenging to fit.

The matrix consisted on minced meat samples scanned in transmission (18 mm path length) on a Foss-Infratec spectrometer. The matrix is quite simple. Basically, meat is a mixture of water, protein and fat. The variation in protein is quite small regarding the two others. There is a strong relationship between water and fat. The matrix can be seen in a first approximation as a ternary mixture and then very few factors would be needed.

The original file contained 240 spectra split in calibration, monitoring, testing, extrapolation for fat and extrapolation for protein. In the current trial only the fat content was kept and the data set was reduced to 224 samples but the way to get these 224 among the 240 is unknown. Anyway, from the 224 spectra, a first selection of 15 spectra was done based only on the 15 highest Mahalanobis distances (GH or Global H obtained by WINISI III package, FOSS-ISI, Port Mathilda, PA, USA). The test set 1 was selected as being outside the calibration spectral space and was to be used to estimate the extrapolation ability of the models. Among the 209 spectra after removing these 15 samples, a second test set containing also 15 samples was selected randomly. Fig. 1 shows the 194 calibration sample spectra and Fig. 2 is the scatter plot of PC1 and PC2 for the calibration set and the two test sets.

3. Deliberate modifications in the calibration data set

The two test sets were uploaded as such on the web site while some spectral data of the calibration set were slightly modified. The modifications did not disturb the models too much but it was to give more additional opportunities for the participants to prove their skills to find these “outliers”.

3.1. First modification

The first spectral modification was to change two consecutive data points of one spectrum to simulate a spike. Data point 60 (968 nm) of spectrum 50 was changed from 3.1661 to 3.17 (0.12%) and data point 61 (970 nm) of the same spectrum from 3.1696 to 3.15 (0.62%) producing a tiny hollow.

Nobody among the nine participants found this “outlier”. Finding it is very simple using a first derivative. The derivative must be done with a gap derivative algorithm with no smoothing and not by a Savitsky–Golay function. The trick is to subtract the consecutive data points. Fig. 3 shows a zoom of the sample n°50 before and after the modification.

3.2. Second modification

The second spectral modification is a global offset applied to one spectrum. Spectrum 85 is multiplied by

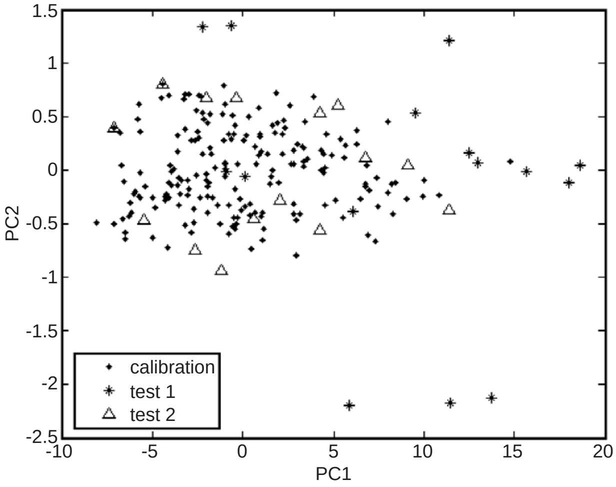

Fig. 2. Scatter plot of PC1 vs. PC2 for the calibration set and both test sets.

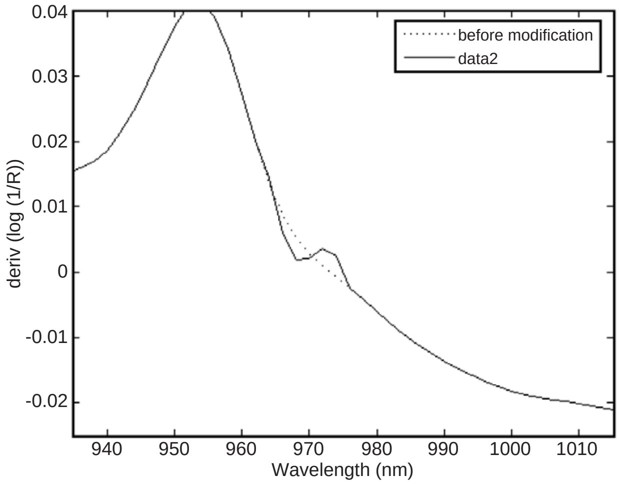

Fig. 3. Zoom of sample 50 before and after first modification.

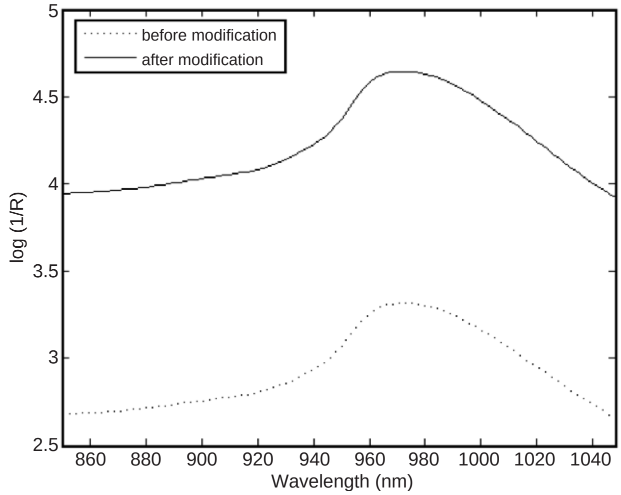

Fig. 4. Spectrum of sample 85 before and after second modification.

1.1 and 1 is added. Fig. 4 shows spectrum n°85 before and after the modification. Two participants found this spectrum as an outlier and removed it from the calibration set. A plot of the scores of the 1st PC from a PCA applied on the raw spectra indicates that this spectrum is particular (Fig. 5). The Mahalanobis distance of this sample is also the largest one of the calibration set.

3.3. Third modification

This modification concerned the spectrum 44. A small heteroscedastic noise was added on spectral data. Heteroscedastic noise is simulated first by generating Gaussian white noise, and then by multiplying this noise by the absorbance values. Noise spectra are reconstructed by adding the heteroscedastic noise to the original signal. This outlier can be found by applying a second derivative as shown in Fig. 6. As the noise is quite small the Mahalanobis distance calculated on the first components was unable to detect this defect. One participant detected this noisy spectrum by looking at the second derivatives.

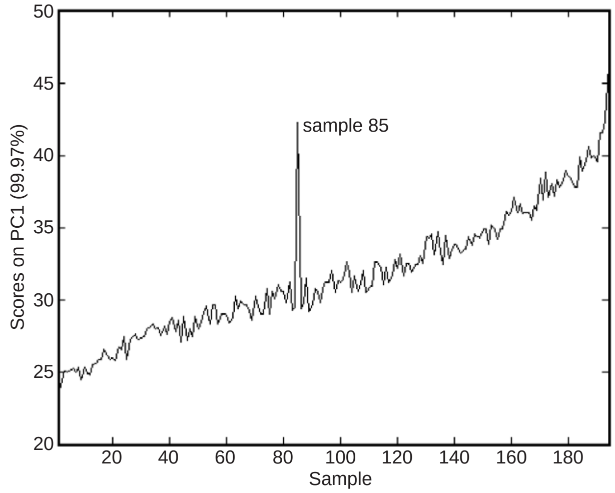

Fig. 5. Scores of the first PC showing a clear difference for sample 85.

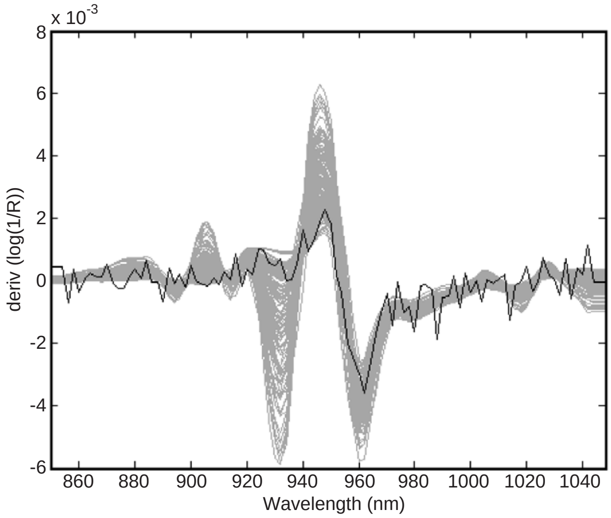

Fig. 6. Spectra corrected using second derivative bringing to the fore spectrum 85.

3.4. Fourth modification

The last change was the swapping of two consecutive fat reference values. The fat content of sample n°103 was reported to the sample n°104 and vice versa. These changes were obvious to see when plotting measured vs. predicted values. Keeping or removing these outliers does not change the RMSEP significantly. Two first approaches were carried out: two PLS models with 194 spectra and with 184 spectra (5 outliers removed), respectively. These models gave a RMSEP of 5.22 and 5.42, respectively, on the test set 1 and of 2.66 and 2.56 on the test set 2.

4. The participants' results

RMSEP, as reported by Fearn [1], is the global statistical parameter to evaluate the quality of the prediction and they are calculated with the values sent by the participants and the actual reference data. Table 1 reports in the second column the rating given by the jury based on the RMSEP and on the quality of the presentation performed during the congress. Columns 3 and 5 are the RMSEP, respectively, for the test set 1 and 2. Columns 4 and 6 are the rating according the RMSEP.

From this table, we can observe the wide range between the results: a factor of >10 was found between the smallest and the largest RMSEP.

Here the different approaches used in this challenge of modelling are presented.

Table 1
Grading of the participant results by the jury and grading according RMSEP for the test sets

Participant	Jury’s grading	RMSEP TEST 1	Grading TEST 1	RMSEP TEST 2	Grading TEST 2
1	5	4.06	3	2.44	7
2	9	35.55	9	8.44	9
3	3	3.39	2	1.48	4
4	8	9.50	7	2.13	6
5	4	6.19	6	1.37	3
6	1	5.16	4	0.72	1
7	7	5.40	5	1.29	2
8	6	10.00	8	2.67	8
9	1	3.26	1	1.81	5
10[a]	–	1.24	–	0.75	–

[a] Author’s results by ANN.

4.1. The linear approaches

Three participants used only linear global methods: n°4, 5 and 8 in Table 1. They used selection of variables and MLR.

4.1.1. Participant n. 4

The first linear approach is summarised by the author himself as follow. “*In the data that was to be analysed for this contest a non-linearity was present. The most logical approach would then be to use a non-linear method. However, one of the test sets contained almost exclusively samples that require extrapolation. It is known that non-linear methods, in particular Neural Network that requires a lot of parameters to be fitted, are less robust in an extrapolation condition than simpler methods. The idea was therefore to use a method as simple as possible like Multiple Linear Regression (MLR). Even though this method would certainly perform less good than non-linear ones when interpolating, it might be very efficient in an extrapolation condition because it is very parsimonious. Thus offset correction was applied to the data, followed by a simple stepwise selection, retaining only 4 variables. Next, a Multiple Linear Regression model with these variables was fitted. Internal validation results were very good in the interpolation condition, even though a non-linearity in the residuals could be detected. Moreover, the quality of cross validated results did not drop even in the extrapolation condition, which was a sign of a good robustness of the approach*”.

4.1.2. Participant n. 5

The second linear approach is very particular and the author did a variable selection among the original variables but also among subtractions or ratios of data points. *“The CORICO method* [2,3] *is characterised by an extension of the interaction notion (the “logical interaction”), a novel usage of partial correlation as a factor selection tool, and a synthetic diagram including the chosen factors and conspicuous instances. The analysis of partial correlation allows to distinguish the effects of various factors, even if they are not independent. The synthetic diagram eliminates redundancies and helps us to better perceive the physical sense of phenomena when the factors are not independent. The method simultaneously allows a screening of factors and the construction of a non-linear model, by a careful study of the interactions. We express the response as a multiple regression. The algorithm is not based on a transformation as in PLS method. Indeed, initial variables are kept. In CORICO, selection of predictors takes place before the model is fitted and the calculation of coefficients is performed. There is no limitation on the number of variables, and we do not have to specify a model a priori: CORICO discovers the model which fits the best with the actual data”.*

4.1.3. Participant n. 8

This participant used PLS after a selection of the variables by bootstrapping according to a modified algorithm proposed by Lazraq et al. [4]. The bootstrap iteration allowed to detect outliers. From 194 spectra, the calibration set was reduced to 125 spectra and the validation set to 18. 100 variables were reduced to 17 given the smallest RMSEP on a validation. The data set was reduced too much and the calibration performances seemed very good but the extrapolation prediction was very poor.

4.2. The classification approaches

Three participants used a classification approach following by linear models: n°1, 3 and 7 in Table 1.

4.2.1. Participant n. 1

This is the summary given by the participant using a classification method. “*In order to predict ’y’ with the two data sets ’test 1’ and ’test 2’, we decided to work on some parameters extracted from the derivatives of the spectra and to estimate ’y’ locally. Firstly, using the Classification And regression Tree introduced by Breiman* [5]*, we created a decision tree determined by binary recursive partitioning so that the groups are homogeneous in terms of the response ’y’. In the MATLAB package, at each step, the selected predictors (i.e., the parameters extracted) were chosen to maximize the reduction of an entropy function. We determined the group of the spectra in ’test 1’ and ’test 2’ with this decision tree. Secondly, a PLS regression in each group was performed with the predictors in order to compute prediction of the response ’y’ for the test samples. This procedure was restarted 5000 times with a training set chosen by re-sampling with replacement in the 191 initial spectra; this approach is similar to the Bootstrap Aggregation Procedure described by Breiman in 1996. This strategy led to an average prediction and a confidence interval for each spectrum in each test sample*”. The procedure worked quite well on test 1 (third place) but get only the 7th place for the test set 2.

4.2.2. Participant n. 3

"*Locally weighted regression (LWR) consists in decomposing a global model in a series of local linear models* [6]. *It is therefore adapted for data sets that exhibit some clustering or some non-linearity that can be approximated by local linear fits. For each point to be predicted, a local PLS model is built using the closest (in terms of Euclidian norm in the X space) calibration points. In this study, the points were given uniform weights in the local model, as described in* [7]. *This model seemed to be appropriate as it has been shown that it can accommodate slight non-linearity, and it is known to be robust in extrapolation conditions* [8], *two difficulties that were present in the data sets to be predicted for the contest."* This approach was the second best RMSEP for the test 1 and the 4th one for the test 2. Due to the quality of the presentation and the simplicity of the approach, the jury gave the contest price to this presentation.

4.2.3. Participant n. 7

This participant used a visual classification. The method is simple and quite efficient. The spectra were sorted according the *Y* values. A second derivative was computed and the spectra plotted. A two group separation was done according the *Y* value at 20% of fat content corresponding to the second derivative at 929 nm crossing the 0 line. MLR and PLS models were built for the 2 groups. The spectra of the test sets were visually classified into the groups according their value of the second derivative at 929 nm. The method obtained the second best score for the test 2 and the 5th for test 1.

4.3. The non-linear approaches

Three participants used non-linear approach: n°2, 6 and 9 in Table 1.

4.3.1. Participant n. 2

This participant applied the ANN approach intensively. "*Intelligent Problem Solver (IPS) of Statistica software was used in order to find the adequate neural network. More than 200 different neural networks were tested like Probabilistic Neural Network (PNN); Generalized Regression Neural Network (GRNN); Multilayer Perceptron (MLP) and Radial Basic Function (RBF) with different algorithms as K-Means Algorithm (KM), K-Nearest Algorithm (KN) and BP (Back propagation). The best result, with a learning error of 0.012, was obtained with a RBF-MLP 6:10:4:1 and back-propagation algorithm."* With the number of samples available this architecture led to overfit the data and the predictions of test sets were very poor.

4.3.2. Participant n. 6

This participant, after a deep exploration of the data using PCA, removed 12 outliers which had Mahalanobis distances higher than 3. A redundancy was found with twice the same spectrum. The data set was reduced from 194 to 181 spectra. The ANN approach was used with a splitting of the data set in 3 parts: 61 spectra for the training, 60 for the validation and 60 as a test set. As the introduction of the 100 variables would have led to overfitting with too many weights to be estimated regarding the number of samples, Genetic algorithm was applied to select a subset of 8 variables. Associated with only 2 nodes in the hidden layer, the model was the best one to predict the test 2 but only the 4th for the test 1. By using smaller validation and test sets, and by removing fewer "outliers", the model would have been better on test 1.

4.3.3. Participant n. 9

This last participant used a methodology based on the Mutual Information to select the variables followed by a classical kNN. *"Modeling data with hundreds of variables, such as spectra, requires a careful selection or projection of the original variables into a subspace of limited dimension. Selection means to extract some variables from the original set, while projection means to build new ones, by linear or nonlinear combinations of the original ones. Working with a limited number of variables is indeed a necessity to avoid over fitting and ill-conditioned models when a limited number of samples is known for learning (what is usually the case in spectrometry applications). A second fact is that in most situations, nonlinear models (built on the reduced set of variables) prove to be more efficient than linear methods. This is the case in most spectroscopic applications too, as it was proven on the challenge dataset. Therefore, if the model to be built is nonlinear, it would be extremely* non-*optimal to use first a linear variable selection or projection method. As nonlinear projection methods are extremely difficult to use (and still not at the level of state-of-the-art usable tools), the only remaining solution is to use a nonlinear selection method in the first stage. Nonlinear selection relies on the use of the mutual information between the independent (spectroscopic) variables and the dependent (to be predicted) one. The method developed for the challenge successively estimates the mutual information between each independent variable and the dependent one; once a variable is selected, it is maintained in the reduced set; then, the next variables are selected according to the same criterion, under the hypothesis that the already selected variables remain in the set. This provides a constructive method that implements a good compromise between the performances and an exhaustive search among all possible reduced sets that would be too time consuming. Another strong advantage of the selection principle (with regards to the projection one) is that it provides variables that are in the original set, i.e., which can be identified to wavelengths and thus interpreted by spectrometry experts".* The Mutual Information algorithm selected 10 variables and the kNN was used with 8 neighbors to reach the best results for the test set 1.

4.4. The authors' approaches

Preliminary trials showed that a classical pre-treatment such as SNV and Detrend followed by a first derivation (1,4,2) gave slightly better results than only the Log(1/*T*). This pre-treatment had been kept unchanged for the subsequent calculations. The calculations were carried out mainly using WINISI® software (Infrasoft International LLC., PA, USA) while MatLab® was used to perform the spectral modifications and the graphics.

When using MLR or PLS, the plot between measured and predicted values showed a scatter of the points with a banana shape indicating obvious non-linearity. It was then decided to start directly with ANN methods based on the PLS scores. As the projection of test 1 showed that we had to do extrapolations, a new test set of spectra was extracted from the 189 (194−5 outliers) spectra based on their GH (GH or Global H is equivalent to the Mahalanobis distance based on the standardised *k* first PCA scores). The 16 spectra with the highest GH were selected as a stop set. We assumed that a model able to predict these extreme spectra will be able to predict the spectra of test 1. Based on the 173 remaining spectra as training set the artificial network is optimised by testing 2 main parameters: the number of PLS terms as inputs and the number of nodes of the hidden layer. The 16 spectra were used as the stop set. The 2 parameters varied respectively from 5 to 12 and 1 to 5. The optimum (minimum of RMSEP) was found with 8 PLS factors and 3 nodes. The final ANN model was recalculated with all the available data (189 spectra) by keeping everything equal even the number of iterations of the training process. This ANN model predicted test 1 with a RMSEP of 1.24 and test 2 with a RMSEP of 0.75. The ANN package used was an experimental ANN software written by Mark Westerhaus from Infrasoft International, LLC. The optimisation took less than 5 min.

A second approach was similar to LWR. The concept invented and patented by Dr Shenk is called "LOCAL". For each sample to be predicted a subset of similar spectra is selected and a local PLS model is built. The final prediction is a weighted prediction according the *X* residuals and the standard deviation (the size) of the regression coefficients (*B* vectors). There are 3 main parameters to be optimised: the number of samples to be selected, the maximum number of PLS factors and the number of the first PLS factors to be ignored (poor prediction with the first PLS terms). There is no need for a test set whereas the algorithm works like a full cross validation (LOO — Leave-One-Out): the spectrum to be predicted is never used in the calibration set. These parameters can be optimised very fast using an experimental package called ISIeval® and provided for testing by Dr. Shenk. The optimisation was set up with 30 to 175 samples by a step of 5 and with 1 to 20 PLS factors by a step of 1. After 20 s of computing the answer was 30 samples to be selected as the closest neighbours, starting the prediction with the 3rd factor and using 8 factors as the maximum. The CVSEP was 0.55. Using this setting the two test sets were predicted. RMSEP were 1.92 for the test set 1 and 0.67 for test set 2.

5. Conclusion

Even if the challenge was the prediction of the two test sets, in the framework of a chemometric symposium, we would have expected to see more developments on the chemical and spectroscopic interpretation of the spectra; like fat and water specific absorbance peaks. But this had not been asked explicitly.

One can observe that the outlier detection step was not very deep. Except the swapping of the reference values which were obvious, few participants mentioned the modifications made on the spectra.

The spectra pre-treatments were not optimised. 2 participants used second derivatives and another used an offset correction, but most of the participants did not optimise the pre-treatment which is a non-negligible part of the calibration process.

The global linear methods were obviously not adapted to the data except after a classification. ANN gave the best results but LWR and "LOCAL" were very close and are less "dangerous" to be implemented.

A final comment is to always use all the data available to create a prediction model which has to be use on future unknown samples. Using a test set is useful and necessary to optimise the parameters of the modelling, but we suggest by experience, keeping everything equal, to recalculate the final model, the one which will be used in real time in the lab or in the plant predicting unknown samples, by using all the information available. It means by merging calibration set, stop set and test set.

The session with the "contest" presentations of these results interested most of the participants and the final conclusion was that it will be repeated during the next conference.

Acknowledgments

We would like to thank all the participants who spent time treating the data and presenting their results:

Dr. Ann Detroyer, L'Oreal Recherche, Aulnay-Sous-Bois, France
Dr. Ludovic Duponchel, LASIR, Villeneuve d'Ascq, France
Dr. Stéphanie Ledauphin, ENITIAA, NANTES, France
Pr. Marc Meurens, BNUT-UCL, Louvain-la-Neuve, Belgium
Dr. Frédéric Estienne, Aventis pharma S.A, Paris, France
Dr. Kieu An Ngo, Université Paul Cézanne, Marseille, France

Dr. Michel Verleysen, MLG-UCL, Louvain-la-Neuve, Belgium
Mr. Michel LESTY, CORYENT Conseil, Versailles, France
Dr. Abdelaziz Faraj, Institut Français du Pétrole, Rueil-Malmaison, France

The data sets are available on http://www.chimiometrie.org/ or can be obtained on request from: dardenne@cra.wallonie.be.

References

[1] T. Fearn, Towards a standard terminology, NIR News 15 (4) (2004).

[2] M. Lesty, Une nouvelle approche dans le choix des régresseurs de la régression multiple en présence d'interactions et de colinéarités. La revue de Modulad, n°22, janvier 1999, pp. 41–77. Inria-Ucis-diffusion, Rocquencourt B.P. 105, 78153 Le Chesnay (1999).

[3] C. Lesty, J. Pleau-Varet, M. Kujas, Geometric method and generalized linear models: two opposite multiparametric approaches illustrated on a sample of pituitary adenomas, J. Appl. Stat. 31 (2) (2004) 191–213.

[4] R. Lazraq, J. Cléoux, P. Gauchi, Selecting both latent and explanatory variables in PLS1 regression model, Chemometr. Intell. Lab. Syst. 66 (2003) 117–126.

[5] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[6] T. Naes, T. Isaksson, The idea behind and algorithm for locally weighted regression, NIR News 5 (4) (1994) 7–8.

[7] V. Centner, D.L. Massart, Optimisation in locally weighted regression, Anal. Chem. 70 (1998) 4206–4211.

[8] J. Verdú-Andrés, D.L. Massart, Comparison of prediction- and correlation-based methods to select the best subset of principal components for principal component regression and detect outlying objects, Appl. Spectrosc. 52 (1998) 1425–1434.