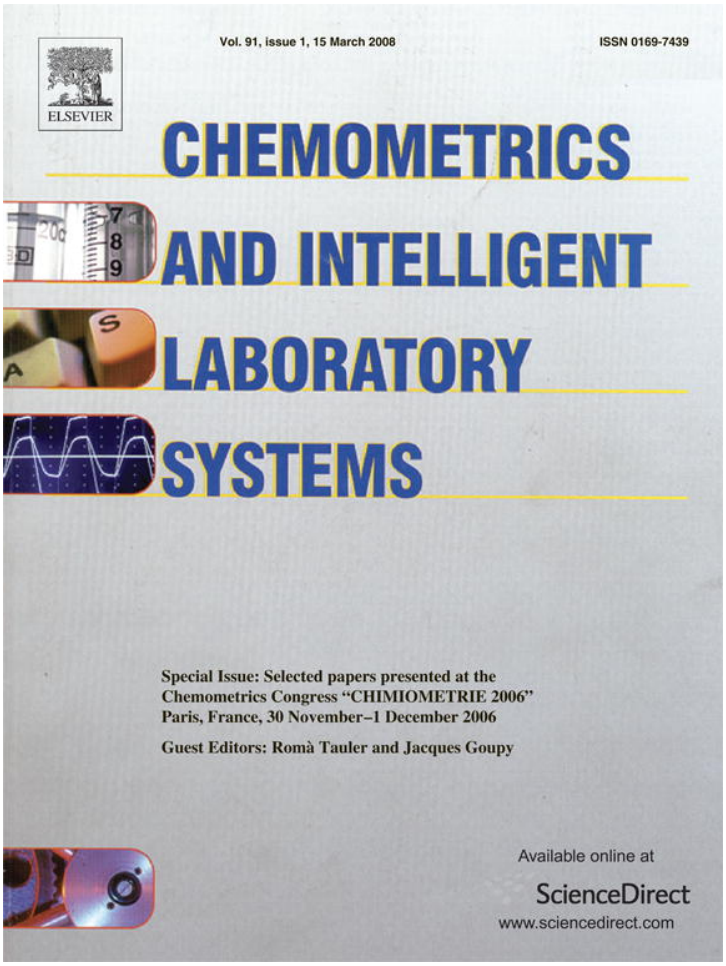

Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 91 (2008) 94–98
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Soil parameter quantification by NIRS as a Chemometric challenge at '*Chimiométrie* 2006'

Juan Antonio Fernández Pierna *, Pierre Dardenne

Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural Products Department, Chaussée de Namur n 24, 5030 Gembloux, Belgium

Received 2 May 2007; received in revised form 8 June 2007; accepted 9 June 2007
Available online 27 June 2007

Abstract

For the third consecutive year and due to the success of the chemometric contests organized within the framework of previous congresses, another data set has been proposed for the organisation committee at the 'Chimiométrie 2006' meeting (http://www.chimiometrie.org/) held in Paris, France (30th November and 1st December). As in the first contest organized in 2004 this data set was selected in order to test the ability of the participants for using regression methods based on NIR data. The data set consists on three different properties characterizing soils coming from the Walloon region in Belgium. This year, unlike previous contests, the data have been not modified by the authors. Only three participants decided to play with the proposed data and presented their own approaches during the conference. As last year's, this paper summarizes the approaches presented during the meeting by the participants and the authors.

Keywords: Chemometrics; Challenge; Soil quantification; NIR

1. The data of the challenge

As last 2 years a chemometric contest has been organized within the framework of the 'Chimiométrie 2006' meeting [1, 2]. This time, two data sets of soils were provided by Dr. Pierre Dardenne, CRA-W (Belgium) and were available on the web site of the conference (http://www.chimiometrie.org/), which included a calibration data set and a test set. The datasets are quite large, typical of datasets encountered in pedometrics or "the application of mathematical and statistical methods for the study of the distribution and genesis of soils" [3]. There are 618 reflectance spectra of dried and sieved soil samples in the calibration (cal) set measured between 1100 and 2498 nm at 2 nm data interval (see Fig. 1). The test set contained 207 spectra measured under the same conditions as the calibration set. All the spectra come from cultivated soil samples collected from all over the Walloon region in Belgium.

There are three parameters associated with these data sets: Nt (Total Nitrogen in g/Kg of dry soil), CEC (Cation Exchange Capacity in meq/100 g of dry soil) and Ciso (Carbon in% of dry soil) (iso means that C has been determined following the ISO14235 method). Several 0s were present in the **y** matrix corresponding to missing values and did not have to be used for calibration. Fig. 2 shows a plot for the **y** matrix showing the missing values for each property. In the case of Nt 22% of the samples are missing values, 11% in the case of Ciso and 46% for CEC. The same situation occurs for the test set and the final test data sets contain 160 for Nt, 184 for Ciso and 113 for CEC.

The aim for the participants was to develop calibrations for the three properties included in the **y** matrix using the calibration set and then to predict the test set blind spectra using any kind of method. Parameter values were not attached to the test set; these values were retained by the judges for evaluating contest entries. The participants sent to the jury a text file or a slide presentation with the proposed methodology and the predicted results. The methodology used (pre-processing, regression algorithm, outlier detection methods, software...) was free.

2. Participant's approaches

2.1. Participant no. 1

Standard Normal Variate (SNV) followed by detrend and second derivative was applied as pre-processing. The corrected data is presented in Fig. 3.

* Corresponding author. Fax: +32 (0) 81 62 03 88.
E-mail address: fernandez@cra.wallonie.be (J.A. Fernández Pierna).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.06.007

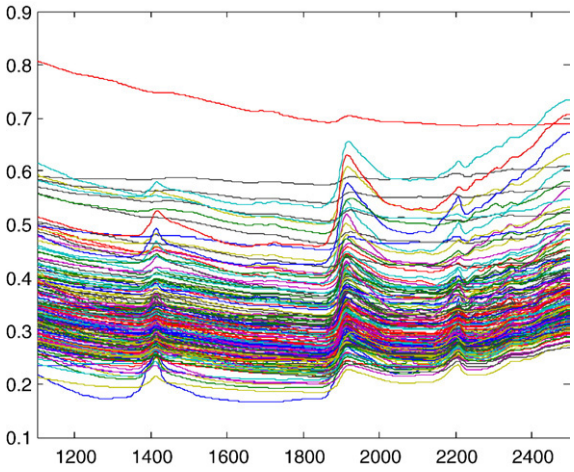

Fig. 1. Soil spectra before pre-treatment.

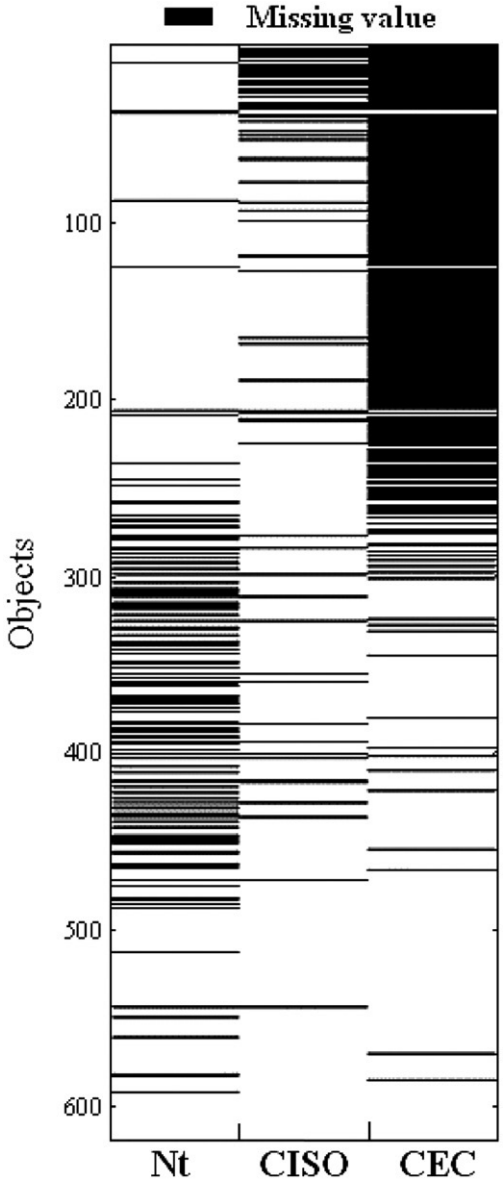

Fig. 2. Plot of the **y** matrix showing the missing values for each property.

The participant has decided to work while taking into account the missing values by replacing those values by possible values calculated using the NORM software [4]. This software replaces each missing value by 20 simulated values. Then 20 full matrices Ycal which are then available can be used to build a regression models between Xcal and the 20 Ycal matrices. The regression method chosen is the LOCAL regression proposed by WinISI [5]. After the 20 local regressions, 20 predicted values are obtained for the unknown test set. The mean of these models is used as final estimation. Several prediction samples (1, 51, 69, 79, 88, 92, 131, 142 and 205) have been considered as outliers due to their high GH, the standardized Mahalanobis distance used by WinISI.

2.2. Participant no. 2

Multiple Scatter Correction followed by auto scaling has been used as pre-processing. Then PCA was used for outlier identification and samples 272, 313, 377, 402 have been removed. For each property, the following steps have been performed:

Elimination of missing values.
Building of the model using Back propagation Neural Network using the first 20 PCs as input, one hidden layer with a hyperbolic transfer function and one output node with a linear transfer function. The hidden layer contains 3 nodes for the Nt and CEC and 4 nodes for Ciso.
Prediction of the test set.

2.3. Participant no. 3

The participant performed a dimension reduction by projection on B-Spline basis. As a first step, each spectrum is whitened (*i.e.* reduction and centering); this preprocessing allows the detection of outliers, which are removed from the database. The problem is then divided into three sub-problems corresponding to the prediction of the three parameters of interest. All following steps are applied separately to each

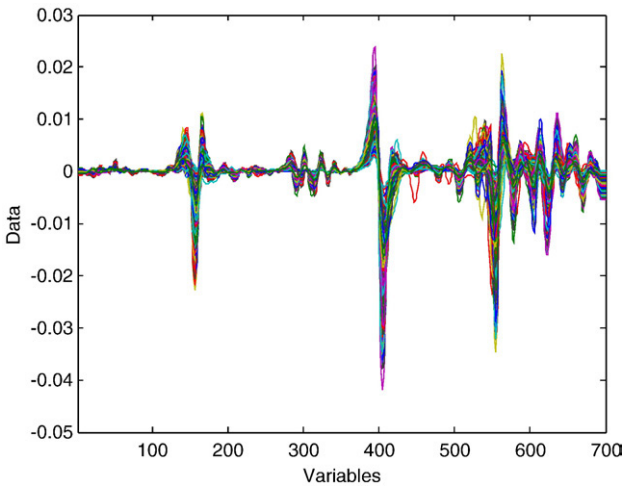

Fig. 3. Soil spectra after SNV followed by detrend and second derivative.

Table 1
Summary of all the results obtained during the challenge

		Participant 1	Participant 2	Participant 3	Author	Author	Author
	Methods	Norm+local	Backpropagation NN	Bsplines+RBFN	PLS	Local	LS-SVM
RMSEP	1-Nt	0.6537	0.7737	0.9655	0.8404	0.5853	0.5575
	2-Ciso	0.5577 (1 sample NaN)	0.7755	1.4846	1.1476	0.4332	0.559
	3-Cec	3.5826 (2 sample NaN)	4.6235	5.3609	4.6051	3.3328	3.4373
R^2	1-Nt	0.7673	0.63	0.438	0.5632	0.7979	0.8172
	2-Ciso	0.8702	0.77	0.5338	0.496	0.9216	0.8908
	3-Cec	0.7148	0.51	0.3474	0.5087	0.7508	0.7294
Mean R^2		0.78	0.64	0.44	0.52	0.82	0.81

sub-problem. The spectra can be considered as high-dimensional vectors. Their dimension has to be reduced in order to avoid problems related to the "curse of dimensionality". Moreover the high degree of co-linearity between wavelengths renders the modeling of the spectra slow and unstable. The dimension reduction is achieved by projecting the spectra on a B-Spline basis as in [6]. Indeed, the spectra can be viewed as the discretization of a continuous function. The order of the B-Splines involved in this process is 4. Their number is chosen by a Leave-One-Out procedure to optimize the reconstruction of the spectra. The B-Splines are selected though their coefficients, which are also whitened to avoid scale problems.

Seven coefficients are subsequently selected by a forward procedure based on the mutual information. The number of selected coefficients is chosen to keep the computational time at a reasonable level. The mutual information criterion and the forward procedure are described in [7]. The estimator used to

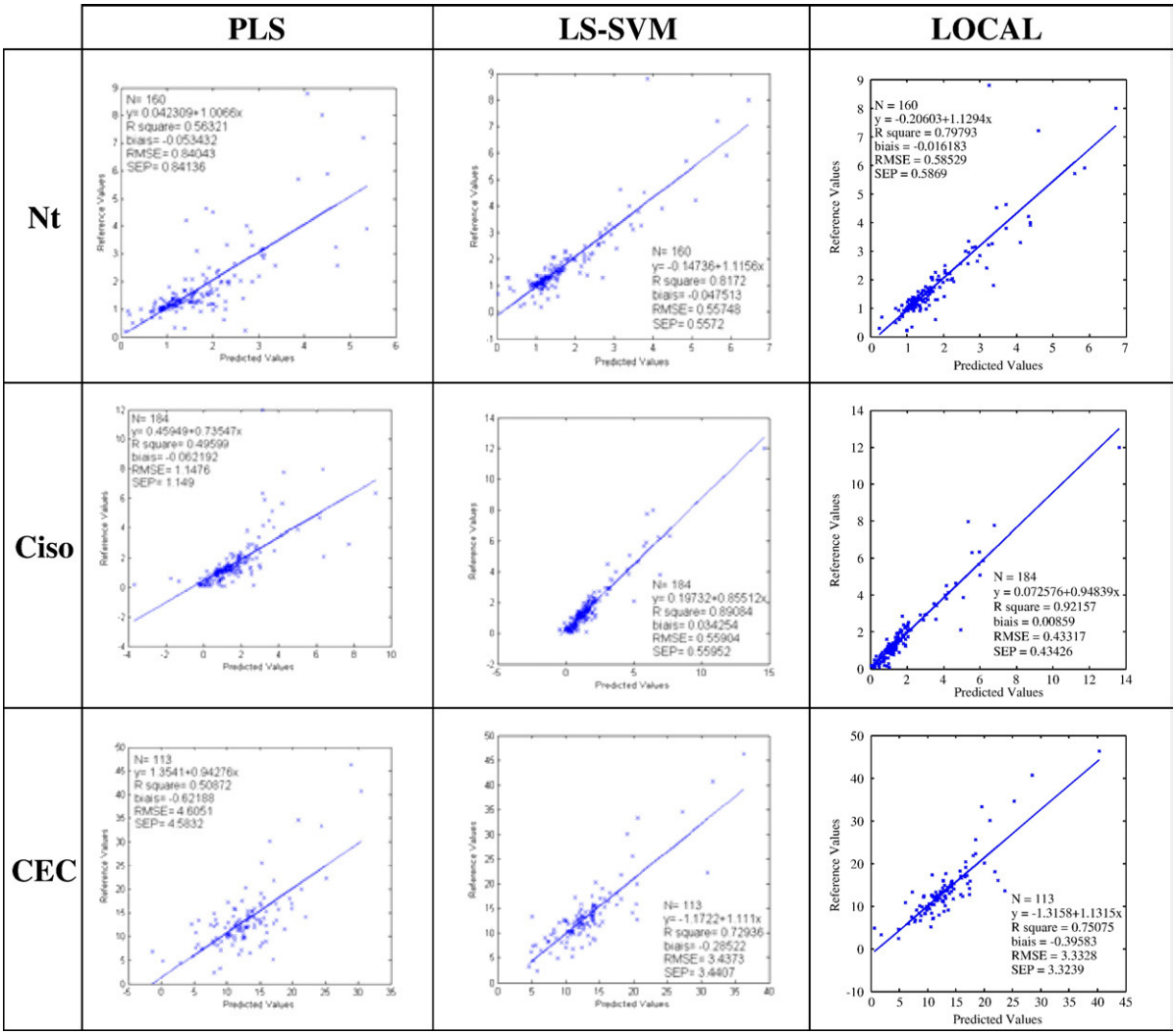

Fig. 4. Prediction results using PLS, LS-SVM and LOCAL for the different properties: Nt (first row), Ciso (second row) and CEC (third row).

estimate the mutual information is described in [8]. Radial Basis Function Networks are then built on each of the 2^7 possible subsets of coefficients. The best model is chosen according to the prediction performances, which are evaluated by an 8-fold Cross-Validation.

3. The author's approaches

The authors tested three approaches: PLS and LS-SVM, using Matlab v7.0 (The Mathworks, Inc., Natick, MA, USA) and LOCAL from the WinISI software [5]. Because PLS and LS-SVM have been widely explained somewhere [9,10], only the main advantages are shown and only the LOCAL regression method will be detailed here.

The main advantages of Support Vector Machines (SVM) are: 1) its ability to deal with ill-posed problems; 2) it shrinks the regression coefficients by imposing a penalty on their values; 3) it leads to global models that are unique; 4) sparse solutions are found and 5) linear and nonlinear regression can be performed. Least-Squares Support Vector Machines (LS-SVM) has the same advantages but it requires solving a set of linear equations (linear programming) instead of requiring a solution of nonlinear equations (quadratic programming), which are difficult to solve.

The LOCAL method of calibration is also called "local calibration". It matches the sample to be predicted with a small homogenous group of samples selected from a product library file. In this situation, a library file for the product is needed containing reference values. Each "unknown" sample (to be predicted) is compared to the library and the closest samples are selected from the library. The similarity index used is simply a correlation coefficient between spectra: two spectra having a correlation of 1 must have the same composition. If the global and neighbourhood H values are acceptable (GH less than 3.0 and NH less than 1.0), the calibration is developed based on the selected library samples and the "unknown" sample is predicted. Temporary **Y** and **X** matrices are arranged and then a specific PLS model is calculated with the N selected references to predict one sample. There are 3 parameters to be optimized before routine operation; the number (N) of the closest samples, the maximum number of PLS factors (Fmax) and the minimum number of PLS factors (Fmin). The final predicted result is a weighted sum of the predicted values from all the models between Fmin and Fmax, values which are weighted according to the standard deviation of the Bcoefficients and to the size of the Xresiduals. This method is the only one (that we know) which takes information of the unknown sample (the spectrum itself with the use of the Xresiduals) to weigh the predicted values and so to improve the accuracy. LOCAL was investigated years ago by Sinnaeve et al. and 15% improvement of accuracy was found [11].

4. Results

Table 1 shows the final results for the challenge 2006. Similar results are obtained when working with LOCAL and LS-SVM with a RMSEP respectively of 0.59 and 0.56 for Nt, 0.43 and 0.56 for Ciso and 3.33 and 3.44 for CEC. Good results are also obtained by participant 1 with a RMSEP of 0.65, 0.56 and 3.58 for Nt, Ciso and CEC respectively and 0.78 as mean R^2. Participants 2 and 3 obtained 0.64 and 0.44 respectively for mean R^2.

Fig. 4 shows the different prediction performances for Nt, Ciso and CEC using PLS, LS-SVM and LOCAL regression.

5. Conclusion

Quantification of soil parameters by NIRS becomes a more and more interesting topic for research in "pedometrics" [12]. The diversity of mineral composition of the soils and the weak level content of organic matter make their quantification by NIR a real challenge. Only three chemometricians dared presenting their results: more had tried, but gave up seeing the difficulties and the relatively poor R^2. The interest of local method is demonstrated and the data sets we used are still too small to find enough neighbours to compute a local model for each unknown spectrum. There is a need to gather internationally data bases from different origins providing the reference methods are correctly aligned.

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 specially the three participants who spent time to treat the data and present their results:

Dr Ludovic DUPONCHEL, Molecular characterization and Chemometrics Group, Laboratory of Raman and Infrared Spectroscopy, University of Lille (France).

Dr. Frédéric ESTIENNE, Sanofi Aventis, Drug Design, Paris (France).

PhD Catherine KRIER and Dr. Michel VERLEYSEN, Catholic University of Louvain, Louvain-la-Neuve (Belgium).

References

[1] P. Dardenne, J.A. Fernández Pierna, A NIR data set is the object of a Chemometric contest at 'Chimiométrie 2004', Chemometrics and Intelligent Laboratory Systems 80 (2006) 236–242.

[2] J.A. Fernández Pierna, P. Dardenne, Chemometric contest at 'Chimiométrie 2005': a discrimination study, Chemometrics and Intelligent Laboratory Systems 86 (2007) 219–223.

[3] IDRC - http://www.idrc-chambersburg.org/.

[4] J.L. Shafer, 'NORM: multiple imputations of incomplete multivariate data under a normal model, version 2.03', software for Windows 95/97/NT, 1999 www.stat.psu.edu/~jls/misoftwa.html.

[5] J.S. Shenk, M.O. Westerhaus, Population definition sample selection and calibration procedures for near infrared reflectance spectroscopy, Crop Science 31 (1991) 469–474.

[6] F. Rossi, D. François, V. Wertz, M. Verleysen, 'Fast Selection of Spectral Variables with B-Spline Compression', Chemometrics and Intelligent Laboratory Systems, Elsevier, in press.

[7] C. Krier, D. François, V. Wertz, M. Verleysen, 'Feature Scoring by Mutual Information for Classification of Mass Spectra', FLINS 2006, 7th International FLINS Conference on Applied Artificial Intelligence, August 29–31, (2006), pp. 557–564, Geneva (Italy).

[8] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information', Physical Review E69 (2004) 066138.
[9] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, J.P. Lewi, J. Smeyers-Verbeke, 'Chemometrics: A Textbook', vol. 2, Elsevier, Amsterdam, 1988.
[10] R.P. Cogdill, P. Dardenne, Least-squares support vector machines for chemometrics: an introduction and evaluation, Journal of Near Infrared Spectroscopy 12 (1) (2004) 93–100.
[11] G. Sinnaeve, P. Dardenne, R. Agneessens, Global or local? A choice for NIR calibrations in analyses of forage quality, Journal of near infrared spectroscopy 2 (1994) 163–175.
[12] D.J. Brown, K.D. Shepard, M.G. Walsh, Global soil characterization with VNIR diffuse reflectance spectroscopy, Geoderma 132 (2006) 273–290.