Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 67 (2005) 736–740
Talanta
www.elsevier.com/locate/talanta

A method for calibration and validation subset partitioning

Roberto Kawakami Harrop Galvão [a], Mário César Ugulino Araujo [b,*], Gledson Emídio José [b], Marcio José Coelho Pontes [b], Edvan Cirino Silva [b], Teresa Cristina Bezerra Saldanha [b]

[a] *Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, São José dos Campos, São Paulo, Brazil*
[b] *Universidade Federal da Paraíba, Departamento de Química, P.O. Box 5093, João Pessoa, Paraiba 58051-970, Brazil*

Received 18 February 2005; received in revised form 24 March 2005; accepted 24 March 2005
Available online 6 May 2005

Abstract

This paper proposes a new method to divide a pool of samples into calibration and validation subsets for multivariate modelling. The proposed method is of value for analytical applications involving complex matrices, in which the composition variability of real samples cannot be easily reproduced by optimized experimental designs. A stepwise procedure is employed to select samples according to their differences in both **x** (instrumental responses) and *y* (predicted parameter) spaces. The proposed technique is illustrated in a case study involving the prediction of three quality parameters (specific mass and distillation temperatures at which 10 and 90% of the sample has evaporated) of diesel by NIR spectrometry and PLS modelling. For comparison, PLS models are also constructed by full cross-validation, as well as by using the Kennard–Stone and random sampling methods for calibration and validation subset partitioning. The obtained models are compared in terms of prediction performance by employing an independent set of samples not used for calibration or validation. The results of *F*-tests at 95% confidence level reveal that the proposed technique may be an advantageous alternative to the other three strategies.

Keywords: Sample subset partitioning; PLS regression; Kennard–Stone algorithm; NIR spectrometry; Diesel analysis

1. Introduction

In multivariate calibration problems involving complex matrices, it can be difficult to reproduce the composition variability of real samples by means of optimized experimental designs [1]. A typical example consists of fuel analysis for the determination of quality parameters such as octane number, cetane index, sulphur content, distillation temperatures, flash point, freezing point, percentage of aromatics and specific mass to name only a few [2–4]. In such cases, a representative calibration set must be extracted from a pool of real samples. Moreover, validation samples should also be selected to assess the quality of the model and to determine model parameters such as the number of latent variables in PLS regression [5].

Several works have addressed the problem of selecting a representative subset from a large pool of samples [6–9]. In this context, random sampling (RS) is a popular technique because of its simplicity and also because a group of data randomly extracted from a larger set follows the statistical distribution of the entire set. However, RS does not guarantee the representativity of the set, nor does it prevent extrapolation problems [10]. In fact, RS does not ensure that the samples on the boundaries of the set are included in the calibration.

An alternative to RS that is often employed is the Kennard–Stone (KS) algorithm. KS is aimed at covering the multidimensional space in a uniform manner by maximizing the Euclidean distances between the instrumental response vectors (**x**) of the selected samples [9–12]. In a neural network classification study by Wu et al. [9], KS was found to be superior to RS, as well as to Kohonen self-organizing mapping [13]. The study also showed that KS leads to classification results similar to those obtained by using the more elaborate and time-consuming D-optimal design method [1].

* Corresponding author. Tel.: +55 83 216 7438; fax: +55 83 216 7437.
E-mail address: laqa@quimica.ufpb.br (M.C.U. Araujo).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.03.025

It is worth noting that the specific problem of partitioning a pool of real samples into calibration and validation sets for multivariate calibration purposes has not been extensively explored in the literature. Kanduc et al. [10] addressed this problem in a case study involving the prediction of colour properties of a titanium dioxide white pigment from other physical and chemical parameters. The study involved the comparison of RS, KS, Kohonen self-organizing mapping and time-dependent sampling. The models obtained in this manner were compared in terms of their generalization performance in a third prediction set not employed in the modelling procedures. The results revealed that the best predictions were achieved by using KS. However, it should be noticed that the choice of the prediction set was not entirely unbiased in that the prediction samples were extracted from the validation set after the calibration/validation partitioning had already been performed. Moreover, the authors emphasize that an investigation of this problem on a case-by-case basis is always recommended.

Despite the comparative advantages of KS over the alternative partitioning methods cited above, a shortcoming of KS in the multivariate calibration context lies in the fact that the statistics of the dependent variable (y) are not taken into account. It could be argued that the inclusion of y-information in the selection process might result in a more effective distribution of calibration samples in the multidimensional space, thus improving the predictive ability and robustness of the resulting model.

In the work of Dantas Filho et al. [14], an approach for considering joint **x**–y statistics in the selection of calibration samples was proposed for the purpose of total sulphur determination in diesel samples by NIR spectrometry. However, such an approach was aimed at extracting a reduced subset from the pool of calibration samples, rather than partitioning the available data into calibration and validation. In fact, the analyst was required to provide the calibration and validation sets as a starting point for the sample selection procedure. For this purpose, the calibration/validation partitioning was carried out in a qualitative manner on the basis of a univariate inspection of the reference parameter values followed by an analysis of the residual **x** and y-variance in the PLS regression.

In the present paper, a method for calibration/validation partitioning is proposed to take into account the variability in both x and y dimensions. The method, termed SPXY (Sample set Partitioning based on joint **x**–y distances), extends the KS algorithm by encompassing both **x**- and y-differences in the calculation of inter-sample distances. For illustration, a multivariate calibration problem involving NIR spectrometric analysis of diesel samples is considered. Three quality parameters are determined, namely specific mass and the distillation temperatures at which 10 and 90% of the sample has evaporated (T10 and T90%). SPXY is compared with KS and RS for the division of modelling data into calibration and validation sets for PLS regression. The performances of the resulting models are compared in terms of root-mean-square errors calculated in prediction sets not included in the modelling procedures. For the purpose of ensuring the independence of such sets, the prediction samples are randomly extracted from the initial pool of experimental data, before the calibration/validation partitioning procedures. In order to improve the robustness of the error statistics, the study is repeated five times by resampling the prediction set. The three strategies (PLS-SPXY, PLS-KS, and PLS-RS) are also compared with PLS employing full cross-validation (PLS-CV).

2. Background and theory

2.1. KS algorithm

The classic KS algorithm is aimed at selecting a representative subset from a pool of N samples. In order to ensure a uniform distribution of such a subset along the **x** (instrumental response) data space, KS follows a stepwise procedure in which new selections are taken in regions of the space far from the samples already selected. For this purpose, the algorithm employs the Euclidean distances $d_{\mathbf{x}}(p, q)$ between the **x**-vectors of each pair (p, q) of samples calculated as

$$d_{\mathbf{x}}(p, q) = \sqrt{\sum_{j=1}^{J} [x_p(j) - x_q(j)]^2}; \quad p, q \in [1, N] \tag{1}$$

For spectral data, $x_p(j)$ and $x_q(j)$ are the instrumental responses at the jth wavelength for samples p and q, respectively. J denotes the number of wavelengths in the spectra.

The selection starts by taking the pair (p_1, p_2) of samples for which the distance $d_{\mathbf{x}}(p_1, p_2)$ is the largest. At each subsequent iteration, the algorithm selects the sample that exhibits the largest minimum distance with respect to any sample already selected. Such a procedure is repeated until the number of samples specified by the analyst is achieved.

2.2. Proposed SPXY algorithm

The proposal of the present paper consists of augmenting the distance defined in Eq. (1) with a distance in the dependent variable (y) space for the parameter under consideration. Such a distance $d_y(p, q)$ can be calculated for each pair of samples p and q as

$$d_y(p, q) = \sqrt{(y_p - y_q)^2} = |y_p - y_q|; \quad p, q \in [1, N] \tag{2}$$

In order to assign equal importance to the distribution of the samples in the **x** and y spaces, distances $d_{\mathbf{x}}(p, q)$ and $d_y(p, q)$ are divided by their maximum values in the data set. In this manner, a normalized **x**y distance is calculated as

$$d_{\mathbf{x}y}(p, q) = \frac{d_{\mathbf{x}}(p, q)}{\max_{p,q \in [1,N]} d_{\mathbf{x}}(p, q)} + \frac{d_y(p, q)}{\max_{p,q \in [1,N]} d_y(p, q)}; \quad p, q \in [1, N] \tag{3}$$

A stepwise selection procedure similar to the KS algorithm can then be applied with $d_{xy}(p, q)$ instead of $d_x(p, q)$ alone.

The Matlab code for implementation of the proposed SPXY algorithm can be found in Appendix A.

3. Experimental

3.1. Samples

The data set consisted of 170 diesel samples that were collected from gas stations in the city of Recife (Pernambuco State, Brazil) and stored in ambar glass flasks.

3.2. Reference methods and apparatus

The reference values for specific mass and distillation temperatures (T10 and T90%) were obtained according to the ASTM (American Society for Testing and Materials) 4615 and D86 methods, respectively.

Specific mass and distillation temperatures were determined by using a Kyoto Electronics DA-130 digital densimeter, and a Herzog HDA 628 automatic distiller, respectively, which were operated according to the recommendations of the manufacturers for optimal working conditions.

3.3. NIR spectra acquisition and pre-processing

The spectra were acquired using a FT-NIR/MIR spectrometer Perkin Elmer GX with a spectral resolution of 2 cm^{-1}, 16 scans and an optical path length of 1.0 cm. Only the NIR region in the range 885–1600 nm was exploited, because at shorter wavelengths (<885 nm) the signal is too close to the baseline, whereas above 1600 nm the signal saturates the detector. In order to circumvent the problem of systematic variations in the baseline, derivative spectra were calculated with a Savitzky–Golay filter using a 2nd-order polynomial and a 11-point window. Each resulting spectrum had 1431 variables.

3.4. Software

Spectrum derivation and PLS modelling were performed with The Unscrambler 7.5 software (CAMO). By using the default settings of the software package, the number of latent variables in the PLS model was determined either by testing on the validation set (PLS-RS, PLS-KS, PLS-SPXY) or by full cross-validation (PLS-CV).

RS, KS, and SPXY routines were implemented in Matlab 6.1. The division of the 170 samples into calibration, validation, and prediction sets was carried out in the following manner. Initially, 50 prediction samples were extracted from the full set in a random manner to simulate the analysis of a batch of real unknown samples. The remaining 120 samples were divided into calibration and validation sets of 70 and 50 elements, respectively, by using the three selection methods to be compared (RS, KS, and SPXY).

In order to improve the statistical significance of the comparison, the extraction of the prediction set and the subsequent partitioning of the remaining samples into calibration and validation by RS, KS, and SPXY was repeated five times. In this manner, the four modelling strategies (PLS-RS, PLS-KS, PLS-SPXY, and PLS-CV) were tested with five different prediction sets.

Because of the random nature of the RS method, special care was taken to improve the statistical significance of the PLS-RS results. For this purpose, five RS calibration/validation partitions were performed for each of the five extractions of the prediction set. In this manner, $5 \times 5 = 25$ evaluations of PLS-RS were carried out.

For each diesel quality parameter, the predictive ability of PLS-RS, PLS-KS, PLS-SPXY, and PLS-CV were compared in terms of an overall root-mean-square error of prediction (RMSEP). Such an RMSEP statistic was defined for PLS-KS, PLS-SPXY, and PLS-CV as

$$\mathrm{RMSEP} = \sqrt{\frac{1}{I \cdot M} \sum_{i=1}^{I} \sum_{m=1}^{M} (y_{i,m} - \hat{y}_{i,m})^2} \quad (4)$$

where $y_{i,m}$ and $\hat{y}_{i,m}$ are the reference and predicted values of the parameter under consideration in the mth prediction sample ($m = 1, \ldots, M$) of the ith prediction set ($i = 1, \ldots, I$). As explained above, $M = 50$, and $I = 5$ were employed in this work. The RMSEP calculation for PLS-RS also embodied the RS repetitions as

$$\mathrm{RMSEP} = \sqrt{\frac{1}{I \cdot K \cdot M} \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{m=1}^{M} (y_{i,k,m} - \hat{y}_{i,k,m})^2} \quad (5)$$

where index $k = 1, \ldots, K$ refers to each of the five ($K = 5$) calibration/validation divisions by RS.

The statistical significance of differences between RMSEP values were assessed by using an F-test for a confidence level of 95%. It is worth noting that the RMSEP calculation for PLS-RS involves five times more degrees of freedom than the respective calculation for each other modelling strategy (PLS-KS, PLS-SPXY, and PLS-CV).

4. Results and discussion

The original spectra of the 170 diesel samples analyzed by NIR spectrometry are presented in Fig. 1a. Such spectra display baseline features that were corrected by derivation with a Savitzky–Golay filter. Fig. 1b shows the resulting derivative spectra, which were employed throughout the work.

Table 1 presents the RMSEP results of the four modelling strategies for each parameter under study.

As regards the comparison of PLS-KS, PLS-RS, and PLS-CV performances, it can be seen that PLS-CV yielded the smallest RMSEP for specific mass, whereas PLS-RS yielded

Table 1
RMSEP results obtained for each modelling strategy

Parameter	PLS-KS	PLS-RS	PLS-CV	PLS-SPXY
Specific mass (830–864 kg m^{-3})	1.8	1.8	1.6	1.7
T10% (186.6–269.9 °C)	5.5	5.4	5.5	5.3
T90% (317.2–385.5 °C)	4.7	4.4	4.5	4.0

The range of each parameter in the data set is indicated in parenthesis.

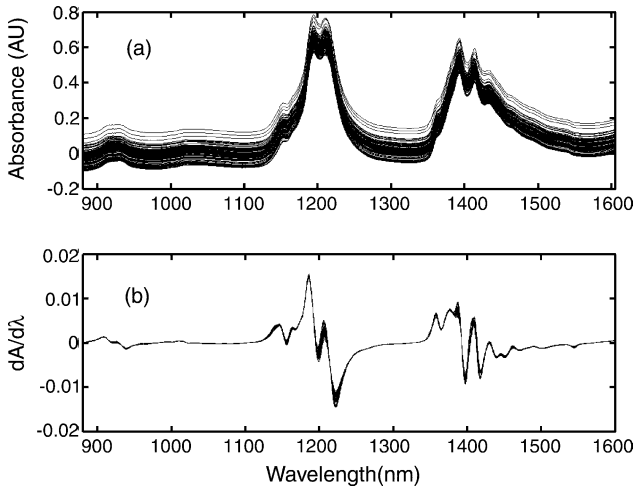

Fig. 1. Original (a) and derivative (b) NIR spectra of the 170 diesel samples.

the smallest RMSEP for T10 and T90%. However, the *F*-test reveals that the only significant differences favour PLS-CV over the other two strategies for specific mass.

On the other hand, the RMSEP of the proposed technique (PLS-SPXY) is smaller than the corresponding values of the other strategies in almost all cases. The only exception consists of PLS-CV for specific mass, but even in this case the difference with respect to PLS-SPXY is not significant at the adopted confidence level (95%) of the *F*-test. In fact, PLS-SPXY is favoured over the other strategies in all significant *F*-test comparisons (against PLS-KS, PLS-RS, and PLS-CV in T90%).

5. Conclusions

This paper proposed a method to divide modelling data into calibration and validation sets for multivariate calibration. The method, termed SPXY, employs a partitioning algorithm that takes into account the variability in both **x**- and *y*-spaces. In this manner, the multidimensional space may be covered more effectively in comparison with partitioning schemes based on **x**-information alone (such as the Kennard–Stone (KS) algorithm) or random sampling (RS). As a result, improvements on the prediction performance of the resulting PLS models may be attained. In terms of computational workload, SPXY is comparable with KS, in the sense that both algorithms employ simple distance calculations. In contrast, full cross-validation, which is often used in PLS calculations, is considerably more demanding in computational terms.

SPXY was successfully employed with PLS regression for NIR spectrometric determination of specific mass and T10, and T90% distillation temperatures in diesel samples. The results showed that the proposed method may be an advantageous alternative to divide modelling data into calibration and validation sets for PLS regression in comparison with KS, RS, or full cross-validation.

Acknowledgments

The authors thank FINEP-CTPETRO (Grant 0652/00) science funding program, PROCAD/CAPES (Grant 0064/01-7), PRONEX/CNPq (Grant 015/98), and FAPESP (Grant 03/09433-5) for partial financial support. The research fellowships granted by the Brazilian agency CNPq are also gratefully acknowledged. The authors are also indebted to the Laboratório de Combustíveis of the Departamento de Engenharia Química of the Universidade Federal de Pernambuco for providing the diesel samples and reference values for this work.

Appendix A. Matlab implementation of the proposed SPXY algorithm

In this Matlab function, X and y are the instrumental response matrix (independent variables) and the column vector of parameter values (dependent variable), respectively. Ncal is the number of objects to be selected for the calibration set. The indexes of the selected objects are returned in vector m.

```
function m = spxy(X,y,Ncal)
dminmax = zeros(1,Ncal); % Inicializes the vector of minimum distances.
M = size(X,1); % Number of objects
samples = 1:M;
Dx = zeros(M,M); % Inicializes the matrix of X-distances.
Dy = zeros(M,M); % Inicializes the matriz de y-distances.
for i = 1:M-1
   xa = X(i,:);
   ya = y(i,:);
   for j = i + 1:M
      xb = X(j,:);
      yb = y(j,:);
      Dx(i, j) = norm(xa–xb);
      Dy(i, j) = norm(ya–yb);
   end
end
```

```
Dxmax = max(max(Dx));
Dymax = max(max(Dy));
D = Dx/Dxmax + Dy/Dymax; % Combines the X and y distances.
% D is an upper triangular matrix.
% D(i,j) is the distance between objects i and j (j > i).
[maxD,index_row] = max(D);
% maxD is a row vector containing the largest element for each column
   of D.
% index_row is the row in which the largest element of the column if found.
[dummy,index_column] = max(maxD);
% index_column is the column containing the largest element of
   matrix D.
m(1) = index_row(index_column);
m(2) = index_column;
for i = 3:Ncal
   pool = setdiff(samples,m);
   % Pool is the index set of the samples that have not been selected yet.
   dmin = zeros(1,M − i + 1);
   % dmin will store the minimum distance of each sample in "pool" with
   respect to the previously selected samples.
   for j = 1:(M − i + 1)
        indexa = pool(j);
        d = zeros(1,i-1);
        for k = 1:(i − 1)
           indexb = m(k);
           if indexa < indexb
              d(k) = D(indexa,indexb);
        else
           d(k) = D(indexb,indexa);
        end
      end
      dmin(j) = min(d);
   end
   % At each iteration, the sample with the largest dmin value is selected.
   [dummy,index] = max(dmin);
   m(i) = pool(index);
end
```

References

[1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.

[2] C.T. Mansfield, B.N. Barman, Anal. Chem. 71 (1999) 81R.

[3] M.D. Judge, Talanta 62 (2004) 675.

[4] M.P. Gomez-Carracedo, J.M. Andrade, M.A. Calvino, D. Prada, E. Fernandez, S. Muniategui, Talanta 60 (2003) 1051.

[5] K.R. Beebe, R.J. Pell, B. Seasholtz, Chemometrics—A Pratical Guide, Wiley, New York, 1998.

[6] M. Daszykowski, B. Walczak, D.L. Massart, Anal. Chim. Acta 468 (2002) 91.

[7] Y. Tominaga, Chemometr. Intell. Lab. Syst. 43 (1998) 157.

[8] F. Sales, A. Rius, M.P. Callao, F.X. Rius, Talanta 52 (2000) 329.

[9] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Chemometr. Intell. Lab. Syst. 33 (1996) 35.

[10] K.R. Kanduc, J. Zupan, N. Majcen, Chemometr. Intell. Lab. Syst. 65 (2003) 221.

[11] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.

[12] E. Bouveresse, C. Hartmann, D.L. Massart, I.R. Last, K.A. Prebble, Anal. Chem. 68 (1996) 982.

[13] L.F. Capitan-Vallvey, N. Navas, M. del Olmo, V. Consonni, R. Todeschini, Talanta 52 (2000) 1069.

[14] H.A.D. Dantas Filho, R.K.H. Galvão, M.C.U. Araújo, E.C. Silva, T.C.B. Saldanha, G.E. José, C. Pasquini, I.M. Raimundo Jr., J.J.R. Rohwedder, Chemometr. Intell. Lab. Syst. 72 (2004) 83.