Talanta 223 (2021) 121693

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Short communication

Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques

Puneet Mishra [a,*], Federico Marini [b], Alessandra Biancolillo [c], Jean-Michel Roger [d,e]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, 00185, Rome, Italy
[c] Department of Physical and Chemical Sciences, University of L'Aquila, Via Vetoio, 67100, Coppito, L'Aquila, Italy
[d] ITAP, INRAE, Institut Agro, University Montpellier, Montpellier, France
[e] ChemHouse Research Group, Montpellier, France

ARTICLE INFO

Keywords:
Multi-block data analysis
Data fusion
Spectroscopy
Preprocessing
Multivariate analysis
Fuel

ABSTRACT

Near-infrared (NIR) spectroscopy of fuels can suffer from scattering effects which may mask the signals corresponding to key analytes in the spectra. Therefore, scatter correction techniques are often used prior to any modelling so to remove scattering and improve predictive performances. However, different scatter correction techniques may carry complementary information so that, if jointly used, both model stability and performances could be improved. A solution to that is the fusion of complementary information from differently scatter corrected data. In the present work, the use of a preprocessing fusion approach called sequential preprocessing through orthogonalization (SPORT) is demonstrated for predicting key quality parameters in diesel fuels. In particular, the possibility of predicting four different key properties, i.e., boiling point (°C), density (g/mL), aromatic mass (%) and viscosity (cSt), was considered. As a comparison, standard partial least-squares (PLS) regression modelling was performed on data pretreated by SNV and 2nd derivative (which is a widely used preprocessing combination). The results showed that the SPORT models, based on the fusion of scatter correction techniques, outperformed the standard PLS models in the prediction of all the four properties, suggesting that selection and use of a single scatter correction technique is often not sufficient. Up to complete bias removal with 50% reduction in prediction error was obtained. The R^2_P was increased by up to 8%. The sequential scatter fusion approach (SPORT) is not limited to NIR data but can be applied to any other spectral data where a preprocessing optimization step is required.

1. Introduction

Non-destructive estimation of fuels properties with near-infrared (NIR) spectroscopy is of key importance as it is a rapid and cost-effective option [1]. Applications of NIR spectroscopy range from petroleum profiling and characterization [2], source and type identification [3], adulteration detection [4], oxidative stability detection [5], production process monitoring [6] and estimation of physical and chemical properties [7].

NIR spectroscopy, as other optical techniques, may suffer from spurious sources of variability in the signal brought by additional unwanted interactions of light with the samples [8]. These effects, the most relevant of which is light scattering, may mask the underlying spectral signal [9], so that scatter correction techniques are commonly used to try to remove them or, at least, reduce their impact [10]. In a traditional approach, a single scatter correction technique [11–13] is usually selected out of a shortlist of potential candidates. However, since the various scatter correction techniques operate differently from one another and, as a consequence, data preprocessed with different scatter correction techniques carry at least partially complementary information, the use of a single scatter correction method may lead to sub-optimal modelling [14,15]. Recently, Mishra et al. [14] showed that a fusion of complementary scatter correction techniques is the best solution as the information captured by differently scatter corrected data can be efficiently used.

Recently, a preprocessing fusion approach called sequential preprocessing through orthogonalization (SPORT) was developed [16]. The SPORT approach exploits the chemometric concept of multi-block data

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.talanta.2020.121693
Received 21 August 2020; Received in revised form 14 September 2020; Accepted 18 September 2020
Available online 24 September 2020

Table 1
A summary of the spectroscopic data sets used in this study.

Fuel properties	Calibration set		Test set	
	Spectroscopic data (samples × variables)	Reference values (mean ± std)	Spectroscopic data (samples × variables)	Reference values (mean ± std)
Boiling point (°C)	237 × 401	257 ± 23	158 × 401	260 ± 16
Density (g/mL)	237 × 401	0.84 ± 0.01	158 × 401	0.84 ± 0.01
Aromatic mass (%)	237 × 401	30 ± 7	158 × 401	31 ± 5
Viscosity (cSt)	237 × 401	2.53 ± 0.58	158 × 401	2.49 ± 0.42

analysis [17] and, in particular, takes inspiration from sequential and orthogonalized partial least-squares (SO-PLS) [18]. In SPORT, data preprocessed by means of different techniques constitute a multi-block data set, which can then be analyzed using SO-PLS. Accordingly, the SPORT approach can be used for the fusion of complementary scatter correction techniques by preprocessing the same data with several scatter correction techniques and later performing multi-block calibration (or classification, depending on the final aim of the study), as explained in detail in Ref. [14].

Starting from these considerations, the aim of the present work is to prove that NIR-based calibration models for predicting fuel properties can benefit from a complementary fusion of scatter correction techniques, in particular using the above-mentioned sequential approach called SPORT. In detail, a case study involving the use of NIR spectroscopy for the prediction of four different key properties in diesel fuels, namely boiling point (°C), density (g/mL), aromatic mass (%) and viscosity (cSt), was considered. To highlight the advantages resulting from the SPORT fusion approach, the outcomes were compared with those of standard PLS modelling on data pretreated by SNV followed by 2nd derivative (which is a common combination of spectral preprocessing for this kind of data).

2. Materials and methods

2.1. Data set

The data set used in the study is a benchmark data set for NIR-based calibration and consists of NIR spectra of diesel fuels along with the reference values of various properties of the samples. The data were obtained at SoutWest Research Institute (SWRI) and made available on the official website of Eigenvector, Inc, where they can be accessed at http://eigenvector.com/data/SWRI/index.html. The data set has been used as benchmark for testing various chemometric approaches, in particular for variable selection: for instance, the data have been analyzed to evaluate the performances of a newly defined criterion for variable importance [19] or of approaches such as spectral clustering-based interval partition (SCIP) [20], moving-window-improved Monte Carlo uninformative variable elimination (MC-UVE-PLS) [21], genetic inverse least squares (GILS) [22]. On the other hand, the same data set has also been used to evaluate the performances of regression approaches other than standard partial least squares, e.g., extreme learning machine (ELM) [23] or wavelet packet consensus interval partial least squares (WpCo-iPLS) [24]. Lastly, the problems connected to preprocessing have only been addressed by investigating the effect of a strategy based on the calculation of fractional order derivatives [25], so that the present study is the first to systematically evaluate how combining multiple preprocessing strategies into a fused approach could improve the quality of calibration models.

A total of 395 samples were analyzed and the spectra were collected in the range 750–1550 nm. Four different properties were considered in this study, i.e, boiling point (°C), density (g/mL), aromatic mass (%) and viscosity (cSt). The boiling point was measured at 50% recovery using ASTM D86 approach. The density was measured at 15 °C using the ASTM D4052 approach. The total aromatic mass was measured using the ASTM D5186. The viscosity was measured at 40 °C. The samples were further portioned into calibration (60%) and test (40%) set using the Kennard-Stone algorithm [26]. A summary of the data is further provided in Table 1.

2.2. Data analysis

2.2.1. Scatter correction methods

Four different scatter correction methods were considered to pretreat the data prior to SPORT fusion. The first method was multiplicative scatter correction (MSC), which models the spectra as a mixture of scattering and absorbance [27]. The second method was the standard normal variate (SNV) [28] transform in which, for a given spectrum, offset correction is achieved by subtracting the mean intensity while reducing the multiplicative effect is achieved by division by the standard deviation. The third method was variables sorting for normalization (VSN) [29] which assumes that not all the bands are equally altered by the unwanted effects and, consequently, assigns to each variable a weight in the range [0,1] corresponding to its probability of being affected only by scattering. VSN calculates the weights based on random consensus (RANSAC) algorithm which estimates to what extent a wavelength is affected by size effects (additive and multiplicative offsets) rather than by shape effects (features ascribable to chemically relevant contributions). In this way, variables that are strongly related to chemical constituents have a low weight and negligible role in the calculation of the size effect. The main benefit of the VSN approach in comparison to MSC is that it does not require any reference spectrum to estimate the weights. In the present work, the weights estimated by VSN (assuming a multiplicative effect and a constant offset and automatically optimizing the RANSAC tolerance based on the maximum variance of the weights criterion described in Ref. [29]) were integrated in SNV leading to a weighted SNV. The fourth method was the calculation of 2nd derivative, which is commonly used to remove both additive and multiplicative effects [10]. Numerical differentiation, i.e, calculation of the second derivative, was performed using the Savitzky-Golay approach (2nd order polynomial + 21-point window). All the pre-processing methods were implemented in MATLAB 2018b (The Mathworks, Natick, MA, USA) using the MBA-GUI [30].

2.2.2. Partial least-squares (PLS) regression

PLS regression is a common latent space-based chemometric method [31] widely used for NIR data modelling [32]. PLS deals with the multicollinearity in the NIR data by projecting the data onto a subspace of latent variables (LVs) which are extracted so to have maximum covariance with the response(s). This guarantees that the scores are at the same time explanatory of the variance in NIR data, and relevant for predicting the response variables. In the study, PLS models have been calculated by means of MATLAB's built-in function 'plsregress', combined with a 10-fold cross validation procedure to select the optimal number of latent variables (LVs).

2.2.3. Sequential preprocessing through orthogonalization (SPORT)

SPORT is a preprocessing fusion approach directly inspired by sequential and orthogonalized partial least squares (SO-PLS) modelling, and it consists in building a multi-block data set made of different versions of the experimental NIR data, each one pretreated according to a specific preprocessing approach, and then using this multi-block data to build a calibration model for the prediction of the property(-ies) of interest by means of SO-PLS [16]. A schematic illustration of the SPORT

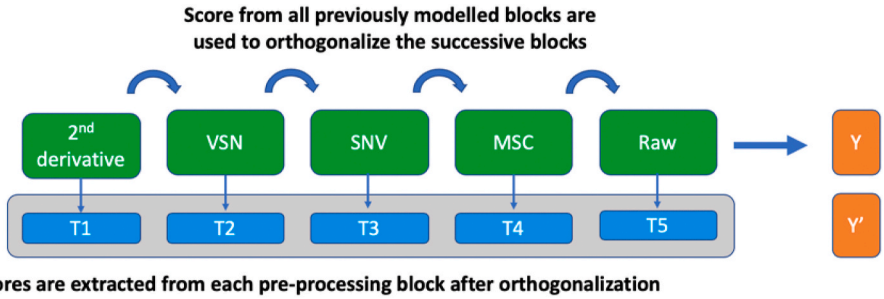

Fig. 1. Schematic illustration of the SPORT approach for sequential fusion of preprocessings. T1 to T5 represent the scores extracted from each block of data. Y and Y′ correspond to the measured and predicted responses, respectively.

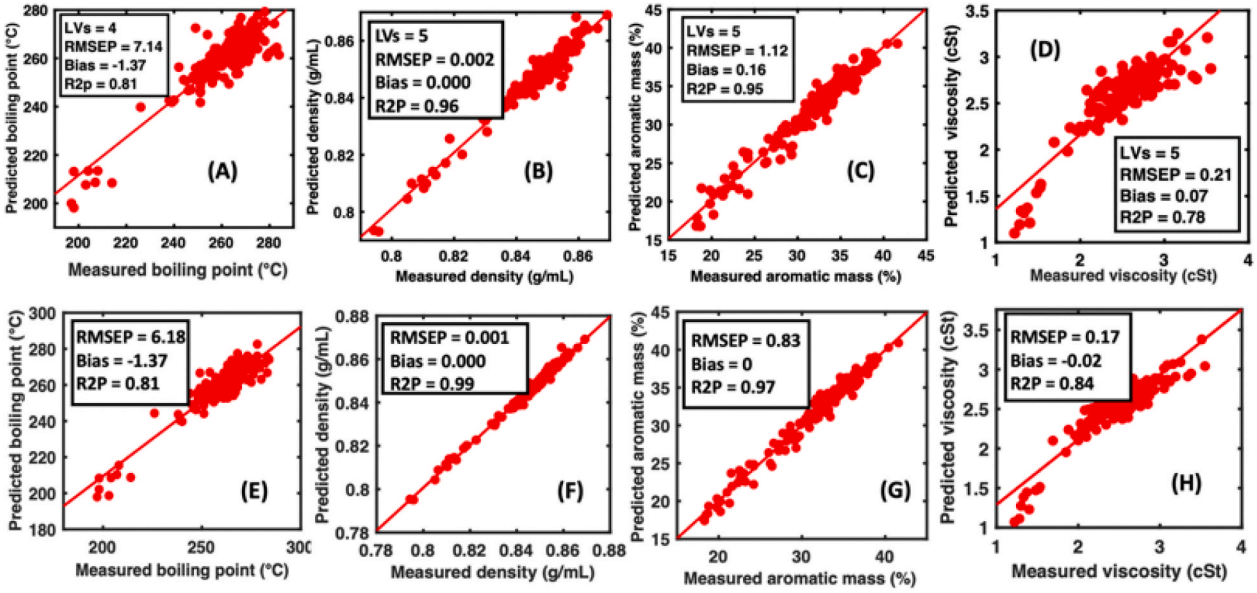

Fig. 2. Graphical representation of the performances of PLS (upper panel) and SPORT (lower panel) on the test set samples for the prediction of the four fuel properties considered in the present study: (A) and (E) boiling point (°C); (B) and (F) density (g/mL); (C) and (G) aromatic mass (%); (D) and (H) viscosity (cSt).

approach is presented in Fig. 1. Initially, a PLS regression model is fitted between the **Y** and the first pre-processed block, obtaining, among other information, the scores for the first block (T1). Then, the second block is orthogonalized with respect to the scores (T1) of the first regression and the residuals of **Y** are fitted to the orthogonalized second block. The procedure is continued for as many blocks as there are pretreatments. The number of LVs to be extracted from each block is usually optimized in cross-validation using a global approach; all possible combinations of LVs are tested and the optimal one is the one resulting in the lowest RMSECV. In the present work, the preprocessing order used for the sequential fusion was 2nd derivative, VSN, SNV, MSC and then raw data (resulting in a total of 5 blocks of predictors), and the SPORT was implemented using the freely available MBA-GUI [30].

3. Results

3.1. PLSR modelling vs SPORT

The results from PLS and SPORT modelling are graphically summarized in Fig. 2. For all the four properties, the SPORT approach attained a higher R^2_p compared to the standard PLS modelling. In the case of boiling point, the improvement was mainly noted in the value of the RMSEP, which was reduced by 14%; on the other hand, as far as density is concerned, the RMSEP was decreased by 50% and the R^2_p was improved by 3%. For the prediction of the aromatic mass, the RMSEP was reduced by 26%, the R^2_p was increased by 2% and the bias was completely removed. Lastly, when considering the viscosity, the RMSEP was reduced by 20%, the R^2_p was increased by 8% and the bias was reduced by 71%.

Table 2
A summary of the optimal number of latent variables extracted from differently scatter-corrected data by SPORT.

Fuel properties	2nd derivative	VSN	SNV	MSC	Raw
Boiling point	3	5	0	0	0
Density	0	8	0	10	0
Aromatic mass	7	1	0	0	0
Viscosity	0	0	4	0	3

3.2. Complementary information captured by SPORT

By examining more in detail the characteristics of the models, it is apparent how the better performances observed when applying the SPORT approach, compared to "standard" PLS calibration, can very likely be ascribed to the fusion of complementary information from differently scatter-corrected data. Indeed, since the orthogonalization steps removes redundancies among the blocks, complementarity is associated to the fact that more than a single block contribute to the final model with a non-zero number of latent variables. In particular, the

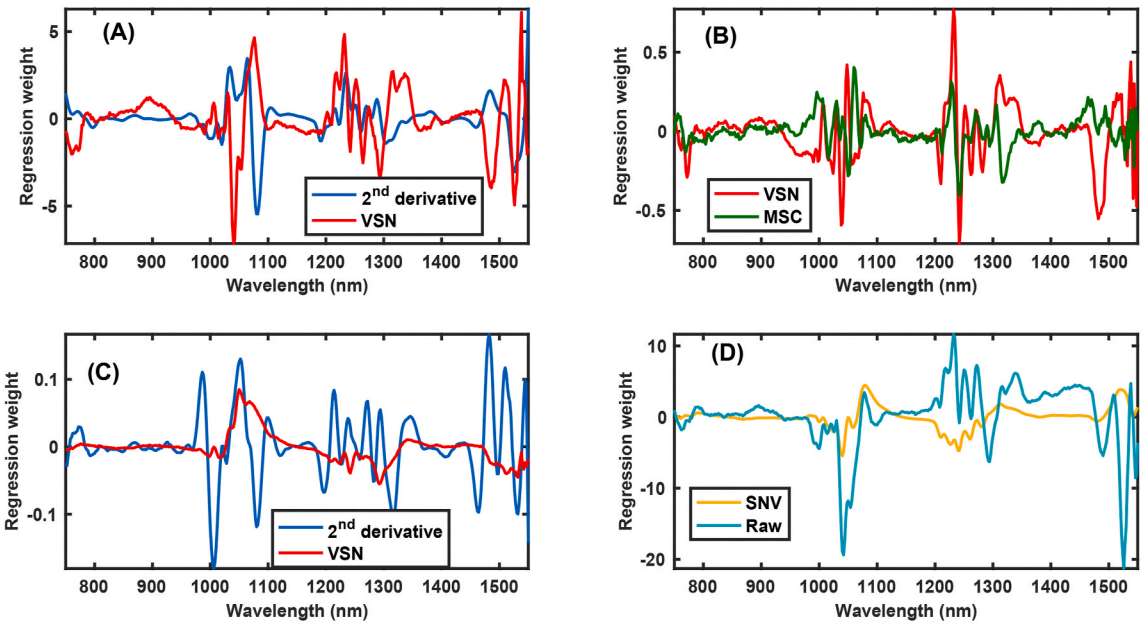

Fig. 3. Regression vectors from SPORT. (A) Boiling point (2nd derivative in blue and VSN in red), (B) density (VSN in red and MSC in green), (C) aromatic mass (2nd derivative in blue and VSN in red), and (D) viscosity (SNV in yellow and raw data in cyan). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

number of LVs from each block corresponding to the optimal models for the prediction of the four diesel fuel properties is shown in Table 2. By looking at the table, it is immediately evident how, for all the properties, the optimal models were obtained by using information extracted from at least two scatter correction techniques: 2nd derivative and VSN in the case of boiling point and aromatic mass, VSN and MSC for density. In the case of viscosity, prediction was anyway improved by integrating the raw data with SNV only.

Such complementarity can also be observed when inspecting the regression vectors of the four SPORT models built for the prediction of the individual fuel properties, which are displayed in Fig. 3. At a first glance, indeed, it can be noted that, for each SPORT model, the regression vectors associated to the individual preprocessing blocks (of course, only those with non-zero LVs) are different both in magnitude and in shape: when comparing the different regression vector for the prediction of a particular property, there are regions with non-zero coefficients only for a preprocessing and not for the other, and there are also wavelength intervals where the regression vectors have similar shape but are slightly shifted. Similar pattern for complementary regression vectors was also noticed by Mishra et al., 2020 [14].

4. Conclusions

The work proves that the fusion of different complementary scatter correction techniques is essential for building optimal models for predicting fuel properties with NIR spectroscopy. In particular, as demonstrated in this study, the SPORT approach allows automatic extraction of the complementary information and, therefore, in the future, its use in NIR modelling is highly recommended to the scientific community. A complementary fusion of scatter correction techniques with SPORT also has the benefit that it takes the user out of the loop of identifying the best preprocessing by developing several models, each based on a candidate option, and then selecting the optimal one, thus, allowing to save time and resources. On the other hand, a drawback of the SPORT approach is that its performances or optimization may be affected by the order in which the preprocessing blocks are arranged. However, in this context it should also be pointed out that while the order of the blocks may influence the selected preprocessings and, in general, the number of latent variables extracted from each matrix, on the other hand, the predictive performances have been shown not to be influenced relevantly [18]. To achieve optimal performances, it is recommended that powerful scatter correction methods such as VSN, SNV, 2nd derivative and MSC should be given priority. If time is a constraint, then user may arrange the faster and model-free approaches such as SNV or 2nd derivative as the first ones, leaving model-based methods such as VSN or MSC to a later stage. Anyway, it is worth to be pointed out that the SPORT approach is not limited to NIR data, but it can be used integrate multiple preprocessing technique with any spectroscopic (or even instrumental) data.

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Data curation, Investigation. **Federico Marini:** Formal analysis, Software, Visualization. **Alessandra Biancolillo:** Formal analysis, Methodology, Software. **Jean-Michel Roger:** Software, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z.S. Baird, V. Oja, Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density, Chemometr. Intell. Lab. Syst. 158 (2016) 41–47.
[2] B.P.O. Lovatti, S.R.C. Silva, N.d.A. Portela, C.M.S. Sad, K.P. Rainha, J.T.C. Rocha, W. Romão, E.V.R. Castro, P.R. Filgueiras, Identification of petroleum profiles by infrared spectroscopy and chemometrics, Fuel 254 (2019) 115670.
[3] R.M. Balabin, R.Z. Safieva, Gasoline classification by source and type based on near infrared (NIR) spectroscopy data, Fuel 87 (2008) 1096–1101.
[4] E.M. Paiva, J.J.R. Rohwedder, C. Pasquini, M.F. Pimentel, C.F. Pereira, Quantification of biodiesel and adulteration with vegetable oils in diesel/biodiesel blends using portable near-infrared spectrometer, Fuel 160 (2015) 57–63.
[5] F.S. Vieira, C. Pasquini, Determination of the oxidative stability of biodiesel using near infrared emission spectroscopy, Fuel 117 (2014) 1004–1009.
[6] R. Sales, N.C. da Silva, J.P. da Silva, H.H. França, M.F. Pimentel, L. Stragevitch, Handheld near-infrared spectrometer for on-line monitoring of biodiesel production in a continuous process, Fuel 254 (2019) 115680.

[7] C.L. Cunha, A.R. Torres, A.S. Luna, Multivariate regression models obtained from near-infrared spectroscopy data for prediction of the physical properties of biodiesel and its blends, Fuel 261 (2020) 116344.
[8] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.
[9] H. Martens, J.P. Nielsen, S.B. Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures, Anal. Chem. 75 (2003) 394–404.
[10] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing methods, in: S. D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Oxford, UK, 2020, pp. 1–75.
[11] J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel, L.M.C. Buydens, Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103.
[12] J. Torniainen, I.O. Afara, M. Prakash, J.K. Sarin, L. Stenroth, J. Toyras, Open-source python module for automated preprocessing of near infrared spectroscopic data, Anal. Chim. Acta 1108 (2020) 1–9.
[13] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M. C. Buydens, Breaking with trends in pre-processing? Trac. Trends Anal. Chem. 50 (2013) 96–106.
[14] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020) 111271.
[15] Puneet Mishra, et al., New data preprocessing trends based on ensemble of multiple preprocessing techniquesAuthor links open overlay panel, Trends Anal. Chemi. (2020), 116045, https://doi.org/10.1016/j.trac.2020.116045. In press.
[16] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometr. Intell. Lab. Syst. 199 (2020) 103975.
[17] A.K. Smilde, I. Måge, T. Næs, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, J. Chemometr. 31 (2017), e2900.
[18] A. Biancolillo, T. Næs, The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, in: M. Cocchi (Ed.), Data Fusion Methodologies and Applications, Data Handling in Science and Technology, vol. 31, Elsevier, Oxford, UK, 2019, pp. 157–177.
[19] J. Zhang, X. Cui, W. Cai, X. Shao, A variable importance criterion for variable selection in near-infrared spectral analysis, Sci. China Chem. 62 (2019) 271–279.
[20] Y. Xiong, R. Zhang, F. Zhang, W. Yang, Q. Kang, W. Chen, Y. Du, A spectra partition algorithm based on spectral clustering for interval variable selection, Infrared Phys. Technol. 105 (2020) 103259.
[21] W. Jiang, C. Lu, Y. Zhang, W. Ju, J. Wang, F. Hong, T. Wang, C. Ou, Moving-window-improved Monte Carlo uninformative variable elimination combining successive projections algorithm for near-infrared spectroscopy (NIRS), J. Spectrosc. (2020) 3590301.
[22] D. Ozdemir, Near infrared spectroscopic determination of diesel fuel parameters using genetic multivariate calibration, Petrol. Sci. Technol. 26 (2008) 101–113.
[23] X.-H. Bian, S.-J. Li, M.-R. Fan, Y.-G. Guo, N. Chang, J.-J. Wang, Spectral quantitative analysis of complex samples based on the extreme learning machine, Anal. Lett. 8 (2016) 4674–4679.
[24] D. Peng, H. Guo, L. Li, Y. Bi, G. Yang, Using consensus strategy and interval partial least square algorithm in wavelet domain for analysis of near-infrared spectroscopy, Adv. Eng. Res. 153 (2018) 113–119.
[25] K.-Y. Zheng, X. Zhang, P.-J. Tong, Y. Yao, Y.-P. Du, Pretreating near infrared spectra with fractional order Savitzky–Golay differentiation (FOSGD), Chin. Chem. Lett. 26 (2015) 293–296.
[26] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[27] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[28] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[29] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: variable sorting for normalization, J. Chemometr. 34 (2020) e3164.
[30] P. Mishra, J.M. Roger, D.N. Rutledge, A. Biancolillo, F. Marini, A. Nordon, D. Jouan-Rimbaud-Bouveresse, MBA-GUI, A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing, Chemometr. Intell. Lab. Syst. 205 (2020) 104139.
[31] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[32] W. Saeys, N.N. Do Trong, R. Van Beers, B.M. Nicolai, Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review, Postharvest Biol. Technol. 158 (2019) 110981.