Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Platforms for Automatic PAT Soft Sensor Development and Analysis

Marco S. Reis *. Tiago J. Rato *

** CIEPQPF - Department of Chemical Engineering, University of Coimbra*
Pólo II, Rua Sílvio Lima, 3030-790 Coimbra, Portugal (e-mail: marco@eq.uc.pt)

Abstract: The performance of soft sensors from spectroscopic Process Analytical Technology data is directly related to the type of modelling methodology and preprocessing technique used during model development. However, their selection is often decoupled and based on simple trial and error procedures. Furthermore, the current modelling methodologies focus solely on selecting the most informative wavebands and do not attempt to enhance the prediction capabilities within each waveband. To overcome these limitations, two frameworks have been proposed for (i) optimal feature selection and (ii) systematic comparison of multiple combinations of modelling methodologies and preprocessing techniques: MR-SS and SS-DAC. The Multiresolution Soft Sensor (MR-SS) optimizes the resolution of each waveband by spectral aggregation, generally leading to models with superior performance than their single-resolution counterparts of the same model class. The Soft Sensor Development, Assessment and Comparison (SS-DAC) framework, selects the best combination of modelling methodologies and preprocessing techniques by use of structured randomization and rigorous statistical analysis of the overall prediction merits of the different models. The analytical and predictive merits of these two proposed methodologies are illustrated on a case study of real near infrared (NIR) spectra.

Keywords: Soft sensors, Automatic model development, Forward stepwise selection, Interval selection, Resolution selection, Spectral preprocessing, AutoML.

1. INTRODUCTION

The development of soft sensors using data collected by Process Analytical Technology (PAT) systems (e.g., near infrared (NIR), Fourier-transform infrared (FT-IR), Raman and fluorescence spectroscopy) plays an important role on bringing quality assessment to real time. They consist of combining spectroscopic data collectors and advanced modelling methods in order to build solutions able to infer properties that otherwise would require complex protocols, expensive equipment and considerable time from highly specialized personnel. However, the development of PAT soft sensors is a challenging task, given the high dimensionality of the predictive space, associated to high levels of redundancy (multicolinarity) and sparsity (not all predictors are relevant). Furthermore, collected spectra are often contaminated by noise and chemical/physical interferences that degrade the accuracy of the models. Therefore, soft sensor development requires a proper selection of the most appropriate modelling methodology and preprocessing technique to address the specificities of the data under analysis. To streamline this selection, a Soft Sensor Development, Assessment and Comparison (SS-DAC) framework was proposed in (Rato and Reis, 2019b). In summary, SS-DAC starts by generating a battery of models by combining different preprocessing techniques and modelling methodologies. Afterwards, the accuracy of each model is assessed on an independent test dataset. Finally, statistical hypothesis tests are used to quantify the relative performance of each model. This step also computes a prediction score that easily identifies the most accurate model as well as the impact of each preprocessing and modelling methodology on the overall performance of each model.

Another critical aspect while developing soft sensors from PAT data is the selection of the relevant subsets of wavelengths to include. In this regard, several variable selection methodologies have been proposed. A review of these methodologies is provided in Refs. (Balabin and Smirnov, 2011; Pasquini, 2018; Wang et al., 2018; Yun et al., 2019). These methodologies are however solely focused on selecting the best subset of wavelengths or wavebands and do not attempt to enhance the prediction capabilities of each interval before applying a modelling methodology. This is a problem that is traversal to all modelling methodologies, such as partial least squares (PLS), support vector regression (SVR) and artificial neural networks (ANN); see e.g. (Soares and Anzanello, 2018; Xu et al., 2018). To overcome this limitation, a Multiresolution Soft Sensor (MR-SS) modelling framework is used to simultaneously select the relevant intervals and the optimal resolution for each interval.

The analytical and predictive capabilities of MR-SS and SS-DAC are illustrated on a case study of real near infrared (NIR) spectra from gasoline samples, with the aim to predict the octane number (the octane number is a standard measure of the performance of an engine or aviation fuel). The synergistic application of the two frameworks allows for an efficient development of robust soft sensors with high prediction performance.

2. METHODOLOGY

2.1 SS-DAC: Soft Sensor Development, Assessment and Comparison

The SS-DAC framework is designed to simultaneously compare models with different preprocessing techniques and modelling methodologies. SS-DAC consists of three main stages addressing (i) the selection of the hyperparameters for each model class, (ii) the assessment of the models' performance and (iii) comparison of the models' relative performance through rigorous statistical hypothesis tests. Afterwards, a series of graphical displays are produced to summarize the results and guide the user on the selection of the best model or subset of models. A schematic representation of the stages involved in SS-DAC is presented in Fig. 1.

In the first stage of SS-DAC, the models' hyperparameters (e.g., the number of retained latent variables, relevant intervals and resolution of each interval) are determined by minimizing the mean squared error (MSE) of Monte Carlo cross-validation (MCCV). MCCV is performed by splitting the training dataset into C random pairs of calibration and validation datasets (see more details in Subsection 2.2).

The second stage of SS-DAC concerns the assessment of the models' accuracy. This assessment is done by resort to random training and test subsets. The random training subsets are built by randomly remove five samples from the original training dataset. These training subsets are then used to retrain the models (using the hyperparameters selected in the first stage of SS-DAC) and thus assess the effects of variations on the models' parameters onto their performance. Likewise, random test subsets are built by randomly divide the original test dataset in folds. Please note that the test dataset (and by consequence the random test subsets) is never used to train the models. Therefore, the models' accuracy and the associated variability due to changes on the response variable is evaluated on completely new data. For this reason, any misspecification that may occur on the first stage of SS-DAC (e.g., overfitting) will be penalized in the assessment stage of SS-DAC. The models' accuracy is then measured through the prediction MSE for each combination of training and test subsets. Thus, for each model a $(Q \times R)$ matrix of prediction MSEs is obtained, where Q is the number of training subsets and R is the number of test subsets.

The resort to multiple test subsets to assess the models' performance is one of the distinguishing characteristics of SS-DAC. Current methodologies, such as those based on an analysis of variance (ANOVA) (Aguado-Sarrió et al., 2017; Galdón-Navarro et al., 2018), use the overall MSE to compare the models. However, the overall MSE may be biased by extreme (either high or low) local performances, which may lead to selecting a model that is very good in a few test subsets, but mediocre in most of the remaining test subsets (the converse situation may also happen). In turn, SS-DAC assesses the models' performance in each test subset and subsequently searches for the model with statistically lower MSE in most of the test subsets.

In the third stage of SS-DAC, the prediction MSEs within each test subset (i.e., each column of the MSE matrix) are compared through a series of paired t-test. This procedure effectively produces a table of paired-wise statistics and p-values that could be used to select the best models. However, since the number of comparisons can be high, this type of assessment is infeasible. To overcome this situation, the outputs of the statistical tests are summarized by a predictive score attributed to each model. To achieve this, SS-DAC records the number of times that a given model leads to statistically lower (*victory*), equal (*tie*) and higher (*loss*) prediction MSEs. Finally, a predictive score for each model is computed by summing its number of victories and ties across all test subsets. As each model is compared against all other models over each test subsets, the maximum score that a model can achieve is $(PM - 1)R$, where P is the number of preprocessing techniques, M is the number of modelling methodologies and R is the number of test subsets. Note that the predictive score relates to the model's ability to produce statistically lower MSE on an independent test dataset. Thus, models with higher scores (i.e., more victories) are deemed more robust and accurate. For a further discussion on the comparison methodology and its relationship with a typical ANOVA comparison (Galdón-Navarro et al., 2018) we refer the reader to Ref. (Rato and Reis, 2019b).

The SS-DAC framework can be applied to any combination of preprocessing technique (Rinnan et al., 2009) and modelling methodology, such as PLS, SVR and ANN (Pasquini, 2018; Rendall et al., 2017).

2.2 MR-SS: Multiresolution Soft Sensor

MR-SS is a modelling framework (Rato and Reis, 2019a) that aims to simultaneously select the relevant wavelength intervals and the optimal resolution for each interval. The interval's resolution relates to the level of aggregation produced by a binning operator that replaces the measurements within a waveband by their average. The highest resolution corresponds to the original signal (i.e., without aggregation), while lower resolution levels correspond to averages over progressively longer wavebands, defined as *support bands*. Without loss of generality, the *support band* is assumed to have a dyadic length of 2^q consecutive wavelengths, where $q \geq 0$ is an integer representing the resolution level. The search for the optimal resolution is made between the highest resolution of the signal ($q = 0$) and a maximum resolution defined by the user (q_{max}). From our accumulated experience, a maximum resolution of 5 is often a good starting point. Furthermore, for parsimony with the dyadic structure, the length of the intervals used by MR-SS should be a multiple of 2^{qmax}. Following these definitions, a multiresolution forward stepwise selection (MR-FSS) algorithm (Rato and Reis, 2019a) is applied to simultaneously select the best intervals and their optimal resolution.

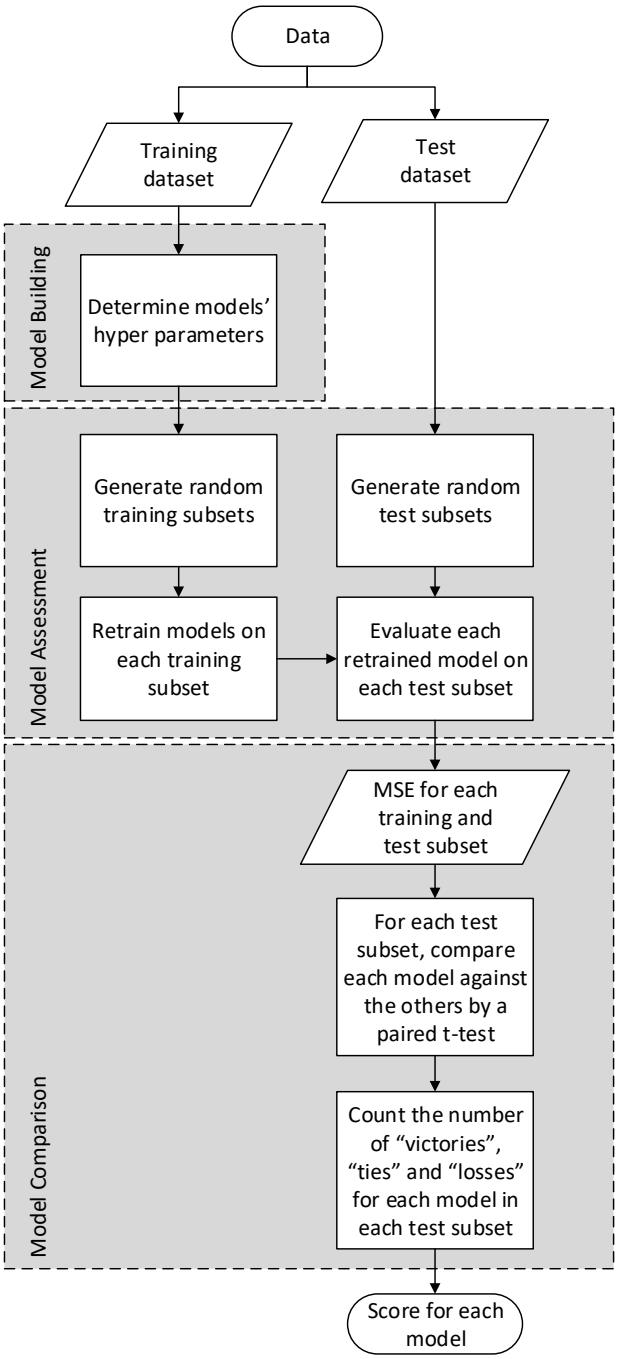

Fig. 1. Diagram of the model building, assessment and comparison stages of SS-DAC.

The MR-FSS algorithm starts by selecting an initial reference model by building tentative models for each combination of interval and resolution level. (The tentative models can be fit through any modelling methodology, e.g., PLS.) For each tentative model, the MSE of MCCV is determined for further comparison. MCCV is performed by randomly divide the training dataset into C pairs of calibration and validation datasets. Each MCCV randomization is made by sorting the response variable and subsequently dividing the samples into g groups with p samples each. Afterwards, a calibration dataset is built by randomly take $k < p$ samples from each group. The respective validation dataset is composed by the remaining samples. In this study, we set $p = 5$ and $k = 4$, thus obtaining a typical 80/20 split of the data. For each MCCV randomization, the model is trained on the calibration dataset and its MSE is evaluated on the validation dataset. This operation produces a $(C \times 1)$ vector of MSEs. Based on this information, the model with lowest median MSE is selected as the initial reference model.

Afterwards, MR-FSS proceeds to an adding stage where intervals are tentatively added to the reference model. Each interval is also tested over all allowed resolution levels. The vector of MSEs of each tentative model is then statistically compared with the MSEs of the reference model. Among the models with statistically lower MSEs, the model with the lowest median MSE is selected as the new reference model (note that this operation inherently selects the best combination of interval and resolution). Next, the selection algorithm tentatively removes one interval from the reference model following the same selection criterion. Finally, the adding and removing stages are repeated until no interval is either added or removed from the reference model.

The main stages of MR-SS are represented in Fig. 2. For further information about MR-SS we refer the reader to Ref. (Rato and Reis, 2019a). This methodology can be used with any type of modelling methodology, but for illustration purposes it is here applied with PLS.

3. CASE STUDY

To show the use of the SS-DAC and MR-SS frameworks, they were applied to the gasoline dataset presented in (Kalivas, 1997). This dataset is readily available in statistical software such as Matlab®, JMP® or R. The dataset consists of 60 gasoline samples with known octane numbers (the response). Their NIR spectra were measured using diffuse reflectance as $\log(1/R)$ from 900 to 1700 nm in 2 nm intervals (leading to measurements over 401 wavelengths). To proceed with the training and comparison of the models, the raw data was randomly divided into a training dataset with 40 samples and a test dataset with 20 samples.

For illustration purposes, SS-DAC is applied over three preprocessing techniques based on mean centering, multiplicative scatter correction (MSC) (Geladi et al., 1985), and standard normal variate (SNV) (Barnes et al., 1989), see Table 1. For further details on the theoretical foundations of these preprocessing techniques we refer the readers to Refs. (Naes et al., 2002; Rinnan et al., 2009). Other preprocessing techniques can also be considered.

As for the modelling methodologies, we focus on methodologies based on PLS due to their widespread use and familiarity to most practitioners. Nevertheless, we highlight that other modeling methodologies can also be used. Among the current approaches, the standard implementation of PLS (Geladi and Kowalski, 1986; Wold et al., 2001) over the entire spectrum and its variants with forward (FiPLS) (Xiaobo et al., 2007) and backward (BiPLS) (Leardi and Nørgaard, 2004; Xiaobo et al., 2007) interval selection are considered (see Table 2). Likewise, a multiresolution interval

PLS (MR-iPLS) model obtained through the application of MR-SS is also considered.

By combining the aforementioned preprocessing techniques ($P = 3$) and modelling methodologies ($M = 4$), a total of 12 models are obtained. To facilitate their identification, a unique compound name is generated through to the following formula: {*Identifier of preprocessing*}{*Identifier of modelling methodology*}:{*Short name of preprocessing*}–{*Short name of modelling methodology*}.

In this study, the hyperparameters of each model (i.e., the number of retained latent variables, relevant intervals and optimal resolution for each interval) were selected by MCCV using $C = 200$ randomizations of the training dataset. These randomizations were made once and used to train all models. For the models with interval selection, the spectral measurements were divided into 12 intervals with 32 wavelengths and one interval with 17 wavelengths. Furthermore, the maximum resolution for MR-iPLS (q_{max}) was set to 5.

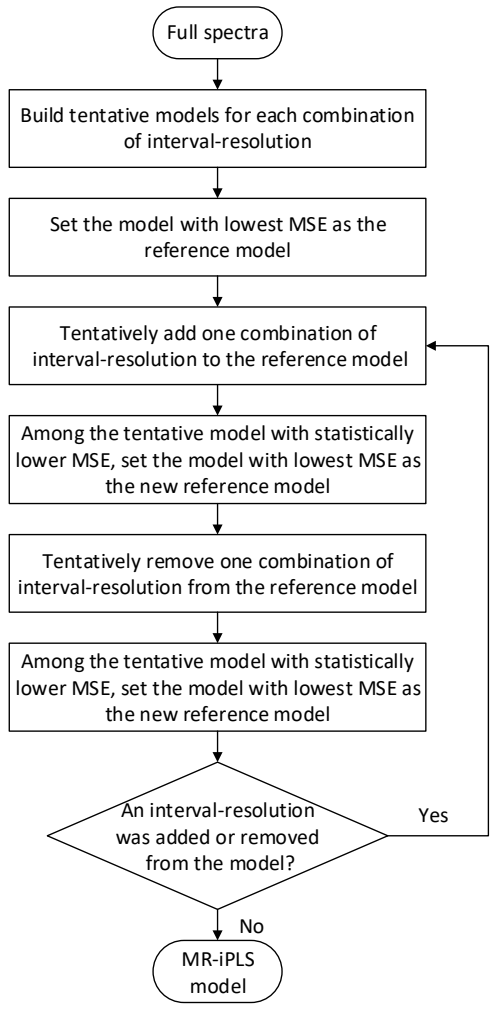

Fig. 2. Diagram of the selection algorithm used by MR-SS.

Table 1. Preprocessing techniques under consideration.

Identifier	Short Name	Type of preprocessing
1	MC	Mean Centering
2	SNV	Standard Normal Variate
3	MSC	Multiplicative Scatter Correction

Table 2. Modelling methodologies under consideration.

Identifier	Short Name	Description
A	PLS	Partial Least Squares
B	FiPLS	Forward interval Partial Least Squares
C	BiPLS	Backward interval Partial Least Squares
D	MR-iPLS	Multiresolution interval Partial Least Squares

After setting the hyperparameters of each model, their relative performance was compared using the SS-DAC methodology. In this study, SS-DAC was applied with $Q = 10\,000$ randomizations of the training dataset and $R = 10$ randomizations of the test dataset. Furthermore, it is noted that for the current comparison settings (i.e., number of models and test subsets), the maximum score that a model can achieve is 110. The results obtained through SS-DAC are presented in Fig. 3 and 4. For comparison purposes, the prediction MSE obtained using the original training and test datasets (i.e., without randomization) are provided in Table 3. Please recall that the prediction MSEs used to compare the models' performance are based on randomizations of an independent test dataset that was never used to train the models. Furthermore, the prediction score of each model corresponds to the number of paired-wise statistical tests where its prediction MSEs were found to be statistically lower (*victory*) or equal (*tie*) than the prediction MSEs of other models. Therefore, a higher prediction score is related to a higher count of statistically lower prediction MSEs.

By analysis of the results provided by SS-DAC it is verified that the best prediction capabilities are obtained with preprocessing 1:MC (see Fig. 4). Furthermore, it is observed that 2:SNV and 3:MSC significantly reduce the prediction capabilities of A:PLS and B:FiPLS. In other words, the use of 2:SNV and 3:MSC in combination with A:PLS and B:FiPLS produces models with statistically higher prediction MSEs, which in turn leads to a lower count of *victories* in the statistical tests of these models.

The low performance of A:PLS is related to its inability to exclude irrelevant intervals. In turn, the low performance of B:FiPLS models is caused by local optima that prevent the inclusion of additional intervals into the models. Compared to C:BiPLS, this behavior suggests that individual intervals are not informative, but a combination of intervals is. However, since B:FiPLS only tests the inclusion of one interval at a time, these intervals are never added to the models.

Table 3. Prediction MSE for the models under consideration. The best model for each preprocessing technique is in bold. The best model for each modelling methodology is underlined.

Modelling Methodology	Preprocessing Technique		
	1:MC	2:SNV	3:MSC
A:PLS	0.0466	0.0634	0.0646
B:FiPLS	0.0510	0.0598	0.0597
C:BiPLS	0.0470	**0.0511**	0.0518
D:MR-iPLS	**0.0451**	0.0549	**0.0510**

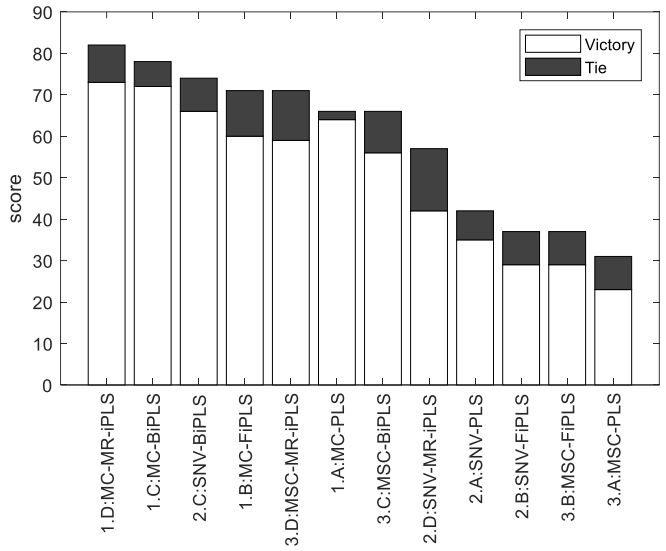

Fig. 3. Predictive scores of the models under consideration.

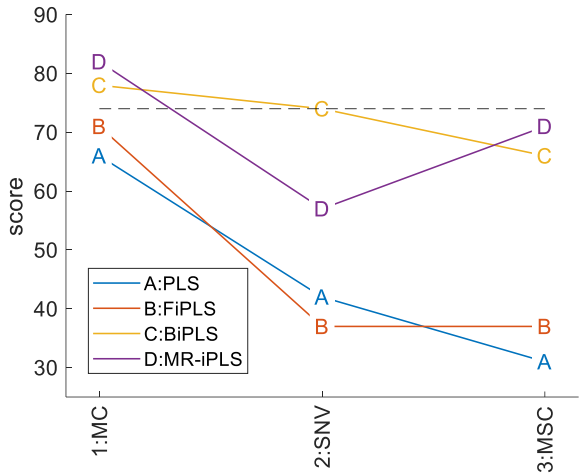

Fig. 4. Predictive score of each modelling methodology stratified by preprocessing technique.

As for D:MR-iPLS, it is verified that it leads to considerably higher predictive scores (i.e., statistically lower prediction MSEs) in most of the preprocessing techniques, being only surpassed by C:BiPLS on preprocessing 2:SNV. In this regard, it is noted that a direct comparison of 2.C:SNV-BiPLS and 2.D:SNV-MR-iPLS using a standard paired *t*-test on the prediction MSEs (see Table 3) already points to the overall performance of the two models. However, by comparing 2.C:SNV-BiPLS and 2.D:SNV-MR-iPLS through the SS-DAC methodology further insights on their local performance are obtained. In this case, it is verified that 2.C:SNV-BiPLS has statistically lower MSEs in 6 out of 10 test subsets (i.e., 2.C:SNV-BiPLS *wins* 6 times against 2.D:SNV-MR-iPLS). Furthermore, these models have statistically equal MSEs in 2 test subsets (i.e., 2.C:SNV-BiPLS and 2.D:SNV-MR-iPLS *tie* in 2 test subsets). Thus, while 2.C:SNV-BiPLS is generally better than 2.D:SNV-MR-iPLS, 2.C:SNV-BiPLS is closely followed by 2D:SNV-MR-iPLS.

Along with a higher prediction performance, selecting the optimal resolution for each waveband also reduces the number of variables in the D:MR-iPLS models. For instance, for the best D:MR-iPLS model (achieved with preprocessing 1:MC) the selected wavebands are all placed at the lowest resolution level (see Fig. 5). (By coincidence, all intervals are at the same resolution, but this is not a requirement: intervals may by at different resolutions from each other.) Therefore, the original 288 wavelengths are replaced by just 9 variables corresponding to the average value within each interval. In contrast, the best model using the standard modelling methodologies (1.C:MC-BiPLS) has 256 variables. A similar situation is observed for other preprocessing techniques as the selected wavebands by D:MR-iPLS are often placed at resolution level 4 or 5.

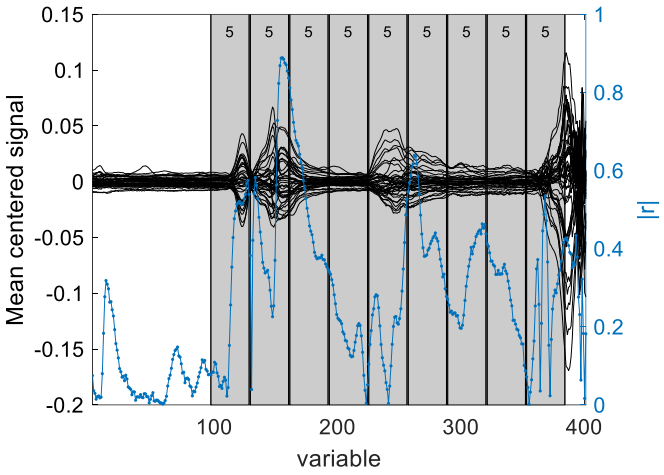

Fig. 5. Selected intervals (in grey) and their resolution for 1.D:MC-MR-iPLS. The absolute value of the correlation between each variable and the response is also shown.

4. CONCLUSIONS

In this work, new process analytic tools are presented for (i) improving the performance of prediction models and (ii) automate the comparison and selection of prediction models.

Regarding the modelling methodologies, it is proposed to enhance the current modelling methodologies by simultaneously select the most informative wavebands and optimize the resolution of each waveband. In this context, a change in resolution corresponds to a binning or spectral aggregation operation, which is a well-known operation to reduce the number of predictors and increase the signal-to-noise ratio. However, this operation was not yet integrated

with waveband selection. As MR-SS only lowers the resolution of an interval if there is statistical evidence favouring lower MSEs, it produces models that are at least as good as their single-resolution counterparts of the same model class. This situation is here illustrated for PLS-based models. Nevertheless, other modelling methodologies (e.g., SVR and ANN) can also be enhanced by optimizing the interval's resolution.

As for model selection, the SS-DAC framework is proposed to systematically assess and compare the relative performance of prediction models. To achieve this, SS-DAC uses a series of randomizations of the training and test datasets to ensure the models are at their optimal performance and to avoid biased comparisons of their accuracy. Afterwards, the results of SS-DAC are summarized through a scoring system that easily identifies the models with consistently better performance (i.e., models with statistically lower prediction MSEs). SS-DAC also highlights trends on the impact of different modelling methodologies and preprocessing techniques.

For the current case study, SS-DAC showed that the best model was achieved by combining MR-iPLS with mean centering. A loss in performance was observed with other preprocessing techniques, especially with the use of the standard PLS and FiPLS modelling methodologies.

It is noted that the obtained results are case dependent. Therefore, for each new application, SS-DAC should be used to determine the best combination of modelling methodology and preprocessing technique.

REFERENCES

Aguado-Sarrió, E., Prats-Montalbán, J.M., Sanz-Requena, R., Garcia-Martí, G., Martí-Bonmatí, L., and Ferrer, A. (2017). Biomarker comparison and selection for prostate cancer detection in Dynamic Contrast Enhanced-Magnetic Resonance Imaging (DCE-MRI). *Chemometrics and Intelligent Laboratory Systems,* 165, 38-45.

Balabin, R.M., and Smirnov, S.V. (2011). Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Analytica Chimica Acta,* 692 (1), 63-72.

Barnes, R.J., Dhanoa, M.S., and Lister, S.J. (1989). Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy,* 43 (5), 772-777.

Galdón-Navarro, B., Prats-Montalbán, J.M., Cubero, S., Blasco, J., and Ferrer, A. (2018). Comparison of latent variable-based and artificial intelligence methods for impurity detection in PET recycling from NIR hyperspectral images. *Journal of Chemometrics,* 32 (1), e2980.

Geladi, P., and Kowalski, B.R. (1986). Partial Least-Squares Regression: a Tutorial. *Analytica Chimica Acta,* 185, 1-17.

Geladi, P., MacDougall, D., and Martens, H. (1985). Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Applied Spectroscopy,* 39 (3), 491-500.

Kalivas, J.H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems,* 37 (2), 255-259.

Leardi, R., and Nørgaard, L. (2004). Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics,* 18 (11), 486-497.

Naes, T., Isaksson, T., Fearn, T., and Davies, T. (2002). *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester (UK).

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Analytica Chimica Acta,* 1026, 8-36.

Rato, T.J., and Reis, M.S. (2019a). Multiresolution interval partial least squares: A framework for waveband selection and resolution optimization. *Chemometrics and Intelligent Laboratory Systems,* 186, 41-54.

Rato, T.J., and Reis, M.S. (2019b). SS-DAC: A systematic framework for selecting the best modeling approach and pre-processing for spectroscopic data. *Computers & Chemical Engineering,* 128, 437-449.

Rendall, R., Pereira, A.C., and Reis, M.S. (2017). Advanced predictive methods for wine age prediction: Part I – A comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods. *Talanta,* 171, 341-350.

Rinnan, Å., Berg, F.v.d., and Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TRAC Trends in Analytical Chemistry,* 28 (10), 1201-1222.

Soares, F., and Anzanello, M.J. (2018). Support vector regression coupled with wavelength selection as a robust analytical method. *Chemometrics and Intelligent Laboratory Systems,* 172, 167-173.

Wang, L.-L., Lin, Y.-W., Wang, X.-F., Xiao, N., Xu, Y.-D., Li, H.-D., and Xu, Q.-S. (2018). A selective review and comparison for interval variable selection in spectroscopic modeling. *Chemometrics and Intelligent Laboratory Systems,* 172, 229-240.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-Regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58, 109-130.

Xiaobo, Z., Jiewen, Z., and Yanxiao, L. (2007). Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models. *Vibrational Spectroscopy,* 44 (2), 220-227.

Xu, S., Lu, B., Baldea, M., Edgar, T.F., and Nixon, M. (2018). An improved variable selection method for support vector regression in NIR spectral modeling. *Journal of Process Control,* 67, 83-93.

Yun, Y.-H., Li, H.-D., Deng, B.-C., and Cao, D.-S. (2019). An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TRAC Trends in Analytical Chemistry,* 113, 102-115.