Flavio Manenti, Gintaras V. Reklaitis (Eds.), Proceedings of the 34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering (ESCAPE34/PSE24), June 2-6, 2024, Florence, Italy

Predicting the coagulation potential of waste lubricant oil (WLO) using multiblock machine learning of NIR and MIR spectroscopy

Rúben Gariso,[a] Tiago J.Rato,[a] Margarida J. Quina,[a] , Licínio Ferreira,[a] Marco S. Reis,[a,*]

[a]University of Coimbra, CIEPQPF, Department of Chemical Engineering, Rua Sílvio Lima, Pólo II – Pinhal de Marrocos, 3030-790 Coimbra, Portugal
**e-mail: marco@eq.uc.pt*

Abstract

Waste lubricant oils (WLO) pose sustainability challenges, necessitating efficient and reliable methods for their treatment. Regeneration is the preferable approach, but WLOs can coagulate in the equipment, causing plant shutdowns for cleaning and maintenance. To avoid this situation, an alkaline treatment is currently used to assess the WLO coagulation potential prior to regeneration. However, this procedure is time-consuming and subjective, as it involves a visual assessment by the analyst. To overcome these limitations, alternative methods that minimize subjectivity and reduce analysis time are needed. To this end, a rapid and reliable method for predicting the coagulation potential of WLOs through multiblock machine learning analysis of near-infrared (NIR) and mid-infrared (MIR) spectroscopy data is introduced in this article. The classification models employ a combination of Partial Least Squares for Discriminant Analysis (PLS-DA) and the Bayesian linear classifier. The models' performance was optimized via extensive search using the AutoML framework called SS-DAC. More specifically, 1755 combinations of preprocessing, block scaling, and modeling methodologies were tested. By automating this process, a comprehensive and accurate prediction of WLO coagulation potential was achieved. The best NIR single-block model presented a classification accuracy of 0.53, while the best MIR single-block model had an accuracy of 0.88. In turn, the best multiblock model combining the NIR and MIR spectra had an accuracy of 0.94. This improvement is linked to an increase in the correct classification of WLOs that do not coagulate, whose miss-classification is the most critical. Thus, our findings reveal that the combined use of NIR and MIR spectra significantly improves the prediction of the coagulation potential of WLOs compared to the use of NIR or MIR alone, namely for the cases where the misclassification is more detrimental.

Keywords: Waste lubricating oil, Multiblock analysis, PLS, Classification.

1. Introduction

Waste lubricant oil (WLO) is a hazardous waste that can cause significant environmental damage if improperly managed. Regeneration is the priority process regarding the waste management hierarchy in the EU. In Portugal, the entire supply chain for the regeneration of WLO is carried out by Sogilub. The goal of regeneration is to obtain base oil again (the main component of virgin lubricant oil) from WLO. However, regeneration is only feasible if the WLO does not coagulate during processing. Otherwise, the process must be stopped for cleaning and subsequent disposal of the entire production. To minimize this risk, the coagulation potential of WLOs is currently assessed through a laboratory analysis using an alkaline treatment. However, this laboratory test is time-consuming, has some safety risks, and is subjective, depending on the analyst's visual interpretation of the results. As an alternative, a process analytical technology (PAT)-based classifier opens the potential to significantly expedite sample processing and enhance the safety and testing capacity of laboratories.

The present work aims to develop a classifier to predict the coagulation potential using a combination of mid-infrared (MIR) and near-infrared (NIR) spectroscopy. The two information sources were combined using multiblock analysis. A variety of PLS-based approaches were developed to handle multiblock data, such as Concatenated PLS, Hierarchical PLS (Wold et al., 1987), multiblock PLS (Wangen and Kowalski, 1989) and Sequential Orthogonalised PLS (Næs et al., 2020). A review of these methodologies can be found in (Campos et al., 2017). In this work, we employ the Concatenated PLS approach, which consists of concatenating all blocks into a single augmented matrix. Afterwards, the standard PLS is applied.

The rest of this article is organized as follows. In Section 2, the methodology is described, including the pre-processing techniques, block scaling, and modeling methodologies. Afterward, in Section 3, the results are presented. Finally, a summary of the conclusions is provided in Section 4.

2. Methodology

The proposed multiblock methodology is composed of three levels to be optimized: (i) spectra preprocessing; (ii) block scaling (iii); and modeling methodology. A simplified diagram of the workflow of the methodology is presented in Figure 1. To find the best combination of levels we resorted to the SS-DAC framework (Rato and Reis, 2019).

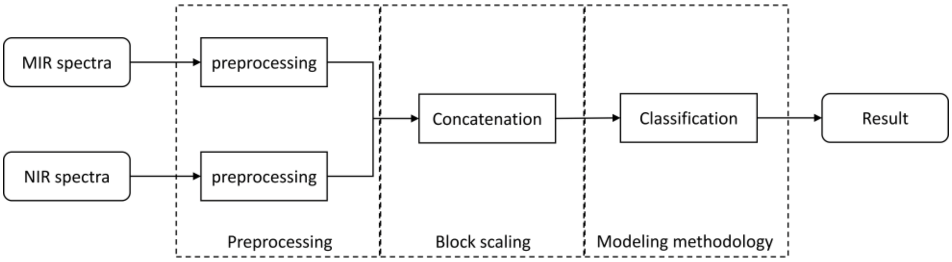

Figure 1 Diagram of the proposed methodology and its three levels to be optimized (spectra preprocessing, block scaling, and modeling methodology).

2.1. Level 1: Spectral preprocessing

The first optimization level of the proposed methodology regards the preprocessing of the spectral data to mitigate potential artifacts that may arise due to unintended interactions between light and the sample under examination. In this work, the focus is directed toward three prominent categories of preprocessing techniques:

- Standard Normal Variate (SNV) (Geladi et al., 1985);
- Multiplicative Scatter Correction (MSC) (Barker and Rayens, 2003);
- Savitzky-Golay differentiation (SGD) (Savitzky and Golay, 1964).

Different combinations of these preprocessing techniques, as well as different parameterizations of SGD (identified as SGD-{derivative order}-{window size}-{polynomial order}) were examined (see Table 1), leading to nine distinct pre-processing variations. Furthermore, it is considered that each data block may have a different preprocessing.

2.2. Level 2: Block scaling

After preprocessing each data block, in the second optimization level, each data block is scaled to ensure that its variability is properly represented within the model.

Block scaling approaches can be generalized by the following equation:

$$\mathbf{X}_{b\ scaled} = (1/K_b) \cdot \mathbf{X}_b \qquad (1)$$

where $\mathbf{X}_b$ represents the original data block, indexed by b (for $b = 1, \ldots, B$), K_b is a block-scaling factor, and $\mathbf{X}_{b\ scaled}$ is the scaled data block.
The block-scaling factors, K_b, are determined using the techniques outlined in Table 1. These techniques can be grouped into two categories:

- Block Scaling (BS) methods, which only consider the number of variables within the block;
- Block Variance Scaling (BVS) methods, which consider the standard deviation of each block.

For detailed formulas and further information, please refer to the work of (Campos and Reis, 2020). After block-scaling, the blocks are concatenated and fed to a modeling methodology.

2.3. Level 3: Modeling methodology

The third and final optimization level of the proposed methodology concerns the fitting of a classification model using the concatenated data blocks. The primary technique employed to fit the models was partial least squares (PLS) for discriminant analysis (DA) (Barker and Rayens, 2003). In the current implementation of PLS-DA, the response variable is an indicator variable that codifies the WLO class (*coagulate* or *does-not-coagulate*) and a PLS-based method is used to extract the latent variables with higher discriminative power. Afterward, the extracted latent variables are fed to the Bayesian Linear Classifier (Hastie et al., 2009) to obtain the final model.
In this study, the full spectra PLS as well as two interval-based extensions of PLS (Nørgaard et al., 2000) were considered to extract the most relevant features:

- PLS (Wold et al., 2001);
- Forward interval PLS (FiPLS) (Xiaobo et al., 2007);
- Backward interval PLS (BiPLS) (Xiaobo et al., 2007).

3. Results

In this study, a total of 107 WLO samples were collected. The WLO coagulation potential was determined through a coagulation test using an alkaline treatment with KOH. Based on the result of this analysis, the samples are classified into *coagulate* or *does-not-coagulate* by the laboratory analyst. Furthermore, the NIR (2530 wavenumbers in the range of 7000 to 3950 cm^{-1}) and MIR (1814 wavenumbers in the range of 4000 to 500 cm^{-1}) spectra of each sample were also collected in triplicate. For the BiPLS and FiPLS-based models, each block was divided into 15 equal intervals.

To evaluate the impact of combining two blocks of information, the SS-DAC framework was used over multiple combinations of preprocessing techniques, block scaling, and modeling methodologies, leading to 1701 multiblock model combinations. Furthermore, the single-block model scenarios were also considered as benchmark, leading to an additional 54 single-block models. Overall, a grand total of 1755 models were tested. These models were labeled using the following nomenclature (see Table 1): {modeling methodology}-{data block}-{MIR preprocessing technique}-{NIR preprocessing technique}-{block scaling technique}.

In the first stage of SS-DAC the raw dataset was randomly split into a training dataset with 80 % of the samples and a test dataset with the remaining 20 % of the samples, maintaining balanced datasets. Furthermore, replicates of the same sample were attributed to the same dataset. The models were then trained on the training dataset using Monte Carlo Cross-Validation (MCCV) for tuning the model's hyperparameters (*i.e.*, the number of retained latent variables, and intervals in the models). Afterward, in the second stage of SS-DAC, the models' performance was assessed on the test dataset using the accuracy (H) as the key performance indicator (KPI):

$$H = \frac{TP + TN}{n} \tag{2}$$

where TP is the number of true positives, TN is the number of true negatives and n is the number of samples in the test dataset. The accuracy varies from 0 to 1, where higher values relate to better classification capabilities.

In the third stage of SS-DAC the models were compared against each other using the Wilcoxon signed-rank test (Wilcoxon, 1945). For each comparison, if a model has a statistically significant higher accuracy against another model it receives a *victory* and if there is no statistically significant difference in accuracy it receives a *tie*. A score is then computed by summing the number of victoried and ties of each model. Models with higher scores (*i.e.*, high count of victories and ties) are deemed to have consistently higher accuracy. The scores for the top 50 models are presented in **Erro! A origem da referência não foi encontrada.**, ranked from highest to lowest performance. For this case, the maximum score a model can achieve is 1754, representing a victory against all other models. For reference, the models' accuracy on the test dataset was also computed.

Table 1 Summary of the data blocks and optimization levels considered in this study. The model's nomenclature is as follows: {modeling methodology}-{data block}-{MIR preprocessing technique}-{NIR preprocessing technique}-{block scaling technique}.

Data block	Optimization levels		
	Modeling methodology	**Spectral preprocessing**	**Block scaling**
Only MIR [A]	PLS [P]	Not used [0]	Not used [0]
Only NIR [B]	FiPLS [F]	Mean centering [1]	No Scaling [I]
MIR&NIR [C] *	BiPLS [B]	SNV [2]	Soft BS [II]
		MSC [3]	Hard BS [III]
		SGD-1-7-2 [4]	Super Hard BS [IV]
		SGD-1-15-2 [5]	Soft BVS [V]
		SGD-2-9-2 [6]	Hard BVS [VI]
		SNV-SGD-1-7-2 [7]	Super Hard BVS [VII]
		SNV-SGD 1-15-2 [8]	
		SNV-SGD 2-9-2 [9]	

* The model is free to select between both data blocks. Depending on the intervals selected by the model the data block is subclassified into: C11 if both blocks are selected; C10 if only the MIR block is selected; C01 if only the NIR block is selected.

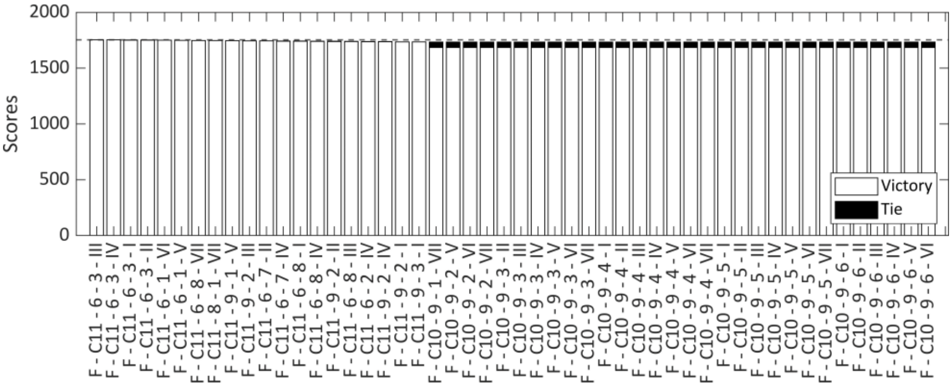

Figure 2 Score of the top 50 models using the SS-DAC framework. The model's nomenclature is as follows (see Table 1): {modeling methodology}-{data block}-{MIR preprocessing technique}-{NIR preprocessing technique}-{block scaling technique}.The dashed horizontal line represents the maximum score value (1754).

Regarding the single-block models, the best model using only the MIR spectra (F-C10-9-1-VII) was ranked in position #21, with an accuracy of 0.88 (with a partial accuracy of 0.78 for the WLO samples that coagulate and 1.00 for the WLO samples that do-not-coagulate), while the best model using only the NIR spectra (B-B-0-4-0) was ranked in position #1708 with an accuracy of 0.56 (with a partial accuracy of 0.44 for the WLO samples that coagulate and 0.67 for the WLO samples that do not coagulate).

As for the multiblock models, it was verified that, due to the interval selection of FiPLS, 478 out of 567 FiPLS-based multiblock models only used the MIR block. Thus, they are, in effect, equivalent to their single-block counterparts, and their performance is independent of the scaling and preprocessing of the NIR block. This also implies a performance tie when compared against each other, which is visible in **Erro! A origem da referência não foi encontrada.** for models in positions #21 (F-C10-9-1-VII) to #50 (F-C10-9-6-VI). The same situation also happens for models with smaller scores (not shown). Nevertheless, the NIR block is still informative when combined with the MIR block as shown on the top four models. These models share a similar structure, being based on FiPLS and SGD for the MIR block and MSC for the NIR block, with block scaling having a small impact on performance. The best model (F-C11-6-3-III) achieved a global accuracy of 0.94 (and a partial accuracy of 0.89 for the WLO samples that *coagulate* and 1.00 for the WLOs that *do-not-coagulate*), which represents an improvement of 6.82 % against the best MIR single-block model.

The top models frequently selected the [734.4 - 966], [1668 - 1900] cm^{-1} intervals from the MIR spectra, and the [4951 - 5150] cm^{-1} interval from the NIR spectra. The interval that appears more predominantly ([1668 - 1900] cm^{-1}) is thought to be related to the presence of esters (Weyer, 2012) in the WLO, a result that was also obtained in (Pinheiro et al., 2017).

4. Conclusions

Among the single-block models, those using the NIR spectra presented the worst performances as even the best NIR single-block model only had an accuracy of 0.56. In turn, the MIR single-block models proved to be more informative, achieving an accuracy of 0.88. The best performance was attained by multiblock models combining FiPLS with variations of SGD. The best multiblock model had an accuracy of 0.94, which represents an improvement of 6.82 % against the best single-block model. For this case, block scaling had a lesser impact on performance since FiPLS tended to select intervals solely from the MIR spectra. Nevertheless, a few intervals from the NIR spectra were also selected by the multiblock models. Thus, it is concluded that incorporating MIR and NIR spectral information significantly enhances the predictive capability of the models. The top models also point to the presence of esters as the most critical factor for WLO coagulation, thus providing crucial insights into the coagulation phenomenon.

Acknowledgments

The authors gratefully acknowledge the financial support of SOGILUB – Sociedade de Gestão Integrada de Óleos Lubrificantes Usados, Lda. The authors also acknowledge the support from the Chemical Process Engineering and Forest Products Research Centre (CIEPQPF), which is financed by national funds from FCT/MCTES (reference UIDB/EQU/00102/2020).

References

M. Barker, W. Rayens, 2003, Partial Least Squares for Discrimination, Journal of Chemometrics, 17, 166–173.

M. Campos, M. Reis, 2020, Data Preprocessing for Multiblock Modelling – A Systematization with New Methods, Chemometrics and Intelligent Laboratory Systems, 199, 103959.

M.P. Campos, R. Sousa, A.C. Pereira, M.S. Reis, 2017, Advanced Predictive Methods for Wine Age Prediction: Part II – A Comparison Study of Multiblock Regression Approaches, Talanta, 171, 132–142.

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

T. Hastie, R. Tibshirani, J. Friedman, 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, Springer, New York, NY.

T. Næs, R. Romano, O. Tomic, I. Måge, A. Smilde, K.H. Liland, 2020, Sequential and Orthogonalized PLS (SO-PLS) Regression for Path Analysis: Order of Blocks and Relations between Effects, Journal of Chemometrics, e3243.

L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, 2000, Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy.

C. Pinheiro, V. Ascensão, M. Reis, M. Quina, L. Gando-Ferreira, 2017, A Data-Driven Approach for the Study of Coagulation Phenomena in Waste Lubricant Oils and Its Relevance in Alkaline Regeneration Treatments, Science of The Total Environment, 599–600, 2054–2064.

T.J. Rato, M.S. Reis, 2019, SS-DAC: A Systematic Framework for Selecting the Best Modeling Approach and Pre-Processing for Spectroscopic Data, Computers & Chemical Engineering, 128, 437–449.

A. Savitzky, M.J.E. Golay, 1964, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Analytical Chemistry, 36, 1627–1639.

L.E. Wangen, B.R. Kowalski, 1989, A multiblock partial least squares algorithm for investigating complex chemical systems, Journal of Chemometrics, 3, 3–20.

J.W.J. Weyer Lois, 2012, Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy, 2nd ed, CRC Press, Boca Raton.

F. Wilcoxon, 1945, Individual Comparisons by Ranking Methods, Biometrics Bulletin, 1, 80–83.

S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, H. Wold, 1987, PLS Modeling with Latent Variables in Two or More Dimensions, Frankfurt am Main.

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

Z. Xiaobo, Z. Jiewen, H. Xingyi, L. Yanxiao, 2007, Use of FT-NIR Spectrometry in Non-Invasive Measurements of Soluble Solid Contents (SSC) of “Fuji” Apple Based on Different

Predicting the coagulation potential of waste lubricant oils (WLO) using multiblock machine learning of NIR and MIR spectroscopy

PLS Models, Chemometrics and Intelligent Laboratory Systems, 87, 43–51.