Microchemical Journal 207 (2024) 112144

Contents lists available at ScienceDirect

Microchemical Journal

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

Another pipeline in local Partial Least Squares Regression (LPLS) methods: Assessing the impact of wavelet transform integration

Antoine Deryck [a,*], Andreas Niemöller [b], Vincent Baeten [a], Juan Antonio Fernández Pierna [a]

[a] Quality and Authentication of Agricultural Products Unit, Knowledge and Valorization of Agricultural Products Department, Walloon Agricultural Research Center, Chaussée de Namur 24, 5030 Gembloux, Belgium
[b] Bruker Optics GmbH & Co. KG., Rudolf-Plank-Straße 27, 76275 Ettlingen, Germany

ARTICLE INFO

Keywords:
Local Partial Least Squares regression
Wavelet transform
NIR spectroscopy
Feed

ABSTRACT

This study evaluates the performance of four Partial Least Squares Regression (PLS) methods, focusing on a new Local Partial Least Squares Regression (LPLS) variant integrating wavelet transformation, named WLPLS, for analyzing feed using Near-Infrared (NIR) spectroscopy. While traditional PLS methods are effective for many spectroscopic applications, their global modeling approach often reduces predictive accuracy in large, heterogeneous datasets. In contrast, LPLS adapts models to local data characteristics, which can enhance prediction but also increase computational demands (power and time). WLPLS seeks to mitigate these demands by incorporating wavelet transformation to reduce data dimensionality while effectively managing spectral variances. This research conducts a comparative analysis of WLPLS against traditional PLS, LPLS, and another LPLS pipeline reducing data dimensionality: the LPLS on global PLS scores (LPLS-S). The performance of these methods were evaluated using a large feed database containing 24,644 samples, analyzing five key constituents: ash, crude fibers, fat, moisture, and proteins. The results demonstrate that local approaches outperformed the global PLS method for this dataset and that the performance of the local methods were relatively similar to each other. The selection of the optimal method therefore depends on the specific requirements of the application, such as dataset characteristics and the required prediction speed. Future studies should broaden this comparative framework to additional datasets and contexts to ensure they are adapted for diverse applications.

1. Introduction

The development of Local Regression (LR) models has been a significant advance in the field of chemometrics, especially for the analysis of spectroscopic data. The expansion of LR stems from its ability and efficiency to handle large and complex datasets, whereas most of the currently used near-infrared (NIR) calibration algorithms fail to do so. Although these methods, such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLS) have proven their efficiency in the fields of agronomy, feed, and food, contributing to their worldwide application, their linear approach often leads to a decrease in their performance when dealing with large datasets, where an increase in overall spectral variance is observed.

This is because these methods rely on a single global model fitted to an entire database to analyze any sample, which limits their ability to account for local variability within the data [1–3]. LR compensates by creating a unique model for each new sample to be predicted. This process starts by identifying a set of k-nearest neighbors (kNN) within a spectral library, pinpointing precisely where the new spectrum fits within the existing data. This step ensures that the selected spectra are similar to the spectrum to be analyzed, and also results in a significantly better-to-calibrate, reduced variance in these spectra compared to what would be observed in a global model approach. This simplifies the calibration step, which is subsequently performed on this specific subset of samples and involves decomposing the variance into latent components that are specifically correlated with the component to be predicted. Consequently, the local model tends to be more robust and accurate, especially in the case of heterogeneous libraries composed of different products or recipes [1]. Numerous LR methods exist, and new ones continue to emerge [4–6], but they all share a common foundation.

The first algorithm of this kind is the Locally Weighted Regression (LWR) proposed by Naes [7–9]. In this case, nearest neighbors are searched for using Euclidean distance and the influence of the neighbors on the regression model (in this case PCR) is then weighted according to

* Corresponding author.
E-mail address: a.deryck@cra.wallonie.be (A. Deryck).

https://doi.org/10.1016/j.microc.2024.112144
Received 15 October 2024; Received in revised form 5 November 2024; Accepted 6 November 2024
Available online 9 November 2024

the distance to the spectrum to be analyzed.

Variations of the LWR quickly emerged and mostly concerned the regression method (PCR [7], PLS [5,10], …), the metrics used to select the kNN (Mahalanobis distance [7], Euclidian distance [5,10], correlation [11], …), and the weight function applied to the calibration objects (uniform [10], cubic [7], …). However, these pipelines have a common drawback: they all require significant computational power, meaning that generating predictions takes considerable time. Other algorithms were developed to solve this issue.

With the LOCAL method of Shenk, Westerhaus, and Berzaghi [1,12], the nearest neighbors are found by correlation calculations of the spectrum to be analyzed with the spectra of the spectral library. In addition, and due to the computing power of the PCs available in 1997, data points were selected at specific intervals of the spectrum (and previously all library spectra) to reduce the number of spectral variables and speed up the calculations. However, this step also reduced the information content of the data set. The rank selection for the local PLS model is either fixed or the result is automatically calculated by a specific averaging of the results of different ranks.

Data compression in the Local Calibration by Percentile Selection (LCPS) [13], in the Local Calibration by Customized Radii Selection (LCCRS) [13], and in the Local Partial Least Squares based on global PLS scores (LPLS-S) [5] approaches takes place via PLS, i.e. PLS modeling on all spectra of a library. The aim is not to perform direct analysis, but to replace high-dimensional spectral data with PLS scores, reducing the size of the spectral library. The kNN of a new sample to be analyzed are therefore found in the new PLS score space (with a unique distance metric for each method) and the PLS model is then performed locally based on these global scores. The disadvantages are that such a method based on global PLS models is specific to a component or property and must always be completely recalculated when the library is expanded. In addition, these global scores are no longer very specific to the local situation of a spectrum to be analyzed. Also, recently, Lesnoff *et al.* have presented a list of different averaging methods that can be easily embedded in pipelines of local PLSR, with the objective of automatized predictions, and thus represent fast and safe alternatives to methods requiring time-consuming calibrations [14].

In this context, a new method combining wavelet transform and LPLS has been proposed: The Wavelet Local Partial Least Squares regression (WLPLS). The goal of this paper is to assess this new WLPLS method in a large and heterogeneous library and to compare its performance with the classical approaches of Partial Least Squares Regression (PLS), Local Partial Least Squares Regression (LPLS), and Local Partial Least Squares Regression based on PLS scores (LPLS-S).

2. Materials and methods

2.1. Introduction to wavelet transform

Wavelet transformation [15–17] is considered a significant breakthrough in mathematical analysis to process, extract, compress, and represent signals and data. In chemistry, it is applied in various fields [18,19] like flow injection analysis, chromatography, spectroscopy (UV, VIS, NIR, IR, NMR, XRF) [20–23], mass spectrometry, and image processing [24]. Application categories include data compression, parsimonious/sparse data representation [16,21], denoising and smoothing [15,16,21], baseline/background removal and correction [15], regression and calibration [21,25,26], and classification and pattern recognition [27–29].

Wavelets are mathematical functions that transform the original signal or data into different frequency components and provide the corresponding location information. This has advantages over the traditional Fourier transform (FT) method, where only the presence and intensity of different frequencies are analyzed, without the information on where these frequencies occur.

There are different types of wavelet transformations: continuous wavelet transformation (CWT), wavelet package decomposition (WPD), and discrete wavelet transformation (DWT). The DWT is the simplest and decomposes the signal by applying a high-pass filter, which is fast and is detailed here:

Unlike the periodic basis functions in the FT, the wavelet basis function has compact support, which enables localization. There are several families of wavelet functions with many possible types and shapes (Fig. 1). For the frequency analysis, the wavelet transform is performed by applying a series of expanding (or shrinking) scales of the wavelet function, which form orthogonal bases. Localization is performed by translating the scaled wavelet functions over the signal. The signal is folded with the series of scaled and translated wavelet functions, resulting in the wavelet coefficients.

In the context of wavelet transformations applied to spectroscopic data, the Haar-Wavelet, or other simple and common types are sufficient because the spectra are decomposed and the coefficients are directly used for local regression. For other tasks where reconstruction of the original signal is required, the wavelet type and its properties (like smoothness) are much more relevant. The transform and the coefficients are orthogonal, which is an excellent property that allows the coefficients to be selected or combined as required.

The decomposition of the spectral data can be better understood by looking at a simulated spectrum containing two absorbance bands, a baseline, and noise (Fig. 2). The coefficients are classified into wavelet bands (j) representing specific frequency ranges that correspond to the expanding scales used in the wavelet transformation (Fig. 3). The band $j = 0$ represents the low-frequency residual part of the decomposition ("approximation coefficients") and is of no use in most cases. The "detailed coefficients" start with the low-frequency band ($j = 1$), where only two coefficients cover the whole spectral range (signal). In the next band, the number of coefficients is doubled, and the analyzed frequency is increased respectively. The highest band(s) show mainly noise unless the signal is smooth.

In the transformed simulated signal, it is visible which coefficients are important or valuable for the two absorbance bands − the relevant information of the spectroscopic application. The valuable coefficients are those in the mid-frequency bands with values above the noise level. The wavelet bands below band 3 represent mainly baseline or unimportant low-frequency information, while the upper bands 7 and 8 represent noise. It is important to note that the coefficients increase with increasing absorbance, which means that Beer's law remains valid.

After the wavelet transformation, the appropriate coefficients are selected manually, either individually or by frequency band, depending on the application. As an alternative to normalizing the spectra, the selected coefficients can be normalized, which is beneficial because the higher frequency bands are normally truncated and noise is removed.

2.2. Local Partial Least Squares Regression (LPLS)

Local Partial Least Squares Regression (LPLS) is a specific instance of Local Weighted Regression (LWR) where all selected samples (nearest neighbors) have equal weights in the regression process. This technique is commonly utilized in the chemometric field [5,30]. Typically, the Mahalanobis or Euclidian distances are used as the metric distance, PLS is the regression method employed, and the weight function is a vector of values $1/k$, where k is the number of nearest neighbors.

2.3. Combining wavelet transform with LPLS

A new wavelet-based Local Regression method, the Wavelet Local Partial Least Squares regression (WLPLS) is proposed. The idea of combining wavelet theory with the LPLS method comes from the possibility of using wavelet coefficients out of selected wavelet bands to span a data space defining the spectral library for the kNN search step. The coefficients are also used directly for the LPLS modeling step for a prediction of an analyzed spectrum.

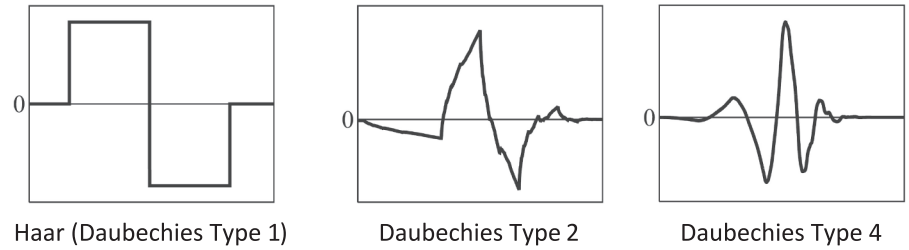


Fig. 1. Examples of wavelet basis functions of the common Daubechies wavelet family including the simple Haar Wavelet (Daubechies Type 1).

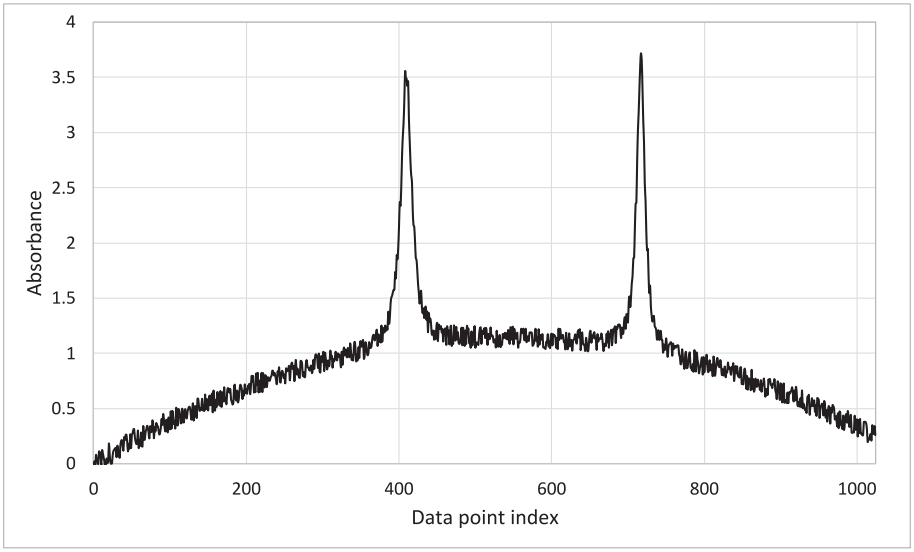


Fig. 2. Simulated spectrum containing two absorbance bands, baseline, and noise added.

2.4. Dataset

For this study, a database of 24,644 samples of feed products (formulations) built by the Walloon Agricultural Research Centre (CRA-W) was used. This database shows high variability as described by Fernández Pierna *et al.* [31,32]. Spectral data for the samples were obtained using a benchtop XDS spectrometer (FOSS Analytics), covering a range from 1100 to 2498 nm with a digital resolution of 2 nm. Reference analytical methods were used to evaluate the content of five constituents within the samples: Ash (ASH), crude fibers (CF), fat (FAT), moisture (MOIST), and protein (PROT). The values of ASH, CF, FAT, MOIST, and PROT are expressed in percentages of the total weight of the sample.

It is important to note that there is variability between the number of samples available for each of these constituents. For instance, the number of samples with associated MOIST content was much higher than the number of samples with associated CF values. This is because the database is composed of several products which underwent various sets of analyses. The related information and statistics of the constituents' values are summarized in Table 1.

2.5. Comparative Methodology: PLS vs LPLS vs LPLS-S vs WLPLS

This section describes how the four techniques (PLS, LPLS, LPLS-S, and WLPLS) were applied to this database. A quick summary of the procedures followed for the local regressions can be visualized in Fig. 4.

Before the models' construction, the database was randomly divided into three datasets: 60 % of the samples constituted the calibration set (14,786 samples), 15 % the optimization set (3,696 samples), and the remaining 25 % formed the validation set (6,162 samples).

To ensure a fair comparison of the models' performance, the same datasets were used across the different methods. However, while the optimization and the validation sets served the same purpose for both global and local techniques, the calibration set had to be used differently. With the global PLS, this set was used specifically to train the PLS models. For each of the LPLS, LPLS-S, and WLPLS, this calibration set was used as a repository in which to find a subset of samples (nearest neighbors), which was then used to build the PLS models. Therefore, the calibration set will be referred to as the spectral library for the local methods.

Two different metrics were used to evaluate the performance of the models. Those metrics are the Root Mean Square Error of Prediction (RMSEP) and the Residual Prediction Deviation (RPD). The RMSEP (Equation (1)) provides a measure of the average magnitude of the errors between predicted and observed values. A lower RMSEP value indicates better predictive accuracy.

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

where n is the number of observations, y_i is the observed value, and $\hat{y}_i$ is the predicted value.

The RPD (Equation (2)) is a dimensionless ratio that compares the standard deviation of the observed values to the RMSEP. Higher RPD values indicate better model performance. According to Chang et al. [33], the interpretation of RPD values is as follows: RPD below 1.4 indicates poor model performance, RPD values between 1.4 and 2 indicate moderate performance, and RPD values greater than 2 have good predictive ability.

$$RPD = \frac{SD}{RMSEP} \quad (2)$$

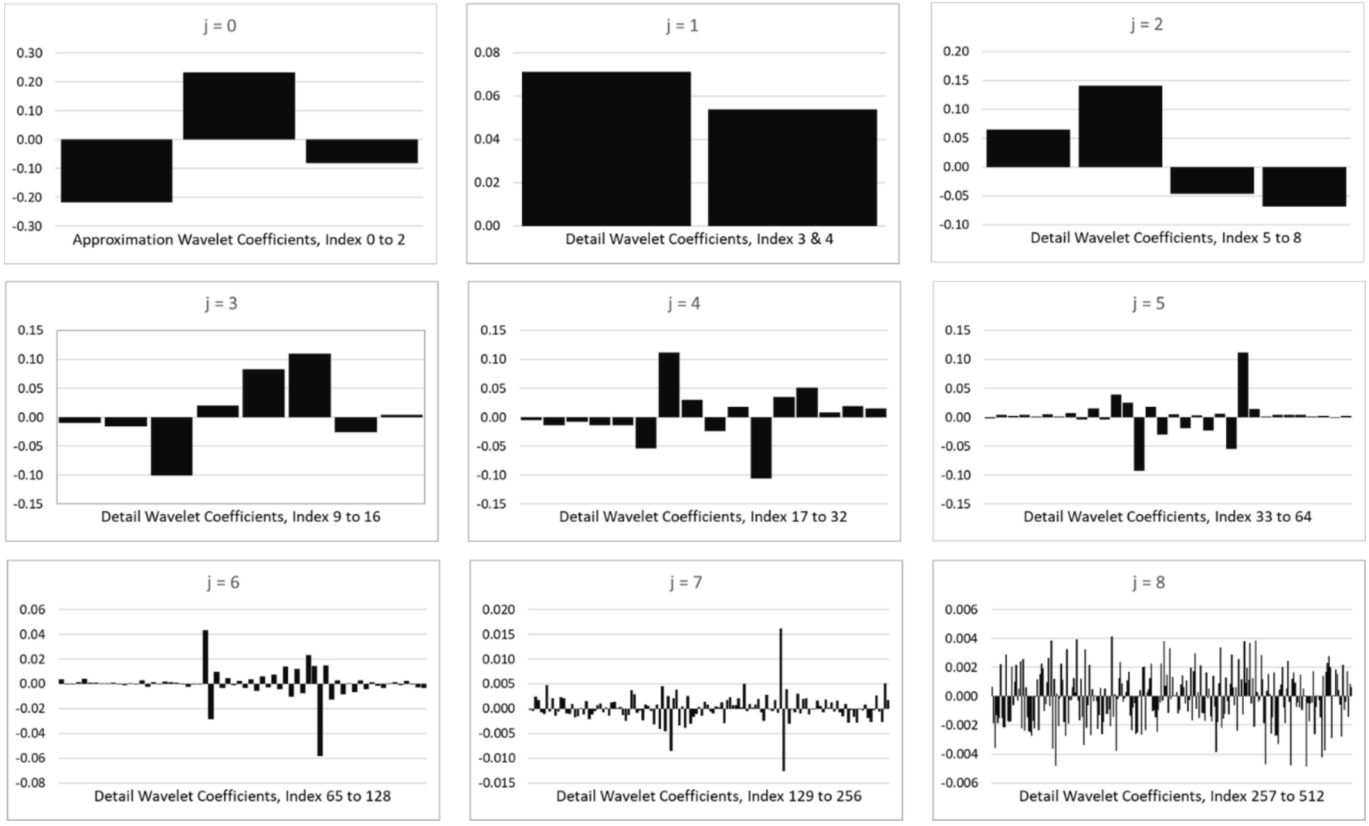


Fig. 3. Wavelet transformation of the simulated spectrum to several frequency bands with wavelet coefficients. The indices represent the ID of the wavelet coefficients.

Table 1
Descriptive statistics of the constituent contents in the database: Number of samples (N), minimum (Min), maximum (Max), mean, and standard deviation (Std). All values, except the number of samples, are expressed in percentages of the total weight of the sample (%).

Constituents	N	Min	Max	Mean	Std
ASH	20,065	0.800	37.000	7.320	3.262
CF	5,423	0.100	22.200	5.319	2.776
FAT	7,640	0.600	31.600	5.253	3.793
MOIST	23,392	2.040	16.700	11.356	1.822
PROT	22,371	6.800	62.300	20.473	8.098

where *SD* is the standard deviation of the observed values.

Using both RMSEP and RPD provides a more comprehensive evaluation of the models' predictive performance. The RMSEP offers a direct measure of prediction error in the same units as the response variable, which is intuitive and easy to interpret. However, it does not account for the variability in the dataset. Conversely, the RPD contextualizes the prediction error relative to the variability in the data, offering a dimensionless measure of predictive power. This enables the qualification of model performance regardless of the units or the scales of the data. By considering both metrics, we obtain a balanced view of model performance, understanding both the absolute error (RMSEP) and the error relative to the data variability (RPD).

For the construction of the PLS models, analyses were carried out on each constituent individually (Ash, CF, FAT, MOIST, and PROT) and samples with no associated constituent value were removed from the datasets (calibration, optimization, and validation). This led to important differences in dataset size between the constituents (Table 2).

The calibration set was then used to train the models (on the raw and pre-processed data). 60 models with different parameters (in this case preprocessing and rank of the models) values (Table 3) were constructed and applied to the optimization dataset.

The RMSEP values of those models on the optimization dataset were assessed and the model giving the lowest RMSEP was selected as the final model. The preprocessing (or absence of preprocessing) associated with the best model was retained as well.

The final model and preprocessing (or none) were then applied to the validation dataset and its performance was assessed with the RMSEP and the RPD.

The LPLS models were then built. As for the PLS models, the analyses were carried out on each constituent individually. However, the samples with no associated constituent value were only removed from the optimization and the validation set. The objective behind keeping samples with no associated constituent value in the spectral library is to be able to compare the LPLS with the WLPLS which applies a constraint related to those specific samples: When a sample to be predicted is presented to this LPLS algorithm, it searches for its kNN based on the Mahalanobis distance. It then removes the neighbors for which no constituent value is found. If less than 40 neighbors remain, no prediction is performed, and a "not available" value (NA) is assigned to the sample. Taking this constraint into account, 360 models with different parameters (preprocessing, number of nearest neighbors, and model rank) values (Table 3) were applied to the samples of the optimization set. The nearest neighbors of the samples within the spectral library were identified and whenever a sample had more than 40 nearest neighbors with reference values, a PLS model was built, and a prediction was performed. The RMSEP values of the models were then calculated on the samples with an assigned prediction. The parameters associated with the lowest RMSEP were retained. The process differs from PLS in that a single LPLS model is built for each sample to be predicted. However, parameters were optimized globally and not for each model.

Predictions were performed on the validation samples with the

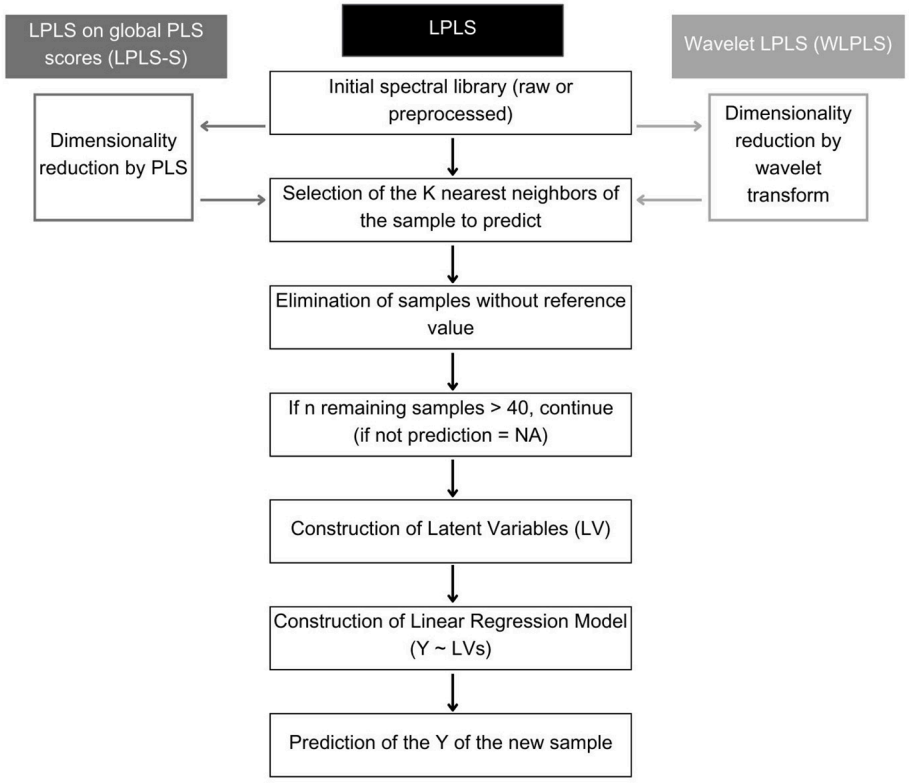


Fig. 4. Visualization of the three local approaches compared in this work.

Table 2
Number of samples with associated constituent values available in each of the datasets.

Datasets	ASH	CF	FAT	MOIST	PROT
Calibration/Spectral Library	12,059	3,327	4,601	14,037	13,394
Optimization	3,003	798	1,138	3,530	3,366
Validation	5,003	1,298	1,901	5,825	5,611

retained pre-processing and parameters. RMSEP and RPD were calculated on the samples with assigned predictions. The number of samples with NA values was also evaluated and considered in the models' comparison.

The LPLS-S models' construction follows the same procedure, except that it starts with a dimensionality reduction of the spectra of the PLS datasets, replacing the spectra with the latent variable scores of the model (Fig. 4). The parameter associated with this additional step (number of latent variables retained after dimensionality reduction) was optimized along with the others (Table 3). A total of 2,970 models with different parameter values were tested.

Finally, the WLPLS construction follows the same procedure as LPLS except that it starts with a discrete wavelet transformation of the spectra of the three datasets (Fig. 4). The parameters associated with this additional step (compression level and preprocessing of the coefficients) were optimized along with the others (Table 3). A total of 2,160 models with different parameter values were tested.

The RMSEP (calculated on the validation sets) of the four algorithms were compared to determine which technique yielded the best performance. Since some models resulted in unpredicted values for a few samples, no statistical tests were carried out for the comparison.

2.6. Software

The PLS, LPLS, and LPLS-S models were built on R version 4.2.2 [34] with the packages rchemo [35] and mdatools [36].

Table 3
Parameters tested for the models' optimization. SNV stands for Standard Normal Variate.

PLS	
Preprocessing	1st derivative – window size of 13/ SNV/None
Number of latent variables (rank)	1 to 20
LPLS	
Preprocessing	1st derivative – window size of 13/ SNV/None
Number of nearest neighbors	100/200/300/400/500/1000
Number of latent variables (rank)	1 to 20
LPLS-S	
Preprocessing	1st derivative – window size of 13/ SNV/None
Number of variables after dimensionality reduction	10 to 20
Number of nearest neighbors	100/200/300/400/500/1000
Number of latent variables (rank)	1 to 20
WLPLS	
Preprocessing	1st derivative – window size of 13/ SNV/None
Number of nearest neighbors	100/200/300/400/500/1000
Compression level (wavelets)	5/6/7
Preprocessing of coefficients (wavelets)	SNV/None
Number of latent variables (rank)	1 to 20

The Wavelet Local Partial Least Squares regression (WLPLS) has been proposed by Bruker Corporation in their module OPUS: QUANT 3 (Bruker Optics GmbH & Co. KG). The WLPLS models were built with this module.

3. Results and discussion

3.1. Performance of the models

Despite the variability of the feed database, the PLS regression allowed high-quality predictions for all the parameters, with RPD higher than 2 (the minimum RPD being 2.73). However, the local methods regressions resulted in even better RPDs (4.37 minimum) (Table 4), and therefore in higher-quality predictions for all constituents. Compared to global PLS and depending on the predicted parameter, LPLS led to a reduction in RMSEP of between 11–39 %, LPLS-S of 2–35 %, and WLPLS 2–39 % (FAT being associated with the lowest reductions and ASH the highest) (Fig. 5).

As expected and already highlighted in several works dealing with large and complex datasets [3,13,37], local regressions enabled a clear improvement in the performance of the prediction of the nutritional values of feed formulations (Fig. 5). The local methods showed smaller differences between them and their performance remain relatively similar, although the performance of the LPLS-S algorithm seem slightly lower and the WLPLS algorithm slightly higher (Fig. 5).

Another observation made across the four evaluated methods concerns their performance when applied to raw versus preprocessed data (Fig. 5). All methods exhibited similar or superior performance with preprocessed data. However, the consistency of the RMSEP differences varied. For the LPLS-S models, performance on preprocessed data was comparable to that of the raw data, with RMSEP variations from 0 to 11 %. Overall, PLS and WLPLS performed slightly better with preprocessed data, with RMSEP variations ranging from 8 % to 33 % for PLS and from 6 % to 20 % for WLPLS. In contrast, LPLS showed considerable improvement with preprocessed data, with RMSEP variations between 41 % and 73 %. This seems to indicate that preprocessing is not as necessary when working with LPLS-S and WLPLS algorithms as with LPLS, however, this could be due to the nature of the feed database.

The wavelet transform in WLPLS likely enhances the model's capability by separating the chemical signal from interfering artifacts, such as baseline shifts, noise, and other distortions common in NIR spectroscopy (e.g., scattering, multiplicative, and additive effects). By breaking down the spectral data into frequency bands, the wavelet transform isolates relevant chemical information, giving less weight to these artifacts in the latent variable construction. This separation reduces the influence of non-chemical artifacts on the final model. In contrast, other local PLS models process all spectral features together, including baseline shifts and noise, although LPLS-S partially reduces these effects in an initial PLS step. While preprocessing usually helps to manage such artifacts, the wavelet transform seems especially effective in minimizing them in this case. This also explains why the preprocessing steps seem to have a lower impact on the WLPLS performance.

3.2. Summarized comparison of the methods

As shown by the results, local methods take into account the local variability of the feed database. This reinforces the assumption that local methods should always be tested when dealing with large and heterogeneous datasets such as those found in spectral studies of feed and forage (samples of different origins, dates, composed of diverse ingredients, …) [5,6]. However, it is essential to keep in mind that each method has its own drawbacks and advantages, and that the choice of algorithm depends on the dataset.

Table 4
Residual Prediction Deviations obtained with the four algorithms applied to five parameters of the feed database. Residual Prediction Deviation values obtained with the best calibration models (built on preprocessed data for all models except for WLPLS on FAT built on raw data) are displayed in this table.

	ASH	CF	FAT	MOIST	PROT
PLS	2.73	4.23	7.67	3.70	9.11
LPLS	4.51	4.75	9.26	5.29	13.88
LPLS-S	4.37	4.41	7.23	4.85	11.44
WLPLS	4.50	4.98	7.78	5.36	13.94

As opposed to the local algorithm, PLS is fast and easy to interpret and implement on instruments for real-time predictions. As only one model is created, global VIP scores can be investigated, which can reinforce the relevance of the model. The number of parameters to optimize remains low.

LPLS is also a relatively straightforward method, but it requires significant computational power and time, making it challenging to implement for real-time analyses. The LPLS-S and WLPLS methods address this limitation by reducing the size of the dataset and speeding up the method without losing relevant information. Dimensionality reduction is especially beneficial for large libraries and method transfer e.g., in a network.

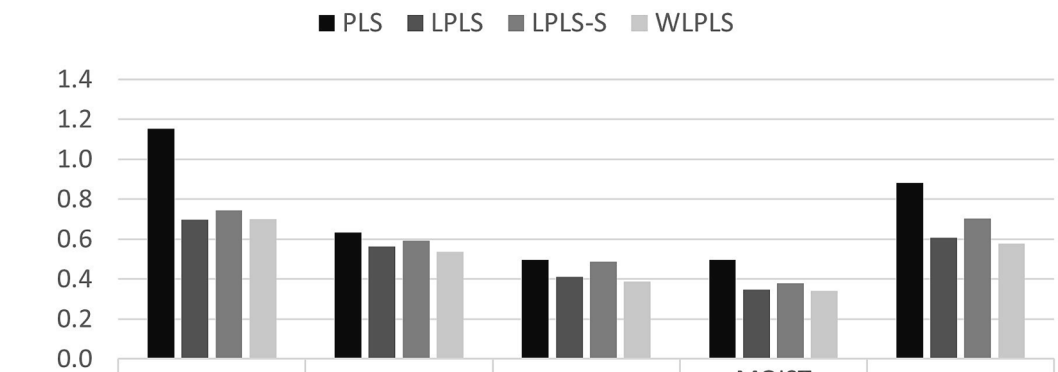
Those methods are, however, more complex and require more extensive optimization. LPLS-S presents the advantages of relying on a simple and widely known technique for its dimensionality reduction step and downsizing the dataset to a few variables e.g. from 10 to 20 latent variables for a spectrum with 700 data points. With the WLPLS approach, as only a few wavelet bands with a low total number of coefficients are used, the spectral data is compressed as well, but the number of transformed variables is larger e.g., from 60 to 100 coefficients for a spectrum with 700 data points. However, as it relies on the wavelet transform technique, WLPLS has some additional distinct features:

- As the spectra are interpolated and transformed one by one, a library can be easily formed and modified without the need for recalculations. This contrasts with approaches like LPLS-S, where a change in the composition of the library requires recalculation of the PLS scores. This can affect the whole data structure of the library and change the performance.
- It is not necessary to select the spectral range or to preprocess the spectral data. The wavelet transform works like a pretreatment in the sense of baseline removal and noise suppression simply by selecting the wavelet bands. Due to the localized character of the wavelet transform, the pretreatment properties are also applied locally. Similarly, in this study, LPLS-S appears to have targeted specific patterns related to the predicted constituent before employing the LPLS algorithm on the transformed data matrix. Transforming the original data matrix into a matrix of latent variable scores using the LPLS algorithm could also minimize unwanted variations and noise.
- Since WLPLS assigns weights to the orthogonal coefficients through the latent variables in the PLS process, there is no need to select coefficients individually within the frequency bands, which makes the method setup straightforward.

4. Conclusion

In the ocean of LPLS pipelines, the integration of wavelet transformations with LPLS is not a silver bullet, but it provides specific advantages. These advantages, which can also be observed with LPLS-S and other methodologies not studied here, suggest a nuanced improvement over traditional methods. However, this paper does not pinpoint the combination of wavelet and LPLS method as the best choice. The optimal methodology depends on the specific dataset to which it is applied and the goal of its users. Therefore, researchers should select the method that best suits their objectives and the unique characteristics of their data.

This study was limited to a single dataset, which restricts the scope of our conclusions. Nonetheless, our analysis has provided a better understanding of the comparative performance of these techniques. The enhanced predictive accuracy and processing efficiency observed in the



	ASH PREPRO	CF PREPRO	FAT PREPRO	MOIST PREPRO	PROT PREPRO
PLS	1.15	0.63	0.50	0.50	0.88
LPLS	0.70	0.56	0.41	0.35	0.61
LPLS-S	0.75	0.59	0.49	0.38	0.70
WLPLS	0.70	0.54	0.39	0.34	0.58

Fig. 5. Root Mean Square Error of Prediction (expressed in percentages of the total weight of the sample) obtained with the four algorithms applied to five parameters of the feed database. Prepro stands for preprocessed.

local methods (LPLS, LPLS-S, and WLPLS) suggest their potential for significant improvements in the handling of complex spectral data, as indicated by other research in the field [1–3,38]. The LPLS-S and WLPLS showed potential for real-time application by reducing computational time while keeping high performance even without preprocessing. These results underline the need for tailor-made analytical strategies in spectroscopic applications, particularly when handling heterogeneous data such as feed and forage libraries.

To fully understand the contexts in which each method can provide the most benefit, further research with a wide range of datasets is essential. Expanding the scope of data types and analytical scenarios will help clarify the conditions under which LPLS-S and WLPLS outperform traditional LPLS methods. This future research will be crucial in guiding users to make informed decisions on the most appropriate methodologies for their specific analytical needs.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used GPT-4o in order to improve the clarity of some sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Antoine Deryck: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andreas Niemöller:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Conceptualization. **Vincent Baeten:** Writing – review & editing, Supervision. **Juan Antonio Fernández Pierna:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization.

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.

Data availability

The data that has been used is confidential.

References

[1] J.S. Shenk, M.O. Westerhaus, P. Berzaghi, Investigation of a LOCAL calibration procedure for NIR instruments, J. Near Infrared Spectrosc. 5 (4) (1997) 223–232, https://doi.org/10.1255/jnirs.115.
[2] E. Fernández-Ahumada, T. Fearn, A. Gómez-Cabrera, J.E. Guerrero-Ginel, D. C. Pérez-Marín, A. Garrido-Varo, Evaluation of Local Approaches to Obtain Accurate Near-Infrared (NIR) Equations for Prediction of Ingredient Composition of Compound Feeds, Appl. Spectrosc. 67 (8) (2013) 924–929, https://doi.org/10.1366/12-06937.
[3] O. Minet, V. Baeten, B. Lecler, P. Dardenne, J.A. Fernández Pierna, Local vs global methods applied to large near infrared databases covering high variability, in: Proceedings of the 18th International Conference on near Infrared Spectroscopy, 2019, pp. 45–49, https://doi.org/10.1255/nir2017.045.
[4] S. Kim, R. Okajima, M. Kano, S. Hasebe, Development of soft-sensor using locally weighted PLS with adaptive similarity measure, Chemom. Intel. Lab. Syst. 124 (2013) 43–49, https://doi.org/10.1016/j.chemolab.2013.03.008.
[5] G. Shen, M. Lesnoff, V. Baeten, P. Dardenne, F. Davrieux, H. Ceballos, J. Belalcazar, D. Dufour, Z. Yang, L. Han, J.A. Fernández Pierna, Local Partial Least Squares Based on Global PLS Scores, Journal of Chemometrics 33 (2019), 5:e3117, https://doi.org/10.1002/cem.3117.
[6] M. Lesnoff, Averaging a local PLSR pipeline to predict chemical compositions and nutritive values of forages and feed from spectral near infrared data, Chemom. Intel. Lab. Syst. 244 (2024) 105031, https://doi.org/10.1016/j.chemolab.2023.105031.
[7] T. Naes, T. Isaksson, B. Kowalski, Locally weighted regression and scatter correction for near-infrared reflectance data, Anal. Chem. 62 (7) (1990) 664–673, https://doi.org/10.1021/ac00206a003.
[8] T. Naes, T. Isaksson, The idea behind and algorithm for locally weighted regression (LWR), NIR News 5 (4) (1994) 7–8, https://doi.org/10.1255/nirn.258.
[9] T. Naes, T. Isaksson, Some modifications of locally weighted regression (LWR), NIR News 5 (5) (1994) 8–9, https://doi.org/10.1255/nirn.269.
[10] V. Centner, L. Massart, Optimization in Locally Weighted Regression, Anal. Chem. 70 (19) (1998) 4206–4211, https://doi.org/10.1021/ac980208r.
[11] J.S. Shenk, M.O. Westerhaus, Calibration system for spectrographic analyzing instruments, United States US5798526A, filed 24 janvier 1997, issued 25 août 1998, https://patents.google.com/patent/US5798526A/en.
[12] P. Berzaghi, J.S. Shenk, M.O. Westerhaus, LOCAL prediction with near infrared multi-product databases, J. Near Infrared Spectrosc. 8 (1) (2020) 1–9, https://doi.org/10.1255/jnirs.258.
[13] F. Allegrini, J.A. Fernández Pierna, W.D. Fragoso, A.C. Olivieri, V. Baeten, P. Dardenne, Regression models based on new local strategies for near infrared spectroscopic data, Analytica Chimica Acta 933 (2016) 50–58, https://doi.org/10.1016/j.aca.2016.07.006.
[14] M. Lesnoff, D. Andueza, C. Barotin, P. Barre, L. Bonnal, J.A. Fernández Pierna, F. Picard, P. Vermeulen, J.-M. Roger, Averaging and Stacking Partial Least Squares Regression Models to Predict the Chemical Compositions and the Nutritive Values

of Forages from Spectral Near Infrared Data, Appl. Sci. 12 (2022) 15:7850, https://doi.org/10.3390/app12157850.
[15] B.K. Alsberg, A.M. Woodward, D.B. Kell, An introduction to wavelet transforms for chemometricians - A time-frequency approach, Chemom. Intel. Lab. Syst. 37 (2) (1997) 215–239, https://doi.org/10.1016/S0169-7439(97)00029-4.
[16] B. Walczak, D.L. Massart, Wavelets - something for analytical chemistry, Trends Anal. Chem. 16 (8) (1997) 451–463, https://doi.org/10.1016/S0165-9936(97)00065-4.
[17] K. Jetter, U. Depczynski, K. Molt, A. Niemoeller, Principles and applications of wavelet transformation to chemometrics, Anal. Chim. Acta 420 (2) (2000) 169–180, https://doi.org/10.1016/S0003-2670(00)00889-8.
[18] B. Walczak, Wavelets in Chemistry, Data Handling in Science and Technology 22 (2000).
[19] F.T. Chau, Y.Z. Liang, J. Gao, X.-G. Shao, J.D. Winefordner, Chemometrics - From Basics to Wavelet Transform, Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications 1 (2004).
[20] R.A. Viscarra Rossel, R.M. Lark, Improved analysis and modelling of soil diffuse reflectance spectra using wavelets, European Journal of Soil Science 60 (2009), 3: 453-464, https://doi.org/10.1111/j.1365-2389.2009.01121.x.
[21] V.D. Hoang, Wavelet-based spectral analysis, Trends Anal. Chem. 62 (2014) 144–153, https://doi.org/10.1016/j.trac.2014.07.010.
[22] M. Vohland, M. Ludwig, M. Harbich, C. Emmerling, S. Thiele-Bruhn, Using variable selection and wavelets to exploit the full potential of visible–near infrared spectra for predicting soil properties, J. Near Infrared Spectrosc. 24 (3) (2016) 255–269, https://doi.org/10.1255/jnirs.1233.
[23] A. Wakiuchi, S. Jasial, S. Asano, R. Hashizume, M. Hatanaka, Y. Ohnishi, T. Matsubara, H. Ajiro, T. Sugawara, M. Fujii, T. Miyao, Chemometrics Approach Based on Wavelet Transforms for the Estimation of Monomer Concentrations from FTIR Spectra, ACS Omega 8 (22) (2023) 19781–19788, https://doi.org/10.1021/acsomega.3c01515.
[24] T. Guo, T. Zhang, E. Lim, M. López-Benítez, F. Ma, L. Yu, A Review of Wavelet Analysis and Its Applications: Challenges and Opportunities, IEEE Trans. Antennas and Propagation 10 (2022) 58869–58903, https://doi.org/10.1109/ACCESS.2022.3179517.
[25] D. Jouan-Rimbaud, B. Walczak, R.J. Poppi, O.E. de Noord, L. Massart, Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration, Anal Chem 69 (21) (1997) 4317–4323, https://doi.org/10.1021/ac970293n.
[26] U. Depczynski, K. Jetter, K. Molt, A. Niemoeller, Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm, Chemom. Intel. Lab. Syst. 47 (2) (1999) 179–187, https://doi.org/10.1016/S0169-7439(98)00208-1.
[27] Marina Cocchi, Maria Corbellini, Giorgia Foca, Mara Lucisano, M. Ambrogina Pagani, Lorenzo Tassi, Alessandro Ulrici, Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra, Analytica Chimica Acta 544 (2005), 1–2:100-107, https://doi.org/10.1016/j.aca.2005.02.075.
[28] M. Vannucci, N. Sha, P.J. Brown, NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection, Chemom. Intel. Lab. Syst. 77 (1–2) (2005) 139–148, https://doi.org/10.1016/j.chemolab.2004.10.009.
[29] L. Wang, Y. Sun, Image classification using convolutional neural network with wavelet domain inputs, IET Image Proc. 16 (8) (2022) 2037–2048, https://doi.org/10.1049/ipr2.12466.
[30] M. Lesnoff, M. Metz, J.-M. Roger, Comparison of Locally Weighted PLS Strategies for Regression and Discrimination on Agronomic NIR Data, Journal of Chemometrics 34 (2020), 5:e3209, https://doi.org/10.1002/cem.3209.
[31] J.A. Fernández Pierna, V. Baeten, P. Dardenne, Screening of compound feeds using NIR hyperspectral data, Chemometrics and Intelligent Laboratory Systems 84 (2006), 1-2:114-118, https://doi.org/10.1016/j.chemolab.2006.03.012.
[32] J.A. Fernández Pierna, B. Lecler, J. P. Conzen, A. Niemoeller, V. Baeten, P. Dardenne, Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products, Analytica Chimica Acta 705 (2011), 1-2:30-34, https://doi.org/10.1016/j.aca.2011.03.023.
[33] C.-W. Chang, D.A. Laird, Near-Infrared Reflectance Spectroscopic Analysis of Soil C and N, Soil Sci. 167 (2) (2002) 110–116, https://doi.org/10.1097/00010694-200202000-00003.
[34] RStudio Team, RStudio: Integrated Development for R, RStudio, PBC, Boston, MA, 2020, .
[35] M. Lesnoff, R package rchemo: Dimension Reduction, Regression and Discrimination for Chemometrics, 2021, https://github.com/mlesnoff/rchemo.
[36] S. Kucheryavskiy, mdatools – R package for chemometrics, Chemom. Intel. Lab. Syst. 198 (2020), https://doi.org/10.1016/j.chemolab.2020.103937.
[37] G. Sinnaeve, P. Dardenne, R. Agneessens, Global or Local? A Choice for NIR Calibrations in Analyses of Forage Quality, Journal of Near Infrared Spectroscopy 2 (1994), 3:163-175, https://doi.org/10.1255/jnirs.43.
[38] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Implementation of LOCAL Algorithm with Near-Infrared Spectroscopy for Compliance Assurance in Compound Feedingstuffs, Appl. Spectrosc. 59 (1) (2005) 69–77, https://doi.org/10.1366/0003702052940585.