Journal of Food Composition and Analysis 128 (2024) 106015

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

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

Yoghurt standardization using real-time NIR prediction of milk fat and protein content

D. Castro-Reigía [a,b], J. Ezenarro [c], M. Azkune [d], I. Ayesta [e], M. Ostra [f], J.M. Amigo [g,h], I. García [b], M.C. Ortiz [a,*]

[a] Universidad de Burgos, Departamento de Química, Facultad de Ciencias, Plaza Misael Bañuelos s/n, Burgos 09001, Spain
[b] Advanced Optical Technologies S.L. (AOTECH), Escuela Ing. de Bilbao, Plaza Ingeniero Torres Quevedo, 1 2º, Bilbao 48013, Spain
[c] Universitat Rovira i Virgili. Chemometrics and Sensorics for Analytical Solutions (ChemoSens) group, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, N4 building, C/Marcel·lí Domingo 1, Tarragona 43007, Spain
[d] Department of Electronic Technology, Engineering School of Bilbao, University of the Basque Country (UPV/EHU), Plaza Ingeniero Torres Quevedo 1, 48013 Bilbao, Spain
[e] Department of Applied Mathematics, Engineering School of Bilbao, University of the Basque Country (UPV/EHU), Plaza Ingeniero Torres Quevedo 1, 48013 Bilbao, Spain
[f] Departament of Applied Chemistry, Faculty of Chemistry, University of the Basque Country (UPV/EHU), Manuel Lardizabal 3, 20018 Donostia, Spain
[g] IKERBASQUE, Basque Society for the Promotion of Science, Plaza Euskadi, 5, Bilbao 48009, Spain
[h] Department of Analytical Chemistry, University of the Basque Country UPV/EHU, Barrio Sarriena S/N, Leioa 48940, Spain

ARTICLE INFO

Keywords:
Partial least squares regression (PLSR)
Near-infrared (NIR)
In-line
Proof of concept
Yoghurt
Fat
Protein

ABSTRACT

A system based on near-infrared (NIR) spectroscopy has been developed for the in-line control of the composition of the milk used as raw material for yoghurt production to control the content of protein and fat in the final product, and, therefore, to reduce variability in the production process. Firstly, after selecting the appropriate method for preprocessing NIR data, Partial Least Squares Regression models were built to predict fat and protein content in milk, obtaining good performances. The variance explained of **y**-block in prediction (R^2P) was 0.99 and 0.80, while the Root Mean Square Error of Prediction (RMSEP), was 0.26 and 0.16 for fat and protein, respectively. With those models, it was possible to determine the fat and protein contents in milk in real time, and therefore, the quantity of milk powder and cream added in the manufacturing process of yoghurt could be readjusted. The presented strategy allows the improvement of the homogeneity of the final product, reducing the variability of the nutritional values in more than 70% with respect to the traditional recipe, and also meet the target values according to yoghurt producers for fat and protein content, that is, 10% of fat and 5% of protein.

1. Introduction

Milk and its derivatives, such as yoghurt, are an essential part of the human diet, with a considerable increase in its global production over the years, according to the Food and Agriculture Organization of the United Nations (FAO) (FAO, 2022a). The worldwide estimated production has increased from 466000 million kilos in 1980 to around 843000 million kilos in 2018, which has a considerable impact from an economic point of view. Due to its nutrient content, nutritional quality, and energy supply, milk is a key food in the diet at any age. Its nutritional relevance lies fundamentally in its lipidic and protein fraction (fat constitutes approximately 3% and 4% of cow milk while protein is around 3.5%) (FAO, 2022b). Therefore, among other parameters, fat and protein content in commercialized milk are essential since it will be decisive in its nutritional value.

Milk has been analyzed for years using conventional methods such as the Gerber (AOAC Official Method, 2000; ISO 19662:2018, 2018) and Kjeldahl (ISO, 8968–1:2014|IDF 20–1:2014, 2014) methods for fat and protein content determination, respectively. The International Dairy Federation (IDF) and the International Organization for Standardization (ISO) have collaborated on all standardizations, relating these two methods of analysis and sampling for milk and its derivatives to improve consumer protection warranties. These reference methods have been proven extremely useful and continue to play a significant role in the dairy industry nowadays because of their good repeatability, among many other characteristics (Margolies and Barbano, 2018). Despite all of

* Corresponding author.
E-mail address: mcortiz@ubu.es (M.C. Ortiz).

https://doi.org/10.1016/j.jfca.2024.106015
Received 15 September 2023; Received in revised form 12 December 2023; Accepted 22 January 2024
Available online 26 January 2024

that, their use is decreasing, as they are time-consuming, and they can determine just one parameter at a time compared with some other instrumental techniques. For instance, chromatographic techniques can be another option (Buzás et al., 2022; Delmonte et al., 2012; Sukhija and Palmquist, 1988). However, they usually involve long analysis times, due to the necessity of sample preparation steps like extraction and preconcentration, followed by a separation process, which significantly increases the time and the economic costs of the analysis (Nielsen, 2010).

In that sense, spectroscopic techniques can be the best option considering their easy and fast handling and the reduction of analysis time and use of chemical products, which imply environmental and economic advantages. Some of the most used spectroscopic techniques for measuring milk or derived dairy products are visible/near-infrared (Vis/NIR) spectroscopy (Aernouts et al., 2011; Melenteva et al., 2016; Surkova et al., 2019), Raman spectroscopy (Mazurek et al., 2015), middle-infrared (MIR) spectroscopy (Soyeurt et al., 2006), and near-infrared (NIR) spectroscopy (Bittante et al., 2022; Melfsen et al., 2012; Růžičková and Šustová, 2006). In this work, a NIR spectrometer has been implemented, allowing the simultaneous and real-time determination of fat and protein content in milk flow.

NIR spectroscopy was accepted in 2007 as an official analysis method by AOAC to measure fat and protein content in certain types of foods, like meat (AOAC, 2007). Regarding milk and dairy, NIR has also been accepted by ISO to measure fat and protein contents in milk and dairy products (ISO 21543:2020, 2020). NIR enables non-destructive measurements with no sample preparation (which implies economic, temporary, and environmental advantages) and can be used in-line in reflectance mode, allowing the determination of multiple parameters simultaneously with no specific installation or derived stream as it is normally done by measuring in transmittance mode. In this way, real-time monitoring of milk quality parameters is achieved without altering the production line. As the NIR spectral bands are less defined, and their interpretation poses an extra difficulty, advanced multivariate analysis methods must be applied to quantify the abovementioned properties. In this regard, the workhorse regression method has been Partial Least Squares Regression (PLSR) (Geladi and Kowalski, 1986; Haaland and Thomas, 1988). Showing that the combination of NIR and chemometrics offered such good results in quantifying properties of the milk and dairy products, the International Standard ISO 21543:2020 "Milk Products-Guidelines for the application of NIR spectrometry" (ISO 21543:2020, 2020) establishes the performance criteria of this type of analysis with multivariate calibration techniques.

Several studies proposed methods that predict fat and protein content in milk using NIR and PLSR (Melendreras et al., 2022; Yang et al., 2020). These investigations were focused on the reliable determination of some milk attributes or the optimization of portable devices. In this work, the determination of fat and protein in milk samples is proposed for the posterior improvement over the real-time control of the yoghurt manufacturing process.

One of the objectives of dairy producers is to manufacture yoghurts with a final percentage of 10% of fat and 5% of protein. To do that with the usual recipe, the quantity of commercial cream (with a known fat content), milk powder (with a known protein content) and raw milk that needs to be mixed to get a final product with the aforementioned nutritional values has to be calculated based on the fat and protein contents of the raw milk. They used the fat and protein contents from the previous batches of raw milk to calculate those contents, getting a final product with considerable variability. As cow-milk nutritional values depend on a big variety of factors (such as food ingestion, lactation status, season of the year, age, illnesses…), and as the employed milk is a mixture coming from many individual animals, nutritional values of the used milk are likely not to be similar to those of the previous batch (FAO, 2022b). This justifies the necessity of determining the fat and protein contents in real time in raw milk to adjust the quantities of the ingredients and homogenize the final product. This is, to avoid large differences of fat and protein contents between yoghurt batches. In this work it was shown that the implementation of in-line NIR reflectance measurements in the milk supply chain leads to a substantial accuracy improvement in the homogenization of the final product. Therefore, a standardization in the manufacturing process of yoghurts is achieved, attaining a higher quality and reducing costs.

Fig. 1. Measuring and sampling system for milk samples. The NIR sensor (blue cylinder) and the outlet for sampling (indicated with a red circle) are installed in the pipe where the milk flows from the storage tank to the pasteurization system.

2. Materials and methods

2.1. Milk and yoghurt samples

The experimental work of this investigation has been carried out in a dairy factory in Spain (Dulce Grado S.L). Milk supplies arrive daily to the factory, coming from different farms and different suppliers, and are then stored in tanks. As they have different origin, as mentioned in the introduction section, the nutritional value of these milk samples will vary according to that. The experimental milk measurements were made in-line in this factory in a five-month period, so the greatest possible variability between milk samples and their nutritional values could be explored and recorded as spectra. Table S1 in the supplementary material shows the number of samples measured each day during the five-month period. This was made to take into account the different conditions that could affect the final product. The fat content ranges in milk samples between 2.65% and 4.06%; while the protein content ranges between 3.17% and 3.46% (w/w).

On the other hand, a total of 6 yoghurt samples were manufactured based on the amount of fat and protein calculated in-line with the models proposed below. Each batch of yoghurt implies 0.2 m^3 of milk (tank volume). Three batches of yoghurt were manufactured taking two samples per batch and their corresponding spectra, which implies a total of 6 yoghurt samples, but 0.6 m^3 of milk. By knowing the fat and protein content of milk in real time, the quantities of cream (to adjust the amount of fat) and milk powder (to adjust the the amount of protein) in the 6 yoghurts were modified, since they depend on the fat and protein of the raw milk. These yoghurts were sent to an external laboratory for their analysis, where the fat content was determined using the gravimetric method and the amount of protein was calculated using the Kjeldahl method (ISO, 8968–1:2014|IDF 20–1:2014, 2014).

2.2. NIR spectroscopy

The procedure was made with the AONIR integrated solution for real-time NIR measurments (AOTECH, 2023), including a NIR sensor coupled to a control and model software platform, which allows recording the NIR diffuse reflectance spectra (from 908 to 1676 nm) of 50 milk samples in real time. The sensor was configured in such a way that the best spectra were obtained with a spectrum reading interval of one second, and 50 readings per spectrum with an integration time of

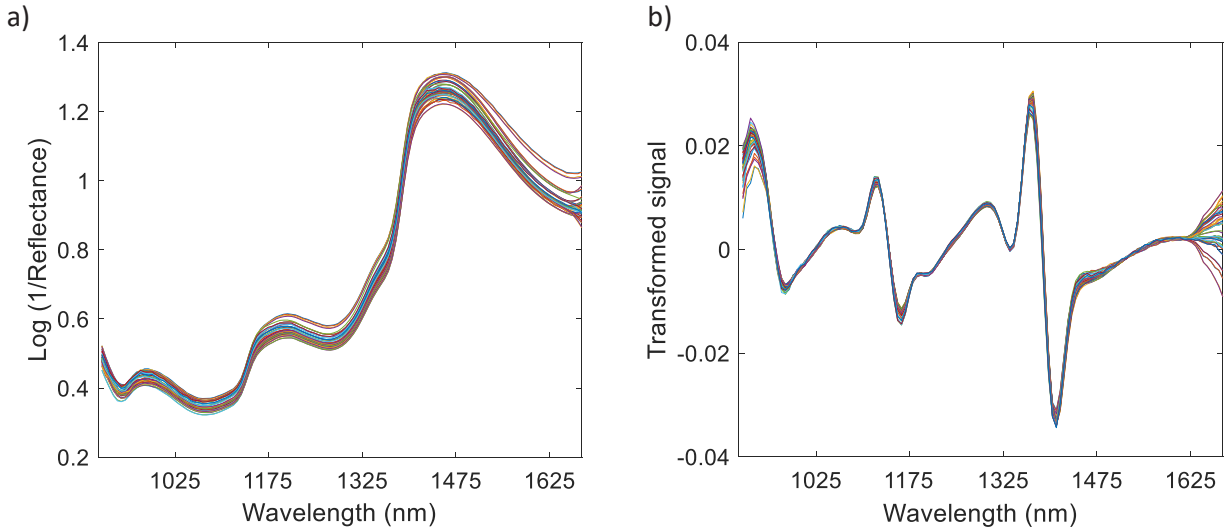

Fig. 2. a) NIR spectra of 50 samples of milk. b) Preprocessed NIR spectra of 50 samples of milk using SNV and S-G with a window width of 11 points using a second-degree polynomial and a 2nd derivative.

0.012 s. Each sample was measured in such a way that both sampling and NIR measurements were performed simultaneously without any specialized member. Fig. 1 represents the measurement and the sampling system whereas Figure S1 in the supplementary material shows the AONIR software output. Daily milk production is stored in large tanks, from which the milk flows through pipes to the pasteurization system. Through that flow, samples are collected and measured. Both the NIR sensor in Fig. 1 (blue cylinder) and the outlet for sampling (indicated with a red circle) are installed in the pipe. Once the pipe is filled, a NIR spectrum is manually recorded and immediately after that, 40 mL of milk is taken to an external laboratory to be analyzed with MILKOSCAN, a certified reference method by AOAC (AOAC, 2016) and ISO and IDF (ISO, 9622:2013|IDF 141:2013, 2013). On the other hand, Figure S1 shows the output of the AONIR software, giving as an example one spectra of milk recorded during this work.

With the 50 samples, calibration models were made to model the milk fat and protein content. Afterwards, six new samples were measured and used to make yoghurts whose fat and protein were also measured by certified laboratories.

2.3. Multivariate regression models

To make the reading of this paper easier, the elements that are going to be used in it are explained below. The relation between the collected NIR spectra (**X**) and the correspondent responses (fat and protein content, **y**) is established using a PLSR. This algorithm builds linear combinations of **X** and **y**, maximizing their covariance, and finding a new set of latent variables (LVs) in **X**-block and **y**-block maximally related to them (Geladi and Kowalski 1986).

PLSR is especially useful when the predictors are highly collinear or when the number of predictors is higher than the number of observations, being widely used in spectroscopic applications (Haaland and Thomas 1988). Nevertheless, to build an optimalmodel, data must be properly preprocessed to reduce as much unwanted variations as possible, such as instrumental or thermal noise, sample background or light scattering effects as these effects are more prominent in diffuse reflectance measurements than in transmittance measurements. The most widely used preprocessing methods can be divided in scatter-correction methods and spectral derivatives, (Chu et al., 2022; Mas et al., 2020; Rinnan et al., 2009; Schoot et al., 2020). Within the first, the Multiplicative Scatter Correction (MSC) and the Standard Normal Variate (SNV) were the ones used in this work, whereas Savitzky-Golay (S-G) derivative calculation was applied as the spectral derivative method. Data were also mean-centred (MC) before modelling.

The SNV normalization is mainly used to correct light scattering effects and changes in the optical path on the NIR reflection spectra (Chu et al., 2022; Rinnan et al., 2009; Schoot et al., 2020) although it could be sensitive to noisy entries in the spectrum, since it does not involve a least square fitting in their parameter estimations. SNV mathematical details can be consulted in the references indicated above.

The purpose of Multiplicative Scatter Correction (MSC) is practically the same as SNV, this is, to remove the effects of particle distribution and size. For a dataset formed by individual spectra **X** (1 x *m*), its average spectrum ($\overline{\mathbf{X}}$) is calculated. Next, a linear regression between each **X** and $\overline{\mathbf{X}}$ is performed, obtaining b and b_0 coefficients by least squares. Then, **X** is corrected by subtracting b_0 and dividing by b. MSC is performed assuming independence from the wavelength and the variations in the composition of samples. It is proved to be correlated with SNV, in fact, the preprocessing results of both methods should be analogous (Chu et al., 2022).

The S-G filter adjusts a polynomial to a moving window through the wavelength points of a spectrum using least squares, then, the central point of the window is predicted using the fitted equation, which can be mathematically derived before the prediction. As the equation does not fit perfectly to the data, this filter has a smoothing effect apart from the derivation. S-G can remove most of baseline interferences and background noise, but proper selection of the window width, the derivative order and the degree of the polynomial must be made. If the window width is too small, the noise is augmented, but, if the difference width is too large, the spectrum becomes excessively smoothed, losing information on the peaks of interest (Chu et al., 2022; Rinnan et al., 2009).

To selecting the appropriate number of LVs Venetian blinds cross-validation method was employed in every case considering the ordering of the samples, the number of objects and the presence of replicate samples in the dataset. The choice of the optimal number of LVs for each preprocessing was obtained by comparing the Root Mean Square Error in Calibration (RMSEC) and the Root Mean Square Error in Cross-validation (RMSECV). The optimal number of LVs was decided by a threshold in the rate of change of the RMSECV between two consecutive ranks.

Permutation tests are another way to help identify an overfit model as well as provide a probability that the given model is significantly different from one built under the same conditions but on random data (Tools: Permutation Test-Eigenvector Research Documentation, 2023). These tests involve repeatedly and randomly reordering the **y**-block, in such a way that the model is rebuilt after each reordering with the

Table 1
PLSR models fitted for fat and protein for each preprocessing. L.V., number of latent variables; R^2C, variance explained of Y block in fitting; R^2CV, variance explained of Y block in cross-validation; R^2P, variance explained of Y block in prediction; RMSEC, Root Mean Square Error in Calibration and RMSECV (%); RMSECV, Root Mean Square Error in Cross-Validation (%); RMSEP, Root Mean Square Error in Prediction (RMSEP). *p*-value is the significance for the cross-validated permutation tests.

	Preprocessing method	LV	R^2C	R^2CV	R^2P	Bias	CV Bias	RMSEC (%)	RMSECV (%)	RMSEP (%)	*p*-value		
											Pairwise Wilcoxon signed rank test	Pairwise signed rank test	Randomisation t-test
Fat	SNV+MC	5	0.895	0.853	0.986	-4.44×10^{-16}	-1.99×10^{-3}	0.10	0.12	0.29	0.000	0.008	0.005
	MSC+MC	4	0.883	0.851	0.986	2.22×10^{-15}	-6.81×10^{-4}	0.11	0.12	0.21	0.001	0.011	0.006
	2nd derivative+MC	6	0.931	0.884	0.951	0.00	3.06×10^{-3}	0.08	0.10	0.40	0.000	0.000	0.005
	2nd derivative + SNV+ MC	6	0.939	0.890	0.944	0.00	3.82×10^{-3}	0.08	0.10	0.34	0.000	0.000	0.005
	*SNV+ 2nd derivative +MC	6	0.960	0.907	0.991	-1.02×10^{-14}	1.42×10^{-3}	0.06	0.09	0.25	0.000	0.001	0.005
Protein	SNV+MC	6	0.826	0.720	0.785	-8.88×10^{-16}	-1.74×10^{-4}	0.04	0.05	0.65	0.000	0.004	0.006
	MSC+MC	6	0.827	0.720	0.797	-8.88×10^{-16}	-2.17×10^{-4}	0.04	0.05	0.65	0.000	0.003	0.006
	2nd derivative+MC	3	0.565	0.422	0.008	0.00	8.57×10^{-4}	0.07	0.08	0.72	0.008	0.064	0.009
	2nd derivative + SNV + MC	4	0.606	0.439	0.029	4.44×10^{-16}	-1.80×10^{-3}	0.06	0.08	0.71	0.004	0.030	0.008
	*SNV+ 2nd derivative +MC	9	0.979	0.861	0.800	0.00	1.27×10^{-3}	0.01	0.04	0.16	0.000	0.001	0.005

* Selected final models

current modelling settings. In this case, three tests were used, the Wilcoxon test, the signed test and the randomisation test. If the modeling conditions are over-fitted, they will often provide a fit to random data which is better than would be expected. If the *p*-value is greater that 0.05, the model will be probably overfitted at a 95% confidence level.

After the PLSR models were built, to demonstrate their veracity, the respective accuracy lines (the predicted content with the PLSR models versus the true concentrations) were built. On the one hand, for the case of protein, the slope is 0.9788 whereas the intercept is 0.0708, and s_{yx}= 0.0138. On the other hand, for the case of fat, the slope is 0.9598 while the intercept is 0.1474, and s_{yx}= 0.0617. Knowing that the ideal situation of an unbiased model is when the slope is equal to one and the intercept zero, the centers of the ellipses, represent the value of the slope and the intercept of the models. The region defined within the ellipse corresponds to the confidence region calculated, in this work, at the 95% confidence level.

2.4. Software

PLS Toolbox v8.8.1 (Wise et al., 2022) for use with MATLAB (R2020b) (MATLAB, 2022) was employed for fitting the PLSR models. A homemade program on MATLAB was used for calculate the confidence ellipses.

3. Results and discussion

3.1. NIR spectra of milk

The obtained spectra are represented in Fig. 2a. Related to water bands, in the milk spectra of the samples, three main bands were observed: one at around 980 nm, which corresponds to the second overtone of the symmetric and the asymmetric stretch of the water molecule, a second one at approximately 1200 nm, which corresponds to the first overtone of the symmetric stretch, the bending mode, and the asymmetric stretch, and the third one, at nearly 1450 nm, that corresponds to the first overtone of the symmetric and asymmetric stretch of the water molecule (Weyer, 2007). It can also be observed a small peak around 1300–1350 nm that correspond to the first overtone of the symmetric stretch of the water molecule. In the case of the protein content, the characteristic bands in NIR will be those associated with the functional groups that define the aminoacids that conform the proteins, such as the N-H and the -COOH groups. Therefore, the most relevant bands are those corresponding to the first and second overtones of N-H stretching (around 1500 nm and between 973 and 1020 nm, respectively). Also, the characteristic bands that correspond to the first N-H overtone of the symmetric and asymmetric combination stretch for primary amides (1470 nm) (Weyer, 2007). The band around 930 nm could indicate the third overtone of C-H stretch vibrations of triglycerides or the third overtone of the O-H bond (Aernouts et al., 2011).

Fig. 2b shows the preprocessed spectra performing the SNV and the 2nd derivative (as this is the optimal preprocessing for the models employed to predict fat and protein contents in milk). With this representation, the most important and previously mentioned bands are highlighted, and if they are really related to fat and protein content, should be similar to the peaks of the loading values that will be obtained with the final models used to predict fat and protein.

3.2. PLSR calibration models for fat and protein

The figures of merit obtained for the PLSR models built with each considered preprocessing method are summarized in Table 1. Considering the previous reasoning in the Section 2.2, the selected preprocessing method was SNV combined with S-G with a window width of 11 points using a second-degree polynomial and a 2nd derivative for both fat and protein. Data were mean-centred before modelling. The number of LVs was chosen using the venetian blinds cross-validation

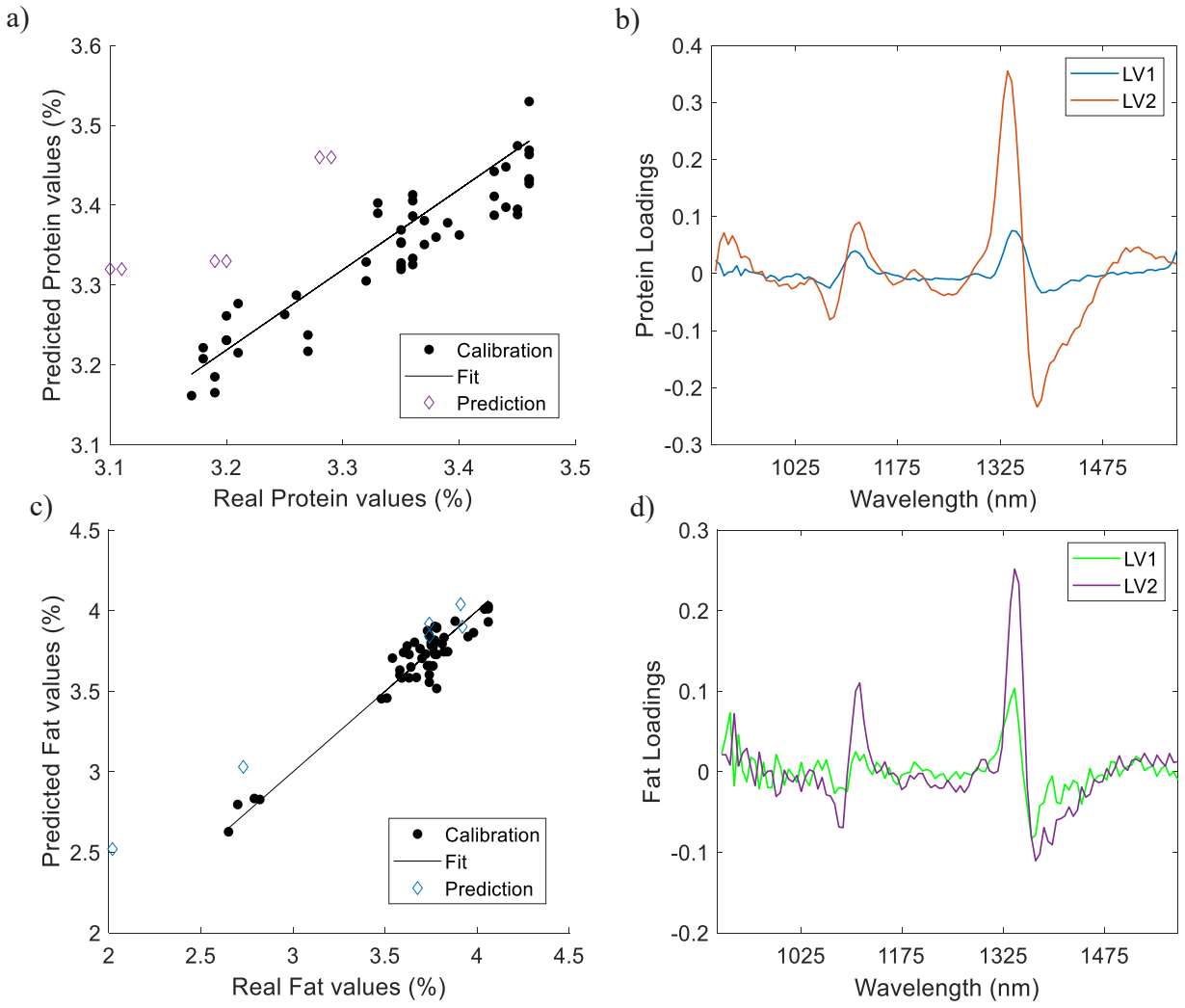

Fig. 3. a) Accuracy line for protein model. b) Loadings of the first two LVs of protein model. In blue the loadings of the first LV and, in orange the loadings of the second LV. c) Accuracy line for fat model. d) Loadings of the first two LVs of fat model. In green the loadings of the first LV and, in purple the loadings of the second LV.

procedure, with six LVs being optimal for the fat prediction model and nine for the protein. The global percentage of explained variance in the calibration set is 96% for fat and 98% for protein, while in the CV set are 91% and 86%, respectively. The absence of overfitting has been evaluated by doing three permutation tests (50 iterations) using the residuals in CV and all *p*-values of the permutation test were lower that 0.05.

Fig. 3a represents the accuracy line for protein model and Fig. 3b shows the correspondent loadings. Observing Fig. 3b, the LV with higher positive correlation with the diffuse reflectance NIR spectra are the first (73% of the **X**-block variance) and the second one (11% of the **X**-block variance). The first LV is more related with water content since the bands at aproximately 1150 nm and 1350 nm correspond to the first overtone of the symmetric stretch, the bending mode and the asymmetric stretch of the water molecule. In the second LV, a gentle peak is observed at approximately 980 nm, that, would correspond to the N-H stretch second overtone, surely related with the presence of proteins. The bands corresponding to the first and second overtones of N-H stretching can be seen around 1500 nm and, the characteristic bands that correspond to the first N-H overtone of the symmetric and asymmetric combination stretch for primary amides (1470 nm) (Weyer, 2007). Besides, high negative loadings are observed at around 1450 nm in the second LV that emphasize that the second variable is not related with water content since that wavelength corresponds to the first overtone of the symmetric and asymmetric stretch of the water molecule.

On the other hand, Fig. 3c represents the results for fat model. Observing Fig. 3d, the fat loadings for the first and second LV, explain 34% and 30% of the **X**-block variance, respectively, being the most important ones. It appears that the loading weights that contribute most to the PLSR models are mainly positive. That indicates a positive correlation between fat composition and the diffuse reflectance of milk samples.

Fig. 4 shows the confidence ellipses at a significant level of 95%, demonstrating the veracity of both regression models by means of the accuracy lines. As it can be seen, the confidence region includes one for the slope and zero for the intercept, respectively, so it can be affirmed that the model is not biased, neither in a proportional way nor in a constant way (slope=1). Also, it can be observed that the protein prediction model is more accurate than the the fat prediction model, because there is a larger residual standard deviation for fat (and therefore, the ellipse is also larger).

With all these results, it can be found the model built for the determination of fat content had the best performance, while the variance explained by the protein model is lower. A possible explanation of this lies in the size of fat particles in milk, since it is similar to the wavelengths of the spectra (around 1000 nm) (Bogomolov et al., 2013), the

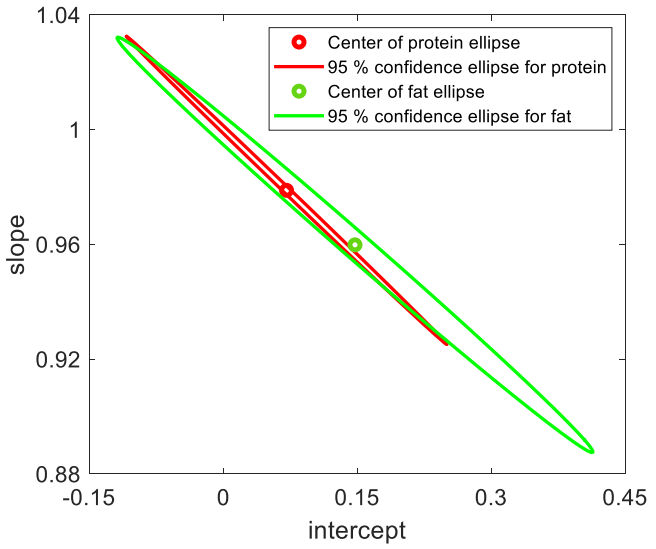

Fig. 4. 95% confidence level ellipse for fat and protein PLSR models. The circles represent the slope and the intercept for the PLSR models, and the ellipse, its confidecnce interval at a 95% confidence level, in red for protein and in green for fat.

Table 2
Proof of concept. Fat and protein percentage in milk samples using a reference method and using AONIR with PLSR models. RMSEP, Root Mean Square Error in Prediction (RMSEP); R^2P, variance explained of **y**-block in prediction.

Sample	Reference fat content (%)	Predicted fat content with AONIR (%)	Reference protein content (%)	Predicted protein content with AONIR (%)
1	3.74	3.92	3.10	3.32
2	3.74	3.84	3.11	3.32
3	2.02	2.52	3.20	3.33
4	2.73	3.03	3.19	3.33
5	3.91	4.04	3.29	3.46
6	3.92	3.90	3.28	3.46
RMSEP	0.26		0.16	
R^2P	0.99		0.80	

fat particles will have stronger capacity to reflect light. On the other hand, since the size of protein micelles, like the casein present in the milk, is normally lower than 200 nm (Ruettimann and Ladisch, 1987), the scattering is dependent on the wavelength, therefore, the intensity will decrease with the energy of light.

3.3. In-line validation of protein and fat content prediction models

With the models proposed in the previous section, new measurements of fat and protein were made on six samples of flowing milk as a proof of concept. The results obtained for fat and protein in milk with AONIR along with the lab results using a reference method (FOSS Milkoscan™ FT equipment) can be found in Table 2. With that data, fit or reference values to predicted values (R^2) is similar to the cross-validation predictions for the fat model, while in the case of protein the external validation fit is a bit lower.

3.4. Correction in the yoghurt manufacture

As the measurements were made in-line and the results of fat and protein content were obtained in real time, yoghurt manufacturing was made according to those results instead of using the usual recipe, which uses values from previous batches. This made possible to reduce the variability in the manufacturing process and to obtain yoghurt with contents of fat and protein that are more in accordance with what is desired.

Table 3
Mean and standard deviation values for fat and protein contents obtained with of the usual recipe of yoghurts.

Sample	Fat content in milk	Protein content in milk	Fat content in yoghurt	Protein content in yoghurt
Mean (%)	4.31	3.24	8.85	4.40
Standard deviation	1.69	0.07	0.91	0.54
RSD (%)	39	2	10	12

Table 4
Fat and protein percentage in raw milk and in yogurths for six samples to calculate the standard deviation in the nutritional content of the final product.

Sample	Fat content in milk (%)	Protein content in milk (%)	Fat content in yoghurt (%)	Protein content in yoghurt (%)
1	3.92	3.32	8.80	4.20
2	3.84	3.32	8.50	4.40
3	2.52	3.33	8.50	4.50
4	3.03	3.33	8.50	4.60
5	4.04	3.46	9.10	4.40
6	3.90	3.46	8.90	4.50
Mean (%)	3.54	3.37	8.72	4.43
Standard deviation	0.62	0.07	0.26	0.14
RSD (%)	18	2	3	3

Table 3 shows the mean values and the standard deviations for fat and protein content obtained with the reference method in six different samples using the traditional recipe, both for milk and yoghurt. As the relative standard deviations (RSD) show, the variaton of fat and protein contents in yoghurt is considerable.

Considering these results, a solution has been proposed based in in-line NIR diffuse reflectance and PLSR models. Once the problem with the variations in fat and protein content of yoghurt obtained using the original recipe was found, it has been considered that a homogenization of the final product and its nutritional values is neccesary. For this purpose, a proof of concept was carried out using six samples. Table 4 represents on the one hand, the fat and protein content in milk calculated in-line using diffuse NIR reflectance combined with the PLSR models developed in Section 3.2. With those results, the quantities of the ingredients (cream and milk powder) that needed to be added to manufacture yoghurts with the correct contents of fat and protein were calculated. On the other hand, fat and protein contents for the manufactured yoghurts (that were calculated by gravimetry and using the kjeldahl method, respectively) are also represented. If the mean values of yoghurts in Table 3 are compared with those in the Table 4, it can be seen that they are similar for the yogurths, nevertheless, observing the standard deviation and the RSD values, it is clear that in the second case, they are smaller. In this way, a homogenization of the final product was achieved, getting similar values of nutritional content in the six yoghurt samples reducing the variability in fat content by 72% and by 75% in the case of protein.

With data from Tables 3 and 4, an interval-based test is applied to decide the non-inferiority of the mean value of the percentage of fat and protein in relation to a target value (as aforementioned, one of the objectives of dairy procedures is to manufacture yoghurts with a final percentage of 10% of fat and 5% of protein) (Ortiz, 2020).

The null hypothesis of this test is $\mu - T < \Delta L$, where μ is the sample mean, T the target value and, ΔL the lower equivalence differencial, that is, the difference allowed to asses the non-inferiority. This non-inferiority test is designed to demonstrate that the mean value obtained with a procedure (reference method or new procedure) is not lower than a target value (10% in the case of fat and 5% for protein). In

Table 5
Non-inferiority tests. Comparison to a target value for fat and protein in yoghurts when the content has been obtained with traditional method or the new procedure (in-line NIR-PLSR).

	Comparison	p-value	Conclusion (alpha=5%)
Fat content in yoghurt (%)	Traditional method versus target value	0.195	Non-inferiority has not been demonstrated
	New procedure versus target value	0.047	Non-inferiority has been demonstrated
Protein content in yoghurt (%)	Traditional method versus target value	0.263	Non-inferiority has not been demonstrated
	New procedure versus target value	0.013	Non-inferiority has been demonstrated

both cases ΔL has been fixed at 15% of the target value (ΔL= −1.5 for fat and ΔL= −0.75 for protein). When using this test procedure, non-inferiority may only be asserted if the *p*-value is less than 0.05. As can be seen in Table 5, only when the new procedure is applied, the non-inferiority has been demonstrated.

These results show that the proof of concept, that needed a total ammount of milk of approximately 0.6 m^3, might provide a promising application for the homogenization of yoghurts as a final product and the standardization of the manufacturing process. This can be considered as an initial approach to analyze more samples and verify industrial variability by using and implementing control charts with more concrete specifications.

4. Conclusions

In this work, a real-time methodology was developed to reduce the manufacturing variability of yoghurts by quantifying fat and protein contents of the raw milk used to produce them. With this new approach, which can be easily automated, the nutritional properties of the final product (yoghurt) are homogenised, improving the standard method without the necessity of qualified staff for analysing the samples. In addition, the feasibility of using a NIR spectrometer combined with PLSR for this purpose was proved. The applied chemometrics strategy allowed the determination of fat and protein in real time with good accuracy, nevertheless, the calibration ranges used in this work are reduced since they respond to real ranges of the industrial process because the production process is already established in an industry and the variability of analytical parameters of sequential batches is small. Therefore, for future calibration maintenance, these ranges should be increased to the maximum extent possible, as is usually done in the pharmaceutical industry. In the case of milk/yoghurt, it is possible to dilute the samples to obtain lower concentrations or to concentrate them (adding powdered milk to modify the amount of protein, or cream in the case of fat). In general, this could improve the precision of the calibrations. The external validation was carried out using six yoghurt samples as a proof of concept, so further measurements are needed to a stronger validation in a future work. Yoghurt manufacturing standardization was accomplished proving that the proposed method can be successfully implemented in the dairy industry since it has been possible to reduce the variability between batches of yoghurt by up to 70% and also meet the target values according to yoghurt producers for fat and protein content, which are 10% of fat and 5% of protein.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824769 and from the European Union-NextGenerationEU under the INVESTIGO programe. Grant URV Martí i Franqués – Banco Santander (2021PMF-BS-12).

CRediT authorship contribution statement

Amigo J.M,: Conceptualization, Formal analysis, Supervision, Validation, Writing – review & editing. **Ostra M.:** Conceptualization, Formal analysis, Supervision, Writing – review & editing. **Ortiz M.C. Cruz:** Data curation, Formal analysis, Supervision, Validation, Writing – review & editing. **García I.:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision. **Ezenarro J.:** Conceptualization, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Castro-Reigía D.:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Ayesta I.:** Conceptualization, Formal analysis, Supervision, Writing – review & editing. **Azkune M.:** Conceptualization, Formal analysis, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial or personal relationships that could have appeared to influence the work reported in this paper. This project reflects the views of the author, and the European Union is not responsible for any use that may be made of the information it contains.

Data availability

The authors do not have permission to share data.

Acknowledgements

Authors appreciate the collaboration of Dulcegrado for providing the milk samples during the project and ALS LIFE SCIENCES GALICIA S.L. for their contribution in the yogurth analysis using reference methods.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jfca.2024.106015.

References

Aernouts, B., Polshin, E., Lammertyn, J., Saeys, W., 2011. Visible and near-infrared spectroscopic analysis of raw milk for cow health monitoring: reflectance or transmittance? J. Dairy Sci. 94 (11), 5315–5329. https://doi.org/10.3168/jds.2011-4354.

AOAC Official Method 972.16. Fat, Lactose, Protein and Solids in Milk, (2016).

AOAC. Official Method 2007.04. Fat, Moisture and Protein in Meat and Meat Products. Mid-Infrared Spectroscopy Method., (2007).

AOAC Official Method 2000.18. Fat Content of Raw and Pasteurized Whole Milk. Gerber Method by Weight, (2004).

AOTECH. Advanced Optical Technologies. (2023). Retrieved from https://www.aotech.es/. Accessed December 11, 2023.

Bittante, G., Patel, N., Cecchinato, A., Berzaghi, P., 2022. Invited review: a comprehensive review of visible and near-infrared spectroscopy for predicting the chemical composition of cheese. J. Dairy Sci. 105 (3), 1817–1836. https://doi.org/10.3168/jds.2021-20640.

Bogomolov, A., Melenteva, A., Dahm, D.J., 2013. Fat globule size effect on visible and shortwave near infrared spectra of milk. J. Infrared Spectrosc. 21 (5), 435–440. https://doi.org/10.1255/jnirs.1076.

Buzás, H., Székelyhidi, R., Szafner, G., Szabó, K., Süle, J., Bukovics, S., Kovács, A.J., 2022. Developed rapid and simple RP-HPLC method for simultaneous separation and quantification of bovine milk protein fractions and their genetic variants. Anal. Biochem. 658 https://doi.org/10.1016/j.ab.2022.114939.

Chu, X., Huang, Y., Yun, Y.H., Bian, X., 2022. Chemometric methods in analytical spectroscopy technology. Chemom. Methods Anal. Spectrosc. Technol. https://doi.org/10.1007/978-981-19-1625-0.

Delmonte, P., Fardin-Kia, A.R., Kramer, J.K.G., Mossoba, M.M., Sidisky, L., Tyburczy, C., Rader, J.I., 2012. Evaluation of highly polar ionic liquid gas chromatographic column for the determination of the fatty acids in milk fat. J. Chromatogr. A 1233, 137–146. https://doi.org/10.1016/J.CHROMA.2012.02.012.

FAO. Food and Agriculture of the United Nations. Global Dairy sector: Status and trends. (2022a).

FAO. Food and Agriculture Organization of the United Nations. (2022b). Gateway to dairy production and products. Retrieved from https://www.fao.org/dairy-production-products/products/milk-composition/en/. Accessed June 6, 2023.
Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 186.
Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60 (11), 1193–1202.
ISO 19662:2018. Milk-Determination of fat content-Acido-butyrometric (Gerber method), (2018).
ISO 21543:2020. Milk and milk products-Guidelines for the application of near infrared spectrometry, (2020).
ISO 8968–1:2014|IDF 20–1:2014. Milk and milk products-Determination of nitrogen content-Part 1:Kjeldahl principle and crude protein calculation, (2014).
ISO 9622:2013|IDF 141:2013. Milk and liquid milk products-Guidelines for the application of mid-infrared spectrometry, (2013).
Margolies, B.J., Barbano, D.M., 2018. Determination of fat, protein, moisture, and salt content of Cheddar cheese using mid-infrared transmittance spectroscopy. J. Dairy Sci. 101 (2), 924–933. https://doi.org/10.3168/jds.2017-13431.
Mas, C., Rubio, L., Valverde-Som, L., Sarabia, L.A., Ortiz, M.C., 2020. Impact of the pretreatment of ATR-FTIR signals on the figures of merit when PLS is used. Chemom. Intell. Lab. Syst. 201, 104006 https://doi.org/10.1016/j.chemolab.2020.104006.
MATLAB (2022). MATLAB version: 9.9.0 (R2020b). The Mathworks Inc., Natick, Massachusetts.
Mazurek, S., Szostak, R., Czaja, T., Zachwieja, A., 2015. Analysis of milk by FT-Raman spectroscopy. Talanta 138, 285–289. https://doi.org/10.1016/j.talanta.2015.03.024.
Melendreras, C., Forcada, S., Fernández-sánchez, M.L., Fernández-colomer, B., Costa-fernández, J.M., López, A., Ferrero, F., Soldado, A., 2022. Near-Infrared sensors for onsite and noninvasive quantification of macronutrients in breast milk. Sensors 22 (4), 1–12. https://doi.org/10.3390/s22041311.
Melenteva, A., Galyanin, V., Savenkova, E., Bogomolov, A., 2016. Building global models for fat and total protein content in raw milk based on historical spectroscopic data in the visible and short-wave near infrared range. Food Chem. 203, 190–198. https://doi.org/10.1016/j.foodchem.2016.01.127.
Melfsen, A., Hartung, E., Haeussermann, A., 2012. Accuracy of in-line milk composition analysis with diffuse reflectance near-infrared spectroscopy. J. Dairy Sci. 95 (11), 6465–6476. https://doi.org/10.3168/jds.2012-5388.
Nielsen, S.S. (2010). *Food analysis* (4th ed.).
Ortiz, M.C., Sarabia, L.A., Sánchez, M.S., Herrero, A., 2020. Quality of analytical measurements: statistical methods for internal validation. Compr. Chemom. Chem. Biochem. Data Anal. 1, 1–52. https://doi.org/10.1016/B978-0-12-409547-2.14746-8.
Rinnan, Å., Berg, F. van den, Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends Anal. Chem. 28 (10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.
Ruettimann, K.W., Ladisch, M.R., 1987. Casein micelles: structure, properties and enzymatic coagulation. Enzym. Microb. Technol. 9 (10), 578–589. https://doi.org/10.1016/0141-0229(87)90109-8.
Růžičková, J., Šustová, K., 2006. Determination of selected parameters of quality of the dairy products by NIR spectroscopy. Czech J. Food Sci. 24 (6), 255–260. https://doi.org/10.17221/3322-cjfs.
Schoot, M., Kapper, C., van Kollenburg, G.H., Postma, G.J., van Kessel, G., Buydens, L.M. C., Jansen, J.J., 2020. Investigating the need for preprocessing of near-infrared spectroscopic data as a function of sample size. Chemom. Intell. Lab. Syst. 204, 104105 https://doi.org/10.1016/j.chemolab.2020.104105.
Soyeurt, H., Dardenne, P., Dehareng, F., Lognay, G., Veselko, D., Marlier, M., Bertozzi, C., Mayeres, P., Gengler, N., 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89 (9), 3690–3695. https://doi.org/10.3168/jds.S0022-0302(06)72409-2.
Sukhija, P.S., Palmquist, D.L., 1988. Rapid method for determination of total fatty acid content and composition of feedstuffs and feces. J. Agric. Food Chem. 36 (6), 1202–1206. https://doi.org/10.1021/jf00084a019.
Surkova, A., Belikova, V., Kirsanov, D., Legin, A., Bogomolov, A., 2019. Towards an optical multisensor system for dairy: global calibration for fat analysis in homogenized milk. Microchem. J. 149, 104012 https://doi.org/10.1016/j.microc.2019.104012.
Tools: Permutation Test-Eigenvector Research Documentation Wiki. 2023. Retrieved from: https://www.wiki.eigenvector.com/index.php?title=Tools:_Permutation_Test. Accessed June 6, 2023.
Weyer, L. (2007). Practical guide to interpretative near-infrared spectroscopy.
Wise, B.M., Gallagher, N.B., Bro, R., Shaver, J.M., Winding, W., Koch, R. 2022. PLS Toolbox 8.8.1. Eigenvector Research Inc.
Yang, B., Zhu, Z., Gao, M., Yan, X., Zhu, X., Guo, W., 2020. A portable detector on main compositions of raw and homogenized milk. Comput. Electron. Agric. 177, 105668 https://doi.org/10.1016/j.compag.2020.105668.