

Near-Infrared Untargeted Metabolomics with Unsupervised and Supervised Multivariate Statistical Analysis of Fatty Acid Profiles in Cheeses

Blanca Nayelli Ocampo-Morales¹, Arturo Hernández-Montes¹, José Enrique Herbert-Pucheta^{2*}

¹Departamento de Ingeniería Agroindustrial, Universidad Autónoma Chapingo, km. 38.5 Carretera México- Texcoco, 56230 Chapingo, Estado de México, México.

²Departamento de Química Orgánica, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, Prolongación de Carpio y Plan de Ayala s/n, Colonia Santo Tomás, Ciudad de México 11340, México.

*Corresponding author: José Enrique Herbert-Pucheta, email: jherbertp@ipn.mx

Received January 31st, 2024; Accepted September 23rd, 2024.

DOI: <http://dx.doi.org/10.29356/jmcs.v69i3.2212>

Abstract. The present work describes a workflow for unsupervised Principal Component (PCA) and supervised Partial Least Squares Discriminant (PLS-DA) multivariate statistical analysis (MSA), to analyze Near Infrared (NIR) data matrixes of cheeses from diverse types and geographical origins, with respect to their NIR saturated fatty acid profile. The data set include (A) acquired NIR absorbance spectra, (B) post-processed first derivative NIR spans and (C) post-processed first derivative frequency-selected NIR spans, within a wavelength range of 12500-3600 cm^{-1} . NIR data inputs were adapted for the first time into a format suitable for the stream-lined metabolomics data analysis “MetaboAnalyst”, by converting spectrophotometer raw data format, into a JCAMP-DX IUPAC standard format family for spectral data exchange, in turn transformed into an editable comma-separated values (.csv) format, suitable for metabolomics studies with MetaboAnalyst. The discriminant regions for the first NIR data matrix were five. For the second matrix, discriminant wave-number regions were reduced to three: 10000 to 8000 cm^{-1} (-CH- overtone), 6000 to 5000 cm^{-1} (-C=O- overtone) and 5000 to 4000 cm^{-1} (-CH- band). Finally, for the third NIR matrix, refined discriminant regions were taken: 9700 to 8265 (-CH- overtone), 6661 to 4655 cm^{-1} (-C=O- overtone) and from 4327 to 4000 cm^{-1} (-CH- band). The PLS-DA model obtained from the first derivative frequency-selected near-infrared spans data matrix showed the best score-plot classification between dairy samples and saturated fatty acid standards. Present results intend to introduce an approach for untargeted and qualitative NIR based metabolomics within a platform with more than 300,000 users to date.

Keywords: Near infrared spectroscopy; NIR based metabolomics; cheeses; untargeted metabolomics; saturated fatty acid (SFA).

Resumen. El presente describe un flujo de trabajo para realizar análisis estadísticos multivariados (MSA) no supervisados por análisis del componente principal (PCA) y supervisados por análisis discriminante por mínimos cuadrados parciales (PLS-DA), para analizar matrices de datos obtenidos por infrarrojo cercano (NIR) de quesos de diversos tipos y orígenes geográficos, con respecto a sus perfiles NIR de ácidos grasos saturados. El conjunto de datos incluye (A) espectros NIR adquiridos en modo absorbancia, (B) espectros NIR post-procesados por primera derivada y (C) espectros NIR post-procesados por primera derivada y con frecuencias seleccionadas, dentro de un intervalo de longitud de onda entre 12500-3600 cm^{-1} . La entrada de datos NIR fue adaptada por primera vez a un formato legible a la plataforma por internet de análisis metabolómico “MetaboAnalyst”, convirtiendo el formato de datos espectrofotométricos sin procesar, al formato IUPAC JCAMP-DX estandarizado para intercambio de datos espectrales, transformados posteriormente hacia un formato de valores separados por comas editable (.csv) apropiado para estudios metabolómicos con

MetaboAnalyst. Las regiones discriminantes para la primera matriz de datos NIR son cinco. Para la segunda matriz, las regiones de número de onda discriminantes se reducen a tres: 10000 a 8000 cm^{-1} (sobretono -CH-), 6000 a 5000 cm^{-1} (sobretono -C=O-) y 5000 a 4000 cm^{-1} (banda -CH-). Finalmente, para la tercer matriz NIR, se tomaron regiones discriminantes refinadas: 9700 a 8265 (sobretono -CH-), 6661 a 4655 cm^{-1} (sobretono -C=O-) y de 4327 a 4000 cm^{-1} (banda -CH-). El modelo PLS-DA obtenido de la matriz de datos de barrido de infrarrojo cercano post-procesados por primera derivada y con frecuencias seleccionadas muestra la mejor clasificación entre los lácteos y los estándares de ácidos grasos saturados. Estos resultados pretenden introducir un método para realizar metabolómica basada en NIR no dirigida y cuantitativa dentro de una plataforma con más de 300000 usuarios al momento.

Palabras clave: Espectroscopía por infrarrojo cercano; metabolómica basada en infrarrojo cercano; quesos; metabolómica no dirigida; ácidos grasos saturados.

Introduction

Metabolomics can be first divided into targeted and untargeted metabolomics. The choice of one of them depends on the objectives of the research, if it is desired to respectively have either the identification and/or quantification of specific compounds or obtaining representative holistic fingerprints, constructed from a metabolic signature of samples defined as complex matrixes, that provide in turn an overview of the expressed metabolites within a system [1]. The most commonly high resolution techniques used to obtain data for metabolomics studies are hyphenated chromatography with mass spectrometry and nuclear magnetic resonance spectroscopy [2]. The metabolomics workflow consists in obtaining a data matrix from an instrumental measurement, that after data processing (such as baseline corrections, both frequencies' alignments and referencing as well as spectroscopic, spectrometric, or chromatographic binning), is suitable to multivariate statistical analysis (MSA). Untargeted metabolomics in food matrices comprises the obtention of holistic chemical foot and/or fingerprints related to their geographical origin, variety, food quality, manufacturing processes, impacts due to external factors such as climate change, counterfeits, amongst others [1,3,4].

Multivariate analyses used in metabolomics studies are broadly based on Principal Component Analysis (PCA) [5] and Partial Least Squares Discriminant Analysis (PLS-DA) [6]. Principal component analysis is an unsupervised technique to produce decreased variable models with maximum variance [7,8], separating classes according to the weight of resulting loadings, wherein higher loading scores have a greater contribution to the separation [9]. In contrast, the supervised Partial Least Squares–Discriminant Analysis (PLS-DA) extracts the information that can predict all possible class memberships from linear combinations of original input data matrix with the use of multivariate regression techniques, whereas class discriminations are assessed by a permutation test between the original data and the permuted class labels via cross-validations [10, 4]. Finally, in terms of quantitative and qualitative metabolomics, the later can be subdivided into unsupervised and supervised pattern recognition methods, whereas supervised methods such as PLS-DA use trained algorithms for classifying samples from data inputs, into predefined groups. The supervised pattern recognition models also reveal variables related to separation amongst groups and how groups behave per analyzed discriminant factor [11]. Typically, cross-validation resampling methods are used in supervised pattern recognition algorithms for evaluating the predictive capacity of a trained independent data set, against new data with an optimum number of factors, by also flagging overestimations and/or biases.

According to the Clarivate Web of Science database, more than 200 reports have been published the use of Near Infrared (NIR) spectroscopy for cheese analysis. These analyses, mainly driven as targeted strategies for identifying and quantifying specific treats in said dairy product, include [12]:

- Gross composition: total weight percentage of fat, protein, salt, pH, Total Nitrogen (TN, in mg/g or g/100g cheese), water soluble nitrogen (g/100g cheese), amino nitrogen with respect to TN
- palm oil content (%wt/wt)
- total antioxidant capacity (in μmol of Trolox / mg of cheese)
- cholesterol (g/ 100g of cheese).
- % Volatile compounds: acetaldehyde, ethanol, 1-propanol, i-propanol, n-propanol, 2- butanol, 2-pentanol, 3-methyl-1-butanol, 2-butanone, 2-pentanone, 2-heptanone, 2-nonanone, and acetone.

- Organic acids: acetoin (mmol/kg), acetic acid (mmol/kg), butyric acid (mmol/kg), pyruvic acid (g/kg), succinic acid (g/kg) and lactic acid (g/kg).
- Free amino acid content (nmol/g) that are responsible of cheeses' taste and also serving as ripening biomarkers.
- Quality traits: Cheeses' appearance, consistency and flavor.
- Descriptive sensory analysis: cheeses' pressure and shear firmness, odor intensity, elasticity, cohesion, pastiness, solubility, dryness, floury, grainy, flavor intensity, aromaticity, maltiness, sweetness, acidity, pungency and bitterness.

However, less than a tenth of said publications relate to the use of NIR based metabolomics or chemometrics for chesses. Most of said reports are coming from 15 countries, whereas at least half of them are coming from 3 European countries (Italy, France and Spain), and in turn being China the most active non-European country that contribute to cheeses' NIR based metabolomics. Up to date, to the best of our knowledge, no report exists regarding the use of NIR based metabolomics for studying Mexican cheeses' assessments such as above-mentioned treats or models to describe geographical origin, quality, authenticity and/or counterfeiting.

NIR metabolomics approaches for cheeses include a model for fingerprinting ageing processes and selected sensory parameters in Cheddars with reflectance NIR coupled with partial-least squares (PLS) multivariate regression of raw, derivatized and scatter-corrected NIR data matrix [13], the use of PCA and modified PLS [14] with cross-validation of raw NIR reflectance data matrix to evaluate diverse visual, taste, texture, flavors and odor attributes in Spanish cheeses [15], as well as a combined NIR Diffuse Reflection with Mid-infrared attenuated total reflection data matrix, treated with PCA and linear discriminant analysis (LDA) as chemometric model to discriminate Swiss, German, French (Bretagne and Savoie), Austrian and Finnish Emmental cheese. To the best of our knowledge, most of the herein mentioned references, do not extensively discuss the details for constructing discriminant infrared data matrices for (un)-supervised multivariate statistical analysis with pre- and post-processing outputs constructed in universal formats such as the Joint Committee on Atomic and Molecular Physical Data IUPAC standard format family for spectral data exchange, known as JCAMP-DX.

The presence of saturated (SFA) or unsaturated fatty acids (UFA) within the raw milk will affect the texture of any produced cheese, whereas softer cheeses are related to a higher degree of unsaturation in fatty acids. In contrast, harder cheeses related to a major medium-sized SFA content are also associated with increased cardiovascular, obesity and some cancer risks, mostly due to the presence of C12:0 (lauric), C14:0 (myristic) and C16:0 (palmitic) fatty acids, claimed as dangerous to human health in high contents [16]. Most common fatty acid in cheeses are C10:0 (capric acid), C14:0, C16:0, C18:0 (stearic acid), and C18:1 *cis* (oleic acid), whereas 60 to 70 % of total fatty acid content in ruminant milk is saturated, 20-30% correspond to monounsaturated fatty acids, and palmitic and oleic acids are the most abundant SFA and UFA, respectively in said dairy matrixes [17,18]. Furthermore, there are differences in medium chain SFA content in cow's, goat's and ewe's milk, whereas goat's and ewe's milk present higher contents of mostly C12:0, C10:0 and C8:0, with respect cow's milk, and thus its ratios in milk are measured as a counterfeit analytical test in goat products adulterated with cow's dairy source [18]. Finally, as it has been demonstrated in previous works [19], cheeses' fatty acid profiles analyzed with NIR based chemometrics serve as a fingerprint for discriminating seasonal origin (winter or summer seasonality), whereas observed specific fatty acid contents varies along ruminants seasonal-dependent feeding regimen, that directly affects milk's fatty acid profile [19,20].

One of the most widely used techniques for the identification of fatty acids in dairy products is gas chromatography. However, its implementation carries on some challenges such as a proper selection of the derivatization method [21], as well as the optimization of classical chromatographic parameters related to the stationary phase such as selectivity for an accurate separation of relevant fatty acids [22], amongst others. In consequence, methods such as Near-Infrared (NIR) spectroscopy represent an excellent alternative in terms of easy implementation, as it does not require exhaustive or even null sample preparation, and furthermore, it is not invasive. With the NIR technique, it is straightforward to identify and even partially quantify saturated fatty acids which are present in cheeses in important amounts, in noticeable contrast to PUFAs that are more difficult to detect [23] with a non-invasive NIR analysis.

Present work introduces a metabolomics approach to construct untargeted fatty acid profiles in a set of Mexican, American, and Italian regional cheeses, with three different data matrixes: i) NIR raw spectra, (ii) first derivative NIR span and (iii) first derivative-selected NIR spans absorbance data inputs, treated with unsupervised PCA and supervised PLS-DA algorithms, trained to compare cheeses' fatty acid profiles with a

set of seven liquid- and solid-state fatty acid standards. Present model details the procedure to obtain three NIR absorbance data inputs, whereas the NIR acquisition raw data was exported from a local instrument format (Bruker OPUS: “.0” File format), into a universal JCAMP-DX format [<https://www.cheminfo.org/Chemistry/Cheminformatics/JcampConverter/index.html>], that allows to produce an editable comma-separated values (.csv) format of NIR matrixes, legible to be submitted to the stream-lined MetaboAnalyst 5.0 user-friendly multivariate statistical analysis platform, as an alternative way to obtain free access NIR metabolomics holistic fingerprints, avoiding the need of imperatively having costly metabolomics software.

Materials and methods

Cheeses and materials

A total of twelve cheeses' samples were purchased from different local markets, nine of which are Mexican artisan cheeses, two cheeses produced in the United States of America, and one produced in Italy. Seven artisanal samples from six Mexican geographical origins (Chiapas, San Luis Potosí, Oaxaca, Hidalgo, Jalisco and Estado de México) were obtained from Mercado de San Juan 19°25'48" N, 99°8'40. 92" W, located in Mexico City:

- I. Q₁= Ocosingo Ball Cheese (Chiapas, Mexico), analyzing the composition of the crust (Q_{1A}) and the cheese's core (Q_{1B}).
- II. Two cheeses with the local denomination “Queso Crema de Chiapas”, respectively: Q₂= Santa Cruz® and Q₃= Vaquero® (Chiapas, México).
- III. A cheese from San Luis Potosí, Mexico denominated and herein mentioned as Q₄= Adobera cheese.
- IV. Q₅= Quesillo de Oaxaca is an artisanal cheese with an origin from Oaxaca, Mexico.
- V. A cheese from the geographical origin Hidalgo, México, herein tagged as Q₆= Oaxaca-type cheese.
- VI. Q₇= Cotija cheese, from “El Mesón del queso Cotija ®” (Jalisco, Mexico)
- VII. Zacazonapan cheese with Q₈=15 days ripening, and Q₉=30 days of ripening (Estado de México).

Cheeses produced in the United States of America were Q₁₀=Cheddar cheese, Tillamook® and Q₁₁= Camembert cheese, Président®. Finally, the Italian cheese provided by Kirkland® was herein tagged as Q₁₂= Grana Padano cheese.

Seven fatty acid standards were purchased from Sigma Aldrich® (Steinheim, Germany) S₁=butyric acid (C4:0; CAS No. 107-92-6), S₂=hexanoic acid (C6:0; CAS No. 142-62-1), S₃=octanoic acid (C8: 0; CAS No. 124-07-2), S₄=decanoic acid (C10:0; CAS No. 334-48-5), S₅=dodecanoic acid (C12:0; CAS No. 143-07-7), S₆=myristic acid (C14:0; CAS No. 544-63-8), S₇=palmitic acid (C16:0; CAS No. 57-10-3).

Near-Infrared absorbance acquisition details

In all cases, one gram of grated cheese was placed in polystyrene integrating sphere sample rotator cups, adapted to the spectrophotometer for maximizing the interaction of the electromagnetic radiation with the inherent cheeses' sample heterogeneity. NIR absorbance spans (NIR data matrix A, Figures 1 to 4) were carried out with a Multipurpose analyzer Bruker Optics spectrophotometer (Rosenheim, Germany) scanning wavelengths between 830 and 2500 nm (wavenumbers between 12000 to 4000 cm⁻¹, respectively), spanning the radiation with a 2.5 nm optical pathlength. Acquisition routines were performed for all samples in absorbance mode with 64 scans for both blanks and sample collections. All acquisitions and data export to JCAMP-DX universal format were carried out using the OPUS 7.8 program (Bruker Optics, Rosenheim, Germany). Subsequently, all data were converted to a comma separated value format (CSV), using the JCAMP-DX to CSV file converter script found at the following link: <https://www.cheminfo.org/Chemistry/Cheminformatics/JcampConverter/index.html> JCAMP-DX converter scripts' outputs were extracted and pasted into Excel spreadsheets as a CSV format. All liquid-state standards' NIR spectra (butyric acid, hexanoic acid, octanoic acid and decanoic acid) were baseline adjusted with respect the solid-state cheeses' NIR absorbances' spans using a standard correction factor. The first derivative NIR Absorbance data matrixes (NIR data matrix B, Fig. 2) were obtained from Raw data with Microsoft Excel. The first derivative selected NIR

spans absorbance data inputs (NIR data Matrix C, Fig. 3) were obtained from NIR data matrix B, by zeroing all NIR frequency regions that were identified as non-relevant inputs to be considered for MSA. Individual CSV files were arranged first in a two-variables input format (mz / into) suitable for MetaboAnalyst 5.0 software. Each NIR raw data and first derivative triplicate were saved in folders defined by its discriminant factor (type of cheese / type of standard). All variable sets were in turn arranged in a proper .zip format as standard Metaboanalyst 5.0 Statistical analysis inputs.

Statistical analysis

Multivariate statistical analysis of raw and first derivative NIR data matrices were performed using unsupervised principal component analysis and Partial Least Squares Discriminant Analysis (PLS-DA), with the software MetaboAnalyst 5.0. Data pre-processing comprising normalization by sum (for adjust differences amongst samples), transformation (Log) and autoscaling (mean centering divided by standard deviation of each variable) were applied to remove any possible variation during experimental phase, in order to make features as comparable as possible [8,24-25]. PLS-DA model validations were done with 100 permutations per analysis. Reliability of each classification per model was evaluated in terms of goodness of fit (R^2) and goodness of prediction (Q^2). The T2 Hotelling's regions depicted by ellipses in score plots of each model define a 95 % confidence interval.

Results and discussion

Table 1. Resumes the expected NIR fingerprints for fatty acids in cheeses, according to recent literature.

Vibration mode	Wavenumber range (cm ⁻¹)	Wavelength range (nm)	Reference
C-H	8331, 7140, 5712, 4327, 4273	1200, 1400, 1750, 2310, 2340	[25]
C-H fat's vibrations	4295-4805	2030-2080	[26]
C-H combination bands	4327- 4273	2310-2340	[25]
C-H first overtone	5812-5681	1720-1760	[25]
C-H saturated acids	5681	1760	[26]
CH ₂ second overtone	8262	1210	[27]
=C-H (cis)	5951, 4651-4563	1680, 2150-2190	[28]
=C-H (C18:1)	5797	1725	[29]
C=O (stretching) acids & esters	5666-5778	1765-1730	[30]

Fig. 1 shows the Near Infrared raw data of twelve cheeses Q₁ to Q₁₂ (Fig. 1(B), see Materials and methods, section 2.1) made from cow's milk, as well as of the seven SFA standards S₁ to S₇ (Fig. 1(A), see Materials and methods, section 2.1), collected with the same acquisition parameters (Materials and methods, section 2.2), whereas the most representative NIR regions for cheeses' analysis obtained from literature [24-29] are highlighted within the stacked NIR plots. The signals correspond to fat (8000 to 9000, 5400 to 6000, and 4000 to 4500 cm⁻¹) and moisture (6000 to 8000 and 4500 to 5400) cm⁻¹, in agreement with previous reported NIR data [19,23,25-31] from cheeses made from different ruminants (cow, ewe and goat milk), seasons (summer and winter) and ripening times (0 to 6 months).

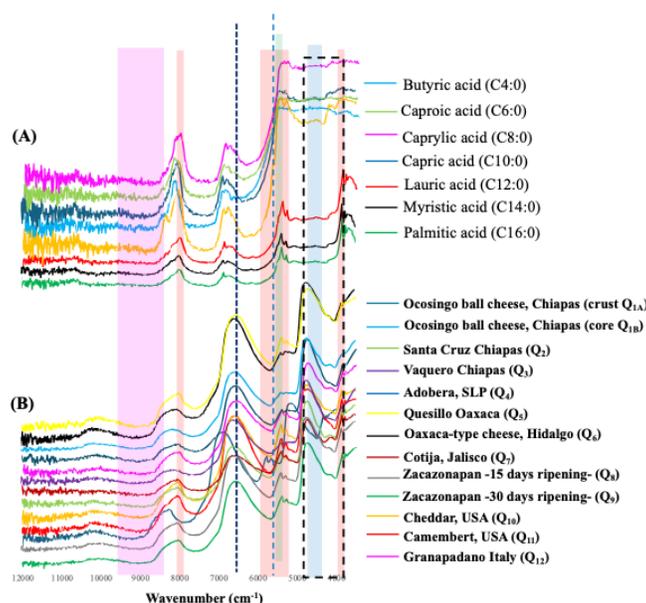


Fig. 1. (A) Near infrared spectra of S₁ to S₇ fatty acid standards. (B) Near infrared spectra of Q₁ to Q₁₂ cheeses.

Curto et al., [32] identified four NIR signals (8264.46, 6896.55, 5780.35, 5181.35 cm⁻¹) relevant to discriminate processing, seasonality (winter and summer), and type of formulation (0 to 100 % of raw milk from cows, ewes and goats) in Spanish cheeses.

In the present study, a wavelength range of 4000 to 10000 cm⁻¹ was used, which allowed the identification of a large number of signals, such as a previous study reported by Bittante et al [33].

Fig. 1(A) shows within the NIR plots elaborated from the raw data matrix, three specific narrow signals associated to (C4:0 to C16:0) SFA, observed for the seven SFA standards:

- 8262 cm⁻¹ (-CH₂ overtone),
- 6665 (-OH- stretching),
- 5666-5778 C=O in fatty acids.

In parallel, for all analyzed Q₁ to Q₁₂ cheeses, Fig. 1(B) shows signals from:

- 4000 to 4500 cm⁻¹
- and 5500 to 6100 cm⁻¹

respectively associated to -CH- vibration and C=O vibration in combination with the -CH- first overtone groups

Fig. 2 shows the NIR data matrix A of Q₁ to Q₁₂ cheeses, with their respective acquisition replicates used as data inputs for PCA and PLS-DA analyses. Identical NIR wavelength ranges were used for postprocessing NIR Data Matrix A into NIR data matrix B (Fig. 3) and NIR data matrix C (Fig. 4).

The resolution of NIR data matrix A raw spectra was improved through a first derivative span postprocessing, albeit the intrinsic loss of signal-to noise ratio [34]. For dairy products, Lobos-Ortega et al., [23] report different data treatments such as applying the second derivative to generate a model that discriminates cheeses of different species (cow, ewe and goat milk) based on their polyunsaturated fatty acid composition. On the other hand, in the model described by Zhao et al., [35], designed to detect fatty acids in cow's milk using the MIR technique, five data pre-processing algorithms were used, revealing the advantages of using first and second derivative postprocessing routines, that produce R² predicting reliability.

Furthermore, Pereira et al., used first and second derivative MIR postprocessing routines as models capable to discriminate butter made with milk's fat, with respect adulterated butter with soybean fat [36], reporting discriminant efficiency between authentic and counterfeited butter by means of the proposed fatty acid model, with outstanding root mean squared (RMSE), relative prediction errors (RE %) and (R²) predictive coefficients values, when MIR spectra were postprocessed with first and second derivative routines.

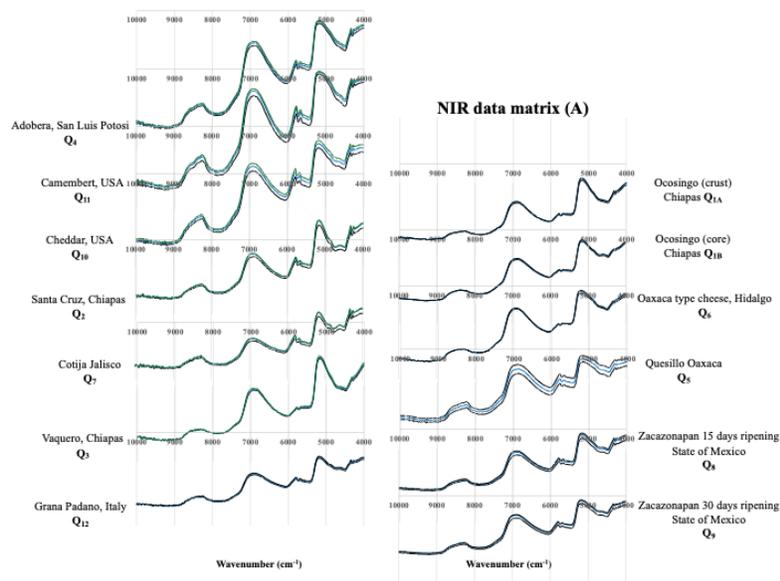


Fig. 2. NIR absorbance raw spectra acquired by triplicate, of the full set of analyzed Q₁ to Q₁₂ cheeses, referred as “NIR data matrix A” (see materials and methods).

For the present study the application of a first derivative routine for producing NIR data Matrix B from NIR data matrix A, three main discriminant regions are observed:

- 8000 to 10000 (-CH- overtone),
- 5000 to 6000 (C=O vibration mode) and

-4000 to 5000 cm⁻¹ (-CH- vibration), as highlighted in Fig. 3. The signals coincide with the three regions reported by Ayraz et al., for fatty acids in Ezine cheese [31].

In terms of acquisition mode as a function of sample type, NIR data of both grated Q₁-Q₁₂ cheeses (Figures 1-4), liquid-state S₁ to S₄ standards and solid-state S₅ to S₇ standards (Figures 1 and 5), were carried out in absorbance mode, finding accurate agreements with equivalent dairy's NIR spectra acquired at reflectance mode [37-38] likewise previous reports describing NIR reflectance acquisition modes for grated fresh freeze-dried cheeses [39], cheeses' slices [40,16,23], and cheese fat extracts [16] to predict fatty acids in all cases. Other reports have used transmittance-mode data acquisition in ground cheeses, to generate predictive fatty acid models [14]. The transmittance method is typically used for liquid materials or solid systems with thin layers [41,42]. Accuracy of NIR absorbance acquisition mode herein used to analyze the pair of grated Q₁-Q₁₂ cheeses and S₁ to S₇ standards, have found also agreements with other cheeses' NIR strategies using reflectance-transmittance [42] or transflectance [43] acquisition modes.

To the best of our knowledge, few reports describe easy and straightforward methods for variable wavelength selection (“data binning” [44]) of near-infrared spectra for performed multivariate analysis. A simple alternative for NIR data binning is herein presented:

1. Relevant frequencies of NIR spectra are left as in the raw input.
2. Those frequencies wanted to not be considered in post-processed MSA data inputs, are simply withdrawn from the NIR raw data, by zeroing those frequency regions not being considered within a variable wavelength selection of a not zeroed NIR data matrix.

Fig. 4 shows the NIR data matrix C of Q₁ to Q₁₂ cheeses, obtained from NIR data matrix B (Fig. 3), zeroed from non-relevant frequency regions. With a higher spectral resolution, three main frequency ranges are appreciated respectively at:

- 9700 - 8265 cm⁻¹
- 6661 - 4655 cm⁻¹, and
- 4327 - 4003 cm⁻¹

Those NIR wavenumber ranges correspond to respectively fatty acids overtones and main vibration modes, in full agreement to previous reports using NIR spectroscopy to identify medium chain C6:0 to C16:0 SFA -mainly palmitic acid in cow products from the 5780.3 – 5665.7 cm^{-1} -C=O- carboxylic vibration mode [16,45], a fatty acid profile model for distinguishing geographical origins of Coalho cheeses, mainly from the 5882.3 – 5404.4 cm^{-1} vibrations modes [46], -CH- overtones respectively at 8403.3 and 5847.9 cm^{-1} from a Ricotta cheese study [24] and a model applied in Abondance and Tomme de Savoie cheeses finding =C(sp²)-H and -OH- key vibration modes at respectively 4664.2 cm^{-1} and 6666.6 cm^{-1} [39].

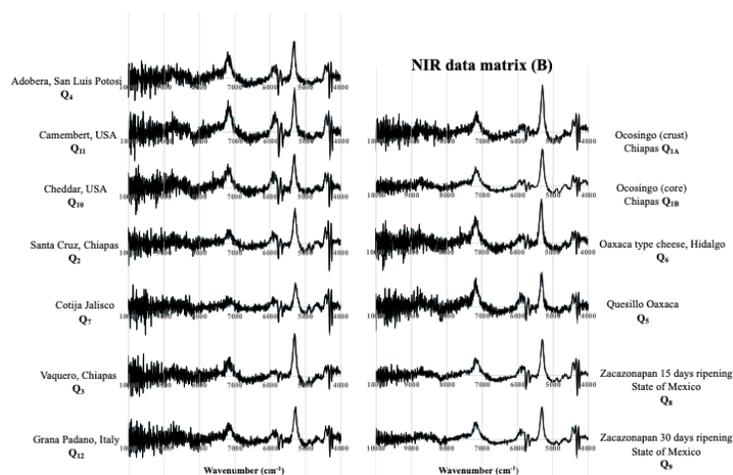


Fig. 3. First derivative NIR absorbance post-processed spectra acquired by triplicate, of the full set of analyzed cheeses, referred as “NIR data matrix B” (see materials and methods).

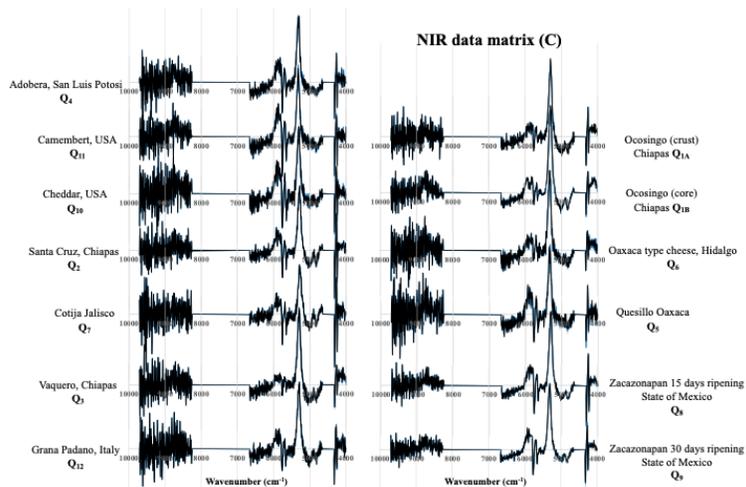


Fig. 4. First derivative-selected NIR spans absorbance spectra acquired by triplicate, of the full set of analyzed cheeses, referred as “NIR data matrix C” (see materials and methods).

Fig. 5 present NIR data matrix A, B and C of seven fatty acid standards S₁=butyric acid (C4:0), S₂=hexanoic acid (C6:0), S₃=octanoic acid (C8:0), S₄=decanoic acid (C10:0), S₅=dodecanoic acid (C12:0),

S₆=myristic acid (C14:0), and S₇=palmitic acid (C16:0). Expected relevant frequencies reported elsewhere, associated to SFA vibration modes include [16,45]:

- 4327 and 4273 cm⁻¹ (-CH-first overtone),
- 5492 and 6171 cm⁻¹ (-CH-),
- 5666 to 5797 cm⁻¹ (C=O),
- 6665 cm⁻¹ (-OH-),
- 8262 cm⁻¹ (-CH₂),
- 8532 - 9689 (-CH- second overtone).

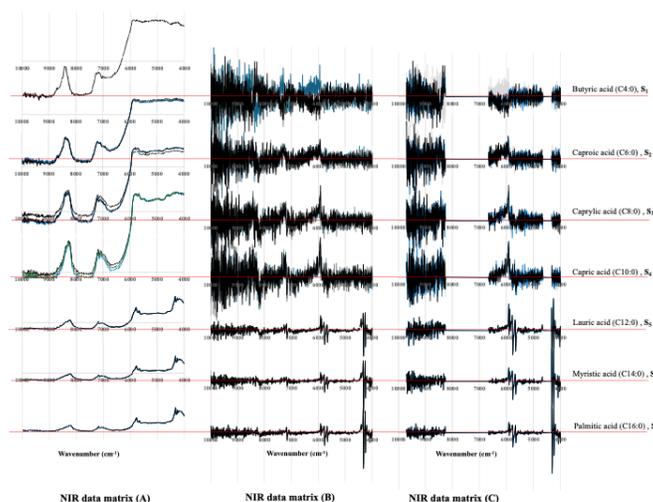


Fig. 5. NIR spectra of S₁ to S₇ fatty acid standards A raw absorbance data. B Post-processed first derivative NIR spans and C Post-processed first derivative selected NIR spans.

Fig. 6 presents the workflow to carry out data arrays of each NIR data matrix A, B and C as inputs for PCA and PLS-DA multivariate statistical analysis done with the stream-lined metabolomics data analysis software *MetaboAnalyst 5.0* (*vide infra*).

First, all NIR acquisitions were exported from the instrument format Bruker OPUS 7.8 program, into a universal Joint Committee on Atomic and Molecular Physical Data (JCAMP-DX) IUPAC standard format. Subsequently, all data were converted to a comma separated value format (CSV), using the JCAMP-DX to CSV file converter script from the cheminformatics department of the Swiss Federal Institute of Technology, available from the following website: <https://www.cheminfo.org/Chemistry/Cheminformatics/JcampConverter/index.html> JCAMP-DX converter scripts' outputs are copied and pasted into Excel spreadsheets as an editable comma-separated values (.csv) format. Individual (.csv) files were arranged first in a two-variables input format (variable "mz" for wavenumbers / variable "into" for absorbances [data matrix A] and Δ Absorbances [data matrix B and data matrix C]) suitable for *MetaboAnalyst 5.0* software. Fig. 6 shows the example of this two-variables array (wavenumber / [Δ] absorbances) plotted for NIR data matrix A (yellow arrow pathway, using Q₂= Santa Cruz® as example within the Fig. 6), NIR data matrix B (red arrow pathway, using Q₄= Adobera, San Luis Potosí cheese as example within the Fig. 6) and NIR data matrix C (green arrow pathway, using Q₁₂= Grana Padano cheese as example within the Fig. 6). Each NIR raw data and first derivative (full and zeroed) triplicates were saved in folders defined by its discriminant factor (Q₁-Q₁₂ cheeses / S₁-S₇ standards). All variable sets were in turn arranged in a proper .zip format as one-factor standard *MetaboAnalyst 5.0* Statistical analysis inputs.

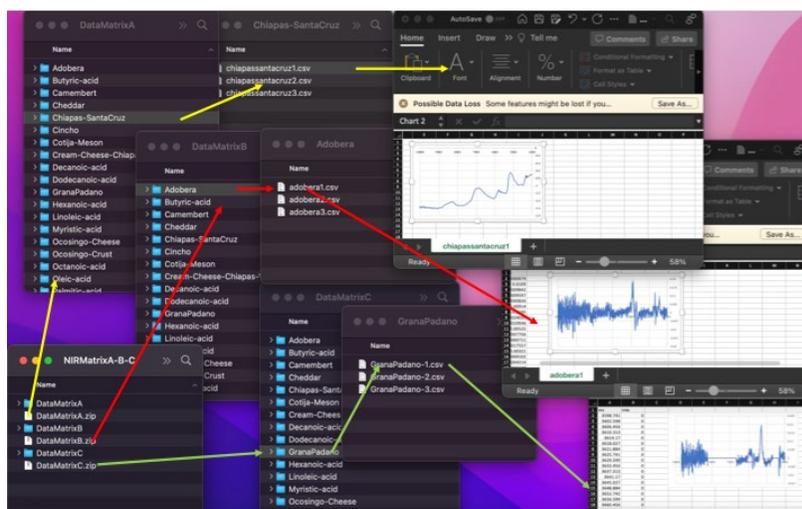


Fig. 6. Workflow to produce NIR data matrixes A (yellow arrows), B (red arrows) and C (green arrows) as inputs for multivariate statistical analysis carried out with the stream-lined metabolomics data analysis software MetaboAnalyst 5.0. (<https://www.metaboanalyst.ca/>).

Fig. 7 depicts unsupervised PCA score (top) and loading (bottom) plots obtained from NIR data matrixes A (left), B (middle) and C (right).

PCA score plot of Data Matrix A present a principal component 1 (PC1) of 80.6 % variance and PC2 of 16.2 % variance, together explaining the 96.8 % of the total variability of the data. PCA score plot of Data Matrix B present a PC1 of 29.2 % variance and PC2 of 13.1 % variance, together explaining the 42.3 % of the total variability of the data. Finally, PCA score plot of Data Matrix C present a principal component 1 (PC1) of 28.4 % variance and PC2 of 12.9 % variance, together explaining the 41.3 % of the total variability of the data. Despite having better variance with NIR data matrix A, first derivative spans (full -NIR data matrix B- and zeroed -NIR data matrix C-) are the inputs that best group the samples, as observed within the loading plots in Fig. 7 bottom, whereas each loading element represent explanatory variables, wherein the more and better distributed they are from the origin, the better is the model to discriminate amongst discriminant factors. This is explained by the fact that loading plots are the multivariate version of scatter plots, thus showing the effect of predictors on variables [47]. The use of the raw data matrix A has limitations, such as the use of few discriminant variables, which may cause an overestimation of the model, as observed in the PCA loading plot produced with the raw NIR absorbance data matrix, shown in Fig. 7 (bottom, yellow dots). On the other hand, the use of a first-derivative NIR absorbance data matrix greatly decreases the variance and has a larger number of discriminant variables (green dots for loadings obtained with PCA from Data Matrix B, magenta dots for loadings obtained with PCA from Data Matrix C, Fig. 7 bottom) and thus a more reliable model for unraveling subtle differences between cheeses and fatty acid standards. Combined PCA and MPLS have been employed to respectively reduce non-relevant NIR signals to produce highly discriminating NIR data matrices and for fatty acid quantification, performing principal components to elucidate metabolomic features related to thawed and fresh cheeses, with acceptable variances of about 72 % to discriminate features related to freshness in cheeses, with only two principal components [16-39].

All NIR data sets produce PCA score plots (Fig. 7, top) defining Q₁-Q₁₂ cheeses at the (PC1, PC2) origin, whilst the S₁- S₇ standards are distributed in different (PC1, PC2) coordinates. The closer the S₁- S₇ coordinates to the Q₁-Q₁₂ origin counterpart, shall suggest the presence of a specific SFA in the cheeses. In all cases, S₅- S₇ C12:0, C14:0 and C16:0 SFA share PCA dimensionality with the full set of Q₁-Q₁₂ cheeses. Data matrix C PCA score plot best represents the PCA equivalence of lauric, myristic and palmitic SFA scores with the full set of analyzed cheeses, strongly suggesting their presence in the dairy products. Myristic (C14:0) and palmitic (C16:0) fatty acids have been reported in fresh cheeses [39] and also in cow's milk which also contains stearic acid (C18:0) [48]. On the other hand, lauric acid is associated with a strong soapy flavor in cheeses [49-

50]. The fact that all cheese samples are clustered near zero in all PCA essays, have also been reported in metabolomics quality control analysis, whereas dairy products have explicitly been mixed with SFA in different ratios [51-52].

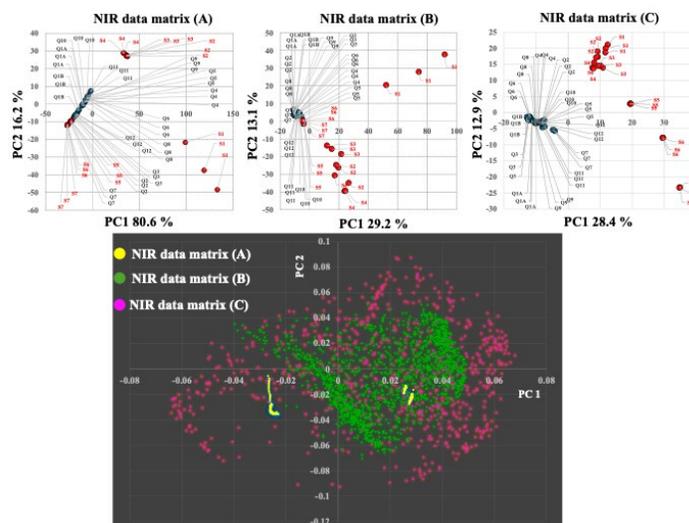


Fig. 7. Unsupervised Principal Component Analysis (PCA) score plots (top) and loading plots between principal components 1 and 2 (bottom) of each multivariate statistical analysis carried out with NIR data matrix A (left), data matrix B (middle) and data matrix C (right) inputs. Cheeses' Q₁-Q₁₂ score plots are highlighted in blue, whilst standards' S₁-S₇ score plots are highlighted in red.

Fig. 8 represents the supervised PLS-DA of the full set of the .csv NIR data matrixes presenting the following variances (%), goodness of prediction (Q^2) and goodness of mathematical fitting (R^2)

- NIR data matrix A PLS-DA outlier: 85.5 % ($Q^2= 0.43$ and $R^2=0.45$)
- NIR data matrix B PLS-DA outlier: 37.6 % ($Q^2= 0.78$ and $R^2=0.82$)
- NIR data matrix C PLS-DA outlier: 37.0 % ($Q^2= 0.83$ and $R^2= 0.9$)

In all cases, the use of 3 PLS-DA components are sufficient to represent a goodness of prediction of respectively $Q^2= 43$ %, 78 %, and 83 % (Fig. 8, bottom). PLS-DA score plot of Data Matrix A presents a principal component 1 (PC1) of 79.6 % variance and PC2 of 5.9 % variance, together explaining the 85.5 % of the total variability of the data. PLS-DA score plot of Data Matrix B presents a PC1 of 25.0 % variance and PC2 of 12.6 % variance, together explaining the 37.6 % of the total variability of the data. Finally, PLS-DA score plot of Data Matrix C presents a principal component 1 (PC1) of 28.2 % variance and PC2 of 8.8 % variance, together explaining the 37.0 % of the total variability of the data.

In comparison to unsupervised PCA, supervised multivariate statistical analysis of NIR data matrixes, produce score plots with equivalent two-component variances and a noticeable variable separation enhancement. PCAs of first derivative NIR spans (Data Matrix B) and first derivative selected NIR spans (Data Matrix C) only produce a pronounced separation between S₁-S₇ standards and Q₁-Q₁₂ cheeses, with respect the PCA score plot of NIR Data Matrix A, being the chesses' scores mostly defined within the PCA origin.

With the most discriminant PLS-DA model obtained with NIR data matrix C, the samples are dispersed along the axes, whereas it is observed a geographical origin separation between Mexican (Q₁-Q₉, with PC2 positive or close to zero) and foreign cheeses (Q₁₀-Q₁₂, negative PC2 component). Mexican cheeses present the same positive PC2 dimensionality as for S₁ to S₅ low chain fatty acid standards (butyric, hexanoic, octanoic, decanoic and dodecanoic acids). In contrast, American and Italian cheeses' PLS-DA score plots analyzed with NIR data matrix C, present the same PC2 negative dimensionality as with S₆ (myristic acid) and S₇ (palmitic acid) SFA standards, strongly suggesting their higher content within their composition.

In terms of the set of Mexican cheeses, Zacazonapan 15-day (Q_8) and 30-days (Q_9) ripening cheeses contain very similar PLS-DA score plots and have therefore no evident discrimination of the ripening time, with the selected multivariate statistical analysis model. Similar trend is observed for the Ocosingo ball cheese obtained from its core (Q_{1B}) and from its crust (Q_{1A}), with no evident separation amongst score plots, strongly suggesting their similar composition, including their SFA profile.

On the other hand, the Oaxaca type cheese (Q_6 , PC1 (-), PC2 (-)) from Hidalgo and its Oaxacan quesillo counterpart (Q_5 , PC1 (-), PC2 (+)) present significant differences in the PLS-DA score plots, with the multivariate model obtained from NIR data matrix C, whereas these differences might be considered as a geographical origin fingerprint, whereas climatic, environmental conditions and livestock feeding are significantly unique from each region. Furthermore, both analyzed Chiapas cheeses' brands (Q_2 Santa Cruz®, PC1 (-), PC2 (-)) and Q_3 Vaquero®, PC1 (-), PC2 (+)) present slight differences in their score plots. The comparative model with S_1 - S_7 SFA standards might suggest that this set of differences are due to the fatty acid content in each analyzed dairy product. Such as with American and Italian samples with an equivalent negative PC2 dimensionality with C14:0 (S_6 , myristic acid) and C16:0 (S_7 , palmitic acid), Mexican cheeses presenting same trend are Q_{1A} / Q_{1B} (Ocosingo, Chiapas), Q_2 (Queso Crema de Chiapas), Q_4 (Adobera, San Luis Potosí) and Q_6 (Hidalgo).

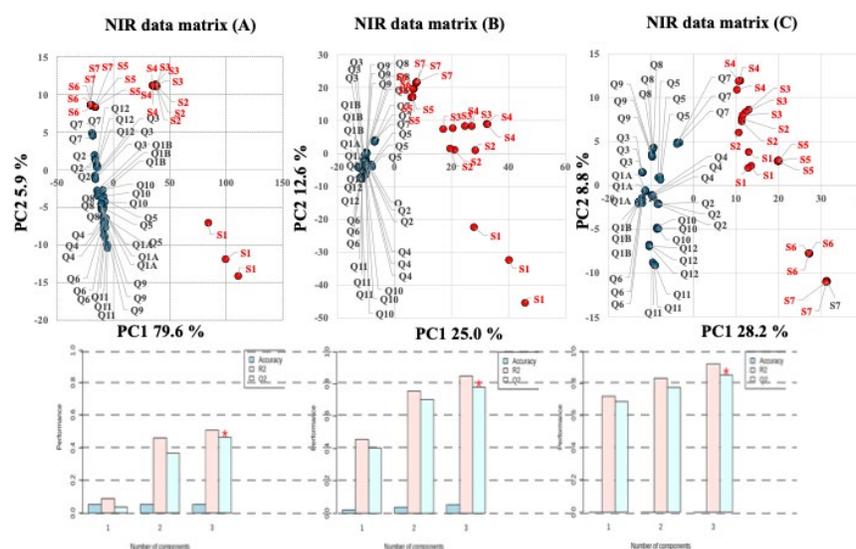


Fig. 8. Partial Least Squares Discriminant Analysis (PLS-DA) score plots (top) and histogram representation (bottom) of the goodness of the MSA fit (R^2) and goodness of prediction (Q^2) of each multivariate statistical analysis carried out with NIR data matrix A (left), data matrix B (middle) and data matrix C (right) inputs. Cheeses' Q_1 - Q_{12} score plots are highlighted in blue, whilst standards' S_1 - S_7 score plots are highlighted in red. Both goodness of prediction (Q^2 , turquoise histograms) and goodness of mathematical fitting (R^2 , pink histograms) were obtained with only three main PLS-DA components, whereas the contribution of each component is highlighted as 1, 2, and 3 sets of histograms.

Data validation by means of hierarchical clustering of NIR raw data matrix A, first derivative NIR data matrix B and first derivative selected NIR spans data matrix C (frequency binning strategy of regions between 9700 to 8265 ($-CH-$ overtone), 6661 to 4655 cm^{-1} ($-C=O-$ overtone) and from 4327 to 4000 cm^{-1} ($-CH-$), shown as dendrograms are depicted in Fig. 9. Data sampling of both Cheeses (Q_1 - Q_{12}) and SFA standards (S_1 - S_7) with NIR data matrix C is the only one presenting two different sets of well separated clusters: S_1 to S_9 SFA standards and Q_1 to Q_{12} cheeses, thus indicating the accuracy of data sampling with the NIR frequency binning strategy herein presented, with NIR data matrix C. In consequence, hierarchical cluster analysis is carried out exclusively from dendrogram obtained with NIR data matrix C (extreme right, Fig. 9).

For SFA standards' clusters, three main sets are observed:

- I. low molecular weight S_1 (C4:0, butyric), S_2 (C6:0, hexanoic), S_3 (C8:0, octanoic) and S_4 (C10:0, decanoic) fatty acids
 - II. Medium S_5 (C12:0, lauric acid) SFA
 - III. Higher molecular weight S_6 (C14:0, myristic) and S_7 (C16:0, palmitic) acids.
- For cheeses' clusters, four main sets are observed:
- I. Q_{11} (Camembert, United States), Q_4 (Adobera, San Luis Potosí, Mexico), Q_5 (Quesillo artesanal cheese, Oaxaca, Mexico) and Q_{12} (Grana Padano, Italy)
 - II. Q_8 (Zacazonapan 15 days ripening, Estado de Mexico), Q_9 (Zacazonapan 30 days ripening, Estado de Mexico), Q_{1A} (Ocosingo crust ball cheese, Chiapas, Mexico) Q_{1B} (Ocosingo core ball cheese, Chiapas, Mexico), Q_2 (Santa Cruz® “queso crema”, Chiapas, México) and Q_3 (Vaquero® “queso crema”, Chiapas, México)
 - III. Q_6 (Oaxaca type cheese, Hidalgo, Mexico) and Q_{10} (cheddar, Tillamook®, United States)
 - IV. Q_7 (Cotija cheese, Jalisco, Mexico)

This NIR data matrix C hierarchical clustering correlates with previous PLS-DA score plot observations, such as the classification of low (S_1 - S_4), medium (S_5) and higher (S_6 - S_7) SFA categories, Mexican cheeses (Q_1 - Q_9) from their American and Italian counterparts (Q_{10} - Q_{12} , except from the lack of hierarchy difference between Q_6 and Q_{10}), cheeses from same type but different geographical origin (Q_5 / Q_6) and the lack of discrimination between samples from the same origin but different ripening processes (Q_8 - Q_9) or chesses' surfaces (Q_{1A} - Q_{1B}). Previous studies using dendrogram clustering include a model to trace cheeses' maturation time by identifying fatty acids profiles in cottage, Dutch, Swiss, blue, and Italian type cheeses in dendrogram clustering through a criterion based on grouping by Euclidean distances and agglomerating by the Ward method [53-55].

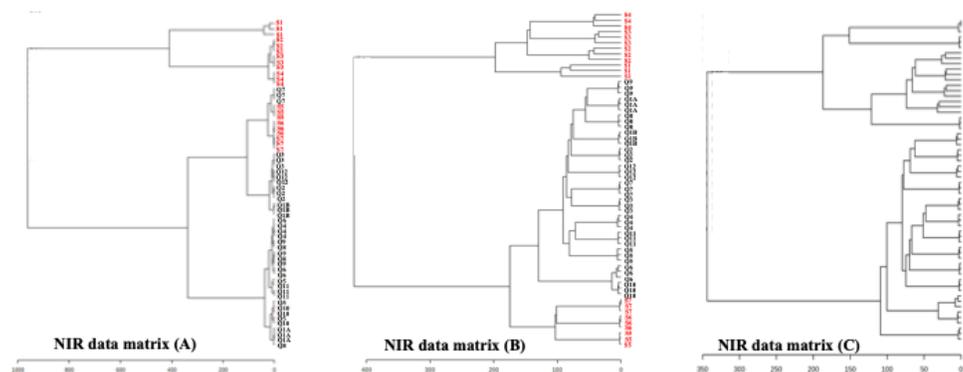


Fig. 9. Hierarchical clustering of NIR raw data matrix A (extreme left), first derivative NIR data matrix B (middle) and first derivative selected NIR spans data matrix C (extreme right), shown as dendrograms, using Euclidian distances and clustering algorithms for distance measurements amongst cheeses (Q_1 - Q_{12} , highlighted with black tags) and SFA standards (S_1 - S_7 , highlighted with red tags).

Conclusions

The use of three different near infrared data matrixes for (un)-supervised multivariate statistical analysis with the stream-lined metabolomics data tool Metaboanalyst -with more than 300,000 users to date- is discussed. Data sampling comprises 19 triplicates of NIR absorbance spectra from nine Mexican, two American and one Italian chesses with seven SFA standards. Absorbance NIR data inputs triad are: i) acquired 12500 to 3600 cm^{-1} spans, ii) first derivative NIR spans of the full acquired frequency range and iii) first derivative selected NIR spans by zeroing all frequency ranges except from 9700 to 8265 cm^{-1} , 6661 to 4655 cm^{-1} and from 4327 to 4000 cm^{-1} as a NIR wavelength binning strategy. All NIR spectra from data matrix A were exported from the instrument format, into a universal JCAMP-DX IUPAC standard format. Subsequently, all data were converted to a comma separated value format (CSV), using the JCAMP-DX to CSV file converter script. The

conversion outputs are copied and pasted into Excel spreadsheets as an editable comma-separated values (.csv) format. Individual (.csv) files were arranged in a two-variables input suitable for both NIR postprocessing with Microsoft Excel to produce NIR data matrixes B and C and for MSA carried out with MetaboAnalyst 5.0 software. NIR data matrix C was the best input for analyzing correlations between cheeses and SFA standards, whereas its supervised PLS discriminant analysis provided the most reliable goodness of mathematical fit ($R^2 = 0.9$) and model prediction ($Q^2 = 0.83$) values.

None of the NIR PLS-DA models were able to discriminate between specific discriminant factors of samples from the same origin such as ripening (Q_8 / Q_9) or cheeses' surface such as core or crust (Q_{1A} / Q_{1B}), but the best selected NIR data matrix C model accurately differentiate between Mexican, American and Italian samples, cheeses from same type but different geographical origin (Q_5 / Q_6), and cheeses from same type and geographical origin but different brand (Q_2 / Q_3), the last also confirmed with hierarchical cluster analysis. The comparative model between Q_1 - Q_{12} cheeses with S_1 - S_7 SFA standards might suggest that the observed set of differences are due to the fatty acid content in each analyzed dairy product. These results overall suggest that a robust cheeses' NIR profiling of specific discriminant factors such as geographical origin, type of cheese, texture (soft / hard), source of raw milk, seasonal origin, gross composition, fatty acids', organic acids' and volatile compounds' profiles as well as sensory and quality traits, shall be more and more possible by means of i) reducing the heterogeneity of the samples (grating or milling with respect analyzing intact samples), ii) selecting an appropriate absorbance – transmittance or reflectance acquisition scheme with a considerable signal-to-noise ratio and iii) considering the use of a NIR wavelength binning post-processing strategy, as herein presented, applied to the highest possible data set.

Acknowledgments

All authors acknowledge the Dirección General de Investigación y Posgrado of Universidad Autónoma chapingo” for all provided scientific support B.N.O.-M. express gratitude to CONAHCYT-México for the doctoral scholarship No. 2020-000013-01NACF-03858). J.E.H.-P thanks financial support from Instituto Politécnico Nacional (IPN) program “Programa Institucional de Contratación de Personal Académico de Excelencia (PICPAE)” and “Estímulos al Desempeño de los Investigadores (EDI)” financial support grant No. F-00318, and the “Secretaría de Investigación y Posgrado (SIP) research grants No. 20231944 and 20241672.

References

1. Cubero-Leon, E.; Peñalver, R.; Maquet, A. *Food Res. Int.* **2014**, 60, 95–107.
2. Marshall, D. D.; Powers, R.; *Prog. Nucl. Magn. Reson. Spectrosc.* **2017**, 100, 1–16.
3. Beleggia, R.; Platani, C.; Papa, R.; Di Chio, A.; Barros, E.; Mashaba, C.; Wirth, J.; Fammartino, A.; Sautter, C.; Conner, S.; Rauscher, J.; Stewart, D.; Cattivelli, L.; *J. Agric. Food Chem.* **2011**, 59, 9366–9377.
4. Herbert-Pucheta, J. E.; Lozada-Ramírez, J. D.; Ortega-Regules, A. E.; Hernández, L. R.; Anaya de Parrodi, C.; *Molecules.* **2021**, 26, 4146.
5. Saccenti, E.; Hoefsloot, H. C. J.; Smilde, A. K.; Westerhuis, J. A.; Hendriks, M. M.; *Metabolomics* **2014**, 10, 361–374.
6. Kalivodová, A.; Hron, K.; Filzmoser, P.; Najdekr, L.; Janečková, H.; Adam, T.; *J. Chemom.* **2015**, 29, 21–28.
7. Gautam, R.; Vanga, S.; Ariese, F.; Umopathy, S.; *EPJ Tech. Instrum.* **2015**, 2, 1–38.
8. Chong, J.; Wishart, D. S.; Xia, J.; *Curr. Protoc. Bioinforma.* **2019**, 68, e86.
9. Worley, B.; Powers, R.; *Curr. Metabolomics.* **2013**, 1, 92–107.
10. Pérez-Enciso, M.; Tenenhaus, M.; *Hum. Genet.* **2003**, 112, 581–592.
11. Rodríguez-Otero, J. L.; Hermida, M.; Centeno, J.; *J. Agric. Food Chem.* **1997**, 45, 2815–2819.
12. De Marchi, M.; Penasa, M.; Zidi, A.; Manuelian, C.L.; *J. Dairy Sci.* **2018**, 101, 10589-10604.

13. Downey, G.; Sheehan, E.; Delahunty, C.; O'Callaghan, D.; Guinee, T.; Howard, V.; *Int. Dairy J.* **2005**, *15*, 701–709.
14. Manuelian, C. L.; Currò, S.; Penasa, M.; Cassandro, M.; De Marchi, M.; *Int. Dairy J.* **2017**, *71*, 107–113.
15. González-Martín, M. I.; Severiano-Pérez, P.; Revilla, I.; Vivar-Quintana, A. M.; Hernández-Hierro, J. M.; González-Pérez, C.; Lobos-Ortega, I. A.; *Food Chem.* **2011**, *127*, 256–263.
16. González-Martín, M. I.; Vivar-Quintana, A. M.; Revilla, I.; Salvador-Esteban, J. T.; *Microchem. J.* **2020**, *156*, 104854.
17. Szumacher-Strabel, M.; Cieślak, A.; Zmora, P.; Pers-Kamczyc, E.; Bielińska, S.; Stanisław, M.; Wójtowski, J.; *J. Sci. Food Agric.* **2011**, *91*, 2031–2037.
18. Markiewicz-Kęszycka, M.; Czyżak-Runowska, G.; Lipińska, P.; Wójtowski, J.; *J. Vet. Res.* **2013**, *57*, 135–139.
19. González-Martín, I.; Hernández-Hierro, J. M.; Salvador-Esteban, J.; González-Pérez, C.; Revilla, I.; Vivar-Quintana, A.; *J. Sci. Food Agric.* **2011**, *91* (6), 1064–1069.
20. Collomb, M.; Bisig, W.; Bütikofer, U.; Sieber, R.; Bregy, M.; Etter, L.; *Dairy Sci. Technol.* **2008**, *88*, 631–647.
21. Hewavitharana, G. G.; Perera, D. N.; Navaratne, S. B.; Wickramasinghe, I.; *Arab. J. Chem.* **2020**, *13*, 6865–6875.
22. Amores, G.; Virto, M.; *Separations* **2019**, *6*, 14.
23. Lobos-Ortega, I.; Hernández-Jiménez, M.; González-Martín, M. I.; Hernández-Hierro, J. M.; Revilla, I.; Vivar-Quintana, A. M.; *Food Anal. Methods* **2021**, *14*, 933–943.
24. Madalozzo, E. S.; Sauer, E.; Nagata, N.; *J. Food Sci. Technol.* **2015**, *52*, 1649–1655.
25. Shenk, J. S.; Westerhaus, M. O.; *Crop Sci.* **1991**, *31*, crops1991.0011183X003100060064x.
26. Šašić, S.; Ozaki, Y.; *Appl. Spectrosc.* **2000**, *54*, 1327–1338.
27. Osborne, B. G.; Fearn, T.; Hindle, P. H., in: *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*; Longman Scientific & Technical, 1993.
28. Garrido-Varo, A.; Carrete, R.; Fernández-Cabanás, V.; *J. Infrared Spectrosc.* **1998**, *6*, 89–95.
29. Hourant, P.; Baeten, V.; Morales, M. T.; Meurens, M.; Aparicio, R.; *Appl. Spectrosc.* **2000**, *54*, 1168–1174.
30. Subramanian, A.; Prabhakar, V.; Rodríguez-Saona, L. *Encycl. Dairy Sci.* **2011**, 115–124.
31. Ayvaz, H.; Mortas, M.; Dogan, M. A.; Atan, M.; Yildiz Tiryaki, G.; Karagul Yuceer, Y.; *J. Food Sci. Technol.* **2021**, *58*, 3981–3992.
32. Curto, B.; Moreno, V.; García-Esteban, J. A.; Blanco, F. J.; González, I.; Vivar, A.; Revilla, I.; *Sensors* **2020**, *20*, 3566.
33. Bittante, G.; Patel, N.; Cecchinato, A.; Berzaghi, P.; *J. Dairy Sci.* **2022**, *105*, 1817–1836.
34. Bou-Orm, N.; AlRomaithi, A. A.; Elrmeithi, M.; Ali, F. M.; Nazzal, Y.; Howari, F. M.; Al Aydarroos, F.; *Planet. Space Sci.* **2020**, *188*, 104957.
35. Zhao, X.; Song, Y.; Zhang, Y.; Cai, G.; Xue, G.; Liu, Y.; Chen, K.; Zhang, F.; Wang, K.; Zhang, M.; Gao, Y.; Sun, D.; Wang, X.; Li, J.; *Molecules* **2023**, *28* (2), 666.
36. Pereira, C. G.; Leite, A. I. N.; Andrade, J.; Bell, M. J. V.; Anjos, V.; *LWT* **2019**, *107*, 1–8.
37. Sørensen, K. M.; van den Berg, F.; Engelsen, S. B. NIR Data Exploration and Regression by Chemometrics—A Primer. In *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*; Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S. B., Eds.; Springer: Singapore, 2021; 127–189.
38. Ikehata, A. NIR Optics and Measurement Methods. In *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*; Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S. B., Eds.; Springer: Singapore, 2021; 211–233.
39. Lucas, A.; Andueza, D.; Ferlay, A.; *Int. Dairy J.* **2008**, *18*, 595–604.
40. Soto-Barajas, M. C.; González-Martín, M. I.; Salvador-Esteban, J.; Hernández-Hierro, J. M.; Moreno-Rodilla, V.; Vivar-Quintana, A. M.; Revilla, I.; Ortega, I. L.; Morón-Sancho, R.; Curto-Diego, B.; *Talanta*. **2013**, *116*, 50–55.

41. Mishra, P.; Roger, J. M.; Rutledge, D. N.; Woltering, E.; *Postharvest Biol. Technol.* **2020**, *168*, 111271.
42. Manuelian, C. L.; Currò, S.; Visentin, G.; Penasa, M.; Cassandro, M.; Dellea, C.; Bernardi, M.; De Marchi, M. T.; *J. Dairy Sci.* **2017**, *100*, 6084–6089.
43. Núñez-Sánchez, N.; Martínez-Marín, A. L.; Polvillo, O.; Fernández-Cabanás, V. M.; Carrizosa, J.; Urrutia, B.; Serradilla, J. M.; *Food Chem.* **2016**, *190*, 244–252.
44. Zhong, L.; Huang, R.; Gao, L.; Yue, J.; Zhao, B.; Nie, L.; Li, L.; Wu, A.; Zhang, K.; Meng, Z.; Cao, G.; Zhang, H.; Zang, H.; *Molecules.* **2023**, *28*, 5672.
45. Ozaki, Y.; Morita, S.; Morisawa, Y. Spectral Analysis in the NIR Spectroscopy. In *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*; Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S. B., Eds.; Springer: Singapore, 2021; 63–82.
46. Silva, L. K. R.; Jesus, J. C.; Onelli, R. R. V.; Conceição, D. G.; Santos, L. S.; Ferrão, S. P. B.; *Int. J. Dairy Technol.* **2021**, *74*, 393–403.
47. Oyedele, O. F.; *J. Appl. Stat.* **2021**, *48*, 1816–1832.
48. Llano Suárez, P.; Soldado, A.; González-Arrojo, A.; Vicente, F.; de la Roza-Delgado, B.; *J. Food Compos. Anal.* **2018**, *70*, 1–8.
49. Khattab, A. R.; Guirguis, H. A.; Tawfik, S. M.; Farag, M. A.; *Trends Food Sci. Technol.* **2019**, *88*, 343–360.
50. Ianni, A.; Bennato, F.; Martino, C.; Grotta, L.; Martino, G. V.; *Molecules* **2020**, *25*, 461.
51. Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B.; *Metabolomics* **2018**, *14*, 72.
52. Dudzik, D.; Barbas-Bernardos, C.; García, A.; Barbas, C.; *J. Pharm. Biomed. Anal.* **2018**, *147*, 149–173.
53. Grassi, S.; Tarapoulouzi, M.; D'Alessandro, A.; Agriopoulou, S.; Strani, L.; Varzakas, T. H.; *Foods* **2022**, *12*, 139.
54. Sztark, A.; Ofiara, K.; Strus, B.; Abdullaev, I.; Ferenc, K.; Sady, M.; Flis, S.; Gajewski, Z. C.; *Foods* **2022**, *11*, 1116.
55. Oliva-Cruz, M.; Mori-Culqui, P. L.; Caetano, A. C.; Goñas, M.; Vilca-Valqui, N. C.; Chavez, S. G.; *Front. Nutr.* **2021**, *8*, 677000.