NOVEL APPLICATIONS OF CHEMOMETRICS IN ANALYTICAL CHEMISTRY AND CHEMICAL PROCESS INDUSTRY

EDITED BY: Alessandra Biancolillo, Angelo Antonio D'Archivio, Federico Marini and Raffaele Vitale

PUBLISHED IN: Frontiers in Chemistry and Frontiers in Analytical Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-297-2
DOI 10.3389/978-2-88976-297-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

NOVEL APPLICATIONS OF CHEMOMETRICS IN ANALYTICAL CHEMISTRY AND CHEMICAL PROCESS INDUSTRY

Topic Editors:
Alessandra Biancolillo, University of L'Aquila, Italy
Angelo Antonio D'Archivio, University of L'Aquila, Italy
Federico Marini, Sapienza University of Rome, Italy
Raffaele Vitale, Université de Lille, France

Citation: Biancolillo, A., D'Archivio, A. A., Marini, F., Vitale, R., eds. (2022). Novel Applications of Chemometrics in Analytical Chemistry and Chemical Process Industry. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-297-2

Table of Contents

frontiers | Frontiers in Chemistry

EDITORIAL
published: 17 May 2022
doi: 10.3389/fchem.2022.926309

Editorial: Novel Applications of Chemometrics in Analytical Chemistry and Chemical Process Industry

*Alessandra Biancolillo [1], Angelo Antonio D'Archivio [1], Federico Marini [2] and Raffaele Vitale [3]**

[1]*Dipartimento di Scienze Fisiche e Chimiche, Università degli Studi dell'Aquila, Coppito, Italy,* [2]*Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza", Rome, Italy,* [3]*Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Lille, France*

Keywords: chemometrics, multivariate statistics, high-dimensional data analysis, analytical chemistry, chemical process industry

Editorial on the Research Topic

Novel Applications of Chemometrics in Analytical Chemistry and Chemical Process Industry

Nowadays, thanks to many ground-breaking technological advances, old and new challenges in chemistry and chemical industry can be constantly addressed by means of cutting-edge analytical platforms, generating massive amounts of complex high-dimensional data. In this regard, chemometric approaches, enabling the extraction of the maximum content of meaningful information such data intrinsically encode, have been playing a key role. The present Research Topic collects a series of articles that actually corroborate this aspect, *i.e.*, how the utilisation of chemometrics could aid practitioners and operators in solving real-world issues in the two aforementioned domains, which, as for most scientific disciplines, are manifold and of rather diverse nature.

Several of these contributions have coped with fundamental methodological problems in the field of Multivariate Statistical Process Control (MSPC), that currently constitutes an undoubtedly *hot topic* given its inherent economic and social implications: Offermans et al. have proposed the use of conditional path modelling to infer the underlying intercorrelations linking different units of a production plant, Rocha de Olivera and De Juan have introduced the application of local Principal Component Analysis (PCA) for the assessment of non-synchronised batch process runs, Paris et al. have explored two different strategies for defining specification regions for raw industrial materals, while Strani et al. have fused near-infrared (NIR) and engineering sensors to construct MSPC control charts for polymerisation reaction monitoring.

Wide attention has also been paid to the world of food manufacturing and quality evaluation. In this sense, Ruiz et al. have developed a diagnostic tool resorting to the principles of Partial Least Squares regression (PLS) for compliant/defective product classification. Nieuwoudt et al. have exploited Analysis of variance-Simultaneous Component Analysis (ASCA) to determine the main sources of variation influencing the performance of various Fourier Transform-InfraRed (FTIR) spectrometers in a milk factory. Astolfi et al. have utilised dedicated chemometric techniques for the authentication of extra-virgin olive oil samples by Inductively Coupled Plasma-Mass Spectrometry (ICP-MS). Finally, Shao et al. have reviewed the state-of-the-art approaches for the electrochemical and biochemical sensor-based characterisation of tea specimens.

New light has also been shed on subjects apparently not yet well-established in the scientific community: Vitale et al., for instance, have addressed the problem of hyperspectral video processing through a hybrid modelling procedure encompassing spatial, spectral and temporal parametrisations of physico-chemical phenomena.

OPEN ACCESS

Edited and reviewed by:
Huangxian Ju,
Nanjing University, China

****Correspondence:***
Raffaele Vitale
raffaele.vitale@univ-lille.fr

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *22 April 2022*
Accepted: *27 April 2022*
Published: *17 May 2022*

Citation:
Biancolillo A, D'Archivio AA, Marini F and Vitale R (2022) Editorial: Novel Applications of Chemometrics in Analytical Chemistry and Chemical Process Industry. Front. Chem. 10:926309. doi: 10.3389/fchem.2022.926309

More theoretical aspects behind the use of chemometrics have been debated by Rutledge et al. who have compared several strategies for the estimation of the optimal complexity of multivariate statistical models.

Last but not least, Mancini and Rinnan as well as Alladio et al. have reported studies bridging elegantly the gap between theory and practice of multivariate statistics applications: the former have designed a solution for estimating waste wood heterogeneity coupling NIR spectroscopy, nested ANalysis Of VAriance (ANOVA) and PCA, the latter have devised a real-time predictive maintenance methodology (that combines Sparse Logistic PCA—SLPCA—and Soft Independent Modelling of Class Analogy—SIMCA) to prevent breakdowns during the evolution of automotive industrial processes.

Overall, as far as the editors are concerned, this Research Topic has surely permitted to stress the importance and relevance that data analysis and, more specifically, chemometrics can have in both basic and applied research scenarios.

AUTHOR CONTRIBUTIONS

AB and RV wrote the first draft of the editorial. All the authors contributed to its revision and approved it for submission.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 13 July 2021
doi: 10.3389/fchem.2021.681958

Handling Variables, *via* Inversion of Partial Least Squares Models for Class-Modelling, to Bring Defective Items to Non-Defective Ones

Santiago Ruiz[1†], *Luis Antonio Sarabia*[1*†], *María Sagrario Sánchez*[1†] *and María Cruz Ortiz*[2†]

[1]*Department Matemáticas y Computación, Facultad de Ciencias, Universidad de Burgos, Burgos, Spain,* [2]*Department Química, Facultad de Ciencias, Universidad de Burgos, Burgos, Spain*

In the context of binary class-modelling techniques, the paper presents the computation in the input space of linear boundaries of a class-model constructed with given values of sensitivity and specificity. This is done by inversion of a decision threshold, set with these values of sensitivity and specificity, in the probabilistic class-models computed by means of PLS-CM (Partial Least Squares for Class-Modelling). The characterization of the boundary hyperplanes, in the latent space (space spanned by the selected latent variables of the fitted PLS model) or in the input space, makes it possible to calculate directions that can be followed to move objects toward the class-model of interest. Different points computed along these directions will show how to modify the input variables (provided they can be manipulated) so that, eventually, a computed 'object' would be inside the class-model, in terms of the prediction with the PLS model. When the class of interest is that of "adequate" objects, as for example in some process control or product formulation, the proposed procedure helps in answering the question about how to modify the input variables so that a defective object would be inside the class-model of the adequate (non-defective) ones. This is the situation illustrated with some examples, taken from the literature when modelling the class of adequate objects.

Keywords: process analytical technology, partial least squares, class-modelling, sensitivity/specificity, latent variables model inversion, authentication, attributes

OPEN ACCESS

Edited by:
Federico Marini,
Sapienza University of Rome, Italy

Reviewed by:
Pierantonio Facco,
University of Padua, Italy
Paolo Oliveri,
University of Genoa, Italy

***Correspondence:**
Luis Antonio Sarabia
lsarabia@ubu.es

[†]*All authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *19 March 2021*
Accepted: *14 June 2021*
Published: *13 July 2021*

Citation:
Ruiz S, Sarabia LA, Sánchez MS and Ortiz MC (2021) Handling Variables, via Inversion of Partial Least Squares Models for Class-Modelling, to Bring Defective Items to Non-Defective Ones. Front. Chem. 9:681958. doi: 10.3389/fchem.2021.681958

INTRODUCTION

Class-modelling techniques (Forina et al., 2008) focus on the ability of the built class-models for recognizing their own objects (sensitivity of the computed class-model) and rejecting all others (specificity). The additional information that the class-models provide about the categories being modelled, as against a pure discriminant rule, is relevant for authentication of products (Rodionova et al., 2016a), for example, to characterize foods or beverages with recognized quality, such as denomination of origin wines or oil (Barbaste et al., 2002; Marini et al., 2006; Forina et al., 2009; Ruisánchez et al., 2021) combined with spectroscopic and chromatographic techniques to characterize green tea (Casale et al., 2018) with near infrared spectroscopy to antibiotic authentication (Chen et al., 2020) to identify bands for functional spectral data (Hermane et al., 2021) for food-authenticity claims (Oliveri and Downey, 2012) for detection of cold chain breaks in tuna (Reguera et al., 2019), or adulterations (Xu et al., 2013a), or nitro explosive vapors (Pablos et al., 2015). Also, a procedure based on band limits are successfully used as probabilistic one-class classifier

(Avohou et al., 2021), among several other applications that can be found in a recent tutorial (Oliveri et al., 2021). In fact, the area is very active: a search in Scopus with key terms "Classification model" OR "Class-modelling" limited to the last five years (2016–2021) and in Chemistry as subject area return 1,013 documents. By reducing the search to ("Classification model" OR "Class-modelling") AND "Chemometrics", there were still 431 resulting documents.

The concept of pattern recognition has evolved since the birth of chemometrics (Brereton, 2015) resulting, more than a decade ago, in the classification of the techniques as either discriminant or one-class classifiers (when modelling the categories independently to one another) (Brereton, 2009). A more flexible taxonomy (Rodionova et al., 2016b) distinguishes between "rigorous" (equivalent to one-class classifiers) and "compliant" class-modelling techniques. To build the class-model only objects of the modelled class are considered in the former case while in the latter, objects of different classes are also used. Alternative denominations make distinction between hard or soft models (Brereton, 2011), as those that do not allow or allow overlap between classes, respectively. This division is also used in ref. (Pomerantsev and Rodionova, 2018) for the particular case of PLS-DA (Partial Least Squares Discriminant Analysis (Ståhle and Wold, 1987; Barker and Rayens, 2003), making a distinction between hard and soft PLS-DA models depending on whether they use LDA or QDA (UNEQ) on the PCA-scores of the PLS-predicted responses.

The assumption under the name "one-class classifier" is that each class is modelled independently of any other class, context that covers the situation where in fact there is a single class, e.g., for authentication purposes (Oliveri and Downey, 2012; Rodionova et al., 2016a). In this case, the quality criterion (figure of merit) of the class-model is only its sensitivity, though it can be possible to estimate the specificity as against other samples by using a different set of objects that do not belong to the modelled class (a so-called specificity set in ref. (Forina et al., 2008)). To obtain an unbiased estimate, the specificity set should be representative of all possible "alternative" classes.

Partial Least Squares for Class-Modelling (PLS-CM), first proposed in ref. (Ortiz et al., 1993), is one class-modelling technique that works by implicitly defining probabilistic class-models with predefined values of sensitivity and specificity or, at least, the closest possible to the desired ones with the data at hand. Unlike PLS-DA that also uses a PLS regression model with binary response, PLS-CM first fits probability density functions to the predicted values, separately in each class, which act as probabilistic class-models. For given values of sensitivity or specificity, a decision threshold can be defined as the critical value computed with the fitted distributions. Since the two class-models are fitted together to estimate both sensitivity and specificity, with the distinction in ref. (Rodionova et al., 2016b), the method would be a "compliant" class-modelling method.

Under the same acronym PLSCM, Xu et al. (Xu et al., 2011) build class-models for a single class. The class-model is a kind of confidence interval of the form $(1-\widehat{\mu}_r) \pm \ z_{1-\alpha/2}\widehat{\sigma}_r$ where $\widehat{\mu}_r, \widehat{\sigma}_r$, are estimates computed with Monte Carlo Crossvalidation, of the mean and standard deviation of the residuals of a PLS model with constant response (response matrix $\mathbf{Y}$ is a vector of ones), assuming that they follow a normal distribution. Accordingly, $z_{1-\alpha/2}$ is the critical value of the standard normal distribution for 1—α confidence. A new sample is inside the class-model if its predicted response $\widehat{y}_{un}$ belongs to the interval. Two years later (Xu et al., 2013b), with the new name OCPLS (one-class partial least squares) classifier, the authors add bounds on the allowed variation of the T^2 statistic as well as on a transformation of the residuals of the regression (difference between one and the predicted responses) to create an outlier identification plot.

Comparing OCPLS (Xu et al., 2013b) with PLSCM (Ortiz et al., 1993), the differences are in the use of one-class (an interval as class-model) or two-classes (probability density functions as class-models) for modelling and that in PLSCM the bounds are imposed as hard constraints in the values of T^2 and Q-residual statistics to reject objects from both class-models.

The membership of an object $\mathbf{x}$ to a given class-model can be posed as a hypothesis test with null hypothesis H_0: object $\mathbf{x}$ belongs to the class-model as against H_1: it does not. With the usual notation, α is the significance level of the test, that is, the probability of type I error (wrongly rejecting the null hypothesis), and β is the probability of type II error (fail to reject the null hypothesis). Then, sensitivity of the class-model is 1—α and specificity is 1—β, the power of the test. The general notion of type I and type II errors (with probabilities α and β, respectively) is usually adapted to the context (Ortiz et al., 2010), and becomes false non-compliance/compliance or false positive/negative, whose meaning is clear once undoubtedly established the hypothesis being tested (the meaning of the "class" we are studying in the class-modelling framework). Speaking in *positive*, the terms sensitivity, specificity, true positive/negative rates, confidence level or power can also be used.

To avoid misunderstanding and facilitate the reading of the paper, in what follows, we will always speak about sensitivity and specificity, which will be estimated as the probabilities that characterize the corresponding class-model, computed with the fitted distributions.

In the illustrative examples in the present work, the class to be modelled is the class of some *adequate* objects, again, understood in a general sense. Besides authentication or fraud detection, another particular situation that fits this framework could be the modelling or monitoring of a process where the class-model of interest is the one for non-defective objects and, clearly, the probability of detecting a defective object (specificity) is important. Furthermore, it can be assumed that the expected failures are known, in other words, that there will be samples representative of the usual defective objects acting as the alternative class. Therefore, the training set for fitting the PLS model has samples representative of both situations: usual defective objects and non-defective ones.

It has been said that with PLSCM, the class-models are defined in the space of the predicted responses. To *backpropagate* them into the input domain requires the inversion of the prediction model. Briefly, the inversion of a model refers to the situation where we have the values of the characteristics we want to achieve

(output space), and the aim is to find the values of the predictor variables (input space) to attain them.

The inversion of PLSCM is a LVMI (Latent Variables Model Inversion), term used more frequently in the field of process industry after the seminal papers by Jaeckle and MacGregor (Jaeckle and MacGregor, 1996; Jaeckle and MacGregor, 2000a). A general formulation for LVMI when the latent variables are computed with PLS is in ref. (Tomba et al., 2012) with a through discussion and also a revision of available literature and applications at that time. There, 95% confidence limits on T^2 and Q statistics are already applied to the PLS model fitted with historical data, so that the operating conditions obtained with PLS model inversion must be interior to it. The region defined with these hard constraints on the solutions was later called PLSbox in ref. (Ruiz et al., 2020) where also the explicit consideration of two existing null spaces (one due to the projection into the latent space and the other from the mapping of the scores onto the responses) in PLS model inversion is described. Some more developments about LVMI can be found in refs. (Tomba et al., 2013; Ottavian et al., 2016; Palací-López et al., 2020), and (Zhao et al., 2019a; Zhao et al., 2019b) where the authors propose a modification called the total projection in latent structures of PLS model inversion to take into account that latent variables of a PLS model may contain information irrelevant to the response. Also, by imposing hard constraints on the input domain further to the PLSbox, a different approach to the inversion is in ref. (Ruiz et al., 2018), similar to the one in Lakshminarayanan et al. (Lakshminarayanan et al., 2000) but for inverting PLS2 models. The use of PLS model inversion for product formulation is also noteworthy, especially in the context of Process Analytical Technology with pharmaceutical processes (Tomba et al., 2014; Bano et al., 2017; Palací-López et al., 2019).

In the present work, with PLS-CM, given values of sensitivity and specificity determine a decision threshold y_d to be imposed in the predicted responses, threshold that acts as the boundary of the class-model. The inversion of the built PLS model for y_d would provide values of predictor variables $\mathbf{x}_\mathrm{d}$ (a vector in the input domain) whose prediction is exactly y_d.

In general, the solution $\mathbf{x}_\mathrm{d}$ is not unique, due to the null space of the PLS model (Jaeckle and MacGregor, 2000b). The null space contains the values of the predictor variables $\mathbf{x}_\mathrm{null}$ (vectors in the input space) that are mapped into zero by the linear model, so that any point $\mathbf{x}_\mathrm{d} + \mathbf{x}_\mathrm{null}$ have the same predicted response y_d.

Since there is a single response (dimension 1), the consideration of the null space when inverting the PLS model would define a (subset inside a) hyperplane in the input space. The objects lying on that hyperplane are at the boundary of the class-model but already in the input space. Moreover, the characterization of this boundary would give indications on how to manipulate or to modify the input variables so that a rejected object can become an accepted one. The details on how to do that are explained in section *Materials and methods*. The computation and possible utility are illustrated in section *Results and discussion* with some data sets taken from the literature. The paper finishes with some conclusions.

TABLE 1 | Settings of the computed plastic pellets following the direction signaled in **Figure 6**.

#	Situation	Size5	Size10	Size15	TGA	DSC	TMA
1	Rejected	14.24	10.07	34.43	622.00	18.73	52.08
2	Rejected	13.51	8.91	32.15	638.47	18.67	53.60
3	Rejected	13.03	8.13	30.63	649.45	18.63	54.61
4	Accepted	12.66	7.55	29.49	657.69	18.60	55.37
5	Accepted	12.29	6.96	28.35	665.92	18.56	56.13
6	Accepted	11.93	6.38	27.21	674.16	18.53	56.89
7	Accepted	11.56	5.80	26.07	682.39	18.50	57.65
8	Accepted	11.20	5.22	24.93	690.63	18.47	58.42
9	Accepted	10.83	4.64	23.79	698.87	18.44	59.18
10	Accepted	10.47	4.05	22.65	707.10	18.41	59.94
11	Accepted	10.10	3.47	21.51	715.34	18.38	60.70

MATERIALS AND METHODS

Partial Least Squares for Class-Modelling

Let $\mathbf{X}$ $(n \times p)$ be a data matrix with p variables measured on n objects, which belong to two categories, class A and B. This set would be the training set, so that it is assumed that it contains representative samples of these two categories or classes.

The PLSCM method consists of fitting a PLS model to a binary response that codifies the categories. If they are coded as −1 and +1, respectively, the n-dimensional vector of responses, $\mathbf{y}$, is made up of as many "−1" as objects belonging to category A and as many "+1" as objects of category B in the training set.

The selection of the proper number of latent variables for the PLS model is based on crossvalidation estimates. Throughout the fitting, objects that surpass the 95% confidence limits on both Q and T^2 statistics, if any, are removed and the model is rebuilt.

During the application phase (i.e. when predicting with the fitted model), the predictions are calculated only for the objects with values in both statistics less than the limits stablished (hard constraints, which are *restrictions that determine the envelope of the subspace of acceptable solutions* (Palací-López et al., 2020). Along the paper, to illustrate the methodology, the usual 95% confidence levels are used. Reducing this level would probably shrink the class-models, or the contrary if it is increased, yet in the present work no sensitivity analysis of the results on the confidence levels has been performed.

As PLS models are regression models for fitting quantitative variables, the individual predicted responses $\hat{y}_i$ are neither −1 nor 1 but different values spreading around −1 and 1. The method then consists on separating these predicted values, according to the class each object belongs to, and probability distributions are fitted independently to each class. Thus, random variable X_A related to PLS prediction for class A follows a F_A distribution and X_B, related to class B, follows a F_B distribution.

Several normality tests are conducted to fit F_A and F_B. If the normal distribution is not adequate, an alternative distribution will be selected, based on the maximum likelihood.

Without loss of generality, let us suppose that we focus on the class-model of class B (coded as '1'). This could be the situation for the particular case of modelling defective/non-defective

objects, for example, where class B would be the category of non-defective objects.

In any case, for a given sensitivity s in [0, 1], we use the cumulative distribution function of F_B to compute the critical value y_c so that $P(X_B \leq y_c) = 1 - s = \alpha$. This critical value will act as a decision threshold, that is, object i-th is assigned to the model of class B when $\widehat{y}_i \geq y_c$ and to class A otherwise. Consequently, y_c defines the boundary of the class-model. It is worth remembering that alien objects (outside both class-modes) are previously removed with the hard constraints imposed on Q and T^2 statistics.

Finally, the specificity sp of the class-model as against class A is given by $P(X_A \leq y_c)$, which is computed with the cumulative distribution function F_A. In this way, as expressed in **Table 1** of (Rodionova et al., 2016a) for class-modelling techniques, PLSCM gives a decision rule for a given α as a result of the modelling, and sensitivity and specificity can be computed as the usual figures of merit.

Inversion of a Partial Least Squares Model

Once fitted a PLS model, its typical use is to predict values of y given $\mathbf{x}$ (p-dimensional vector of predictor variables). The reverse situation, looking for the values of $\mathbf{x}$ whose prediction is a predefined y requires the inversion of the regression model.

In the context of process control or product formulation with a PLS prediction model, its *direct* use means predicting quality characteristics of the product manufactured with given settings $\mathbf{x}$ of input variables (process variables, characteristics of material including their amounts mixed, environmental variables, etc.). Thus, the inversion of the PLS model would refer to the situation where we have the desired characteristics and need to find the settings of the input variables, if any, to attain them.

In the following, we will introduce the inversion of the PLS model for a single response, which is the only situation that applies here. With the notation stablished in the previous section, $\mathbf{X}$ ($n \times p$) is the matrix of predictor variables and $\mathbf{y}$ is the response vector with the n binary values. In the class-modelling situation, the PLS model fitted to $\mathbf{X}$-$\mathbf{y}$ leads to defining different threshold values y_d, each one related to a pair (sensitivity, specificity) that qualifies the corresponding class-model.

Consequently, by defining y_d as the target value, the inversion of a PLS-CM model would provide values of the predictor (input) variables that are mapped exactly into y_d via the PLS model, i.e., the characteristics of the objects that are directly projected into the class-model boundary. Therefore, setting aside the uncertainty in the prediction of any data-driven model, these objects would represent the boundary of the class-model already in the input space. Since PLS is a linear model, the boundary thus constructed is also linear. These ideas are developed in a more precise way in the following lines.

With a single response in the response space, like in this case, the inversion of the PLS model with a latent variables can be computed algebraically because it consists on solving **Eq. 1** in $\mathbf{x}$.

$$\widehat{\mathbf{y}} = \mathbf{T}\mathbf{Q}^T = \mathbf{x}^T\mathbf{W}\mathbf{Q}^T \tag{1}$$

where $\mathbf{T}$ ($n \times a$) is the matrix of common scores, $\mathbf{W}$ ($p \times a$) is the weights matrix and $\mathbf{Q}$ ($1 \times a$) is the $\mathbf{y}$-loadings matrix (which is a row vector in this case). As usual, superscript T means transposing.

The input space of predictor variables has dimension p and the dimension of the output (response) space is one. Therefore (Lay et al., 2016), the kernel of the PLS model (null space of $\mathbf{QW}^T$), which is the set of points with null response, has dimension p—1 > 0 unless p = 1, which would be a very unrealistic situation. Therefore, the null space is a hyperplane in the input space passing through zero (p-dimensional vector of null coordinates), that is, a linear subspace.

Because of their own definition, any vector in the null space adds variability in the input space without modifying the predicted value. That means that, given a desired y_d, for any p-dimensional solution of the inversion, that is, any vector $\mathbf{x}_d$ with $\mathbf{x}_\mathbf{d}^T\mathbf{W}\mathbf{Q}^T = y_d$, all the remaining solutions of **Eq. 1** can be written as

$$\{\mathbf{x}_d + \mathbf{x}_0 \ : \mathbf{x}_0^T\mathbf{W}\mathbf{Q}^T = 0\} \tag{2}$$

Hence, the inversion has infinitely many solutions for y_d, although it suffices to consider one of them and characterize the null space.

A sequential alternative for the inversion starts by finding the vector of scores $\mathbf{t}_d$ (a-dimensional) such that

$$y_d = \mathbf{t}_d^T\mathbf{Q}^T \tag{3}$$

In this sequential approach, the dimension of the latent space spanned by $\mathbf{T}$ is a so the null space inside the latent space has dimension a—1 (which is positive for more than one latent variable), i.e., for $a > 1$ the null space is also a hyperplane, but inside the latent space.

Because of this null space, the solution of **Eq. 3** is not unique either, there are infinitely many solutions described from any particular one, $\mathbf{t}_d$, in the set in **Eq. 4**.

$$\{\mathbf{t}_d + \mathbf{t}_0 \ : \mathbf{t}_0^T\mathbf{Q}^T = 0\} \tag{4}$$

All a-dimensional vectors belonging to the set in **Eq. 4**, solutions of **Eq. 3**, lie on a hyperplane in the a-dimensional latent space that, contrary to the null space, does not contain the null vector (unless, of course, $y_d = 0$).

This property about null spaces of linear models has been already used in ref. (Largoni et al., 2015). to divide the latent space into two subspaces, one for on-spec batches and the other for off-spec batches, depending on an end-point product quality.

In the present context with PLSCM, given the threshold value y_d, the hyperplane in **Eq. 4** is in fact the decision boundary of the class-model in the latent space. Moreover, via the $\mathbf{X}$-loadings matrix $\mathbf{P}$ ($p \times a$), **Eq. 5** gives the objects in the input space whose projection are the scores in **Eq. 4**.

$$\widehat{\mathbf{x}}_d = (\mathbf{t_d} + \mathbf{t}_0)\mathbf{P}^T \text{ with } \mathbf{t}_\mathbf{d}^T\mathbf{Q}^T = y_d, \text{and } \mathbf{t}_0^T\mathbf{Q}^T = 0 \tag{5}$$

Because all the scores in **Eq. 4** lie on the same hyperplane, the corresponding input objects computed with **Eq. 5** also belong to

a subspace of dimension $a-1$ inside the p-dimensional input space.

However, once in the input space and if $p > a$ (which is usually the case), there are still some more solutions of the inversion, additional to the ones computed with **Eq. 5**. They correspond to a $(p—a)$-dimensional subspace obtained when adding points (p-dimensional vectors) that belong to what we have called the **W**-null space (Ruiz et al., 2020), spanned by the loadings of the latent variables discarded when building the PLS model.

Consequently, the solutions in **x** of **Eq. 1** for $\hat{y} = y_d$, described in **Eq. 2**, are also described as in **Eq. 6**, where $\hat{\mathbf{x}}_\mathrm{d}$ is defined in **Eq. 5**.

$$\left\{\hat{\mathbf{x}}_d + \mathbf{x}_{\mathrm{w0}} \ : \mathbf{x}_{\mathrm{w0}}^T \mathbf{W} = 0\right\} \tag{6}$$

A final consideration is worth mentioning. Although the PLS prediction for all the points in either **Eq. 2** or **Eq. 6** will be y_d, not all of them define a *feasible* object or, in general, a valid solution of the inversion. The valid solutions are those that belong to the PLSbox (Ruiz et al., 2020), which is the region of applicability of the model, characterized by the limits imposed on both the Q and T^2statistics when fitting the PLS model; and that also belong to a given domain D inside the input space, that accounts for the characteristics of the input variables in each particular application. This domain should be explicitly defined since it imposes additional hard constraints for the valid solutions of the inversion.

For the present work, the PLSbox is defined with the limits at 95% confidence level. The domain D on its part is defined with the range of the variables in the training set, which at least describes the physical bounds on the predictor input variables (Tomba et al., 2012).

In what follows, we will only consider *valid* (*feasible*) solutions of the inversion, that is, points whose prediction is y_d and that belong to both D and the PLSbox.

If the situation were one that fits any form of process control, or product formulation, the general principle in model inversion problems is to manipulate the variables that can be manipulated (in a process control sense or compositional variables) to obtain a product as close as possible to the required specification (Dunn, 2020).

The *specification* in the situation being discussed is related to sensitivity and specificity of the class-model, and the solutions of the inversion give the boundary of the class-model. Thus, different directions of *manipulation* (of scores inside the latent space or of variables in the input space) can be defined, any of them crossing the boundary at some point so that following the direction allows moving in or out of the class-model.

In the latent space, the most easily computable direction is the one defined by the normal vector of the boundary hyperplane (i.e., the vector perpendicular to the hyperplane) which is $\mathbf{Q}^T$. This direction does not depend on the inversion of the model but the precise position of the hyperplane does, that is, at least one solution of the inversion is needed to have the boundary that allows deciding whether a given object is inside or outside the class-model.

The same idea can be applied directly in the domain D to define a direction of movement/manipulation of the input variables. In this case, it would be the straight line whose director vector is $\mathbf{QW}^T$, orthogonal to the global null space of the fitted PLS model and, thus, to any hyperplane computed as in **Eq. 2** or **Eq. 6**, that positions the boundary of the class-model in the input space.

Data Sets

Two different data sets are considered to illustrate the proposed method. The first one does not come from a process with attributes data but illustrate other situations, provide some of the variables can be manipulated. The second one will emulate the use of historical data to fit a model that helps in process control and/or product formulation.

The first data set[a] contains samples of 128 red young wines from Spanish DOC (*Denominación de Origen Calificada*) Rioja (Ortiz et al., 1995). The wines are characterized by six variables related to physical-chemical measures of color, namely red/green chromaticity (a), yellow/blue chromaticity (b), lightness (L), chroma (C), hue (H), and saturation (S). Expert tasters visually assess the color of each wine and divide the objects into two categories, acceptable or non-acceptable wines because of their color.

The second data set[b] contains six characterizing measurements for batches of plastic pellets, which will be the predictor input variables, with 24 rows. The first three characteristics, coded for confidentiality, are related to the percentage material in the mixture with different size range (size5, size10 and size15). The last three characteristics are measurements from TGA (thermal gravimetric analysis), DSC (differential scanning calorimetry) and TMA (thermomechanical analysis) devices. The outcome when using this material is either Poor or Adequate.

RESULTS AND DISCUSSION

Rioja Red Wines

Predictor matrix **X** is 128 × 6 and response **y** is a vector with binary values, namely −1 for non-acceptable wines and one for the acceptable ones. With autoscaled **X** and **y** and leave-one-out crossvalidation, a three latent variables PLS-model is fitted that explains 91.01% of variance in **X** with 72.86% in **y** (70.88% in crossvalidation).

The predicted responses corresponding to non-acceptable wines are fitted to a normal distribution with mean -0.95 and standard deviation 0.45 (the smallest p-value among the tests performed was greater than or equal to 0.10, thus, the idea that the values come from a normal distribution cannot be rejected with 90% or greater confidence). On the contrary, the responses corresponding to acceptable wines are not compatible with a normal distribution. The minimum log likelihood was similar for a beta distribution with four parameters and to a highly asymmetric triangular distribution with three. This was the one selected with lower limit −0.57, center point 1.16 and upper limit 1.17.

[a]Available in RIUBU, at http://hdl.handle.net/10259/5753

[b]Available at http://openmv.net/info/raw-material-characterization

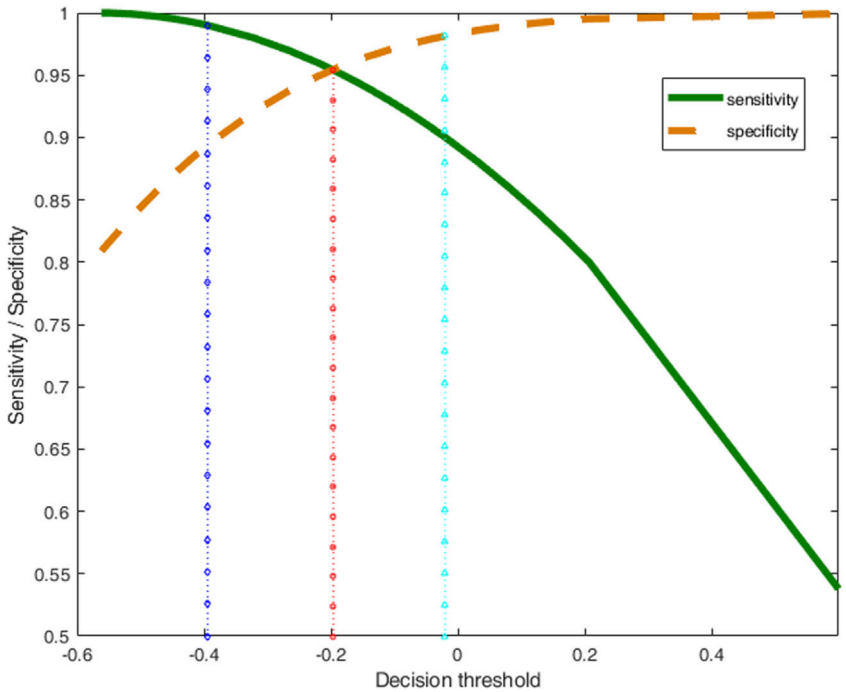

FIGURE 1 | Sensitivity (green thick line) and specificity (brown dashed line) of the class-model of 'acceptable' wines as a function of the decision threshold. The dotted vertical lines mark the decision thresholds for obtaining different class-models in terms of their sensitivity and specificity.

Without loss of generality, let us focus in the class of acceptable wines, codified as 1. The fitted probability distributions allow setting different decision thresholds y_d which, in turn, are related to different values of sensitivity and specificity for the class-model of the acceptable wines.

These values are depicted in **Figure 1** (green continuous line for sensitivity, brown dashed line for specificity) as a function of the decision threshold. It is clear how larger threshold values results in an increase of specificity (dashed line), linked to a decrease of sensitivity (continuous line).

From the set of possible class-models computed with PLS-CM, the more balanced one is the one indicated with the vertical red dotted line, with little squares in **Figure 1**, for which we expect the same values of sensitivity and specificity, 0.954 in this case, that corresponds to $y_{\mathrm{d}} = -0.196$.

By using this y_d as target value, the inversion of the PLS model would provide points in the input space (where the objects vary) whose predicted response will be exactly the decision threshold y_{d}, according to **Eq. 1** with $\hat{y} = y_d = -0.196$. Working sequentially, the solutions of **Eq. 3** are scores in the latent space, some of them depicted in **Figure 2A** as red squares.

By using the loadings as in **Eq. 5**, the corresponding points in the input space are in six dimensions. Therefore, the usual Cartesian representation is not available. Extensions to visualize data in greater dimension includes the so-called matrix plot, which consists of a set of two-by-two Cartesian plots for any two variables. This matrix plot is usually more informative when representing the scores of a PCA (Principal Component Analysis) that better describe the internal correlation structure of data.

Another alternative, whatever we are visualizing, the Parallel Coordinates Plot also helps in describing the joint behavior of the variables (the "coordinates" of the points). The value of each coordinate is plotted as height above the ordinate axis, against its position in the vector. Then, the values are linked together by a broken line to follow each point. Therefore, rather than its usual meaning, the abscissa axis only accommodates as many slots as coordinates in the point. Although with this disposition there is no limit to the dimension of the points depicted as Parallel Coordinates Plot, it becomes messier when increasing the number of coordinates.

In any case, the points in the input space that correspond to the red squares in **Figure 2A** are depicted also in red in **Figure 2C** in the form of a parallel coordinates plot. In both cases, we are seeing points falling on the boundary of the class-model, whether scores in **Figure 2A** or raw variables in **Figure 2C**.

If the requirements on the class-model change, the decision threshold y_{d} also changes. To illustrate this property, the inversion procedure is repeated for another two different threshold values, in blue and cyan vertical lines in **Figure 1**, that correspond to the class-models with the usual 0.99 and 0.90 sensitivity, respectively.

Figure 2 also shows some new valid solutions predicting every threshold in both the input and latent spaces. **Figure 2B** and **Figure 2D** depict the raw variables in the input space (in the domain D defined by the range in $\mathbf{X}$) in the form of a parallel coordinates plot. **Figure 2A** is the plot of their projection (scores)

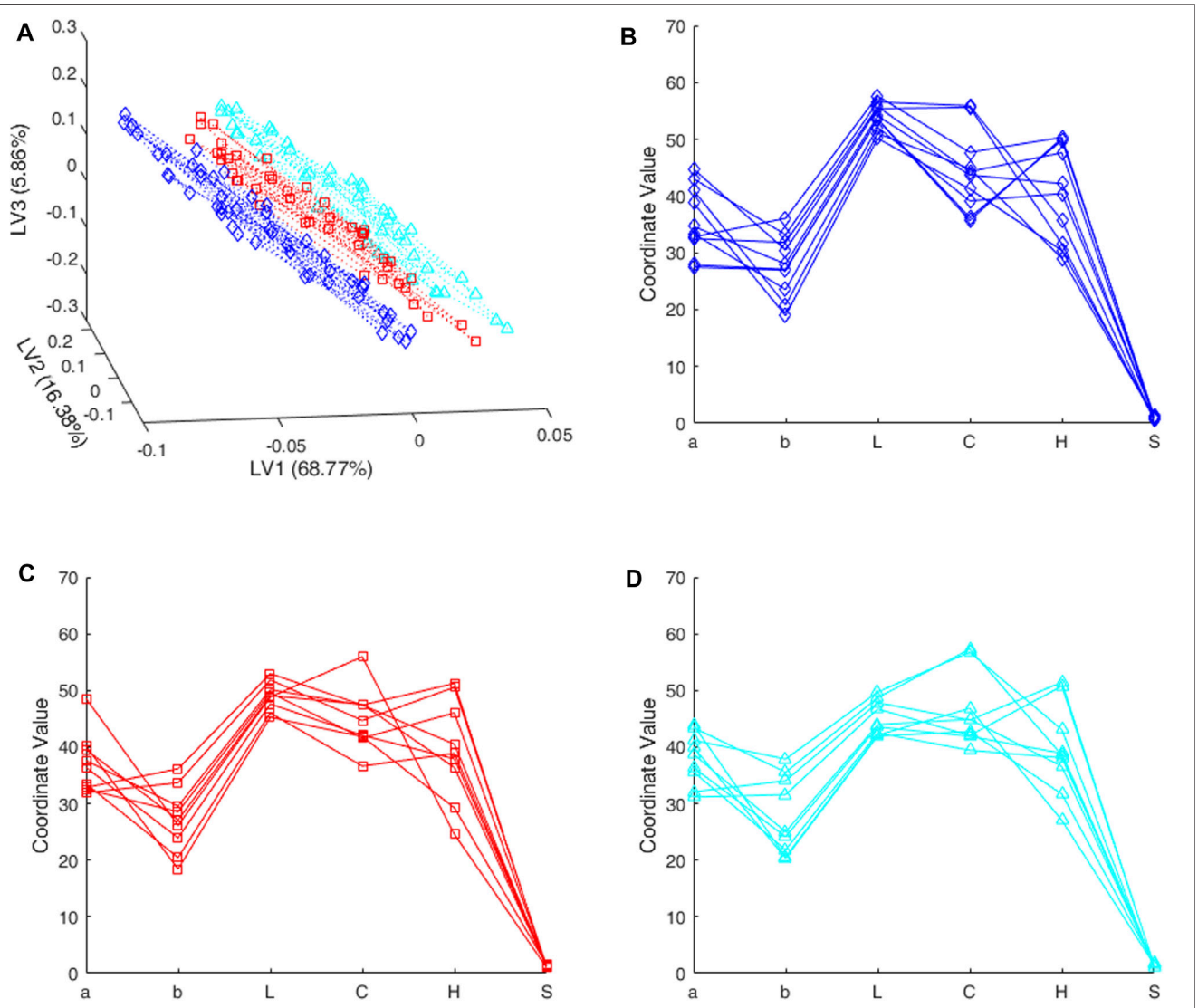

FIGURE 2 | Rioja wines. Boundary objects for different class-models **(A)** latent space **(B–D)** input space. Blue lines and rhombuses are for the class-model with sensitivity 0.99 and 0.89 of specificity; red lines and squares are for the balanced class-model with both sensitivity and specificity equal to 0.954; cyan lines and triangles are for the class-model with sensitivity 0.90 and 0.98 of specificity.

in the 3D-latent space. In both cases, the solutions in blue (lines and rhombuses) are for the class-model with sensitivity 0.99 (with 0.894 of specificity, see **Figure 1**); cyan lines and triangles are for the class-model with sensitivity 0.90 (specificity 0.981).

As we have a single response, the null space in the latent space is a plane because we have three latent variables. Consequently, the projection of the computed solutions into the latent space will be in the corresponding 2-dimensional subspace. The dotted lines in **Figure 2A** are meant to help observing how the points of the same color lie on the same plane, and different colors and symbols define different parallel planes in the latent space.

It is less clear but the corresponding objects in the **X**-space in **Figures 2B–D** are in a two-dimensional subspace inside the boundary of the different class-models, and thus they correspond to some kind of prototype discriminating objects. To make graphs clearer, only around fifty points were calculated for each threshold. However, any convex combination of any pair of points in **Figure 2** is also a valid solution and therefore belongs to the boundary of the class-model at hand.

In any case, the solutions depicted have different values for the variables, in particular, we see how the boundary objects for the balanced class-model in red, that clearly occupy an intermediate position among scores in **Figure 2A**, have not so clear differences in **Figure 2C**, when comparing with **Figures 2B,D**.

Finally, there are some more possibilities that do not come from the latent space or, in other words, that predict the same threshold value but are projected into the origin of the latent space. All points together, added to a particular solution as in **Eq. 6**, define the boundary of the class-model (a hyperplane) in the domain D of the input variables.

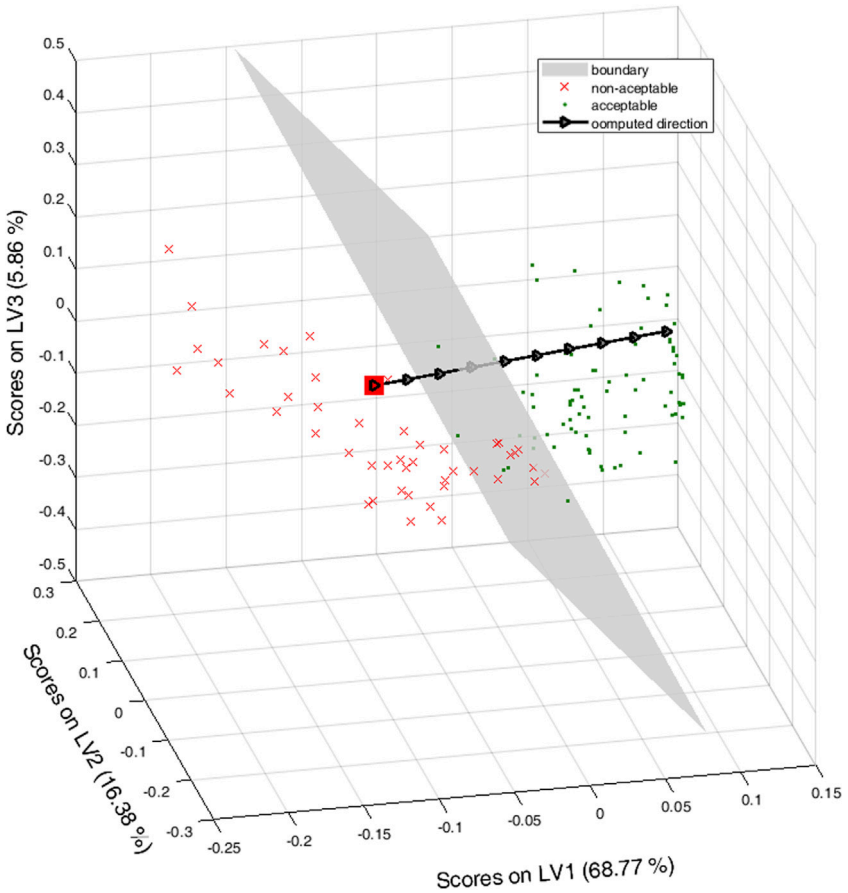

FIGURE 3 | Latent space for the Rioja wines. Green points are for acceptable wines, red crosses and the red filled square for non-acceptable ones. The grey plane is the boundary plane for the class-model when sensitivity and specificity both equal to 0.954. The black triangles are along the direction of improving the color of the wine.

From the practical point of view, it is probably more interesting to notice that the probability of being inside the class-model of accepted wines increases when moving in the latent space, graphically in **Figure 2A**, from scores near the blue rhombuses (which, in fact, define a plane), traversing the red squares toward scores 'above' the cyan triangles which define another plane.

Obviously, each wine is projected into a unique position in the latent space and its acceptance or rejection depends on the sensitivity and specificity selected to make the decision. However, for a given class-model, we can compute scores (ideal scores not necessarily corresponding to any of the wines in the training set) moving in the direction of improving the color toward the acceptance of the wine.

For example, let us consider the balanced class-model (in red lines or squares in **Figure 2** with sensitivity and specificity both equal to 0.954) and let us take one of the wines rejected with the class-model, $\mathbf{x}_d$, which is outside the class-model of the acceptable wines, with a 0.046 probability (4.6%) of being wrongly rejected.

Its projection into the latent space is the filled red square in **Figure 3**, where the boundary plane is depicted in grey extending the convex hull of the red squares in **Figure 2A** to better illustrate the indeterminacy due to the null space. For reference, the scores of the training set are also depicted, red crosses for the non-acceptable wines, green points for the acceptable ones.

Filled black arrows in the black line in **Figure 3**) mark an ideal direction of improving the color, discretized by taking 10 points equally spaced along the line segment orthogonal to the plane and starting in $\mathbf{x}_d$. Graphically, it is clear that, at some point, the computed score crosses the plane and then, the corresponding object would be inside the class-model of acceptable wines.

The objects in the input space whose projections are the ten scores along the black line in **Figure 3** are the colored lines in the Parallel Coordinates Plot in **Figure 4**, from the continuous red line (that corresponds to the non-acceptable wine $\mathbf{x}_d$) to the dash-dotted and dashed red lines, both still for rejected objects.

Following further the same direction pointed in **Figure 3**, we have the continuous green line, already inside the class-model and the remaining green lines (dot dashed, dashed, dotted and thinner continuous, dot-dashed and dashed green lines) depicting objects that would be "more and more clearly" inside the defined class-model and, hence, accepted. For reference, the light grey lines in **Figure 4** are the wines of the training set accepted with the

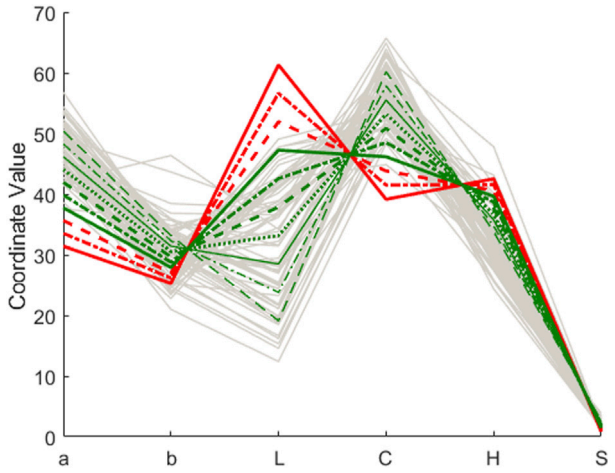

FIGURE 4 | Rioja wines. Parallel Coordinates Plot for objects in the input space. In grey the wines inside the class-model with sensitivity and specificity 0.954. The red continuous line is $\mathbf{x}_{cl}$, the remaining colored lines are for the points computed, rejected in red, and accepted in green.

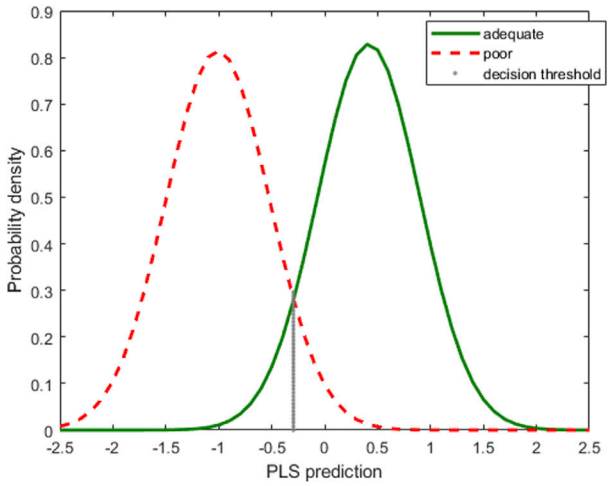

FIGURE 5 | Plastic pellets. Probability density functions of the normal distributions fitted to the PLS predictions, red and dashed for the poor category, green for the adequate one. The vertical dotted line marks the decision threshold for equal sensitivity and specificity.

class-model. It is clear that the green lines are, more and more, among the real values of the acceptable wines.

We have already said that, except for the red continuous line, the remaining colored lines in **Figure 4** are computed points. Nevertheless, they show how the movement along the line in **Figure 3** is related to a systematic variation of the input variables. Following the different lines in **Figure 4**, we see that to improve the color of the wine toward its acceptance, it is necessary to increase *a* and (to a lesser extent) *b*, decrease *L*, increase also *C*, decrease *H* and slightly increase *S*, but always maintaining the exact relation (relative systematic variation) shown in **Figure 4**.

Although there is more than one direction to exert the same effect, with the one selected, it is clear that moving the colorimetric parameters in the adequate range and relation, which is viable for an expert oenologist by mixing different wines, it is possible to get closer to and eventually inside the class-model of acceptable wines, based on their color.

Plastic Pellets

In this case, matrix **X** of predictor variables is 24 x 6. The outcome when using the corresponding material, either poor or adequate, is coded into −1 and 1, respectively, to form the matrix of binary responses to be predicted.

With autoscaled predictors in **X** and binary responses in **y**, also autoscaled, a PLS model is fitted with two latent variables that explain 73.64% of the variance in **X** and 66.17% of the variance in **y**, with R^2_{cv} = 56.65% (obtained with venetian blinds, ten data splits, one sample per blind).

The low predictive ability of the model could be due to the small number of samples at our disposal. This implies that the conclusions obtained can carry great uncertainty, which is one the reasons why the results should be experimentally validated, whenever possible. However, the example is still valid to show how to proceed.

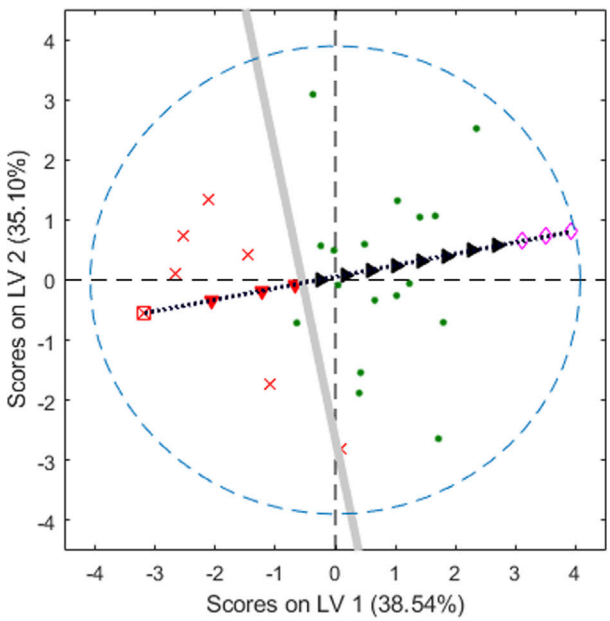

FIGURE 6 | Plastic pellets. Second vs. first latent variables. Green points for adequate objects, red crosses, and the square, for the poor ones. The thicker grey line is the decision line with 93% sensitivity and specificity and the black dotted line is the direction of 'repairing' the *poor* object to become *adequate*. Red filled downward-pointing triangles for poor, black filled right-pointing triangles for adequate. Empty rhombuses are outside the domain. The blue dashed line is the 95% confidence limit of the PLSbox inside the latent space.

The PLS-predictions for the class adequate are fitted to a $N(0.42, 0.48)$, with the smallest *p*-value for several normality hypothesis tests being greater than 0.10. The small number of samples in the class poor prevent testing the normality, though

the points are well aligned in the 'normal probability plot'. Therefore, for the computation of sensitivity and specificity the N (−1.01, 0.49) is used for the poor class.

The corresponding probability density functions of the fitted distributions are depicted in **Figure 5**, red dashed line for the poor category, green continuous line for the adequate one. Again, we focus on the class of adequate pellets, coded as 1, that mimics the situation of a process control with attributes data: one minus the sensitivity of the class-model would be the probability of false alarm and the specificity would be the power to detect a true defective (poor) object.

Choosing a threshold value for PLS predictions, for instance the one marked with the vertical dotted line in **Figure 5**, means defining a class-model whose sensitivity is the probability under the green curve to the right of the line, whereas the specificity would be the probability under the red dashed curve to the left of the black vertical dotted line.

In fact, usually, first the sensitivity and specificity required for the decision are set, and then, taking into account the fitted distributions, the decision limit y_d is computed. In the illustration of **Figure 5**, the value y_d = -0.2913 corresponds to the class-model with the same sensitivity as specificity, namely 92.9%.

As we have already pointed out, the inversion of y_d up to the latent variables space has infinitely many solutions, all obtained when adding points belonging to the null space (Jaeckle and MacGregor, 2000b), precisely, in what we have called the **Q**-null space (Ruiz et al., 2020). Therefore, the set of solutions defined in **Eq. 4** is a subspace (a hyperplane) in the latent space, the grey straight line in **Figure 6**, representing the boundary line for the chosen class-model.

Graphically, all the objects whose scores are "to the right" of the grey line will be inside the class-model of adequate objects. On the contrary, those whose projections are "to the left" of the grey line will be predicted as poor (or, more precisely, they are predicted to be outside the class-model of adequate pellets).

However, it is clear that if the scores move along, for example, the black dotted line (orthogonal to the decision line), eventually, they will fall inside the class-model of the adequate objects. This is the situation illustrated with the different symbols superimposed on the line that starts at one of the poor pellets, the empty square, followed by (computed) scores, red filled downward-pointing triangles, still rejected by the class-model, up to the black filled right-pointing triangles corresponding to points inside the class-model.

Undoubtedly, we can go on moving along the line in the mentioned direction. However, only the valid solutions should be considered, that is, those scores corresponding to objects inside the PLSbox (whose boundary in terms of the 95% confidence level for the T^2 statistic is depicted as the blue dashed line in **Figure 6**) and inside the input domain. For example, the three empty rhombuses in **Figure 6** follow the right direction, but their corresponding points in the input space, though inside the PLSbox, are outside the domain defined with the range of the variables in the training set, and they should be discarded.

By multiplying by the loading on **P**, as in **Eq. 5**, the valid scores can be seen in the domain inside the space of the input variables where some of them can be manipulated. The computed solutions are written in **Table 1**, whose rows follow the order along the direction of improvement in **Figure 6**. Accordingly, the first three computed objects are rejected by the class-model, the remaining objects are accepted, i.e., inside the class-model of the adequate pellets.

In general, when seeing the computed values in the order of **Table 1**, in each individual variable, it is shown that to improve the characteristics of the poor object to become adequate the percentage material of all sizes should be reduced as well as the DSC measurements and, at the same time, the TGA and TMA measurements should increase.

Table 1 shows that, following the selected direction from a poor pellet (rejected by the class-model) to an accepted object (inside the class-model) by theoretically modifying its formulation, there is also bounds for these six variables for adequate pellets, namely, Size5 must be less than 12.66, the upper bound of Size10 is 7.55 and 29.49 for Size15, whereas the DSC measurements slowly decrease from 18.60. Similarly, from row four in **Table 1**, TGA measurements should be greater than 657.69 and TMA measurements start from 55.37. Taking into account the actual domain, defined with the data at hand, the restriction of being in both the PLSbox and the domain also imposes upper bounds for TGA and TMA measurements and lower bounds for the other four variables.

In any case, the variables cannot be varied in the sense of **Table 1** independently of each other, they should follow the relation shown in the different rows of **Table 1**, or any convex combination of any of those rows.

A principal component analysis (PCA) on **X** (autoscaled) shows that the first two principal components, depicted in **Figure 7A**, also contain information to reasonably distinguish the two classes, in green the adequate pellets and in red crosses the poor ones. It is seen that, qualitatively, to improve the characteristics of the poor objects to become adequate ones is to move in this plane to the left and up, that is, decrease the scores on the first principal component and increase the ones on the second principal component.

Figure 7B shows the loadings on the two principal components, blue for the first, orange for the second. Similar to the previous analysis with **Table 1**, with the loadings in the first three variables (percentage in the three different size ranges), the *manipulation* should be done clearly decreasing the values of the three variables. The loadings on the last three variables (measurements in different devices) is less clear, but, as the loadings on the second principal component are larger (in absolute value), TGA and TMA should be increased, and DSC decreased.

Nevertheless, questions still remain, such as how much of any one, in which proportion, whether any given relation must be maintained among variables, etc. These questions are answered in the solutions in **Table 1**, which define the joint combination among all input variables that guarantee a given property.

CONCLUSION

PLS-CM models are computed by setting a threshold decision limit in the space of predictions obtained when fitting a binary response that codifies the categories. This limit is selected based on the sensitivity and specificity that are needed in each specific application.

For one of such threshold values, the inversion of the fitted PLS model with a single response defines hyperplanes in both the

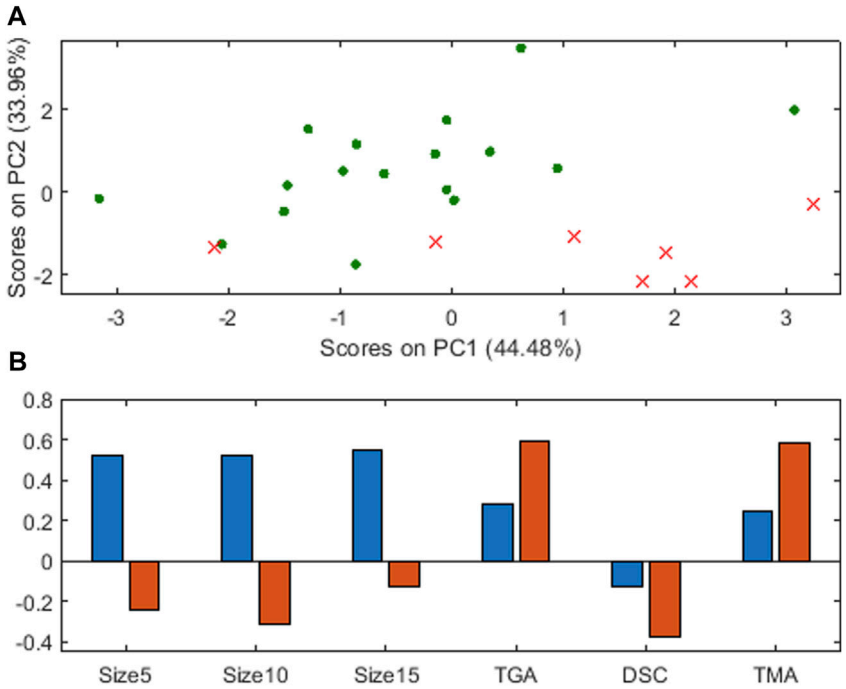

FIGURE 7 | Principal Component Analysis for plastic pellets **(A)** scores on the two first principal components, red crosses for poor, green points for adequate objects **(B)** loadings, blue for the first principal component, orange for the second.

latent and input spaces that, when observed in the input space, correspond to a kind of prototype of the object belonging to the boundary of the class-model being computed.

For cases where the classes are 'fail/no fail', ('defective/non-defective') a vector normal to the boundary hyperplane in the latent space defines one direction to move the scores along, exiting the 'fail' class to enter the other. In that case, the computed points in the domain corresponding to these scores provide information on how to modify the input variables to improve defective objects. Alternatively, if there is no need of working in the latent space, a direction with the same properties can be obtained directly in the domain by using the boundary hyperplane in the input space.

In that sense, the proposed procedure can be used as a diagnostic tool since it gives the characteristics of the predictor variables (input space) that allow the valid objects to be separated from the invalid ones. The characteristics are precisely those of the objects on the boundary hyperplane of the corresponding class-model. With PLS, contribution plots are common descriptive tools, that allow identification of the variables with the greatest relative influence to discriminate objects of a class in relation to the other. With respect to them, the boundary computed in the latent space with the proposed procedure provides, additionally, estimations of sensitivity and specificity. Furthermore, by "moving" this boundary to the input space, the information about the predictor variables is direct, for example, about how to modify them together pursuing a given goal.

The paper shows some possibilities of acting in specific situations, based on theoretical properties of both the fitted model and its inversion. The theoretical solutions developed in the present work apply in class-modelling contexts, where at least one 'alternative' class is adequately represented in the training set together with the target class, and the input variables (at least some of them) can be manipulated. In addition, good predictive PLS models need to be fitted and validated and, whenever possible, the predicted solutions should be experimentally validated.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://openmv.net/info/raw-material-characterization http://hdl.handle.net/10259/5753.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

The authors acknowledge the funding from the Spanish Agencia Estatal de Investigación (AEI, MINECO) through the research project CTQ 2017-88894-R, and from the Junta de Castilla y León *via* the research project BU052P20, both of them co-financed with European funds (FEDER).

REFERENCES

Avohou, T. H., Sacré, P. Y., Lebrun, P., Hubert, P., and Ziemons, E. (2021). A Probabilistic Class-Modelling Method Based on Prediction Bands for Functional Spectral Data: Methodological Approach and Application to Near-Infrared Spectroscopy. *Analytica Chim. Acta* 1144, 130e149. doi:10.1016/j.aca.2020.11.039

Bano, G., Facco, P., Meneghetti, N., Bezzo, F., and Barolo, M. (2017). Uncertainty Back-Propagation in PLS Model Inversion for Design Space Determination in Pharmaceutical Product Development. *Comput. Chem. Eng.* 101, 110–124. doi:10.1016/j.compchemeng.2017.02.038

Barbaste, M., Medina, B., Sarabia, L., Ortiz, M. C., and Pérez-Trujillo, J. P. (2002). Analysis and comparison of SIMCA models for denominations of origin of wines from de Canary Islands (Spain) builds by means of their trace and ultratrace metals content. *Analytica Chim. Acta* 472, 161–174. doi:10.1016/S0003-2670(02)00979-0

Barker, M., and Rayens, W. (2003). Partial Least Squares for Discrimination. *J. Chemometrics* 17, 166–173. doi:10.1002/cem.785

Brereton, R. G. (2009). *Chemometrics for Pattern Recognition*. United Kingdom: John Wiley & Sons. doi:10.1002/9780470746462

Brereton, R. G. (2011). One-class Classifiers. *J. Chemometrics* 25, 225–246. doi:10.1002/cem.1397

Brereton, R. G. (2015). Pattern Recognition in Chemometrics. *Chemometrics Intell. Lab. Syst.* 149, 90–96. doi:10.1016/j.chemolab.2015.06.012

Casale, M., Pasquini, B., Hooshyari, M., Orlandini, S., Mustorgi, E., Malegori, C., et al. (2018). Combining Excitation-Emission Matrix Fluorescence Spectroscopy, Parallel Factor Analysis, Cyclodextrin-Modified Micellar Electrokinetic Chromatography and Partial Least Squares Class-Modelling for green tea Characterization. *J. Pharm. Biomed. Anal.* 159, 311–317. doi:10.1016/j.jpba.2018.07.001

Chen, H., Lin, Z., and Tan, C. (2020). Application of Near-Infrared Spectroscopy and Class-Modeling to Antibiotic Authentication. *Anal. Biochem.* 590, 113514. doi:10.1016/j.ab.2019.113514

Dunn, K. (2020). *Process Improvement Using Data*. Release. Availableat:https://learnche.org/pid/PID.pdf.

Forina, M., Oliveri, P., Jäger, H., Römisch, U., and Smeyers-Verbeke, J. (2009). Class Modeling Techniques in the Control of the Geographical Origin of Wines. *Chemometrics Intell. Lab. Syst.* 99, 127–137. doi:10.1016/j.chemolab.2009.08.002

Forina, M., Oliveri, P., Lanteri, S., and Casale, M. (2008). Class-modeling Techniques, Classic and New, for Old and New Problems. *Chemometrics Intell. Lab. Syst.*, 93, 132–148. doi:10.1016/j.chemolab.2008.05.003

Hermane, A. T., Pierre-Yves, S., Pierre, L., Philippe, H., and Eric, Z. (2021). A Probabilistic Class-Modelling Method Based on Prediction Bands for Functional Spectral Data: Methodological Approach and Application to Near-Infrared Spectroscopy. *Analytica Chim. Acta* 1144, 130–149. doi:10.1016/j.aca.2020.11.039

Jaeckle, C., and MacGregor, J. (1996). Product Design through Multivariate Statistical Analysis of Process Data. *Comput. Chem. Eng.* 20, S1047–S1052. doi:10.1016/0098-1354(96)00182-2

Jaeckle, C. M., and MacGregor, J. F. (2000). Industrial Applications of Product Design through the Inversion of Latent Variable Models. *Chemometrics Intell. Lab. Syst.* 50, 199–210. doi:10.1016/S0169-7439(99)00058-1

Jaeckle, C. M., and MacGregor, J. F. (2000). Product Transfer between Plants Using Historical Process Data. *Aiche J.* 46, 1989–1997. doi:10.1002/aic.690461011

Lakshminarayanan, S., Fujii, H., Grosman, B., Dassau, E., and Lewin, D. R. (2000). New Product Design via Analysis of Historical Databases. *Comput. Chem. Eng.* 24, 671–676. doi:10.1016/S0098-1354(00)00406-3

Largoni, M., Facco, P., Bernini, D., Bezzo, F., and Barolo, M. (2015). Quality-by-Design Approach to Monitor the Operation of a Batch Bioreactor in an Industrial Avian Vaccine Manufacturing Process. *J. Biotechnol.* 211, 87–96. doi:10.1016/j.jbiotec.2015.07.001

Lay, D. C., Lay, S. R., and McDonald, J. J. (2016). *Linear Algebra and its Applications*. Fifth edition. Harlow, England: Pearson Education, Inc.

Marini, F., Magrì, A. L., Bucci, R., Balestrieri, F., and Marini, D. (2006). Class-Modeling Techniques in the Authentication of Italian Oils from Sicily with a Protected Denomination of Origin (PDO). *Chemometrics Intell. Lab. Syst.* 80, 140–149. doi:10.1016/j.chemolab.2005.05.002

Oliveri, P., and Downey, G. (2012). Multivariate Class Modeling for the Verification of Food-Authenticity Claims. *Trac Trends Anal. Chem.* 35, 74–86. doi:10.1016/j.trac.2012.02.005

Oliveri, P., Malegori, C., Mustorgi, E., and Casale, M. (2021). Qualitative Pattern Recognition in Chemistry: Theoretical Background and Practical Guidelines. *Microchemical J.* 162, 105725. doi:10.1016/j.microc.2020.105725

Ortiz, M. C., Herrero, A., Sánchez, M. S., Sarabia, L. A., and Íñiguez, M. (1995). The UNEQ, PLS and MLF Neural Network Methods in the Modelling and Prediction of the Colour of Young Red Wines from the Denomination of Origin 'Rioja'. *Chemometrics Intell. Lab. Syst.* 28, 273–285. doi:10.1016/0169-7439(95)80063-FAvailableat:http://hdl.handle.net/10259/5753 † Available at http://openmv.net/info/raw-material-characterization.

Ortiz, M. C., Saez, J. A., and Palacios, J. s. L. p. (1993). Typification of Alcoholic Distillates by Multivariate Techniques Using Data from Chromatographic Analyses. *Analyst* 118, 801–805. doi:10.1039/an9931800801

Ortiz, M. C., Sarabia, L. A., and Sánchez, M. S. (2010). Tutorial on Evaluation of Type I and Type II Errors in Chemical Analyses: From the Analytical Detection to Authentication of Products and Process Control. *Analytica Chim. Acta* 674, 123–142. doi:10.1016/j.aca.2010.06.026

Ottavian, M., Tomba, E., and Barolo, M. (2016). "Advanced Process Decision Making Using Multivariate Latent Variable Methods," in *Process Simulation and Data Modeling in Solid Oral Drug Development and Manufacture, Methods in Pharmacology and Toxicology*. Editors M. G. Ierapetritou and R. Ramachandran (New York, NY: Humana), 159–189. doi:10.1007/978-1-4939-2996-2_6

Pablos, J. L., Sarabia, L. A., Ortiz, M. C., Mendía, A., Muñoz, A., Serna, F., et al. (2015). Selective Detection and Discrimination of nitro Explosive Vapors Using an Array of Three Luminescent Sensory Solid Organic and Hybrid Polymer Membranes. *Sensors Actuators B: Chem.* 212, 18–27. doi:10.1016/j.snb.2015.01.103

Palací-López, D., Villalba, P., Facco, P., Barolo, M., and Ferrer, A. (2020). Improved Formulation of the Latent Variable Model Inversion–Based Optimization Problem for Quality by Design Applications. *J. Chemometrics*, e3230. doi:10.1002/cem.3230

Palací-López, D., Facco, P., Barolo, M., and Ferrer, A. (2019). New Tools for the Design and Manufacturing of New Products Based on Latent Variable Model Inversion. *Chemometrics Intell. Lab. Syst.* 194, 103848. doi:10.1016/j.chemolab.2019.103848

Pomerantsev, A. L., and Rodionova, O. Y. (2018). Multiclass Partial Least Squares Discriminant Analysis: Taking the Right Way-A Critical Tutorial. *J. Chemometrics* 32, e3030, doi:10.1002/cem.3030

Reguera, C., Sanllorente, S., Herrero, A., Sarabia, L. A., and Ortiz, M. C. (2019). Detection of Cold Chain Breaks Using Partial Least Squares-Class Modelling Based on Biogenic Amine Profiles in Tuna. *Talanta* 202, 443–451. doi:10.1016/j.talanta.2019.04.072

Rodionova, O. Y., Oliveri, P., and Pomerantsev, A. L. (2016). Rigorous and Compliant Approaches to One-Class Classification. *Chemometrics Intell. Lab. Syst.* 159, 89–96. doi:10.1016/j.chemolab.2016.10.002

Rodionova, O. Y., Titova, A. V., and Pomerantsev, A. L. (2016). Discriminant Analysis Is an Inappropriate Method of Authentication. *Trac Trends Anal. Chem.* 78, 17–22. doi:10.1016/j.trac.2016.01.010

Ruisánchez, I., Jiménez-Carvelo, A. M., and Callao, M. P. (2021). ROC Curves for the Optimization of One-Class Model Parameters. A Case Study: Authenticating Extra virgin Olive Oil from a Catalan Protected Designation of Origin. *Talanta* 222, 121564. doi:10.1016/j.talanta.2020.121564

Ruiz, S., Ortiz, M. C., Sarabia, L. A., and Sánchez, M. S. (2018). A Computational Approach to Partial Least Squares Model Inversion in the Framework of the Process Analytical Technology and Quality by Design Initiatives. *Chemometrics Intell. Lab. Syst.* 182, 70–78. doi:10.1016/j.chemolab.2018.08.014

Ruiz, S., Sarabia, L. A., Ortiz, M. C., and Sánchez, M. S. (2020). Residual Spaces in Latent Variables Model Inversion and Their Impact in the Design Space for Given Quality Characteristics. *Chemometrics Intell. Lab. Syst.* 203, 104040. doi:10.1016/j.chemolab.2020.104040

Ståhle, L., and Wold, S. (1987). Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem: A Monte Carlo Study. *J. Chemometrics* 1, 185–196. doi:10.1002/cem.1180010306

Tomba, E., Barolo, M., and García-Muñoz, S. (2012). General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products. *Ind. Eng. Chem. Res.* 51, 12886–12900. doi:10.1021/ie301214c

Tomba, E., Facco, P., Bezzo, F., and García-Muñoz, S. (2013). Exploiting Historical Databases to Design the Target Quality Profile for a New Product. *Ind. Eng. Chem. Res.* 52, 8260–8271. doi:10.1021/ie3032839

Tomba, E., Meneghetti, N., Facco, P., Zelenková, T., Barresi, A. A., Marchisio, D. L., et al. (2014). Transfer of a Nanoparticle Product between Different Mixers Using Latent Variable Model Inversion. *Aiche J.* 60, 123–135. doi:10.1002/aic.14244

Xu, L., Cai, C.-B., and Deng, D.-H. (2011). Multivariate Quality Control Solved by One-Class Partial Least Squares Regression: Identification of Adulterated Peanut Oils by Mid-infrared Spectroscopy. *J. Chemometrics* 25, 568–574. doi:10.1002/cem.1402

Xu, L., Shi, P.-T., Ye, Z.-H., Yan, S.-M., and Yu, X.-P. (2013). Rapid Analysis of Adulterations in Chinese lotus Root Powder (LRP) by Near-Infrared (NIR) Spectroscopy Coupled with Chemometric Class Modeling Techniques. *Food Chem.* 141, 2434–2439. doi:10.1016/j.foodchem.2013.05.104

Xu, L., Yan, S.-M., Cai, C.-B., and Yu, X.-P. (2013). One-class Partial Least Squares (OCPLS) Classifier. *Chemometrics Intell. Lab. Syst.* 126, 1–5. doi:10.1016/j.chemolab.2013.04.008

Zhao, Z., Wang, P., Li, Q., and Liu, F. (2019). Product Design for Batch Processes through Total Projection to Latent Structures. *Chemometrics Intell. Lab. Syst.* 193, 103808. doi:10.1016/j.chemolab.2019.07.007

Zhao, Z., Wang, P., Li, Q., and Liu, F., (2019). Input Trajectory Adjustment within Batch Runs Based on Latent Variable Models, *Ind. Eng. Chem. Res.* 58, 15562–15572. doi:10.1021/acs.iecr.9b03262

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fchem.2021.722090

Study of Variability of Waste Wood Samples Collected in a Panel Board Industry

*Manuela Mancini and Åsmund Rinnan**

Department of Food Science, Faculty of Science, University of Copenhagen, Frederiksberg C, Denmark

Waste wood is becoming an appealing alternative material to virgin wood, and the main drivers are the increased demand for waste wood by the panel industry, the introduction of renewable energy policies, and the waste framework directive. In fact, the use of waste wood as a secondary resource is favored over both landfills and combustion. The best reuse and cascading use of the material are linked to its characteristics. That is why it is important to know the chemical composition and the variation in the properties of such a heterogeneous material. In this article, a sampling study was carried out in a panel board company located in the northern part of Italy. In order to investigate the heterogeneity of waste wood, all samples have been analyzed by near-infrared spectroscopy. Nested analysis of variance and principal component analysis have been used to evaluate the heterogeneity and the variation in sample properties. The approach gives information about how to ensure representative measurements and efficiently describe the variability of the material. The results suggest that it is important to have replicates or at least two subsamples for each lot and then measure each of these with at least 100 scans, in order to get representative measurements and describe the variability of the material. The determination of waste wood composition and variability is the focal point for improving the sorting process and increasing the reuse of waste wood, avoiding expensive landfills and risks for human health and the environment.

Keywords: sampling, variability, NIR spectroscopy, nested analysis of variance, heterogeneity, PCA

OPEN ACCESS

Edited by:
Federico Marini,
Sapienza University of Rome, Italy

Reviewed by:
Rosalba Calvini,
University of Modena and Reggio Emilia, Italy
Ingunn Burud,
Norwegian University of Life Sciences, Norway

****Correspondence:***
Åsmund Rinnan
aar@food.ku.dk

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *08 June 2021*
Accepted: *28 June 2021*
Published: *26 July 2021*

Citation:
Mancini M and Rinnan Å (2021) Study of Variability of Waste Wood Samples Collected in a Panel Board Industry. Front. Chem. 9:722090. doi: 10.3389/fchem.2021.722090

INTRODUCTION

Wood is one of the oldest and highly exploited resources in several sectors (e.g., pulp, construction, and energy), but it is also a limited resource (Rettenmaier et al., 2008). Since the 1970s, wood consumption has increased continuously, and it is expected to do so in the future (FAO/ECE, 2012). At the end of the life cycle, wood utilization produces waste wood (WW). The term indicates wood or wood-containing post-consumer and post-use products from different sectors (packaging, furniture, construction and demolition, and industrial and commercial sectors) (Edo et al., 2016). A study has estimated that the European Union generates 50 million cubic meters of wood waste each year (Mantau, 2012), and nowadays, a large amount remains unused (Hakala, 2012).

The most relevant drivers of the growth of the waste wood trade are related to the increased demand for waste wood by the panel board industry (Mazzanti and Zoboli, 2013) (Bergeron, 2016). The European Union is promoting the reuse and recycling of the materials over the landfill (Waste Framework Directive, 2008/98/EC, European Parliament 2008) (Commission of the European

Communities, 2008) and has introduced European renewable energy policies for mitigating greenhouse gas emissions (Röder and Thornley, 2018).

Because of the various sources of origin, WW composition presents high heterogeneity (Huron et al., 2017). In addition, it should be taken into account that its chemical composition, quality classes definition, and degree of contamination also change according to the countries and their different laws (Edo et al., 2016). Consequently, identifying the best-suited application and possible end-users is related to the assessment of the WW composition and quality characteristics.

Some studies have already examined the characterization of waste wood materials. Edo et al. have investigated the waste wood variability across time (Edo et al., 2016). They collected five hundred samples from an industrial heating plant during nine years and performed lab analysis to assess the material heterogeneity. The concentrations of the examined contaminants varied according to the sampling method, demonstrating the variability of the material. In another study, Moreno and Font have carried out a complete characterization of furniture waste wood and studied the differences in thermochemical conversion by performing pyrolysis tests (Moreno and Font, 2015). Huron et al. have performed an extensive characterization of various treated waste wood to evaluate their heterogeneity and assessment of suitability with combustion processes. Different samples were collected, including waste wood mixtures, specific waste wood classes, and untreated wood for comparison. Some parameters, such as heating value and composition in C, H, and O, did not vary significantly compared to those of untreated wood, while minor elements showed differences in relation to the chemical treatments of waste wood (Huron et al., 2017). Faraca et al. have investigated the quality of wood waste and pointed out the importance of physical and chemical impurities in waste wood to improve recyclability (Faraca et al., 2019). In some other studies, waste wood has been extensively characterized for properties relevant to combustion, and the suitability of waste wood as feedstock in combustion units has also been tested (Tatàno et al., 2009) (Gehrmann et al., 2020). It was demonstrated that waste wood contained higher ash content and metals than natural virgin wood and that the chemical and physical characteristics of the different types of waste wood play a role in choosing the best use of the material as a feedstock for energy recovery. To the best of our knowledge, there are no studies examining the variability of waste wood samples using fast analytical technologies, such as Near-Infrared Spectroscopy (NIRS). In fact, Vrancken et al. have listed and reviewed different studies where sensors and modern sorting technologies were developed for recycling plants to improve/optimize the material sorting and/or measure critical waste characteristics (Vrancken et al., 2017). The optical sensors could be used to obtain real-time information about waste characteristics, which helps in selecting the best waste processes, proving to be a useful tool for stakeholders.

As it can be seen by the references cited above regarding the heterogeneity of WW, the assessment of waste wood variability is of utter importance for improving the waste management in terms of sorting and related best reuse of the material and avoiding health and environmental issues at the end of the life cycle of wood utilization. Consequently, in the current study, WW samples have been collected during a sampling in a panel board industry located in the northern part of Italy. All of the samples have been analyzed using NIRS following strict sampling protocols. Our aim is to show how the variability of WW can be characterized, both within and between each sample. Furthermore, we will show how this information can directly be implemented and used for the increased reuse of WW. Throughout the manuscript, we have decided to include information about the bound water content. This is a very important quality attribute for waste wood and is one of the most important parameters influencing the NIR analysis.

To address this issue, the following data analyses have been carried out: 1) nested analysis of variance for investigating the variability at each sampling level; 2) Principal Component Analysis (PCA) as a rapid tool for the assessment of the material variability; 3) repeated nested analysis of variance considering a subset of the original data. The first two give a good overview of the variability in and between the lots, while the latter is a good procedure for finding the most suitable sampling procedure. Obtaining information about the number of samples and replicates to be performed during sampling is fundamental to guarantee an accurate and successful application of a NIR sensor classification tool, especially when dealing with heterogeneous material. In fact, efficient quality control with a high degree of accuracy is imperative for its use in the industry. In order to meet these requirements, it is essential to have detailed information on how to perform the sampling procedure in practice, out in the field.

MATERIALS AND METHODS

Collection of Waste Wood Samples

Waste wood samples were collected in a large panel board company located in the northern part of Italy (Lombardy region) over two days of sampling (February 18–19, 2020). The material was collected in the earliest phases of the production stream, precisely after the first step of cleaning (removal of stone, iron, and other heavy materials by washing) and grinding (reducing the particle size of the material to around 5 cm).

In order to get representative samples, a sampling plan has been defined based on the EN-15442:2011 standard (CEN, 2011). The sampling was carried out from a static lot. The material was taken every hour from the production stream in an external unloading tank for a total of 16 lots. As the incoming material is of variable quality, it is also assumed that the quality and variability within the 16 lots are different. For each lot, four representative samples were randomly taken from different locations in the lot (Mancini and Rinnan, 2021). The samples were collected using a sampling scoop for a total volume of 10 L; afterward, they were sent to the lab for the next lab and near-infrared analyses. In short, a total of 64 samples (16 lots x 4 samples from each lot) were obtained.

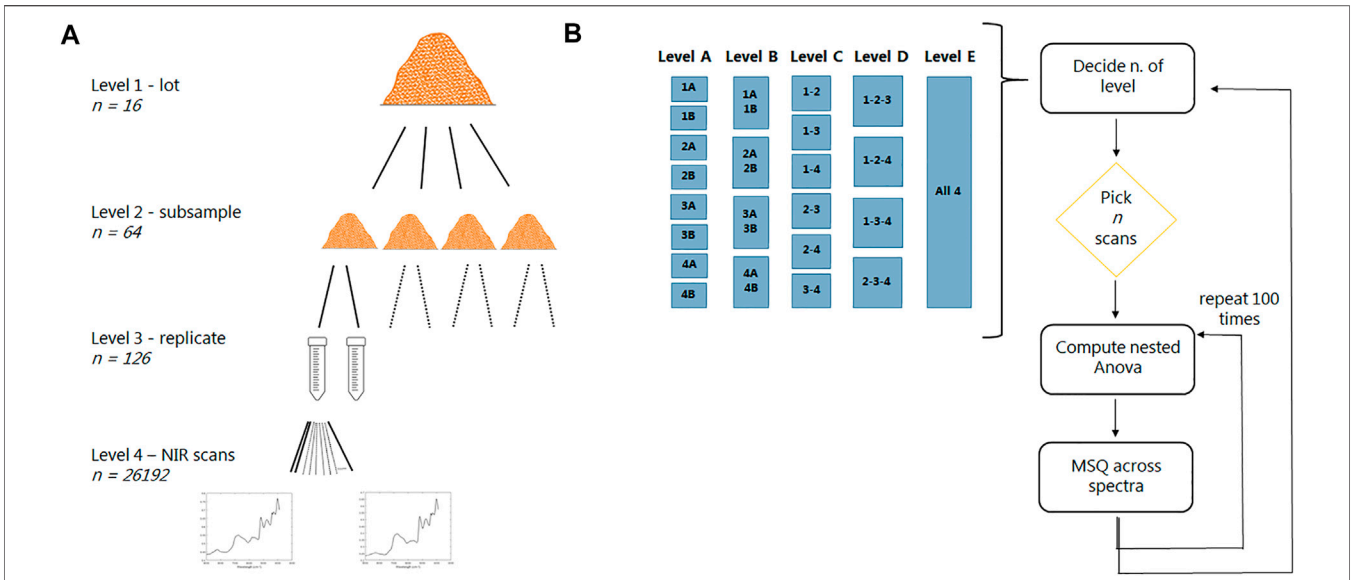

FIGURE 1 | A) Schematic representation of the hierarchical sampling procedure. Please note that there "only" are 126 samples at level 3 due to some unfortunate problems during the NIR analysis. **(B)** Schematic representation of the data analysis procedure (*n* is the number of random scans selected). Level A, each subsample and replicate are represented equally; level B, each subsample is represented equally; level C, groups of two subsamples are represented equally; level D, groups of three subsamples are represented equally; level E scans are selected freely among all subsamples and replicates.

The hierarchical sampling procedure from lot level down the individual NIR scans is presented in **Figure 1A**.

Sample Preparation

The samples have been prepared for the successive lab analysis using the technical standard UNI 15443. The sample preparation consists of a combination of sample division and particle size reduction, carefully avoiding loss in representativeness and sample composition during each step of the preparation.

Firstly, the sample has been stabilized by drying for at least 24 h not exceeding 40°C. The dried samples have been divided using a quartering process. The quartering process means that the sample is piled, divided into four, and the two opposite fractions are combined. The process of piling, dividing, and combining is repeated until the wanted sample size is achieved. Subsequently, the sample particle size has been reduced to below 5 mm using a cutting mill (mod. SM 2000; RETSCH). This material has been used for the near-infrared analysis. Finally, part of the material was further reduced to under 1 mm for the bound water content analysis. Before the NIR and lab analyses, the samples have been stored in hermetically closed plastic bags in a room with controlled temperature and humidity.

Bound Water Content

The analytical methodology adopted for the determination of bound water content (BWC) follows the standard ISO 18122: 2015. The parameter has been determined using a thermo-gravimetric analyzer (mod. 701 Leco). In detail, the sample has been air-dried to a controlled temperature (105 ± 10°C) using a muffle furnace and has been weighted until constant mass is achieved. The loss in mass has been used to calculate BWC. Each BWC value was estimated twice per sample, and the average of these estimates was used in the subsequent data analysis.

The BWC parameter has been chosen because it is easy to determine and it is important for investigating the influence of moisture in the variability of waste wood material.

Near-Infrared Data

All waste wood samples were analyzed using a Quant FT-NIR spectrophotometer (Q-Interline A/S, Tølløse, Denmark) provided with the patented spiral sampler (Spiral Sampler, Q-Interline A/S, Tølløse, Denmark). The spiral sampler scans a total of 375 cm^2 surface, improving the representativeness of heterogeneous samples.

The instrument is equipped with a quartz halogen lamp as a light source and an InGaAs detector. The samples were acquired in diffuse reflectance mode and were kept in rotation during the acquisition by means of the spiral sampler. Near-infrared spectra were recorded in the range from 14,885 to 3,700 cm^{-1} (equals to 670–2,700 nm) with a maximum of 210 scans per sample/tube and a spectral resolution of 8 cm^{-1}. Instead of averaging all scans, each scan was stored individually, meaning that we get a good estimate of the heterogeneous nature of each sample. It is important to note that the start of each measurement had to be performed manually for each sample. Thus, some of the scans at the beginning of one series had air/plastic lids instead of the wood sample, which needed to be removed before data analysis. Random effects associated with the instrument or environment were removed by acquiring a blank spectrum, by measuring Spectralon, at the beginning of the analysis session. (However, we later realized that we should have measured this Spectralon sample several times during the measurement session, despite the whole process only taking approximately 6 h; see *Nested Analysis*

TABLE 1 | The degree of freedom computation for the nested ANOVA. N_4 is the total number of scans.

Levels	Degrees of freedom (D)	Computed degrees of freedom
Lot	$D_1 = N_1–N_0$	$D_1 = 16–1$
Subsample	$D_2 = N_2–N_1$	$D_2 = 64–16$
Replicate	$D_3 = N_3–N_2$	$D_3 = 126–64$
Scan	$D_4 = N_4–N_3$	$D_4 = 26{,}192–126$
Total	$D_{Tot} = N_4–N_0$	$D_{Tot} = 26{,}192–1$

of Variance.) Spectra were collected at room temperature and in duplicate for each sample in random order. The resulting dataset consists of 26,192 observations at 1,091 wavenumbers, as two tubes were only measured once due to an unfortunate computer error[1] only realized after arriving back at the University. Consequently, level 3 of the replicate consists of 126 objects instead of 128 (see **Figure 1A**). The measurements were completed on the same day, taking a total of approximately 6 h.

Nested Analysis of Variance

Considering the multi-stage approach of the sampling procedure, a nested analysis of variance (ANOVA) was computed in order to investigate the statistical differences between 1) the different lots (level 1); 2) the subsamples within each lot (level 2); 3) the two replicates within each subsample (level 3); 4) the scans within each subsample replicate (level 4).

For each sampling level, the sum of squares (SSQ) and the average of the sum of squares (MSQ) were computed (Sahai and Ageel, 2000). In detail, SSQ was computed as follows:

$$SSQ_{lvl} = \sum_{n=1}^{N_{lvl}} \left(x_{n,lvl} - \overline{x}_{lvl-1}\right)^2.$$

Moreover, MSQ was computed as follows:

$$MSQ_{lvl} = SSQ_{lvl} / \left(N_{lvl} - N_{lvl-1}\right),$$

where lvl is the current level, $x_{n,lvl}$ corresponds to the observations/average at the current level, and N_{lvl} is the number of unique measurement points at each level (e.g., number of lots for the uppermost level). The term $(N_{lvl}–N_{lvl-1})$ thus corresponds to the degrees of freedom within each level, where *lvl-1* refers to the previous sampling level. In this way, both SSQ and MSQ are calculated to represent the individual contributions from each level of the sampling. **Table 1** summarizes the computation of the degrees of freedom at each level. The MSQ was calculated for each wavenumber independently in order to investigate which wavenumbers are causing the variability at each level.

Before any variance analysis, the NIR spectra have been preprocessed by Multiplicative Scatter Correction (MSC) (Martens et al., 1983) in order to reduce the light scattering effects (Rinnan et al., 2009).

[1]For two sample replicates, the computer froze without saving the collected data.

Deciding the Best Sampling Procedure

In order to find the best sampling procedure to describe the variability of waste wood material, the nested analysis of variance was computed again considering the setup reported in **Table 2**. Based on the total number of scans for each of the tested levels, the nested analysis of variance was computed again, taking n random selected scans, and the procedure was repeated one hundred times for each of the new levels.

This is important, as how to perform the sampling procedure in the real world is of utter importance for the usefulness of applying advanced sensors to the system of WW reuse. Here, we investigated how the variability of the lot is described by increasing the number of subsamples and/or scans. We have decided to perform this at different levels of constraints, efficiently showing the effect of each of these constraints on the subsequent sampling conclusion. In detail, at level A, each subsample and replicate are represented with the same number of scans; at level B, each subsample is represented with the same number of scans; at level C, two subsamples are grouped together; at level D, three subsamples are grouped, while level E picks scans at random across all subsamples and replicates. Differences and similarities between these different approaches will aid in finding the optimal sampling procedure, with regard to both the number of subsamples and replicates and number of scans necessary to cover the variability. A schematic representation of the data analysis procedure is displayed in **Figure 1B**.

Multivariate Data Analysis

Principal Component Analysis (PCA) (Wold et al., 1987) has been computed using two different datasets: the mean-centered MSQ values of the nested analysis of variance and the preprocessed NIR absorbance values of the waste wood samples.

The former was performed in order to investigate similarities in the variability among the lots at the different sampling levels. We are well aware that this is an untraditional use of PCA, but it gives a nice and quick overview of how the variability varies between the lots. The latter was performed in order to explore the variability of waste wood and search for differences/groupings among the lots at each sampling level. In this latter case, the computation was carried out on the MSC pretreated and mean-centered data. In order to search for differences among the lots and investigate the variability within each lot, a confidence ellipse is drawn around each lot. This ellipse is calculated based on a local PCA on the scores, indicating the direction and extent of variability for each lot individually. Each ellipse was calculated using the mean score values as the center, and the standard error of each variability direction as the radius of the ellipse. The loading plot of the two first PCs was investigated to identify the compounds associated with the variability of the waste wood samples and the variability within the lots.

Both the multivariate data analysis and the nested analysis of variance have been computed using Matlab software (ver. MATLAB R2019b, The MathWorks) with in-house functions based on existing algorithms.

TABLE 2 | Setup for the computation of the nested analysis of variance for deciding the best sampling procedure.

	Setup	Total n. of scans	n. of randomly selected scans (*n*)
Level A	A single subsample with replicates as two different subsamples	210	25, 50, 75, 100, 125, 150
Level B	A single subsamples with replicates together	420	25, 50, 75, 100, 150, 200, 250, 300
Level C	Two subsamples	840	25, 50, 100, 150, 200, 300, 400, 600
Level D	Three subsamples	1,260	25, 50, 100, 150, 250, 400, 600, 900
Level E	All 4 subsamples	1,680	25, 50, 100, 150, 250, 500, 750, 1,000

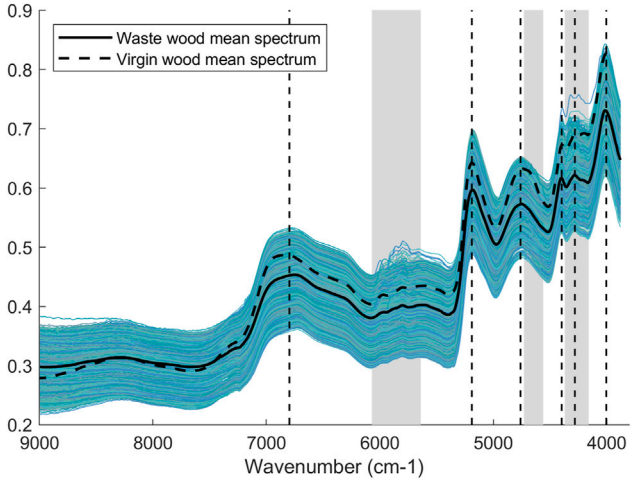

FIGURE 2 | All the spectra of waste wood samples with the mean spectrum of all the waste wood samples highlighted with a solid black line, and the mean spectrum of virgin wood samples highlighted with a dotted black line. Vertical dotted lines refer to the most relevant wavenumbers and are also reported in **Table 3**. The grey areas highlight the spectral areas mostly associated with glue compounds.

RESULTS AND DISCUSSION

Spectra

A total of 55 spectra was detected as either being due to the plastic lid or air, and was deleted before any further data analysis. Furthermore, wavenumbers lower than 3,880 cm^{-1} and greater than 9,000 cm^{-1} were removed as the data were either deemed noisy or containing very limited information. The new dataset thus consists of 26,192 scans measured at 664 wavenumbers. **Figure 2** illustrates the plot of all the spectra of waste wood samples and their mean spectrum highlighted with a solid black line. Because of the light scattering, all the spectra have been preprocessed with MSC before any further data analysis. In addition, in order to investigate the differences between waste wood and virgin wood, the mean spectrum of virgin wood samples was added to **Figure 2** as a dotted black line. The virgin wood samples have been acquired during a previous study (Toscano et al., 2017). The most relevant wavenumbers in the two spectra are marked with vertical dotted lines and reported in **Table 3**. As it can be noted, the same spectral wavenumbers selected for the mean spectrum of waste wood samples can also be found in the mean spectrum of virgin wood samples, demonstrating similar chemical composition. By inspecting the waste wood spectra, we can clearly see that some spectral areas include observations with deviating trends: 6,070–5,640 cm^{-1}, 4,730–4,560 cm^{-1}, and 4,370–4,160 cm^{-1}, strongly indicating that it will be possible later to classify the samples between virgin wood and treated wood. These spectral areas are probably associated with glue compounds related to the composite wood materials or plastic materials contained in the waste wood.

As reported by Lian et al., the band at 5,911 cm^{-1} corresponds to the characteristic absorption peak of C-H in methyl glycol,

TABLE 3 | Near-infrared absorption band assignment associated with the most important wavenumbers (str.: stretching; def.: deformation; OT: overtone; L: lignin; H: hemicellulose; C: cellulose).

Measured wavenumber (cm^{-1})	Bibliography wavenumber (cm^{-1})	Compound	Assignment
6,797	6,790	C	1st OT O-H str. Schwanninger et al. (2011)
	6,800	H	1st OT O-H str. Schwanninger et al. (2011)
5,189	5,220–5,150	Water	O-H asymmetric str. + O-H def. Of H_2O Schwanninger et al. (2011)
4,760	4,762	C	O-H and C-H def. + O-H str. Sandak et al. (2010)
	4,780–4,760	C	O-H and C-H def. + O-H str. Schwanninger et al. (2011)
	4,890–4,620	C	O-H str. + C-H def. Schwanninger et al. (2011)
4,397	4,392	C	O-H str. + C-C str. and/or C-H str. + C-H def. Schwanninger et al. (2011)
4,281	4,288	H	C-H str. + C-H def. Schwanninger et al. (2011)
	4,280	C	C-H str. + C-H def. Schwanninger et al. (2011)
	4,280	L	C-H str. + C-H_2 def. Schwanninger et al. (2011)
	4,282	C	C-H str. + C-H_2 def. combination band (and 2nd OT of C-H_2 str.) Hein et al. (2011)
4,004	4,014	L	C-H str. + C-C str. Schwanninger et al. (2011)

while the peak at 5,996 cm^{-1} corresponds to C-H on the benzene ring (Lian et al., 2020). In general, the spectral range between 6,700 and 6,330 cm^{-1} corresponds to the characteristic absorption of methyl glycol, indicating that it is related to glue/plastic compounds. Furthermore, these results were confirmed in a study by Workman and Weyer, where the assigned peaks at 5,847 and 5,975 cm^{-1} are attributed to C-H from methyl of glue, while the band at 5,624 cm^{-1} was assigned as the second overtone of CH methylene of glue (Workman and Weyer, 2007). The band at 5,805 cm^{-1} was assigned to the 1st overtone of C–H stretching of methyl and methylene structures of glue (Tomlinson et al., 2006). Regarding the second spectral area, the absorption band at 4,440 cm^{-1} is related to the CH_2 combination of methylol group (Dessipri et al., 2003). In another study, Hein et al. have investigated the physical and mechanical properties of agro-based particleboards by NIR spectroscopy and assigned the peak at 4,587 cm^{-1} to symmetric NH stretching and NH_2 rocking and/or 2nd overtone of amide I and amide III (Hein et al., 2011). Moreover, the relationship between this spectral region and wood composite materials is confirmed by the peak at 4,617 cm^{-1}, associated with NH_2 species from urea (Dessipri et al., 2003), and 4,550 cm^{-1} assigned to NH symmetrical stretching and NH bending combination bands (Henriques et al., 2012). Lastly, the region from 4,370 to 4,160 cm^{-1} is assigned to the combination band of NH_2 and CH bonds.

The knowledge of the chemical composition of the waste wood and the inspection of the spectra are important steps for defining the waste wood quality and, accordingly, the best reuse of the material. The difference between the mean spectra of virgin wood and waste wood indicates that some absorption bands of the two materials are not exactly the same, suggesting that a classification model for separating the material according to its best reuse would perform well.

Bound Water Content Analysis

A descriptive statistic of the BWC has been carried out. The 64 waste wood samples analyzed have a mean = 8.0%, standard deviation = 0.7%, max value = 11.1%, and min value = 7.0%. Thus, the parameter has a range of 4.1%. An outlier sample in BWC values has been detected using Tukey's test. The test identifies the possible outliers of the samples falling outside the Q1 - 1.5 · IQR (interquartile range) or the Q3 + 1.5 · IQR limits; Q1 and Q3 are first and third quartiles, respectively. For this study, limits that are more conservative have been used: Q1 - 3.0 · IQR or Q3 + 3.0 · IQR. The lot with the highest variability in BWC was lot 12 (range of 2.92%), and the one with the lowest was lot 15 (range of 0.24%). The average lot variability in BWC was 0.79%. The reported results are useful for the discussion of the successive outcomes (see *Nested Analysis of Variance* and *PCA*).

A nested analysis of variance was also computed. The MSQ value is higher at lot level (MSQ = 3.20), decreases considerably at subsample level (MSQ = 0.39), and drops even further at the replicate level (MSQ = 7.11 e-4). The results confirm that by increasing the number of samples, the variability in their moisture content also decreases.

Nested Analysis of Variance

The nested ANOVA was computed on the dataset consisting of 26,192 observations and 664 wavenumbers. The analysis of variance has been computed on the spectra preprocessed with MSC on all the sampling levels. **Figure 3** (A, B, and C) shows the plot of the MSQ values plotted against the wavenumbers at the different sampling levels. As expected, the variability is higher at the lot level (**Figure 3A**) and lowest at the scan level (**Figure 3C**). Unexpectedly, the variability at the subsample level is lower than at the replicate level (**Figure 3B**) and will therefore be investigated further. The subsample and lot lines have a similar trend indicating that the variability is affected by the same wavenumbers. To better investigate this, **Figure 3D** shows the normalized MSQ values at the lot and subsample levels. The two lines differ for some wavenumbers. In detail, the lot level has two higher and sharper peaks at 5,609 cm^{-1} and 4,791 cm^{-1}. The former is assigned to 1st overtone of CH_2 stretching of cellulose, while the latter is related to OH stretching + OH and CH deformation of cellulose and hemicellulose (Schwanninger et al., 2011). Both lines have a high absorption band at 5,177 cm^{-1} (O-H stretching and O-H deformation of H_2O) and 6,943 cm^{-1} (first overtone O-H stretching of water), indicating that the bound water content plays a role in the variability of the waste wood material, as also confirmed by the results reported in *Bound Water Content*. The subsample level has two noisy areas: between 7,400 and 7,050 cm^{-1} and between 5,500 and 5,200 cm^{-1}. Finally, we can observe small "vibrations" in the areas 6,070–5,640 cm^{-1} and 4,370–4,160 cm^{-1}, confirming our previous conclusions (see **Figure 2**).

Figure 4 shows the plot of the MSQ values for each of the 16 lots at the subsample and replicate levels, respectively. Basically, the nested ANOVA has been computed again for each of the 16 lots individually, and the MSQ values have been estimated at both the subsample and replicate levels of sampling. This gives an indication about the variability among the different lots. In **Figure 4A**, it can be noted that the lots with higher variability are lots 12, 14, and 11. In detail, lot 12 has a higher variability at wavenumber 5146 cm^{-1}, while MSQ values of lots 14 and 11 are higher on all the other wavenumbers. The band at 5,146 cm^{-1} is assigned to O-H asymmetric stretching and O-H deformation of H_2O (Schwanninger et al., 2011), indicating that the higher variability of the lot is probably related to a higher BWC in some samples. In fact, lot 12 contains the sample with the highest BWC value (11.1%) (see *Bound Water Content*).

Figure 4B reports the variability between the two replicates of the subsamples within each lot. Lots 16, 10, and 7 (in descending order) have higher MSQ values. The MSQ values of lot 11 are quite different, resulting in a particular shape/trend of the variance line, more similar to a spectrum. All the other lots show higher variability in the wavenumbers between 7,400 and 7,050 cm^{-1} and between 5,500 and 5,200 cm^{-1}. The two spectral regions are quite noisy and the peaks do not probably contain relevant information. However, they could be related to the detector drift since, unfortunately, only one reference spectrum at the very beginning of the analysis was acquired (see *Bound Water Content*). The differences in the variability among the lots could be explained by calculating the distance in

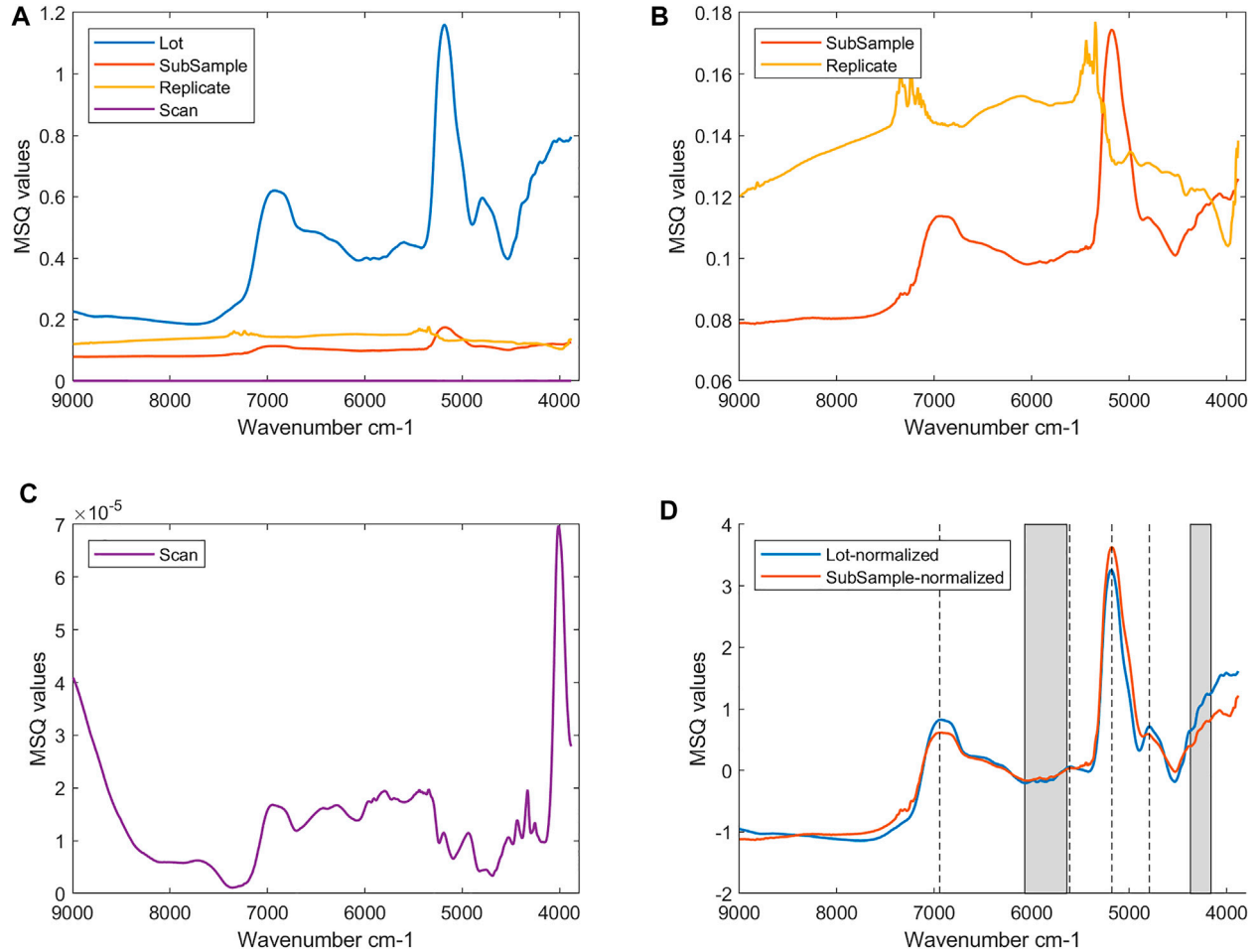

FIGURE 3 | The MSQ values of the nested ANOVA: **(A)** all sampling levels, **(B)** only subsample and replicate levels, **(C)** only scan level, and **(D)** the normalized MSQ values at the lot and subsample sampling levels (the grey areas highlight the spectral areas mostly associated with glue compounds).

the PCA score plot (see *PCA* section) between the two replicates at the subsample level. **Figure 4C** shows the lots colored according to the replicates distance and we can conclude that the longer the distance between the two replicates in the PCA score plot, the higher the MSQ values and, consequently, the variability at the replicate level.

PCA

In order to get a quick overview of how the variability changes between the different lots, a PCA was carried out using the MSQ values of the nested analysis of variance, computed individually for each lot, at both subsample and replicate levels of sampling. The score plot confirms the results of the nested ANOVA, but with increased clarity. At the subsample level (**Figure 5A**), the lots with the most deviating scores are 11, 12, and 14, while at the replicate level (**Figure 5B**), lots 7, 10, 11, and 16 deviate the most compared to the remaining lots.

The loadings were investigated to understand what variables are responsible for the separation of the lots in the PCA scores plots. At the subsample level (**Figure 5C**), both the first and second PCA loadings show two main bands at around 6,950 cm^{-1} and 5,150 cm^{-1}. Both bands are related to the overtone of O-H stretching bonds (Schwanninger et al., 2011), confirming the results of the nested analysis of variance and what already was stated during the discussion of **Figure 4**. At the replicate level (**Figure 5D**), the first loading shows the same noisy areas (i.e., between 7,400 and 7,050 cm^{-1} and between 5,500 and 5,200 cm^{-1}), as shown in **Figure 4**. The second loading contains information related to the variability of lot 11. It is important to note that the PCA analysis on MSQ values confirmed the outcomes of the nested analysis of variance and is an efficient alternative, giving a nice and quick overview of how the variability varies among the different lots.

To explain why some replicates/subsamples present a higher variability than others, the average spectra at the subsample level, after preprocessing with MSC, have been taken into account. A PCA was computed based on these 16 x 8 spectra, and as it can be noted in the PCA score plot (**Figure 6A**), the samples seem to be spread across the whole score space without any clear groupings between them. However, by closer inspection, there is some trend

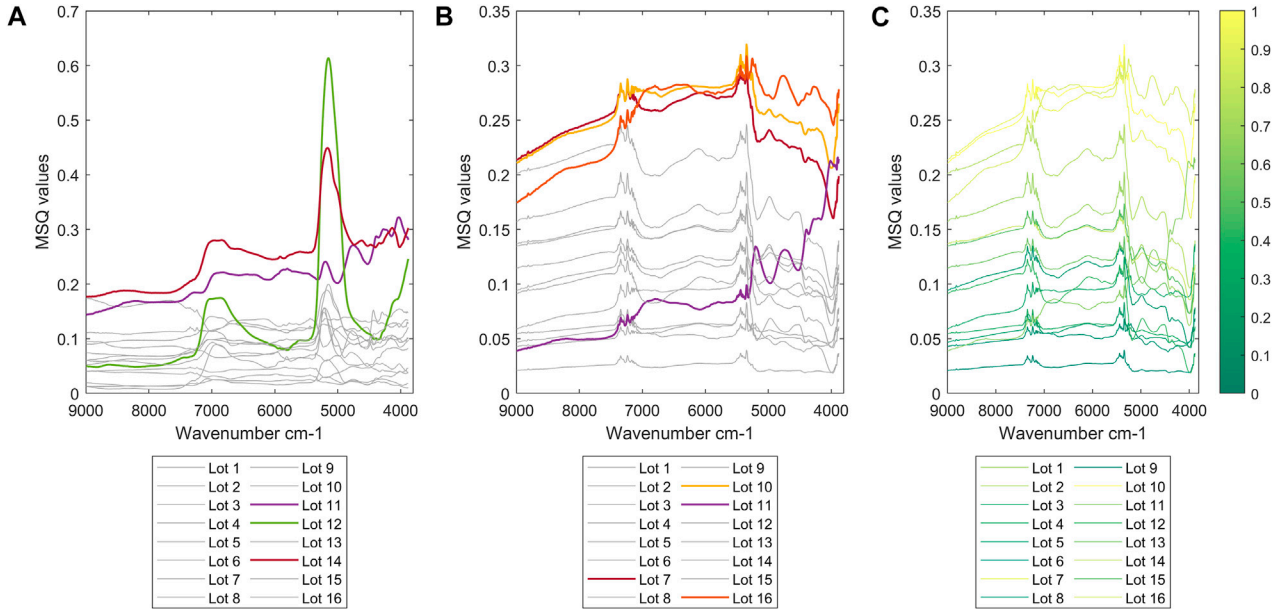

FIGURE 4 | The MSQ values of the nested analysis of variance computed within each lot **(A)** at subsample level and **(B)** replicate sampling level. MSQ values at the replicate sampling level are also colored according to the distance in the PCA score plot between the two replicates of a sample **(C)**.

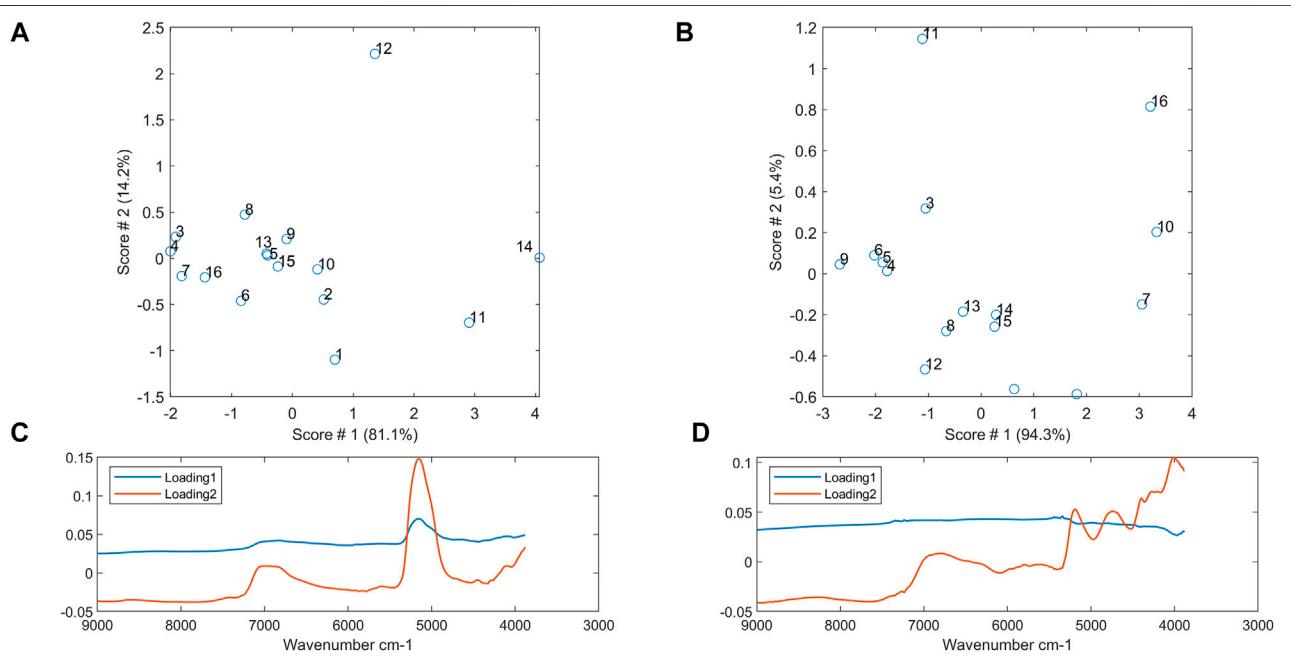

FIGURE 5 | PCA score plot of MSQ values of the nested analysis of variance at **(A)** subsample level and **(B)** replicate level. PCA loading plot of MSQ values of the analysis of variance **(C)** at subsample level and **(D)** replicate level.

in the distribution of the samples according to the lot; i.e., the samples with higher BWC are located in the bottom right part of the score plot (i.e., lots 4 and 5 and some samples of lot 12).

In order to get a clearer picture of the differences among the lots, confidence ellipses were computed using the standard error for each lot. The score plot of the two first PCs clearly shows some groupings among the lots (**Figure 6B**, please note that this is the same plot as **Figure 6A**, but now with confidence ellipses instead of each individual subsample being plotted). Lot 15 is clearly different from the others (lower range in BWC). Lots 7, 13, 14,

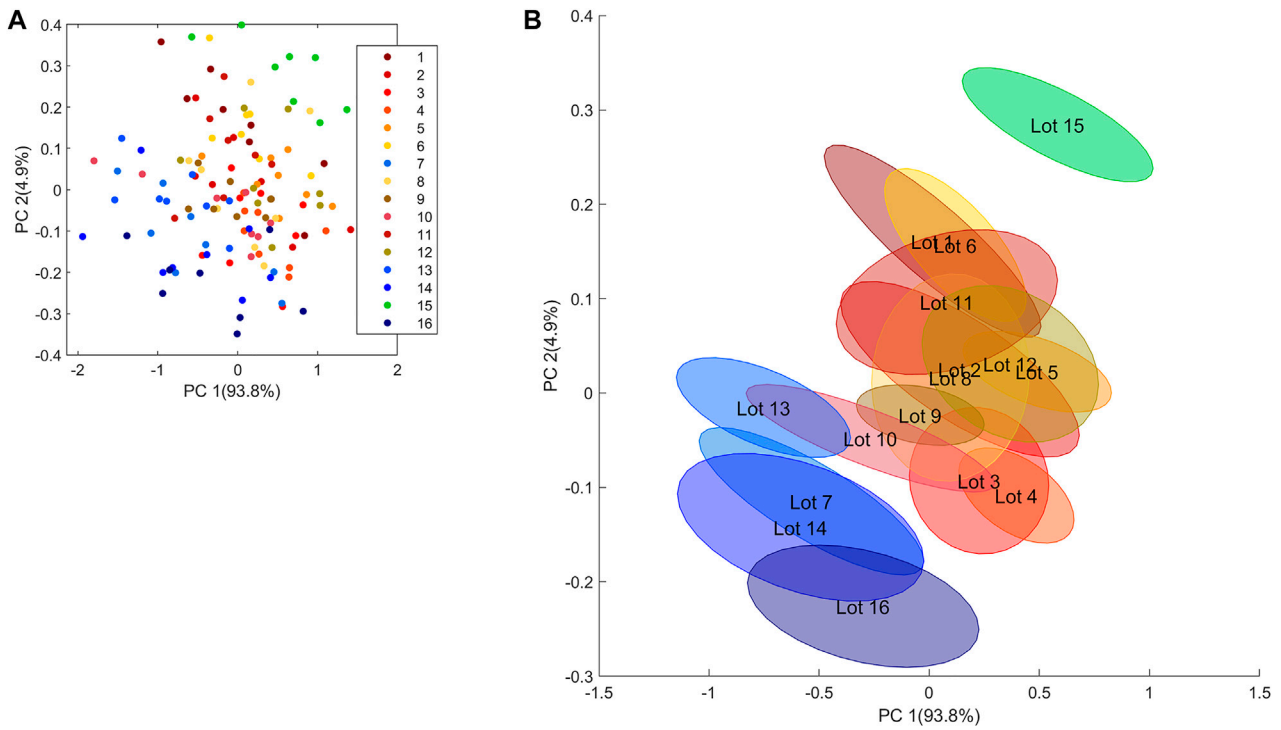

FIGURE 6 | PCA on the average spectra at the subsample level. Each lot has been colored in shades of blue, red, and green. The color scheme is based on the position of the lots in the score space. **(A)** The first two scores colored according to the lot number and **(B)** PCA score plot of the waste wood sample with standard error ellipses for each lot.

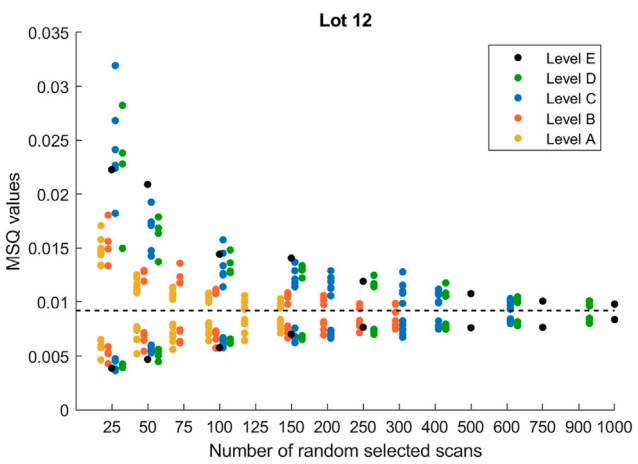

FIGURE 7 | Variability in the MSQ values of nested analysis of variance changing the number of randomly selected scans and the number of subsamples. The dotted horizontal black line represents the estimated overall variability of lot 12. Level A, each subsample and replicate are represented equally; level B, each subsample is represented equally; level C, groups of two subsamples are represented equally; level D, groups of three subsamples are represented equally; level E scans are selected freely among all subsamples and replicates. For a detailed description of the different levels, please refer to **Figure 1B**.

and 16 are located at the bottom left part of the PCA score plot. All the other lots are close to each other and located in the central part of the score plot, indicating that their composition/variability is very similar. The size of the ellipses confirms that the lots with the highest variability are lots 8, 11, and 12, and 14. Lots 4 and 9 have the lowest variability, confirming once again the results of the nested ANOVA.

Deciding the Best Sampling Procedure

For the practical implementation of a NIR sensor classification tool in the WW industry, it is imperative to know how to actually perform the NIR measurements in order to ensure representative and reliable measurements of the heterogeneous WW material. In this section, we will give strong indications in this regard by describing the variability of waste wood material with a nested ANOVA with resampling. The analysis was performed on all the 16 lots and all showed similar results. However, in order to simplify the discussions, we will focus our analysis on one lot only. We have decided to report lot 12 as an example because our earlier results indicated that this is the one with the highest variability. **Figure 7** shows the variability in the MSQ values of nested ANOVA at each of the aforementioned levels (see *Deciding the Best Sampling Procedure*). As noted, the variation decreases with increasing number of scans, as expected. This trend is the same in all five levels.

These results provide good indications regarding the optimal sampling procedure to carry out in terms of the number of subsamples and scans to be performed to describe the variability in the waste wood materials. In fact, the variability in the MSQ values reached almost constant values at 125 scans for

level A; 250 scans for level B; 400 scans for level C; 600 scans for level D; around 500 for level E. It means that the same variability can be obtained by increasing the number of subsamples and decreasing the number of scans or by decreasing the number of subsamples and increasing the number of scans.

As seen from **Figure 7**, levels A and B give lower variability than the remaining sampling procedures, clearly indicating that it is insufficient to investigate one subsample only. This is confirmed by the MSQ values located slightly below the horizontal black line, which is the estimated overall variability of lot 12 and is deemed to be the true estimated variability of the lot. The three other sampling schemes are all very similar, giving indications that taking out two subsamples, splitting them into two replicates, and then measuring each of them with at least 100 scans seem to provide reliable and representative variability estimates of the lots (around 10 m^3 of fairly heterogeneous waste wood material).

CONCLUSION

Waste wood samples were collected in a panel board industry located in the northern part of Italy. All samples were analyzed using FT-NIR provided with a spiral sampler to investigate their variability and heterogeneity. A nested analysis of variance was computed to investigate the statistical differences for each level of the sampling procedure, i.e., lot, subsample, replicate, and scan levels. According to the results, waste wood has the highest variability at the lot level and lowest at the scan level.

PCA analysis on the MSQ values of the nested analysis of variance confirms the results of the nested ANOVA with increased clarity and shows how some lots deviate more from the others. The score plot clearly shows groupings among the lots and the loading plot displays that the main bands responsible for such separation are related to the overtone of O-H stretching bonds, which we also were able to confirm through reference analysis.

The knowledge of waste wood variability and composition is a key point for enhancing the sorting and related best reuse of the material with related positive effects in terms of economic, health, and environmental issues. NIRS proves to be a useful technique for rapidly obtaining this information. The definition of the most appropriate sampling procedure is essential for improving waste wood management and moving NIRS into real industrial applications. In fact, having a number of samples, replicates, and scans able to describe the variability of the material translates into reliable analytical results and accurate classification models for sorting the material based on the best reuse, especially when dealing with heterogeneous material. This study has proved that by taking at least two subsamples, splitting them into two replicates, and measuring each of them with at least 100 NIR scans, it is possible to describe the variability of around 10 m^3 of waste wood material. In future studies, this result can be used as the starting point for developing classification models, essential for more accurate and sustainable waste wood management.

These results have a large potential impact on the waste management sector, representing the first steps for moving NIR sensors to industrial waste management applications. In fact, the methodology used in this study can be applied not only to any other NIR spectrophotometers but also to other waste sources. When working with waste in general, the big challenge is the heterogeneity of the material. Thus, having a protocol that ensures efficient and reliable sampling will lead to the success of the subsequent classification of the waste according to waste categories, which will improve the sorting and, as a consequence, the reuse of the material.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in an online repository. The name of the repository and accession number can be found as follows: the dataset is uploaded to Zenodo repository with data DOI's assignment (http://doi.org/10.5281/zenodo.4896579).

AUTHOR CONTRIBUTIONS

M.M. and Å.R. were responsible for conceptualization and funding, contributed to the methodology, were responsible for computation using software, validated the data, conducted the formal analysis and investigation, acquired the resources, wrote and prepared the original, and reviewed and edited the manuscript. M.M. performed data curation, was responsible for visualization. Å.R. supervised the study and was responsible for project administration. All authors have read and agreed to the published version of the article.

FUNDING

This research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant, agreement No. 838560.

ACKNOWLEDGMENTS

The authors would like to thank Gruppo Saviola for allowing the sample collection and helping with the sampling organization and Q-Interline A/S (particularly Nicklas Bøg Pedersen) for the support in analyzing all the waste wood samples.

REFERENCES

Bergeron, F. C. (2016). Energy and Climate Impact Assessment of Waste wood Recovery in Switzerland. *Biomass and Bioenergy*. 94, 245–257. doi:10.1016/J.BIOMBIOE.2016.09.009

Commission of the European Communities (2008). Innovative and Sustainable Forest-based Industries in the EU: a Contribution to the EU's Growth and Jobs Strategy. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52008DC0113&from=EN (Accessed October 16, 2020).

Dessipri, E., Minopoulou, E., Chryssikos, G. D., Gionis, V., Paipetis, A., and Panayiotou, C. (2003). Use of FT-NIR Spectroscopy for On-Line Monitoring of Formaldehyde-Based Resin Synthesis. *Eur. Polym. J.* 39, 1533–1540. doi:10.1016/S0014-3057(03)00073-9

Edo, M., Björn, E., Persson, P.-E., and Jansson, S. (2016). Assessment of Chemical and Material Contamination in Waste wood Fuels - A Case Study Ranging over Nine Years. *Waste Management* 49, 311–319. doi:10.1016/J.WASMAN.2015.11.048

FAO/ECE (2012). Forest Production Annual Market Review 2011-2012. Available at: http://www.unece.org/fileadmin/DAM/timber/publications/FPAMR_2012.pdf

Faraca, G., Boldrin, A., and Astrup, T. (2019). Resource Quality of wood Waste: The Importance of Physical and Chemical Impurities in wood Waste for Recycling. *Waste Management* 87, 135–147. doi:10.1016/j.wasman.2019.02.005

Gehrmann, H.-J., Mätzing, H., Nowak, P., Baris, D., Seifert, H., Dupont, C., et al. (2020). Waste wood Characterization and Combustion Behaviour in Pilot Lab Scale. *J. Energ. Inst.* 93, 1634–1641. doi:10.1016/j.joei.2020.02.001

Hakala, J., and Deroubaix, G. (2012). "Used Wood Sorting, Utilization and Recycling in Different Value Chains", in NWBC 2012: The 4th Nordic Wood Biorefinery Conference. Editor K. Niemelä (VTT Technical Research Centre of Finland, VTT Technology), No. 53, pp. 45–50.

Hein, P. R. G., Campos, A. C. M., Mendes, R. F., Mendes, L. M., and Chaix, G. (2011). Estimation of Physical and Mechanical Properties of Agro-Based Particleboards by Near Infrared Spectroscopy. *Eur. J. Wood Prod.* 69, 431–442. doi:10.1007/s00107-010-0471-5

Henriques, A., Cruz, P., Martins, J., Ferra, J. M., Magalhães, F. D., and Carvalho, L. H. (2012). Determination of Formaldehyde/urea Molar Ratio in Amino Resins by Near-Infrared Spectroscopy. *J. Appl. Polym. Sci.* 124, 2441–2448. doi:10.1002/app.35128

Huron, M., Oukala, S., Lardière, J., Giraud, N., and Dupont, C. (2017). An Extensive Characterization of Various Treated Waste wood for Assessment of Suitability with Combustion Process. *Fuel* 202, 118–128. doi:10.1016/J.FUEL.2017.04.025

Lian, X., Zhang, M., Sun, X., Song, C., Yuan, H., Guo, L., et al. (2021). Online Real Time Determination of Free Formaldehyde Content during Polymerization Process of Phenolic Resin by NIR Spectra and a Modeling-free Method. *Polym. Test.* 93, 106584. doi:10.1016/j.polymertesting.2020.106584

Mancini, M., and Rinnan, Å. (2021). Near Infrared Technique as a Tool for the Rapid Assessment of Waste wood Quality for Energy Applications. *Renew. Energ.* 177, 113, 123. doi:10.1016/j.renene.2021.05.137

Mantau, U. (2012). Wood Flows in Europe (EU27). Available at: http://www.unece.org:8080/fileadmin/DAM/timber/meetings/20150311/Wood_flows_in_Europe_Mantau.pdf.

Martens, H., Jensen, S. A., and Geladi, P. (1983). "Multivariate Linearity Transformations for Near Infrared Reflectance Spectroscopy," in *Nordic Symposium Applied Statistics*. Editors O. H. J. Christie and S. Forlag (Stavanger, Norway), 205–234.

Mazzanti, M., and Zoboli, R. (2013). "International Waste Trade: Impacts and drivers," in *Waste Management in Spatial Environments*. Editors A. D'Amato, M. Mazzanti, and A. Montini (Routledge Studies in Ecological Economics), 99–136.

Moreno, A. I., and Font, R. (2015). Pyrolysis of Furniture wood Waste: Decomposition and Gases Evolved. *J. Anal. Appl. Pyrolysis*. 113, 464–473. doi:10.1016/j.jaap.2015.03.008

Rettenmaier, N., Schorb, A., and Köppen, S. (2008). *Status of Biomass Resource Assessments Version 1*. IFEU, Heidelberg: Biomass Energy Europe Project (D3.2).

Rinnan, Å., Berg, F. v. d., and Engelsen, S. B. (2009). Review of the Most Common Pre-processing Techniques for Near-Infrared Spectra. *Trac Trends Anal. Chem.* 28, 1201–1222. doi:10.1016/j.trac.2009.07.007

Röder, M., and Thornley, P. (2018). Waste wood as Bioenergy Feedstock. Climate Change Impacts and Related Emission Uncertainties from Waste wood Based Energy Systems in the UK. *Waste Management* 74, 241–252. doi:10.1016/J.WASMAN.2017.11.042

Sahai, H., and Ageel, M. I. (2000). *The Analysis of Variance*. Boston: Birkhäuser. doi:10.1007/978-1-4612-1344-4

Sandak, A., Sandak, J., Zborowska, M., and Prądzyński, W. (2010). Near Infrared Spectroscopy as a Tool for Archaeological wood Characterization. *J. Archaeological Sci.* 37, 2093–2101. doi:10.1016/j.jas.2010.02.005

Schwanninger, M., Rodrigues, J. C., and Fackler, K. (2011). A Review of Band Assignments in Near Infrared Spectra of Wood and Wood Components. *J. Near Infrared Spectrosc.* 19, 287–308. doi:10.1255/jnirs.955

Tatàno, F., Barbadoro, L., Mangani, G., Pretelli, S., Tombari, L., and Mangani, F. (2009). Furniture wood Wastes: Experimental Property Characterisation and Burning Tests. *Waste Management* 29, 2656–2665. doi:10.1016/j.wasman.2009.06.012

Tomlinson, S. K., Ghita, O. R., Hooper, R. M., and Evans, K. E. (2006). The Use of Near-Infrared Spectroscopy for the Cure Monitoring of an Ethyl Cyanoacrylate Adhesive. *Vibrational Spectrosc.* 40, 133–141. doi:10.1016/j.vibspec.2005.07.009

Toscano, G., Rinnan, Å., Pizzi, A., and Mancini, M. (2017). The Use of Near-Infrared (NIR) Spectroscopy and Principal Component Analysis (PCA) to Discriminate Bark and Wood of the Most Common Species of the Pellet Sector. *Energy Fuels* 31, 2814–2821. doi:10.1021/acs.energyfuels.6b02421

Vrancken, C., Longhurst, P. J., and Wagland, S. T. (2017). Critical Review of Real-Time Methods for Solid Waste Characterisation: Informing Material Recovery and Fuel Production. *Waste Management* 61, 40–57. doi:10.1016/j.wasman.2017.01.019

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal Component Analysis. *Chemometrics Intell. Lab. Syst.* 2, 37–52. doi:10.1016/0169-7439(87)80084-9

Workman, J., and Weyer, J. L. (2007). "Appendix 4a: Spectra-Structure Correlations for Near Infrared," in *Practical Guide to Interpretative Near-Infrared Spectroscopy*. Editors J. Workman and J. L. Weyer (Boca Raton: CRC Press), 239–264.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Analytical Science

ORIGINAL RESEARCH
published: 23 August 2021
doi: 10.3389/frans.2021.721657

Improved Understanding of Industrial Process Relationships Through Conditional Path Modelling With Process PLS

Tim Offermans[1], *Lynn Hendriks*[1], *Geert H. van Kollenburg*[1], *Ewa Szymańska*[2], *Lutgarde M. C. Buydens*[1] *and Jeroen J. Jansen*[1]*

[1]Institute for Molecules and Materials, Radboud University, Heyendaalseweg, Netherlands, [2]FrieslandCampina, Amersfoort, Netherlands

Understanding how different units of an industrial production plant are operationally related is key to improving production quality and sustainability. Data science has proven indispensable in obtaining such understanding from vast amounts of historical process data. Path modelling is a valuable statistical tool to obtain such information from historical production data. Investigating how relationships within a process are affected by multiple production conditions and their interactions can however provide an even deeper understanding of the plant's daily operation. We therefore propose conditional path modelling as an approach to obtain such improved understanding, demonstrated for a milk protein powder production plant. For this plant we studied how the relationships between different production units and steps are dependent on factors like production line, different seasons and product quality range. We show how the interaction of such factors can be quantified and interpreted in context of daily plant operation. This analysis revealed an augmented insight into the process that can be readily placed in the context of the plant's structure and behavior. Such insights can be vital to identify and improve upon shortcomings in current plant-wide monitoring and control routines.

Keywords: path modelling, process PLS, industry, relationships, experimental design

OPEN ACCESS

Edited by:
Raffaele Vitale,
Université de Lille, France

Reviewed by:
Daniel Gonzalo Palací-López,
IFF-Benicarlo, Spain
Carl Duchesne,
Laval University, Canada
Marco Reis,
University of Coimbra, Portugal

***Correspondence:**
Jeroen J. Jansen
chemometrics@science.ru.nl

Specialty section:
This article was submitted to Chemometrics, a section of the journal Frontiers in Analytical Science

Received: *07 June 2021*
Accepted: *09 August 2021*
Published: *23 August 2021*

Citation:
Offermans T, Hendriks L, van Kollenburg GH, Szymańska E, Buydens LMC and Jansen JJ (2021) Improved Understanding of Industrial Process Relationships Through Conditional Path Modelling With Process PLS. Front. Anal. Sci. 1:721657. doi: 10.3389/frans.2021.721657

INTRODUCTION

Industrial (bio)chemical processes need to be monitored and controlled well to guarantee sustainable and high-quality production despite variations in external factors such as raw materials, weather, plant operators, equipment maintenance and customer wishes. A deep understanding of how the production plant operates under and responds to these conditions is crucial for the development of accurate process monitoring and control strategies. To considerable extent, such understanding follows from first-principle knowledge. In practice, however, influences of external factors on the production, daily operation of the plant cannot be described completely by these first principles. Multivariate statistical analysis of historical production data can therefore reveal an augmented insight into the process, as this data does reflect the daily and real operation rather than the engineered operation.

Examples of statistical modelling methods that are widely used for this purpose are Principal Component Analysis (PCA), Partial Least Squares (PLS), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) (MacGregor & Kourti, 1995; Qin, 1997; Kourti, 2005; Cuentas

et al., 2017). These methods are often employed for process fault diagnosis through multivariate control (Shewhart) charts and for predicting difficult-to-measure production indicators, such as product quality, from easy-to-measure process variables (soft-sensoring) (Bersimis et al., 2007; Kadlec et al., 2009). Though these methods can be used to quantify the relationships between individual process parameters and variables, they provide limited higher-level insight into the relationships between different production units, as limited higher-level structural knowledge about the plant is employed.

The use of path analysis or structural equation modelling methods to industrial data analysis is therefore becoming increasingly popular, as these methods explicitly model the valuable information about relationships and can be considered explainable artificial intelligence (Höskuldsson et al., 2007; Gade et al., 2019). In general, path analysis methods estimate the directional statistical relationship between groups of measured variables. For industrial data, grouping process variables by the production unit in which they are measured thus allows for the estimation of how much operations of different production units are mutually related. This incorporates the physical structure of the production plant in the analysis of the data, of which the results in turn can be interpreted in the context of that structure (van Kollenburg G. H. et al., 2020).

Different methods for path analysis exist, including PLS-path modeling (Hair et al., 2011), sequential and orthogonalized PLS-path modeling (Romano et al., 2019), sequential multi-block PLS (Lauzon-Gauthier et al., 2018), multiblock kernel PLS (Zhang et al., 2010) and network PCA (Codesido et al., 2020). PLS-PM in particular is a well-established method in social sciences, but its high value for modelling industrial production data is also already demonstrated (van Kollenburg G. H. et al., 2020). Another path analysis method that has been developed very recently, is Process PLS (van Kollenburg et al., 2021). This method improves upon the mathematical limitations of PLS-PM and is better suited to model the complexity and heterogeneity of industrial production data as a network.

Process PLS is more appropriate for path modelling industrial data than alternative methods for three main reasons. Firstly, it can model multiple latent variables per group of process variables, in contrast to for instance PLS-path modeling. It can thus describe multiple sub-processes per production step, which are present for most industrial processes. Secondly, it can cope with the multicollinearity that the process variables of production steps often show (Guo et al., 2019). This gives rise to a more accurate estimation and better interpretability of the relationships between the production steps. Lastly, Process PLS (like PLS-path modeling but unlike for instance sequential and orthogonalized PLS-path modeling) does not require any a priori (importance) ranking to be imposed on the production steps, which in practice is difficult to do even for process experts (van Kollenburg et al., 2021a).

The relationships estimated with path modeling give much insight into the structure of the plant. Their sizes may even be related to an external production factor that is not directly included in the model, such as production cost (van Kollenburg G. H. et al., 2020). An even more exhaustive understanding of a plant's behavior can however be obtained by quantifying how the process relationships are affected by multiple, possibly interacting operating conditions, such as production season, year, parallel lines or product quality ranges. Such an analysis yields an elaborate insight into how the plant's operation is different under different combinations of production conditions. This allows process operators and engineers to even better steer the plant to cope with production variations caused by those multilevel conditions.

This paper presents a systematic approach for performing such a conditional path analysis on historical production data, using Process PLS. The work focuses on the use of Process PLS for such modelling, and a comparison to conditional modelling using alternative path modelling methods is out of scope for the current work. A large dataset from an industrial-scaled milk protein powder production plant is separated based on one or more operating conditions, after which each data subset is modelled and quantitatively compared. A thorough discussion of how the results of the analysis can be visualized, interpreted and communicated with and among process operators and engineers is provided.

METHODS AND DATA

Process PLS

A Process PLS model comprises two user-defined parts: the *inner* (*structural*) and *outer* (*measurement*) model. A production plant's structure can be modelled by grouping of the process variables (X) in the outer model according to the production units (or production steps). A group of variables is then called a block. The inner model defines which directional relationships are estimated between which production steps. For each unit, one or more latent variables (LV) are constructed to represent the major sources of covariance between the process variables of blocks which are connected in the inner model. The contribution of a process variable to specific latent variables for that unit are called weights (R, in some literature also referred to as W). Effects of the latent variable on other latent variable in the inner model are represented as explained variances (P^2, i.e. 'rho-squared'). The design of a Process PLS model is similar to that of a PLS-PM model, and is visualized in **Figure 1** for an example process. The relationships in the inner model may represent for instance a direct physical connection (piping), indirect connection between similar variables being measured at different locations), or feedforward control loops. As only recursive (non-cyclic) pathways can be modelled, feedbacks of either (intermediate) product or operation control actions cannot be directly modelled, but the process set points of a control scheme and/or the level of (intermediate) product feedback may for instance be used as a variable in the Process PLS outer model.

Estimation of a Process PLS model is done by iteratively optimizing a network of PLS-models using the SIMPLS-algorithm (de Jong, 1993). First, the dimensionality of the blocks is reduced to obtain estimates for the latent variables which maximize the covariance between interconnected blocks

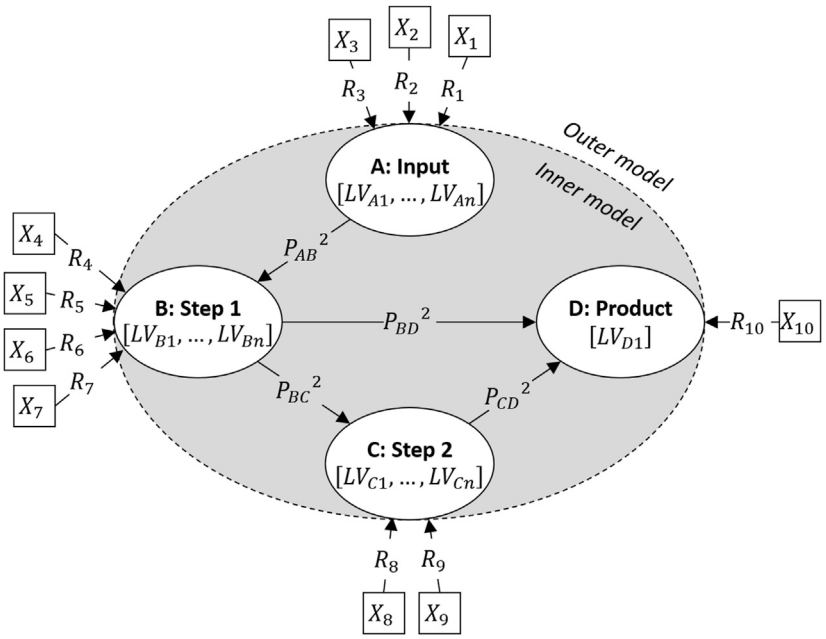

FIGURE 1 | The design of a Process PLS model for an example two-step production process. The input and product are also modelled as steps in order to estimate their relationships to the two production steps.

TABLE 1 | Number of samples and variables of the data collected for each of the three production lines, after synchronization and cleaning as explained in *Data preparation*.

Dimensions	Line A	Line B	Line C
Samples	1,569	560	924
Variables		51	
Milk		*1*	
Heating		*2*	
Precipitation 1		*5*	
Precipitation 2		*4*	
Washing		*21*	
MeltMaking		*7*	
Drying		*10*	
Product		*1*	

through a set of PLS2 regressions, one for each block of variables. To estimate the latent variables of a given step with PLS2, the process variables of that step are used as predictors and the process variables of all steps that step has a relationship to are used as responses. Only when a step has only incoming relationships, the process variables of the steps that have a relationship to that step are used as predictors and the process variables of the step itself are used as responses. The number of latent variables per block can be manually fixed if desired or optimized by internal cross-validation (which is the default in the software implementation used for the results in this paper, see *Software*). The process variable weights (R) are effectively the contributions of the variables to the relationships modelled by these PLS models. After the latent variables are estimated, a second set of PLS regressions is performed to estimate the relations in the inner model. The strengths of these relationships (P^2) are calculated from the PLS2 regression coefficients and represent the fraction of variance that the latent variables in a predictor block can explain in the response block. As Process PLS does not take into account process dynamics like mechanistic modelling approaches, knowledge about the kinetics of the process are not required for modelling. More details on the Process PLS method may be found in (van Kollenburg et al., 2021a).

Demonstrator Process

The industrial production facility investigated is a well-controlled plant that produces milk protein powder from skim milk. The skim milk is heated, after which it is subjected to two precipitation steps. The resulting curd is washed, dissolved in an alkali solution, and finally dried to a powder. The critical product quality indicator for the protein powder is the mineral content, which should be as low as possible. More details on milk powder production can be found in the dairy processing handbook (Bylund, 1995).

Data Collection

The data used in this study corresponds to three parallel production lines and three consecutive production years, and was not originally collected for other purposes than the current study. The data comprises 51 process variables, which are the same for the different production lines and are distributed across the processing steps as given in **Table 1**. All variables represent physical measurements, and not setpoints or production status values. Only data from effective production time was used in the current analysis. The variable representing

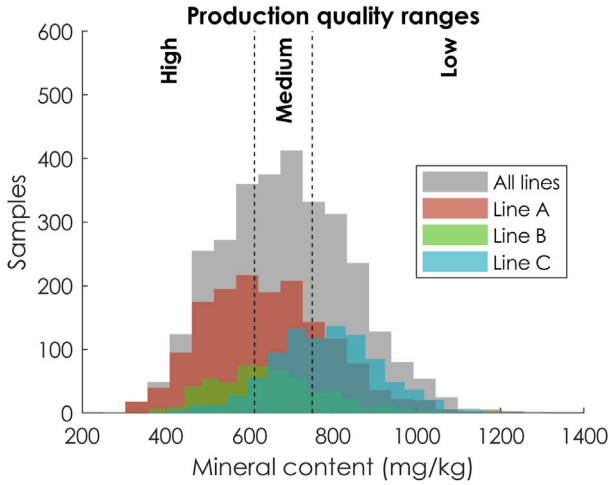

FIGURE 2 | Ranges for the product mineral content measurements used to separate the data based on product quality.

the product quality is the mineral content mentioned earlier, which is measured at-line at a relatively low frequency (hourly basis). The variable on incoming milk is also measured at similar frequency. All other variables are process variables such as temperatures, pressures and flow rates, and are measured in- or on-line at high frequency. The specific identities of these variables will not be disclosed as they are not relevant for the conclusions in this paper.

Data Preparation

Because the process variables are measured at separate locations and at different time intervals, the collected data had to be synchronized to obtain a multivariate dataset that can readily be analyzed. The high-frequency process variables were synchronized to the low-frequency product quality variable using median-filtering with a 3 h wide window, systematically selected as optimal synchronization (Offermans et al., 2020). This method also allows for a small degree of process dynamics to be included in the modelling procedure, as each synchronized sample represents the measurements done in the 3 hours before its sampling time. Time-lags between individual process variables are not taken into account. For the relative low-frequency measurements on incoming milk, the most recent measured value was matched to each mineral content sample. Missing values can be and were present after the synchronization procedure, and were imputed by replacing them by the median of the values that were present (Souza et al., 2016). This was done per production line and per production variable. Outlying samples were detected per production line using the multivariate Hotelling's T^2- and Q-statistics calculated from PCA models explaining at least 70% variance of the autoscaled data. Samples for which at least one statistic was over three standard deviations removed from the median were removed (Varmuza and Filzmoser, 2016). The number of samples obtained after the data collection, synchronization and cleaning are given in **Table 1**.

Path Modelling Conditional to Single Operation Conditions

The first part of the study focused on investigating the effects of the individual production conditions separately on the process relationships. The three (multilevel) conditions that were explored are production line, production season and product quality. All data was for instance only separated according to the three production lines. For separating the data into seasons, meteorological seasons were used as these are identical for each year. The mineral content values were used to separate the data into three relative product quality ranges. The boundaries of these ranges were set at the 1st and 2nd tertiles to ensure comparable sample sizes for all models, as is illustrated in **Figure 2**. As mentioned before, a low mineral content value indicates a high-quality production.

Each data subset was individually modelled with Process PLS, using the same *inner* and *outer* model specification for each model. The directional relationships between the production steps that were estimated using Process PLS are illustrated in **Figure 3**. The *inner model*, shown in **Figure 3**, was specified according to two criteria introduced by van Kollenburg *et al.* (van Kollenburg G. H. et al., 2020). Firstly, relationships of each step on the subsequent step are included (counter-clockwise, starting from the top, in **Figure 3**). These represent the physical architecture of the plant and the flow of the process (piping). Secondly, direct relationships of each production step on the product-variables and thus the product quality are included. The *outer model*, which relates the process variables to the different production steps, was specified based on the physical location of each process variables. The number of variables per step thus are reported in **Table 1**.

The number of latent variables considered for each block/step was optimized using the default cross-validation procedure in the Process PLS implementation used (*'pathmodelr'*). Before modelling, all individual process variables were autoscaled to

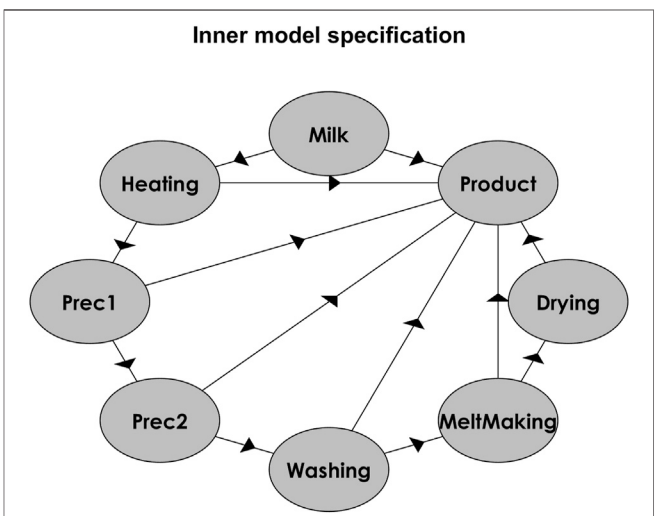

FIGURE 3 | Inner model specification used for path modelling of the milk protein powder production process.

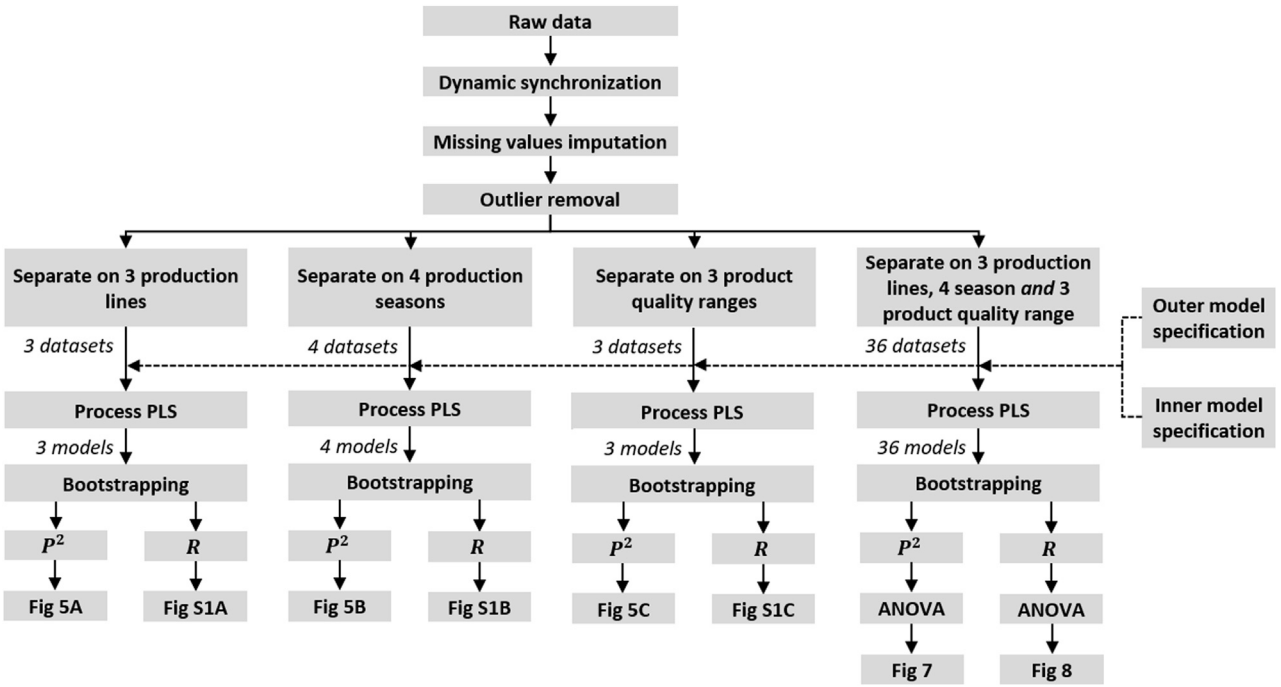

FIGURE 4 | Schematic overview of the conditional path modelling analyses presented.

have zero mean and unit standard deviation, after which the process variables are collectively but per step rescaled so that each step has a sum of squares of 1. This is the default procedure by *pathmodelr*. All remaining modelling settings were also kept at their default values. To estimate the precision of the modelled process relationships, each Process PLS model was subjected to a non-parametric bootstrap with 200 replicates (Johnson, 2001).

Path Modelling on Multiple Production Conditions

For the second part of the study, the full data was separated on all production conditions at once, following a full factorial design. Each data subset was modelled using Process PLS, to calculate the process relationships for each possible combination of production conditions. Three-way ANOVA analyses were used to estimate the main and interaction effects of the production conditions on each separate process relationship and process variable weight (Huitson et al., 1976). This allows for the investigation of interactions between the production conditions on the process relationships, for instance between production season and line. The boundaries for the quality ranges were, as before, set relatively at the 1st and 2nd tertiles. They were set per combination of line and season, to ensure sufficient samples in each experiment for reliable modelling. The design matrices for the experimental design and the sample sizes for each experiment (and thus Process PLS model) are shown in **Supplementary Table S1** in the supplemental material.

The modelling and bootstrapping procedure for each data subset (full factorial design experiment) was identical to that used before while investigating the separate production conditions. The three-way ANOVA analyses were performed on the mean results found after bootstrapping. A False Discovery Rate (FDR) correction was applied to the *p*-values obtained with ANOVA using the method proposed by Benjamini and Hochberg to adjust for multiple testing errors (Benjamini and Hochberg, 1995). This because the relationships and dependencies identified with the proposed analysis may require further investigation by plant personnel, which is time and cost intensive. As such, false positives (type I) errors are more harmful and less desirable than false negatives (type II) errors.

A schematic overview of the different data preparation, separation, modelling and interpretation steps performed as part of the presented study on conditional path modelling is shown in **Figure 4**.

Software

Data preparation was done using MATLAB R2017a (*MATLAB*, 2017). Modelling data with Process PLS was done in R, using the *pathmodelr* package version 0.1.2 (Team R Development Core, 2018; van Kollenburg G. H. et al., 2020).

RESULTS AND DISCUSSION

Path Modelling Conditional to Single Operation Conditions

Figures 5A–C show the primary modelling results found after partitioning the complete data only on either production line, production season or product quality range (respectively). Shown

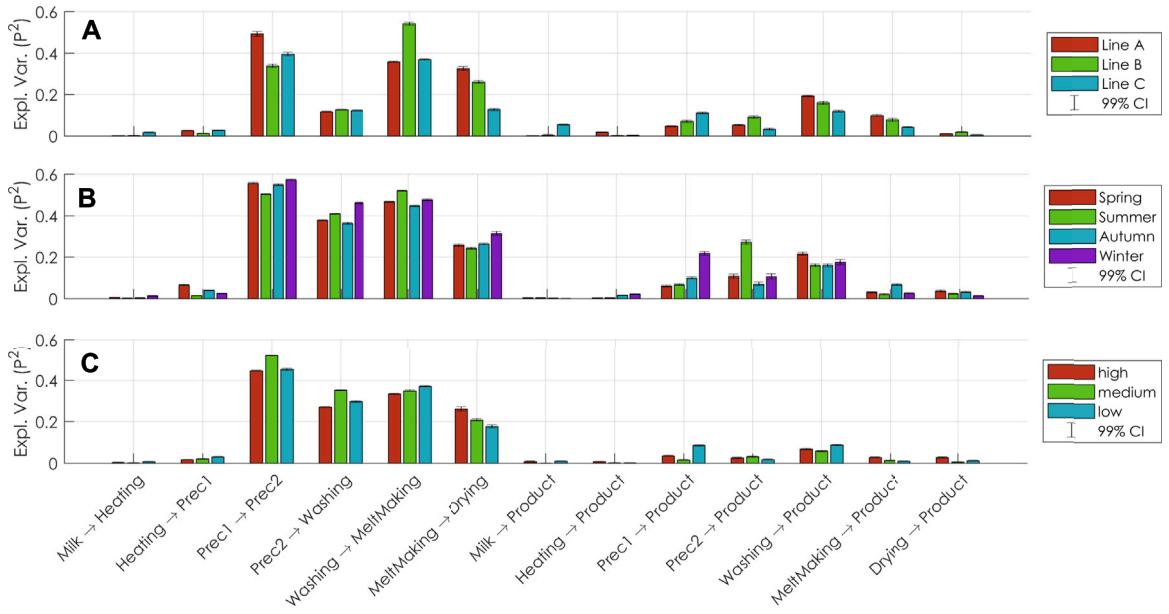

FIGURE 5 | (A–C): Size of process relationships in terms of fractions of explained variance (P^2), as found when using Process PLS modelling on either separate production lines (**A**), or production season (**B**), or product quality range (**C**). The bars represent the means and the whiskers represent the 99% confidence intervals over 200 bootstrap replicates.

are the proportions of variance explained (P^2) for each relationship in the inner model (as shown in **Figure 3**). These values quantify the directional relationship between the production steps. Shown per relationship are the mean values over the 200 bootstrapping replicates. The 99% confidence intervals are plotted as error whiskers but are for many results too small to discern. This indicates that the results have high precision and attests that Process PLS is a robust method for statistical modelling of industrial data.

The results in **Figures 5A,B** give insights into the relationships within the process, and how they differ under various production conditions. Firstly, they show which relationships are overall strongest. For this process, the relationship from *Prec1* to *Prec2* is in general the strongest, irrespective of production line, season, or product quality range. These steps are likely strongly related because they have a similar function in the process. From all the production steps, *Washing* relates strongest to *Product* under most conditions. This indicates that *Washing* may be the most influential step for the product quality, and future optimization efforts should be directed to this step. Importantly, *Milk* in general only relates to *Product*. Though this may sound counter-intuitive, it indicates that variations in *Milk* do not influence the production quality. In turn, this supports the notion that the process is well-controlled and that stable production quality is achieved despite raw material variations.

Results from the conditional modelling show that the relationship between *Prec1* and *Prec2* is weaker for production line B than for the other production lines (**Figure 5A**). This indicates that the operation of *Prec2* is less related to that of *Prec1* in line B than in the other lines. Additionally, the relationship between *Prec2* and *Product* is stronger for line B than for the other lines, indicating that variations in *Prec2* are related to variations in *Product*. In a production process with a focus on constant quality, this results may be an important focus for follow-up investigations.

Separating the data only on production season (**Figure 5B**) reveals that the *Prec1* relates stronger to *Product* in the winter, while *Prec2* relates stronger to *Product* in the summer. This indicates that the focus of process control is different for the seasons, for instance because seasonal variation manifested in the raw material or weather influences the *Prec1* and *Prec2* steps differently. This is supported by *Prec1* → *Prec2* being lower in summer and higher in winter.

When looking at the different product quality ranges (**Figure 5C**), it is interesting that *Washing* → *MeltMaking* increases and *MeltMaking* → *Drying* decreases with decreasing product quality. This suggests that higher quality product is obtained when the operation of *MeltMaking* is more aligned with that of *Drying* (the step after it) than with that of *Washing* (the step before it). This should be further investigated, as it could indicate that aligning the *MeltMaking* settings with that of *Drying* instead of *Washing* leads to structurally higher production quality.

The results in **Figures 5A–C** give already much insight into the process but understanding of the process can be augmented by evaluating the weights (R) of the process variables in the Process PLS models. As an example, **Figure 6** shows the weights for the variables corresponding to *Prec1* and *Prec2* in the models obtained after separating the data on production line alone. These weights represent the contributions of the process variables on the latent variables of their respective block. As previously discussed, the relationship between *Prec1* to *Prec2* is weaker for line B than for lines A and C (**Figure 5A**). Because *Prec2 V2* has a particular high weight in the model of line B, plant operators and engineers

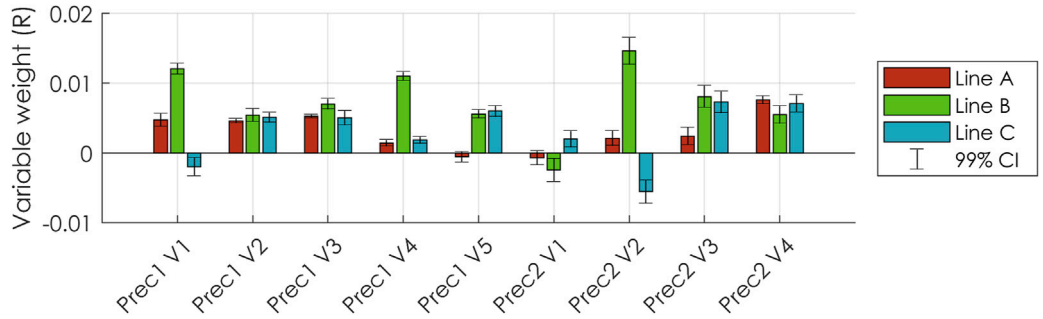

FIGURE 6 | Weights (R) of the process variables of Prec1 and Prec2 in the different Process PLS models trained per production line. The bars represent the means and the whiskers represent the 99% confidence intervals over 200 bootstrap replicates.

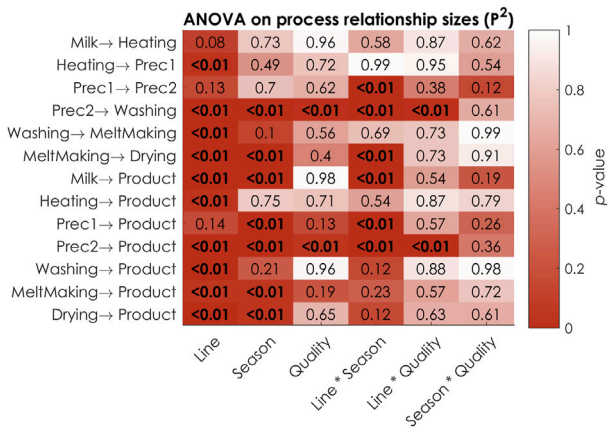

FIGURE 7 | *FDR-corrected p-values obtained by performing three-way ANOVA on the fractions of explained variance (P^2) found with Process PLS according to the full factorial conditional path modelling approach. A low p-value signifies a low probability that the process relationship size is comparable under different (combinations of) production conditions.*

could be advised to investigate the operation of this variable further. It likely has a characteristic behavior unique in line B that causes the operation of *Prec2* to be less related to *Prec1* which, as discussed earlier, may influence the product quality.

This example illustrates how variable weights should be interpreted, and how investigating these may aid process operators and engineers in optimizing monitoring and control of a production plant. The variable weights can provide much more information, but discussing all of them for the process in this paper is of limited value, as their identities are disclosed. The weights of all variables for all models are given in the supplementary materials in **Supplementary Figures S1A–C** for the interested reader but are not discussed further here.

Path Modelling Conditional to Multiple Operation Conditions

Figure 7 displays the results of analyzing each combination of the three production conditions according to a full-factorial experimental design with the same Process PLS model and analyzing variations in the model parameters using an ANOVA. Note that this experimental design is applied to data that is already measured, and that is no further measurements are collected according to that design. As many PLS regressions are calculated during this experiment, 936.000 to be exact (36 production condition combinations, 13 inner relationships, 10 cross-validation repeats and 200 bootstrap repeats), it should be noted that the computation time for obtaining the results as presented in this manuscript is around 18 min when using a desktop computer with an Intel Core i7-7900 K processor. Although significant, this computation time should not be limiting for the use of the proposed methodology as a tool for off-line exploration of historical data. The number of cross-validation repeats and/or bootstrap repeats could be reduced to save computation time on slower systems, but the robustness of the models should be checked with additional care.

Shown in **Figure 7** are the FDR-corrected *p*-values of each three-way ANOVA that was performed per modelled process relationship size (in terms of mean explained variance, P^2, over bootstrap replicates). These results thus represent the inner path model. The *p*-values quantify the probability of the relationships sizes being identical regardless of a certain condition (e.g. *'Line'*) or interaction of conditions (e.g. *'Line*Season'*). Thus, a very low *p*-value indicates that relationship is significantly different for at least one (combination of) production conditions. This visualization offers a comprehensive view of the conditional path modelling results, while also quantifying statistical significance as it is not subjective to visual interpretation.

The results of the first part of the study (discussed above) showed that the individual production conditions do effect the process relationships. The results in **Figure 7** confirm such primary effects. All but three process relationships are, for instance, different for at least one production line. The ANOVA results however also show that there are many interactions of these production conditions. The relationship size of *MeltMaking* to *Drying* is for instance dependent on both the production season and line individually (*p*-values < 0.01), but there is also a significant interaction of these two operation conditions for that relationship. This indicates that the relationship size between *MeltMaking* and *Drying* not only differs

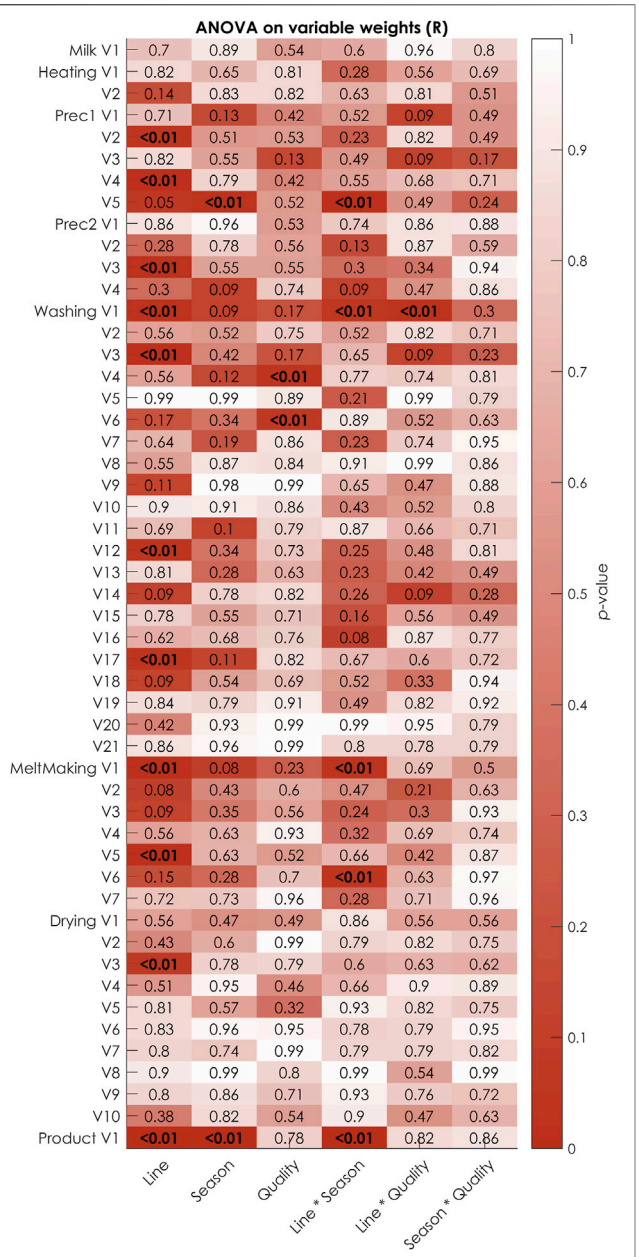

FIGURE 8 | *FDR-corrected p-values obtained by performing three-way ANOVA on the process variable weights (R) found with Process PLS according to the full factorial conditional path modelling approach. A low p-value signifies a low probability that the process variable weight is comparable under different (combinations of) production conditions.*

for the seasons, but that the way in which they differ for the seasons in turn also differs for the production lines.

The results found for *Prec1* → *Prec2* when separating the data on single conditions, which were elaborately discussed in *Path modelling conditional to single operation conditions*, seem to contradict the main effects for the single conditions found with ANOVA when separating the data on all conditions. *Prec1* → *Prec2* was concluded to be different for the production lines and seasons (**Figures 5A,B**), but these conditions show relative high *p*-values for *Prec1* → *Prec2* in **Figure 7** (0.13 and 0.7, respectively). The results in **Figure 7** thus suggest that *Prec1* → *Prec2* is not likely different for at least production line or for at least one production season. Such apparent contradictions are caused by the interactions of the production conditions: the ANOVA results do suggest a large interaction between production line and season, signified by a relative low *p*-value (<0.01). This means that the production line and season are affecting this relationship, but that they are not doing so independently. Such information is highly valuable, as future efforts to make this step more robust against seasonal variations should thus be done per production line. Being able to quantify such interactions underlines the value of conditional path modelling while separating the data on all combinations of production conditions.

Figure 8 gives the results of the three-way ANOVAs performed on the individual process variable weights (*R*, averaged over bootstrap replicates), when modelling the data while separated on all production conditions simultaneously (full-factorial). These *p*-values are also FDR-corrected. The results represent the outer path model and can be similarly interpreted as the results in **Figure 7**, and supplement those results to extract more process-specific information. For instance, the relationship size of *Washing* to *Product* was found to be relatively strong in general (**Figures 5A–C**), and was found to be highly dependent on the production line (**Figure 7**). This makes *Washing* an interesting step to investigate further, or even experiment with. That analysis could then be advised to focus on variable *Washing V1*, of which the operation is dependent on the production line alone, but also on the interactions of both the production season and quality range with the production line. This variable is thus likely largely responsible for the dependencies of *Washing* → *Product* on the production conditions. This observation and the ones discussed above exemplify the insight that conditional path modelling gives into the relationships within a production process. Much more process-specific information can however still be extracted from these results, especially by or while consulting with process operators and engineers that are experienced in controlling the process on a daily basis.

For this demonstration, data was available for each combination of production conditions, but this may not be the necessarily hold for other production facilities. One parallel line may for instance never be used during winter, leading to a missing experiment in the design. In such cases, ANOVA may still be used to analyze the path modelling results, but Type I sums of squares should be used rather than Type III sums of squares. Alternatively, if including one operation condition causes too many missing experiments, it may be better to remove it altogether from the analysis. A parallel line that is only used during winter is for instance less insightful to include, and could be excluded from the analysis. Another solution could be to adapt the Process PLS model specification and include the operation condition as a process variable. It should furthermore be ensured that enough samples are present for each of the experiments to enable a reliable estimation of the process relationships with Process PLS for the corresponding combination of production conditions. A minimum of 30 samples is used for the

demonstration given and is advisable, but the robustness of the fitted process relationships should in any case be assessed by analyzing the bootstrapping results as the minimum number of samples required will be process-specific.

CONCLUSION

This study presented a systematic approach for conditional path modelling of industrial production data using Process PLS, and demonstrated its value for a milk powder production facility. The approach consists of separating historical data based on one or more operation conditions, and modelling and comparing each of those datasets. This can be used to investigate how the statistical relationships between the production steps of a plant vary for, for instance, different production lines, seasons and quality ranges, and which of the measured process variables in those steps are most correlated to this behavior. An unprecedented high level of process expert knowledge on the structure and operation of the plant can thus be incorporated in the analysis of large historical datasets. Results for conditional modelling on a single production condition at a time and on all production conditions simultaneously were presented. The latter requires more data for stable modelling, was shown to be preferred as it allows for the quantification of interaction effects of the production conditions on the process relationships. Such interactions were present for the demonstrator process, and interpreting them gave a very detailed insight into the plant operation. These insights can both confirm and expand the current understanding of the process. This is of high value to process operators and engineers, who can use this improved understanding to pinpoint shortcomings in the current process monitoring and control strategy. Although only demonstrated on a continuous process in the current work, conditional path modelling may also be of great value for (batch-like) process with multiple production stages by considering those stages as a production condition. Ultimately, conditional path modelling can help in making production plants less prone to variations in external operating conditions, and in increasing product quality even for production plants that are already considered well-controlled.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

TO: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review and editing, Visualization; LH: Conceptualization, Methodology, Formal analysis, Investigation, Writing—review and editing; GK: Conceptualization, Methodology, Writing—review and editing, Supervision; ES: Conceptualization, Methodology, Resources, Writing—review and editing, Visualization, Supervision, Project administration; LB: Supervision, Project administration, Funding acquisition; JJ: Conceptualization, Methodology, Writing—review and editing, Visualization, Supervision, Project administration, Funding acquisition.

FUNDING

This project is co-funded by TKI-E and I with the supplementary grant 'TKI- Toeslag' for Topconsortia for Knowledge and Innovation (TKI's) of the Ministry of Economic Affairs and Climate Policy. The authors thank all partners within the project 'Integrating Sensor Based Process Monitoring and Advanced Process Control (INSPEC)', managed by the Institute for Sustainable Process Technology (ISPT) in Amersfoort, Netherlands.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frans.2021.721657/full#supplementary-material

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* 57 (1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bersimis, S., Psarakis, S., and Panaretos, J. (2007). Multivariate Statistical Process Control Charts: An Overview. *Qual. Reliab. Engng. Int.* 23 (5), 517–543. doi:10.1002/qre.829

Bylund, G. (1995). "Dairy Processing Handbook," in *Tetra Pak Processing Systems*, Vol. G3. Tetra Pak Processing Systems AB. Available at: http://www.ales2.ualberta.ca/afns/courses/nufs403/PDFs/chapter15.pdf.

Codesido, S., Hanafi, M., Gagnebin, Y., González-Ruiz, V., Rudaz, S., and Boccard, J. (2020). Network Principal Component Analysis: a Versatile Tool for the Investigation of Multigroup and Multiblock Datasets. *Bioinformatics* 37, 1297–1303. doi:10.1093/bioinformatics/btaa954

Cuentas, S., Peñabaena-Niebles, R., and Garcia, E. (2017). Support Vector Machine in Statistical Process Monitoring: a Methodological and Analytical Review. *Int. J. Adv. Manuf Technol.* 91 (1–4), 485–500. doi:10.1007/s00170-016-9693-y

de Jong, S. (1993). SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics Intell. Lab. Syst.* 18 (3), 251–263. doi:10.1016/0169-7439(93)85002-X

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., and Taly, A. (2019). "Explainable AI in Industry," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 3203–3204. doi:10.1145/3292500.3332281

Guo, S., Pang, K., and Qin, S. (2019). Least Angle Regression and Partial Least Squares Regression on Process Data and High Collinearity. *Foundations Process Analytics Machine Learn.* 57, 201682944. https://api.semanticscholar.org/CorpusID:201682944.

Hair, J. F., Ringle, C. M., and Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *J. Marketing Theor. Pract.* 19 (2), 139–152. doi:10.2753/MTP1069-6679190202

Höskuldsson, A., Rodionova, O., and Pomerantsev, A. (2007). Path Modeling and Process Control. *Chemometrics Intell. Lab. Syst.* 88 (1), 84–99. doi:10.1016/j.chemolab.2006.09.010

Huitson, A., Dunn, O. J., and Clark, V. A. (1976). Applied Statistics: Analysis of Variance and Regression. *The Statistician* 25 (Issue 3), 236, 1976 . Wiley. doi:10.2307/2987845

Johnson, R. W. (2001). An Introduction to the Bootstrap. *Teach. Stat.* 23 (Issue 2), 49–54. CRC press. doi:10.1111/1467-9639.00050

Kadlec, P., Gabrys, B., and Strandt, S. (2009). Data-driven Soft Sensors in the Process Industry. *Comput. Chem. Eng.* 33 (4), 795–814. doi:10.1016/j.compchemeng.2008.12.012

Kourti, T. (2005). Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry. *Int. J. Adapt. Control. Signal. Process.* 19 (4), 213–246. doi:10.1002/acs.859

Lauzon-Gauthier, J., Manolescu, P., and Duchesne, C. (2018). The Sequential Multi-Block PLS Algorithm (SMB-PLS): Comparison of Performance and Interpretability. *Chemometrics Intell. Lab. Syst.* 180, 72–83. doi:10.1016/J.CHEMOLAB.2018.07.005

MacGregor, J. F., and Kourti, T. (1995). Statistical Process Control of Multivariate Processes. *Control. Eng. Pract.* 3 (3), 403–414. doi:10.1016/0967-0661(95)00014-L

MATLAB (2017). *The Math Works* (Natick, Massachusetts: Inc). R2017a ed.

Offermans, T., Szymańska, E., Buydens, L. M. C., and Jansen, J. J. (2020). Synchronizing Process Variables in Time for Industrial Process Monitoring and Control. *Comput. Chem. Eng.* 140, 106938. doi:10.1016/j.compchemeng.2020.106938

Qin, S. J. (1997). "Neural Networks for Intelligent Sensors and Control - Practical Issues and Some Solutions," in *Neural Systems for Control*. Editors O. Omidvar and D. L. Elliott (Academic Press), 213–234. doi:10.1016/b978-012526430-3/50009-x

Romano, R., Tomic, O., Liland, K. H., Smilde, A., and Næs, T. (2019). A Comparison of twoPLS-based Approaches to Structural Equation Modeling. *J. Chemometrics* 33 (3), e3105. doi:10.1002/cem.3105

Souza, F. A. A., Araújo, R., and Mendes, J. (2016). Review of Soft Sensor Methods for Regression Applications. *Chemometrics Intell. Lab. Syst.* 152, 69–79. doi:10.1016/j.chemolab.2015.12.011

Team R Development Core (2018). "A Language and Environment for Statistical Computing," in *R Foundation for Statistical Computing*, Vol. 2. 3.6.3. Available at: https://www.R-project.org..

van Kollenburg, G., Bouman, R., Offermans, T., Gerretzen, J., Buydens, L., van Manen, H.-J., et al. (2021). Process PLS: Incorporating Substantive Knowledge into the Predictive Modelling of Multiblock, Multistep, Multidimensional and Multicollinear Process Data Manuscript Revision Printed in Blueblue. *Comput. Chem. Eng.* 154, 107466. doi:10.1016/J.COMPCHEMENG.2021.107466

van Kollenburg, G. H., Bouman, R., Offermans, T., and Jansen, J. (2020b). Data, Software and Scripts Related to the Process PLS Methodology Manuscript. *Mendeley Data*. doi:10.17632/9x9h7fr4kn.1

van Kollenburg, G. H., van Es, J., Gerretzen, J., Lanters, H., Bouman, R., Koelewijn, W., et al. (2020a). Understanding Chemical Production Processes by Using PLS Path Model Parameters as Soft Sensors. *Comput. Chem. Eng.* 139, 106841. doi:10.1016/j.compchemeng.2020.106841

Varmuza, K., and Filzmoser, P. (2016). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Taylor & Francis Group, LLC. doi:10.1201/9781420059496

Zhang, Y., Zhou, H., Qin, S. J., and Chai, T. (2010). Decentralized Fault Diagnosis of Large-Scale Processes Using Multiblock Kernel Partial Least Squares. *IEEE Trans. Ind. Inf.* 6 (1), 3–10. doi:10.1109/TII.2009.2033181

Conflict of Interest: ES was employed by FrieslandCampina.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/fchem.2021.734132

The "DOLPHINS" Project: A Low-Cost Real-Time Multivariate Process Control From Large Sensor Arrays Providing Sparse Binary Data

Eugenio Alladio[1], Marcello Baricco[1], Vincenzo Leogrande[2], Renato Pagliari[2], Fabio Pozzi[3], Paolo Foglio[3] and Marco Vincenti[1]*

[1]Dipartimento di Chimica, Università Degli Studi di Torino, Torino, Italy, [2]RADA Snc–Soluzioni Informatiche, Rivoli, Italy, [3]CNH Industrial–Lungo Stura Lazio, Torino, Italy

OPEN ACCESS

Edited by:
Angelo Antonio D'Archivio, University of L'Aquila, Italy

Reviewed by:
Gianpiero Adami, University of Trieste, Italy
Mohammad Sharif Khan, Wake Forest Baptist Medical Center, United States

***Correspondence:**
Eugenio Alladio
eugenio.alladio@unito.it

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *30 June 2021*
Accepted: *06 August 2021*
Published: *03 September 2021*

Citation:
Alladio E, Baricco M, Leogrande V, Pagliari R, Pozzi F, Foglio P and Vincenti M (2021) The "DOLPHINS" Project: A Low-Cost Real-Time Multivariate Process Control From Large Sensor Arrays Providing Sparse Binary Data. Front. Chem. 9:734132. doi: 10.3389/fchem.2021.734132

The "DOLPHINS" project started in 2018 under a collaboration between three partners: CNH Industrial Iveco (CHNi), RADA (an informatics company), and the Chemistry Department of the University of Turin. The project's main aim was to establish a predictive maintenance method in real-time at a pilot plant (CNHi Iveco, Brescia, Italy). This project currently allows maintenance technicians to intervene on machinery preventively, avoiding breakdowns or stops in the production process. For this purpose, several predictive maintenance models were tested starting from databases on programmable logic controllers (PLCs) already available, thus taking advantage of Machine Learning techniques without investing additional resources in purchasing or installing new sensors. The instrumentation and PLCs related to the truck sides' paneling phase were considered at the beginning of the project. The instrumentation under evaluation was equipped with sensors already connected to PLCs (only on/off switches, i.e., neither analog sensors nor continuous measurements are available, and the data are in sparse binary format) so that the data provided by PLCs were acquired in a binary way before being processed by multivariate data analysis (MDA) models. Several MDA approaches were tested (e.g., PCA, PLS-DA, SVM, XGBoost, and SIMCA) and validated in the plant (in terms of repeated double cross-validation strategies). The optimal approach currently used involves combining PCA and SIMCA models, whose performances are continuously monitored, and the various models are updated and tested weekly. Tuning the time range predictions enabled the shop floor and the maintenance operators to achieve sensitivity and specificity values higher than 90%, but the performance results are constantly improved since new data are collected daily. Furthermore, the information on where to carry out intervention is provided to the maintenance technicians between 30 min and 3 h before the breakdown.

Keywords: predictive maintenance, machine learning, sparse binary data, multivariate data analysis, principal component analysis, soft independent modeling by class analogy

INTRODUCTION

The current and future sustainable economic growth of companies worldwide are today, more than ever, increasingly based on the value and the information created by data. In the field of industry, the features of Industry 4.0 are showing a growing impact on the productive processes, since the companies are financially encouraged to move towards industrial automation that integrates some new production technologies aimed at improving working conditions, creating new business models, and increasing the productivity and product quality of their plants. Furthermore, the governments of several countries are promoting business plans and strategies focused on Industry 4.0 to offer the companies the tools aimed at seizing the opportunities of innovation and digital instruments related to the current fourth industrial revolution (Gentner, 2016; Enyoghasi and Badurdeen, 2021; Gallo et al., 2021; Ghobakhloo et al., 2021). In this context, Big Data and Data Analytics themes play a strategic role since data are indeed considered the lifeblood of economic development of the industrial (but not only) companies nowadays. Data are the basis for evaluating the quality of the products and generating gains in productivity and resource efficiency, making it possible to optimize the production process and enhance the whole plant's efficiency. Consequently, many companies face the necessity of implementing strategies capable of collecting and interpreting the data robustly and systematically alongside their productive process (Cugno et al., 2021; Goldman et al., 2021; Jeske et al., 2021; Lee and Lim, 2021). Various Multivariate Data Analysis (MDA) models and Machine Learning (ML) approaches have been gradually introduced within the production plants to develop competitive strategies, such as process control, quality control, and predictive maintenance (Elsisi et al., 2021; Jamwal et al., 2021; Lee and Lim, 2021; Wankhede and Vinodh, 2021). The last topic is fascinating for the companies since, if historical data have been already stored in databases, predictive maintenance substantially requires the computation of ML algorithms to predict the necessity of a repair or, eventually, a replacement, which can be therefore programmed and performed the way it turns to be most effective. Predictive maintenance was originally performed using user-defined alerts or expert-defined thresholds involving Supervisory Control And Data Acquisition (SCADA) systems. However, this approach does not consider the presence of correlations, patterns, and similarities among the collected features and the available signals detected from the sensors on the machinery. On the other hand, MDA and ML tools perform a multivariate interpretation of the stored data, which can belong to even different kinds of databases (e.g., sensors, SCADA, and history data) and origins (e.g., IT data, shop floor information, and manufacturing processes) (Ghobakhloo et al., 2021; Lee and Lim, 2021). The current work focuses on developing and testing several Machine Learning approaches at a pilot automotive plant (CNHi Iveco, Brescia, Italy) for predictive maintenance purposes. In particular, the goal of the "DOLPHINS" project was to build a low-cost edge digital twin capable of performing real-time predictive maintenance starting from data already collected and available at the plant level. This project was settled in 2018 under a collaboration among CNH Industrial Iveco (CHNi), RADA (an informatics company), and the Department of Chemistry of the University of Turin. In more detail, the goal of the DOLPHINS project was to develop a software application—in the tangible form of a dashboard working in real-time as a statistical digital twin of a shopfloor asset—by implementing a twin statistical model of the equipment under examination to deliver behavioral predictive warnings to the maintenance technicians in order to intervene on the investigated machinery preventively. Fundamental targets of the DOLPHINS project were as follows: 1) to provide the technicians an approach showing robust predictive capabilities of performing real-time maintenance; 2) to diminish as much as possible the occurrences of breakdowns, stops, and micro-stops, aspiring to a near-zero downtime goal; 3) to develop a low-cost implementation of this approach since training data for ML and MDA approaches were already collected and stored in programmable logic controller (PLC) devices. By referring to the last DOLPHINS target, a relevant advantage of this project is that ML models were built on data already available by the PLC equipment itself from large sensor arrays. Hence, no additional sensors were needed since the multivariate models were trained on the historical data and then tested on those acquired recently, reducing the impact on the company in terms of implementation costs and time. Since data are stored by the PLCs in the form of sparse binary matrices, several ML algorithms were tested during the development stage of DOLPHINS. Therefore, various MDA classification models were evaluated to predict the occurrence of failures within a given time window and their performance was monitored to choose the optimal model to perform constant and real-time processing of the data. Finally, once new data and signals are collected, they are interpreted by the developed ML model to monitor the performance of the machinery under examination and predict its evolution by detecting any significant drift and variation over time. The real-time results are expressed in terms of the probability of malfunctions and severity of the signals recorded by the PLCs to allow the maintenance technicians to work promptly on specific machinery sections. This approach diminishes the occurrences of stops and breakdowns sensitively and provides further knowledge on the behavior of the machinery itself. The final goal of the DOLPHINS project is to extend this approach to other shopfloor systems by raising the amount of cost savings, diminishing the periods of downtime, and improving the efficiency and the predictability of the productive process.

MATERIALS AND METHODS

Framework and Project Development

The development of the DOLPHINS project started as a proof-of-concept study on evaluating the data acquired at the CNH Industrial Iveco (CHNi) Plant of Brescia (Italy). The working area selected for the project consisted of the Welding Operative Unit and two types of machinery were monitored [namely, 0P10–External Door Compartment Ring (AVPE) and 0P10–Internal Door Compartment Ring (AVPI)]. Signals registered from AVPE and AVPI machinery (**Figure 1**) were historically stored into PLCs but not interpreted in ML modeling for predictive maintenance. In detail, the data of AVPE and AVPI consisted of 176 and 153 sensors

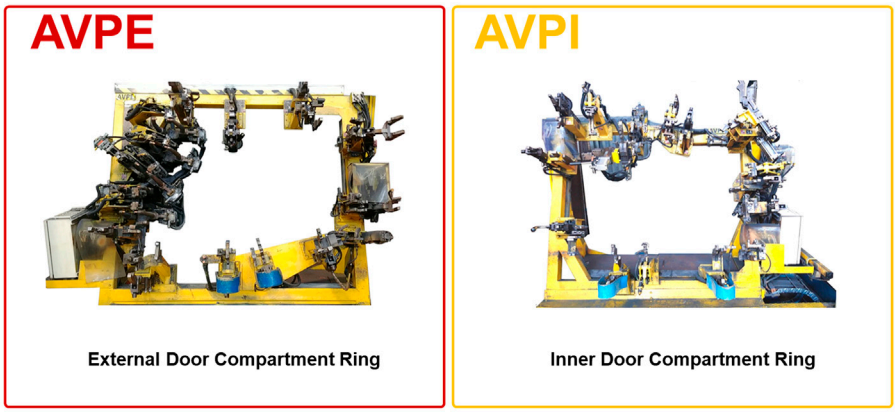

FIGURE 1 | Graphical representation of the External Door Compartment Ring (AVPE) and Inner Door Compartment Ring (AVPI) machinery under examination of the CNH Industrial Iveco (CHNi) Plant of Brescia (Italy).

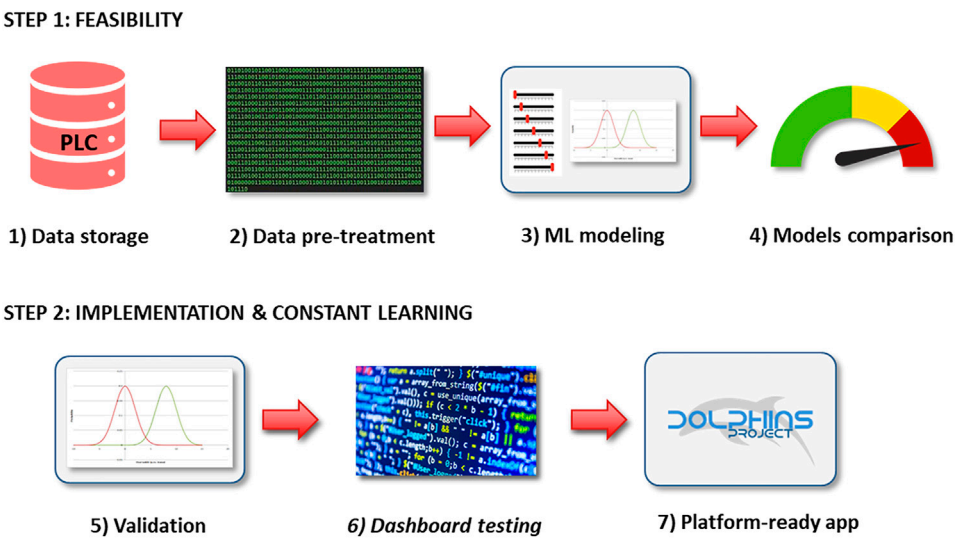

FIGURE 2 | Working steps of the DOLPHINS project.

connected to the PLCs, respectively. The signals were registered into the database in the form of sparse binary output (i.e., ON/OFF, with a prevalence of OFF results), indicating if the specific threshold values for each sensor are exceeding (i.e., ON) or not (i.e., OFF). Then, a pre-treatment step involving the binarization of the collected data (i.e., 0 means that the signal of a specific collected variable is OFF, whereas 1 means that the signal of the variable is ON) was performed. Furthermore, a categorical binary output indicating the status of the machinery (i.e., "Working" or "Stop") for each collection record was available, too.

In the present proof-of-concept study, the records collected from September 1, 2020, up to November 15, 2020, are shown as an example of two matrices of dimensions 210,307 × 177 and 199,077 × 154 for AVPE and AVPI machinery, respectively. Records are collected on the PLCs with a frequency of one second per record during the different work shifts. The whole study was composed of two developmental steps: the first step assessed the feasibility of the study, involving the acquisition of the data, their pre-pretreatment, the evaluation of several ML models, and the comparison of their performances, while the second step focused on the real-time implementation of the developed model within the plant, by testing the elected ML model on newly acquired data, updating the model with a scheduled frequency (approx. one month), and programming dashboards and platform-ready applications to be employed by the maintenance technicians during their everyday work. A

graphical representation of the developmental steps of the DOLPHINS project is reported in **Figure 2**.

Machine Learning Strategies

Several classification ML models were tested on the collected data to decide which algorithm best discriminates the records labeled as "Working" conditions from those labeled as "Stop" conditions of both the AVPE and the AVPI machinery. For this purpose, a benchmark analysis was performed by involving the following classification algorithms: k-Nearest Neighbors (kNN) (Massart et al., 1997), Logistic Regression (LogReg) (Cruyff et al., 2016), Linear Discriminant Analysis (LDA) (Massart et al., 1997; Martinez and Kak, 2001), Quadratic Discriminant Analysis (QDA) (Srivastava et al., 2007), Partial Least Squares–Discriminant Analysis (PLS-DA) (Ballabio and Consonni, 2013), Soft Independent Modelling of Class Analogies (SIMCA) (Wold and Sjostrom, 1977; Vanden Branden and Hubert, 2005), Naive Bayes (NB) (Cassidy, 2020), Support Vector Machine (SVM) (Hearst et al., 1998; Vapnik, 2000), Decision Trees (DT), Random Forest (RF) (Fratello and Tagliaferri, 2019), and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016; Guang et al., 2020). Since the acquired data are in the form of sparse binary matrices, the Sparse Logistic Principal Components Analysis (SL-PCA) approach was performed on the datasets before computing different ML approaches such as kNN, LogReg, LDA, QDA, and SIMCA. Since these algorithms can not evaluate sparse binary data properly, they were calculated on the Principal Components (PCs) provided by SL-PCA modeling. The SL-PCA algorithm introduced by Lee et al. (2010) involves an iterative weighted least squares algorithm and the calculated PCs were then used as new variables for the cited ML algorithms.

Both external validation and cross-validation were performed in the study. All the MDA models were tuned and trained on the data from September 1, 2020, up to October 31, 2020, using a repeated k-fold cross-validation strategy. For benchmark and tuning purposes, each tested algorithm got the same training set since the data had the same partitioning for every model and every cross-validation step. This data partitioning strategy was employed to compare the performance of the various models properly.

Grid search analysis was made to tune the number of components and the values of the hyperparameters of all the algorithms effectively. The use of an exhaustive grid search analysis (involving cross-validation, too) was performed to find the combination of hyperparameters that performed best for each ML model. Grid search analysis (rather than random search or sequential search) allowed us to monitor many values within the hyperparameters' space when looking for the best-performing values. Despite grid search being time-consuming and expensive, we decided to exploit it to achieve the best tuned and cross-validated ML models. SL-PCA tuning grid search involved evaluating the optimal number of k components (from 1 up to 30) and λ penalty parameter (from 0 up to 0.01). The best compromise for the goodness-of-fit and the model complexity was achieved by minimizing the Bayesian Information Criterion (BIC) (Lan et al., 2012). Grid search was performed for PLS-DA and SIMCA to find the optimal number of k components and latent variables in terms of Root Mean Square Error in Cross-Validation (RMSECV) (Massart et al., 1997). The optimal value of k-nearest neighbors for the kNN algorithm was varied from 1 up to 10. No tuning grid search was required for LDA, QDA, LogReg, and NB algorithms. In contrast, SVM tuning involved the grid search evaluation of four hyperparameters: *kernel* (involving the use of polynomial, radial, or sigmoid kernels), *degree* (related to the shape of the SVM decision boundaries for polynomial kernels, from 1 up to 3), *gamma* (describing the influence of the records on the location of the SVM decision boundaries, from 0.1 up to 10), and C (influencing the penalization of the records arranged within the margin of SVM boundary, from 0.1 up to 10) (Vapnik, 1995). DT tuning involved the grid search approach on four hyperparameters: *minsplit* (describing the minimum amount of records to be included into a node before splitting, from 1 up to 20), *minbucket* (defining the maximum depth of the calculated decision tree, from 1 up to 10), *cp* (indicating the minimum improvement in the performance of a node to allow a further split, from 0.01 up to 0.1), and *maxdepth* (describing the minimum amount of records that can be included into a leaf, from 1 up to 10). RF algorithm also involved the grid search tuning evaluation of four hyperparameters: *ntree* (expressing the number of trees in the forest model, from 10 up to 300), *mtry* (representing the number of variables to be randomly sampled at each node, from 5 up to 40), *nodesize* (defining the minimum number of records to be included into a node, 1 up to 10), and *maxnodes* (establishing the maximum number of leaves allowed in the model, from 2 up to 30) (Bischl et al., 2016; Fratello and Tagliaferri, 2019). XGBoost tuning grid search evaluated seven hyperparameters: *eta* (indicating the learning rate to avoid overfitting, from 0 up to 1), *gamma* (describing the minimum amount of splitting for a node, from 0 up to 20), *max_depth* (indicating how deeply each evaluated tree can grow, from 1 up to 5), *min_child_weight* (defining the level of impurity that is maintainable for a node, from 1 up to 10), *subsample* (describing the proportion of samples to be randomly selected when evaluating each tree, from 0 up to 1), *colsample_bytree* (evaluating the proportion of variables selected by each tree, from 0.1 up to 1), and *nrounds* (defining the number of trees that can be sequentially calculated within the model, from 10 up to 100) (Bischl et al., 2016; Guang et al., 2020). The tuning of the kNN, SVM, DT, RF, and XGBoost methods was evaluated in terms of mean misclassification error (MMCE), which represents the ratio between the number of records classified as belonging to a specific class different from their actual class (i.e., "Stop" or "Working") (Bischl et al., 2016; Probst et al., 2017). This parameter was calculated for all the ML algorithms and, therefore, the best tuning scenarios selected turned to be those providing the lowest MMCE value. A repeated k-fold cross-validation strategy involving a 10-fold CV approach repeated five times was performed when performing the grid search analysis. As a result, in summary, the best models were selected in average terms of Bayesian Information Criterion (BIC) for SL-PCA, Root Mean Square Error in Cross-Validation (RMSECV) for SIMCA and PLS-DA, and mean misclassification error (MMCE) for

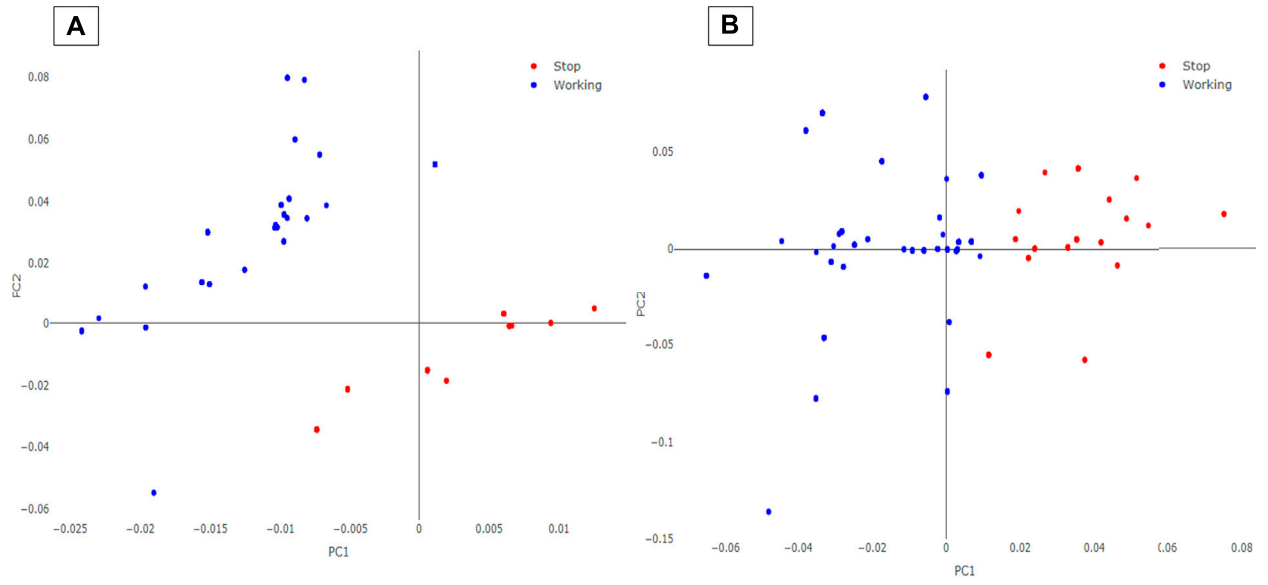

FIGURE 3 | SL-PCA scores plot for AVPE **(A)** and AVPI **(B)** machinery. The blue circles represent the records labeled as "Working" on the PLCs, while the red circles are the records acquired as "Stop."

kNN, SVM, DT, RF, and XGBoost. This approach, in our opinion, validates our entire model-building procedure, including the hyperparameter-tuning step.

The external validation was made by removing the records from November 1, 2020, up to November 15, 2020 from AVPE and AVPI original datasets. These data, consisting of matrices of dimensions 42,062 × 177 for AVPE and 33,180 × 154 for AVPI, were employed as a test set. Therefore, the results and the performance of the ML algorithms on the external validation test set were expressed using several metrics such as precision, recall, specificity, and F_{score}. In the present studio, the records classified as "Working" were considered positive samples (i.e., indicating proper functioning of the tested machinery). In contrast, the records classified as "Stop" were considered negative samples (i.e., indicating a malfunction or a breakdown of the machinery under examination). The "models" performance metrics were calculated as follows:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{TN + FP},$$

$$F_{score} = \frac{2 \times Recall \times Precision}{Recall + Precision},$$

where TP and FP represent the number of true positive and false positive records, whereas TN and FN indicate the number of true negative and false-negative records (Bischl et al., 2016). Finally, the model showing the best compromise among the different performance metrics was employed to develop the dashboards for predictive maintenance and the real-time evaluation of the new data collected in the plant. Nevertheless, an update of the ML models was scheduled with a frequency of 1 month (in parallel with the real-time analysis of the new data) to monitor the performance of the ML models on a larger amount of data.

Software

R statistical environment (version 4.0.2) (R Core Team, 2020) and R Studio Desktop IDE (version 1.4.1717) (RStudio Team, 2020) were used in this study. In addition, the following R packages were employed: *caret* (Kuhn, 2020), *dplyr* (Wickham et al., 2020), *ggplot2* (Wickham, 2016), *mdatools* (Kucheryavskiy, 2020), *mixOmics* (Rohart et al., 2017), *mlr* (Bischl et al., 2016), *parallel* (R Core Team, 2020), *parallelMap* (Bischl et al., 2020), *plotly* (Sievert, 2018), and *tidyverse* (Wickham et al., 2019). PLS-DA modeling was performed using the R codes available at (Github, 2013).

RESULTS AND DISCUSSION

Tuning and Benchmark Analysis

SL-PCA modeling indicated, as optimum, a tuning of 6 PCs with a λ value of the penalty parameter equal to 0.0025 for the AVPE data 5 PCs and λ equal to 0.0020 for the AVPI data. Examples of the scores plots of the SL-PCA models on the AVPE and AVPI machinery training datasets are reported in **Figure 3**. As it can be seen, a distinct separation is observed between the "Working" and the "Stop" samples in the space modeled by the new PCs. Since PCA is an exploratory data analysis algorithm, these results suggest that the classification task focused on predicting the operative conditions and the behaviors of the machinery

TABLE 1 | MMCE values of all the tested ML algorithms for AVPE and AVPI machinery training datasets.

ML models	AVPE (MMCE)	AVPI (MMCE)
KNN	0.073	0.075
LogReg	0.114	0.093
LDA	0.099	0.127
QDA	0.085	0.106
PLS-DA	0.171	0.216
SIMCA	0.034	0.052
NB	0.210	0.194
SVM	0.052	0.081
DT	0.226	0.102
RF	0.083	0.135
XGBoost	0.097	0.143

TABLE 2 | SIMCA performance metrics for AVPE and AVPI machinery test datasets.

Machinery	Precision	Recall	Specificity	F_{SCORE}
AVPE	0.977	0.944	0.844	0.960
AVPI	0.985	0.962	0.899	0.973

under examination might be feasible. Therefore, the calculated PCs were used as new features for the following ML classification algorithms: kNN, LogReg, LDA, QDA, and SIMCA.

The results for all the evaluated ML algorithms are expressed in MMCE for AVPE and AVPI machinery in **Table 1**. Further details about the tuning results for all the models are reported in the Supplementary Material (**Supplementary Table S1**). As shown in **Table 1**, SIMCA modeling (preceded by SL-PCA processing) provided the lowest results in MMCE. Therefore, this approach was selected for further testing with the external validation data and the implementation within an on-purpose developed dashboard to be used at the shopfloor level by the maintenance technicians of the plant.

SIMCA Model

The external validation dataset involving the AVPE and the AVPI data from November 1, 2020, up to November 15, 2020, were predicted by the developed SIMCA model. Hotellings T^2 vs. Q residuals plots for AVPE and AVPI test sets can be observed in **Figure 4**. Again, a satisfactory separation is observed between the "Working" and the "Stop" records collected by the PLCs during the period under examination. Although some records are still misclassified (mainly false negatives, i.e., false "Stop" predictions), the performance of the SIMCA model appears robust for both AVPE and AVPI machinery, thus suggesting the use of this approach for predictive maintenance purposes.

SIMCA prediction results are expressed in precision, recall, specificity, and F_{score} for both the types of machinery under examination, as reported in **Table 2**. These evaluations were made for all the ML algorithms, but the results turned to be lower than those obtained by the SIMCA model (results not reported here). SIMCA model provided optimal results for all the metrics under examination. However, specificity turned to be the metric with the lowest value; this result may be due to the lower number of "Stop" occurrences collected by the PLCs. The machinery under examination does not stop frequently, and several recorded "Stop" instances can be defined as micro-stops since they show a downtime lower than 1 minute. Moreover, the number of "Stop" records collected by the PLCs is only around 5% of the data. Our opinion is that the model's performance might be improved further by updating the training sets in a scheduled way (approx. one month) and collecting new data, especially those related to "Stop" records. Since the approach involving SL-PCA and SIMCA algorithms provided optimal and robust

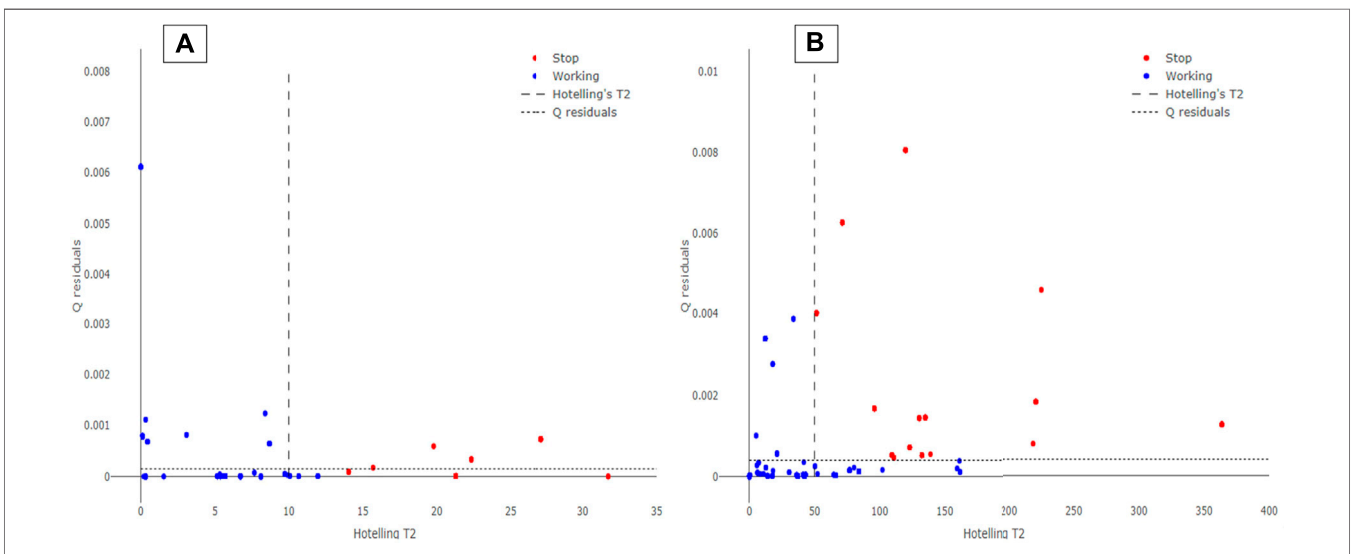

FIGURE 4 | Hotelling's T^2 vs. Q residuals plot for AVPE **(A)** and AVPI **(B)** machinery. The blue circles represent the records labeled as "Working" on the PLCs, while the red circles are the records acquired as "Stop." The dotted line indicates the 95% Hotelling's T^2 limit, while the dashed line represents the 95% Q residuals limit.

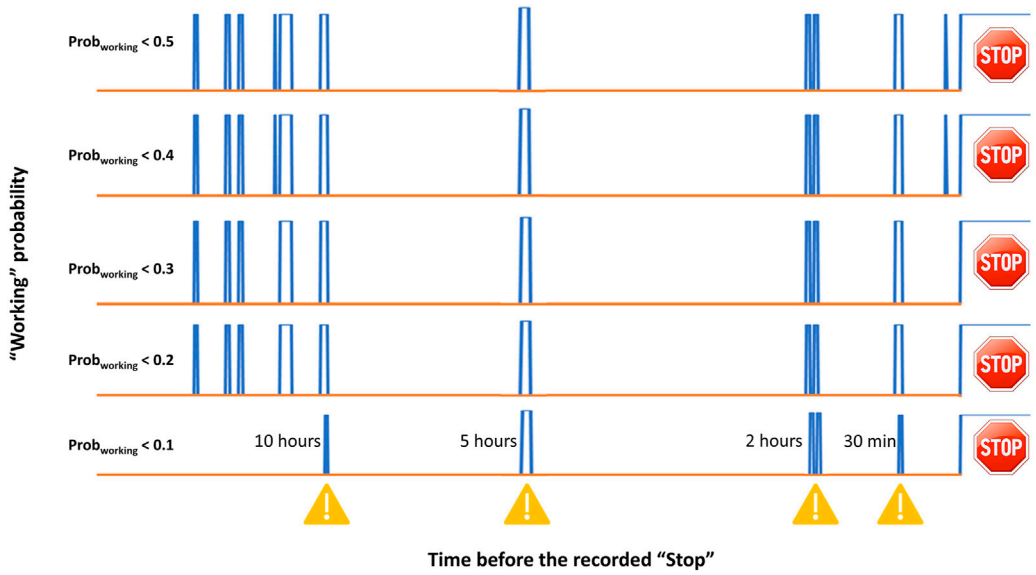

FIGURE 5 | Transient of SIMCA $Prob_{working}$ monitored over the time for the AVPE machinery before a breakdown. The x-axis represents the time before a breakdown event (occurring on the right side of the plot), while the blue lines represent the coded transient at different $Prob_{working}$ thresholds (i.e., 0.5, 0.4, 0.3, 0.2 and 0.1).

classification performance, this method was implemented into a dashboard to perform real-time predictive maintenance in the plant.

Dashboard Implementation

SIMCA algorithm provides several advantages for the development of a real-time predictive maintenance approach. Firstly, no assumptions are made on the probability distributions of the features under examination, which allows analyzing the new PCs from the SL-PCA model reliably. Secondly, since each class (i.e., "Working" and "Stop" records) is modeled independently, it is possible to obtain and predict the information about the classification probability of a certain record when introduced into the trained model. Thirdly, the SIMCA approach allows the maintenance technicians to identify the sensor or the single component of the machinery to consider for intervention before the occurrence of an incoming fault. Thanks to the evaluation of Hotelling's T^2 and Q residuals contribution plots provided by the combined SL-PCA and SIMCA algorithms, it is possible to recognize the critical signals recorded on the PLCs. An indication of the severity of the recorded signals was also implemented by calculating the logarithm (base 10) of the maximum Hotelling's T^2 and Q residuals contribution (in absolute value) and normalizing it on a scale from 0 up to 100%. However, this approach is still under evaluation since the amount of recorded breakdown events is relatively low.

Finally, the probability of classifying a new record as "Working" ($Prob_{working}$) is inferred for all the new records collected on the PLCs. **Figure 5** displays the transient of $Prob_{working}$ over time. The example reported in **Figure 5** shows the fluctuation of such probability before a specific breakdown occurred (on the right part of the plot). The x-axis represents the time before the occurrence of the stop of the machinery (in this case, AVPE), while the y-axis shows a binary output related to $Prob_{working}$. As a rule of thumb, it was established that if $Prob_{working}$ turns higher (or equal) than 0.5, the record is classified as "Working" and the transient is set to 0. On the other hand, if $Prob_{working}$ turns to be lower than 0.5, the record is classified as "Stop" and the transient is set to 1. The indication of a probable malfunction of the AVPE machinery was observed, in this case, 10 h before the breakdown. Other alerts were predicted 5, 2 h, and 30 min before the adverse event. However, the number of false "Stop" occurrences (i.e., false-negative records) might be rather high, as also remarked by the specificity values reported in **Table 2**. Again, this might be ascribed to the necessity of collecting new data and updating the SIMCA models (or the other tested ML algorithms). Nevertheless, further tuning of the employed decoding was tested. As it can be seen in **Figure 5**, different thresholds of $Prob_{working}$ were evaluated (e.g., 0.4, 0.3, 0.2, and 0.1 thresholds) to diminish the number of false "Stop" occurrences and improved sensitivity values (approx. 0.89 and 0.94 for AVPE and AVPI, respectively) were found using a $Prob_{working}$ threshold of 0.1. This further refining of the algorithms is still under examination and will be monitored over time. Furthermore, this approach allows providing the maintenance technician a tool capable of predicting a breakdown event before its occurrence. In fact, by analyzing the transients of $Prob_{working}$ monitored over time, it was observed that reliable alerts occurred in the range between 30 min and 3 h before the breakdown. At the current stage, it is still not trustworthy to provide $Prob_{working}$ with a confidence interval in terms of time before the occurrence of the stop event since the number of

"Stop" records is still limited. However, further analyses will be made on $Prob_{working}$ over time to estimate such a parameter reliably. An example of the developed dashboard is shown in **Supplementary Figure S2**.

CONCLUSION

DOLPHINS project represents a proof-of-concept and low-cost tool to perform reliable real-time predictive maintenance. It combines ML technology-driven algorithms with the evaluation of historical datasets that have never been interpreted using a multivariate data analysis approach. The algorithm involving SL-PCA and SIMCA has now been implemented by the automotive plant of CNHi Iveco (Brescia, Italy) at the shop floor level efficiently, and the number of failures and breakdown events has significantly diminished since the commissioning of the project.

This project allowed the development of an automated dashboard that shows the operator, in real-time, and the current instrumentation's operating conditions and, if signals arrive at the PLC, indicates the severity and probability that these lead to a stop. This predictive maintenance approach has numerous advantages, including 1) a meager impact in terms of costs (data already available are used); 2) the possibility of physically interpreting the information; 3) the possibility of not having to stop the production process; 4) the transversality of the application of Machine Learning also to other components and instrumentation within the plant.

At the current stage, the DOLPHINS algorithm can run on edge and cloud systems and conventional plant infrastructure. For this reason, the future perspectives of this project will focus on converting the DOLPHINS algorithm into a multiplatform application to raise its scalability on other types of machinery and plants. However, DOLPHINS is now equipment-oriented, and all the steps involving the tuning, training, and testing of the ML algorithms are required to develop a robust real-time predictive maintenance strategy. Therefore, a constant and frequent update of the databases and the ML models have to be scheduled to obtain reliable results and reach the goal of near-zero downtime.

.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors without undue reservation.

AUTHOR CONTRIBUTIONS

EA contributed to the conceptualization, methodology, visualization, investigation, models validation, and preparation of the original draft. VL and RP were responsible for the conceptualization, visualization, data curation, and software validation. FP contributed to the data curation and software validation. PF was responsible for the conceptualization, methodology, supervision, visualization, and reviewing and editing of the manuscript. MB contributed to the conceptualization, methodology, supervision, and reviewing and editing of the manuscript. MV was responsible for the supervision, reviewing and editing of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2021.734132/full#supplementary-material

REFERENCES

Ballabio, D., and Consonni, V. (2013). Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* 5, 3790. doi:10.1039/c3ay40582f

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al. (2016). {mlr}: Machine Learning in R. *J. Mach. Learn. Res.* 17, 1–5. Available at: https://jmlr.org/papers/v17/15-066.html.

Bischl, B., Lang, M., and Schratz, P. (2020). parallelMap: Unified Interface to Parallelization Back-Ends. Available at: https://cran.r-project.org/package=parallelMap (Accessed June 30, 2021).

Cassidy, C. (2020). Parameter Tuning Naïve Bayes for Automatic Patent Classification. *World Patent Inf.* 61, 101968. doi:10.1016/j.wpi.2020.101968

Chen, T., and Guestrin, C. (2016). "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, August 13–17, 2016 New York, NY, USA: ACM, 785–794. doi:10.1145/2939672.2939785

Cruyff, M. J. L. F., Böckenholt, U., van der Heijden, P. G. M., and Frank, L. E. (2016). A Review of Regression Procedures for Randomized Response Data, Including Univariate and Multivariate Logistic Regression, the Proportional Odds Model and Item Response Model, and Self-Protective Responses. *Handbook Stat.* 34, 287–315. doi:10.1016/bs.host.2016.01.016

Cugno, M., Castagnoli, R., and Büchi, G. (2021). Openness to Industry 4.0 and Performance: The Impact of Barriers and Incentives. *Technol. Forecast. Soc. Change* 168, 120756. doi:10.1016/j.techfore.2021.120756

Elsisi, M., Tran, M.-Q., Mahmoud, K., Lehtonen, M., and Darwish, M. M. F. (2021). Deep Learning-Based Industry 4.0 and Internet of Things towards Effective Energy Management for Smart Buildings. *Sensors* 21, 1038. doi:10.3390/s21041038

Enyoghasi, C., and Badurdeen, F. (2021). Industry 4.0 for Sustainable Manufacturing: Opportunities at the Product, Process, and System Levels. *Resour. Conservation Recycling* 166, 105362. doi:10.1016/j.resconrec.2020.105362

Fratello, M., and Tagliaferri, R. (2019). "Decision Trees and Random Forests," in *Encyclopedia of Bioinformatics and Computational Biology* (Amsterdam: Elsevier), 374–383. doi:10.1016/B978-0-12-809633-8.20337-3

Gallo, T., Cagnetti, C., Silvestri, C., and Ruggieri, A. (2021). Industry 4.0 Tools in Lean Production: A Systematic Literature Review. *Proced. Comput. Sci.* 180, 394–403. doi:10.1016/j.procs.2021.01.255

Gentner, S. (2016). Industry 4.0: Reality, Future or Just Science Fiction? How to Convince Today's Management to Invest in Tomorrow's Future! Successful Strategies for Industry 4.0 and Manufacturing IT. *Chim. Int. J. Chem.* 70, 628–633. doi:10.2533/chimia.2016.628

Ghobakhloo, M., Fathi, M., Iranmanesh, M., Maroufkhani, P., and Morales, M. E. (2021). Industry 4.0 Ten Years on: A Bibliometric and Systematic Review of

Concepts, Sustainability Value Drivers, and success Determinants. *J. Clean. Prod.* 302, 127052. doi:10.1016/j.jclepro.2021.127052

Github (2013). SparseLogisticPCA. Available at: https://github.com/andland/SparseLogisticPCA (Accessed June 30, 2021).

Goldman, C. V., Baltaxe, M., Chakraborty, D., and Arinez, J. (2021). Explaining Learning Models in Manufacturing Processes. *Proced. Comput. Sci.* 180, 259–268. doi:10.1016/j.procs.2021.01.163

Guang, P., Huang, W., Guo, L., Yang, X., Huang, F., Yang, M., et al. (2020). Blood-based FTIR-ATR Spectroscopy Coupled with Extreme Gradient Boosting for the Diagnosis of Type 2 Diabetes. *Medicine (Baltimore).* 99, e19657. doi:10.1097/MD.0000000000019657

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines. *IEEE Intell. Syst. Their Appl.* 13, 18–28. doi:10.1109/5254.708428

Jamwal, A., Agrawal, R., Sharma, M., Kumar, V., and Kumar, S. (2021). Developing A Sustainability Framework for Industry 4.0. *Proced. CIRP* 98, 430–435. doi:10.1016/j.procir.2021.01.129

Jeske, T., Würfels, M., and Lennings, F. (2021). Development of Digitalization in Production Industry - Impact on Productivity, Management and Human Work. *Proced. Comput. Sci.* 180, 371–380. doi:10.1016/j.procs.2021.01.358

Kucheryavskiy, S. (2020). Mdatools - R Package for Chemometrics. *Chemometrics Intell. Lab. Syst.* 198, 103937. doi:10.1016/j.chemolab.2020.103937

Kuhn, M. (2020). Caret: Classification and Regression Training. Available at: https://cran.r-project.org/package=caret (Accessed June 30, 2021).

Lan, W., Wang, H., and Tsai, C.-L. (2012). A Bayesian Information Criterion for Portfolio Selection. *Comput. Stat. Data Anal.* 56, 88–99. doi:10.1016/j.csda.2011.06.012

Lee, C., and Lim, C. (2021). From Technological Development to Social advance: A Review of Industry 4.0 through Machine Learning. *Technol. Forecast. Soc. Change* 167, 120653. doi:10.1016/j.techfore.2021.120653

Lee, S., Huang, J. Z., and Hu, J. (2010). Sparse Logistic Principal Components Analysis for Binary Data. *Ann. Appl. Stat.* 4, 1579–1601. doi:10.1214/10-AOAS327SUPP

Martinez, A. M., and Kak, A. C. (2001). PCA versus LDA. *IEEE Trans. Pattern Anal. Machine Intell.* 23, 228–233. doi:10.1109/34.908974

Massart, D. L., Vandeginste, B. G. M., Buydens, J. M. C., de Jong, S., Lewi, P. J., Smeyers-Verberke, J., et al. (1997). *Handbook of Chemometrics and Qualimetrics: Part B. First Edit*. Netherlands: Elsevier Science Amsterdam.

Probst, P., Au, Q., Casalicchio, G., Stachl, C., and Bischl, B. (2017). Multilabel Classification with R Package Mlr. Available at: http://arxiv.org/abs/1703.08991 (Accessed June 30, 2021). doi:10.32614/rj-2017-012

R Core Team (2020). R: A Language and Environment for Statistical Computing. Available at: https://www.r-project.org/ (Accessed June 30, 2021).

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration. *Plos Comput. Biol.* 13, e1005752. doi:10.1371/journal.pcbi.1005752

RStudio Team (2020). RStudio: Integrated Development Environment for R. Available at: http://www.rstudio.com/ (Accessed June 30, 2021).

Sievert, C. (2018). Plotly for R. Available at: https://plotly-r.com (Accessed June 30, 2021).

Srivastava, S., Edu, S. W., Gupta, M. R., Edu, G. W., and Frigyik, B. A. (2007). Bayesian Quadratic Discriminant Analysis. *J. Mach. Learn. Res.* 8, 1277–1305.

Vanden Branden, K., and Hubert, M. (2005). Robust Classification in High Dimensions Based on the SIMCA Method. *Chemometrics Intell. Lab. Syst.* 79, 10–21. doi:10.1016/j.chemolab.2005.03.002

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. New York, NY: Springer New York. doi:10.1007/978-1-4757-3264-1

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Wankhede, V. A., and Vinodh, S. (2021). Analysis of Industry 4.0 Challenges Using Best Worst Method: A Case Study. *Comput. Ind. Eng.* 159, 107487. doi:10.1016/j.cie.2021.107487

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Softw.* 4, 1686. doi:10.21105/joss.01686

Wickham, H., François, R., Henry, L., and Müller, K. (2020). Dplyr: A Grammar of Data Manipulation. Available at: https://cran.r-project.org/package=dplyr (Accessed June 30, 2021).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. Available at: https://ggplot2.tidyverse.org.

Wold, S., and Sjöström, M. (1977). "SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy," in *Chemometrics, Theory and Application*. Editor B. R. Kowalski (Washington, DC: ACS Symp. Ser.), 52, 243–282. doi:10.1021/bk-1977-0052.ch012

Conflict of Interest: Authors VL and RP were employed by RADA Snc. Authors FP and PF were employed by CNH Industrial.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 07 October 2021
doi: 10.3389/fchem.2021.733331

Routine Monitoring of Instrument Stability in a Milk Testing Laboratory With ASCA: A Pilot Study

Michel K. Nieuwoudt[1,2,3,4,5]*, *Cannon Giglio*[1,2,3], *Federico Marini*[6], *Gavin Scott*[5] and *Stephen E. Holroyd*[7]*

[1]The Photon Factory, The University of Auckland, Auckland, New Zealand, [2]School of Chemical Sciences, The University of Auckland, Auckland, New Zealand, [3]The MacDiarmid Institute for Advanced Materials and Nanotechnology, Wellington, New Zealand, [4]The Dodd-Walls Centre for Photonic and Quantum Technologies, Dunedin, New Zealand, [5]Fonterra On-farm R and D, Hamilton, New Zealand, [6]Dipartimento di Chimica, Universita di Roma "La Sapienza", Rome, Italy, [7]Fonterra Research and Development, Palmerston North, New Zealand

Mid-infrared spectroscopy has been developed as a reliable and rapid tool for routine analysis of fat, protein, lactose and other components in liquid milk. However, variations within and between FTIR instruments, even within the same milk testing laboratory, present a challenge to the accuracy of measurement of particularly minor components in the milk, such as individual fatty acids or proteins. In this study we have used Analysis of variance–Simultaneous Component Analysis (ASCA), to monitor the spectral variation between and within each of four different FOSS FTIR spectrometers over each week in an independent milk testing laboratory over 4 years, between August 2017 and March 2021 (223 weeks). On everyday of each week, spectra of the same pilot milk sample were recorded approximately every hour on each of the four instruments. Overall, variations between instruments had the largest effect on spectral variation over each week, making a significant contribution every week. Within each instrument, day-to-day variations over the week were also significant for all but two of the weeks measured, however it contributed less to the variance overall. At certain times other factors not explained by weekday variation or inter-instrument variation dominated the variance in the spectra. Examination of the scores and loadings of the weekly ASCA analysis allowed identification of changes in the spectral regions affected by drifts in each instrument over time. This was found to particularly affect some of the fatty acid predictions.

Keywords: ASCA, quality control, milk testing, instrument stability, standardization, FTIR spectroscopy, analysis of variance with simultaneous component analysis

OPEN ACCESS

Edited by:
Cosimino Malitesta, University of Salento, Italy

Reviewed by:
Ofélia Anjos, Instituto Politecnico de Castelo Branco, Portugal
Alessandro Ulrici, University of Modena and Reggio Emilia, Italy

***Correspondence:**
Michel K. Nieuwoudt
m.nieuwoudt@auckland.ac.nz
Stephen E. Holroyd
steve.holroyd@fonterra.com

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *30 June 2021*
Accepted: *13 September 2021*
Published: *07 October 2021*

Citation:
Nieuwoudt MK, Giglio C, Marini F, Scott G and Holroyd SE (2021) Routine Monitoring of Instrument Stability in a Milk Testing Laboratory With ASCA: A Pilot Study. Front. Chem. 9:733331. doi: 10.3389/fchem.2021.733331

INTRODUCTION

The goal of quantitative mid-infrared (MIR) analysis is to reproduce the analytical results achieved with accepted standard reference methods. The quantitative analysis of milk components from MIR spectra is based on the direct proportionality between the intensities of the absorbance bands for each component and their concentrations and the path length through the sample. The accuracy of this measurement requires routine calibration of the spectrometers with pre-analysed milk (with chemical reference tests). Signal variations in the interferometer within an instrument over time, between different instruments and between different types of instruments can alter the shapes, intensities and relative intensities of the vibrational mode bands (Pelletier, 2003) which can affect the

prediction accuracy, particularly of minor milk components such as individual fatty acids. The accuracy of a predictive calibration is affected both by instrumental factors (Smith et al., 1993) and by the characteristics of the materials used to calibrate the instruments (Kaylegian et al., 2006). Already in the early use of MIR spectral techniques, inherent issues in the stability of predictions between instruments and over time were shown (Biggs, 1978). However, differences in results obtained from different laboratories can also occur because of differences between the reference methods used, and because of failure to achieve good calibrations (Biggs, 1978). In a report on performance of the older generation of milk analysers, it was found that the main problems affecting calibration and accuracy of predictions were inaccurate reference tests, air incorporation, homogenizer inefficiency, mechanical wear, sample cell and filter system, electronics and mechanical maintenance and operator errors (Young, 1978). The newer milk analysers have been engineered with improved designs to minimize these factors; however, some still persist. The small variations in spectra caused by variations in spectrometer parameters such as light source intensity, detector sensitivity and laser stability, and in laboratory environment such as temperature, vibrations, humidity, are minimized by a procedure called Zero-setting (Foss Electric, Hillerød, Denmark) (Hansen, 2020), and by weekly calibration adjustment for fat, protein, lactose and total solids. Manufacturers of other MIR spectrometers used for routine milk analysis similarly incorporate one or more methods to reduce variations within and between instruments.

Differences between the spectrometers, even from the same manufacturer and model in the same laboratory are minimized by routine calibration adjustments. Weekly adjustments on the calibration models are performed to correct or adjust the prediction models used on the different instruments in the laboratory. These adjustments compensate for any week-to-week changes in path length, temperature and humidity variations, mechanical wear, sample handling and minor changes in detector, source and the mechanical and electronic performance. (Young, 1978). Such changes can result in changes in peak intensity or band shape which would render the prediction results inaccurate.

Standardization of MIR spectrometers is particularly necessary for exchange of MIR spectral databases across laboratories and countries. The standardization procedure corrects for systematic variations in intensity due to random variations in linearity of the detectors, or in the relative intensity across the wavelength range from different instrument manufacturers and models. Within the same instrument, the standardization procedure also corrects for path length changes with time, due to erosion of sample cell windows made of CaF_2 and due to window contamination in the case of diamond sample cell windows. It also corrects for shifts in frequency (or wavenumber), however these are random and if present, would occur on a very minor scale as this is an effect of laser fluctuation, source variation and detector instability, all of which are usually minor compared to other instrument variations.

For the FOSS MilkoScan™ FT1, FT2, FT120, FT + , MilkoScan™ 7 and FT6000 milk analysers, a patented standardization procedure has been developed for regular use which applies a slope and intercept adjustment to the spectra recorded on an instrument to correct for wavenumber (frequency) shift, changes in intensity and changes in linearity due to instrument variations over time. The procedure involves recording a spectrum of a standardization liquid and comparing the intensities and wavenumber positions at two selected wavelengths with those in a standard spectrum ("Master or Gold equalizer spectrum") stored on the instrument. Any differences between the spectrum of the standardization liquid and gold equalizer spectrum are corrected for by applying four correction factors: A and B for intensity variation, and α and β for wavenumber shifts (Hansen, 2014).

An alternative and non-instrumental standardization method called Piecewise Direct Standardization (PDS) (Wang et al., 1991) was recently used to standardize spectra of samples measured with different makes of instruments (Delta, Bentley, and FOSS) inside a European dairy network. (Grelet et al., 2015a; Grelet et al., 2015b). This standardization aimed to allow spectra from different sources to be pooled and matched to physiological data in a common database to create calibrations predicting cow fertility, health and environmental and feeding indicators. The application of PDS on spectra recorded on 21 different instruments in ten laboratories was found to significantly reduce the RMSE (Grelet et al., 2015b). However, this procedure requires a large amount of post-processing of the spectra and provides retrospective rather than time-based monitoring of instrument performance.

In this paper we describe an innovative approach that covers a different aspect of instrument standardization, namely the routine monitoring of faults or discrepancies in the MIR spectrometers in a single milk testing laboratory over time. This would have a complementary function to the calibration transfer and instrument standardization approaches by monitoring with time the instrument performance. The method relies on measuring spectral variations over time of a pilot sample of milk that is recorded on all the instruments in the laboratory at the same time. A spectrum of the same pilot milk sample is recorded approximately every hour on all instruments in the same laboratory over a period of a week; this is repeated every week using a fresh pilot sample. The effects of day-to-day variation within the individual spectrometers over the week and the variations between the individual laboratory spectrometers, are measured using ASCA (Analysis of variance - simultaneous component analysis) of the spectra (Jansen et al., 2005; Smilde et al., 2005). ASCA is a method used to determine which factors within a fixed effects experimental design are significant relative to the residual error and permits an ANOVA-like analysis even when there are many more variables than samples, as in the case of spectra (Smilde et al., 2005). In this study, the contributors to variation in the spectral intensities over the week were assumed to be instrument and weekday as the main factors, and the interaction between them. The purpose of the study was to explore whether ASCA could provide a useful tool for monitoring and comparing the performance of four MIR milk analysers in a single milk test laboratory in New Zealand. The ability of ASCA to measure changes or differences in the actual

spectral output from each instrument allows identification of the source of variation on a weekly basis and thus enable timeous and appropriate intervention.

MATERIALS AND METHODS

Pilot Milk Samples

The pilot milk sample was prepared by combining randomly selected milk samples from three or more different farms from different regions. Every week a new pilot sample was made up as a fresh sample. The aim of sourcing milk samples from different suppliers was to provide the most representative pilot sample, by averaging out milk compositions from different regions of NZ. The pilot sample was then preserved with bronopol and stored in a refrigerator to be used as the single pilot sample for all instruments over that week. For each subsequent week a new pilot sample was prepared. Approximately every hour, a sample of milk from the same pilot sample was introduced to each of the instruments in use, in between measurement of the routine milk samples. This occurred on each MIR analyzer in the laboratory over a period of 1 week from aliquots of the same pilot milk sample. Every week, a fresh pilot sample was prepared for the following week's measurements.

MIR Spectra

The MIR spectra of aliquots of the same pilot sample were recorded on between two and four FOSS milk spectrometers at any one time, named in this study as MS1, MS2 (FT6000 models with diamond sample cuvette windows), MS3, MS4 (FT + models with CaF_2 sampling windows), and MS7 and MS8 (MilkoScanTM 7 with CaF_2 sampling windows). Although MS1 and MS2 are the same FT6000 model type, MS3 and MS4 are of the same FT+ model type, and similarly MS7 and MS8 are both MilkoScanTM 7, each are individual instruments and will show differences in variations arising from their optical components. For example, the globar light source intensity, detector sensitivity/noise, homogenizer function or interferometer function. Even slight variations in these components of the optical bench will affect the spectral intensities to different extents. These differences are minimized by regular (approximately 6-weekly) instrument standardization procedures (Hansen, 2014). Variations in the predicted milk components are minimized by weekly calibration adjustments by the milk test laboratory of the slope and bias of fat and protein calibration models for a calibration set.

All spectra were recorded between 929 cm^{-1} and 5,000 cm^{-1} at spectral resolution 16 cm^{-1}, and ratioed against a water background. The spectra were transformed by an inverse log from transmittance to absorbance. Although the spectrum is measured over 929 to 5,000 cm^{-1}, many of these regions were not usable for measurement of milk components. This is mainly because of the intense absorption by water at specific frequencies, although subtracted out, results in random noise between 3,600–3,000 cm^{-1} and between 1,693–1723 cm^{-1}. These regions were exluded from the ASCA analysis. In addition, the region between 1785 cm and 1 to 2,600 cm^{-1} was also excluded weak inteference fringes are visible in the region, arising from internal reflection between the inside windows of the samplng cuvette. This region also includes absorption bands by atmospheric CO_2. These regions were therefore excluded in order to be able to identify variations as being due to changes in instrument parameters within or between instruments or due to other factors such as laboratory environment conditions.

Fat, Protein, C16:0 and C18:0 MIR Measurements in Pilot Milk Samples

In order to assess the influence of spectral variation on the MIR predictions of the components in the pilot milk MIR predictions, fat measurements (ranging between 3.13–6.54 g/ 100 ml) and true protein (3.17–4.61 g/100 ml) were selected as examples of major milk components. Also, two fatty acids were selected as minor milk components: the more abundant C16:0 (ranging from 0.94 to 2.04 g/ g milk) and C18:0 (ranging from 0.33 to 0.75 g/100 ml).

Data Analysis

The first step in ASCA is a decomposition of the variation for every variable (wavenumber) through ANOVA (Jansen et al., 2005; Zwanenburg et al., 2011). We set up a data matrix, X, for each week that contains the spectra of each instrument, and a design matrix that defines the instrument and weekday for each spectrum. An ANOVA is performed for every wavenumber in the FTIR spectra of each pilot sample (week) to determine whether the variation in the spectral data matrix is due to a weekday effect (milk changing or instrument varying over the week), instrument effect (difference between instruments), interactions between instrument and weekday, or other reasons such as noise not described by any of these effects (residual variation). So, for every variable (wavenumber) we define a main effect (the mean), factor effects (instrument and weekday), interaction effects (between instrument and weekday) and a noise or residual term. This results in the definition of different effect matrices:

$$\boldsymbol{X} = \boldsymbol{X}_{mean} + \boldsymbol{X}_{w} + \boldsymbol{X}_{i} + \boldsymbol{X}_{wi} + \boldsymbol{X}_{res} \quad (1)$$

where w = weekday, i = instrument, wi = interaction between instrument and weekday and res = residuals.

These matrices are made of identical copies of the mean profiles calculated by averaging all the replicates at the different levels of each factor or interaction. For instance, if a factor has two levels, half of the rows of the corresponding effect matrix will contain identical copies of the mean profile of the experiments in which the factor was at level 1; the other half will be made of the average of the remaining signals (i.e., those corresponding to level 2).

Once this decomposition has been done, the effect of the individual design terms is calculated as the sum of squares (SSQ_j) of the corresponding effect matrix, $\boldsymbol{X}_j$:

$$SSQ_j = \|\boldsymbol{X}_j\|^2 \quad j = i, w, wi \quad (2)$$

Accordingly, the portion of the total variance in $\boldsymbol{X}$, after centering, accounted for by any of the design terms can be

calculated by dividing SSQ_j by the sum of squares of the mean-centered data ($\boldsymbol{X} - \boldsymbol{X}_{mean}$). The contributions of a factor in the ASCA model can be summarized by dividing the sum of squares of a factor effect matrix by the sum of squares of the mean-centered data.

If a factor/interaction is found to have a significant effect (e.g., by means of permutation tests), PCA is then performed on the corresponding matrix $\boldsymbol{X}_j$ to correlate the effect to the variations observed in the spectroscopic profiles.

$$\boldsymbol{X}_j = \boldsymbol{T}_j \boldsymbol{P}_j^T + \boldsymbol{E}_j \qquad j = i, w, wi \tag{3}$$

where $\boldsymbol{T}_j$, $\boldsymbol{P}_j$ and $\boldsymbol{E}_j$ are the matrices of PCA scores, loadings and residuals, respectively, while the superscript ^{T} indicates matrix transposition. Additionally, to carry out multiple comparisons, when the number of levels for a factor is higher than two or, in general, to graphically visualize the significance of the effect of a design term, it is customary to calculate a new set of scores $\boldsymbol{T}_{j+res}$ by projecting the residual matrix onto the PC subspace of the factor/interaction of interest:

$$\boldsymbol{T}_{j+res} = \left(\boldsymbol{X}_j + \boldsymbol{X}_{res}\right)\boldsymbol{P}_j \qquad j = i, w, wi \tag{4}$$

For all the models, to evaluate the statistical significance of the effects, the calculated values of the sum of squares of the corresponding effect matrices were compared to their null distributions, non-parametrically estimated by means of permutation tests (Zwanenburg et al., 2011). Permutation tests were run on the spectra from each week to evaluate the significance of the effect of the different instruments and days in the week, and of their interaction, with 1,000 randomization per model; effects with $p < 0.05$ were deemed significant. The PCA scores and loadings of the corresponding effect matrices were used to highlight differences in the spectra, or changes over time, as influenced by instrument differences or weekday. Box plots were used to monitor differences in values predicted by the calibration models with time.

For each week's worth of data, outlier removal was performed prior to ASCA calculation, as there were often a small number of spectra with highly anomalous behavior. After calculating PCA models, outliers were identified based on the values of Hotelling T^2 and Q residuals using the R package "mdatools" (Kucheryavskiy, 2020; Kucheryavskiy, 2021). A square cutoff option was used in which samples with $T^2 > T^2_{\text{lim}}$ or $Q > Q_{\text{lim}}$ (Pomerantsev, 2008), with T^2_{lim} were calculated using the Hotelling T^2-distribution and Q_{lim} being calculated at a 99% confidence level based on the corresponding null distributions. These thresholds were chosen so as to include enough spectra to enable comparison of the number of outliers from each instrument, while excluding extreme values to avoid unduly influencing the results of ASCA modeling and sum of squares of the effects.

All computations were performed using the R programming language version 4.0.5 (R Foundation for Statistical Computing, Vienna) and the RStudio integrated development environment (RStudio Team, Boston). The "MetStaT" package was used for ASCA calculations and the package "ggplot2" was used for generating figures.

RESULTS

ASCA

A total of 223 weeks' worth of data comprised the full dataset which spanned from December 2016 to March 2021. Particularly in the winter season (June-August) when fewer milk samples were analysed, and in other periods when one or more instruments were under maintenance, there were not enough instruments active or not enough days in the week with sufficient measurements to perform ASCA. These were excluded from the analysis so that ASCA was performed on the remaining 177 of the 223 weeks.

The total sum of squares (TSSQ) for each of the 177 weeks, obtained from the ANOVA calculation of the ASCA algorithm are plotted in **Figure 1** over the time period December 2016 to March 2021. The TSSQ was adjusted for sample size to TSSQ (adj), as the number of samples over the measurement period varied each week between 202 and 1,324, depending on time of year (fewer samples in winter season) or whether instruments were undergoing maintenance. The plot of TSSQ (adj) in **Figure 1** shows the overall variance for every week for all the instruments in the laboratory and serves as a useful monitor of instrument performance and/or laboratory stability.

The shaded regions in the plot indicate changes in which laboratory instruments were used. Between weeks 1 and 83, the four instruments MS1, MS2, MS3, and MS4 were active, while over weeks 84–142 only instruments MS3, MS4 and MS7 were active. From weeks 143–223, the four instruments MS3, MS4, MS7, and MS8 were active. The dashed lines indicate the thresholds for one (0.0022), two (0.0044) and three (0.0066) standard deviations of all 177 weeks' TSSQ (adj) values. These thresholds can be selected to flag when the overall spectral variance deviates from the norm. The mean TSSQ (adj) (0.0015) is also indicated on **Figure 1** as a green dotted line.

For 130 of the 177 weeks (74%) the TSSQ (adj) was below the mean. In 28 of the 177 weeks (15%) the TSSQ (adj) was above one SD (σ) of the 177 weeks' TSSQ (adj) values. Of these 28 weeks, seven exceeded two SD's (4% of the 177 weeks) and six exceeded three SD's (3% of the 177 weeks), with 15 exceeding only one SD (8% of the 177 weeks). The weeks with TSSQ (adj) exceeding one, two or three SD's are labelled according to the major contribution from one or more of weekday effect (W), instrument effect (I), or residual effect (R). There was no correlation between the TSSQ (adj) exceeding one, two or three times the SD of the 177-weeks TSSQ (adj) values with time of year or with season.

Figure 2 shows the percentage contribution to the total SSQ (representing the total variance) for each week, by the instrument effect (black trace), weekday effect (green trace), weekday/instrument interaction (blue trace) and residual factors/noise (grey trace). Evident from the graph is that weekday changes originating from the sample itself or within each instrument, and weekday/instrument interactive effects contribute very little to the overall variance. Differences between instruments and residual variations form the major contributions; only three of the 177 weeks showed greater contribution from weekday variation than instrument effects. 53 of the weeks have

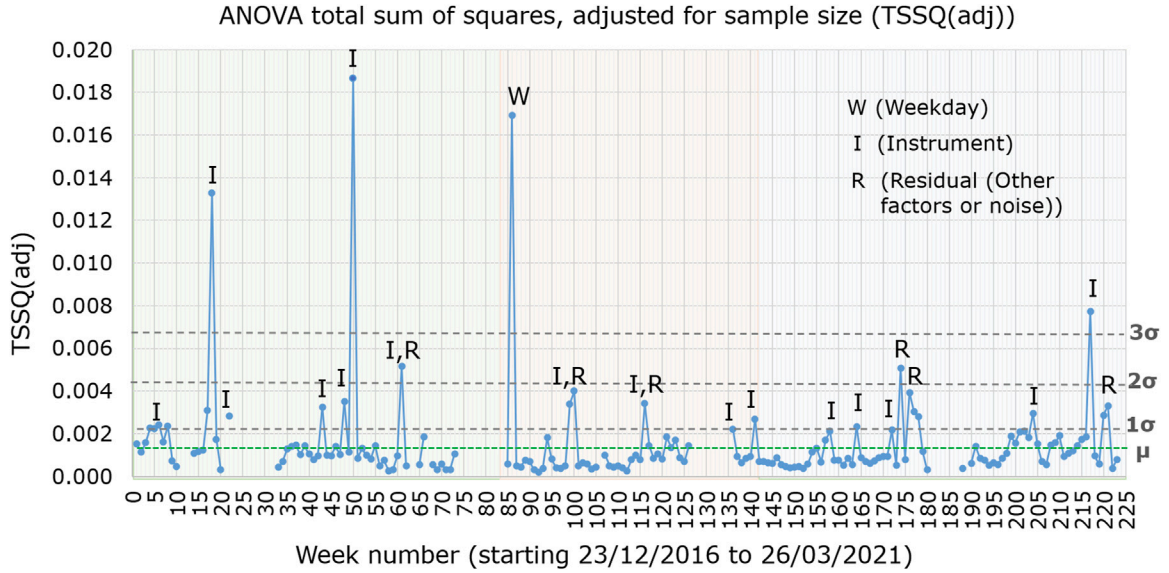

FIGURE 1 | Total sum of squares adjusted for sample size (TSSQadj), for instruments MS1, MS2, MS3, and MS4 from weeks 1–83 (green shaded region), for instruments MS3, MS4, and MS7 over weeks 84–142 (orange shaded region) and for instruments MS3, MS4, MS7, and MS8 over weeks 143–223 (blue shaded region). Weeks in which the TSSQ (adj) exceeded thresholds of one, two and three times the SD (σ) of the 177 weeks TSSQ (adj) values are labelled with the main contributing effects. The mean TSSQ (adj), labelled as μ, is represented by a green horizontal dotted line.

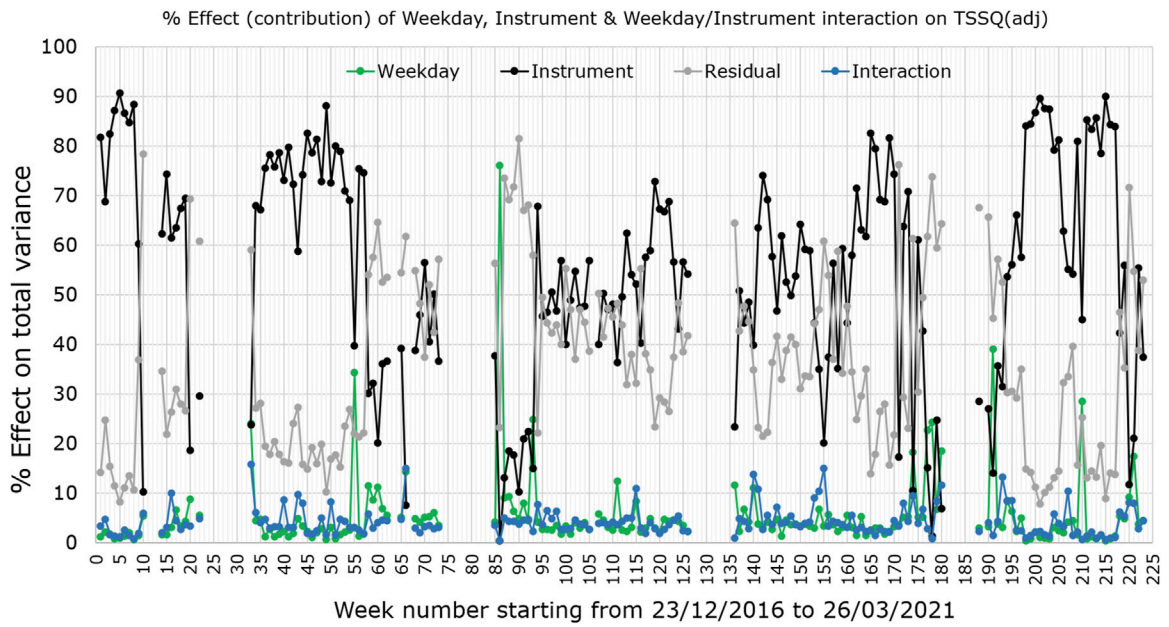

FIGURE 2 | Weekly percentage contribution of instruments (black trace), weekdays (green trace) and weekday/instrument interactions (blue trace), to the total variance (TSSQ, adjusted for sample numbers) in the pilot sample spectra. Also plotted is the residual effect (grey).

residual effects contributing more than either instrument or weekday effects to the TSSQ (adj).

In all 177 weeks measured with ASCA the contribution of the instrument effects to the TSSQ (adj) was significant, having $p < 0.05$. The weekday effects were also significant except for two of the weeks: with only weeks 137 and 188 having $p > 0.05$. Interaction effects between weekday and instrument were significant for all weeks except for 43 and 177. Those weeks exceeding one, two and three standard deviations of the 177 weeks' TSSQ (adj) values are listed in **Table 1**, with the number of SD's indicated in the fifth column from the right. Also listed are the total number of samples (N) for each week and the

TABLE 1 | List of weeks in which the TSSQ adjusted for sample number [TSSQ (adj)] exceeded one, two or three times the standard deviation of the 177 weeks, given in the column fifth before the last. N is the total number of samples in each week. The percentage that each of the factors contribute to each weekly total SSQ [TSSQ (adj)] are given as SSQ (day), SSQ (instr) and SSQ (inter) (weekday/instrument interaction). The residual variance due to other factors is SSQ (resid). The number of standard deviations (SDs) by which the SD's in each of MIR predicted values of fat, protein, C16:0 and C18:0 exceed the SD's of the 177 weeks by 1, 2, or 3 SD's are given on the four rightmost columns of the table.

Week No	No. samp	SSQ (day)	SSQ (instr)	SSQ (inter)	SSQ (resid)	TSSQ	TSSQ (adj)	No of SDs	SDs[a] (fat)	SDs[a] (prot)	SDs[a] (C16: 0)	SDs[a] (C18: 0)
4	853	0.8	87.2	1.35	11.46	1.92	0.0023	1	1	3	3	3
5	712	0.88	90.65	1.17	8.17	1.57	0.0022	1	–	1	2	3
6	765	1.39	86.65	2.53	11.1	1.82	0.0024	1	–	1	3	3
8	907	0.71	88.44	0.82	10.61	2.12	0.0023	1	1	1	2	3
17	578	6.67	63.58	4.5	30.9	1.69	0.0029	1	1	1	2	3
18	725	3.36	67.46	2.7	28.01	9.48	0.0131	3	–	3	3	3
22	252	5.46	29.55	4.81	60.75	0.71	0.0028	1	3	1	3	3
43	1125	4.89	58.81	9.67	27.24	3.63	0.0032	1	–	2	2	3
48	1164	2.87	72.83	4.93	19.81	4.06	0.0035	1	–	2	3	3
50	1306	3.03	72.61	8.24	16.91	24.18	0.0185	3	–	3	3	3
61	469	6.86	36.08	4.57	52.6	2.42	0.0052	2	1	3	1	2
86	514	76.04	0.37	0.39	23.3	8.69	0.0169	3	3	1	3	3
99	1122	1.72	56.96	2.63	40.03	3.75	0.0033	1	1	1	1	3
100	1002	3.41	40.05	2.7	55.25	3.95	0.0040	1	1	1	2	3
116	848	2.21	40.25	2.96	55.26	2.87	0.0034	1	3	1	3	3
141	671	3.76	63.54	10.86	23.3	1.77	0.0026	1	–	1	1	3
164	747	2.24	35.11	4.03	58.76	2.22	0.0023	1	1	2	2	3
174	512	18.28	10.49	9.51	61.35	2.6	0.0051	2	–	1	2	3
176	504	5.16	42.66	6.79	49.39	1.94	0.0039	1	–	1	2	3
177	372	22.71	15.09	2.85	61.74	1.11	0.0030	1	–	–	2	3
178	334	24.32	1.35	0.75	73.85	0.94	0.0028	1	–	1	1	3
204	1144	3.12	79.21	5.77	13.06	3.34	0.0029	1	–	1	3	3
217	793	1.32	83.94	1.07	13.74	6.13	0.0077	3	–	2	3	2
220	749	9.16	11.72	8.14	71.66	2.13	0.0028	1	–	1	2	3
221	578	17.37	21.05	7.97	54.69	1.89	0.0033	1	–	1	2	3

[a]*Number of standard deviations exceeding the average SD of 177 weeks, by the MIR-predicted values of fat, protein, C16:0 and C18:0 in each of the 25 weeks having TSSQ (adj) > 1 SD of the 177 weeks.*

ASCA output of percentage contribution from each of the effects: weekday (SSQ day), instrument (SSQ instr), weekday/instrument interaction (SSQ inter), and the residual variance due to other factors and noise (SSQ resid).

Of interest is how the TSSQ (adj) affects the accuracy of the milk component MIR predictions of fat, true protein, C16:0 and C18:0. The columns on the right-hand side of **Table 1** give the number of SD's by which each of the MIR predicted values of fat, protein, C16:0 and C18:0 exceed the average 177-weeks SD of each component: i.e., by one, two or three SD's. Of the 25 weeks shown in **Table 1** having TSSQ (adj)>1SD, the number of weeks in which the SDs of the fat predictions overall exceed one or more SDs is 10. For true protein, 24 of the 25 weeks have predicted values with SD's exceeding the average SD by one or more, while for C16:0 and C18:0 this occurs in all 25 weeks. This is not unexpected, as predictions of more minor components would be expected to be more sensitive to variations between instruments or within each instrument through the week, whether this is due to changes in the pilot sample, other effects such as laboratory environment or instrument variations.

PCA of Weeks Exceeding Three Standard Deviations

The instrument scores and loadings plots from the ASCA can provide more information about the spectral variations over the weeks with TSSQ (adj) exceeding one or more standard deviations (SD). Of particular interest are the weeks showing TSSQ (adj) exceeding the three SD threshold. The scores and loadings of three of these weeks with such TSSQ (adj): weeks 50, 86, and 217, are shown as examples in **Figure 3** (week 50), **Figure 4** (week 86), and **Figure 5** (week 217). Also plotted are the spectra after mean centering and boxplots for the MIR-predicted components fat, true protein, C16:0 and C18:0. The mean and standard deviations are indicated in the boxplots for each component for that week, with a number of asterisks that indicate the number of SDs by which each predicted component exceeds the SD of the 177 weeks TSSQ (adj) values (* for 1 SD, ** for 2 SDs and *** for 3SDs). Breaks in the plots of the spectra and loadings show the spectral regions excluded from the ASCA analysis. The scales of the intensity axis of the box plots have been expanded over reduced regions to exclude extreme outliers, in order to more clearly compare the medians inter quartile ranges (IQR) and whiskers.

According to **Table 1**, the instrument effect in week 50 contributed 72.6% of the TSSQ (adj). The mean centered spectra in **Figure 3A** show clear differences in the spectral intensities, particularly between the MS3, MS4 instruments and MS1, MS2 instruments, and particularly in the spectral region 930–1,200 cm^{-1}. The PCA scores in **Figure 3C** show clear separation along PC1 for the two sets of instruments.

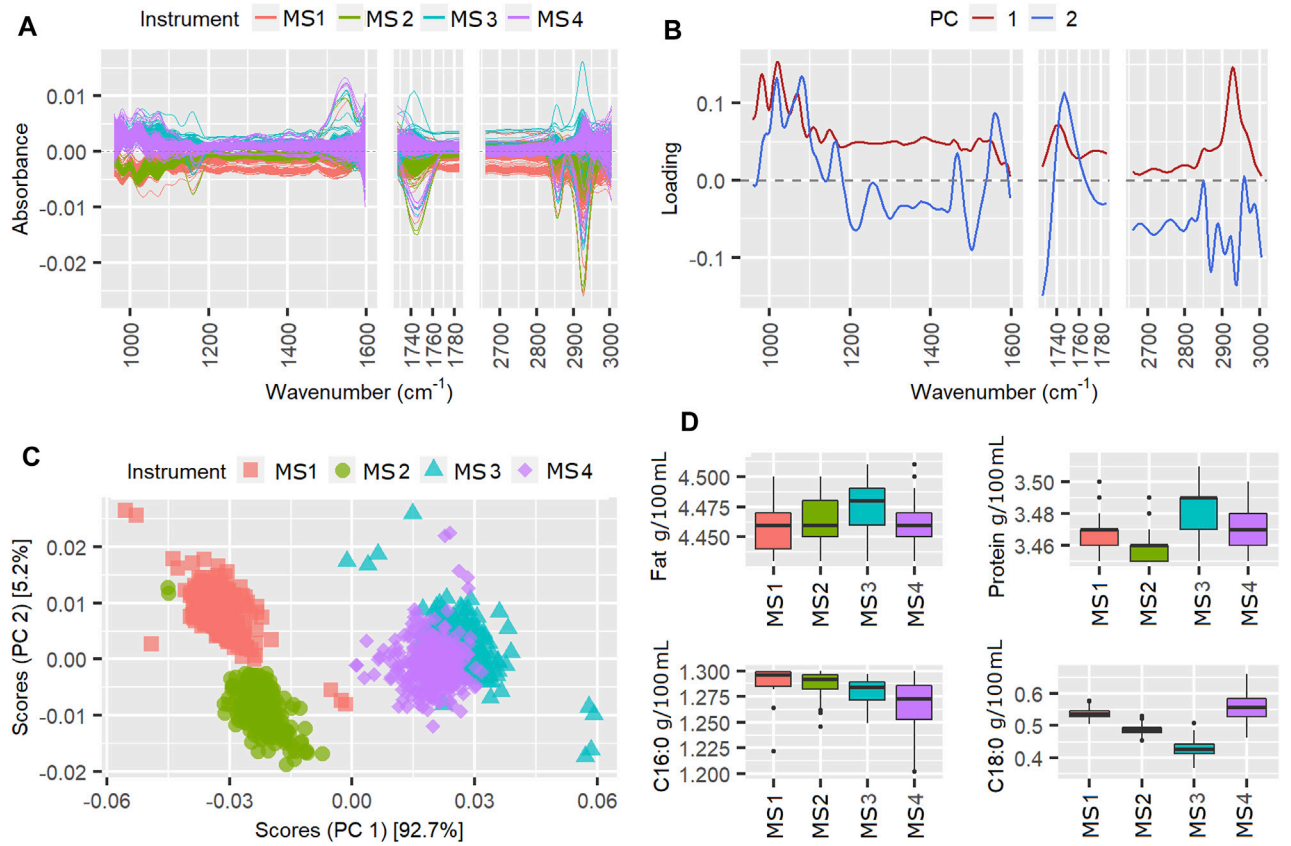

FIGURE 3 | Mean centered spectra **(A)**, instrument loading plots **(B)** and scores **(C)** for pilot spectra from week 50, in which the TSSQ (adj) exceeded 3 standard deviations. Boxplots **(D)** show the predicted fat, protein, C16:0 and C18:0 values from each instrument and their mean values and standard deviations (SDs), marked * for 1SD, ** for 2SDs and *** for 3SDs greater than the SD of the 177 weeks TSSQadj values.

This can be explained by differences in FOSS instrument models; MS3 and four are FT6000 models with CaF_2 windows while MS1 and MS2 are FT + models with diamond windows. Additionally, along PC2, MS1 and MS2 are more separated than MS3 and MS4. Ideally all four scores should overlap as the six-weekly standardization procedure adjusts the slave spectra of each instrument to match a master spectrum. The PC1 loading in **Figure 3B** shows that the main difference between the two spectrometer models are overall intensity, possibly due to pathlength differences, with the CaF_2 windows in MS3 and MS4 possibly eroded at this point to a slightly wider pathlength. The regular wavelike features in the loadings may be due to interference patterns from internal reflectance in the cell windows. These interference patterns were also seen in the spectral regions between 1730 and 2,650 cm^{-1} which were excluded from the analysis for this reason (besides this region displaying atmospheric CO_2 bands).

The boxplots of the MIR predicted components in **Figure 3D**, however, do not correspond with the PCA observations of higher fat for MS3 and MS4 compared with MS1 and MS2. This may be because the weekly calibration adjustments for fat have compensated for the spectral differences. The instrument differences for week 50 do, however, result in the true protein, C16:0 and C18:0 values for this week showing SDs three times higher than the 177-weeks average SD, while the fat predictions were not affected.

The results for week 86 are given in **Figure 4**. In week 86, the weekday effect contributed 76% to the TSSQ (adj). Being the winter season, only two spectrometers, MS3 and MS4 were active in this week.

The mean-centered spectra clearly show a subset of spectra from three of the weekdays that markedly differ from the others. This is also seen in the weekday scores plot. PC loading 1 is mostly represented by fat bands (C-H stretching of lipids 2,550–2,962 cm^{-1}, C=O stretching of fatty acids at 1745 cm^{-1} and C-O-C stretching of fatty acid esters at 1,160 cm^{-1}) (Grelet et al., 2015b). The separation of the scores according to these differentiates some of spectra in weekdays 1, 2, and three from the rest of the spectra in days 1, 2, and 3, and all the spectra in days 4–7. The weekday effect accounted for 76% of the variance compared to around 0.4% from instrument effects and weekday/instrument interactions. This implies that all both instruments, MS3 and MS4 underwent changes in weekdays one to three that resulted in a bigger effect than any differences between the

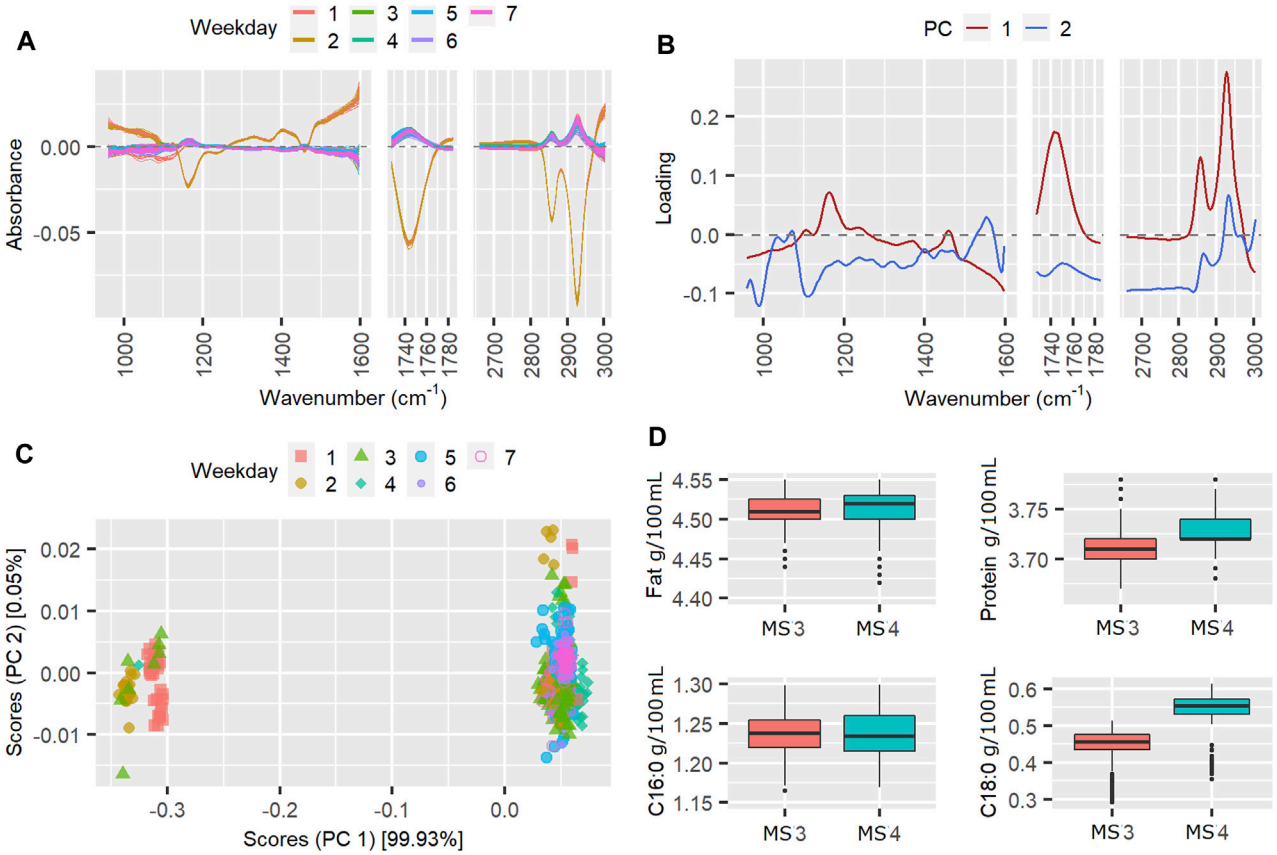

FIGURE 4 | Mean cenered spectra **(A)**, instrument loading plots **(B)** and scores **(C)** for pilot spectra from week 86, in which the TSSQ (adj) exceeded 3 standard deviations. Boxplots **(D)** show the predicted fat, protein, C16:0 and C18:0 values from each instrument and their mean values and standard deviations (SDs), marked * for 1SD, ** for 2SDs and *** for 3SDs greater than the SD of the 177 weeks TSSQ (adj) values.

instruments. These weekday differences result in the MIR-predictions of fat, C16:0 and C18:0 having SD's more than three times the SD of the 177-weeks TSSQ (adj). The protein was less affected, exceeding only one SD this week; this is also evident in the loadings which represent mainly fat and fatty acids.

The results for week 217 are shown in **Figure 5**. During week 217 three instruments were active and the instrument variation contributed 83.9% to the TSSQ (adj). The separation of instrument scores in the scores plot in **Figure 5C** shows that the spectra of instrument MS3 are consistently different from those of the MS4 and MS7 instruments. All three instruments are the same model, however, the higher negative PC1 spectral loadings appear to show that MS3 has generally higher spectral intensities than the other two (**Figure 5A**). This difference translates into higher predicted values for fat and C16:0 as shown in the boxplots (**Figure 5D**), but the MS3 C18:0 values are lower. The differences in the spectra have likely been compensated for by the weekly fat and protein calibration adjustments, given that the SDs for fat and protein are below one SD of the 177-weeks average. The MS3 spectral differences do affect the variation in measurement for true protein, however, and also affect the C16:0 and C18:0 measurements, with the SD at three times and twice the 177-weeks average, respectively.

PCA of a Series of Four Successive Weeks: 193 to 196

Of interest for routine monitoring of instrument performance are changes in contribution from instrument effect on TSSQ (adj) over successive weeks. An example of such a change is seen in the sharp increase in instrument effect (black trace) on the TSSQ (adj) in **Figure 2** between weeks 193 and 196. These weeks were selected as an example because over this period, the same four instruments were in use and the TSSQ (adj) was well below the mean. A PCA could show useful insight into the observed increase in instrument effect, while the corresponding boxplots would show how this affects the MIR-predicted component values. The PCA scores and loadings are plotted for weeks 193–196 in **Figure 6**, and the corresponding boxplots for the four MIR-predicted components are given in **Figure 7**.

The scores can be seen to drift slightly over the first 3 weeks, with the biggest change between instrument scores and loadings occurring between weeks 194 and 195, while those for weeks 195 and 196 are similar. Differences in scores between week 193 and

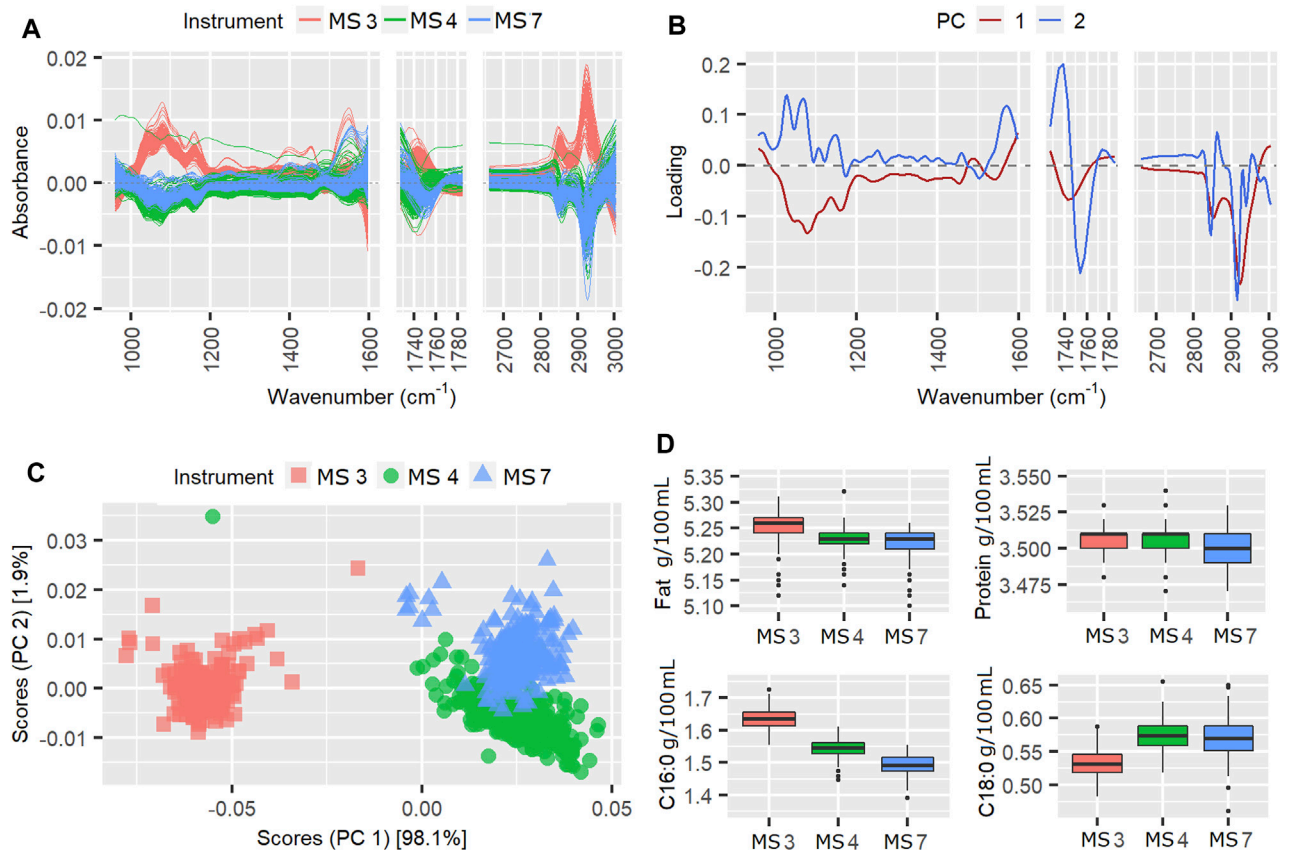

FIGURE 5 | Mean centered spectra **(A)**, instrument loading plots **(B)** and scores **(C)** for pilot spectra from week 217, in which the TSSQ (adj) exceeded 3 standard deviations. Boxplots **(D)** show the predicted fat, protein, C16:0 and C18:0 values from each instrument and their mean values and standard deviations (SDs), marked * for 1SD, ** for 2SDs and *** for 3SDs greater than the SD of the 177 weeks TSSQ (adj values.

194 are mainly along PC1 and mainly for MS3 (**Figures 6A,C**), with the main difference in the PC1 loadings between weeks 193 and 194 being in the relative intensities of the lipid C-H stretch modes at 2,856 and 2,926 cm^{-1} (**Figures 6B,D**), and the protein amide II band intensity around 1,550 cm^{-1}. This is consistent with an increase in the instrument effect from 31.5% in week 193 to 53.7% in week 194, and an increase the weekday effect from 3.1% in week 193 to 8.5% in week 194. The difference observed in the loadings affect the protein prediction by MS3 (**Figure 7B**), however, it does not result in a SD above the 177-weeks average.

Between weeks 194 and 195 there is a noticable drift in the scores for MS4 away from those of the other instruments, and further up PC2 (**Figures 6E,G**). The loadings show that these differences are due to clear changes in relative intensities between the C-H stretching region at higher wavenumbers 2800–3,000 cm^{-1} typical for fat and the C-O stretching and C-H deformations at lower wavenumbers between 1,000–1,100 cm^{-1}, representing mostly lactose (**Figures 6F,H**). These changes result in a slight increase in fat prediction for MS3 relative to the other instruments (**Figures 7K,L**), and an increase in SD from one to two SD's above the 177-weeks mean in predicted values for C16:0, indicated by single asterisks in the boxplots in **Figures 7K,L** for weeks 195 and 196. These changes also correspond with a small increase in instrument effect, from 53.6% in week 194 to 56.1% in week 195 (**Figure 2**). At the same time the weekday effects decrease from 8.5% in week 194 to 6.3% in week 195, corresponding with reduced spread of scores for MS3 along PC1 (**Figures 6C,E**). The changes in scores plots and loadings from weeks 195–196 are small, however, the contribution of instrument effects increases from 56.1 to 66.1% between weeks 195 and 196, while the weekday effects decrease from 6.3 to 2.3%. The change in relative contribution of these effects can be seen in a small decrease in spread over PC1 of the scores in **Figure 6G** for week 196.

DISCUSSION

Plotting the TSSQ (adj) for each week with time (**Figure 1**) presents an overview of the overall variance of the active instruments in the laboratory over the time period December 2016 to March 2021. This 4 year time record enables a robust measure of the SD expected over all seasons, and can be used to monitor instrument performance and/or laboratory environment stability with time. When the TSSQ (adj) is flagged as exceeding one, two or three times the SD of the

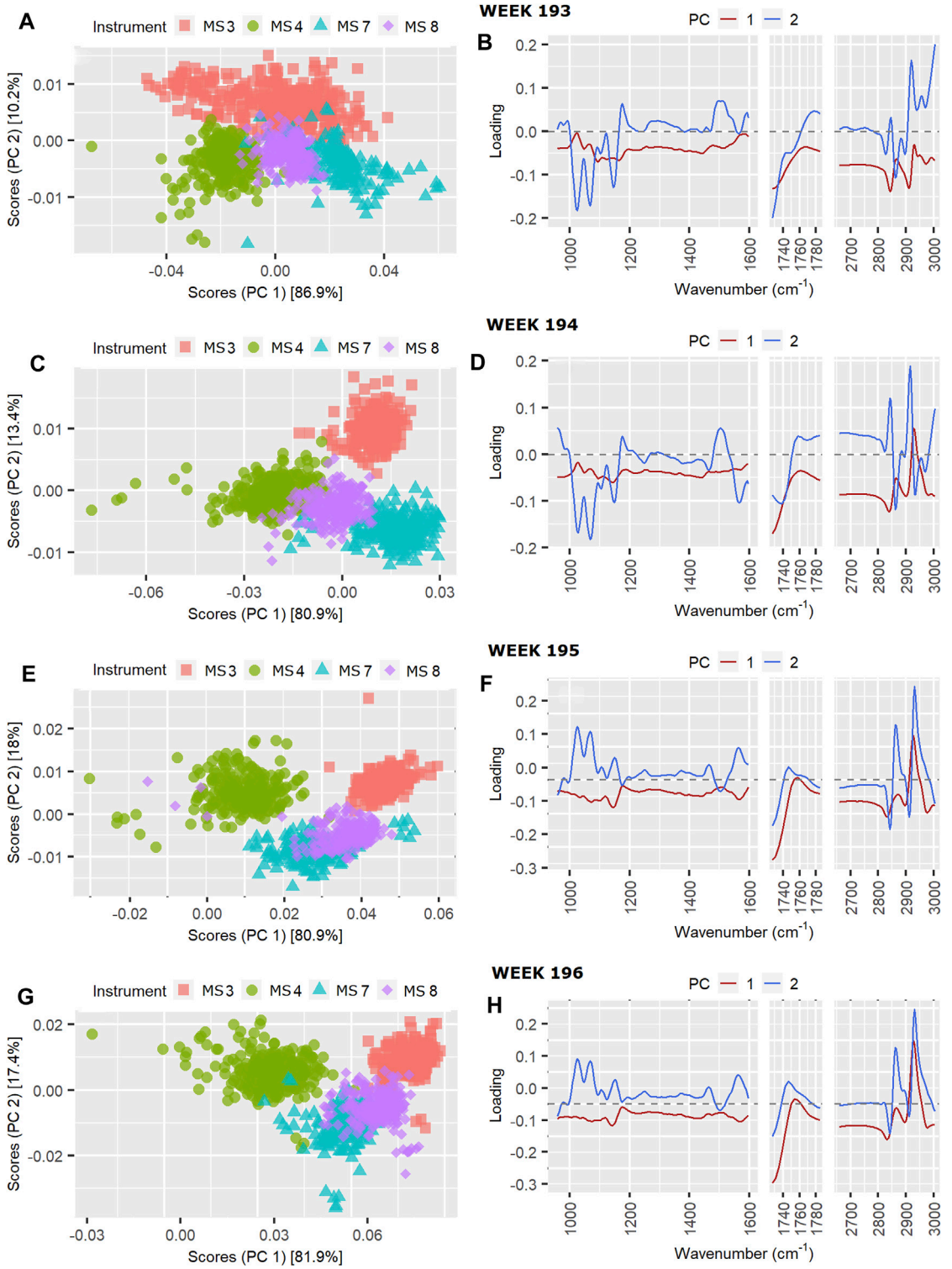

FIGURE 6 | Instrument scores and loading plots, respectively, for pilot spectra from four successive weeks: **(A)** and **(B)** week 193, **(C)** and **(D)** week 194, **(E)** and **(F)** week 195 and **(G)** and **(H)** week 196.

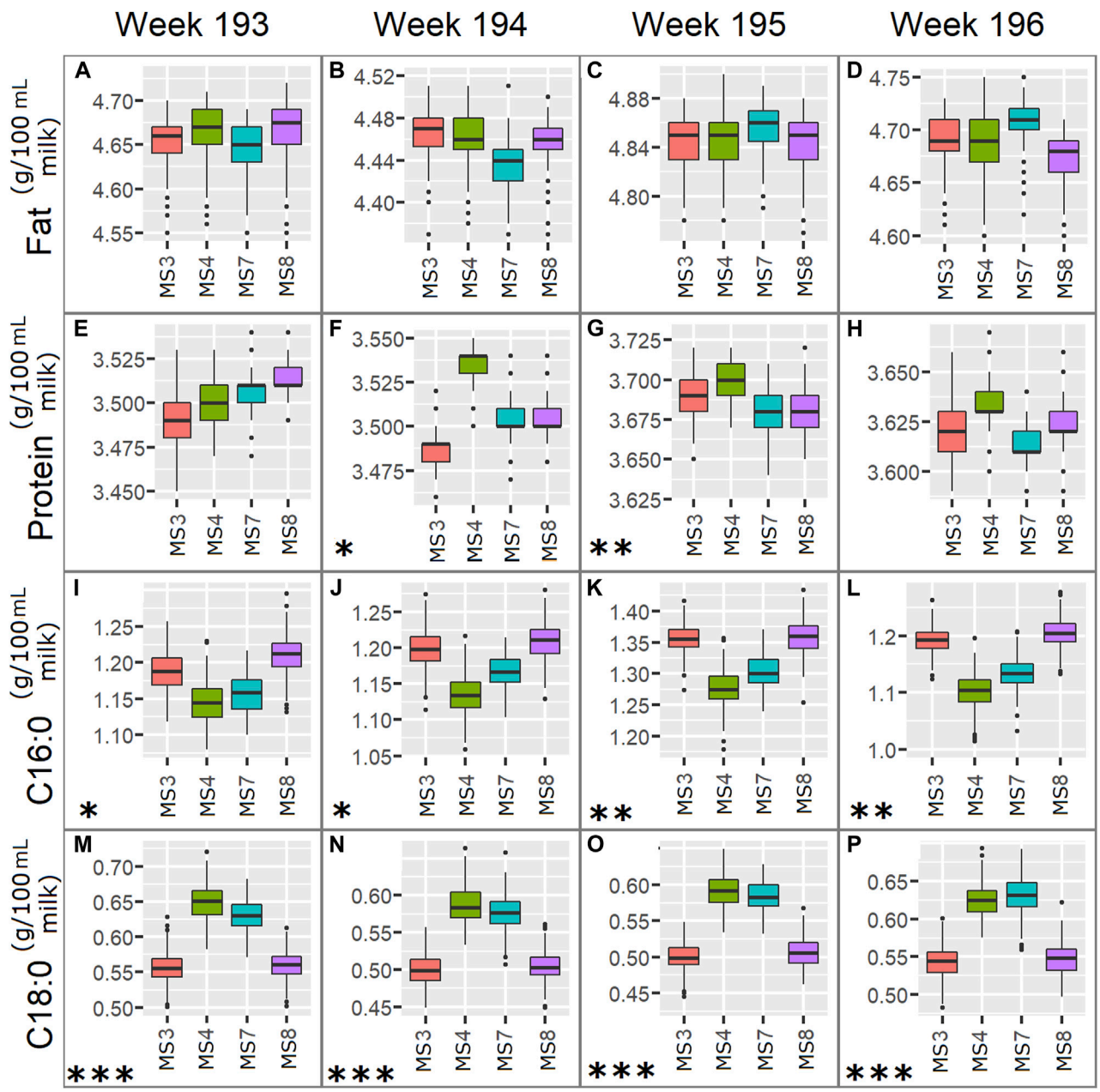

FIGURE 7 | Boxplots for fat, true protein, C16:0 and C18:0 predicted from pilot spectra using instruments MS3, MS4, MS7, and MS8, over successive weeks that had low TSSQ: week 193: **(A)** Fat, **(E)** True protein, **(I)** C16:0, **(M)** C18:0; week 194: **(B)** Fat, **(F)** True protein, **(J)** C16:0, **(N)** C18:0; week 195: **(C)** Fat, **(G)** True protein, **(K)** C16:0, **(O)** C18:0 and week 196: **(D)** Fat, **(H)** True Protein, **(L)** C16:0, **(P)** C18:0. The plots for each week show the median, IQR and variability outside the upper and lower quartiles for each component. Asterisks represent the number of SD's, marked * for 1SD, ** for 2SDs and *** for 3SDs, by which the predicted component exceeds the average of all 177-weeks average SD.

177 week period, the contribution of instrument, weekday, interaction between these or residual effects can be examined to identify the source of variation. The plot showing the contribution from these effects in **Figure 2** shows that over the 4 years, the major contributions to the TSSQ (adj) of each week are differences between instruments and residual effects.

Factors contributing to residual effects include ground vibrations, electronic gain settings, cell temperature, instrument temperatures and operator changes (Young, 1978). Controlling the lab temperature and humidity aims to minimize variations in these. Instrument effects have multiple sources of variations. The detectors used in the MIR milk instruments are DTGS (deuterium triglyceride sulphate) thermal detectors that convert thermal energy to electrical signal; they respond to temperature by changing their capacitance which is measured as a voltage change. Again, controlling the lab temperature and

humidity minimizes variations in these detectors, however, noise from the IR source due to random photons and thermodynamic noise in interaction with these photons can also affect both the intensity of, and noise in the signal (King et al., 2004) and so contribute to residual variation. In addition, pathlength changes are commonly caused by build-up of protein and foreign material on the windows of diamond sample cells, or gradual erosion of sample cell windows made of CaF_2. This results in changes in path length, which also greatly affect the IR signal/noise ratio in aqueous systems, (Jensen and Bak, 2002), and can cause a calibration shift.

The homogenizers on each individual instrument wear at different rates, depending on the number of samples through each instrument. The effects of variations in these instrument components on the spectra are minimized by routine standardization procedures, typically every 6 weeks, and by monitoring a homogenizer index which measures the efficiency as an approximate prediction of one of the fat globule distribution parameters (FOSS, private communication).

Considering these possible sources of variation, instrument differences would therefore be expected to have the greatest effect on variations in the pilot milk spectra over the week, with residual effects also contributing to a large extent. The *p*-values>0.05 confirming significance of the instrument effects throughout all 177 weeks confirms this, while the weekday variations within each instrument were significant for only two of the 177 weeks. Recent work on minor milk components such as milk urea has also shown the impact of inter instrument differences on IR predicted results (Wood et al., 2020; Portenoy et al., 2021).

Of the 25 weeks in **Table 1** with TSSQ (adj) exceeding one or more of the 177-weeks average SD, only two of the 177 weeks (1% of the time) had TSSQ (adj) values above 2SD's and in only four (2%) of the time) the TSSQ (adj) exceeded 3SD's (**Table 1**). The TSSQ (adj) in the other 19 weeks the exceeded only one SD above the 177-weeks mean (11% of the time). This is a relatively low rate which shows that the weekly calibration adjustments and regular instrument standardization procedures are effective in adjusting for instrument drift and maintaining the inter-instrument and intra-instrument variances below one SD of the 177-weeks mean, during 152 of the 177 weeks (86% of the time). Of the 25 weeks in **Table 1**, the instrument effect dominated the TSSQ (adj) in 13 weeks (52%), while the weekday effect dominated once (4% of the time) while residual effects dominated in 11 weeks (44% of the time).

The 25 week period in which the TSSQ (adj) > 1SD of the 177-weeks average was found to affect the mean and SD of the four components fat, protein, C16:0 and C18:0 to different extents. In this period, the predicted fat SD exceeded the 177-weeks average by one or more in only 10 of the 25 weeks, while for the true protein this occurred in 24 of the 25 weeks. The prediction of the less abundant fatty acids, C16:0 and C18:0, was affected in all 25 weeks, with C18:0, the least abundant consistently showing SD's three times higher than the 177-weeks average. Predictions of other fatty acids, not discussed here, were also found to show differing extents of SD's over this period, and greater than those shown by the major milk components fat, protein, lactose and total solids. We thus note the relevance of this more sensitive monitoring approach considering the recent trend towards deployment of predictive models focused on greater use of IR data of milk (Grelet et al., 2017). Recent work on minor milk components such as milk urea has also shown the impact of inter instrument differences on IR predicted results. (Wood et al., 2020; Portenoy et al., 2021).

The PCA scores and loadings obtained from the ASCA analysis of the spectra are useful for monitoring instrument drift with time. This was shown in the example of four successive weeks 193–196, during which a marked increase in instrument effect from 31 to 66% was observed, while at the same time the residual variation contribution decreased from 53 to 29%, while weekday or within-week variations showed no trend, instead signaling spread of weekday scores along PC1. The scores and loadings can be monitored to signal drifts beginning to occur in individual instruments week by week. Weekly calibration adjustments of the instruments allow adjustment of bias and slope of the fat, protein, total solids and lactose calibrations in the laboratory and thus compensate for differences in all the milk component predictions that may arise through weekly changes in instrument performance. The ASCA scores are especially sensitive to differences in spectral intensities of the different instruments on different weekdays, and to changes with time in relative intensities over the spectral region. Monitoring the ASCA scores and loading plots could provide a useful indicator of the extent of instrument drift, and signal when standardization of the instruments would be necessary rather than adjusting the calibration to compensate for these changes. We suggest such an approach could be used in conjunction with recent advances in instrument standardization that have allowed calibrations to be deployed across networks of instruments from different manufacturers (Grelet et al., 2021).

Monitoring the boxplots of predicted components could be useful for testing the effectiveness of the calibration adjustments of the major components and whether these improve the predictions of less abundant components such as individual fatty acids or indirectly-measured traits. Comparison of the boxplots with the score plots and loadings are also useful for evaluating when calibration adjustments are compensating for instrument differences to an extent that instrument signal standardization is necessary.

CONCLUSION

We have described the novel use of ASCA on the spectra of pilot test milk samples over time as a new approach for routine monitoring of instrument performance in a milk testing laboratory. Plotting of the scores and loadings derived from the ASCA effect models, the mean centered spectra and boxplots of the MIR-predicted components provides a useful overview of the weekly performance of the spectrometers in the laboratory, in terms of day-to-day variations in spectral intensities, differences arising between spectrometers and to what extent the spectral variance shows residual effects, not explained by these two effects, such as changes in laboratory environment or unexplained noise. This can be particularly useful to flag unexpected laboratory environment changes or weekly instrument changes that may affect the accuracy of the MIR-predicted milk components. Weekly monitoring of these plots can also serve as an indicator for when instrument standardization of one or more instruments is necessary, and can evaluate when the weekly calibration adjustments may be

compensating for instrument differences. Comparison of the boxplots with the score plots and loadings is also useful to signal the effectiveness of instrument standardization and the weekly calibration adjustments, especially with the trend towards greater use of IR data for predicting milk components and other relevant traits present in lower levels.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Conceptualization: MN and SH; methodology, CG, FM, and MN; software, CG, MN, and FM; investigation, MN, GS; resources, SH, GS; data curation, MN and CG; writing—original draft preparation, MN; writing—review and editing, all authors; supervision, SH, GS, and MN; project administration, SH and GS All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Fonterra On-farm Research and Development, as part of a collaborative project with the University of Auckland.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the help of Scott Middleton and Callie Smith from MTNZ, Hamilton, New Zealand for contribution of the spectral data.

REFERENCES

Biggs, D. A. (1978). Instrumental Infrared Estimation of Fat, Protein, and Lactose in Milk: Collaborative Study. *JAOAC* 61, 1015–1034. doi:10.1093/jaoac/61.5.1015

Grelet, C., Dardenne, P., Soyeurt, H., Fernandez, J. A., Vanlierde, A., StevensGengler, F. N., et al. (2021). Large-scale Phenotyping in Dairy Sector Using Milk MIR Spectra: Key Factors Affecting the Quality of Predictions. *Methods* 186, 97–111. doi:10.1016/j.ymeth.2020.07.012

Grelet, C., Fernández Pierna, J. A., Dardenne, P., VBaeten, V., and Dehareng, F. (2015a). Standardization of Milk MIR Spectra, Development of Common MIR Equations. *Biotechnol. Agron. Soc. Environ.* 19 (2), 99–100.

Grelet, C., Fernández Pierna, J. A., Dardenne, P., Baeten, V., and Dehareng, F. (2015b). Standardization of Milk Mid-infrared Spectra from a European Dairy Network. *J. Dairy Sci.* 98 (4), 2150–2160. doi:10.3168/jds.2014-8764

Grelet, C., Pierna, J. A. F., Dardenne, P., Soyeurt, H., Vanlierde, A., Colinet, F., et al. (2017). Standardization of Milk Mid-infrared Spectrometers for the Transfer and Use of Multiple Models. *J. Dairy Sci.* 100 (10), 7910–7921. doi:10.3168/jds.2017-12720

Hansen, P. W. (2020). *Digging into the Zero Setting Gold Mine in White Paper from FOSS*. Hilleroed, Denmark: FOSS.

Hansen, P. W. (2014). *FOSS Analytical, Method of and Apparatus for Correcting Intensity Deviations in a Spectrometer*. USA: Patent WO2000017611A1.

Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., Westerhuis, J. A., and Smilde, A. K. (2005). ASCA: Analysis of Multivariate Data Obtained from an Experimental Design. *J. Chemometrics* 19 (9), 469–481. doi:10.1002/cem.952

Jensen, P. S., and Bak, J. (2002). Near-infrared Transmission Spectroscopy of Aqueous Solutions: Influence of Optical Pathlength on Signal-To-Noise Ratio. *Appl. Spectrosc.* 56 (12), 1600–1606. doi:10.1366/000370202321115878

Kaylegian, K. E., Houghton, G. E., Lynch, J. M., Fleming, J. R., and Barbano, D. M. (2006). Calibration of Infrared Milk Analyzers: Modified Milk versus Producer Milk. *J. Dairy Sci.* 89, 2817–2832. doi:10.3168/jds.s0022-0302(06)72555-3

King, P. L., Ramsey, M. S., McMillan, P. F., and Swayze, G. (2004). "Laboratory Fourier Transform Infrared Spectroscopy Methods for Geologic Samples," in *Molecules to Plants: Infrared Spectoscopy in Geochemistry, Exploration Geochemistry and Remote Sensing* (Canada: MIneral Association of Canada).

Kucheryavskiy, S. (2020). Mdatools - R Package for Chemometrics. *Chemometrics Intell. Lab. Syst.* 198, 103937. doi:10.1016/j.chemolab.2020.103937

Kucheryavskiy, S. (2021). Mdatools: Multivariate Data Analysis for Chemometrics. Available at: https://mdatools.com/docs/pca-distances-and-limits.html.

Pelletier, M. J. (2003). Quantitative Analysis Using Raman Spectrometry. *Appl. Spectrosc.* 57 (1), 20A–42A. doi:10.1366/000370203321165133

Pomerantsev, A. L. (2008). Acceptance Areas for Multivariate Classification Derived by Projection Methods. *J. Chemometrics* 22 (11-12), 601–609. doi:10.1002/cem.1147

Portenoy, M., Coon, C., and Barbano, D. M. (2021). Infrared Milk Analyzers: Milk Urea Nitrogen Calibration. *J. Dairy Sci.* 104, 7426–7437.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R.-J. A. N., van der Greef, J., and Timmerman, M. E. (2005). ANOVA-simultaneous Component Analysis (ASCA): a New Tool for Analyzing Designed Metabolomics Data. *Bioinformatics* 21 (13), 3043–3048. doi:10.1093/bioinformatics/bti476

Smith, E. B., Barbano, D. M., Lynch, J. M., and Fleming, J. R. (1993). Performance of Homogenizers in Infrared Milk Analyzers: A Survey. *J. AOAC Int.* 76 (5), 1033–1041. doi:10.1093/jaoac/76.5.1033

Wang, Y., Veltkamp, D. J., and Kowalski, B. R. (1991). Multivariate Instrument Standardization. *Anal. Chem.* 63 (23), 2750–2756. doi:10.1021/ac00023a016

Wood, E. M., Portenoy, M., Barbano, D. M., and Reed, K. F. (2020). Precision and Accuracy of Mid-infrared Spectroscopy for Milk Urea Nitrogen Analysis. *J. Dairy Sci.* 103, 4.

Young, R. S. (1978). Calibration and Standardization of the Infrared Milk Analyzer. The California Experience. *J. Dairy Sci.* 61 (9), 1279–1283. doi:10.3168/jds.s0022-0302(78)83718-7

Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., and Smilde, A. K. (2011). ANOVA-principal Component Analysis and ANOVA-Simultaneous Component Analysis: a Comparison. *J. Chemometrics* 25 (10), 561–567. doi:10.1002/cem.1400

Author Disclaimer: The contents of this manuscript are the authors' opinions and should not be considered as opinions or policy of Fonterra. The mention of trade names and manufacturers is for technical accuracy and should not be considered as endorsement of a specific product or manufacturer.

Conflict of Interest: We note that authors MN, GS, and SH are employees of Fonterra Co-operative Group Ltd., which supported this work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Analytical Science

ORIGINAL RESEARCH
published: 18 October 2021
doi: 10.3389/frans.2021.754447

Different Methods for Determining the Dimensionality of Multivariate Models

Douglas N. Rutledge[1,2,3]*, *Jean-Michel Roger*[1,4] and *Matthieu Lesnoff*[1,5,6]

[1]ChemHouse Research Group, Montpellier, France, [2]INRAe, AgroParisTech, UMR SayFood, Université Paris-Saclay, Paris, France, [3]National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, NSW, Australia, [4]UMR ITAP, INRAe, Montpellier Institut Agro, Univ Montpellier, Montpellier, France, [5]SELMET, CIRAD, INRAe, Institut Agro, Univ Montpellier, Montpellier, France, [6]CIRAD, UMR SELMET, Montpellier, France

A tricky aspect in the use of all multivariate analysis methods is the choice of the number of Latent Variables to use in the model, whether in the case of exploratory methods such as Principal Components Analysis (PCA) or predictive methods such as Principal Components Regression (PCR), Partial Least Squares regression (PLS). For exploratory methods, we want to know which Latent Variables deserve to be selected for interpretation and which contain only noise. For predictive methods, we want to ensure that we include all the variability of interest for the prediction, without introducing variability that would lead to a reduction in the quality of the predictions for samples other than those used to create the multivariate model.

Keywords: multivariate models, dimensionality, latent variables, regression, cross validation (min5-max 8)

OPEN ACCESS

Edited by:
Hoang Vu Dang,
Hanoi University of Pharmacy, Vietnam

Reviewed by:
Jahan B Ghasemi,
University of Tehran, Iran
Ludovic Duponchel,
Université de Lille, France

***Correspondence:**
Douglas N. Rutledge
rutledge@agroparistech.fr

Specialty section:
This article was submitted to Chemometrics, a section of the journal Frontiers in Analytical Science

Received: *10 August 2021*
Accepted: *06 October 2021*
Published: *18 October 2021*

Citation:
Rutledge DN, Roger J-M and Lesnoff M (2021) Different Methods for Determining the Dimensionality of Multivariate Models. Front. Anal. Sci. 1:754447. doi: 10.3389/frans.2021.754447

In the case of predictive methods such as PLS, the most common procedure to determine the number of Latent Variables for use in the model is Cross Validation which is based on the difference between the vector of observed values, $\mathbf{y}$, and the vector of predicted values, $\hat{\mathbf{y}}$.

In this article, we will first present this procedure and its extensions, and then other methods based on entirely different principles. Many of these methods may also apply to exploratory methods.

These alternatives to Cross Validation include methods based on the characteristics of the regression coefficients vectors, such as the Durbin-Watson Criterion, the Morphological Factor, the Variance or Norm and the repeatability of the vectors calculated on random subsets of the individuals. Another group of methods is based on characterizing the structure of the $\mathbf{X}$ matrices after each successive deflation.

The user is often baffled by the multitude of indicators that are available, since no single criterion (even the classical Cross-Validation) works perfectly in all cases. We propose an empirical method to facilitate the final choice of the number of Latent Variables. A set of indicators is chosen and their evolution as a function of the number of Latent Variables extracted is synthesized by a Principal Components Analysis. The set of criteria chosen here is not exhaustive, and the efficacy of the method could be improved by including others.

INTRODUCTION

A tricky aspect in the use of all multivariate analysis methods is the determination of the number of Latent Variables, both for exploratory methods such as Principal Components Analysis (PCA) and Independent Components Analysis (ICA), and predictive methods such as Principal Components Regression (PCR), Partial Least Squares regression (PLS) or PLS Discriminant Analysis (PLS-DA). For exploratory methods, we want to know which Latent Variables deserve to be selected for

interpretation and which contain only noise. For predictive methods, we want to ensure that we include all the variability of interest for the prediction, without introducing variability that would lead to a reduction in the quality of the predictions for samples other than those used to create the multivariate model.

Whatever the type of method (exploratory or predictive), the most common procedure consists in examining the evolution of a criterion, as a function of the number of Latent Variables calculated. In the case of predictive methods such as PLS, the most common criterion is the Cross Validation error, which is based on the difference between the vector of observed values, **y**, and the vector of predicted values, $\hat{\mathbf{y}}$. But many other criteria can be used. In this article, we will first present the cross-validation procedure and its extensions, and then other methods based on entirely different principles. The objective of this article is not to make an exhaustive review of these criteria, but to present some of those of most interest for chemometrics.

Principal Components Analysis is based on the mathematical transformation of the original variables in the matrix **X** into a smaller number of uncorrelated variables, T.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{R} \tag{1}$$

where the matrices **T** and **P** represent, respectively, the vectors of factorial coordinates ("scores") and factorial contributions ("loadings") derived from **X**.

This method is interesting because, by construction, the PCs are uncorrelated and it is not possible to have more PCs than the rank of **X**, i.e., min ($N_{individuals}$, $N_{variables}$) if the data are not centered and min ($N_{individuals}$-1, $N_{variables}$) otherwise. In addition, since the first PCs correspond to the directions of greatest dispersion of the individuals, it is possible to retain only a small number of PCs, **T***, in the calculation of the coefficients of a PCR regression model.

$$\mathbf{B} = \left(\mathbf{T}^{*\mathrm{T}}\mathbf{T}^{*}\right)^{-1}\mathbf{T}^{*\mathrm{T}}\mathbf{Y} \tag{2}$$

The values of new objects are then be predicted by the classical equation:

$$\widehat{\mathrm{Y}} = \mathbf{TB} = \mathbf{XPB} \tag{3}$$

PLS regression (Partial Least Squares regression) also allows to link a set of dependent variables, **Y**, to a set of independent variables, **X**, when the number of variables (independent and dependent) is high.

The independent variables, **X**, and dependent variables, **Y**, are decomposed as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{4}$$

$$\mathbf{y} = \mathbf{U}\mathbf{R}^{\mathrm{T}} + \mathbf{F} \tag{5}$$

where the **P** and **R** represent the vectors of the factorial contributions ("loadings") and **T** and **U** are the factorial coordinates ("scores") of **X** and **Y**, respectively.

PLS is based on two principles:

1) the **X** factor coordinates, **T**, are good predictors of **Y**;
2) there is a linear relationship between the scores **T** and **U.**

In the case of PLS, the model's regression coefficient matrix is given by:

$$\mathbf{B} = \mathbf{X}^{\mathrm{T}}\mathbf{U}^{*}\left(\mathbf{T}^{*\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{U}^{*}\right)^{-1}\mathbf{T}^{*\mathrm{T}}\mathbf{Y} \tag{6}$$

In the case of PCR and PLS, successive scores and loadings are calculated after removing the contribution of each vector of scores from the **X** matrix, a process called deflation.

To present the different methods of determining the number of Latent Variables to use in the regression models, we use a dataset consisting of the near-infrared (NIR) spectra of 106 different olive oils (**Supplementary Figure S1A**) and the variable to be predicted is the concentration of oleic acid (**Supplementary Figure S1B**) determined by the classical method (gas chromatography) (Galtier et al., 2007).

It should be stressed that this article is not an exhaustive review of the possible methods that can be used to determine the dimensionality of multivariate models, as was for example the article by Meloun et al. (2000). Here, a limited number of criteria have been chosen, but based on very different criteria that characterize the multivariate models. Since these criteria may not always indicate the same dimensionality, rather than just examining them all and deciding on a value somewhat subjectively, we propose here the idea of applying a Principal Components Analysis (PCA) to the various criteria so as to have a consensus value.

DIMENSIONALITY

The problem of optimizing model dimensionality comes down to introducing as many as possible of the Latent Variables containing variability of interest, and none that contain "detrimental variability", which is often due to contributions from outliers or just different types of noise (gaussian, spike, . . .).

Already a PCA on the spectra shows that the loadings of the later components are noisier than those of the earlier ones (**Supplementary Figure S2**). It is clear that when including more than a certain number of Latent Variables into a PLS regression model there is a risk of including more noise than information.

When establishing a prediction model based on Latent Variables extracted from a multivariate data table, we must ensure that we have extracted neither too many nor too few.

Determining the number of Latent Variables can be done using a number of criteria that could be classified into two categories: prediction error or model characteristics.

CRITERIA BASED ON PREDICTION ERROR

The methods most often used are based on the quality of the predictions for individuals which were not used to create the model - either an independent dataset (test-set validation) or for individuals temporarily removed from the dataset (cross validation).

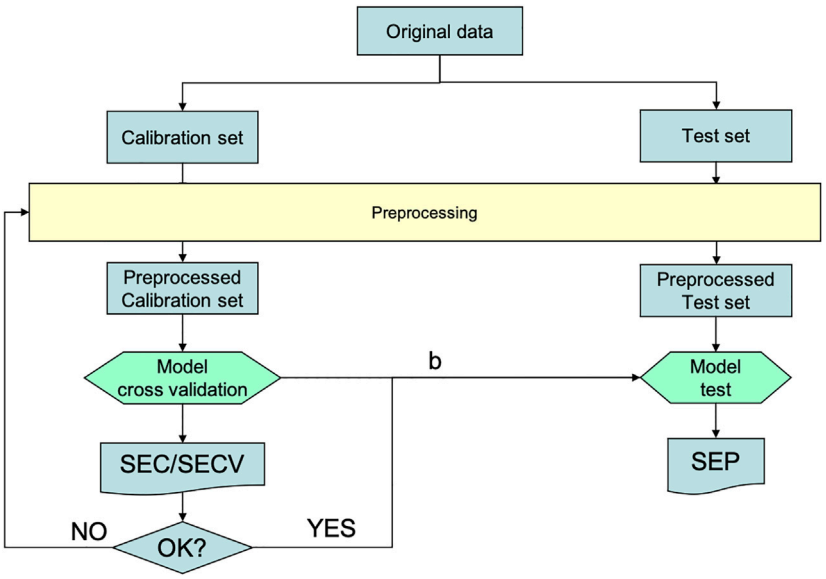

FIGURE 1 | –The process of calibration (creating and adjusting the model) by cross validation, followed by its validation with a separate test set. "**b**" is the model calculated on the calibration set.

The term "validation" as it is used in "cross validation" is incorrect, because the objective here is not to validate the model, but to adjust its parameters optimally. In **Figure 1**, the "Calibration" branch contains the "Cross Validation" step that does this model tuning, while the "Test" branch is for the true validation of the final model.

The model is adjusted by creating models with an increasing number of Latent Variables extracted from one set of individuals and observing the evolution of the differences between observed and predicted values for another set of individuals. This evolution can be followed by plotting the sum of squared residuals (RESS Residual Error Sum of Squares) or the square root of the mean sum of squares (RMSE). When this tuning is done with another single set of individuals (test-set validation), we have the SEV and RMSEV; when it is done by removing, with replacement, a few individuals from the data set (cross validation), we have the SECV and the RMSECV.

$$\mathrm{RESS} = \sum_{1}^{n} \left(\hat{y}_i - y_i \right)^2 \tag{7a}$$

$$\mathrm{RMSEVorRMSECV} = \sqrt{\frac{\sum_{1}^{n} \left(\hat{y}_i - y_i \right)^2}{\mathrm{n}}} \tag{7b}$$

Calculating the model and applying it on the entire dataset provides an estimation of $\mathbf{Y}$ ($\hat{\mathbf{Y}}$), which is used to calculate the RMSEC:

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{1}^{n} \left(\hat{y}_i - y_i \right)^2}{\mathrm{n} - (\mathrm{nLVs} + 1)}} \tag{7c}$$

The RMSEC is intended to estimate the standard deviation of the fitting error, σ. The division by [n-(k+1)] instead of n (the number of individuals) is intended to take into account the fact that the number of degrees of freedom for the estimate of σ is decreased by the inclusion of *k* Latent Variables plus the intercept. The use of this correction is valid in PCR regression, but subject to much criticism in the case of PLS where the **Y** matrix influences the calculation of the Latent Variables (Krämer and Sugiyama, 2011; Lesnoff et al., 2021). It is nevertheless sometimes used as a "naïve estimate of the RMSEC".

The principle of cross-validation is presented in **Figure 2**. Blocks of individuals are removed from the dataset and are used as a test set while the remaining individuals form the calibration dataset to create models which are used to predict the values ($\hat{\mathbf{Y}}$) for the test set individuals. The differences between the observed values ($\mathbf{Y}$) and predicted values ($\hat{\mathbf{Y}}$) are calculated for the different models. The test set individuals are then put back in the calibration dataset and another block of individuals is moved to be the test set. This process is repeated until all individuals have been used in the test set. If the size of the blocks is small (large number of blocks), the number of individuals tested each time is low and the number used to create the models is high. The limiting case is called Leave-One-Out Cross Validation (LOO-CV), where the number of blocks is equal to the total number of individuals. In this case, the result tends to be optimistic (small RMSECV) but simulates well the final model, because each prediction is made using a model calculated with a collection of samples close to that in the final model.

On the other hand, using large blocks allows us to better assess the predictive power of the model. In all cases, in order not to distort the results, it is necessary to ensure that repetitions of samples (e.g., triplicates) are kept together in the same block.

A fundamental hypothesis of theories on machine learning from empirical data assumes that the training and future datasets are generated from the same probability distribution (e.g., Faber,

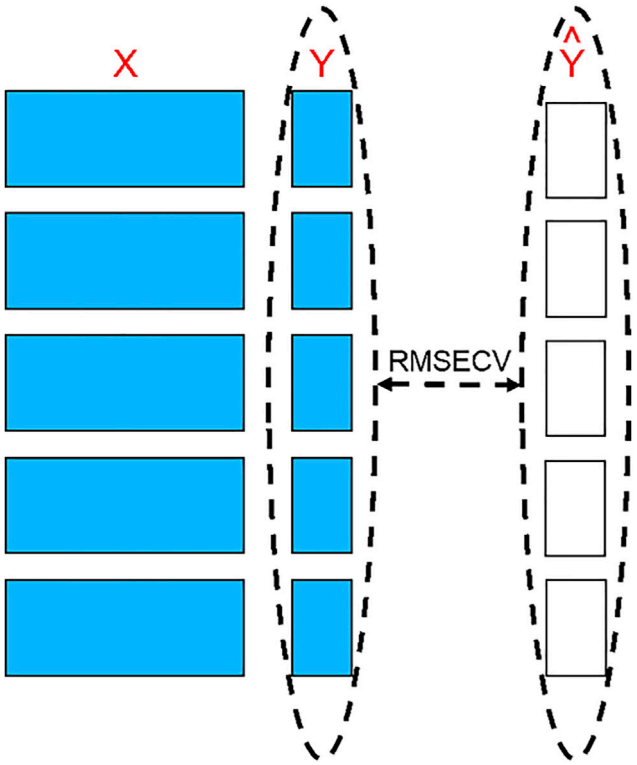

FIGURE 2 | The principle of cross-validation. Blocks of individuals are removed from the dataset to be used as tests to measure the differences between their observed values and values predicted by the models created using the remaining individuals.

1999; Denham, 2000; Vapnik, 2006; Lesnoff et al., 2021). Under this hypothesis, it is known that leave-one-out cross-validation has low bias but can have high variance for the prediction errors (i.e., variable prediction if the training set would be replicated) (Hastie et al., 2009). On the other hand, when K is smaller, cross-validation has lower variance but higher bias. Overall, five-or tenfold cross-validations are recommended as a good compromise between bias and variance (Hastie et al., p. 284).

There are many ways to build blocks, the choice being based on the organization of individuals in the matrix.

Consecutive Blocks: (1, 2, . . ., 10) (11, 12, . . ., 20) (21, 22, . . ., 30).

Venitian Blind: (1, 4, 7, . . ., 28) (2, 5, 8, . . ., 29) (3, 6, 9, . . ., 30).

Random Blocks

Predefined Blocks: for example, to manage measurement repetitions.

Figure 3 presents the evolution of the RMSECV (red circles) and the "naïve" RMSEC (blue squares) based on the number of Latent Variables used to create the prediction model. The "naïve" RMSEC, which quantifies the residual errors for the samples used to create the models, tends to zero. On the other hand, the RMSECV often has a minimum, more or less marked depending on the amount of noise in the data, which corresponds to the balance between information and noise, indicating the optimal number of Latent Variables.

Although the minimum in the RMSECV curve is for 6 LVs, this value is not much lower than that for 3 LVs. Parsimony could imply retaining only 3 LVs. To visualize more clearly the point corresponding to the minimum of RMSECV, one can use a rule that says that, on the one hand, the prediction error (here estimated by RMSECV) should be close to the fitting error (here estimated by RMSEC) and on the other hand, the RMSEC curve may present a break. A way of implementing that rule is to plot the RMSECV against the RMSEC (Bissett, 2015).

In **Figure 3** and many subsequent figures, a vertical line indicates the number of LVs resulting from a consensus found by the procedure we propose, i.e., by applying a PCA to the various very different criteria presented here.

To get a better indication of variability in the estimation of the optimal number of Latent Variables, repeated cross-validation is often used. In this case, several cross-validations are made with few blocks (here 2 blocks) containing randomly selected individuals each time. It is thus possible to calculate an average RMSECV and its variability (**Figure 4**).

Another related procedure is to plot the proportions of variability extracted from the **Y** vectors, R^2, for the calibration samples, and Q^2, for the samples removed during the cross validation, as a function of the number of Latent Variables. In **Figure 5** one can see that the difference between R^2 and Q^2 is close to zero for from 4 to 6 LVs.

Other criteria can be calculated based on the values predicted by cross-validation.

Wold's R criterion (Wold, 1978; Li et al., 2002) is given by:

$$\mathrm{PRESS}(\mathrm{k}) = \sum_{1}^{n} \left(\hat{y}_i - y_i\right)^2 \tag{8a}$$

$$\mathrm{Wold's\ R} = \frac{\mathrm{PRESS}(2\colon \mathrm{k})}{\mathrm{PRESS}(1\colon \mathrm{k}-1)} \tag{8b}$$

where PRESS(k) is the predicted residual sum of squares for k LVs; and Wold's R is a vector of the ratios of successive PRESS values. The usual cutoff for Wold's R criterion is when R is greater than unity. In **Figure 6** it can be seen that the maximum R is at 6 LVs but the value is already greater than 1 for 3 LVs.

More recently, Osten proposed the criterion (Osten, 1988), given by:

$$\mathrm{Osten's\ F}(\mathrm{k}) = \frac{\mathrm{PRESS}(1\colon \mathrm{k}-1) - \mathrm{PRESS}(2\colon \mathrm{k})}{\mathrm{PRESS}(2\colon \mathrm{k})/(\mathrm{N} - (\mathrm{k}+1))} \tag{9}$$

Figure 6 also shows that Osten's F confirms the results for Wold's R: F is less than 0 at 3 LVs but reaches a minimum at 6 LVs.

When doing a PCA, Cattell's Residual Percent Variance (RPV) criterion (Cattell, 1966) assumes that the residual variance should level off, as in **Figure 6**, after a suitable number of factors have been extracted. RPV for the model with k LVs is given by:

$$\mathrm{RPV}(k) = \frac{\sum_{i=k+1}^{K} \lambda_i}{\sum_{i=1}^{K} \lambda_i} \tag{10}$$

where λ_i is the eigenvalue for the ith PC. Here, in the case of PLS, we have replaced the eigenvalues by the variances of the scores for each LV.

There are other methods, such as Mallow's Cp (Mallows, 1973) and Akaike's Information Criterion (AIC) (Akaike, 1969), that are

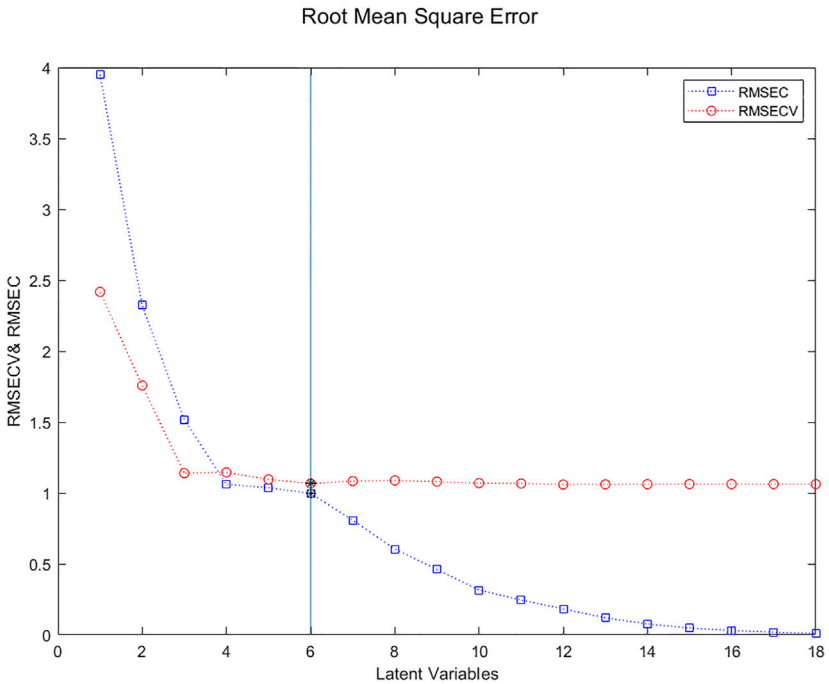

FIGURE 3 | Evolution of the RMSECV (red circles) and the naïve RMSEC (blue squares) based on the number of Latent Variables used to create the prediction model. The minimum for 6 Latent Variables is clearly visible.

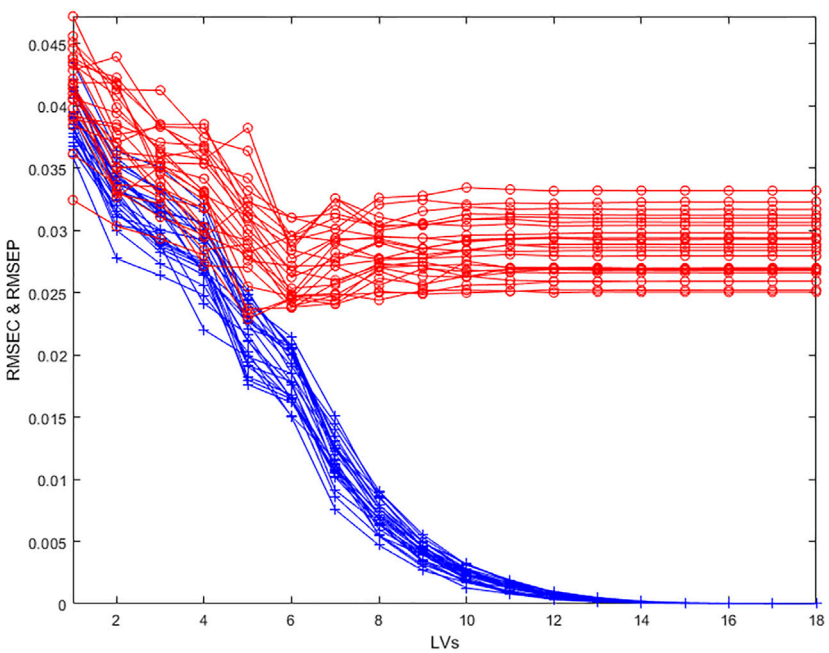

FIGURE 4 | Evolution of the RMSECV (red circles) and naïve RMSEC (blue crosses) as a function of the number of Latent Variables in the model for 25 repetitions of a 2 random blocks cross validation.

commonly used to select the dimensionality of regression models, as an alternative to cross-validation (CV). However, the calculation of Cp and AIC requires the determination of the effective number of degrees of freedom of the model, which as mentioned above, is not straightforward in the case of PLS (Lesnoff et al., 2021). For that reason, these criteria will not be considered here.

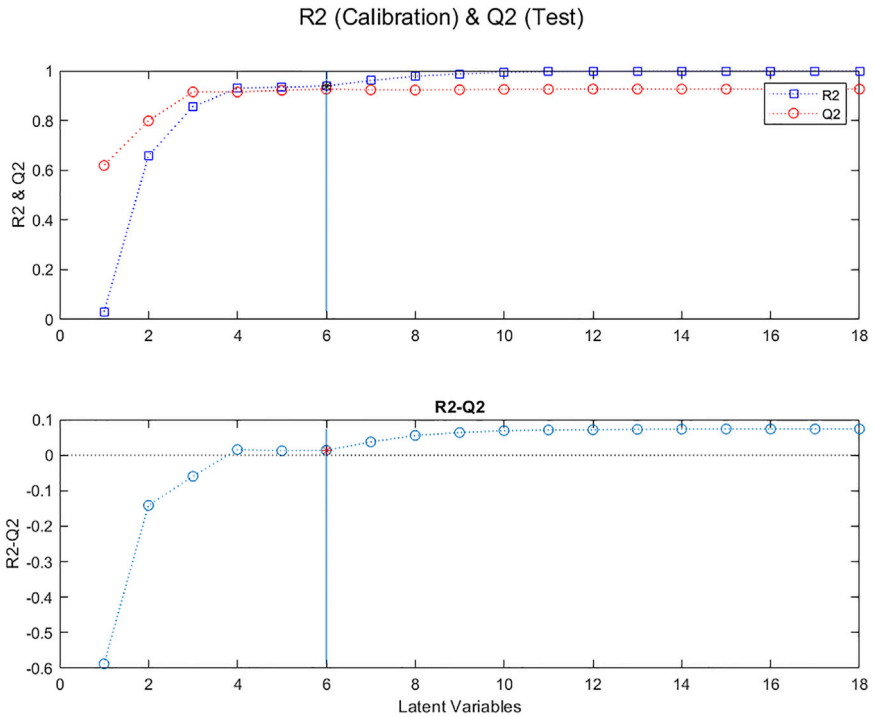

FIGURE 5 | Evolution of R^2 (blue squares) and Q^2 (red circles), for "calibration" samples and "test" samples, respectively, as a function of the number of Latent Variables in the model; Evolution of the difference between R^2 and Q^2.

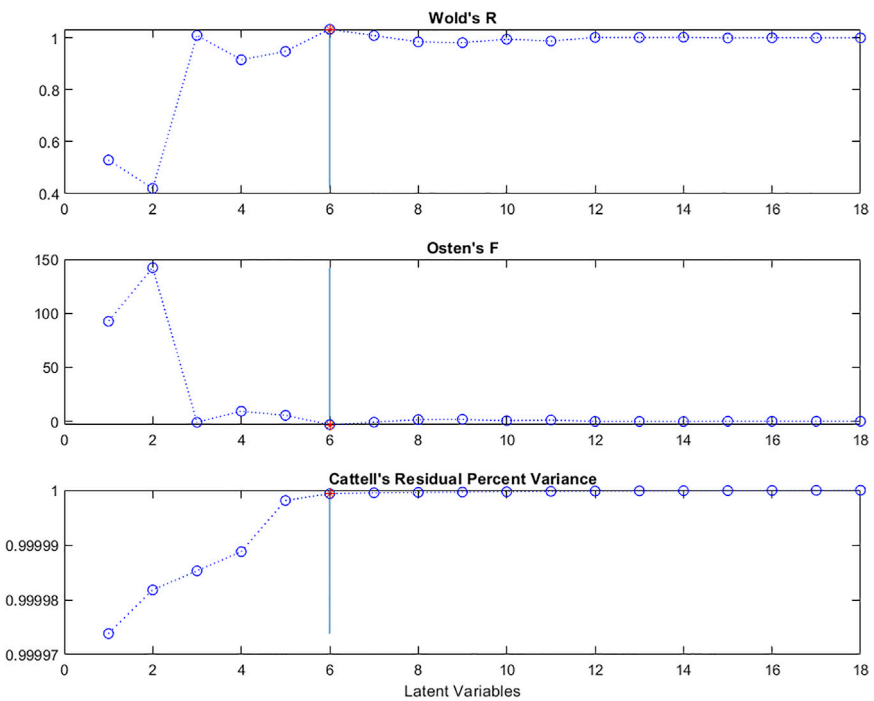

FIGURE 6 | Evolution of Wold's R; Osten's criterion and Cattell's Residual Percent Variance (RPV) criterion, as a function of the number of Latent Variables in the model.

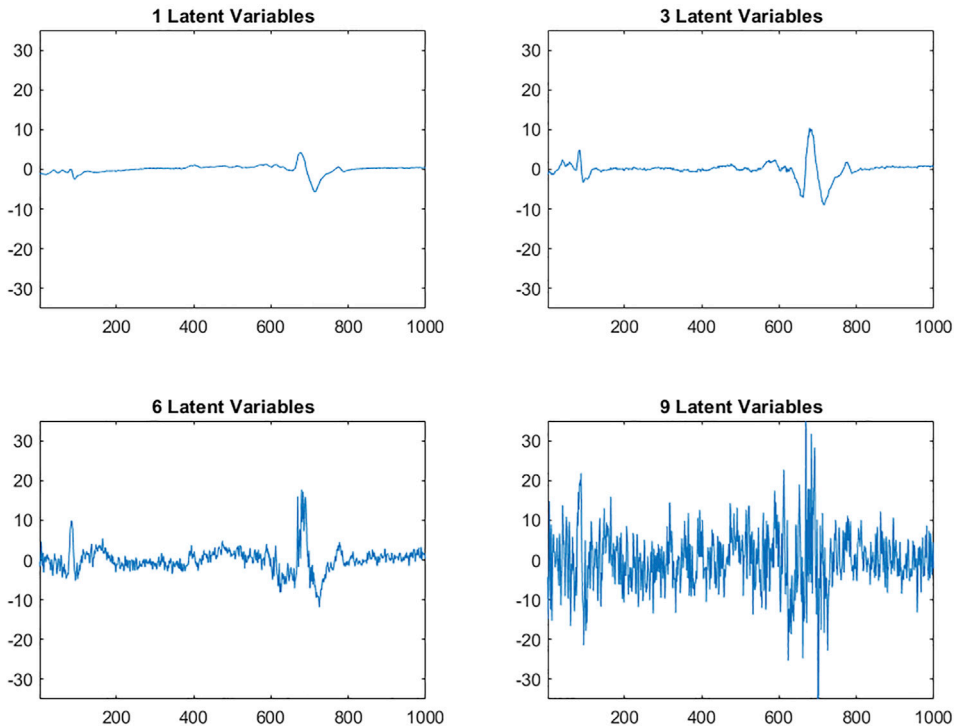

FIGURE 7 | PLS regression coefficient vectors, **b**, based on 1, 3, 6 and 9 Latent Variables plotted with a constant ordinate scale (abscissa: variable numbers).

CRITERIA BASED ON OTHER PROPERTIES OF THE MODELS

Cross-validation is sometimes difficult to perform, for example when there are many individuals and/or variables, so the calculation time can be excessive. And even when the calculation is feasible, one does not always observe a clear minimum in the RMSECV curve (as in **Figure 3**) or maximum in the Q^2 curve (as in **Figure 5**), which makes it difficult to choose the number of LVs.

As well, as indicated by Wiklund et al. (2007) CV handles "the available data economically, but like any data-based statistical test gives an interval of results and hence sometimes gives either an under-fit or an over-fit, that is they reach the minimum RMSEV for a lower or higher model rank than would be achieved using an infinitely large independent validation set". They also stressed the fact that "One area where CV works poorly both for PLS and PCR is design of experiments, where exclusion of data has large consequences for modeling". To solve these problems, they proposed carrying out permutation tests on the **Y** vector and then comparing the correlations between the scores of each latent variable and the true **Y** vector with the correlations between the scores obtained for the permuted **Y**s and the corresponding true **Y**s.

It should be noted that all these criteria are based on comparing the observed and predicted **Y** vectors. It could therefore be helpful to use other criteria based on entirely different characteristics of the models to facilitate the choice of the number of latent variables.

We will now see a set of such complementary methods, based on the characteristics of the regression coefficients vectors, **b**, and on the characteristics of the **X** matrix after each deflation.

Characteristics of the Regression Coefficients Vectors, b

As the number of Latent Variables used to calculate the regression coefficients vector, **b**, increases, more and more noise is included. When the **X** matrix contains *structured signals*, such as the near infrared spectra in **Supplementary Figure S1**, **b** coefficients are initially structured and gradually become random, as can be seen in **Figure 7**.

In the case of **b**-vectors calculated from structured signals in the rows of the **X** matrix, a "signal-to-noise ratio" can be calculated using the Durbin-Watson (DW) criterion (Durbin and Watson, 1971; Rutledge and Barros, 2002). This criterion is given by:

$$\mathbf{DW} = \frac{\sum_{i=2}^{n} (\boldsymbol{b}_i - \boldsymbol{b}_{i-1})^2}{\sum_{i=1}^{n} \boldsymbol{b}_i^2} \tag{11}$$

where b_i and $b_{(i-1)}$ are the values for successive points in a series of **b**-coefficients values. DW is close to zero if there is a strong

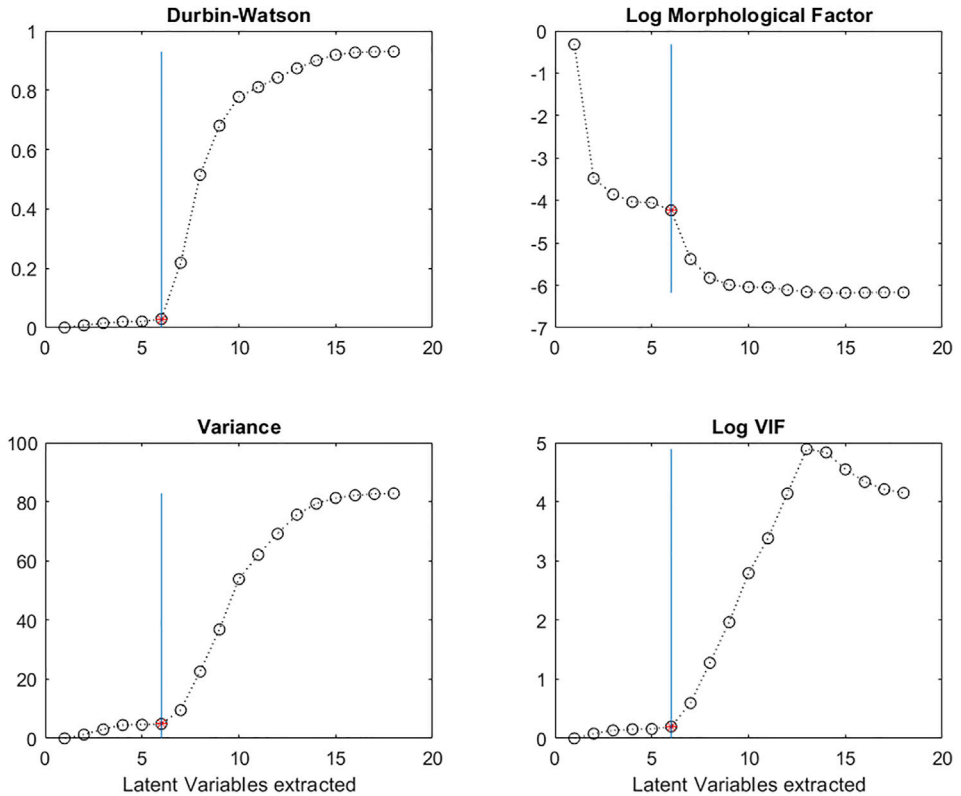

FIGURE 8 | Evolution of the Durbin-Watson (DW) criterion; the log of the Morphological Factor; the Variance; the Variance Inflation Factor (VIF) calculated on the regression vectors, **b**, as a function of the number of Latent Variables in the PLS models.

correlation between successive values. On the other hand, if there is a low correlation (i.e., a random distribution), the value of DW tends to 2.0. DW can therefore be used to characterize the degree of correlation between successive points, and thus give an objective measure of the non-random behavior of the **b** coefficients vectors. However, if the noise in the data has been reduced by smoothing, the transition will not be as clear and DW will not increase as much.

Figure 8 shows the evolution of DW calculated for a succession of regression coefficients vectors, as a function of the increasing number of LVs used in the PLS model. It is clear that there is a very sudden increase in DW after 6 LVs.

The Morphological Factor (MF) (Wang et al., 1996) is based on the same phenomenon as the DW criterion, noisy vectors are less structured than non-noisy vectors. On the other hand, the mathematical principle is different:

$$\mathrm{MF}(\mathbf{b}) = \|\mathbf{b}\| / (\|\mathbf{MO}(\mathbf{b})\|.\mathrm{ZCP}(\mathbf{MO}(\mathbf{b}))) \quad (12a)$$

$$\mathbf{MO}(\mathbf{b}) = \mathbf{b_{i+1}} - \mathbf{b_i} \,(\text{for i} = 1, 2, \ldots \text{n} - 1) \quad (12b)$$

where **b** is a vector of regression coefficients; **MO(b)** the vector of differences in intensity between successive points in b; ZCP (**MO**(**b**)) the number of times **MO(b)** changes signs, and the operator ||o|| is the Euclidian norm.

In the case of a noisy vector, **MO(b)** will contain bigger values and there will be more sign changes than in the case of a smooth vector, resulting in lower MF values. **Figure 8** shows the evolution of MF as a function of the number of Latent Variables extracted. The log of MF evolves in a similar way to the DW criterion with a decrease after 6 Latent Variables.

In the case of an **X** matrix that does not contain structured signals (e.g., physical-chemical data or mass spectra) DW or MF should not be used. But other characteristics of the regression vectors can be used instead.

It can be seen that the range of **b** vector values initially remains relatively stable, but beyond a certain number of LVs, the b-coefficient values increase enormously (**Figure 7**). By plotting the variance of the regression vectors it is possible to see the point at which this phenomenon appears (**Figure 8**) for both structured and non-structured data matrices. This is also true for the standard deviation or the norm of the vectors.

The Variance Inflation Factor of a variable *i* in a matrix **X** (VIFi) (Marquardt, 1970; Ferré, 2009) is equal to the inverse of (1-Ri^2), where Ri^2 is the coefficient of determination of the regression between all the other predictor variables in the matrix and the variable *i*. VIFi quantifies the degree to which that variable can be predicted by all the others. The closer the Ri^2 value to 1, the higher the multicollinearity with independent variable *i* and the higher the value of VIFi.

As the number of LVs included in a regression model increases, the structure of the b-coefficients vectors changes due to the inclusion of more sources of variability, initially

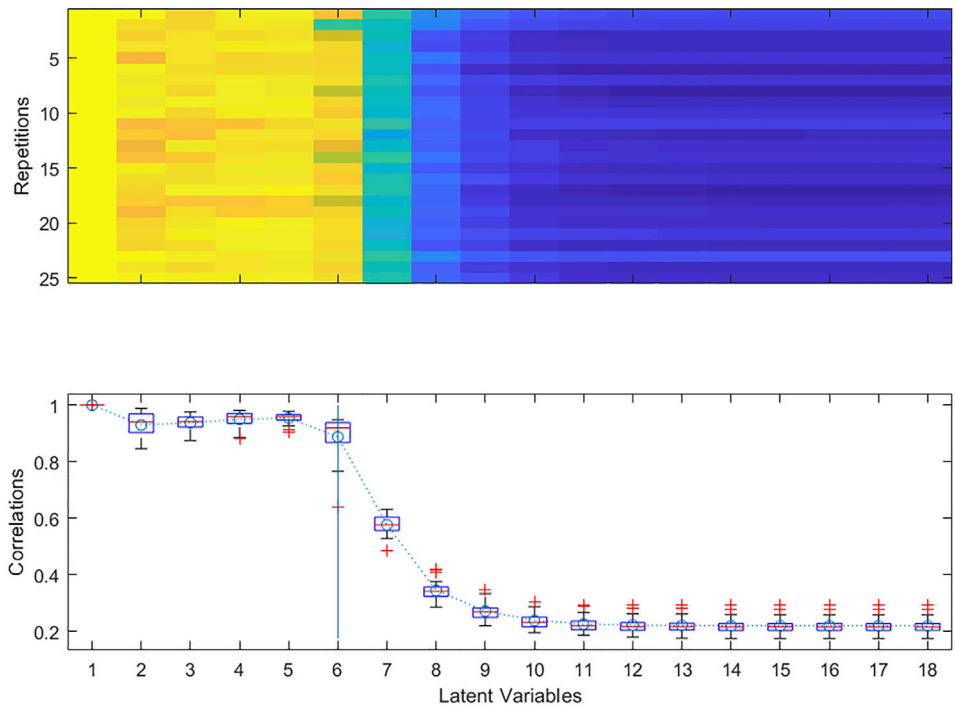

FIGURE 9 | Evolution of the correlations between b-coefficients calculated for 25 randomly selected pairs of subsets of samples for increasing numbers of Latent Variables.

corresponding to information, and later to noise. There are initially significant changes in the b coefficient vectors, due to the fact that the loadings are very different, reflecting different sources of information. Subsequent loadings correspond more and more to noise and change less the shape of the **b**-vectors.

It can therefore be interesting to quantify the correlations between the columns of a matrix **B** containing vectors of b-coefficients calculated with increasing numbers of LVs.

To detect the number of LVs at which point the multi-collinearities increase, we can plot the VIF values of the b-coefficient vectors as a function of the number of LVs. In **Figure 8**, we see that the VIF values remain low up to 6 LVs, and then increase.

In a way similar to the Random_ICA method (Kassouf et al., 2018), one can study whether similar b-coefficients vectors are extracted from two random subsets of the **X** and **Y** matrices. PLS regressions are performed with increasing numbers of LVs on the two subsets. Too many LVs have been extracted when there is no longer a strong correlation between the pair of b-coefficients vectors. To avoid the possibility of a bias being introduced by a particular distribution of the rows into the two blocks, the whole procedure is repeated k times resulting in different sets of blocks, producing a broader perspective for the selection of the number of LVs (**Figure 9**).

Structure of the X Matrix After Each Deflation Step.

Most multivariate analysis methods contain a deflation step where the contribution of each Latent Variables is removed from the matrix before extracting the next Latent Variables. This is true for PCA, PCR and PLS. This process of deflation means that the rows in the deflated matrices contain less and less information and more and more noise. As well, since the remaining variability corresponds more and more to Gaussian noise, the distribution of individuals in the space of the variables gradually approaches that of a hypersphere.

Several criteria can be used to characterize the evolution of the signal/noise ratios in the rows and the sphericity of the deflated matrices so as to determine when all the interesting information has been removed.

Again, the DW criterion can be used, this time to measure the signal-to-noise ratio in each row of the matrix following the successive deflations. **Figure 10** shows the evolution of the distribution of DW values calculated as in **Equation 13**, for each row of the **X** matrix, as a function of the number of Latent Variables extracted.

$$\mathbf{DW} = \frac{\sum_{i=2}^{n} (\boldsymbol{x}_i - \boldsymbol{x}_{i-1})^2}{\sum_{i=1}^{n} \boldsymbol{x}_i^2} \tag{13}$$

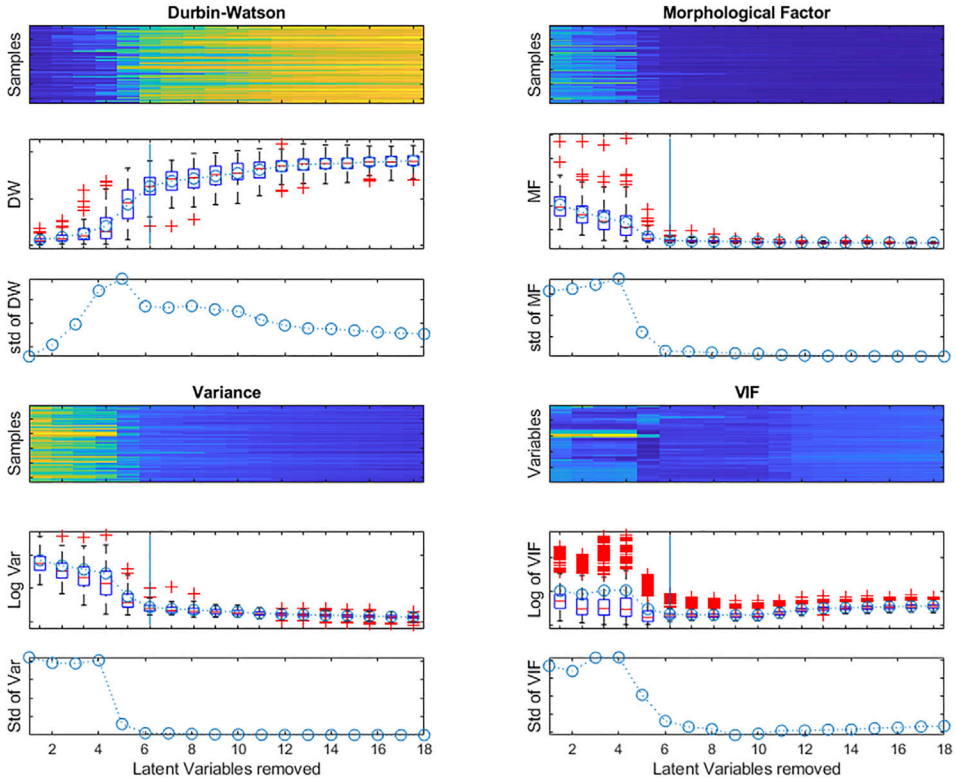

FIGURE 10 | Evolution of the Durbin-Watson (DW) criterion; the Morphological Factor; the Variance calculated for each row of the X matrix during deflation and the log of the VIF for all X-matrix variables after each deflation.

There is a sharp increase in the median value and interquartile interval when 5 latent variables are extracted. The heatmap and boxplot show that not all rows (samples) evolve in the same way, some becoming noisy later than most. This is reflected in the size of the boxplots of the DW values and also in the standard deviation of the values.

As with the DW criterion, the Morphological Factor can be calculated for each row of the matrix after deflation. **Figure 10** also shows the evolution of the distribution of the MF values, as a function of the number of Latent Variables extracted. The values stabilize with the elimination of 6 Latent Variables.

For non-structured data, the variance (or the standard deviation or the Norm) of the matrix rows can be used (**Figure 10**).

As the **X**-matrix is deflated, the sources of variability corresponding to information are eliminated, leaving behind only random noise, so that there are less and less correlations between the variables in the deflated **X**-matrix. To detect the moment when there are no more multi-collinearities between the variables, we can do linear regressions between each variable and all the others and then examine the corresponding R^2 for all successive models. If the R^2 of a variable is close to 1, there is still a linear relationship between this variable and the others.

The VIF is equal to the inverse of $(1\text{-}R^2)$. If the VIF of a variable is greater than 4, there may be multi-collinearities; if the VIF is greater than 10, there are significant multi-collinearities.

To determine whether all information has been eliminated from the **X**-matrix, the VIFs of all the variables can be plotted as a function of the number of LVs extracted, as in **Figure 10**, where only a few variables still have high VIFs after eliminating 6 LVs.

As the X-matrix is deflated, the dispersion of the samples in the reduced multivariate space tends to become spherical, as all the directions of non-random dispersion are progressively removed. Sphericity tests can therefore be applied to the deflated matrices to determine how many LVs are required to remove all interesting dispersions.

Bartlett's test for Sphericity (Bartlett, 1951) compares a matrix of Pearson correlations with the identity matrix. The null hypothesis is that the variables are not correlated. If there is redundancy between variables, it can be interesting to proceed with the multivariate analysis. The formula is given by:

$$\chi^2 = -[(n-1) - (2k+5)6]^* log|\mathbf{R}| \qquad (14)$$

where:

n is the number of observations, k the number of variables, and **R** the correlation matrix of the data in **X**. $|\mathbf{R}|$ is the determinant of **R**.

Bartlett's test in **Figure 11** shows that the deflated matrices are very non-spherical until after 6 LVs have been removed.

Similarly, Hartley and Cochran proposed F-tests based on the ratio of the maximum variance/minimum variance (Hartley, 1950)

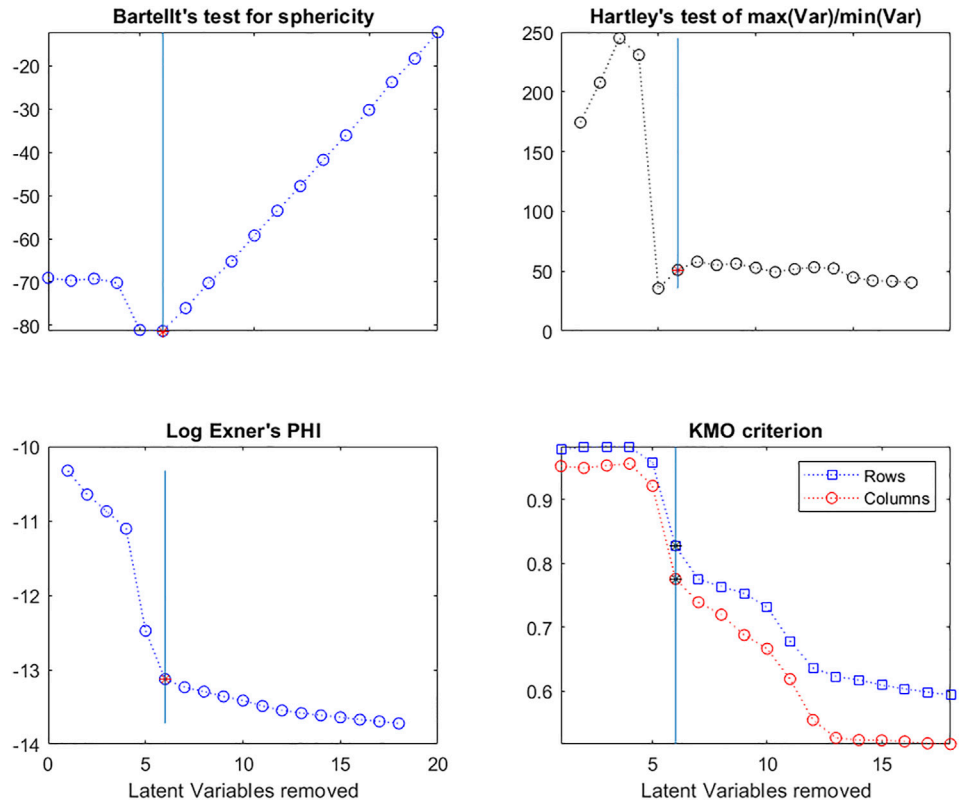

FIGURE 11 | Evolution of Bartlett's test for Sphericity; Hartley's F-test; the Log of Exner's Phi criterion and the KMO criterion for **X**-matrix rows and columns, as a function of the number of Latent Variables in the model.

and the maximum variance/mean variance (Cochran, 1941), respectively. The Hartley criterion in **Figure 11** shows that the deflated matrices are very spherical once 5 LVs are removed.

Exner proposed the Ψ criterion (Exner, 1966; Kindsvater et al., 1974) as a measure of fit of a set of predicted data to a set of experimental data, given by the equation:

$$\psi = \sqrt{\left[\frac{\sum_{i=1}^{nc}\left(X_i - \hat{X}_i\right)^2}{\sum_{i=1}^{nc}\left(X_i - \bar{X}\right)^2}\frac{nc}{nc-k}\right]} \tag{15}$$

where X_i is a data point in the matrix, $\hat{X}_i$ is that data point reproduced using k LVs, n and c are the number of rows and columns in the data matrix and $\bar{X}$ is the grand mean of **X**.

Here Exner's criterion (**Figure 11**) is calculated between the original **X** matrix and each successive deflated matrix to determine at what point there is no longer any similarity between them.

The KMO (Kaiser-Meyer-Olkin Measure of Sampling Adequacy) criterion (Kaiser, 1970; Kaiser, 1974) was developed to determine whether it was useful to conduct a multivariate analysis of a data matrix. For example, if the variables are uncorrelated, it is no use to do a PCA.

The KMO index is given by:

$$\mathbf{KMO} = \frac{\sum_i \sum_{j\neq i} r_{ij}^2}{\sum_i \sum_{j\neq i} r_{ij}^2 + \sum_i \sum_{j\neq i} a_{ij}^2} \tag{16a}$$

where r_{ij} is the correlation between variables i and j, and a_{ij} is the partial correlation, defined as:

$$a_{ij} = \frac{v_{ij}}{\sqrt{v_{ij} + v_{ij}}} \tag{16b}$$

v_{ij} being an element of the inverse of the correlation matrix ($v_{ij} = r_{ij}^{-1}$).

The value of the KMO index varies between 0 (no correlation between variables, thus useless to do a multivariate analysis) and 1 (correlated variables, thus useful to do a multivariate analysis). A KMO value of 0.5 is usually considered the cutoff point below which there is no interest in doing a multivariate analysis. Here this index was calculated for the variables (columns) and for the individuals (rows) in each matrix. We can see (**Figure 11**) that the values are close to 1 until 6 LVs are removed from the matrix and that there is a second decrease after removing 11 LVs. This means that much of the information shared by the original variables and individuals has been removed by 6 LVs, but there is still some present to a lesser extent up to 11 LVs.

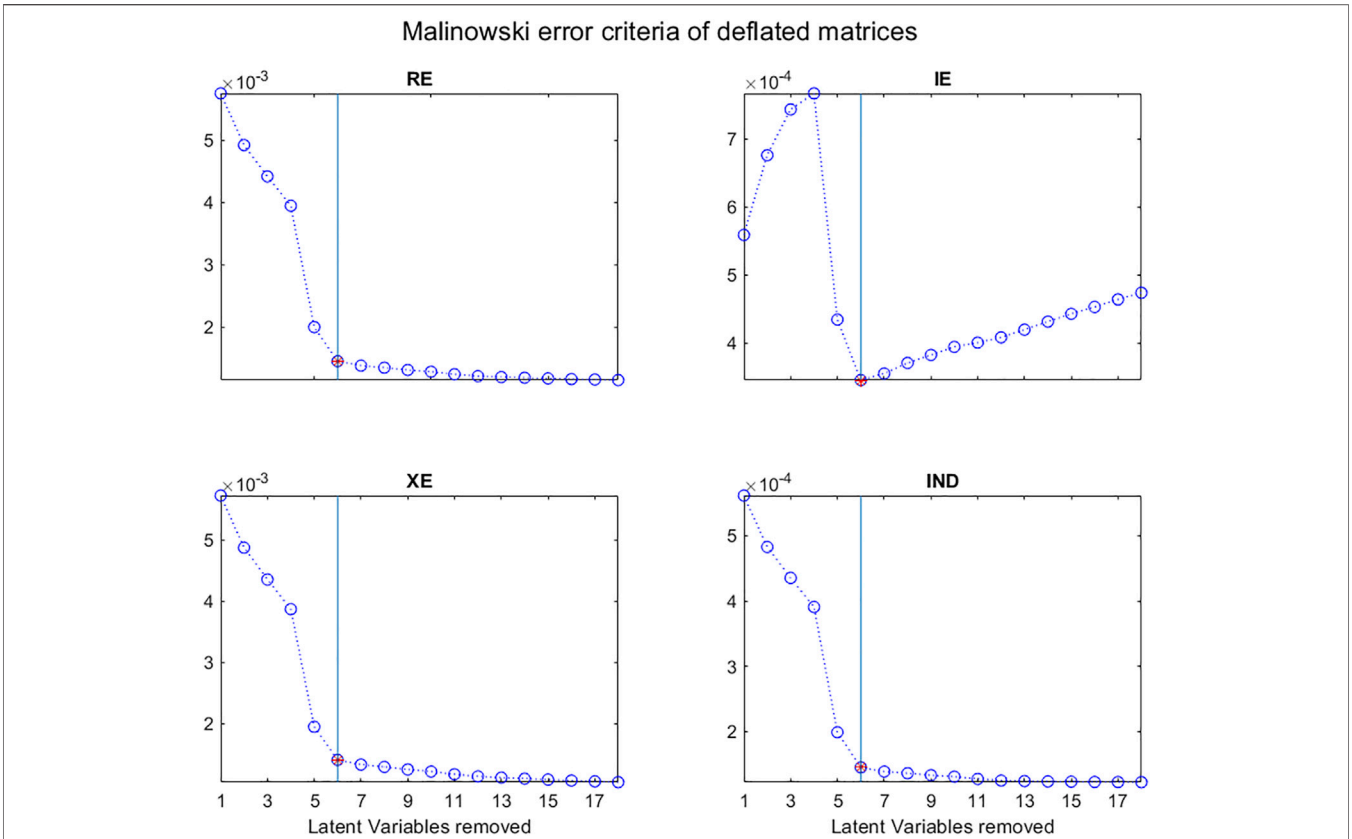

FIGURE 12 | Evolution of the 4 criteria proposed by Malinowski after each deflation.

In 1977, Malinowski (1977a) developed the idea that there were two types of Factors (or Latent Variables) "a primary set which contains the true factors together with a mixture of error and a secondary set which consists of pure error". He also showed that there were three types of errors: RE, real error; XE, extracted error; and IE, Imbedded error, which can be calculated "from a knowledge of the secondary eigenvalues, the size of the data matrix, and the number of factors involved", the secondary eigenvalues being those associated with pure noise.

He considered that if k, the number of LVs associated with the "pure data" is known, the real error is the difference between the pure data and the raw data, that is the Residual Standard Deviation (RSD) given by:

$$RE = RSD = \sqrt{\left[\frac{\sum_{i=k+1}^{c} \lambda_i}{n(c-k)}\right]} \tag{17}$$

where, n and c are the respective number of rows and columns in the data matrix; k the number of factors used to reproduce the data; and λi is the ith eigenvalue.

He stressed that "it was assumed that $n > c$. If the reverse is true, i.e., $n < c$, then n and c must be interchanged in these equations".

He also proposed that the imbedded error (IE) is the difference between the pure data and the data approximated by the multivariate decomposition:

$$IE = \sqrt{\frac{k}{c}} RSD \tag{18}$$

and that the extracted error (XE) is the difference between the data approximated by the multivariate decomposition and the raw data:

$$XE = \sqrt{\frac{c-k}{c}} RSD \tag{19}$$

Malinowski then proposed another empirical criterion to determine the number of Latent Variables in a data matrix (Malinowski, 1977b). This indicator function (IND) is closely related to the error functions described above:

$$IND = \frac{RSD}{(c-k)^2} \tag{20}$$

As can be seen in **Figure 12**, a plot of these criteria as a function of k, the number of LVs, can help to distinguish "pure data" from "error data".

Several criteria have been proposed to estimate the correlation between matrices. Here 3 of them (Dray, 2008) will be used to compare the original **X** matrix with each deflated matrix, the assumption being that these correlations will decrease as the information is being removed.

The RV coefficient (Escoufier, 1973; Robert and Escoufier 1976) is a measurement of the closeness between two matrices and is defined by:

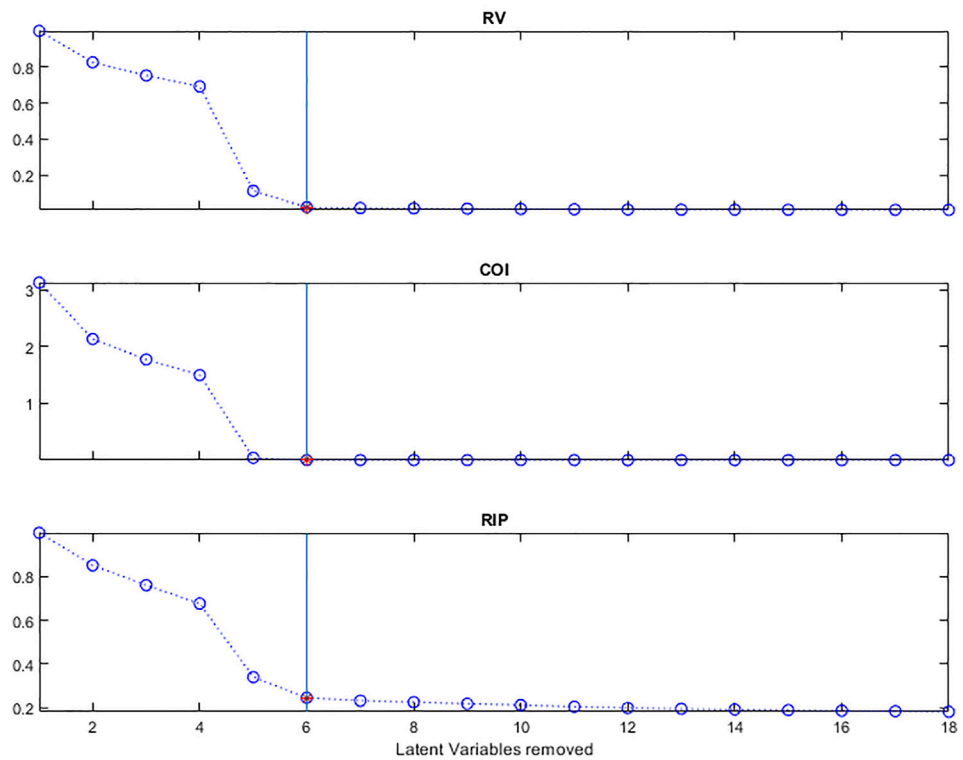

FIGURE 13 | Evolution of RV, COI and RIP of the **X**-matrix after each deflation.

$$RV = \frac{trace\left(X_1 X_1^T X_k X_k^T\right)}{\sqrt{trace\left(X_1^T X_1 X_1^T X_1\right) trace\left(X_k^T X_k X_k^T X_k\right)}} \quad (21)$$

In our case, $\mathbf{X_1}$ is the original matrix, $\mathbf{X_k}$ is the deflated matrix after removing k LVs.

The numerator of the RV coefficient is the co-inertia criterion (COI) (Dray et al., 2003) which is also a measurement of the link between the two matrices:

$$COI = trace\left(X_1 X_1^T X_k X_k^T\right) \quad (22)$$

According to Ramsay et al. (1984) and Kiers et al. (1994), the most common matrix correlation coefficient is the 'inner product' matrix correlation coefficient, which we will call RIP, defined as:

$$RIP = \frac{trace\sqrt{\left(X_1^T X_k\right)}}{\sqrt{trace\left(X_1^T X_1\right) trace\left(X_k^T X_k\right)}} \quad (23)$$

Figure 13 shows the evolution of these 3 measures of the correlation between the original **X** matrix and the matrices after deflation.

CONSENSUS NUMBER OF LATENT VALUES

Given all the criteria that can be calculated, one needs to find a consensus value for the number of LVs to retain in the PLS regression model. Some criteria (RMSEC and RMSECV in **Figure 3**; R^2 and Q^2 in **Figure 5**; Wold's R, Osten's F and Cattell's RPV in **Figure 6**) characterize the proximity of the predicted values to the observed values, but they can be subject to errors due to the particular choice of the calibration and test sets. Others characterize the regression coefficients (B-DW, B_Morph, B_VIF in **Figure 8**) which should not be excessively noisy or of too high a magnitude (B_Var in **Figure 8**). As well, similar B-coefficients vectors should be extracted from subsets of the data matrix (mean of the correlations between regression coefficients vectors in **Figure 9**). Still others characterize the noisy structure of the residual variability in the deflated matrices (mean and standard deviations of DW_X, Morph_X, Var_X and VIF_X in **Figure 10** as well as Malinowski's RE, IE, XE and IND in **Figure 12**).

These deflated matrices should also tend towards a spherical structure (Bartlett_X, Hartley_X, Exner_X, KMO_X_rows, KMO_X_columns in **Figure 11**). As well, as successive components are removed, the correlations between the original matrix and the deflated matrices should decrease (RV, COI and RIP in **Figure 13**).

To create a consensus of all these different types of information, we propose to apply a Principal Components Analysis to the various criteria.

All the criteria were concatenated so that each row corresponded to a number of Latent Values and the columns contained the criteria. Criteria such as DW were used *as is* while for criteria like RMSECV the inverse was used, so that in all cases, earlier LVs are associated with lower values.

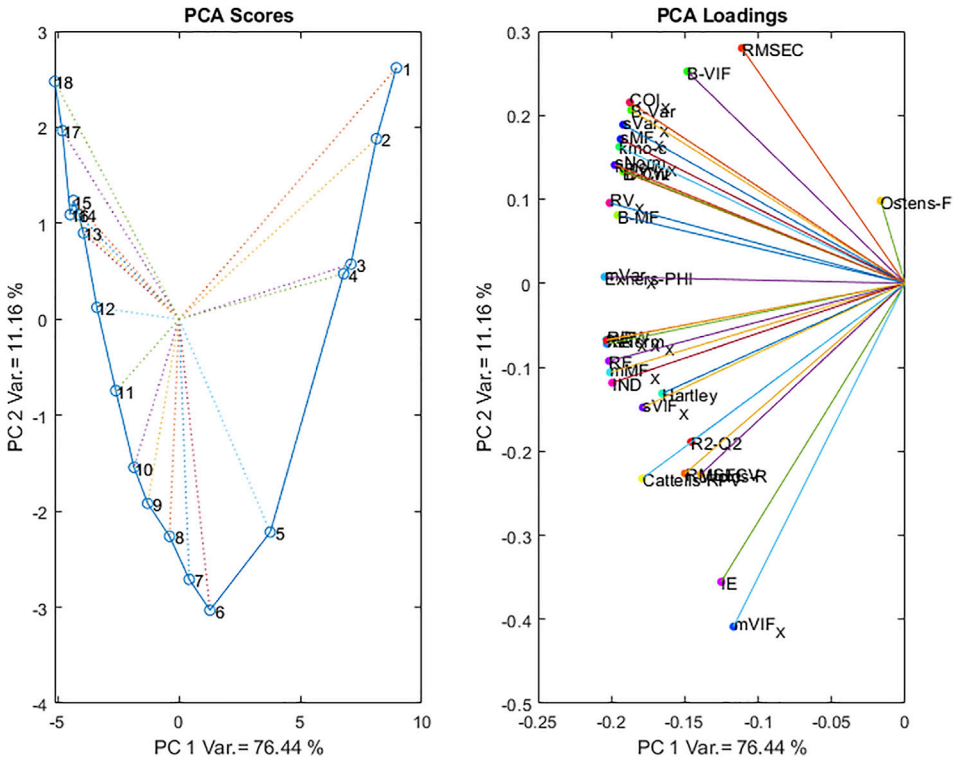

FIGURE 14 | Plot of PC1-PC2 scores and loadings after applying a PCA to the standardized criteria.

The matrix was then z-transformed by subtracting the column means and dividing by the column standard deviations.

The resulting PC1-PC2 Scores plot and Loadings plot are presented in **Figure 14**.

The scores plot shows a clear evolution from low dimensionality models to high dimensionality along PC1, reflecting the increase in all values as the number of LVs increases. The evolution along PC2 corresponds to another phenomenon since the scores are highly positive for both small and large numbers of LVs, with a very clear negative minimum for a model at 6 LVs. The loadings plots shows an opposition between RMSEC, COI, std_Var_X, std_Morph_X and most of the criteria based on the B-coefficients vectors on the positive side; while mean_VIF_X, IE, RMSECV, Wold's R, Cattell's RPV, R2-Q2 and most of the criteria based on the deflated X matrices are on the negative side. This contrast between the criteria based on the B-coefficients vectors and those based on the deflated X matrices shows their complementary nature.

Only the first 2 PCs are presented as the following scores (corresponding to models with increasing numbers of LVs) did not have any interpretable structure.

CONCLUSION

PLS regression is a high-performance calibration and prediction method to link predictive **X**-variables to the **Y**-variables to be predicted, even when variables are highly correlated and in very large numbers.

However, adjusting the number of latent variables in the model is crucial. This adjustment should be done on the basis of several criteria.

To do this, various methods can be used:

The most common method is to observe the evolution of calibration errors (RMSEC) and validation or cross validation errors (RMSEV or RMSECV); One can also examine the evolution of the vectors of regression coefficients. This also provides information on the role of the variables or spectral components in the model; Finally, the evolution in the structure of the rows and columns as well as the sphericity of the X-matrix after each deflation step, can be examined.

To do this we have proposed applying a Principal Components Analysis to a collection of criteria characterizing the different aspects of models obtained with increasing numbers of Latent Variables. The set of criteria used in the present study is far from exhaustive, and the efficacy of the method may even be improved by including others.

Matlab function to calculate most of the non-trivial criteria are to be found at: https://github.com/DNRutledge/LV_Criteria.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Interested readers should contact the authors of the article cited as producers of the data. Requests to access these datasets should be directed to nathalie.dupuy@imbe.fr.

AUTHOR CONTRIBUTIONS

DR: Conception, calculations, writing J-MR: Corrections, calculations, writing ML: Corrections, calculations, writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frans.2021.754447/full#supplementary-material

REFERENCES

Akaike, H. (1969). Fitting Autoregressive Models for Prediction. *Ann. Inst. Stat. Math.* 21, 243–247. doi:10.1007/BF02532251

Bartlett, M. S. (1951). The Effect of Standardization on A X2 Approximation in Factor Analysis. *Biometrika* 38 (3/4), 337–344. doi:10.1093/biomet/38.3-4.337

Bissett, A. C. (2015). *Improvements to PLS Methodology, PhD*. Manchester: University of Manchester. Available at: http://www.manchester.ac.uk/escholar/uk-ac-man-scw:261814.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behav. Res.* 1 (2), 245–276. doi:10.1207/s15327906mbr0102_10

Cochran, W. G. (1941). The Distribution of the Largest of a Set of Estimated Variances as a Fraction of Their Total. *Ann. Eugenics* 11 (1), 47–52. doi:10.1111/j.1469-1809.1941.tb02271.x

Denham, M. C. (2000). Choosing the Number of Factors in Partial Least Squares Regression: Estimating and Minimizing the Mean Squared Error of Prediction. *J. Chemometrics* 14 (4), 351–361. doi:10.1002/1099-128X(200007/08)14:4<351::AID-CEM598>3.0.CO;2-Q

Dray, S., Chessel, D., Chessel, D., and Thioulouse, J. (2003). Co-inertia Analysis and the Linking of Ecological Data Tables. *Ecology* 84, 3078–3089. doi:10.1890/03-0178

Dray, S. (2008). On the Number of Principal Components: A Test of Dimensionality Based on Measurements of Similarity between Matrices. *Comput. Stat. Data Anal.* 52, 2228–2237. doi:10.1016/j.csda.2007.07.015

Durbin, J., and Watson, G. S. (1971). Testing for Serial Correlation in Least Squares Regression. III. *Biometrika* 58, 1–19. doi:10.2307/2334313

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* 29, 751–760. doi:10.2307/2529140

Exner, O. (1966). Additive Physical Properties. I. General Relationships and Problems of Statistical Nature. *Collect. Czech. Chem. Commun.* 31, 3222–3251. doi:10.1135/cccc19663222

Faber, N. M. (1999). Estimating the Uncertainty in Estimates of Root Mean Square Error of Prediction: Application to Determining the Size of an Adequate Test Set in Multivariate Calibration. *Chemometrics Intell. Lab. Syst.* 49 (1), 79–89. doi:10.1016/S0169-7439(99)00027-1

Ferré, J. (2009). "Regression Diagnostics," in *Comprehensive Chemometrics*. Editors S. D. Brown, R. Tauler, and B. Walczak (Amsterdam: Elsevier), 33–89. 9780444527011. doi:10.1016/B978-044452701-1.00076-4

Galtier, O., Dupuy, N., Le Dréau, Y., Ollivier, D., Pinatel, C., Kister, J., et al. (2007). Geographic Origins and Compositions of virgin Olive Oils Determinated by Chemometric Analysis of NIR Spectra. *Analytica Chim. Acta* 595, 136–144. doi:10.1016/j.aca.2007.02.033

Hartley, H. O. (1950). The Maximum F-Ratio as a Short-Cut Test for Heterogeneity Op Variance. *Biometrika* 37, 308–312. doi:10.1093/biomet/37.3-4.308

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

Kaiser, H. F. (1970). A Second Generation Little Jiffy. *Psychometrika* 35, 401–415. doi:10.1007/BF02291817

Kaiser, H. F. (1974). An index of Factorial Simplicity. *Psychometrika* 39, 31–36. doi:10.1007/BF02291575

Kassouf, A., Jouan-Rimbaud Bouveresse, D., and Rutledge, D. N. (2018). Determination of the Optimal Number of Components in Independent Components Analysis. *Talanta* 179, 538–545. doi:10.1016/j.talanta.2017.11.051

Kiers, H. A. L., Cléroux, R., and Ten Berge, J. M. F. (1994). Generalized Canonical Analysis Based on Optimizing Matrix Correlations and a Relation with IDIOSCAL. *Comput. Stat. Data Anal.* 18, 331–340. doi:10.1016/0167-9473(94)90067-1

Kindsvanter, J. H., Weiner, P. H., and Klingen, T. J. (1974). Correlation of Retention Volumes of Substitutued Carboranes with Molecular Properties in High Pressure Liquid Chromatography Using Factor Analysis. *Anal. Chem.* 46, 982–988. doi:10.1021/ac60344a032

Krämer, N., and Sugiyama, M. (2011). The Degrees of Freedom of Partial Least Squares Regression. *J. Am. Stat. Assoc.* 106 (494), 697–705. doi:10.1198/jasa.2011.tm10107

Lesnoff, M., Roger, J. M., and Rutledge, D. N. (2021). Monte Carlo Methods for Estimating Mallows's Cp and AIC Criteria for PLSR Models. Illustration on Agronomic Spectroscopic NIR Data. *J. Chemometrics*. doi:10.1002/cem.3369

Li, B., Morris, J., and Martin, E. B. (2002). Model Selection for Partial Least Squares Regression. *Chemometrics Intell. Lab. Syst.* 64, 79–89. doi:10.1016/S0169-7439(02)00051-5

Malinowski, E. R. (1977b). Determination of the Number of Factors and the Experimental Error in a Data Matrix. *Anal. Chem.* 49 (4), 612–617. doi:10.1021/ac50012a027

Malinowski, E. R. (1977a). Theory of Error in Factor Analysis. *Anal. Chem.* 49 (4), 606–612. doi:10.1021/ac50012a02710.1021/ac50012a026

Mallows, C. L. (1973). Some Comments onCp. *Technometrics* 15 (4), 661–675. doi:10.1080/00401706.1973.10489103

Marquardt, D. W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics* 12 (3), 591–612. doi:10.1080/00401706.1970.1048869910.2307/1267205

Meloun, M., Čapek, J., Mikši´k, P., and Brereton, R. G. (2000). Critical Comparison of Methods Predicting the Number of Components in Spectroscopic Data. *Analytica Chim. Acta* 423, 51–68. doi:10.1016/S0003-2670(00)01100-4

Osten, D. W. (1988). Selection of Optimal Regression Models via Cross-Validation. *J. Chemometrics* 2, 39–48. doi:10.1002/cem.1180020106

Ramsay, J. O., Ten Berge, J., and Styan, G. P. H. (1984). Matrix Correlation. *Psychometrika* 49, 403–423. doi:10.1007/BF02306029

Robert, P., and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Appl. Stat.* 25, 257–265. doi:10.2307/2347233

Rutledge, D. N., and Barros, A. S. (2002). The Durbin-Watson Statistic as a Morphological Estimator of Information Content. *Analytica Chim. Acta* 446, 279–294. doi:10.1016/S0003-2670(01)01555-0

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*. 2nd ed. New York: Springer.

Wang, J.-H., Liang, Y.-Z., Jiang, J.-H., and Yu, R.-Q. (1996). Local Chemical Rank Estimation of Two-Way Data in the Presence of Heteroscedastic Noise: A Morphological Approach. *Chemometrics Intell. Lab. Syst.* 32, 265–272. doi:10.1016/0169-7439(95)00072-0

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., and Faber, K. (2007). A Randomization Test for PLS Component Selection. *J. Chemometrics* 21, 427–439. doi:10.1002/cem.1086

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* 20 (4), 397–405. doi:10.2307/126763910.1080/00401706.1978.10489693

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fchem.2021.748723

Fusing NIR and Process Sensors Data for Polymer Production Monitoring

Lorenzo Strani[1], *Erik Mantovani*[2], *Francesco Bonacini*[2], *Federico Marini*[3] and *Marina Cocchi*[1]*

[1]Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Modena, Italy, [2]Research Center, Versalis (ENI) S.p.A., Mantvoa, Italy, [3]Department of Chemistry, University of Roma La Sapienza, Roma, Italy

Process analytical technology and multivariate process monitoring are nowadays the most effective approaches to achieve real-time quality monitoring/control in production. However, their use is not yet a common practice, and industries benefit much less than they could from the outcome of the hundreds of sensors that constantly monitor production in industrial plants. The huge amount of sensor data collected are still mostly used to produce univariate control charts, monitoring one compartment at a time, and the product quality variables are generally used to monitor production, despite their low frequency (offline measurements at analytical laboratory), which is not suitable for real-time monitoring. On the contrary, it would be extremely advantageous to benefit from predictive models that, based on online sensors, will be able to return quality parameters in real time. As a matter of fact, the plant setup influences the product quality, and process sensors (flow meters, thermocouples, etc.) implicitly register process variability, correlation trends, drift, etc. When the available spectroscopic sensors, reflecting chemical composition and structure, consent to monitor the intermediate products, coupling process, and spectroscopic sensor and extracting/fusing information by multivariate analysis from this data would enhance the evaluation of the produced material features allowing production quality to be estimated at a very early stage. The present work, at a pilot plant scale, applied multivariate statistical process control (MSPC) charts, obtained by data fusion of process sensor data and near-infrared (NIR) probes, on a continuous styrene-acrylonitrile (SAN) production process. Furthermore, PLS regression was used for real-time prediction of the Melt Flow Index and percentage of bounded acrylonitrile (%AN). The results show that the MSPC model was able to detect deviations from normal operative conditions, indicating the variables responsible for the deviation, be they spectral or process. Moreover, predictive regression models obtained using the fused data showed better results than models computed using single datasets in terms of both errors of prediction and R^2. Thus, the fusion of spectra and process data improved the real-time monitoring, allowing an easier visualization of the process ongoing, a faster understanding of possible faults, and real-time assessment of the final product quality.

Keywords: process monitoring, data fusion, MSPC charts, NIR online, styrenic polymers

OPEN ACCESS

Edited by:
Paolo Oliveri,
University of Genoa, Italy

Reviewed by:
Prats-Montalbán José Manuel,
Polytechnic University of Valencia, Spain
Rodrigo Rocha de Oliveira,
University of Barcelona, Spain

***Correspondence:**
Marina Cocchi
marina.cocchi@unimore.it

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *28 July 2021*
Accepted: *09 September 2021*
Published: *22 October 2021*

Citation:
Strani L, Mantovani E, Bonacini F, Marini F and Cocchi M (2021) Fusing NIR and Process Sensors Data for Polymer Production Monitoring. Front. Chem. 9:748723. doi: 10.3389/fchem.2021.748723

INTRODUCTION

A large number of sensors, such as thermocouples, pressure gauges, and flow indicators, which generate an enormous amount of data, are normally installed in petrochemical production plants. The plant operators use these process sensors to control production and monitor operating conditions (Kourti and MacGregor, 1995). The aim is to reduce production faults and defects resulting from accidental plant malfunctions, changes in product characteristics (molecular weight, particle size, etc.), and nonoptimal conditions, caused by the complexity of the process or by its tendency to get contaminated that generates frequent maintenance needs. The collected data are used for the control and optimization of processes and also for extracting significant information to predict the properties that define the quality of the final product in real time. Furthermore, in all industrial processes, energy saving, efficient use of raw materials, and optimal production planning are essential. The measurements made by the sensors in the plants can be used for these needs. The production control in the petrochemical industry, as well as in many others, is based on the knowledge and experience of the technical operators and is mainly supported by single univariate control charts developed for a few selected sensors and monitoring points (Chaudhry and Higbie, 1989). The control is carried out by verifying that the values of the selected parameters fall within a predetermined and carefully chosen confidence interval. As a process always presents variability, it is fundamental to define the standard operating conditions, according to which the process can be considered stable around its natural variability and therefore within the confidence limits of the monitored process parameters (Ferrer-Riquelme, 2009). The plant operators are perfectly aware of the optimal values of the parameters and their confidence intervals, but since more than one variable is used to monitor the entire process, it results in a large number of control charts to pay attention to. When the process encounters an anomaly and goes out of the range of standard operating conditions, it is very likely that several parameters would change simultaneously, due to the correlation that exists between the variables, and it would be very difficult for operators to identify the source of the problem. The sources of variability during production can be related to impurities, defective sensors, plant aging, leaks, and many other possible causes.

Multivariate statistical process control, instead of focusing on individual variables, focuses on the entire group of process variables and their correlation (Kourti, 2006). In this way, the plant operators can identify anomalies, reset the plant parameters, change the raw material, and, in general, properly fix all the other possible events that cause a change in the conditions of the process. This method allows for monitoring the production through few multivariate control charts. It is based on the concept of benefiting from the correlation structure of the process variables, which allows the compression of the responses of a large number of sensors into a few components (the latent variables). In this way, it will be possible to parsimoniously describe the sources of variability in the process (Kourti, 2009) and its time evolution by a few selected trajectories and establishing confidence limits in order to show how far the current condition is from the desired or normal operating situation.

Process sensors that typically measure temperature, pressure, flow, etc., provide information of the process ongoing, but they do not allow the operators to directly know the status of the product. In order to obtain chemical and physical information of the product in real time, near-infrared (NIR) spectroscopic probes are often installed in crucial steps of the process. NIR spectroscopy performs fast, and it is nondestructive and low-invasive on/inline measurements, making it perfectly suitable for being used as a process analyzer. The fusion of NIR data with process sensors data to build multivariate statistical process control (MSPC) charts provided successful results in the three different examples proposed by de Oliveira et al., 2020, in the pharmaceutical and petrochemical fields. In general, some studies conducted in collaboration with petrochemical companies reported the use of multivariate statistical control methods, showing numerous successes (Skagerberg et al., 1992; Macho and Larrechi, 2002; Kourti, 2005; Ferrer, 2007; Bonacini et al., 2013; de Oliveira et al., 2017), suggesting that in recent years, industries have opened up to the use of multivariate techniques, taking advantage of them.

In this context, the present work aimed at building PCA-based MSPC charts from the data fusion of spectroscopic data collected by two NIR probes (located at an early reaction step and close to the final stage, respectively) with process sensors data on a continuous styrenic polymers production process. Furthermore, PLS regression was used for the real-time prediction of selected quality parameters.

MATERIALS AND METHODS

Plant Description

The monitoring of the styreneacrylonitrile (SAN) production has been carried out in the Versalis (ENI) company industrial pilot plant, operating continuously. A schematic representation of the plant is shown in **Figure 1**. The most relevant plant sectors for the present study are the two reactors (R1 and R2), where the polymer formation occurs, and the cutting zone (CZ), a final section where the finished product, i.e., the polymer, is reduced by cutting in small pieces. A total of 52 process sensors are installed throughout the process lines, of which 32 are for measuring the temperature, 11 for the pressure, 7 for the flow, and 2 for the motor speed. Furthermore, two NIR probes were installed in crucial steps of the process: one between R1 and R2 (NIR1) and the other right before the CZ (NIR2).

The monitoring of the SAN production occurred from February 4 to February 23, 2016, and the data were collected every 5 min. In this period, there was a deliberate variation of settings for some of the process sensors at the end of February 11, a pause and restart of the production during the morning on February 12, and a change in the formulation of the product on February 15 (i.e., an increase of the chain transfer amount). The settings variation was carried out in order to test how the plant

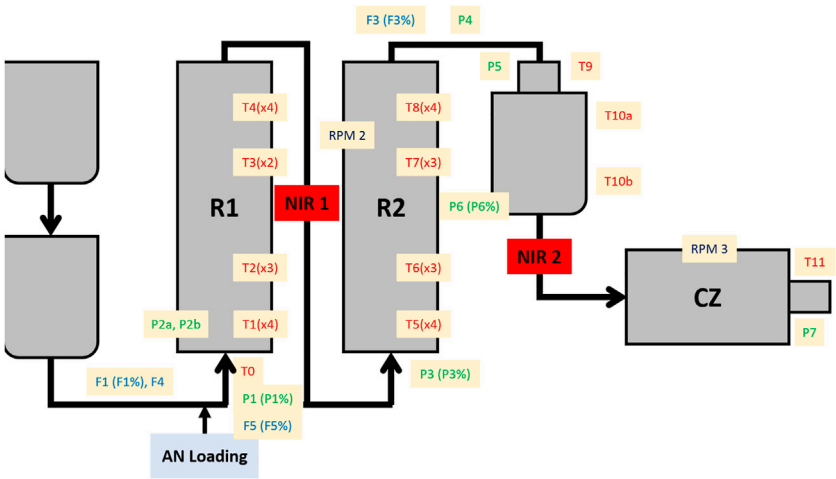

FIGURE 1 | Schematic representation of the SAN production plant. R1 = first reactor; R2 = second reactor; NIR1 = first NIR probe; NIR2 = second NIR probe; CZ = cutting zone; AN = acrylonitrile; T = temperature sensors; P = pressure sensors; F = flow sensors; RPM = motor speed sensors. A percentage symbol after the sensor name indicates the opening extent of a valve linked to the specific sensor.

would react to this kind of modification in view of the chain transfer amount increase.

Reference Analysis

With the aim of assessing the quality of SAN polymer, two different parameters were evaluated: Melt Flow Index (MFI) and percentage of bound acrylonitrile. SAN samples were immediately collected after being cut and brought to the laboratory for the offline analyses.

MFI is an analysis that indicates the fluidity of a molten polymer, providing information about the fluid dynamic behavior of the material. The analysis is carried out by measuring the quantity of matter in grams that passes through a capillary (with a known and standard section) at a temperature of 220°C under the pressure of a weight of 10 kg in 10 min. The results generally range from 4 g, which denotes a very hard product, to 30 g, indicating a highly fluid product, and depend on the molecular weight and on the possible presence of fluidifying agents (Shenoy and Saini, 1986). In this study, 196 MFI analyses were carried out, ranging from 3.1 to 18 g and covering homogeneously the considered time range.

The amount of bonded acrylonitrile (%AN) in SAN samples is measured in order to define how much chemical and thermal resistance the material has. To determine %AN amount in the SAN copolymer, an NIR analysis is performed offline with a Matrix FT-NIR spectrometer (Bruker Optics, Milan, Italy). The sample, in the form of a granule, is analyzed with an integrating sphere, and two NIR spectra are recorded for each of them. A Vario El Elementar (Waltham, MA, United States) CHNS elemental analyzer, used as a reference method, calibrated the NIR spectrometer (the multivariate calibration curve was previously established by PLS regression). In total, 218 %AN analyses were performed ranging from 13.37 to 16.6%, covering homogeneously the considered time range.

NIR Spectroscopy

The on-line monitoring of SAN production was carried out with a Matrix FT-NIR spectrometer (Bruker Optics, Milan, Italy), connected with a probe (HT immersion probe, Drawing-no. 661.2350_1, Hellma GmbH and Co. KG, Müllheim, Germany) via optical fibers (length: 50 m, diameter: 600 µm). These special polymer fibers are directly coupled to the process pipe in high temperature and stress conditions. Spectra were collected in transmission mode (path length: 5 mm) every 5 min in the whole NIR spectral range (12,500–4,000 cm^{-1}) for a total of 5,434 acquisitions, with a resolution of 4 cm^{-1} and 64 scans for both background and spectra.

Data Analysis

Datasets

Data were arranged in three different datasets: two containing the spectra collected by NIR1 and NIR2, and a third one containing the process sensor data (PS). These datasets were analyzed both singularly and merged together, applying low- and mid-level data fusion techniques. A schematic representation of data arrangement is shown in **Figure 2**. The low-level data fusion was achieved simply concatenating NIR1 and NIR2 datasets row-wise, obtaining a single dataset with the same number of rows (data points) as the previous ones, but with twice the number of columns (wavenumbers). For the mid-level data fusion, two steps were required: first, the information contained in NIR1 and NIR2 datasets was extracted via PCA, selecting the proper number of PCs with the aim of retaining just the relevant information contained in the data. Then, the features (scores) obtained in this way were concatenated with the PS dataset, creating a single dataset containing the information of both NIR probes and process sensors data (NPS). The datasets assembly was performed taking into account the residence time according to the position of each sensor and NIR probe along the process line and the process itself. In this way, each data point present in the

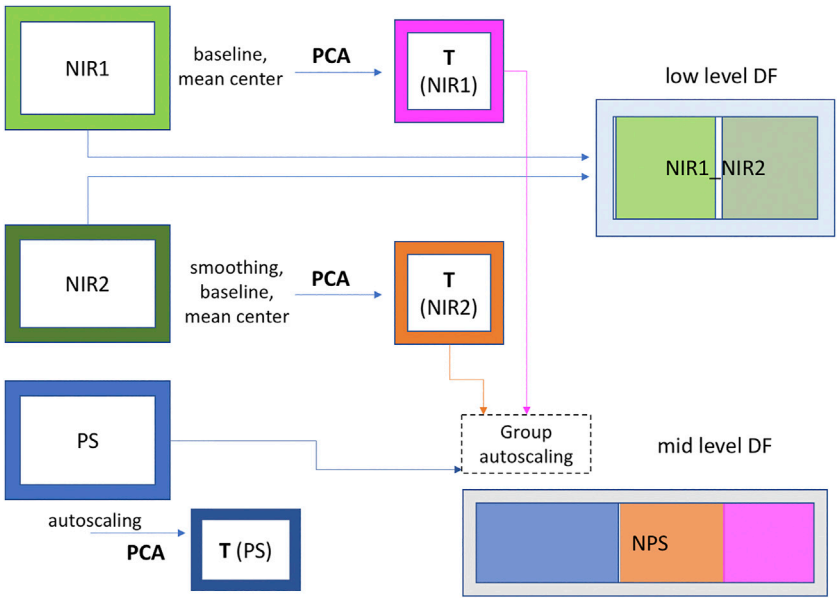

FIGURE 2 | Schematic representation of data sets assembly.

datasets, which contains information collected at different times, was referred to the same material.

The spectral range considered for the data analysis was 6,200–4,700 cm^{-1}, as other regions were characterized by high noise and baseline regions, i.e., no bands linked to either reactant or product are present. Spectra were pretreated in order to improve the quality of the analysis. In particular, automatic weighted least square method has been used for the baseline correction, followed by mean centering. Furthermore, only for the spectra acquired by NIR 2, prior to baseline correction, smoothing (SavitzkyGolay method, filter width seven points, polynomial order 1) was applied with the purpose of reducing noise. Autoscaling followed by block scaling was applied on the NPS dataset in order to avoid that a single block of data (NIR1 and NIR2 features and PS) could contribute more than the others just for containing a greater number of variables.

PCA and MSPC Charts

PCA, described by **Eq. 1**, was used both to perform the initial exploratory data analysis and to build MSPC charts.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{T} + \boldsymbol{E}. \tag{1}$$

Here, **X** is a data matrix composed of *m* rows (samples) and *n* columns (variables). The scores matrix **T** describes how each sample relates to each other, whereas the loading matrix **P** contains information about the influence of the measured variables on the model and their correlation structure. **E** is the residual matrix, which contains the unmodeled variation, has the same dimensions of **X**, and it is obtained by subtraction of the reconstructed (by the PCA model) data (TP^{T}) from **X**. Thus, the original data is compressed into a fewer number of independent variables, i.e., principal components (PCs), orthogonal to each other. Therefore, a new projection space is created, smaller in size, whose coordinates are represented by the PCs.

The PCA-based MSPC chart models were built using the data from February 4 to February 10, before the variation of some of the process settings, whereas data from February 11 to February 15 were used to validate the model. Data points acquired after February 15, corresponding to the formulation change, were not used in this part of the work. The cross-validation scheme used for the internal validation of the models was contiguous blocks with ten cancelation groups, in order to mimic the routine situation in which the monitoring MSPC model is going to be applied.

MSPC is based on two distinct monitoring charts reporting as function of time the distance in PCA scores space (T^2) and the squared residuals (Q), respectively:

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{ia}^2}{\lambda_a} \tag{2}$$

$$Q_i = \sum_{m=1}^{M} e_{im}^2 \tag{3}$$

where t_{ia} is the score value for the *ath* component of a given sample (time point observation *i*), λ_a the corresponding eigenvalue, and e_{im} its residual value for a given variable *m*.

The T^2 and Q acceptance limits are calculated based on Hotelling-T^2 (Nomikos and Mac Gregor 1995) and χ^2 statistics, calculated with the Jackson and Mudholkar approximation, respectively.

The T^2 parameter indicates the distance of a sample in the model space, which means that a sample with a high T^2 value has a distance from the center of the model larger than what is usually

expected but is still described properly by the model. On the other hand, the Q parameter describes the distance of a sample from the model space, indicating an anomalous condition with respect to the optimal operating conditions, i.e., the conditions under which the model was built.

Once an anomalous sample is detected to assess the sensors responsible for the deviation, the T^2 or Q contribution plots, depending on which chart detected it, can be displayed. The contributions to T^2 for the *i*th sample, $\boldsymbol{t}_{con,i}$, are a vector calculated from:

$$\boldsymbol{t}_{con,i} = t_i \lambda^{-1/2} \mathbf{p}^T. \quad (4)$$

Here, **P** is the loading matrix (n° of variables x n° of components) and $\mathbf{P}^T$ its transpose.

While the $\mathbf{Q}_i$ contribution is simply a vector holding the *i*th sample squared residuals for each sensor multiplied by its sign, the contribution plots can aid fault diagnosis. In a T^2 contribution, a high absolute value of the contribution of a given variable denotes a problem with that specific variable which assumes an extreme value, higher or lower depending on the sign of the contribution, with respect to the other ones. The interpretation of Q contribution is less straightforward because it signals that the correlation structure of the variables (with a high absolute value of the contribution) has changed. Thus, if, e.g., two variables have a high positive and negative contribution value, respectively, it could be that for the corresponding out-of-control observations, these variables are inversely correlated, while for the normal operative conditions observation, they were directly correlated. Inspection of scatter plot of one variable vs. the other may be used to have a confirmation (Westerhuis et al., 2000).

Predictive Models

PLS regression was used with the aim of developing predictive models of SAN quality in real time. Venetian blinds cross-validation with ten cancelation groups was used to establish the number of PLS components. The external validation of the PLS models was performed using a test set whose sample was not used for the model computation. Since MFI and %AN reference analyses were not always performed on the same samples and the number of the two kinds of analyses was not the same, PLS models and predictions were carried out as follows: the models were calculated using the 130 samples on which both analyses were made, whereas the predictions were performed using samples on which only one of the two determinations was carried out, i.e., 66 for MFI and 88 for %AN. To evaluate the reliability of the models both RMSECV and RMSEP, i.e., the root mean square error in cross-validation and in prediction, respectively, and the corresponding values of the coefficient of determination (R2) were taken into account. A total of 4 PLS models were computed, three using as X block each of the three datasets NIR1, NIR2, and PS individually, and the last one using the fused NPS dataset. Besides, the Y block contains the results of the MFI and %AN analysis together (PLS2 models); even if, for the reasons mentioned above, predictions on the validation samples were evaluated separately. Autoscaling was applied on Y block, since it contains values obtained with different techniques, having different ranges and scales.

Software

Data elaboration has been carried out by using PLS Toolbox (version 8.9, Eigenvector Research Inc. WA, United States) (MathWorks, MA, United States).

RESULTS AND DISCUSSION

Exploratory Data Analysis

Each different data block was analyzed with PCA in order to visualize and extract features and relevant information on the process. The first PCA analysis was carried out on the NIR1 dataset, choosing five PCs for the model computation that explain 95% of the total variance. **Figure 3A** represents the scores on the first PC as a function of time. It is possible to observe a slow but constant decrease of the scores over time, until the temporary stop of the production, highlighted by the red bar. During the last 2 days of production, samples start to increase their score values, behaving differently from the previous ones. The spectral bands responsible for the data variation are shown in the loadings plot (**Figure 3B**). Bands at 6,130, 6,000, and 4,720 cm^{-1} can be ascribed to the styrene monomer, whereas the band at 5,900 cm^{-1} is related to the forming SAN polymer (Takeuchi et al., 1968). These bands present higher intensity in samples with positive scores and lower intensity in samples with negative scores, suggesting a slow decrease overtime of their intensity until the production stops. **Figure 3C** shows the scores of the first PC as a function of time related to the PCA performed on the NIR2 dataset. Also, in this case, five PCs were selected for the model computation, explaining 99.8% of the total variance. At this final stage, a general more stable trend over time is observed, with the exception of three distinct moments: 20 h before and 4 h after the production stops, and at the very end of the period taken into account. Looking at the corresponding loadings plot (**Figure 3D**), it can be observed how these extreme samples have negative scores, meaning that with respect to the other time points, they are characterized by a less intense band at 5,900 cm^{-1}, suggesting a lower extent of polymer formation. Finally, PCA was also carried out on the PS dataset (**Figure 4**). In this respect, the model was computed considering three PCs explaining 84.9% of the total variance. The scores of PC1 as a function of time (**Figure 4A**) provide a different trend than those of the PCA performed on spectral data, as in this case, the measurements made after the production stop, with positive scores, resulted clearly different from the others without returning to the stable range of values before the stopping. The loadings plot (**Figure 4B**) explains how samples collected after the production pause show, among others, high values for temperature sensors linked to the two reactors (T1–T8).

MSPC Charts

From these PCA models, it is clear how each data block provides different information about the processes; therefore, two different

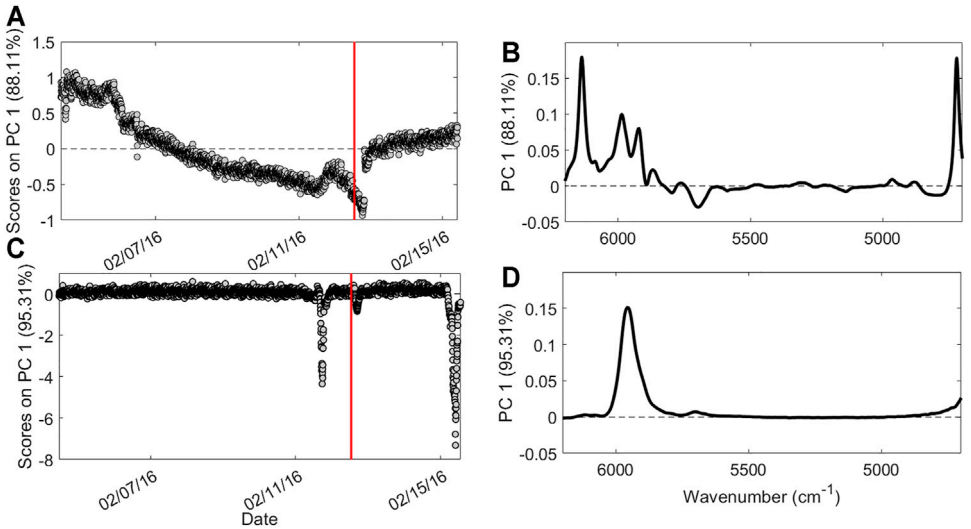

FIGURE 3 | Results of the Exploratory Data Analysis performed on spectral data. Scores as a function of time **(A)** and loadings **(B)** on PC1 for NIR1 dataset; scores as a function of time **(C)** and loadings **(D)** on PC1 for NIR2 dataset. Red bar indicates the moment of the production pause.

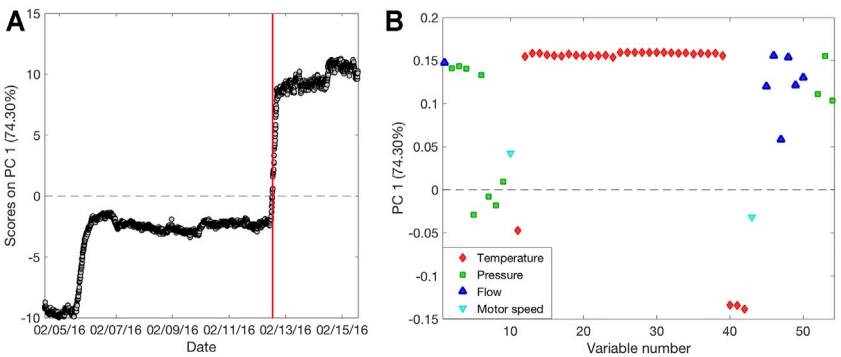

FIGURE 4 | Results of the Exploratory Data Analysis performed on process sensors data. Scores as a function of time **(A)** and loadings **(B)** on PC1 for PS dataset. Red bar indicates the moment of the production pause.

data fusion approaches were applied. The low-level data fusion approach was performed merging NIR1 and NIR2 datasets, in order to gather the spectral information collected in the two key steps of the process, namely between the two reactors and before the CZ. The PCA carried out on this dataset (data not shown for the sake of brevity) confirmed what already showed by PCA performed on NIR1 and NIR2 datasets separately. Scores and loadings profiles related to PC1 and PC2 are almost identical to the ones obtained by NIR2 and NIR1 PCA models, respectively. Furthermore, MSPC charts based on T^2 and Q were built with the modality described in chapter 2.4.2. The results obtained were good, but it would be difficult for the plant operators to understand the nature of an occurring problem, being the spectral interpretation above their expertise. For this reason, a mid-level data fusion approach was applied, considering also the information contained in the process sensors data, i.e., PS dataset. Hence, NPS dataset was created merging the scores obtained from PCA performed on NIR1 and NIR2 datasets together with PS data. A further PCA was carried out, using three PCs to build the model. Also, in this case, MSPC charts were computed as described in chapter 2.4.2.

Figure 5A shows the MSPC chart related to the T^2 parameter, which describes the distance of each sample from the origin within the model space. **Figure 5B** is a zoom of **Figure 5A** close to the confidence interval area. Black circles represent the calibration samples used to build the model, as they can efficiently represent optimal operative conditions according to plant experts, whereas red diamonds indicate the validation samples projected on the model. The calibration samples are almost all inside the 95% confidence interval, with some isolated exceptions of samples falling just outside the interval. Since neither consecutive set of calibration samples outside the confidence interval nor samples falling too far away from it were present, these isolated samples were kept in the model.

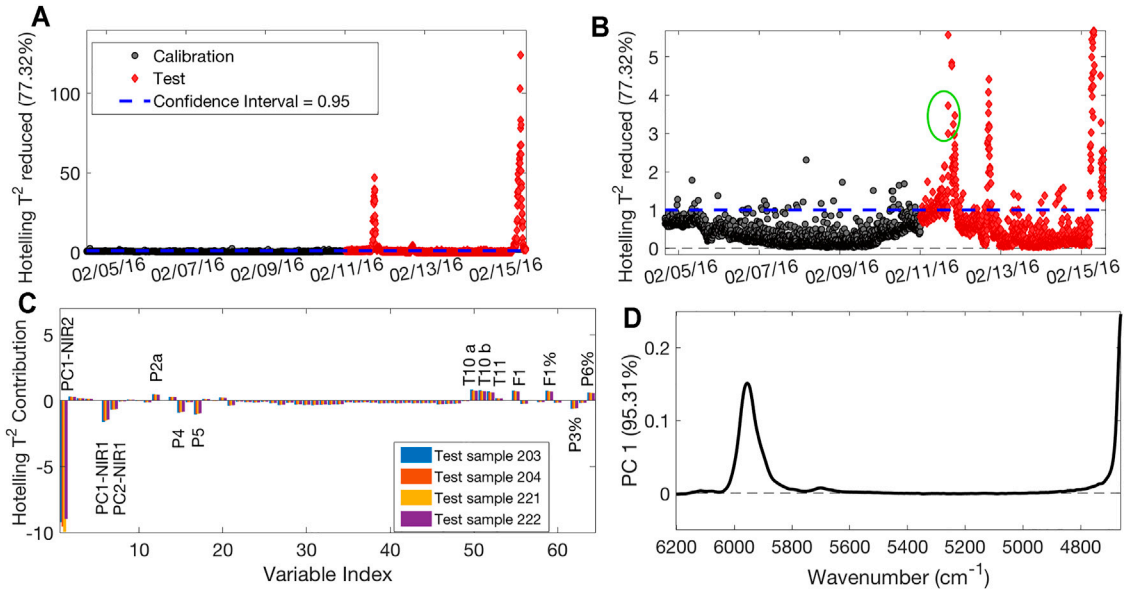

FIGURE 5 | T^2-based MSPC chart **(A)**; zoom on the confidence limit area **(B)**; contribution plot of the four circled samples of **Panel B (C)**; loadings plot on PC1 of PCA performed on NIR2 dataset **(D)**.

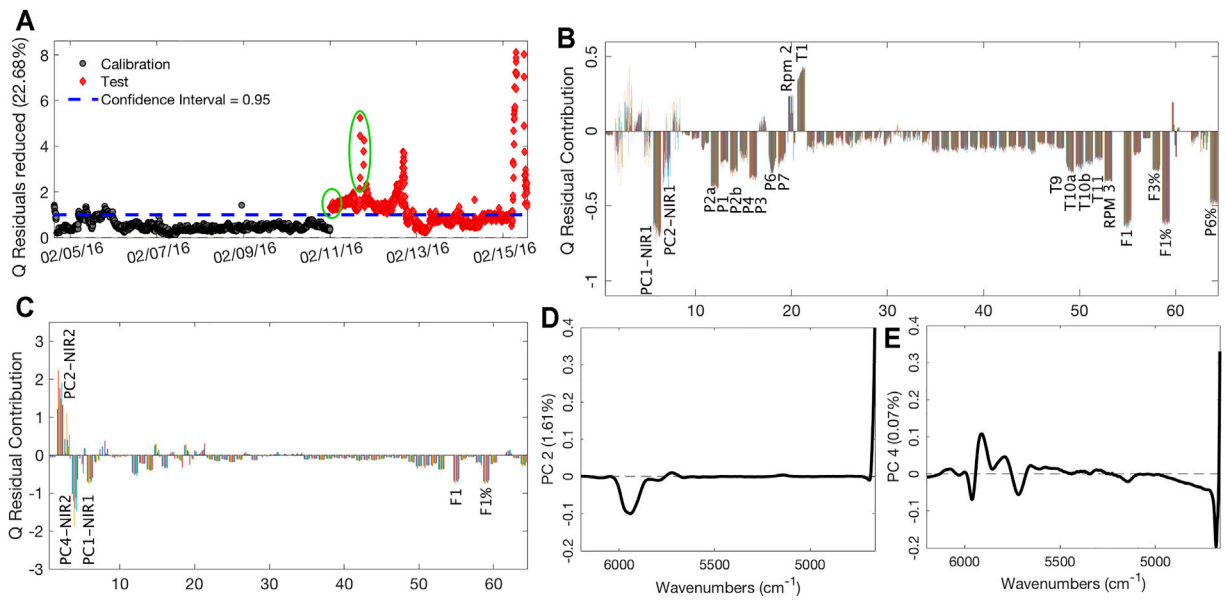

FIGURE 6 | Q-based MSPC chart **(A)**; contribution plot of the first group of circled samples of **Panel A (B)**; contribution plot of the second group of circled samples of **Panel A (C)**; loadings plots on PC2 **(D)** and PC4 **(E)** of PCA performed on NIR2 dataset.

There are three different clusters of validation samples that are outside the confidence interval: a first group corresponds to time observations taken 20 h before the production stop, and a second group corresponds to time observations taken 4 h after it, as described by the first PC of NIR2 PCA. The third group is observed at the end of the monitored time. The T^2 contribution plot of the samples of the first two clusters reveals that the PC1 scores linked to the second NIR probe are the variable mostly responsible for this behavior. As an example, the contribution plot of the four circled samples in **Figure 5B** is shown in **Figure 5C**. Since the loadings related to this PC can be ascribed to the SAN band at 5,900 cm^{-1} (**Figure 5D**), it follows that the anomalous samples present a lower polymer conversion. However, also PC1 scores related to

TABLE 1 | Results of PLS regression.

X block	LVs	Analysis	Calibration		Cross-validation		Prediction	
			R^2c	RMSEC	R^2cv	RMSECV	R^2p	RMSEP
NIR1	5	**MFI (g)**	0.94	1.27	0.86	1.99	0.89	1.6
		%AN	0.88	0.31	0.83	0.37	0.82	0.36
NIR2	5	**MFI (g)**	0.94	1.23	0.84	2.12	0.86	1.92
		%AN	0.87	0.32	0.81	0.39	0.75	0.45
PS	5	**MFI (g)**	0.97	0.83	0.93	1.4	0.92	1.4
		%AN	0.88	0.3	0.8	0.39	0.76	0.42
NPS	3	**MFI (g)**	0.96	1.05	0.95	1.14	0.96	1.2
		%AN	0.95	0.18	0.94	0.21	0.92	0.25

the first NIR probe were found relevant to explain this difference, proving that a probe that collects spectra of an intermediate product is useful, as it could provide information on possible faults well in advance with respect to the second one. Furthermore, it is possible to detect the process sensors linked to the sample's abnormality which can suggest possible reasons for the deviations. In this case, among others, sensors T10a, T10b, and T11 in the zone between R2 and CZ registered values higher than the ones registered for calibration samples, indicating a possible problem in that specific zone.

The zoom on the confidence limit area of the Q residuals MSPC chart, which describes the distance of each sample from the model space, accordingly providing information on the samples not described properly by the model, is reported in **Figure 6A**. It is observed that the changes in process settings, performed on February 11, caused the samples to initially fall outside the confidence interval. The contribution plot linked to these initial samples, reported in **Figure 6B**, shows that the PC1 scores related to the first NIR probe are the variable that mainly causes the difference with the calibration samples. In this case, it is clear and visually immediate that the process sensors concurring to explain this difference are many, suggesting plant operators to take action. The presence of the cluster of samples that present a very high Q values (**Figure 6C**), occurring just before the production pause, confirms the plant production problem, as highlighted by PC2-4 scores related to the second NIR probe (**Figures 6D,E**, respectively); thus, at this time, variations in the final product also occurred. The related loadings can be ascribed to the SAN and AN bands, suggesting, also in this case, a lower conversion of the polymer and a lower presence of AN in the final product. These observations highlight that changes in the process settings first are reflected on the intermediate product, as depicted by the NIR1 probe, and later on, the final product quality started to be nonoptimal and an intervention was operated (stop/restart), if an MSPC monitoring, like the one we analyzed retrospectively, would have been in place and a much earlier warning would have been given to the plant operators.

After the stop and the restart of the production, during which the operators worked to fix the problems, it is possible to observe a last little cluster of samples with high Q values that finally drop below the confidence limits after few hours. After that, samples remain inside the confidence interval until the moment of the formulation changes, observable by the last huge cluster of samples with high Q values.

Predictive Models

PLS regression was used to create models capable of predicting in real time the selected quality parameters for the SAN polymer, i.e., MFI and %AN. In this part of the work, the data collected from February 15th to February 23rd, corresponding to a different formulation, was also used aiming at general predictive models. The results obtained by the four different PLS models computed as described in *Data analysis* are reported in **Table 1**.

For the models computed using NIR1, NIR2, and PS datasets, five latent variables (LV) were selected, whereas only three LV were considered to build the PLS model with NPS dataset. Considering the first three models, it is observable how better MFI prediction was obtained considering PS dataset, providing a prediction error of 1.4 vs. 1.6 and 1.92 g obtained using data from the first and the second NIR probes, respectively. On the other hand, %AN is slightly better predicted using the NIR1 dataset (RMSEP = 0.36%, R^2p = 0.82) rather than the other two. However, further considering the model computed using the NPS dataset, which contains both NIR and process sensors data, it is clear how it presents the best predictions for both MFI and %AN. Both internal and external validation errors, i.e., RSMECV and RMSEP, respectively, were lower than the corresponding values obtained using any of the individual datasets, whereas the related R^2 values are higher. In detail, MFI was predicted with an error of prediction equal to 1.2 g, with an R^2p = 0.96, a better prediction accuracy compared to the one obtained using the process sensors data only. This result suggests that the information NIR probes provide is important for the prediction of this quality parameter, even if the data block most significant is the one related to the process sensors. Regarding %AN, the obtained model provided an RMSEP of 0.25% and a R^2p equal to 0.92, significantly better than prediction errors and determination coefficients obtained with the other models.

CONCLUSION

The current work demonstrated that the mid-level data fusion strategy, performed on the SAN polymer production process, using both NIR spectra and process sensors data, improved the quality of process control as well as the prediction ability of PLS

regression models. In fact, the extraction of the features from PCA models performed on NIR data allowed to add a different and valuable kind of information to the one provided by process sensor data. T^2- and Q-based MSPC charts computed with the NPS dataset were able to correctly detect the moments in which the process deviates from the normal operative conditions, providing at the same time information on which the sensors and/or the spectral features are linked to the problem. Furthermore, better PLS prediction of MFI and %AN parameters were obtained, in terms of RMSEP and R^2p, using the NPS dataset rather than the ones obtained using single blocks of data.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of confidential agreement restrictions with the company. Requests to access the datasets should be directed to Erik.Mantovani@versalis.eni.com.

AUTHOR CONTRIBUTIONS

Conceptualization, EM, FB, FM, and MC; methodology, LS, EM, FB, FM, and MC; software, LS; validation, EM, FB, FM, and MC; investigation, LS, EM, FB, FM, and MC; resources, EM and FB; data curation, LS, EM, and FB; writing—original draft preparation, LS; writing—review and editing, LS, EM, FB, FM, and MC; supervision, MC and FM; project administration, EM and FB. All authors have read and agreed to the published version of the article.

ACKNOWLEDGMENTS

The authors acknowledge Angelo Ferrando of Versalis (ENI) Company for supplying data used for the current study and fruitful discussion of the results. LS acknowledges Emilia Romagna region for funding his grant under POR FSE project "Data analytics per la REAlizzazione di sistemi predittivi e Monitoraggio real TIME di processi produttivi in industria 4.0 (DREAMTIME)" PA n° 2019-13551/RER.

REFERENCES

Bonacini, F., Ferrando, A., Mantovani, E., Sappino, C., Arcidiacono, G., Ardizzone, D., et al. (2013). Fourier Transform Near Infrared Application for Advanced Process Control of an Ethylene Cracking Plant. *NIR news* 24 (6), 9–11. doi:10.1255/nirn.1387

Chaudhry, S. S., and Higbie, J. R. (1989). Practical Implementation of Statistical Process Control in a Chemicals Industry. *Int. J. Qual. Reliability Mgmt* 6 (5), 37–48. doi:10.1108/02656718910134322

de Oliveira, R. R., Avila, C., Bourne, R., Muller, F., and de Juan, A. (2020). Data Fusion Strategies to Combine Sensor and Multivariate Model Outputs for Multivariate Statistical Process Control. *Anal. Bioanal. Chem.* 412 (9), 2151–2163. doi:10.1007/s00216-020-02404-2

de Oliveira, R. R., Pedroza, R. H. P., Sousa, A. O., Lima, K. M. G., and de Juan, A. (2017). Process Modeling and Control Applied to Real-Time Monitoring of Distillation Processes by Near-Infrared Spectroscopy. *Analytica Chim. Acta* 985, 41–53. doi:10.1016/j.aca.2017.07.038

Ferrer, A. (2007). Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. *Qual. Eng.* 19 (4), 311–325. doi:10.1080/08982110701621304

Ferrer-Riquelme, A. J. (2009). "Statistical Control of Measures and Processes," in *"Statistical Control of Measures and Processes" in Comprehensive Chemometrics*. Editors S. D. Brown, R. Tauler, and B. Walzak (Amsterdam, Netherlands: Elsevier), 97–126. doi:10.1016/b978-044452701-1.00096-x

Kourti, T. (2005). Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry. *Int. J. Adapt. Control. Signal. Process.* 19 (4), 213–246. doi:10.1002/acs.859

Kourti, T., and MacGregor, J. F. (1995). Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods. *Chemometrics Intell. Lab. Syst.* 28 (1), 3–21. doi:10.1016/0169-7439(95)80036-9

Kourti, T. (2009). Quality by Design in the Pharmaceutical Industry: Process Modelling, Monitoring and Control Using Latent Variable Methods. *IFAC Proc. Volumes* 42 (11), 36–41. doi:10.3182/20090712-4-TR-2008.00007

Kourti, T. (2006). The Process Analytical Technology Initiative and Multivariate Process Analysis, Monitoring and Control. *Anal. Bioanal. Chem.* 384 (5), 1043–1048. doi:10.1007/s00216-006-0303-y

Macho, S., and Larrechi, M. S. (2002). Near-infrared Spectroscopy and Multivariate Calibration for the Quantitative Determination of Certain Properties in the Petrochemical Industry. *Trac Trends Anal. Chem.* 21 (12), 799–806. doi:10.1016/S0165-9936(02)01202-5

Nomikos, P., and MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37 (1), 41–59. doi:10.2307/126915210.1080/00401706.1995.10485888

Shenoy, A. V., and Saini, D. R. (1986). Melt Flow index: More Than Just a Quality Control Rheological Parameter. Part I. *Adv. Polym. Technol.* 6 (1), 1–58. doi:10.1002/adv.1986.060060101

Skagerberg, B., MacGregor, J. F., and Kiparissides, C. (1992). Multivariate Data Analysis Applied to Low-Density Polyethylene Reactors. *Chemom. Intell. Lab. Syst.* 14 (1-3), 341–356. doi:10.1016/0169-7439(92)80117-M

Takeuchi, T., Tsuge, S., and Sugimura, Y. (1968). Near-infrared Spectrophotometric Analysis of Styrene-Acrylonitrile Copolymer. *J. Polym. Sci. A-1 Polym. Chem.* 6 (12), 3415–3417. doi:10.1002/pol.1968.150061220

Westerhuis, J. A., Gurden, S. P., and Smilde, A. K. (2000). Generalized Contribution Plots in Multivariate Statistical Process Monitoring. *Chemometrics Intell. Lab. Syst.* 51 (1), 95–114. doi:10.1016/s0169-7439(00)00062-9

Conflict of Interest: Authors FB and EM are employed by Versalis (ENI) SpA.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Analytical Science

ORIGINAL RESEARCH
published: 08 November 2021
doi: 10.3389/frans.2021.729732

Establishing Multivariate Specification Regions for Incoming Raw Materials Using Projection to Latent Structure Models: Comparison Between Direct Mapping and Model Inversion

Adéline Paris[1], *Carl Duchesne*[1]* *and Éric Poulin*[2]

[1]Department of Chemical Engineering, Université Laval, Québec, QC, Canada, [2]Department of Electrical and Computer Engineering, Université Laval, Québec, QC, Canada

Increasing raw material variability is challenging for many industries since it adversely impacts final product quality. Establishing multivariate specification regions for selecting incoming lot of raw materials is a key solution to mitigate this issue. Two data-driven approaches emerge from the literature for defining these specifications in the latent space of Projection to Latent Structure (PLS) models. The first is based on a direct mapping of good quality final product and associated lots of raw materials in the latent space, followed by selection of boundaries that minimize or best balance type I and II errors. The second rather defines specification regions by inverting the PLS model for each point lying on final product acceptance limits. The objective of this paper is to compare both methods to determine their advantages and drawbacks, and to assess their classification performance in presence of different levels of correlation between the quality attributes. The comparative analysis is performed using simulated raw materials and product quality data generated under multiple scenarios where product quality attributes have different degrees of collinearity. First, a simple case is proposed using one quality attribute to illustrate the methods. Then, the impact of collinearity is studied. It is shown that in most cases, correlation between the quality variable does not seem to influence classification performance except when the variables are highly correlated. A summary of the main advantages and disadvantages of both approaches is provided to guide the selection of the most appropriate approach for establishing multivariate specification regions for a given application.

Keywords: multivariate specifications, direct mapping, PLS-model inversion, projection to latent structures, quality control

OPEN ACCESS

Edited by:
Alessandra Biancolillo,
University of L'Aquila, Italy

Reviewed by:
Marco Reis,
University of Coimbra, Portugal
Martina Foschi,
University of L'Aquila, Italy

***Correspondence:**
Carl Duchesne
carl.duchesne@gch.ulaval.ca

Specialty section:
This article was submitted to Chemometrics, a section of the journal Frontiers in Analytical Science

Received: *23 July 2021*
Accepted: *15 October 2021*
Published: *08 November 2021*

Citation:
Paris A, Duchesne C and Poulin É (2021) Establishing Multivariate Specification Regions for Incoming Raw Materials Using Projection to Latent Structure Models: Comparison Between Direct Mapping and Model Inversion. Front. Anal. Sci. 1:729732. doi: 10.3389/frans.2021.729732

1 INTRODUCTION

For many manufacturing industries, reaching market standards in terms of product quality is a priority to ensure sales. Product quality is influenced by different factors, but one of the most important is the variability in raw material properties. If no corrective action is applied, these fluctuations propagate directly to final product quality. This is a real problem for many industries especially those processing bio-based materials using raw materials extracted from natural resources.

Ensuring good quality control may attenuate the impact of raw material variability. This can be performed in three ways: defining specifications for raw material properties, choosing adequate operating conditions, and characterizing final products for quality (Amsbary, 2013). A particular attention should be paid to the first as it deals directly with the source of the problem. Defining specifications and acceptance criteria for incoming lots of raw materials is key to achieve high and consistent quality final product. This is a useful tool to determine whether a lot of raw materials is processable, and indicates the risk of not reaching desired quality.

The main approach commonly used in the industry is to determine the acceptability of lots of raw materials based on a set of univariate specifications, past experiments, and/or the properties of the best suppliers (Duchesne and MacGregor, 2004). As the properties of any material are often highly correlated, univariate limits may lead to misclassification (De Smet, 1993; Duchesne and MacGregor, 2004). If the multiple univariate specifications are set large enough to accept all past good lots of raw materials, the risk of accepting bad quality lots increases. To mitigate this, univariate specification limits can be tightened to minimize acceptance of poor quality raw materials. However, this increases the rejection rate of good lots of materials, which typically leads to higher purchasing costs. Thus, the correlation structure between the raw material properties needs to be considered to minimize the risk of inadequate decisions. Establishing multivariate specification regions to select incoming lots of raw materials is a solution to this problem. The concept was first introduced by De Smet (1993). It consists of building a Projection to Latent Structures model first to relate the raw material properties to the final quality attributes. Then, each lot of raw materials is projected in the latent space of the PLS model. Its class assignment (e.g., good or bad quality) is inherited from the corresponding final product quality assessment, hence the name Direct Mapping (DM) approach. Finally, a boundary is established to discriminate the two classes by balancing type I and II errors or by minimizing one. The resulting region is then used to decide whether a new incoming lot of raw materials should be accepted or rejected.

As the impact of process control actions, changes in process operating conditions and disturbances on final product quality were not considered by De Smet (1993), Duchesne and MacGregor (2004) extended the previous approach. They proposed a framework for different scenarios based on how process variability affects final product quality, and its level of collinearity with raw material properties. The methods are illustrated using simulated and industrial data from a film blowing process (Duchesne and MacGregor, 2004). Tessier and Tarcy (2010) have also applied the technique in the context of the aluminum production.

Further improvements were then proposed. To increase the size of the dataset and to include more variations in the context of pharmaceutical process scale-up, García-Muñoz (2009) introduced a new step prior to the Duchesne and MacGregor technique to take into account data collected from multiple scales. Later, Azari et al. (2015) suggested using the Sequential Multi-Block PLS algorithm (SMB-PLS) instead of PLS as a more efficient method to establish multivariate specifications when raw material properties and process operating conditions are correlated. This approach allows to clearly identify the variation in raw material properties uncompensated by control actions. Finally, to establish specifications in situations where several different types of raw materials are used, MacGregor et al. (2016) have proposed a new approach based on Monte Carlo simulations to calculate the risk of accepting a new lot.

A similar concept to multivariate specifications called Design Space (DS) was introduced by the Internal Conference of Harmonization (2009) mainly for the pharmaceutical industry. The goal is to determine: "the multidimensional combination and interaction of input variables (e.g., material attributes and process parameters) that have been demonstrated to provided assurance of quality." Essentially, the general objective of establishing a design space is to reduce product quality variability by design rather than by inspection techniques aiming at characterizing final product properties (MacGregor and Bruwer, 2008; Godoy et al., 2017). One main advantage of this approach is that modifications applied to the process or raw material variability within the DS are not considered as a change for the regulatory agencies as Food and Drug Administration (FDA) (ICH, 2009; Lawrence et al., 2014).

Even if the two concepts (raw material specifications and DS) aim at improving product quality control, differences exist between them. The DS is typically defined during the product development stage using raw material properties and process conditions simultaneously. Multivariate specifications, however, are built using larger sets of industrial historical data, and require that variability introduced by process variables be removed prior to defining the specification region. In addition, even if both concepts are based on PLS models, they use different mathematical approaches to determine the acceptance region. Defining a DS in latent space is mostly performed using PLS model inversion of a single desired quality attribute (Facco et al., 2015; Bano et al., 2017; Palací-López et al., 2019) while, in the past, multivariate specification regions were obtained using direct mapping of final product quality based on several correlated attributes. As suggest by Garcia-Muñoz et al. (2010), the inversion technique could be an alternative to DM for developing raw material multivariate specifications. Applying PLS model inversion using multivariate product quality attributes was demonstrated by Jaeckle and MacGregor (1998) and Jaeckle and MacGregor (2000) in the context of product development problems.

The objective of this paper is to compare the two approaches for establishing multivariate specification regions, namely PLS model inversion and direct mapping, in terms of classification performance for a given application, and to determine their advantages and drawbacks. It also shows how to establish multivariate specification regions by PLS inversion for a multivariate set of quality attributes, and assess the influence of different levels of correlation between them for both techniques. Such a comparison for one or multiple quality attributes has not been attempted in the past, to the best knowledge of the authors. The proposed paper should be considered as a guide to support the development of

multivariate specifications using the most appropriate technique for a given application.

This work is quite ambitious since many scenarios need to be considered and several decisions had to be made to ensure a fair comparison. First, simulated data is used to allow multiple scenarios to be generated. A simple model involving four raw material properties and two final product quality attributes was developed to facilitate the comparisons and interpretations. The shape of the final product quality acceptance region was selected to be elliptical to reflect the correlation structure between the quality attributes. When building the PLS models between final product quality attributes and raw material properties, the number of components retained in both approaches is chosen as that maximizing classification performance for the PLS inversion approach. This choice was made to avoid introducing biases in the comparison since the direct mapping approach has more flexibility. For each combination of final product quality attributes, a single PLS model is built and used to define the specification regions with both approaches. Finally, the classification performance is assessed without considering the uncertainty back propagation (Bano et al., 2017).

The paper is organized as follows. First, the simulator used to generate the datasets is presented. Then, the proposed methodology is exposed. The section includes a brief description of PLS regression, how to establish multivariate specifications using direct mapping and PLS inversion, as well as the classification metrics used to calculate classification performance. The results are then presented and discussed. Thereafter, the main conclusions are drawn.

2 DATASET GENERATION

Within the scope of the study, to simplify the comparison between the two techniques, multivariate specifications are developed under the hypothesis that process variables do not influence the quality of the product (i.e., the process is under control). However, how to cope with process variations in establishing multivariate specifications and design spaces was already extensively studied (Duchesne and MacGregor, 2004; Azari et al., 2015; Facco et al., 2015; MacGregor et al., 2016). The comparative analysis proposed in this study is generic, and is applicable in scenarios where process variations significantly affect product quality. Hence, in this study, only two blocks of data $\mathbf{X}$ ($N\times M$) and $\mathbf{Y}$ ($N\times K$) are involved when building PLS models. The first contains M raw material properties characterized in the laboratory or on-line using spectroscopy techniques, for instance, and the second K quality attributes of the final product collected for N observations or lots of raw materials. The data contained in these matrices are generated by simulations using analytical equations as described in the following subsection to facilitate the generation of combinations of y-variables spanning the full range of correlation. In addition, for the N observations included in the dataset, the quality of the final product is assigned to a class using a binary variable (i.e., good/bad quality) which is used to assess classification performance. The methods used to establish multivariate specification regions are then presented.

TABLE 1 | Noise percentage and nominal signal values.

	x_1	x_2	x_3	x_4	y_1	y_2
ε [%]	1	0.5	2	3	1	0.5
$\overline{x_i}$	22.01	8.13	12.18	11.99	N/A	N/A

2.1 Simulated Process

The X-dataset is inspired from the model proposed by De Smet (1993). A total of four equations are used to generate variations in raw material properties:

$$x_1 = 22 + h_1 \tag{1}$$

$$x_2 = \sqrt{0.1 + 2h_2 + 3x_1} \tag{2}$$

$$x_3 = 1.5 + 0.3x_1 + 0.5x_2 + h_3 \tag{3}$$

$$x_4 = 12 + 0.5h_4 \tag{4}$$

where h_i are random numbers following a standard normal distribution $N(0,1)$. Correlation exists between properties 1–3 while the fourth is independent of the others.

For each lot of raw materials, i.e. an observation in $\mathbf{X}$, two quality attributes are calculated using the following equation:

$$y_j = \sum_{i=1}^{4} k_{i,j} g_{i,j} x_i \tag{5}$$

and the values are stored in the $\mathbf{Y}$ matrix. The binary variables $k_{i,j}$ determine if the i^{th} raw material property affects the j^{th} quality attribute while $g_{i,j}$ consists of random integers between −5 and 5. These were used to generate different magnitude for the effect of each x-variable on the y-variables. As the objective of this article is to compare the performance of two approaches for defining specification regions under different levels of correlation between both y-variables, the same $\mathbf{X}$ dataset is used throughout the analysis to generate different combinations of y-variables by changing parameters $k_{i,j}$ and $g_{i,j}$. When the product of these two parameters results in similar values for both y-variables, a high level of correlation is obtained. Conversely, very different values for this product leads to a low correlation. The span of different levels of correlations is owing to the random values generated for $g_{i,j}$. It should be noted that each combination is obtained randomly and not by smoothly increasing the correlation level between the y-variables.

Noise is added to all variables. The measured values $y_{m,j}$ and $x_{m,i}$ are obtained using the following equations:

$$y_{m,j} = y_j + \left(\varepsilon_{y,j}\overline{y}_j\right)e_{y,j} \tag{6}$$

$$x_{m,i} = x_i + \left(\varepsilon_{x,i}\overline{x}_i\right)e_{x,i} \tag{7}$$

where $e_{y,j}$ and $e_{x,i}$ represent the errors added to the y- and x-data. These random errors also follow a standard normal distribution $N(0,1)$. Their magnitude is characterized by the error standard deviation set as a percentage ε of the mean $\overline{x}$ or $\overline{y}$ for each variable. In all simulations, noise was generated in the same way. The values of ε and $\overline{x}$ are presented in **Table 1** while $\overline{y}$ are not shown since they vary from one dataset to another. The mean values are obtained using the calibration dataset which contained 500 observations.

In addition to the calibration set, two other datasets are generated. The first is used to determine the number of PLS components needed while the classification performance of the specification regions is assessed using the second. Each of these datasets contains 10,000 observations. This number was selected in such a way that stable classification performance for each metric is obtained. Note that a large number of data points were generated as a mean to compare the direct mapping and inversion methods using a fair and sound statistical approach. However, both methods have already been demonstrated as effective on smaller datasets collected on simulated and industrial processes (Duchesne and MacGregor, 2004; Facco et al., 2015).

2.2 Definition of Product Acceptance

Establishing multivariate specification regions using a data-driven approach begins with identifying past lots of products of good and poor quality. This involves a product acceptance region in the Y-space. As the data used in this work are obtained from simulations, an indicator associated with the final quality of the product needs to be defined to identify good and bad products. The acceptance limit used in this study has an elliptical shape:

$$(\mathbf{y}-\overline{\mathbf{y}})\mathbf{\Sigma}_{\mathbf{y}}(\mathbf{y}-\overline{\mathbf{y}})' \leq \zeta \tag{8}$$

where $\overline{\mathbf{y}}$ $(1 \times K)$ is the vector containing the means of each y-variable, and $\Sigma_y (K \times K)$ is the y-covariance matrix. Parameter ζ is adjusted to specify the size of the region and to control the proportion of data assigned to good and bad quality. Once this parameter is selected, a binary variable was used to assign each observation to good and bad classes. In this work, ζ was chosen to ensure a proportion of good/bad products of 4:1. Even if the ratio of bad product is quite high compared to what is usually observed in industry, this choice was made to reduce the impact of class imbalance. There is no specific rule stating that a dataset should not be used as it is too imbalanced. However, in practice, ratios ranging from 2:1 to 10:1 are considered to be between marginally and modestly imbalanced (Weiss, 2013). Therefore, a choice was made to find a compromise between a realistic situation and balanced classes. When using industrial data, the ratio should be adjusted to obtain a more balanced dataset by oversampling the smallest class or under-sampling the most populated one (He and Garcia, 2009).

3 METHODS

This section presents the direct mapping and PLS inversion-based approaches used to define multivariate specifications regions. As both techniques are based on PLS regression, a brief overview of this latent variable method is provided. Finally, the classification metrics used to quantify the performance are described.

3.1 Projection to Latent Structure Regression

Before building PLS models between **X** and **Y**, the data are mean-centered and scaled to unit variance. As the **X** and **Y** matrices contained collinear data, latent variable modelling techniques are suitable approaches. PLS regression is retained as it builds the best linear relationships between the **X** and **Y** while modelling the variability contained in both spaces.

Variability is extracted using a group of A orthogonal latent variables known as scores **T** $(N \times A)$. PLS regression is defined mathematically by the following set of equations:

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} = \mathbf{T}\mathbf{P}' + \mathbf{E} \tag{9}$$

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{F} = \mathbf{T}\mathbf{C}' + \mathbf{F} \tag{10}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} = \mathbf{X}\mathbf{W}\left(\mathbf{P}'\mathbf{W}\right)^{-1} \tag{11}$$

where **E** $(N \times M)$ and **F** $(N \times M)$ are the model residuals. **C** $(K \times A)$ and **P** $(M \times A)$ are the loadings of the **Y** and **X** spaces, respectively. The loadings and the score values are computed using the NIPALS algorithm (Wold et al., 2001). It also provides the weight matrices **W** $(M \times A)$ and $\mathbf{W}^{*}$ $(M \times A)$ allowing to make predictions of **Y** based on **X**.

Prior applying PLS to new X-data, it is important to ensure that they are consistent with historical data used to build the model. This is achieved by computing the squared prediction error SPEX and the Hotelling's T^2, and verifying that they fall below their respective statistical limits. The SPEX is used to check consistency of the correlation structure of new data. It is defined as follows:

$$\mathrm{SPEX_i} = \mathbf{e_i}\mathbf{e_i'} \tag{12}$$

where $\mathbf{e}_i$ $(1 \times K)$ is the X-residual vector for the i^{th} observation:

$$\mathbf{e}_i = \mathbf{x}_i - \mathbf{t_i}\mathbf{P} \tag{13}$$

As the SPEX values follow approximately a χ^2 distribution with $\frac{2m^2}{\nu}$ degrees of freedom (Nomikos and MacGregor, 1995), a (1- α) upper control limit (UCL) can be obtained:

$$\mathrm{SPE_{UCL}} = \frac{\nu}{2m}\chi^2_{\frac{2m^2}{\nu},\ \alpha} \tag{14}$$

where ν and m are respectively the variance and the SPE mean calculated during the model calibration.

The Hotelling's T^2 is used to measure the distance of projected new observations from the origin of the latent variable space. It is typically used to confirm whether a new observation falls within the so-called knowledge space (KS). The KS represents the space spanned by historical data in the latent variable space of the PLS model. The T^2 value for the i^{th} observation is obtained as follows:

$$\mathrm{T_i^2} = \sum_{a=1}^{A}\left(\frac{t_{a,i}}{s_a}\right)^2 \tag{15}$$

where $t_{a,i}$ is the score values obtained for the a^{th} component and s_a its standard deviation calculated in calibration.

The T^2 values are known to follow a Fischer distribution approximately (Jackson and Edward, 1991). A (1-α) upper control limits as proposed by Weirda (Wierda, 1994) can be calculated using the number of points in the calibration dataset N and the number of components retained A using the following equation:

$$T^2_{UCL} = \frac{A(N^2-1)}{N(N-A)} F_{A,N-A,\alpha} \tag{16}$$

where $F_{A,N-A,\alpha}$ is the value of the Fischer distribution for A and (N-A) degrees of freedom. This elliptical-shaped limit is typically drawn in the scores space. The length of each axis r_a is equal to:

$$r_a = \sqrt{\frac{A(N^2-1)}{N(N-A)} s_a^2 F_{A,N-A,\alpha}} \tag{17}$$

which is deduced from **Eq. 15** and **Eq. 16**.

One important step in the model development is to select the optimal number of components. The appropriate method depends on how the model will be used. If the objective is to build PLS models for making predictions, criteria such as cumulative predicted variance Q^2Y or the root mean squared errors of prediction (RMSEP) in cross-validation or calculated on an external dataset should be used. For a classification problem, such as defining multivariate specification regions, the optimal number of components should be the one that maximizes the classification performance on an external dataset. Classification performance is obtained by using the accuracy as defined in a following section.

The same PLS model is used to establish multivariate specification regions using both DM and inversion techniques. The number of components maximizing classification performance may be different for both approaches, but a single value of A needs to be selected for the comparative study. As the direct mapping is based on a compromise between type I and type II errors, which is an additional degree of freedom compared to inversion, using direct mapping might introduce a bias when choosing the number of components. To overcome this issue, the number of components is determined by maximizing classification performance obtained with the inversion approach, and this number of components is also used for DM.

3.2 Direct Mapping Approach

Defining multivariate specifications using direct mapping is performed in two steps. First, a PLS model is built using the quantitative y-data. Second, the specification limit in the latent space is defined by mapping product quality in the scores space. In other words, the class assigned to the score values (i.e., good/bad) corresponds to that of the final product obtained for the same lot of raw materials. The goal is to define a region that allows the separation of the two classes. Note that the quality classes are only used to assess classification performance in the latent space and not for building discriminant PLS models (i.e., PLS-DA). The shape of the region is defined by the user. In this study, a similar shape as that of the product quality acceptance region is chosen for both methods. Since the limit in the Y-space is elliptical and the PLS model is linear, the region obtained in the score space by inversion is also elliptical. For this reason, the following elliptical-shaped specification region was selected for the DM approach:

$$\mathbf{t}\boldsymbol{\Lambda}\mathbf{t}' \leq \eta \tag{18}$$

where $\boldsymbol{\Lambda}$ $(A \times A)$ is the score covariance matrix. The value of η is used to adjust the size of the elliptical region. The strategy used to select η depends on the context in which the specification will be used, and the consequence of each type of misclassification. One may prefer minimizing type I or type II error while another could seek a compromise between both. By definition, type I error represents a sample predicted as bad quality when it is good while a type II error is a sample of truly bad quality predicted as good. For this work, as there is no specific context or limitation, the value of η is chosen to be the one leading to the same percentage of type I and type II errors.

Prior to using the specified region for incoming new lots of raw materials, the correlation structure of each observation needs to be assessed to ensure the model validity for this lot. This is done by defining an upper control limit on SPEX during the PLS model calibration as discussed in the previous section. If a given lot violates the limit, it should be flagged as having an inconsistent correlation structure compared with historical data, and should be rejected unless it is desired to process it, and used it to update the model and/or improve the specification region definition.

3.3 Projection to Latent Structure Model Inversion

Alternatively, multivariate specification regions in the score space can be established by inverting the PLS model for each point lying on the final product quality acceptance limit. In other words, instead of adjusting a limit within the score space using product quality class assignments, the limit is propagated from the Y-space acceptance region using the model structure.

As the limit in the Y-space is elliptical in this study, its parametric equation is used to generate combinations of quality attributes (y_1, y_2) lying on the ellipse to use for the inversion. The transformation for the matrix to the parametric equation is the following:

$$[\mathbf{y}_1, \mathbf{y}_2] = \mathbf{V}\sqrt{\mathbf{D}}[\cos(\boldsymbol{\theta}),\ \sin(\boldsymbol{\theta})] \tag{19}$$

where $\mathbf{D}$ $(K \times K)$ is a diagonal matrix containing the eigenvalues of $\zeta\boldsymbol{\Sigma}_y$ and $\mathbf{V}$ $(K \times K)$ the corresponding eigenvectors while $\boldsymbol{\theta}$ contains value between 0 and 2π.

For each combination (y_1, y_2) which is named $\mathbf{y}_{des}$ $(1 \times K)$, the PLS inversion method proposed by Jaeckle and MacGregor (1998) and Jaeckle and MacGregor (2000) allows calculating the corresponding score vector $\mathbf{t}_{des}$ $(1 \times A)$. Computations begin with the PLS model equation for the Y-space:

$$\mathbf{y}_{des} = \mathbf{t}_{des}\mathbf{C}' \tag{20}$$

where the dimensions of the loading matrix $\mathbf{C}$ yields three possible cases depending upon the number of PLS components A and the number of y-variables K, as described in the following subsections.

3.3.1 Case 1: $A = K$

This case is the simplest one since there is a unique solution (i.e., number of equations equal to number of unknown parameters). As $\mathbf{C}$ is a square matrix, solving for $\mathbf{t}_{des}$ from **Eq. 20** yields the following result:

$$\mathbf{t}_{des} = \mathbf{y}_{des}\left(\mathbf{C}'\mathbf{C}\right)^{-1} = \mathbf{y}_{des}\left(\mathbf{C}'\right)^{-1} \tag{21}$$

which directly provides the score vector associated with a combination of y-variables lying on the product acceptance region. The two terms are equivalent since $\mathbf{C}$ is a square matrix.

3.3.2 Case 2: $A < K$

In this case, since the number of unknown parameters is lower than the number of equations, there is no solution. As the matrix $\mathbf{C}$ is not square, to obtain $\mathbf{t}_{des}$ from **Eq. 20** a right inverse is used. The resulting equation is the following:

$$\mathbf{t}_{des} = \mathbf{y}\mathbf{C}\left(\mathbf{C}'\mathbf{C}\right)^{-1} \tag{22}$$

In fact, the solution is the result of an ordinary least squares prediction between $\mathbf{y}$ and $\mathbf{C}$ where the prediction error of $\mathbf{y}$ is minimized (Jaeckle, 1998).

3.3.3 Case 3: $A > K$

For DS estimation, this case is the one that happens the most frequently (Facco et al., 2015; Palací-López et al., 2019). Since there are more unknown parameters than equations, the number of solutions is infinite. To obtain all of the possible solutions, Jaeckle and MacGregor (1998), Jaeckle and MacGregor (2000) proposed the following approach. As $\mathbf{C}'\mathbf{C}$ is singular, solving for $\mathbf{t}_{des}$ requires using the Moore-Penrose inverse. Prior to the inversion, Jaeckle and MacGregor (2000) suggested to transform the score vector $\mathbf{t}$ into two new matrices to facilitate proper scaling:

$$\mathbf{y}_{des} = \mathbf{t}\mathbf{C}' = \mathbf{u}\mathbf{S}\mathbf{C}' \tag{23}$$

where $\mathbf{u}$ $(1 \times A)$ is an orthonormal vector and $\mathbf{S}$ $(A \times A)$ is a diagonal matrix where the diagonal values are equal to $\sqrt{\mathbf{T}'\mathbf{T}}$. Then, using the Moore-Penrose inverse for a combination of y-variables stored in $\mathbf{y}_{des}$, the predicted value $\mathbf{t}_{pred}$ is obtained:

$$\mathbf{t}_{pred} = \mathbf{y}_{des}\left(\mathbf{C}\mathbf{S}'\mathbf{S}\mathbf{C}'\right)^{-1}\mathbf{C}\mathbf{S}'\mathbf{S} \tag{24}$$

which is the solution that is the closest to the origin of the PLS model plane. The other possible solutions $\mathbf{t}_{des}$ are distributed along the null space:

$$\mathbf{t}_{des} = \mathbf{t}_{pred} + \mathbf{t}_{null} \tag{25}$$

where $\mathbf{t}_{null}$ spans an orthogonal subspace of $A - K$ dimensions. To obtain $\mathbf{t}_{null}$ values, singular value decomposition is applied on $\mathbf{S}\mathbf{C}'$ to extract the left singular vectors. Only the $(A\text{-}K)$ vectors associated with null singular values are kept in matrix $\mathbf{G_2}$ $(A \times (A - K))$. The $\mathbf{t}_{null}$ vector is then calculated as follows:

$$\mathbf{t}_{null} = \boldsymbol{\lambda}\mathbf{G}_2'\mathbf{S} \tag{26}$$

by specifying a $(A\text{-}K)$ vector of constants $\boldsymbol{\lambda}(1 \times (\boldsymbol{A} - \boldsymbol{K}))$ that represents a position along the null space.

As the specification region is defined using an infinite number of equations (i.e. one for each point of the ellipse in the y-space), determining whether an observation falls within the specification limits or not is not simple. Geometrical approaches such as triangularization or visual inspection of score plots when $A < 4$ are needed to determine the position of one observation towards the region. When $A > 3$, more complex manipulations and calculations are necessary to determine the position of the scores with respect to the specification limits. Hence, in this study, it was decided to limit the number of PLS components to $A \leq 3$. Also, before projecting a new lot into the specification region, the same approach using the SPEX limit needs to be performed to ensure that the model is valid for new observations.

3.4 Classification Metrics

As the main objective of this study is to compare two methods for developing multivariate specification regions, metrics are needed to compare their classification performance. Five different metrics are considered. They are based on the elements of the confusion table, which is schematically represented in **Figure 1A**.

The figure shows the relationship between the ground truth for good (G) and bad (B) final product, and the predicted class labels $\widehat{\mathrm{G}}$ and $\widehat{\mathrm{B}}$. In summary, a true positive TP is a good product well classified while false negative FN is a good product predicted as bad. On the other hand, a bad product which is misclassified is considered a false positive FP, and a true negative TN when it is well classified. It should be noted that FN and FP correspond to type I and II errors, respectively.

The first performance metric used is accuracy (ACC), which consists of the ratio of well-classified samples over the whole population:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{27}$$

The next four metrics are shown in (**Figure 1B**) illustrated as the element of the confusion matrix. This allows a better visualization of the calculated ratios. Precision, also known as positive predictive value (PPV), is defined as:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{28}$$

which is the ratio of predicted good products to all the good observations. Recall, or true positive rate (TPR), is defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{29}$$

It is the proportion of the well classified good product. False positive rate (FPR):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{30}$$

is used to quantify the percentage of misclassified bad products. The last metric is the false omission rate (FOR) which represents the percentage of errors made in assigning bad quality products to the right class:

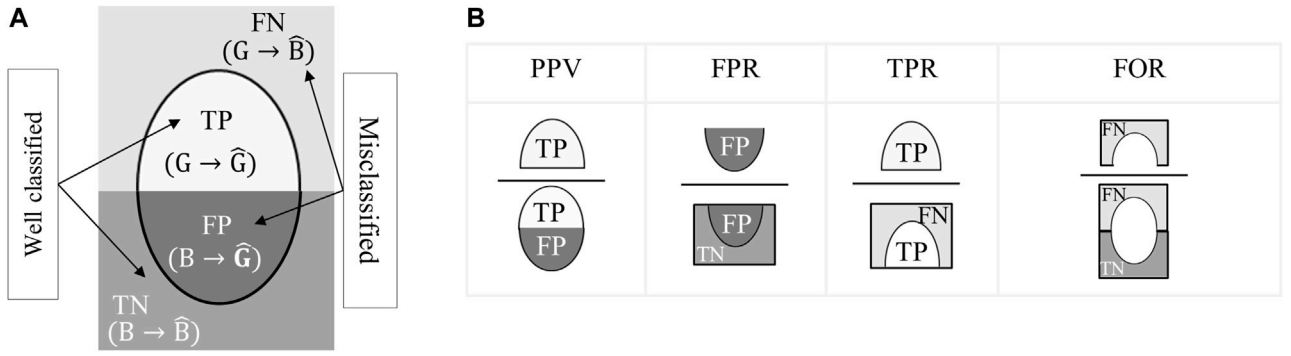

FIGURE 1 | Schematic representation of the classification metrics: **(A)** confusion matrix **(B)** classification metrics.

TABLE 2 | Summary of the parameters involved in establishing specification regions with DM and PLS inversion, as well as some performance statistics for the different scenarios investigated.

Scenario	*Y*-Space definition			Validation performance			DM constant η	
	Coefficients in eq. 5 ($k_i \times g_i$)	|r| [%]	Quality constant ζ	ACC inversion [%]	Q^2Y [%]	*A*	Value	Error type I and II [%]
1	[3, -2, -1, 1]	–	1.75	88.3	84	2	3.51	8.8
2A)	[-5, 0, 3, -3] [0, 0, 2, -1]	30	3.35	90.3	89	2	3.54	4.4
2B)	[-1, 0, -1, -4] [0, 0, 0, 2]	–66	3.25	86.5	79	2	3.34	5.8
2C)	[0, 0, 2, 1] [-1, 0, 5, 0]	95	3.25	86.2	92	2	3.23	10
4	[0–1 0 1] [3 0 0 0]	–40	3.29	88.9	87.5	3	4.78	7.8

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{TN} + \mathrm{FN}} \tag{31}$$

4 RESULTS AND DISCUSSION

The results are presented in three parts. First, a simple example considering a single quality attribute is shown to illustrate the methodologies, and to explain the main criteria used for comparing both techniques. Then, the impact of collinearity between the two quality attributes on the shape and size of the specification regions is presented. Finally, the main advantages and disadvantages of both techniques are highlighted based on the observations made during the analysis.

For ease of presentation, **Table 2** summarizes all the information used to generate the different scenarios. The table is divided in three parts. First, the columns identified as Y-space definition show the values of the simulation model parameters selected for generating the datasets. This includes the coefficients ($k_i \times g_i$) needed to define the y-variables, and the level of correlation between them, as well as the quality constant ζ that allows obtaining a 4:1 good/bad class ratio. The second part provides an overview of the PLS model performance in validation. The accuracy (ACC) obtained when inverting the model, which was used to choose the number of Latent Variables (LV) retained *A*, as well as the cumulative predicted explained variance Q^2Y are shown in the table. The last part provides the values of the DM constant adjusting the size of the specification region, and the resulting percentage of type I and II error.

4.1 Scenario 1–Illustration Using a Simple Example

The first scenario proposed is obtained by using one quality attribute. The output is simulated with all raw material properties affecting the quality attribute (i.e., $k_i \neq 0$) with a different value of g_i for each x-variable as shown in **Table 2**. Then, the product quality acceptance zone is defined. Since the Y-space is univariate, the product acceptance region consists of lower and upper bounds using **Eq. 8** where $\zeta = 1.75$.

After mean-centering and scaling the data using the calibration dataset, the PLS model is built, and the number of components is selected to maximize classification accuracy for PLS inversion. **Table 2** shows that an optimal accuracy of 88.3% is obtained using 2 components. The resulting model predicts 84% of the y-variance (Q^2Y) based on the validation set. This model is then

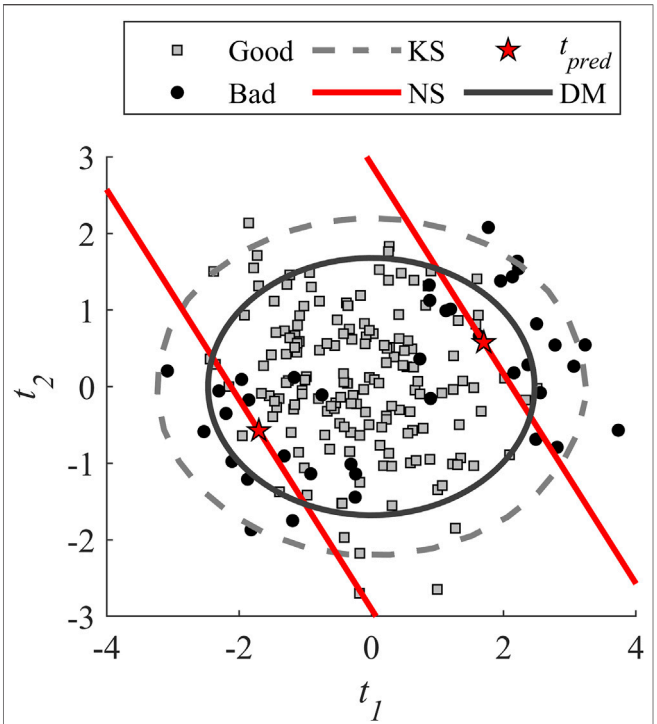

FIGURE 2 | Score plot showing the multivariate specification regions obtained using both Direct Mapping and PLS Inversion for Scenario 1.

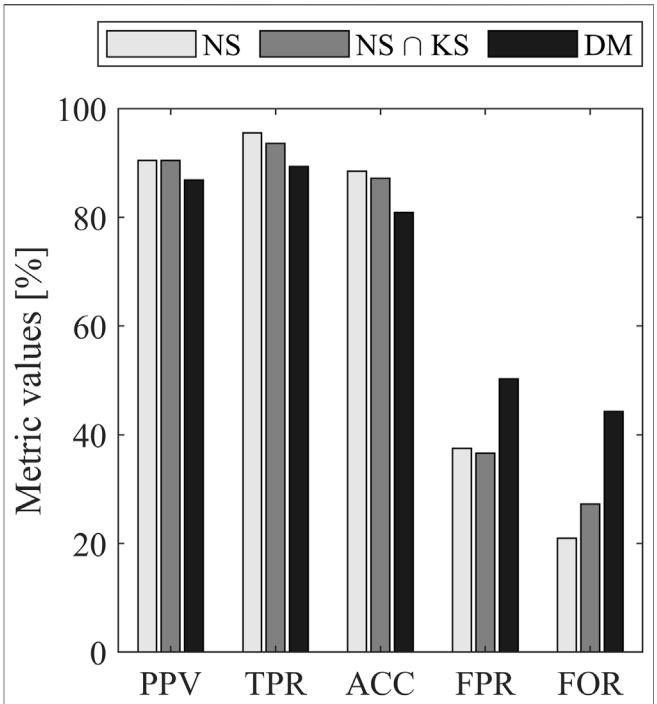

FIGURE 3 | Classification performance metrics for Scenario 1.

used to define the DM specification region by finding the value of η that gives the same percentage of type I and type II errors. The obtained value of 3.51 leads to 8.8% of both types of errors.

At this point, both specification regions are defined and drawn in the latent space. For ease of visualization, **Figure 2** shows a subsampling of the testing dataset where the proportion of each class is preserved. The solid black line represents the DM region obtained previously.

Since the number of components is higher than the number of quality attributes ($A > K$), the specification region was determined using the inversion case 3 which considers the presence of a null space. The lower and upper y-limits are inverted to obtain the corresponding t_{pred} values represented by red stars. The null-space (NS) is calculated and shown by the solid red line. Thus, all score values falling between these two lines are associated with good quality final product as per the inversion approach. However, this region is opened which may lead to misclassification as the predicted score values outside the knowledge space (KS) extrapolate. Therefore, the solution is constrained by the 95% upper Hotelling's T^2 limit as advocated in some papers (Tomba et al., 2012; Facco et al., 2015; Bano et al., 2017). The gray dash line represents the KS.

It is observed in **Figure 2** that the DM is already included inside the KS. This was expected because, the DM ellipse is designed to discriminate the classes using the calibration dataset which is the same used to define the KS. In addition, the inversion seems slightly better compared to direct mapping. Better performance might have been obtained if another shape was chosen for DM regions (i.e., the shape is an additional degree of freedom for DM).

Based on these observations, the performance in classification is analyzed using the classification metrics described in section *Classification Metrics*. Three different specification regions are considered. The first is the region obtained with the inversion alone (NS). The second is the NS region constrained by the KS (NS ∩ KS). The third is the DM region.

For all the metrics, the performance obtained with PLS inversion approaches are quite similar except for the false omission rate that is higher when considering the KS limit. Constraining the region within the KS generates more good samples predicted as bad, which increase the number of FN as shown in **Figure 2**.

When comparing direct mapping and inversion coupled with KS, **Figure 3** shows that the performance are better for inversion for all the metrics. A particular attention should be paid to FPR and FOR for the DM as the difference is higher compared to other metrics. For the FPR, it can be seen in **Figure 2** that the edge of the ellipse allows accepting more lots of bad quality which is not the case for the inversion. The higher FOR metric is caused by the bounding of the region with the KS limit.

Globally, Scenario 1 allowed to illustrate the methodology with a simple example using a univariate quality attribute. The basis is set to analyze more complex cases with multiple quality attributes. For the proposed example, the inversion is slightly better compared to direct mapping based on the five metrics. Also, the acceptance region is more restrictive for the direct mapping since its area is smaller compared to inversion. The performance might have been better if the shape of the DM

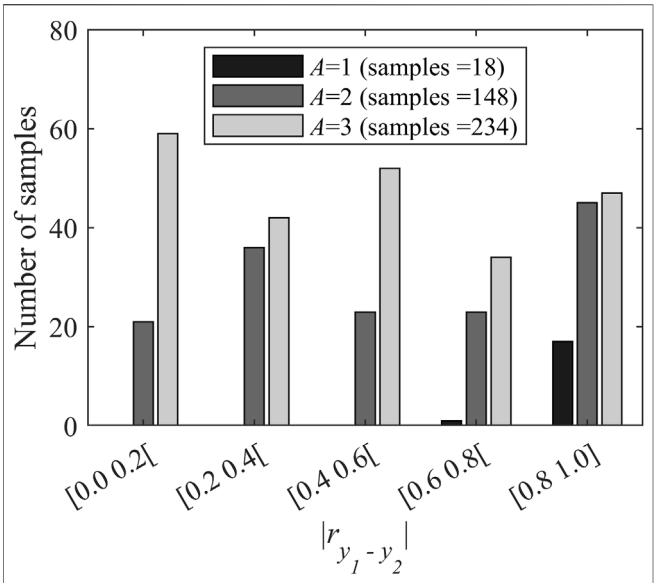

FIGURE 4 | Distribution of the combinations of y-variables and the corresponding number of PLS components as a function of level of collinearity.

regions would have been modified to exploit this additional degree of freedom.

4.2 Scenarios 2, 3 and 4: Impact of Collinearity Between Quality Attributes on the Specification Regions

The impact of collinearity between the two quality attributes is studied with respect to the three inversion cases (i.e., $A < K$, $A > K$ and $A = K$). Initially, 400 combinations of two quality attributes were generated using the simulator (**Eq. 5**). For each of them, the number of components was chosen based on maximizing classification accuracy for PLS inversion. **Figure 4** shows the number of y-combinations for the different levels of correlation, and the number of components retained when building the PLS model. Note that both negative and positive correlations were obtained, but the absolute value is shown in the figure.

As it can be observed in **Figure 4**, 58.5% of the combinations require three components and they cover the full range of correlations. The samples associated with two components also spanned the entire range. This is not the case for the datasets where a single component is selected. Less than 5% of the combinations fall in this category and they concentrate in the zone of high levels of correlation i.e., with a value of $|r_{y_1-y_2}|$ greater than 80%. This was expected as when correlation coefficient tends toward unity, fewer components are needed since both y-variables are almost the same, and so is **X**.

It should be noted that the number of components retained depends strongly on the selected performance criteria. If another metric would have been selected or if the performance had been calculated using the direct mapping, the number of combinations associated with each inversion cases and their distribution relative to the level of correlation between both y-variables might have been different.

4.2.1 Scenario 2: Impact of Collinearity When $A = K$

In this scenario (involving inversion Case 1), the 148 combinations associated with $A = 2$ in **Figure 4** are considered to analyze the impact of correlation between both y-variables. For each of them, the specification regions were defined with both techniques. **Figure 5** shows the performance calculated with the test dataset for the different metrics. It should be noted that the FOR metric was not shown as in previous analysis, because it provides redundant information with TPR. To facilitate interpretation of the figure, the data were filtered using a moving average and a window of five samples to minimize the stochastic variations introduced by random generation of the model parameters in **Eq. 5**.

First, the accuracy is analyzed as it gives an overview of classification performance since it measures the proportion of well-classified samples. Classification performance is judged against the so-called no-skill line (NSL). The latter represents the accuracy that would be obtained if the samples were randomly assigned to a class. The performance of a useful classifier needs to be above the NSL. As the ratio of good to bad samples is 4:1 in this study, the NSL is set at 80%. Except for a few regions obtained from direct mapping with combination of highly-correlated quality attributes, the accuracy is above the no-skill line. This shows that both methods performed better than making random decisions. Also, for low to moderate levels of correlation (i.e., up to 60%) accuracy is almost the same for both methods. To discriminate both methods in this zone, other metrics need to be analyzed.

It is possible to observe that a distinction exists between both methods at all correlation levels. The PPV is greater for direct mapping which means the classifier has a better precision. However, the TPR rate is lower because the predictions for the positive class is better with the inversion. Usually, a compromise between TPR and PPV needs to be achieved to identify the best classifier. Also, it can be observed that the FPR is lower for direct mapping. This is considered an advantage for DM when the goal is to minimize the risk of producing bad quality products, since the probability of accepting a bad lot is lower.

For levels of correlation higher than 60%, the gap between the two methods widens especially for the TPR metric. The DM technique becomes more restrictive and generate more rejection of good lots of raw materials whereas the region obtained with inversion leads to accepting all the good lot as it tends toward 100%. For the FPR, a large increase is observed for both methods. However, even if the rate doubles and seems more drastic compared to the other metrics, it is normal to have higher values since there are fewer bad lots than good ones. Based on the ratio of bad and good samples, an increase of one FP leads to an increase of 4% of the FPR, while an increase of one FN causes a decrease of 1% of the TPR.

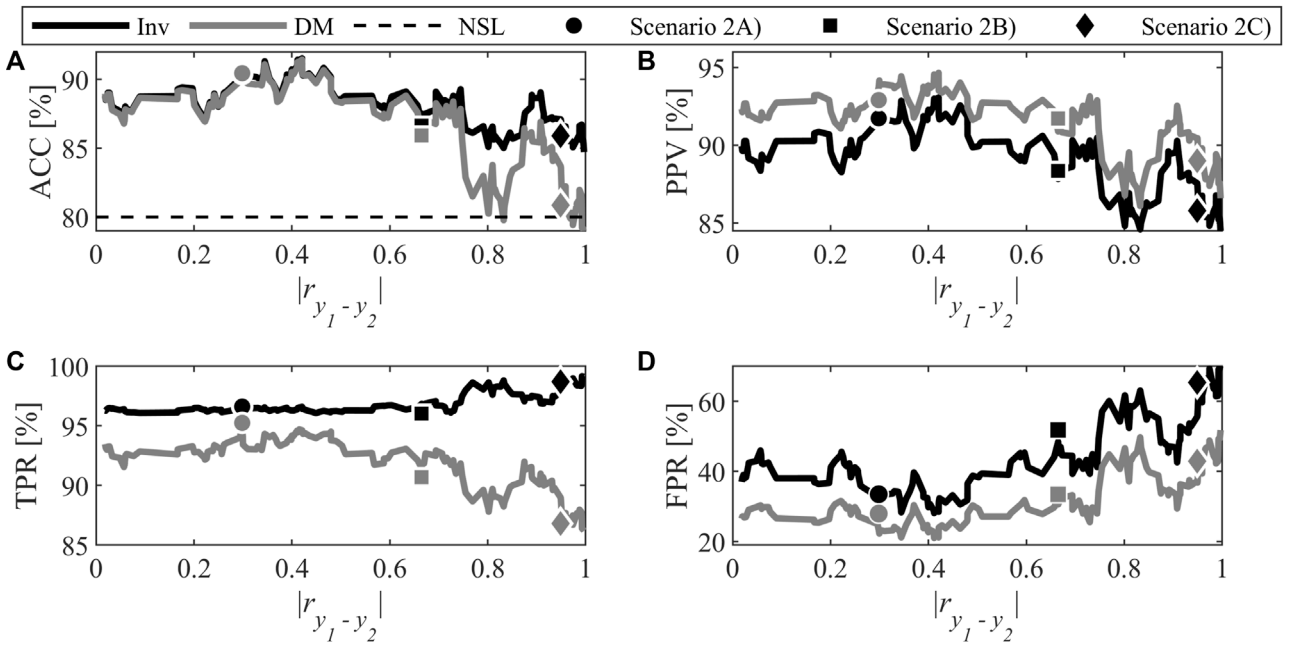

FIGURE 5 | Classification performance for Scenario 2: **(A)** Accuracy, **(B)** Positive Predictive Value, **(C)** True Positive Rate, and **(D)** False Positive Rate. The symbols mark the specific combinations investigated further, and their color indicate the method (DM or inversion).

To better understand what happens when the level of correlation increases, three examples were drawn from the set of 148 combinations to compare the acceptance regions obtained with both techniques as collinearity between quality attributes increase. The simulator's parameters used for these examples and their respective level of correlation is presented in **Table 2** (Scenarios 2A–C). The classification metrics for all three examples are shown in **Figure 5** using markers. The marker shape discriminates the level of collinearity, and its color is associated with the methods (DM or inversion).

As shown in **Figure 6A**, at low levels of correlation (here 30%), the two regions are almost the same. This explains why the accuracy was quite identical for DM and inversion. When collinearity increases to 66% (**Figure 6B**), a slight difference between the regions is observed. The largest region obtained by inversion increases acceptance of good lots at the expense of bad lots. The same observation can be made from **Figure 6C** when the correlation level is very high, i.e., 95%.

The three examples need to be compared together to explain the cause of increasing FPR with collinearity. As the level of correlation increases, good and bad products in the score space overlap to a greater extent which increases the difficulty to obtain distinct classes. This can also be observed in **Table 2** through the compromise between type I and II errors used for choosing the η constant in direct mapping. The percentage of classification errors increases with collinearity to achieve the desired balance between the two types of errors. This does not seem to be caused by the model performance in prediction since at high levels of correlation the model has a Q^2Y value of 90% as shown in **Table 2**, which is the case at low levels of correlation. The most likely cause for this behavior is that bad lots projecting near the origin of the scores space (i.e., generating a FP) are associated with observations in the y-space located in close to the edge of the product acceptance limit, but near the origin.

A particular attention should be paid to changes in the trends of TPR for both methods at high levels of correlation, which differ from those of other metrics. For DM, the TPR decreases and this may be explained similarly as for the increase in FPR. As the overlapping of the two product classes in score space is more important, the specification region needs to be more restrictive for good lots which generates more FN to obtain the same performance in terms of type I and II errors.

For PLS inversion, however, the trend is very different. The TPR increases to 100%, which means accepting all good lots of raw materials. Scenario 2C) in **Figure 6C** illustrates this situation. The ellipse obtained by inversion is stretched over the latent space which results in an acceptance region that includes a larger area where there are no or very few points (i.e., there is a risk of model extrapolation). The reason behind this behavior originates from the inversion of the $\mathbf{C}'$ matrix. When the correlation increases between the two y-variables, this is reflected in the y-loading matrix **C**, which eventually becomes ill-conditioned. Inverting this loading matrix increases of the norm of the scores and results in a larger ellipse. This is just like what happens to ordinary least squares regression parameters when highly correlated predictors are used.

Globally, Scenario 2 allowed showing that high correlation levels between both y-variables (i.e., higher than 80%) influences the classification performance of both methods. This may be caused by the proximity of observations to the product quality attribute acceptance limit in the y-space, the increasing overlap between both product classes in score space and model extrapolation for inversion. Concerning the classification performance itself, a distinction between both methods is observed for all the metrics. Direct mapping obtains a

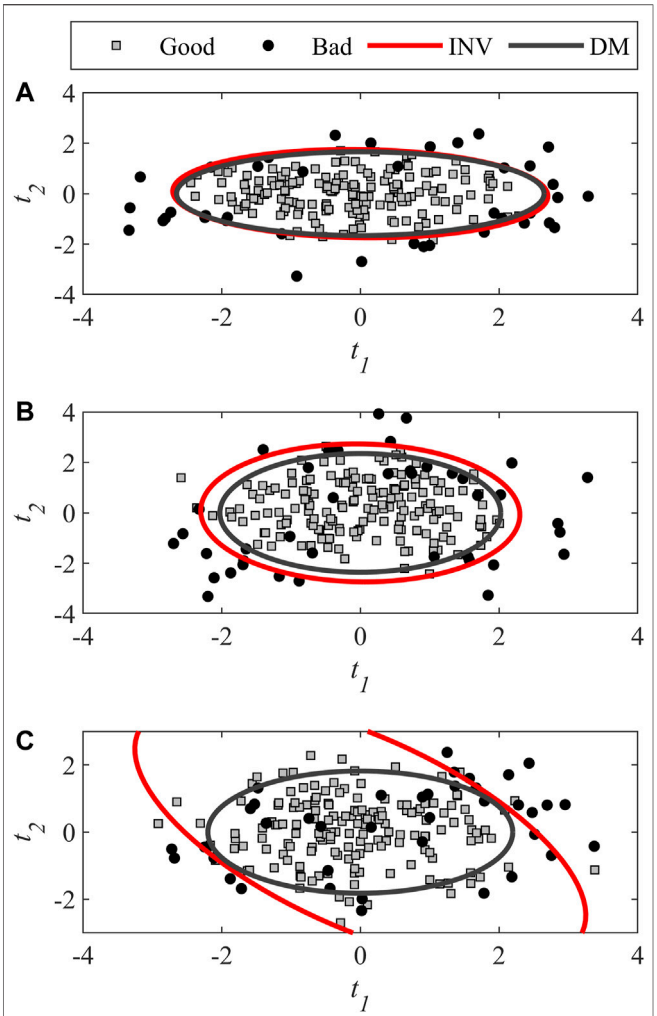

FIGURE 6 | Score plot showing the specification regions obtained with DM and inversion for different levels of correlation between the quality attributes (Scenario 2): **(A)** 30%, **(B)** 66%, and **(C)** 95%.

better FPR at the expense of TPR compared to the inversion where the relationship is opposite. Which one is best depends on the specific context and the relative cost of FPR vs. TPR.

4.2.2 Scenario 3: Impact of Collinearity when $A < K$

The third scenario illustrates the inversion Case 2 in which the number of PLS components is smaller than the number of y-variables. As the model investigated further in this section has only one component, the multivariate specification region in the latent space boils down to univariate limits (i.e., lower and upper bounds). Applying PLS inversion to several points on the product acceptance ellipse results in scores evolving between a minimum and a maximum value. These are used to define the univariate limits.

The simulations used to generate data in this study only led to a few combinations where $A < K$, and in all of those cases, $A = 1$ (see **Figure 4**). The 18 occurrences generated concentrate in the high correlation levels (i.e., mostly above 0.9). The classification performance is presented in **Figure 7**. Compared to **Figure 5**, the classification metrics are noisier due to the fact that the moving average was not apply due to the low number of samples.

Determining the impact of correlation is more difficult for this scenario since no information are available for the level of correlation ranging between 0 and 0.75. For the available data, a distinction between both methods can be observed in **Figure 7** for each metric, and is comparable to Scenario 2. For the same range of correlation, the direct mapping provides similar performance for both scenarios. For the inversion, using one component leads to PPV and FPR that are slightly worse compared to what is obtained with two components. For the TPR, the same behavior is observed where the values tend toward 100%. This was expected since the inversion cases 1 and 2 are obtained by minimization of prediction errors (e.g., for case 1, the resulting objective function value is 0).

For the same range of levels of correlation, the conclusions drawn for Scenario 3 are similar to those of Scenario 2. However, if FPR in inversion had been chosen as the criteria to determine the number of components it might be expected that some of these samples would have been moved to Scenario 2 ($A = K$) since for the same range of correlation level, the FPR is lower when using $A = K$. This shows that the criteria used for determining the number of components influence the distribution of the sample between the three inversion cases.

4.2.3 Scenario 4: Impact of Collinearity When $A > K$

The last scenario considers the situations where $A > K$. In the context of this study, this means that three PLS components leads to the best accuracy in inversion. In contrast with Scenario 2, the specification regions obtained by inversion are not bounded due to the existence of a null space. For this reason, the specification regions were established in three ways and compared: inversion alone (NS), inversion constrained by the KS (NS ∩ KS), and DM. For the different levels of correlation, the performance of the methods is presented in **Figure 8**. As for Scenario 2, a moving average window was applied to remove noise and make the interpretation clearer.

For accuracy and TPR, a large gap exists in the inversion results when constraining the region to be within the KS or not. This makes sense since adding a limit on the knowledge space tightens the specification region, and makes it more restrictive. The chance of rejecting a good lot is increased, which leads to a reduced number of well-classified good lots. Considering these two metrics, when bounded, the inversion technique gives similar performance compared to the direct mapping.

However, for PPV and FPR, the performance of PLS inversion using both approaches are very similar. The KS bounding does not seem to have an impact or only a slight one on the misclassification of bad lots. The difference observed in ACC is then mainly caused by misclassification of good lots. By comparing the inversion and direct mapping, the PPV in **Figure 8** shows that inversion is slightly better mainly because of lower FPR for this technique. However, the gap between the two techniques is smaller compared to Scenario 2. For the FPR in direct mapping for low to moderate correlation levels, adding one PLS component seems to double the rate when **Figure 8** and **Figure 5** are compared. This suggests that if the number of components had been selected using the FPR obtained by direct mapping, the partition of combinations might

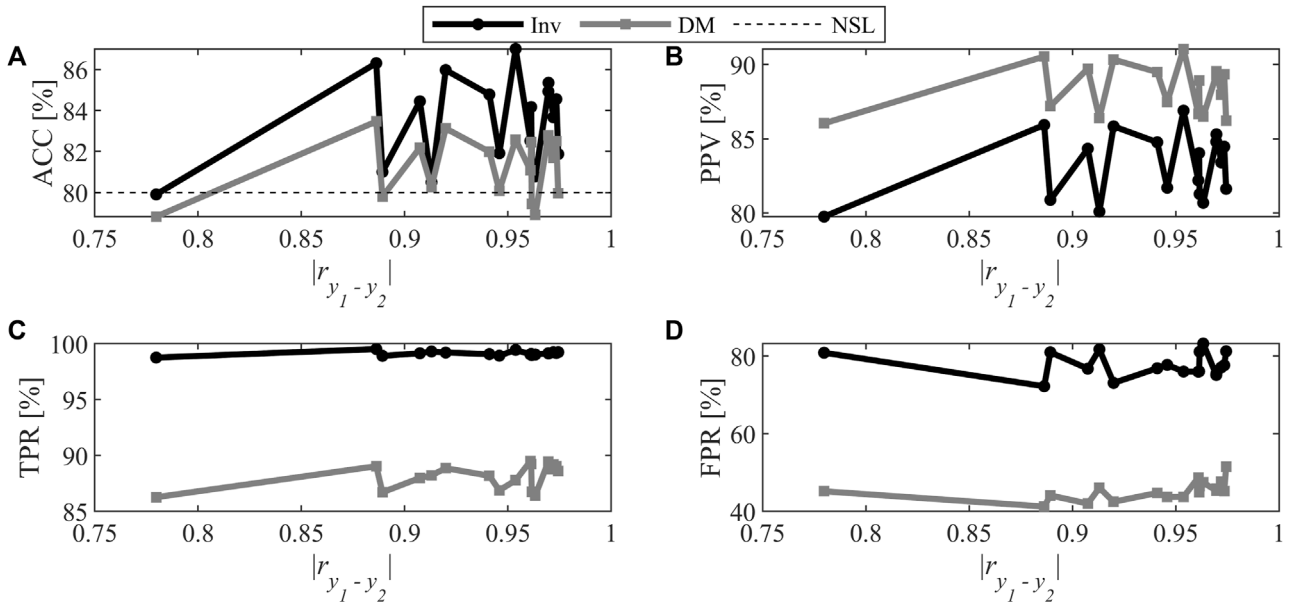

FIGURE 7 | Classification performance for Scenario 3: **(A)** Accuracy, **(B)** Positive Predictive Value, **(C)** True Positive Rate, and **(D)** False Positive Rate.

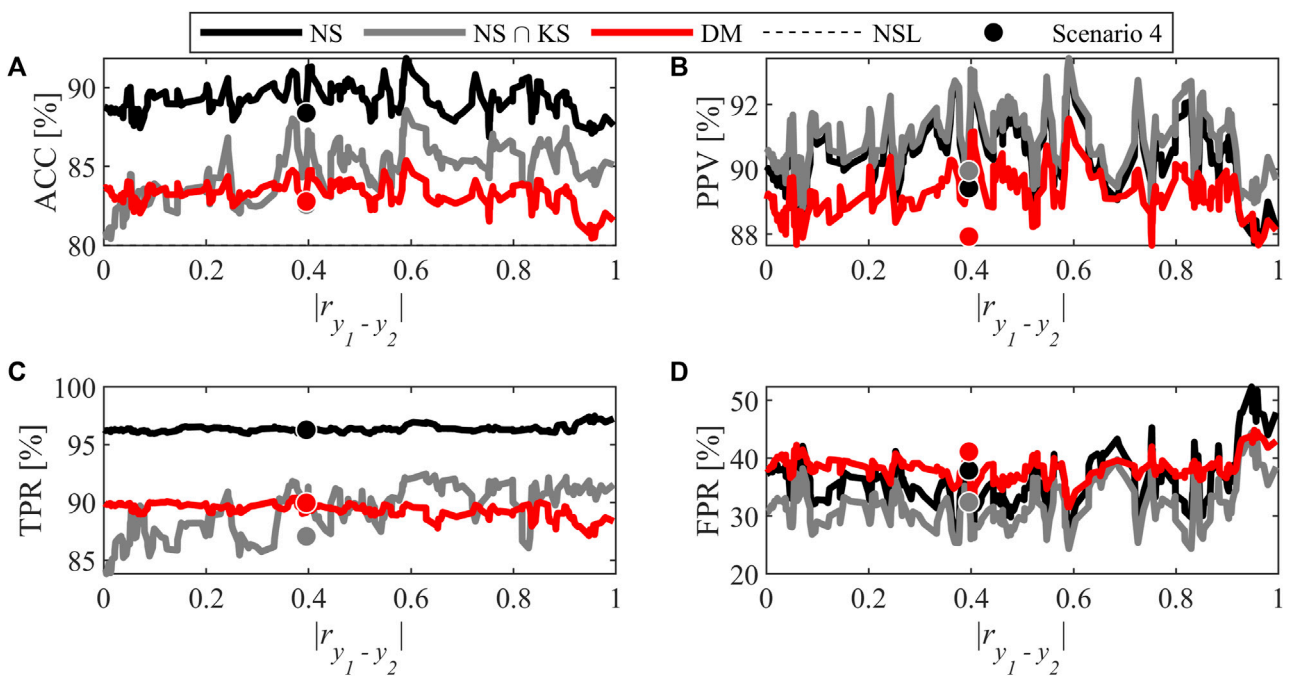

FIGURE 8 | Classification performance for Scenario 4: **(A)** Accuracy, **(B)** Positive Predictive Value, **(C)** True Positive Rate, and **(D)** False Positive Rate. The symbols mark a specific combination investigated further, and their color is used to identify the method (DM or inversion).

have been different. When testing this hypothesis, for almost all combinations, the number of components minimizing FPR is achieved using two components (i.e., $A = K$).

Figure 8 also allows interpreting the impact of the collinearity between the y-variables. Compared with Scenario 2, the correlation does not seem to have an impact on performance. Even when some fluctuations are present, the performance are relatively stable, and no systematic trend is observed in the different classification metrics. In addition, the performance at high levels of correlation does not degrade as observed in Scenario 2. In the latter, a unique solution exists for all combinations. For Scenario 4, the system of equations to solve is under-determined because the

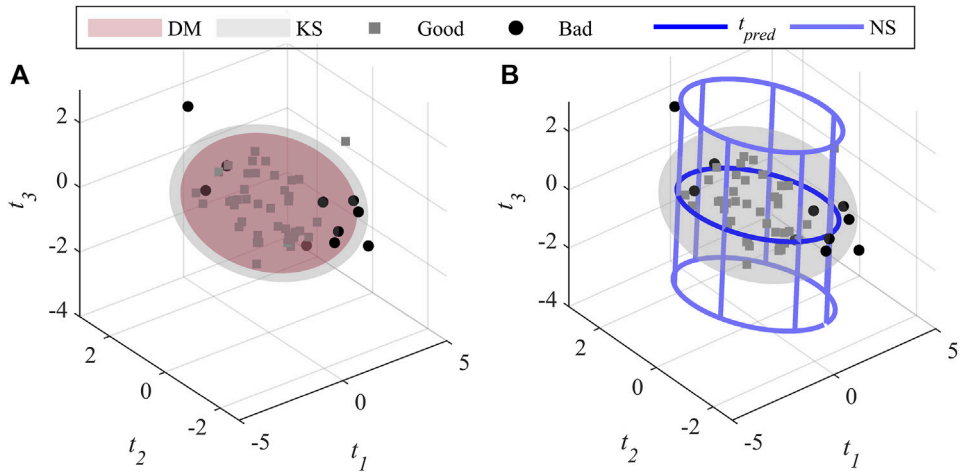

FIGURE 9 | Example of an acceptance region for the case $A > K$ using a 3 components PLS model obtained by **(A)** direct mapping, and **(B)** inversion.

number of components (i.e. scores) is greater than the number of equations. The solution provided by **Eq. 24** results from the minimization of the Euclidian norm of the score vector under the hard constraint imposed by **Eq. 20**. This forces the solution to be close to the origin of the latent space and results in a tighter and bounded specification region. The impact of collinearity between y-loading (i.e., c's) seems less important, and the TPR tend to be more stable (i.e., no increase as for Scenario 2).

In addition, Scenario 4 allows showing that specifications in three dimensions are more difficult to use compared with Scenario 2. To illustrate the situation, an example named Scenario 4, is drawn from the different combinations of y-variables requiring three PLS components. **Table 2** shows the parameters used to build the specification region while **Figure 8** shows the performance metrics of the selected combination using a makers (dots). This example is representative of the average performance across all levels of correlation. **Figure 9** shows the difference between inversion and DM in terms of the size of the specification regions. For ease of interpretation, the direct mapping and inversion are presented in different plots but using the same scale.

As in Scenario 2, the DM technique shown in **Figure 9A**) leads to a smaller region included in the knowledge space compared to inversion. **Figure 9B**) presents the predicted score vector $\boldsymbol{t}_{pred}$, the one that minimises the distance to the origin of the latent space for all the combination of y-variables. The null space representation is shown using a light color to provide a clearer image. In fact, the real representation is an elliptical cylinder where the periphery is modelled by an infinity of NS lines. If the region is unbounded, the new prediction needs to fall within the cylinder. When bounded, the point should fall at the intersection of the KS ellipsoid and the cylinder to be classified as a good lot. Thus, it is necessary to test the limit of the Hotelling T^2 first, and then to determine if the observation falls within the cylinder. Since the equation representing the specification region is unknown, it is more difficult to assess the position of a new observation using an automatic approach compared to DM.

4.3 Advantages and Drawbacks of the Methods

The various scenarios investigated allowed to identify the main advantages, and drawbacks of the two methods used for defining multivariate specification regions. This section wraps-up all the observations made through previous analyses and highlights the most important points to consider when choosing the method used to define the regions in **Table 3**.

Globally, the direct mapping approach is more restrictive in terms of volume/area compared with the inversion as the selected region is always included within the knowledge space. This can also be seen as an advantage since the user does not need to define a second limit to be within the KS. Furthermore, the DM allows a higher level of flexibility regarding the choice of the specification region shape. The inversion technique forces a similar shape to the product acceptance region in the y-space.

The type of classifier resulting from both approaches is different. Direct mapping provides a soft classifier since a choice is made by the user to set the limit. The limits can be adjusted by using the most relevant or important classification metric based on the specific objective of the case considered, for example to minimize acceptance of bad lots (i.e., FP). On the other hand, with inversion, no degree of freedom is available to adjust the position of the region based on the classification performance. The only exception is when choosing the number of components to use in the model. However, if the region is restricted to lie within the KS, the classifier becomes soft since the user needs to specify the confidence level of the T^2 limit.

The previous results have shown that it is easier to calculate the performance in classification and the location of a new sample against the specification region with direct mapping since it involves solving a simple inequality. To calculate performance using inversion, the equation of the resulting region is difficult to obtain, at the least. For example, the elliptical cylinder shown in **Figure 9** is constructed with a series of points. The current technique to determine whether a point falls within the specification region

TABLE 3 | Summary of the main features of direct mapping and inversion techniques.

	Direct mapping	Inversion ($A = K$)	Inversion ($A > K$)
Specification region shape	No restriction	Same as the y-space acceptance region	Same as the y-space acceptance region extending along the null space
Multivariate specification equation	Inequation	Area based on points in space. No direct equation to determine the position of a new point	
Ease of use on new data	Results obtained directly from the equation	Requires the use of triangularization or graphical tools if A is lower than 4 dimensions. Otherwise, calculation becomes more complicated	
Classifier Type	Soft	Hard	Soft/Hard
Permissiveness	More restrictive	More permissive Better classification of good sample	
Position of the MVspecs and KS	Always inside	Might be partially outside	When unbounded, always partially outside
Impact of correlation and performance ($A = K$)	PPV and FPR performance decrease at high levels of correlation Higher precision and lower FPR	Higher recall	—
Impact of correlation and performance ($A > K$)	No impact of correlation Worst or equal performance for all metrics	—	No impact of correlation Unbounded specifications give better performance

obtained by inversion requires performing triangularization of the area. This leads to more complex calculation compared to direct mapping where it is straightforward to use the ellipsoid equations to determine if a new prediction is included or not in the acceptance region. For 2-dimensional cases, an easier way would be to use a graphical tool to check where the point fall compared with the region. The same approach could be used for 3 dimensions, but it would be more difficult to determine if the predicted point is within the specification region volume. For more than 4 components, further research is needed to find the best way to calculate the positioning of a new lot automatically.

Based on these analyses, identifying the best approach for defining specification regions is not straightforward and depends on the user's objective. As classification performance is not superior for all the metrics for either method, one of them cannot be discarded. A compromise needs to be made during the development stage. PLS model inversion should be used when the cost of false negatives (FN) is higher than that of false positives (FP), and maximizing recall (or TPR) should be prioritized, and/or when the user prefers defining the shape of the specification regions using the PLS model structure. Otherwise, the direct mapping approach should be considered. Also, a careful attention should be paid when the y-variable are very correlated. This may lead to degradation of the classification performance. As a solution, using fewer y-variables to reduce redundancy or performing PCA on the y-space and using the scores to define the specifications could provide simple alternatives (Jaeckle, 1998).

5 CONCLUSION

The variability of raw materials is increasing, and affects the quality of the final product in many industries. To mitigate the situation, efforts are made to improve quality control. A key solution is to establish specifications regions for the properties of incoming lots of raw materials to detect unsuitable materials before processing it. In this work, a comparative analysis of two data-driven approaches for establishing multivariate specification regions using PLS models is proposed, namely the direct mapping and PLS inversion. Their classification performance is compared using multiple metrics. A focus was made on assessing the impact of collinearity in the y-space on the region classification performance.

It was shown that classification performance of bad quality lots of raw materials are poorer when quality attributes are highly correlated, when the number of PLS components is less than or equal to the number of y-variables. At low to moderate levels of correlation, the performance is slightly better for direct mapping when minimizing the false positive rate (TPR) or, alternatively Type II errors, is prioritized (i.e., reducing the risk of accepting poor quality raw materials). For the case where the PLS model has more components than the number of quality attributes, the performance is quite stable across the range of correlation levels. Both methods give similar classification performance when the specification region obtained by inversion is included within the knowledge space.

This study has shown that the decision of choosing a method for defining multivariate specification regions for raw materials depends on different factors. None of the method is superior in all possible cases. Direct mapping offers a higher degree of flexibility in the definition of the multivariate specification compared to inversion since the user can choose the shape of the region, and adjust its size/volume based on the most relevant criteria for a given industrial application. This technique is also advantageous in terms of computing resources as it requires solving an inequality to determine whether a new observation falls inside

the region or not instead of the more complex approaches required with inversion. All in all, the work presented should be considered as a guide for establishing multivariate specifications regions for incoming raw materials. Knowing the main advantages/drawbacks, and selecting the most relevant classification metric for their application will help users choosing the most appropriate approach for defining their specification regions.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

AP led this research, performed the simulations, analysed the results and wrote the manuscript. CD and ÉP supervised the work, and reviewed the manuscript.

ACKNOWLEDGMENTS

The authors acknowledge financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2019-04800 and ESD3-547424-2020) and the Fonds de Recherche Nature et Technologie (FQRNT) (Scholarship 287725).

REFERENCES

Amsbary, R. (2013). Raw Materials: Selection, Specifications, and Certificate of Analysis. Quality Assurance & Food Safety [Online]. Available at: https://www.qualityassurancemag.com/article/aib0613-raw-materials-requirements/ (Accessed June 20, 2021).

Azari, K., Lauzon-Gauthier, J., Tessier, J., and Duchesne, C. (2015). Establishing Multivariate Specification Regions for Raw Materials Using SMB-PLS. *IFAC-PapersOnLine* 48 (8), 1132–1137. doi:10.1016/j.ifacol.2015.09.120

Bano, G., Facco, P., Meneghetti, N., Bezzo, F., and Barolo, M. (2017). Uncertainty Back-Propagation in PLS Model Inversion for Design Space Determination in Pharmaceutical Product Development. *Comput. Chem. Eng.* 101, 110–124. doi:10.1016/j.compchemeng.2017.02.038

De Smet, J. (1993). *Development of Multivariate Specification Limits Using Partial Least Squares Regression*. Master. Hamilton, ON, Canada: McMaster University.

Duchesne, C., and MacGregor, J. F. (2004). Establishing Multivariate Specification Regions for Incoming Materials. *J. Qual. Technol.* 36 (1), 78–94. doi:10.1080/00224065.2004.11980253

Facco, P., Dal Pastro, F., Meneghetti, N., Bezzo, F., and Barolo, M. (2015). Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development. *Ind. Eng. Chem. Res.* 54 (18), 5128–5138. doi:10.1021/acs.iecr.5b00863

García-Muñoz, S., Dolph, S., and Ward, H. W. (2010). Handling Uncertainty in the Establishment of a Design Space for the Manufacture of a Pharmaceutical Product. *Comput. Chem. Eng.* 34 (7), 1098–1107. doi:10.1016/j.compchemeng.2010.02.027

García-Muñoz, S. (2009). Establishing Multivariate Specifications for Incoming Materials Using Data from Multiple Scales. *Chemometrics Intell. Lab. Syst.* 98 (1), 51–57. doi:10.1016/j.chemolab.2009.04.008

Godoy, J. L., Marchetti, J. L., and Vega, J. R. (2017). An Integral Approach to Inferential Quality Control with Self-Validating Soft-Sensors. *J. Process Control.* 50, 56–65. doi:10.1016/j.jprocont.2016.12.001

Haibo He, H., and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* 21 (9), 1263–1284. doi:10.1109/TKDE.2008.239

ICH (2009). "Pharmaceutical Development Q8(R2). ICH Harmonised Tripatite Guideline [Online]. Available at: https://database.ich.org/sites/default/files/Q8%28R2%29%20Guideline.pdf (Accessed June 20, 2021).

Jackson, J., and Edward, A. (1991). *User's Guide to Principal Components*. New York: John Willey Sons. Inc., 40.

Jaeckle, C. M., and MacGregor, J. F. (1998). Product Design through Multivariate Statistical Analysis of Process Data. *Aiche J.* 44 (5), 1105–1118. doi:10.1002/aic.690440509

Jaeckle, C. M., and MacGregor, J. F. (2000). Industrial Applications of Product Design through the Inversion of Latent Variable Models. *Chemom. Intell. Lab. Syst.* 50, 199–210. doi:10.1016/S0169-7439(99)00058-1

Jaeckle, C. M. (1998). *Product and Process Improvement Using Latent Variable Methods*. PhD. Hamilton, ON, Canada: McMaster University.

MacGregor, J. F., and Bruwer, M.-J. (2008). A Framework for the Development of Design and Control Spaces. *J. Pharm. Innov.* 3 (1), 15–22. doi:10.1007/s12247-008-9023-5

MacGregor, J. F., Liu, Z., Bruwer, M.-J., Polsky, B., and Visscher, G. (2016). Setting Simultaneous Specifications on Multiple Raw Materials to Ensure Product Quality and Minimize Risk. *Chemom. Intell. Lab. Syst.* 157, 96–103. doi:10.1016/j.chemolab.2016.06.021

Nomikos, P., and MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37 (1), 41–59. doi:10.1080/00401706.1995.10485888

Palací-López, D., Facco, P., Barolo, M., and Ferrer, A. (2019). New Tools for the Design and Manufacturing of New Products Based on Latent Variable Model Inversion. *Chemom. Intell. Lab. Syst.* 194, 103848. doi:10.1016/j.chemolab.2019.103848

Tessier, J., and Tarcy, G. P. (2010). Multivariate Specifications of Raw Materials: Application to Aluminum Reduction Cells. *IFAC Proc. Vol.* 43 (9), 1–6. doi:10.3182/20100802-3-ZA-2014.00001

Tomba, E., Barolo, M., and García-Muñoz, S. (2012). General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products. *Ind. Eng. Chem. Res.* 51 (39), 12886–12900. doi:10.1021/ie301214c

Weiss, G. M. (2013). "Foundations of Imbalanced Learning," in *Imbalanced Learning*. Editors H. He and Y. Ma (Piscataway, NJ: IEEE Press).

Wierda, S. J. (1994). Multivariate Statistical Process Control-Recent Results and Directions for Future Research. *Stat. Neerland* 48 (2), 147–168. doi:10.1111/j.1467-9574.1994.tb01439.x

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-Regression: a Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* 58 (2), 109–130. doi:10.1016/S0169-7439(01)00155-1

Yu, L. X., Amidon, G., Khan, M. A., Hoag, S. W., Polli, J., Raju, G. K., et al. (2014). Understanding Pharmaceutical Quality by Design. *Aaps J.* 16 (4), 771–783. doi:10.1208/s12248-014-9598-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 November 2021
doi: 10.3389/fchem.2021.769620

Multielement Characterization and Antioxidant Activity of Italian Extra-Virgin Olive Oils

Maria Luisa Astolfi[1]*, *Federico Marini*[1], *Maria Agostina Frezzini*[2], *Lorenzo Massimi*[2], *Anna Laura Capriotti*[1], *Carmela Maria Montone*[1] and *Silvia Canepari*[2]

[1]Department of Chemistry, Sapienza University of Rome, Rome, Italy, [2]Department of Environmental Biology, Sapienza University of Rome, Rome, Italy

Food product safety and quality are closely related to the elemental composition of food. This study combined multielement analysis and chemometric tools to characterize 237 extra-virgin olive oil (EVOO) samples from 15 regions of Italy, and to verify the possibility of discriminating them according to different quality factors, such as varietal or geographical origin or whether they were organically or traditionally produced. Some elements have antioxidant properties, while others are toxic to humans or can promote oxidative degradation of EVOO samples. In particular, the antioxidant activity of oils' hydrophilic fraction was estimated and the concentrations of 45 elements were determined by inductively coupled plasma mass spectrometry (ICP-MS). At first, univariate and multivariate analyses of variance were used to compare the element concentrations, and statistically significant differences were found among samples from different regions. Successively, discriminant classification approaches were used to build a model for EVOO authentication, considering, in turn, various possible categorizations. The results have indicated that chemometric methods coupled with ICP-MS have the potential to discriminate and characterize the different types of EVOO, and to provide "typical" elemental fingerprints of the various categories of samples.

Keywords: authenticity, chemometrics, inductively coupled plasma mass spectrometry, olive oil, statistical analysis, trace elements, traceability

OPEN ACCESS

Edited by:
Paolo Oliveri,
University of Genoa, Italy

Reviewed by:
Itziar Ruisánchez,
University of Rovira i Virgili, Spain
Alegria Carrasco-Pancorbo,
University of Granada, Spain

***Correspondence:**
Maria Luisa Astolfi
marialuisa.astolfi@uniroma1.it

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *02 September 2021*
Accepted: *19 October 2021*
Published: *16 November 2021*

Citation:
Astolfi ML, Marini F, Frezzini MA, Massimi L, Capriotti AL, Montone CM and Canepari S (2021) Multielement Characterization and Antioxidant Activity of Italian Extra-Virgin Olive Oils. Front. Chem. 9:769620. doi: 10.3389/fchem.2021.769620

1 INTRODUCTION

The elemental composition of foods is of toxicological and nutritional interest and can be considered an important quality parameter (Astolfi et al., 2021a; 2020a; 2020b). In particular, the concentrations of trace elements in extra-virgin olive oil (EVOO) are also one of the criteria for the assessment of the quality regarding storable period and freshness (Choe and Min, 2006). In fact, some elements, such as Ca, Co, Cu, Fe, Mg, Mn, Ni, and Sn, can promote the oxidative degradation of this important component of the Mediterranean diet appreciated among consumers for its nutritional properties and specific flavor (Choe and Min, 2006). Other elements (such as As, Cd, Cr, Cu, Hg, and Pb) present in EVOO are of great concern because they are toxic and potentially carcinogenic to humans even at low concentration (Tchounwou et al., 2012). The International Olive Council has established, as a quality criterion, a maximum residue level (MRL) for the content of As, Cu, Pb (0.1 mg kg^{-1}), and Fe (3 mg kg^{-1}) in olive oils and olive–pomace oils (International Olive Council, 2019), and the maximum levels of Cu and Fe in other vegetable oils have been also legislated (Codex Stan 33-1981, 2021), varying from 0.1 up to 5.0 mg kg^{-1}. Recently, element determination in EVOO samples has

gained importance for oil geographical traceability and authentication (Cordella et al., 2002; Dugo et al., 2004; Benincasa et al., 2007; Cabrera Vique et al., 2012; Camin et al., 2010; Beltran et al., 2015; Bajoub et al., 2018; Aceto et al., 2019; Damak et al., 2019; Zaroual et al., 2021). In particular, elements are useful in the characterization of protected designations of origin (PDOs) or protected geographical indications (PGIs) (European Union (EU), 2012), and they can also contribute to determine EVOO geographical provenance of non-PDO oils (Beltran et al., 2015; Aceto et al., 2019). In fact, the presence of metals in EVOO varies according to their origin and can be due to natural contamination from the soil, environment, fertilizers, and genotype of the plant or to the production process and contact with storage materials (Zeiner et al., 2005; Chatzistathis et al., 2009; Kabata-Pendias, 2010; Lepri et al., 2011; Yaşar et al., 2012; Bakircioglu et al., 2013). A suitable statistical treatment of trace element data could allow a geographical characterization of different EVOO samples. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) (Gumus et al., 2017; Luka and Akun, 2019; Russo et al., 2020; Savio et al., 2014), linear discriminant analysis (LDA) (Benincasa et al., 2007; Cabrera-Vique et al., 2012; Beltran et al., 2015; Damak et al., 2019), classification trees (CTs) (Gumus et al., 2017), and artificial neural networks (ANNs) (Farmaki et al., 2012; Gonzalez-Fernandez et al., 2019) have been used most.

Several beneficial implications of EVOO are derived from its antioxidant content (Dugo et al., 2020; Hannachi and Elfalleh, 2020). Intake of antioxidant compounds from oil, such as phenols, phenolic acids, and flavonoids (Capriotti et al., 2014), is usually related to health well-being. As well known, natural antioxidants play a key role in contrasting reactive species activity in living organisms, thus preventing oxidative stress-related diseases, such as cardiovascular and neurodegenerative illness and many other chronic disorders (Pérez-Jiménez et al., 2008; Cioffi et al., 2010; Šarolić et al., 2014). Moreover, antioxidants prevent lipid oxidations that cause quality degradation and unpleasant taste formation in edible oils (Christodouleas et al., 2015). Therefore, estimation of antioxidant capacity is crucial for evaluating oil's healthy and organoleptic properties. One of the most widely used *in vitro* procedures to routinely and globally estimate oil antioxidant power is the 2,2-diphenyl-1-picrylhydrazyl spectrophotometric assay (DPPH) that has the possibility of being easily applied to a high number of samples, allowing a great level of reliability (Kedare and Singh, 2011; Frezzini et al., 2019). The assay is based on the quantitative measurement of the decrease of absorbance due to the scavenging capacity of antioxidants present in the sample toward DPPH free radicals (Christodouleas et al., 2015).

All the described aspects making trace element determination, as well as the antioxidant activity of EVOO samples, are very important for both economic and health contexts (Zaroual et al., 2021; Bajoub et al., 2018). In particular, the European Union is the first producer, consumer, and exporter of olive oil in the world (Eurostat, 2019; IOC, 2018a,b). Italy follows Spain, the first world producer with an average of 20% of the total European olive oil production. About two-thirds of total Italian production is represented by EVOO (Carbone et al., 2018). Therefore, the use of a rapid and accurate analytical method for trace element analysis in EVOO has a great importance in quality control and food analysis (Llorent-Martínez et al., 2011; Shah and Soylak, 2021). Unfortunately, the determination of trace elements in EVOO samples is particularly difficult to perform, as some of them are present at very low concentrations and due to high complexity of the matrix (Shah and Soylak, 2021; Trindade et al., 2015). Sample preparation of EVOO samples is a critical step, and the determination of trace elements in EVOO requires very sensitive instrumental techniques such as inductively coupled plasma–mass spectrometry (ICP-MS) (Astolfi et al., 2021b).

The main purpose of this study is to evaluate the most significant relationships between element levels in EVOO and different categorizations, mostly related to the geographical origin using chemometric tools coupled with the ICP-MS method. For this purpose, 45 elements from a total of 237 EVOO samples from 15 Italian regions were analyzed. Also, the antioxidant activity of oils' hydrophilic fraction (HF) was estimated by the DPPH assay. The corresponding data set constituted the basis for building and validating classification models for the discrimination of the samples according to specific categorizations, which reflect possible quality attributes of the oils (and for which there could be a statistically significant number of individuals available). In particular, discriminant classification models were built using partial least square discriminant analysis (PLS-DA) to account for the possibility of dealing with correlated variables and low samples to variable ratios; moreover, to evaluate model stability and, at the same time, their reliability in an unbiased way, also in cases where the available number of samples per category was not too large, a repeated double cross-validation strategy (rDCV) was adopted.

2 MATERIALS AND METHODS

Sample collection

EVOO samples (N = 237) were collected between 2017 and 2018 from 15 production regions of Italy and different cultivars. In particular, a total of 64 EVOO samples were with PDOs and 21 with PGIs (European Union (EU), 2012). **Table 1** summarizes the number of EVOO samples according to their geographical provenances in terms of the regions. All samples (~100 mg) were kept in screw-capped glass vials in the dark at room temperature until analysis.

Chemicals

All the solutions were prepared with deionized water (18.3 MΩ cm resistivity) obtained from an Arioso (Human Corporation, Seoul, Korea) Power I RO-UP Scholar UV deionizer system. HNO_3 at 67% (suprapure; Carlo Erba Reagents, Milan, Italy), H_2O_2 at 30% (suprapure; Merck KgaA, Darmstadt, Germany), and Ar, He, and H_2 gases at 99.9995% (SOL Spa, Monza, Italy) were used.

For ICP-MS analysis, all calibration standard solutions were prepared from a 1,000-mg l^{-1} multielement standard solution (VWR International, Milan, Italy) by dilution with 10% (v/v) HNO_3 and H_2O_2 (2:1 v/v). Single standard solutions of In, Rh, Sc,

TABLE 1 | Number of samples of extra virgin olive oil for each considered category.

	Region	All samples	Oil production			Cultivar (number of samples)
			Organically	Non-organically	Not reported	
Northern Italy	Trentino Alto Adige	7	2	4	1	Blend (3); Casaliva (1); Coratina (2)
	Liguria	6	0	4	2	Lavagnina (1); Taggiasca (4)
	Lombardy	3	0	2	1	Blend (1); Casaliva (1); Leccino (1)
	Veneto	3	1	2	0	Blend (2); Grignano (1)
	Emilia Romagna	1	0	1	0	Careggiolo (1)
Central Italy	Abruzzo	14	3	11	0	Blend (8); Dritta (3); Intosso (3)
	Lazio	24	6	8	10	Blend (5); Canino (3); Frantoio (1); Itrana (2); Leccino (2); Rosciola (1)
	Marche	7	2	5	0	Ascolana (2); Blend (1); Leccino (1); Orbetana (1); Raggiola (2)
	Toscany	79	33	42	4	Blend (38); Arancino (1); Coratina (1); Frantoio (10); Leccino (4); Moraiolo (7); Nocellara (1); Olivastra Seggianese (1); Pendolino (1); Raggiolo (1)
	Umbria	8	0	8	0	Blend (8)
Southern Italy	Apulia	33	6	18	9	Blend (3); Coratina (18); Frantoio (1); Leccino (1); Ogliarola (2); Olivastra (1); Peranzana (4); Pichioline (2)
	Calabria	12	5	6	1	Blend (3); Carolea (2); Nocellara (1); Ottobratica (4)
	Campania	7	0	7	0	Blend (2); Cammarotana (1); Ortice (1); Ravece (1); Salella (1)
	Sardinia	12	1	11	0	Bosana (3); Blend (6); Semidana (1)
	Sicily	21	4	14	3	Blend (4); Biancolilla (2); Cerasuola (1); Leccio del Corno (1); Nocellara (9); Tonda Iblea (3)

and Th (at 0.010 mg l^{-1} from 1,000 ± 5 mg l^{-1}; Merck KGaA, Darmstadt, Germany) and Y (at 0.005 from 1,000 ± 2 mg l^{-1}; Panreac Química, Barcelona, Spain) were used as internal standards. A multielemental solution containing Ba, Be, Ce, Co, In, Pb, Mg, Tl, and Th (at 0.005 mg l^{-1} from 10.00 ± 0.05 mg l^{-1}; Spectro Pure, Ricca Chemical Company, Arlington, TX, USA) was used to check the instrument performance.

For the estimation of the antioxidant activity of EVOO samples, DPPH was purchased from Sigma Aldrich Co. (St. Louis, MO, USA).

Sample preparation and analysis

2.1.1 Analysis of elements

Duplicate samples (~0.5 g) of each EVOO variety were accurately weighed in 10-ml disposable graduated tubes (Artiglass, Due Carrare, PD, Italy). Then, 5 ml reagent mixture of 10% (v/v) HNO_3 and H_2O_2 (2:1 v/v) was added to each tube and heated in a water bath (WB12, Argo Lab, Modena, Italy) at 95°C for 40 min (Astolfi et al., 2021b). The lower aqueous phase was transferred into a clean tube and subjected to the ICP-MS (820-MS; Bruker, Bremen, Germany) analysis without further dilutions. The elements were monitored in standard and collision–reaction interface (CRI) modes to check and reduce possible polyatomic interference, and the following isotopes were used: ^{7}Li, ^{9}Be, ^{11}B, ^{23}Na, ^{24}Mg, ^{27}Al, ^{28}Si, ^{31}P, ^{39}K, ^{44}Ca, ^{49}Ti, ^{51}V, ^{52}Cr, ^{55}Mn, ^{57}Fe, ^{59}Co, ^{60}Ni, ^{65}Cu, ^{66}Zn, ^{71}Ga, ^{75}As, ^{78}Se, ^{85}Rb, ^{88}Sr, ^{90}Zr, ^{93}Nb, ^{98}Mo, ^{107}Ag, ^{112}Cd, ^{118}Sn, ^{121}Sb, ^{125}Te, ^{133}Cs, ^{137}Ba, ^{139}La, ^{140}Ce, ^{141}Pr, ^{146}Nd, ^{159}Tb, ^{163}Dy, ^{182}W, ^{205}Tl, ^{208}Pb, ^{209}Bi, and ^{238}U. CRI was used with He (30 ml min^{-1}) and H_2 (70 ml min^{-1}) as cell gases. The ICP-MS operating conditions and parameters were as follows: radiofrequency power 1,400 W; plasma Ar flow rate 18 l min^{-1}; auxiliary Ar flow rate 1.8 l min^{-1}; nebulizer gas flow rate 0.9 l min^{-1}; peak hopping scanning mode; steady-state analysis mode; dwell time between 50 and 100 ms, pump rate 3 rpm; five scans/replicate; and three replicates/sample. For the quantitative analysis of EVOO samples, calibration curves were built on seven different concentrations between 0.00025 and 0.05 mg l^{-1} and 0.0125 and 5 mg l^{-1} for all trace and major elements, respectively.

2.1.2 Estimation of antioxidant activity

DPPH assay was performed according to the procedure described by Šarolić et al. (2014) with slight modifications. In detail, ~0.5 g of each EVOO sample was mixed with 1 ml of 80:20 (v/v) CH_3OH:H_2O, and the mixture was blended in an ultrasonic bath (PROCLEAN 10.0 ultrasonic cleaner; Ulsonix, Berlin, Germany) for 15 min at 30°C. When the two phases appeared, the hydrophilic phase was collected, and the extraction was repeated another two times. Then, the hydrophilic extracts were combined to get a homogeneous sample. To perform DPPH assay, 50 µl of the HF sample was added to 2 ml of methanolic DPPH (0.04 mM), then the mixture was shaken for 30 min by rotating agitation (60 rpm; rotator; Glas-Col, Terre Haute, IN, USA) at room temperature in the dark and analyzed by UV-Vis spectrophotometry (Varian Cary 50 Bio UV-Vis; Varian Inc., Palo Alto, CA, USA) set at 517 nm, by measuring the sample absorbance decrease against the control (blank solution). Solutions were prepared daily and used fresh, and three replicates of each type of oil were performed. The DPPH radical scavenging activity was calculated in terms of percentage reduction of DPPH according to the following equation:

$$\text{DPPH [\%]} = \frac{(A_0 - A_S)}{A_0} \times 100$$

where A_0 represents the absorbance of the blank solution and A_S is the absorbance of the sample.

Quality assurance

The method accuracy for element determination was checked by recovery assays in the EVOO samples adding element at the low (0.005 and 0.02 mg l^{-1}) and high (0.2 and 1 mg l^{-1}) spike concentrations for all trace and major elements (B, Ca, K, Mg, Na, P, Si, and Sr) and always in the linear calibration range. In addition, accuracy was tested by a certified reference material (Conostan S-21; lot number: 21550100) obtained from SCP SCIENCE (Baie D'Urfé, Canada). The recoveries fell within 20% of the expected value and reproducibility lower than 20% (Astolfi et al., 2021b). The method detection and quantification limits (MDL and MQL, respectively) were in the range 0.004–510 and 2.5–5,000 μg kg^{-1}, respectively. Only the Ca, Cr, Mg, Mn, Ni, P, Rb, Ti, and Zn levels in the EVOO samples were 100% greater than the MDL. The possible instrumental drift for the ICP-MS analysis was checked and corrected using an internal standard solution of In, Rh, Sc, Th, and Y (Astolfi et al., 2021b; 2020c). Blank samples and control standards were tested every 20 samples in each run, and recalibration was performed every 100 samples.

Statistical analysis

The data were statistically evaluated according to the procedures of the software SPSS Statistics 25 (IBM Corp., Armonk, NY, USA) for univariate analysis. Analytical replicates were averaged prior to the successive elaboration. Non-parametric tests (Kruskal–Wallis and pairwise *post-hoc*) were applied because of the unequal numbers of samples per group and the not normal distribution (Soliani, 2003). The element concentrations measured below MDL were substituted by its half value (MDL/2) for the statistical elaboration (Farmaki et al., 2012). A *p*-value lower than 0.05 was considered statistically significant.

Partial least square discriminant analysis (PLS-DA; Ståhle and Wold, 1987; Barker and Rayens, 2003) implemented through in-house written functions running under the Matlab environment (R2015b, v.8.6, The MathWorks Inc., Natick, MA, USA) was used to build multivariate classification models. PLS-DA is a regression-based classification model which operates by coding class belonging by means of a dummy binary response matrix (or vector, when the problems involve only pairs of classes, as in the present study). In particular, if discrimination is sought between two categories, class belonging of the training samples is described by the vector y, having 1 in correspondence of all the individuals from the first class and 0 in all the remaining positions (i.e., those corresponding to the second group). A PLS model (Wold et al., 1983) is then built between the experimental data X and the dummy vector y, and the predicted value of the response ($\hat{y}$) constitutes the basis for the classification of the samples: since the predicted responses are real-valued, an optimal threshold y_{thres} has to be calculated so that, if the predicted response is greater than y_{thres}, the sample is predicted as class 1, otherwise as class 2. In the present study, the threshold was calculated by applying LDA on the predicted responses calculated on the training samples (Perez et al., 2009).

The reliability of the classification models was evaluated by means of a repeated double-cross-validation (rDCV) procedure (Filzmoser et al., 2009). Double cross-validation (DCV) is a validation strategy which involves two nested loops of cross-validation: an inner loop for model selection (i.e., for choosing the optimal number of latent variables) and an outer loop which mimics an external (i.e., not involved in any model building and/

TABLE 2 | Method detection limits (MDL; μg kg^{-1}) and element levels [median, minimum (min) and maximum (max); μg kg^{-1}] in extra-virgin olive oils (EVOO; *n* = 237) from all over Italy.

Element	MDL	Italian EVOO samples			
		%N > MDL	Median	Min	Max
Ag	0.06	27	<0.06	<0.06	0.86
Al	9	85	34	<9	1,300
As	0.3	28	<0.3	<0.3	4.0
B	20	16	<20	<20	770
Ba	0.7	49	<0.7	<0.7	175
Be	0.004	43	<0.004	<0.004	0.431
Bi	0.1	23	<0.1	<0.1	1.0
Ca	510	100	4,090	1,230	35,700
Cd	0.07	67	0.09	<0.07	0.97
Ce	0.1	65	0.2	0.1	3.5
Co	0.05	70	0.12	<0.05	2.16
Cr	0.3	100	5	0.4	839
Cs	0.007	55	0.008	<0.007	0.101
Cu	0.6	99	3.2	<0.6	41.6
Dy	0.005	19	<0.005	<0.005	0.055
Fe	12	99	77	<12	582
Ga	0.06	15	<0.06	<0.06	0.69
K	40	24	<40	<40	939
La	0.05	70	0.10	<0.05	0.79
Li	0.06	32	<0.06	<0.06	6.07
Mg	10	100	91	21	723
Mn	0.5	100	2.4	1.1	43.5
Mo	0.3	20	<0.3	<0.3	2.0
Na	25	98	110	<25	585
Nb	0.04	7	<0.04	<0.04	0.11
Nd	0.03	52	0.03	<0.03	13.8
Ni	0.5	100	5.6	2.1	49.7
P	60	100	272	127	650
Pb	0.3	99	0.9	<0.3	22.1
Pr	0.008	40	<0.008	<0.008	1.65
Rb	0.06	99	0.24	<0.06	1.77
Sb	0.02	16	<0.02	<0.02	0.37
Se	0.6	48	<0.6	<0.6	7.8
Si	270	1	<270	<270	3,340
Sn	0.06	63	0.08	<0.06	1.94
Sr	1	65	3	1	58
Tb	0.006	19	<0.006	<0.006	1.28
Te	0.03	3	<0.03	<0.03	0.06
Ti	0.4	100	1.9	0.8	10.7
Tl	0.06	0	<0.06	<0.06	<0.06
U	0.005	30	<0.005	<0.005	0.050
V	0.08	98	0.53	<0.08	1.40
W	0.3	38	<0.3	<0.3	5.1
Zn	20	100	111	54	749
Zr	0.1	57	<0.1	<0.1	2.3

or optimization stage) test set, to be used for estimating the prediction and generalization ability. In order to avoid that the performances of the model depend on a particular sample splitting scheme, the procedure is repeated a sufficient number of times, changing the distribution of the individuals across the different cancelation groups, hence the name "repeated" DCV. Repeating the double-cross-validation procedure allows also having multiple predictions for the same samples, which translates to the possibility of estimating confidence intervals for all the classification figures of merit and model parameters.

3 RESULTS AND DISCUSSION

Levels of elements

Table 2 shows the concentration of the elements in EVOO from all over Italy. The content of Tl was below the respective MDL (0.06 μg kg^{-1}) in all the samples. Si, Te, and Nb were found above the MDL (270, 0.03, and 0.04 μg kg^{-1}, respectively) only in 1%, 3%, and 7% of all samples. Only the Ca, Cr, Mg, Mn, Ni, P, Ti, and Zn levels in the EVOO samples were 100% greater than the MDL. The maximum concentrations for As (4.0 μg kg^{-1}), Cu (41.6 μg kg^{-1}), Fe (582 μg kg^{-1}), and Pb (22.1 μg kg^{-1}) were lower than the MRLs established by the IOC for olive and pomace-olive oils, which are 100 μg kg^{-1} for As, Cu, and Pb and 3,000 μg kg^{-1} for Fe (International Olive Council, 2009). Calcium showed the highest concentration ranging from 1,230 to 35,700 μg kg^{-1}, whereas from 10- to 50-fold lower levels were found for Fe, Mg, Na, P, and Zn (median = 77, 91, 110, 272, and 111 μg kg^{-1}, respectively).

Concentrations of elements obtained in this study were compared to levels measured in EVOO from several other Mediterranean countries (**Supplementary Tables S1–S4**). Levels of many elements showed wide variability even within the same country. The Ag, Ba, P, and Sn data were not considered because these elements are not completely extracted with the method used. As regards the content of B, Be, Dy, Nd, Pr, Si, Tb, and Te, we could not find other data for EVOO in the literature. Our results were similar to those reported by another study on Italian EVOO (Benincasa et al., 2007); on the contrary, they differed significantly from other data concerning most of the elements investigated in the EVOOs of Spain (Beltran et al., 2015; Llorent-Martínez et al., 2014), Croatia (Pošćić et al., 2019), Tunisia (Damak et al., 2019), and Turkey (Gumus et al., 2017). The concentrations of Ca (1,230–35,700 μg kg^{-1}), Cr (0.4–839 μg kg^{-1}), Mg (21–723 μg kg^{-1}), and Ni (2.1–49.7 μg kg^{-1}) found in this study were in the same range to that found in other Italian EVOO (Ca = 1850–26,900 μg kg^{-1}; Cr = 116–437 μg kg^{-1}; Mg = 56–1,030 μg kg^{-1}; and Ni = nd-46.9 μg kg^{-1}) as reported by Benincasa et al. (2007), but from 10 to 100 times higher than the levels reported in Croatian (Pošćić et al., 2019) and Turkish EVOO (Gumus et al., 2017). Fe concentrations (<12–582 μg kg^{-1}) varied from 100 times lower to 100 times higher than the level of Fe quantified in EVOO from Turkey (1–14,670 μg kg^{-1}) by Gumus et al. (2017) and Croatia (0.19–2.57 μg kg^{-1}) by Pošćić et al. (2019) or Spain (0.5–1.2 μg kg^{-1}) by Beltran et al. (2015), respectively. This variability in the concentrations of the elements present in EVOO samples may depend on various factors related to the geochemistry of the provenance soil but also to physiological aspects typical of the species from which a particular EVOO derives (Giaccio and Vicentini, 2008).

Grouping the data according to geographic origin as north (Emilia Romagna, Liguria, Lombardy, Trentino Alto Adige, and Veneto), center (Abruzzo, Lazio, Marche, Tuscany and Umbria), and south (Apulia, Calabria, Campania, Sardinia and Sicily) of Italy, it is possible to identify elements that differ significantly from one group to another (**Table 3**). In particular, the EVOO samples from northern Italy had significantly higher levels of Cs, Fe, Na, P, and Pr than those from central Italy and Fe, Pr, and U than those from southern Italy. Both Fe and Pr appear to provide a good tool for tracing the EVOO production chain in accord with other authors (Aceto et al., 2019; Damak et al., 2019). Iron is common in silicates and carbonates present in soil (Pohl, 2011); however, some authors reported that Fe may be present in edible oils as a result of storage and processing contaminations (Mendil et al., 2009; Zeiner et al., 2010). Praseodymium and the other lanthanides do not have a defined role in the metabolism of plants; therefore, their distribution remains almost unchanged in the passage from the soil to the fruits (Aceto et al., 2019). For this reason, these elements can be used as fingerprints to discriminate the geographic origin of the EVOO samples (Farmaki et al., 2012; Aceto et al., 2019). In addition, the analysis of some elements in EVOO, such as Cs and Rb, which can be easily mobilized in the soil, can be linked to a geogenic source rather than an anthropogenic origin (such as extraction process or cultivation practices) and can help in the geographical traceability of EVOO samples (Kelly, Heaton, & Hoogewerff, 2005).

By comparing the concentrations of the elements in the EVOO samples from each region (**Supplementary Tables S5–S7**), the number of elements that differ significantly increases. **Table 4** shows a summary of all the elements that differ significantly according to the region. Emilia Romagna was not considered for the comparison because there was only one EVOO sample to consider. EVOOs from Lombardy did not have levels of elements that are significantly different from those of oils from all other regions. Considering the other oils of northern Italy, the EVOOs from Trentino and Liguria differed significantly from the EVOOs from Marche only for the content of Na, which in the EVOOs from Marche (median = 38 μg kg^{-1}) was about four times lower, while the EVOOs from Veneto had a higher content of Fe (median = 218 μg kg^{-1}) than the oils from Abruzzo (median = 15 μg kg^{-1}) and a higher content of Fe and Na (median = 218 and 174 μg kg^{-1}, respectively) compared to the Marche. The EVOO samples from Lazio differed significantly for a large number of elements (Ba, Ca, Cd, Ce, Cs, Dy, Ga, La, Mg, Na, Nd, Pr, Pb, Rb, Sb, Sr, Tb, Ti, U) compared to Tuscany, Abruzzo, Campania, and Marche. In all cases, levels of Cd (median = 0.14 μg kg^{-1}), La (median = 0.20 μg kg^{-1}), and Rb (median = 0.48 μg kg^{-1}) were higher than those of oils from other regions mentioned above.

Antioxidant activity

Following the extraction and storage of EVOO, it is inevitable that an oxidation process occurs, which leads to a deterioration of the oil (Bendini et al., 2007). Some factors such as temperature, light,

TABLE 3 | Element levels [median, minimum (min) and maximum (max); μg kg^{-1}] in extra-virgin olive oils from north (n = 20), central (*n* = 132) and south (*n* = 85) Italy.

Element	North Italy[a]				Central Italy[b]				South Italy[c]			
	%N > MDL	Median	Min	Max	%N > MDL	Median	Min	Max	%N > MDL	Median	Min	Max
Ag	25	<0.06	<0.06	0.26	30	<0.06	<0.06	0.86	25	<0.06	<0.06	0.22
Al	75	32	<9	615	86	34	<9	1,291	86	34	<9	1,298
As	25	<0.3	<0.3	2.8	28	<0.3	<0.3	4.0	28	<0.3	<0.3	2.2
B	20	<20	<20	85	15	<20	<20	734	16	<20	<20	770
Ba	50	0.6	0.4	99.5	49	2.9	<0.7	175	48	<0.7	<0.7	147
Be	40	<0.004	<0.004	0.431	37	<0.004	<0.004	0.272	49	0.004	<0.004	0.061
Bi	40	<0.1	<0.1	0.2	23	<0.1	<0.1	0.4	19	<0.1	<0.1	1.0
Ca	100	4,590	1,480	9,170	100	3,648	1,229	35,709	99	4,278	1,432	24,122
Cd	70	0.12	<0.07	0.33	66	0.09	<0.07	0.97	67	0.09	<0.07	0.61
Ce	70	0.2	<0.1	0.7	69	0.2	0.1	3.5	60	0.2	0.1	1.0
Co	90	0.11	<0.05	0.59	68	0.13	<0.05	1.23	69	0.08	<0.05	2.16
Cr	100	3.7	0.5	839	99	5.0	0.4	123	100	4.1	0.5	533
Cs	80	0.013[a]	<0.007	0.080	52	0.007[a]	<0.007	0.084	55	0.008	<0.007	0.101
Cu	100	4.6	<0.6	20.7	99	3.0	<0.6	40.9	100	3.3	<0.6	41.6
Dy	20	<0.005	<0.005	0.010	20	<0.005	<0.005	0.026	17	<0.005	<0.005	0.055
Fe	100	158[a,b]	<12	495	99	70[a]	<12	403	100	86[b]	14	582
Ga	35	<0.06	<0.06	0.33	11	<0.06	<0.06	0.69	16	<0.06	<0.06	0.59
K	40	<40	<40	293	20	<40	<40	673	26	<40	<40	939
La	80	0.13	<0.05	0.41	70	0.08	<0.05	0.79	67	0.11	<0.05	0.71
Li	40	<0.06	<0.06	1.67	29	<0.06	<0.06	6.07	35	<0.06	<0.06	4.42
Mg	100	97	37	262	100	90	21	723	99	96	28	613
Mn	100	2.7	1.5	7.1	99	2.3	1.1	18.6	100	2.6	1.4	43.5
Mo	30	<0.3	<0.3	1.3	19	<0.3	<0.3	1.7	20	<0.3	<0.3	2.0
Na	100	131[a]	87	331	99	102[a]	<25	585	100	114	<25	513
Nb	15	<0.04	<0.04	0.05	5	<0.04	<0.04	0.06	8	<0.04	<0.04	0.11
Nd	80	0.06	<0.03	1.39	47	<0.03	<0.03	6.43	53	0.03	<0.03	13.8
Ni	100	5.2	2.5	29.5	100	6.0	2.1	40.6	100	5.4	2.4	49.7
P	100	309[a]	220	650	99	269[a]	127	522	100	272	189	548
Pb	100	1.2	<0.3	4.3	100	0.8	<0.3	22.1	100	1.1	<0.3	8.7
Pr	70	0.012[a,b]	<0.008	0.359	38	<0.008[a]	<0.008	1.58	35	<0.008[b]	<0.008	1.65
Rb	95	0.29	<0.06	1.10	100	0.24	0.06	1.77	100	0.26	<0.06	1.36
Sb	10	<0.02	<0.02	0.04	17	<0.02	<0.02	0.37	16	<0.02	<0.02	0.14
Se	55	0.6	<0.6	6.8	49	<0.6	<0.6	6.9	44	0.6	<0.6	7.8
Si	0	<270	<270	<270	1	<270	<270	3,344	1	<270	<270	442
Sn	80	0.10	<0.06	0.45	59	0.06	<0.06	0.60	65	0.09	<0.06	1.94
Sr	80	3	1	7	64	3	1	34	63	3	1	58
Tb	25	<0.006	<0.006	0.112	20	<0.006	<0.006	1.13	16	<0.006	<0.006	1.28
Te	5	<0.03	<0.03	0.05	1	<0.03	<0.03	0.05	6	<0.03	<0.03	0.06
Ti	100	2.2	1.2	8.1	99	1.8	0.8	5.6	100	2.1	1.1	10.7
Tl	0	<0.06	<0.06	<0.06	1	<0.06	<0.06	0.08	0	<0.06	<0.06	0.03
U	55	0.006[a]	<0.005	0.044	31	<0.005	<0.005	0.044	24	<0.005[a]	<0.005	0.050
V	100	0.50	<0.08	1.04	99	0.52	<0.08	1.21	100	0.55	<0.08	1.40
W	40	<0.3	<0.3	2.0	37	<0.3	<0.3	2.3	39	<0.3	<0.3	5.1
Zn	100	145	55	283	99	98	54	749	99	143	57	672
Zr	55	0.1	0.1	0.6	53	0.1	0.1	1.8	63	0.1	0.1	2.3

[a]*North Italy groups the following regions: Emilia Romagna, Liguria, Lombardy, Trentino Alto Adige, and Veneto.*

[b]*Central Italy groups the following regions: Abruzzo, Lazio, Marche, Tuscany, and Umbria.*

[c]*South Italy groups the following regions: Apulia, Calabria, Campania, Sardinia, and Sicily. For each element, numbers in bold with the same superscript indicate significant differences ($p < 0.05$).*

oxygen and other chemical elements, unsaturated fatty acid composition, and the presence of antioxidants can affect the oxidation process differently (Frankel 1985). Phenolic compounds have antioxidant capacities in EVOO since they can eliminate peroxyl and alkoxy radicals and chelate transition metal ions present in traces (Visioli et al., 1998). Several elements are known for their antioxidant properties (Perna et al., 2012; Thiruvengadam et al., 2020). Indeed, in the present study (**Supplementary Table S8**), significant correlations were observed between the antioxidant activity and elements. A positive and significant moderate correlation (r = 0.500–0.768, $p = 0.05$) was observed between Al, Ca, Fe, V, and Zr in EVOOs from Abruzzo, Ba in EVOOs from Apulia, and B, Mn, Se, and V in EVOOs from Sardinia and the DPPH% data. Conversely, a low and positive correlation (r < 0.4) was recorded between Ba and Ni and the antioxidant activity of all samples. Other elements (Ag, Cr, Cu, Li, Sb, Si, and Tl) might not affect the antioxidant properties as non-significant correlations were observed between them.

Supplementary Tables S5–S7 show the antioxidant activity measured by the DPPH assay (DPPH%) in the EVOO samples

TABLE 4 | Summary of significant differences within medians of the 45 selected elements and antioxidant activity (DPPH%) among all samples from Italian regions by Kruskall–Wallis and pairwise *post-hoc* tests. A *p*-value lower than 0.05 was considered statistically significant.

	Trentino	Liguria	Veneto	Lazio	Tuscany	Umbria	Calabria	Apulia	Sardinia	Sicily
Toscany		DPPH%		Cd, Cs, Dy, Ga, La, Na, Nd, Pr, Rb, Sb, Tb, Ti, U	-					
Umbria				Dy, U		-				
Apulia					Ti,Zr		-			
Sardinia				La, Tb, U				Al	-	
Sicily				U	Be,DPPH%			DPPH%	Be	-
Abruzzo		DPPH%	Fe	Ba, Ca, Cd, Ce, Dy, La, Mg, Nd, Rb, Sr, Tb	Ni	Fe,Se	Fe	Ce,La,Ni,Zn,Zr		As,Ba,Ca,Ce,Fe,La Ni, Zn, DPPH%
Campania				Ba, Cd, La, Mg, Na, Rb, U			Na	La	Na	Ba,La
Marche	Na	Na	Fe,Na	Ba, Cd, Ce, Cs, La, Na, Nd, Mg, Pb, Pr, Rb, Ti, U		Fe	Cd,Fe,Na	Na,Rb,Ti	Na	Ba,Be,Rb

TABLE 5 | PLS-DA discrimination between pairs of geographical origin. Figures of merit estimated on the outer loop of the rDCV procedure (expressed as mean ± standard deviation).

Class1	Class2	% accuracy	Mean % correct classification rate	% sensitivity (Class1)	% sensitivity (Class2)
Lazio	Abruzzo	76.2 ± 3.9	77.1 ± 4.2	73.1 ± 3.9	81.0 ± 7.7
Lazio	Sicily	79.4 ± 3.1	79.0 ± 2.9	81.9 ± 4.9	76.1 ± 2.8
Lazio	Apulia	68.8 ± 5.2	68.9 ± 5.2	64.6 ± 6.2	73.2 ± 7.6
Lazio	Tuscany	75.2 ± 1.8	69.2 ± 2.3	57.8 ± 4.2	80.6 ± 2.1
Lazio	Calabria	61.7 ± 5.1	54.4 ± 5.5	71.9 ± 5.9	36.9 ± 8.5
Abruzzo	Calabria	81.4 ± 6.1	81.0 ± 7.1	82.9 ± 4.3	79.1 ± 12.2
Abruzzo	Sicily	75.0 ± 4.3	75.5 ± 4.4	81.2 ± 7.6	69.7 ± 5.9
Abruzzo	Tuscany	58.2 ± 3.4	54.6 ± 5.6	49.3 ± 11.6	59.9 ± 4.0
Abruzzo	Apulia	54.3 ± 6.6	54.2 ± 6.8	53.6 ± 9.6	54.8 ± 7.5
Sicily	Tuscany	69.5 ± 2.7	65.8 ± 4.1	59.9 ± 7.6	71.8 ± 2.8
Sicily	Apulia	70.7 ± 4.3	70.2 ± 4.2	65.4 ± 4.9	74.9 ± 6.5
Tuscany	Apulia	64.6 ± 3.1	55.1 ± 4.9	72.5 ± 2.9	37.7 ± 9.3

from each region. EVOOs from central and southern Italy showed higher antioxidant activity than oils from northern Italy. In particular, **Table 4** shows that EVOOs from Sicily had a significantly lower DPPH% (median = 18.2%) than oils from Abruzzo (median = 47%), Apulia (median = 37.6%), and Tuscany (median = 36.2%), while the EVOOs from Liguria had significantly lower DPPH% (median = 15.7%) compared to Tuscany. The highest data of DPPH% (67.3%) was found in the oils of Campania. Cioffi et al. (2010) demonstrated that oils from Campania have antioxidant properties, which are very likely due to the presence of high contents of phenolic compounds.

Classification of EVOOs according to geographical origin

At first, the possibility of discriminating the different EVOOs according to their geographical origin was considered. In particular, due to the unavailability of the information about the origin of all the samples and to the unbalancedness in the distribution of samples per class, when considering the oils of known origin, several two-class models (i.e., comparing two regions at a time) were built and validated. Here it must be further stressed that all the regions for which the available number of certified individuals was too low to be considered representative have not been included in the comparison.

In all cases, PLS-DA analysis was carried out on the matrix made up of the concentrations of the elements presenting at least 70% of the values above the limit of detection (so to avoid possible artifacts related to data imputation) and including also TEAC and DPPH. Models were built after autoscaling and validated by means of an rDCV procedure with 50 runs, 10 cancelation groups in the outer loop (the one mimicking the external test set) and 5 in the inner loop (the one used for model selection, i.e., definition of the optimal number of latent variables). The results obtained are summarized in **Table 5**, where the accuracy, the mean correct classification rate, and the sensitivities for the two compared classes are reported. Since two-class discriminant models were calculated, due to symmetry the sensitivity of a class (true positive rate) is the specificity (true negative rate) of the other category; this is why sensitivities only have been reported. Moreover, since the number of samples per class was, in some cases, highly unbalanced (**Table 1**), we have decided to report both classification accuracy (percentage of correctly classified

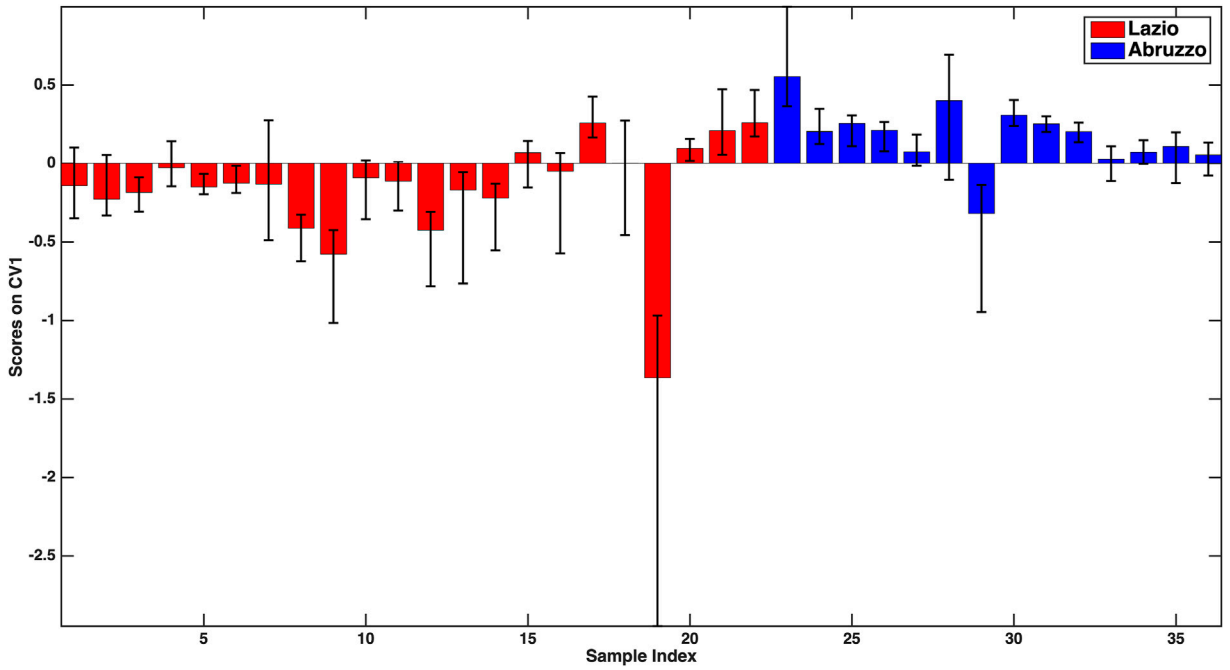

FIGURE 1 | –PLS-DA model for the discrimination between Lazio and Abruzzo samples: mean scores of the rDCV outer loop samples along the only canonical variate of the model together with their 95% confidence intervals. Legend: red bars–Lazio; blue bars–Abruzzo.

samples irrespectively of the category over the total number of samples) and the mean correct classification rate, which is the average of the specificities of the two classes.

As anticipated in the materials and methods section, the use of repeated double cross-validation allows obtaining not only a point estimate of the figures of merit on the validation (outer loop) samples but also their confidence intervals, so as to be able to evaluate the consistency of the results.

By looking at **Table 5**, it is evident how the different models result in different reliabilities, with some presenting rather low classification performances. On the other hand, there are some models which result in an overall accuracy higher than 75%, with a comparable mean correct classification rate (suggesting that the classification performances are not affected by the numerosity of the samples; **Table 1**). Additionally, the standard deviation of the figures of merit for these models is relatively low (corresponding roughly to one more sample being correctly or wrongly classified with respect to the reported averages), confirming the consistency of the obtained classification. Based on these considerations, only the best models will be discussed in detail in the remainder of this section, namely, Lazio vs. Abruzzo, Lazio vs. Sicily, Abruzzo vs. Calabria and Abruzzo vs. Sicily.

The first model to be examined is the one discriminating Lazio samples from the oils from Abruzzo, for which an overall 76.2 ± 3.9% classification accuracy on the outer loop samples was registered. By looking at the individual sensitivities together with their confidence intervals (73.1 ± 3.9% for Lazio and 81.0 ± 7.7% for Abruzzo), it can be stated that the two categories are predicted comparably well. These results can also be graphically appreciated in **Figure 1**, where the mean scores of the outer loop samples along the only canonical variate of the model together with their 95% confidence intervals are displayed. It is evident from **Figure 1** how almost all the Abruzzo samples have positive scores, while the large majority of Lazio samples are characterized by negative coordinates on the component, indicating a good separation between the categories.

For the sake of interpretation, another advantage of the rDCV procedure is that confidence intervals can also be calculated for model parameters, so as to be able to identify which are the variables that contribute significantly to the discrimination (e.g., by inspecting the values of the associated regression coefficients or of the VIP scores). Moreover, investigating the sign of the regression coefficients also allows postulating whether the associated predictor is more or less concentrated in a category with respect to the other. In particular, the variables found to significantly contribute to the discriminant model were V, Fe, Zn, Rb, antioxidant capacity (all higher in Lazio samples), and Ni and antioxidant activity in the DPPH assay (higher in the oils from Abruzzo).

As far as the Lazio vs. Sicily model is concerned, a slightly higher accuracy was obtained (79.4 ± 3.1%), the individual sensitivities being 81.9 ± 4.9% for Lazio and 76.1 ± 2.8% for Sicily. Analogously to that described above, the discrimination between the two classes can also be visually appreciated in **Figure 2**, where the mean scores of the outer loop samples along the only canonical variate of the model together with their 95% confidence intervals are displayed.

In this case, based on the values of the PLS-DA regression coefficients, all the variables found to significantly contribute

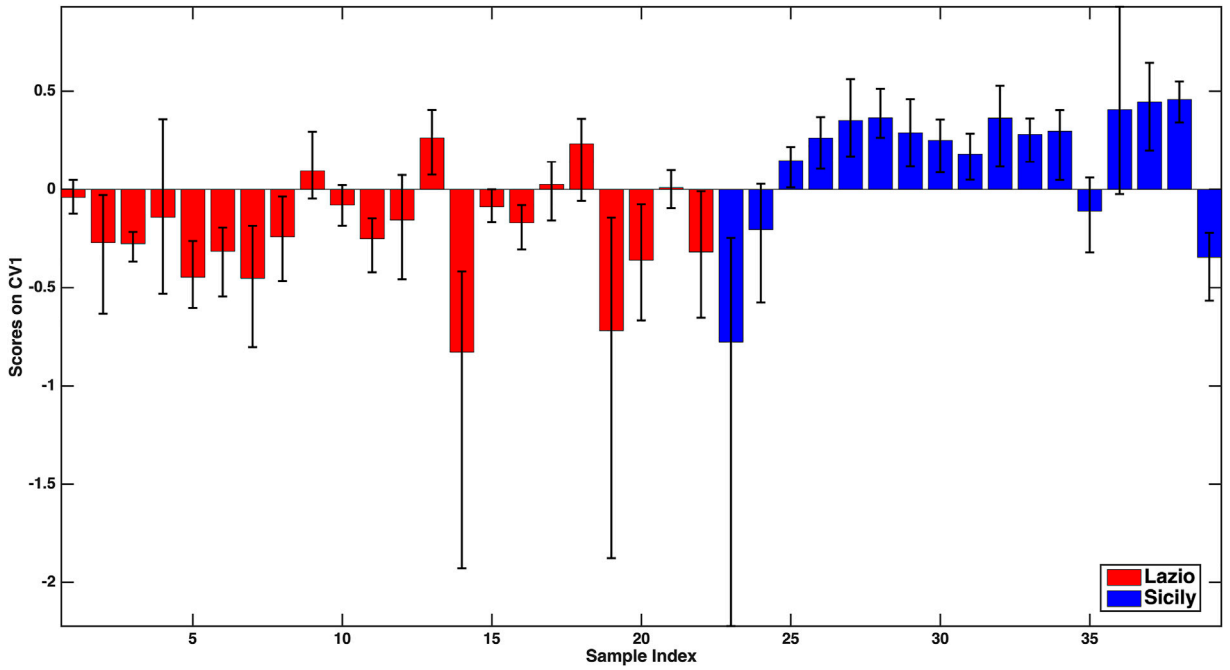

FIGURE 2 | –PLS-DA model for the discrimination between Lazio and Sicily samples: mean scores of the rDCV outer loop samples along the only canonical variate of the model together with their 95% confidence intervals. Legend: red bars–Lazio; blue bars–Sicily.

to the discriminant model (Na, Mg, P, Ti, Rb, antioxidant activity in the DPPH assay) should be, on average, higher in the oils from Lazio. When considering the model discriminating Abruzzo oils from the Calabrian ones, an overall 81.4 ± 6.1% accuracy on the outer loop samples was obtained, the mean correct classification rate (81.0 ± 7.1%) and the individual sensitivities for the two categories (82.9 ± 4.35% for Abruzzo and 79.1 ± 12.2% for Calabria) being almost equal. In particular, the higher standard deviation of the sensitivity for Calabria is due to the very limited number of samples in that class. When looking at the significant predictors, only five variables (P, V, Fe, Zn, and antioxidant activity) were identified, and the coefficients indicate that they all should be, on average, higher for the Abruzzo samples.

Lastly, the Abruzzo vs. Sicily model resulted in an accuracy of 75.0 ± 4.3%, the sensitivities being 81.2 ± 7.6% for Abruzzo and 69.7 ± 5.9% for Sicily. Inspection of the model parameters led to identifying as significant the contribution of Na, Ni, and antioxidant activity (in the DPPH assay with higher data in the oils from Abruzzo) and of V and Fe, more concentrated in the samples from Sicily.

Classification of EVOOs according to cultivar and to whether it was organically produced.

In a second stage of the study, the possibility of discriminating oil samples according to their cultivar was explored. In this case, given the available information about the samples and the fact that only a relatively small fraction of the analyzed oils was monovarietal, the investigation was restricted to the comparison of Coratina (21 samples) and Frantoio (12 samples) (**Table 1**). The PLS-DA classification approach was validated through an rDCV strategy as described in the previous section and resulted in an overall accuracy of 68.9 ± 6.2%, and 80.7 ± 8.8% and 61.9 ± 7.6% sensitivities for Frantoio and Coratina, respectively, corresponding to a mean correct classification rate of 71.3 ± 6.2%. Investigation of the model parameters suggested that five variables only, namely, P, Ti, Zn (higher in Coratina), Fe, and Ni (more concentrated in Frantoio), significantly contributed to the discriminant model.

Lastly, the possibility of discriminating whether the oil was organically produced or not was also attempted, but the classification model resulted in a very poor accuracy (close to 50%) suggesting that, at least for the investigated samples, organic cultivation has little impact on the elemental composition with respect to non-organic production.

4 CONCLUSION

This study showed that the As, Cu, Fe, and Pb levels in the analyzed samples were far below the MRLs, which certifies the high quality of Italian EVOO.

The element concentrations allow to distinguish well some geographical origins of the EVOO samples and also, although slightly less well, the two cultivars Coratina and Frantoio. On the other hand, given the high heterogeneity of the data set, it is not possible to distinguish organic oils from non-organic ones. This is

probably due to the fact that within the two classes the variability related to geographical origin and cultivar is added.

This study can be used to create datasets for element levels in EVOOs for each production region to support geographic origin authentication. In the future, other information will have to be considered together with the elemental profile of EVOO such as climatic factors and bioavailable fraction of the total content of elements to further corroborate the use of the elements as a marker of provenance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ML-A: conceptualization, investigation, methodology, validation, formal analysis, visualization, data curation, writing—original draft preparation, writing—reviewing and editing and supervision. FM: methodology, validation, formal analysis, writing—original draft preparation and writing—reviewing and editing. MA-F: investigation. LM: formal analysis. ALC: resources. CMM: resources. SC: resources and supervision.

FUNDING

This work was partially supported by the "Agroalimentare e Ricerca" (AGER) program, project AGER2-Rif.20160169, "Valorization of Italian Olive products through INnovative analytical tools–VIOLIN".

ACKNOWLEDGMENTS

We thank Dr. Elisabetta Marconi, and Dr. Giulia Vitiello for their excellent support in the treatment and classification of the samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2021.769620/full#supplementary-material

REFERENCES

Aceto, M., Calà, E., Musso, D., Regalli, N., and Oddone, M. (2019). A preliminary study on the authentication and traceability of extra virgin olive oil made from Taggiasca olives by means of trace and ultra-trace elements distribution. *Food Chemistry* 298, 125047. doi:10.1016/j.foodchem.2019.125047

Astolfi, M. L., Conti, M. E., Marconi, E., Massimi, L., and Canepari, S. (2020a). Effectiveness of Different Sample Treatments for the Elemental Characterization of Bees and Beehive Products. *Molecules* 25, 4263. doi:10.3390/molecules25184263

Astolfi, M. L., Marconi, E., Protano, C., and Canepari, S. (2020b). Comparative elemental analysis of dairy milk and plant-based milk alternatives. *Food Control* 116, 107327. doi:10.1016/j.foodcont.2020.107327

Astolfi, M. L., Marconi, E., Vitiello, G., and Massimi, L. (2021b). An optimized method for sample preparation and elemental analysis of extra-virgin olive oil by inductively coupled plasma mass spectrometry. *Food Chemistry* 360, 130027. doi:10.1016/j.foodchem.2021.130027

Astolfi, M. L., Marotta, D., Cammalleri, V., Marconi, E., Antonucci, A., Avino, P., et al. (2021a). Determination of 40 Elements in Powdered Infant Formulas and Related Risk Assessment. *Ijerph* 18, 5073. doi:10.3390/ijerph18105073

Astolfi, M. L., Protano, C., Marconi, E., Massimi, L., Brunori, M., Piamonti, D., et al. (2020c). A new rapid treatment of human hair for elemental determination by inductively coupled mass spectrometry. *Anal. Methods* 12, 1906–1918. doi:10.1039/c9ay01871a

Bajoub, A., Bendini, A., Fernández-Gutiérrez, A., and Carrasco-Pancorbo, A. (2018). Olive oil authentication: A comparative analysis of regulatory frameworks with especial emphasis on quality and authenticity indices, and recent analytical techniques developed for their assessment. A review. *Crit. Rev. Food Sci. Nutr.* 58, 832–857. doi:10.1080/10408398.2016.1225666

Bakircioglu, D., Kurtulus, Y. B., and Yurtsever, S. (2013). Comparison of extraction induced by emulsion breaking, ultrasonic extraction and wet digestion procedures for determination of metals in edible oil samples in Turkey using ICP-OES. *Food Chemistry* 138, 770–775. doi:10.1016/j.foodchem.2012.10.089

Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. *J. Chemometrics* 17, 166–173. doi:10.1002/cem.785

Beltrán, M., Sánchez-Astudillo, M., Aparicio, R., and García-González, D. L. (2015). Geographical traceability of virgin olive oils from south-western Spain by their multi-elemental composition. *Food Chemistry* 169, 350–357. doi:10.1016/j.foodchem.2014.07.104

Bendini, A., Cerretani, L., Carrasco-Pancorbo, A., Gómez-Caravaca, A., Segura-Carretero, A., Fernández-Gutiérrez, A., et al. (2007). Phenolic Molecules in Virgin Olive Oils: a Survey of Their Sensory Properties, Health Effects, Antioxidant Activity and Analytical Methods. An Overview of the Last Decade Alessandra. *Molecules* 12 (8), 1679–1719. doi:10.3390/12081679

Benincasa, C., Lewis, J., Perri, E., Sindona, G., and Tagarelli, A. (2007). Determination of trace element in Italian virgin olive oils and their characterization according to geographical origin by statistical analysis. *Analytica Chim. Acta* 585, 366–370. doi:10.1016/j.aca.2006.12.040

Cabrera-Vique, C., Bouzas, P. R., and Oliveras-López, M. J. (2012). Determination of trace elements in extra virgin olive oils: A pilot study on the geographical characterisation. *Food Chemistry* 134 (1), 434–439. doi:10.1016/j.foodchem.2012.02.088

Camin, F., Larcher, R., Perini, M., Bontempo, L., Bertoldi, D., Gagliano, G., et al. (2010). Characterisation of authentic Italian extra-virgin olive oils by stable isotope ratios of C, O and H and mineral composition. *Food Chemistry* 118, 901–909. doi:10.1016/j.foodchem.2008.04.059

Capriotti, A. L., Cavaliere, C., Crescenzi, C., Foglia, P., Nescatelli, R., Samperi, R., et al. (2014). Comparison of extraction methods for the identification and quantification of polyphenols in virgin olive oil by ultra-HPLC-QToF mass spectrometry. *Food Chemistry* 158, 392–400. doi:10.1016/j.foodchem.2014.02.130

Carbone, A., Cacchiarelli, L., and Sabbatini, V. (2018). Exploring quality and its value in the Italian olive oil market: a panel data analysis. *Agric. Econ.* 6, 6. doi:10.1186/s40100-018-0102-8

Chatzistathis, T., Therios, I., and Alifragis, D. (2009). Differential uptake, distribution within tissues, and use efficiency of manganese, iron, and zinc by olive cultivars Kothreiki and Koroneiki. *horts* 44, 1994–1999. doi:10.21273/HORTSCI.44.7.1994

Choe, E., and Min, D. B. (2006). Mechanisms and factors for edible oil oxidation. *Comp. Rev. Food Sci. Food Saf.* 5, 169–186. doi:10.1111/j.1541-4337.2006.00009.x

Christodouleas, D. C., Fotakis, C., Nikokavoura, A., Papadopoulos, K., and Calokerinos, A. C. (2015). Modified DPPH and ABTS assays to assess the antioxidant profile of untreated oils. *Food Anal. Methods* 8 (5), 1294–1302. doi:10.1007/s12161-014-0005-6

Cioffi, G., Pesca, M. S., De Caprariis, P., Braca, A., Severino, L., and De Tommasi, N. (2010). Phenolic compounds in olive oil and olive pomace from Cilento (Campania, Italy) and their antioxidant activity. *Food Chemistry* 121 (1), 105–111. doi:10.1016/j.foodchem.2009.12.013

Codex Stan 33-1981 (2021). *Standard for Olive Oils and Olive Pomace Oils. Adopted in 1981Revision: 1989, 2003, 2015*. Amendment: 2009, 2013. URL: http://www.fao.org/input/download/standards/88/CXS_033e_2015.pdf (accessed on April, 2021).

Cordella, C., Moussa, I., Martel, A.-C., Sbirrazzuoli, N., and Lizzani-Cuvelier, L. (2002). Recent developments in food characterization and adulteration detection: Technique-oriented perspectives. *J. Agric. Food Chem.* 50, 1751–1764. doi:10.1021/jf011096z

Damak, F., Asano, M., Baba, K., Suda, A., Araoka, D., Wali, A., et al. (2019). Interregional traceability of Tunisian olive oils to the provenance soil by multielemental fingerprinting and chemometrics. *Food Chemistry* 283, 656–664. doi:10.1016/j.foodchem.2019.01.082

Dugo, L., Russo, M., Cacciola, F., Mandolfino, F., Salafia, F., Vilmercati, A., et al. (2020). Determination of the phenol and tocopherol content in italian high-quality extra-virgin olive oils by using LC-MS and multivariate data analysis. *Food Anal. Methods* 13, 1027–1041. doi:10.1007/s12161-020-01721-7

European Union (EU) (20122012). *The Commission Regulation No. 1151/2012 of 21 November 2012 on quality schemes for agricultural products and foodstuffs OJ L 343/1, 14.12*. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012R1151&from=EN (accessed on August, 2021).

Eurostat (2019). Olive trees - Area by age and density classes (area in ha). Available online: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=orch_olives3&lang=en (accessed on August, 2021).

Farmaki, E. G., Thomaidis, N. S., Minioti, K. S., Ioannou, E., Georgiou, C. A., and Efstathiou, C. E. (2012). Geographical characterization of Greek olive oils using rare earth elements content and supervised chemometric techniques. *Anal. Lett.* 45, 920–932. doi:10.1080/00032719.2012.655656

Filzmoser, P., Hron, K., and Reimann, C. (2009). Principal Component Analysis for Compositional Data With Outliers. *Environmetrics: The Official Journal of the International Environmetrics Society* 20 (6), 621–632. doi:10.1002/env.966

Frankel, E. N. (1985). "Chemistry of autoxidation: mechanism, products and flavor significance," in *Flavor Chemistry of Fats and Oils*. Editors D. B. Min and T. H. Smouse (Champaign, IL (USA): AOCS Press), 1–37.

Frezzini, M. A., Castellani, F., De Francesco, N., Ristorini, M., and Canepari, S. (2019). Application of DPPH assay for assessment of particulate matter reducing properties. *Atmosphere* 10 (12), 816. doi:10.3390/atmos10120816

Giaccio, M., and Vicentini, A. (2008). Determination of the geographical origin of wines by means of the mineral content and the stable isotope ratios: A review. *J. Comm. Sci. Technol. Qual.* 47 (I–IV), 267–284.

Giacomo, D., Pera Lara, L., Daniele, G., Francesco, S., and Turco Vincenzo, L. (2004). Influence of the olive variety and the zone of provenience on selenium content determined by cathodic stripping potentiometry (CSP) in virgin olive oils. *Food Chemistry* 88, 135–140. doi:10.1016/j.foodchem.2003.12.036

Gonzalez-Fernandez, I., Iglesias-Otero, M. A., Esteki, M., Moldes, O. A., Mejuto, J. C., and Simal-Gandara, J. (2019). A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. *Crit. Rev. Food Sci. Nutr.* 59 (12), 1913–1926. doi:10.1080/10408398.2018.1433628

Gumus, Z. P., Celenk, V. U., Tekin, S., Yurdakul, O., and Ertas, H. (2017). Determination of trace elements and stable carbon isotope ratios in virgin olive oils from Western Turkey to authenticate geographical origin with a chemometric approach. *Eur. Food Res. Technol.* 243, 1719–1727. doi:10.1007/s00217-017-2876-4

Hannachi, H., and Elfalleh, W. (2020). Enrichment of olive oil with polyphenols from oleaster leaves using central composite design for the experimental measurements. *Anal. Lett.* 54, 590–607. doi:10.1080/00032719.2020.1774599

International Olive Council (IOC) (2018a). IOC data for the 2017/18 crop year show a year-on-year increase in the production of olive oil. Available online: http://www.internationaloliveoil.org/news/view/698-year-2018-news/1049-ioc-data-for-the-2017-18-crop-year-show-a-year-on-year-increase-in-the-production-of-olive-oil (accessed on August, 2021).

International Olive Council (IOC) (2019). *Trade standard applying to olive oils and olive-pomace oils*. COI/T.15/NC No. 3/Rev. 15. Available online: https://www.internationaloliveoil.org/wp-content/uploads/2020/07/Trade-standard-T15-NC3-Rev15-EN.pdf (accessed on April, 2021).

International Olive Council (IOC) (2018b). World Olive Oil Figures. Available online: http://www.internationaloliveoil.org/estaticos/view/131-world-olive-oilfigures (accessed on August, 2021).

Kabata-Pendias, A. (2010). *Trace elements in soils and plants*. 4th ed. Boca Raton: CRC Press. doi:10.1201/b10158

Kedare, S. B., and Singh, R. P. (2011). Genesis and development of DPPH method of antioxidant assay. *J. Food Sci. Technol.* 48 (4), 412–422. doi:10.1007/s13197-011-0251-1

Kelly, S., Heaton, K., and Hoogewerff, J. (2005). Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends Food Sci. Tech.* 16, 555–567. doi:10.1016/j.tifs.2005.08.008

Lepri, F. G., Chaves, E. S., Vieira, M. A., Ribeiro, A. S., Curtius, A. J., DeOliveira, L. C. C., et al. (2011). Determination of Trace Elements in Vegetable Oils and Biodiesel by Atomic Spectrometric Techniques-A Review. *Appl. Spectrosc. Rev.* 46, 175–206. doi:10.1080/05704928.2010.529628

Llorent-Martínez, E. J., Fernández-de Córdova, M. L., Ortega-Barrales, P., and Ruiz-Medina, A. (2014). Quantitation of Metals During the Extraction of Virgin Olive Oil From Olives Using ICP-MS After Microwave-Assisted Acid Digestion. *J. Am. Oil Chem. Soc.* 91, 1823–1830. doi:10.1007/s11746-014-2511-5

Llorent-Martínez, E. J., Ortega-Barrales, P., Fernández-de Córdova, M. L., and Ruiz-Medina, A. (2011). Analysis of the legislated metals in different categories of olive and olive-pomace oils. *Food Control* 22, 221–225. doi:10.1016/j.foodcont.2010.07.002

Luka, M. F., and Akun, E. (2019). Investigation of trace metals in different varieties of olive oils from northern Cyprus and their variation in accumulation using ICP-MS and multivariate techniques. *Environ. Earth Sci.* 78, 578. doi:10.1007/s12665-019-8581-9

Mendil, D., Uluözlü, Ö. D., Tüzen, M., and Soylak, M. (2009). Investigation of the levels of some element in edible oil samples produced in Turkey by atomic absorption spectrometry. *J. Hazard. Mater.* 165, 724–728. doi:10.1016/j.jhazmat.2008.10.046

Pérez, N. F., Ferré, J., and Boqué, R. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometrics Intell. Lab. Syst.* 95, 122–128. doi:10.1016/j.chemolab.2008.09.005

Pérez-Jiménez, J., Arranz, S., Tabernero, M., Díaz- Rubio, M. E., Serrano, J., Goñi, I., et al. (2008). Updated methodology to determine antioxidant capacity in plant foods, oils and beverages: Extraction, measurement and expression of results. *Food Res. Int.* 41 (3), 274–285. doi:10.1016/j.foodres.2007.12.004

Perna, A., Simonetti, A., Intaglietta, I., Sofo, A., and Gambacorta, E. (2012). Metal content of southern Italy honey of different botanical origins and its correlation with polyphenol content and antioxidant activity. *Int. J. Food Sci. Technol.* 47, 1909–1917. doi:10.1111/j.1365-2621.2012.03050.x

Pohl, W. L. (2011). *Economic geology: Principles and practice*. New Jersey: Wiley-Blackwell. 978-1-444-33663-4. doi:10.1002/9781444394870

Pošćić, F., Furdek Turk, M., Bačić, N., Mikac, N., Bertoldi, D., Camin, F., et al. (2019). Removal of pomace residues is critical in quantification of element concentrations in extra virgin olive oil. *J. Food Compost. Anal.* 77, 39–46. doi:10.1016/j.jfca.2019.01.002

Russo, G., Beritognolo, I., Bufacchi, M., Stanzione, V., Pisanelli, A., Ciolfi, M., et al. (2020). Advances in biocultural geography of olive tree (Olea europaea L.) landscapes by merging biological and historical assays. *Sci. Rep.* 10, 7673. doi:10.1038/s41598-020-64063-8

Šarolić, M., Gugić, M., Tuberoso, C., Jerković, I., Šuste, M., Marijanović, Z., et al. (2014). Volatile Profile, Phytochemicals and Antioxidant Activity of Virgin Olive Oils from Croatian Autochthonous Varieties Mašnjača and Krvavica in Comparison with Italian Variety Leccino. *Molecules* 19 (1), 881–895. doi:10.3390/molecules19010881

Shah, N. S., and Soylak, M. (2021). Advanced methodologies for trace elements in edible oil samples: a review. *Crit. Rev. Anal. Chemistry* 15, 1–20. doi:10.1080/10408347.2021.1895710

Ståhle, L., and Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemometrics* 1, 185–196. doi:10.1002/cem.1180010306

Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K., and Sutton, D. J. (2012). Heavy Metal Toxicity and the Environment. *Exp. Suppl.* 101, 133–164. doi:10.1007/978-3-7643-8340-4_6

Thiruvengadam, M., Ghimire, B. K., Kim, S.-H., Yu, C. Y., Oh, D.-H., Chelliah, R., et al. (2020). Assessment of mineral and phenolic profiles and their association with the antioxidant, cytotoxic effect, and antimicrobial potential of Lycium chinense Miller. *Plants* 9 (8), 1023. doi:10.3390/plants9081023

Trindade, A. S., Dantas, A. F., Lima, D. C., Ferreira, S .L., and Teixeira, L. S. (2015). Multivariate Optimization of Ultrasound-Assisted Extraction for Determination of Cu, Fe, Ni, and Zn in Vegetable Oils by High-Resolution Continuum Source Atomic Absorption Spectrometry. *Food Chem.* 15 (185), 145–150. doi:10.1016/j.foodchem.2015.03.118

Visioli, F., Bellomo, G., and Galli, C. (1998). Free radical-scavenging properties of olive oil polyphenols. *Biochem. Biophysical Res. Commun.* 247 (1), 60–64. doi:10.1006/bbrc.1998.8735

Wold, S., Martens, H., and Wold, H. (1983). "The multivariate calibration problem in chemistry solved by the PLS method," in *Matrix Pencils. Lecture Notes in Mathematics*. Editors B. Kågström and A. Ruhe. 1st ed. (Germany: Springer, Berlin/Heidelberg), 973, 286–293. doi:10.1007/BFb0062108

Yaşar, S. B., Baran, E. K., and Alkan, M. (2012). "Metal determinations in olive oil, Olive oil - Constituents, quality, health properties and bioconversions," in *InTech*. Editor Dr. Dimitrios Boskou. Available online: http://www.intechopen.com/books/olive-oil-constituents-quality-healthproperties-and-bioconversions/metal-determinations-in-olive-oil (accessed on July, 2021).

Zaroual, H., Chénè, C., El Hadrami, E. M., and Karoui, R. (2021). Application of new emerging techniques in combination with classical methods for the determination of the quality and authenticity of olive oil: a review. *Crit. Rev. Food Sci. Nutr.* 1, 1–24. doi:10.1080/10408398.2021.1876624

Zeiner, M., Juranovic-Cindric, I., and Škevin, D. (2010). Characterization of extra virgin olive oils derived from the Croatian cultivar Oblica. *Eur. J. Lipid Sci. Technol.* 112, 1248–1252. doi:10.1002/ejlt.201000006

Zeiner, M., Steffan, I., and Cindric, I. J. (2005). Determination of trace elements in olive oil by ICP-AES and ETA-AAS: A pilot study on the geographical characterization. *Microchemical J.* 81, 171–176. doi:10.1016/j.microc.2004.12.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Chemistry

REVIEW
published: 14 January 2022
doi: 10.3389/fchem.2021.818461

Electrochemical Sensors and Biosensors for the Analysis of Tea Components: A Bibliometric Review

Jinhua Shao, Chao Wang, Yiling Shen, Jinlei Shi and Dongqing Ding*

School of Chemistry and Bioengineering, Hunan University of Science and Engineering, Yongzhou, China

Tea is a popular beverage all around the world. Tea composition, quality monitoring, and tea identification have all been the subject of extensive research due to concerns about the nutritional value and safety of tea intake. In the last 2 decades, research into tea employing electrochemical biosensing technologies has received a lot of interest. Despite the fact that electrochemical biosensing is not yet the most widely utilized approach for tea analysis, it has emerged as a promising technology due to its high sensitivity, speed, and low cost. Through bibliometric analysis, we give a systematic survey of the literature on electrochemical analysis of tea from 1994 to 2021 in this study. Electrochemical analysis in the study of tea can be split into three distinct stages, according to the bibliometric analysis. After chromatographic separation of materials, electrochemical techniques were initially used only as a detection tool. Many key components of tea, including as tea polyphenols, gallic acid, caffeic acid, and others, have electrochemical activity, and their electrochemical behavior is being investigated. High-performance electrochemical sensors have steadily become a hot research issue as materials science, particularly nanomaterials, and has progressed. This review not only highlights these processes, but also analyzes and contrasts the relevant literature. This evaluation also provides future views in this area based on the bibliometric findings.

Keywords: electrochemical sensor, tea, antioxidant, caffeic acid, gallic acid, tea polyphenols, analytical chemistry

OPEN ACCESS

Edited by:
Federico Marini,
Sapienza University of Rome, Italy

Reviewed by:
Masoumeh Ghalkhani,
Shahid Rajaee Teacher Training University, Iran
Mani Govindasamy,
National Taipei University of Technology, Taiwan

****Correspondence:***
Jinhua Shao
hnkjxysjh@huse.edu.cn

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *19 November 2021*
Accepted: *28 December 2021*
Published: *14 January 2022*

Citation:
Shao J, Wang C, Shen Y, Shi J and Ding D (2022) Electrochemical Sensors and Biosensors for the Analysis of Tea Components: A Bibliometric Review. Front. Chem. 9:818461. doi: 10.3389/fchem.2021.818461

INTRODUCTION

Tea is one of the most popular natural health drinks and is deeply ingrained in people's lives. Tea trees are grown in around 30 nations throughout the world, yielding roughly 2.5 million tons of tea every year. The introduction of tea trees, on the other hand, can be easily influenced by soil and climate conditions (Feng et al., 2010; Wambu et al., 2017; Kottawa-Arachchi et al., 2019; Beringer et al., 2020). Furthermore, the nutritional composition and taste of tea fluctuate significantly due to variances in processing processes. Tea leaves are classified in a variety of ways based on various factors (Liu et al., 2015). Green tea, yellow tea, white tea, oolong tea, black tea, and dark tea are the most often used criterion for categorizing tea based on the degree of fermentation (de Carvalho Couto et al., 2021). Green tea is produced without the use of fermentation. Yellow tea is fermented to a degree of 10–20 percent. White tea is fermented to a percentage of 10–30%. Oolong tea is fermented to a degree of 20–60%. Black tea is fermented to a degree of 80–90 percent. Dark tea is a post-fermented tea, meaning it has undergone the most fermentation. The material components in tea leaves are changed into diverse forms as a result of varying degrees of fermentation (Zheng et al., 2016; Marx et al., 2017; Seth et al., 2019). Unfermented green tea, for example, preserves more of the

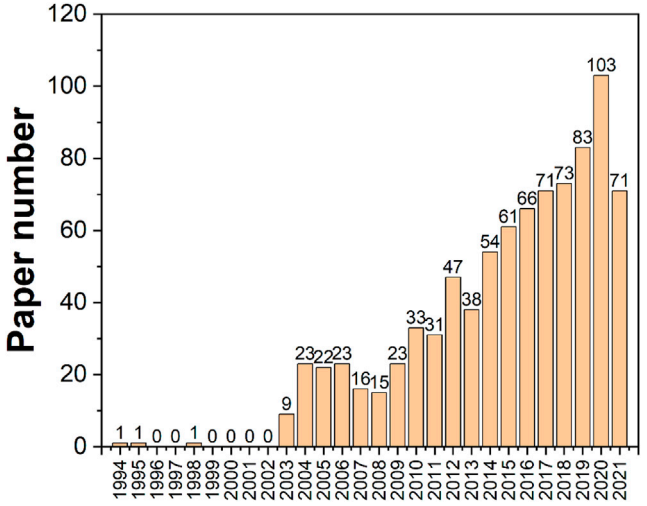

FIGURE 1 | Statistical histogram of research on tea using electrochemistry from 1994 to 2021.

natural components of the fresh leaves. These nutrients offer unique therapeutic properties in the human body, including anti-aging, anti-cancer, anti-inflammatory, and antiseptic properties (Kochman et al., 2021). Oolong tea processing, on the other hand, employs alternating mechanical force and stacking (Ng et al., 2018). The external force damages the cellular tissue of the leaf edge, while the polyphenols are oxidized and undergo other chemical changes as a result of the mechanical activity.

Tea leaves contain about 700 chemicals that have been extracted and identified. The secondary metabolic components of tea, such as tea polyphenols, amino acids, alkaloids, aromatic chemicals, pigment molecules, and so on, are primarily responsible for its distinctive flavor (Chapagain and Hoekstra, 2007; McCants, 2008; Tanui et al., 2012; Mzembe et al., 2016; Wang et al., 2020; Ying et al., 2020; Zhang et al., 2020; Zhou et al., 2020). They have a close association with specific pharmacological effects and impact the quality and flavor of tea (Koch et al., 2018; Wei et al., 2018). Based on the characteristics of the components, tea component analysis may be separated into two primary types: flavor component and quality component (Xu et al., 2019a). The flavor component is linked to the color, aroma, and taste of tea, and its detection focuses on determining the flavor qualities of tea quality (Xu et al., 2019b). The detection of the quality component is primarily for quality control and inspection purposes (Wang et al., 2019). The most often utilized techniques are colorimetric and spectroscopic approaches (Zhi et al., 2017). However, in the recent decade, the rapid development of electrochemical sensing techniques has enticed many scientists to experiment with electrochemical biosensing approaches to assess tea components. High sensitivity, wide linear response, outstanding stability, and reproducibility are all advantages of electrochemical sensors. Furthermore, the low cost of electrochemical measurements is a significant advantage. Electrochemical sensors consist of an electrochemical cell with at least two electrodes to form a closed electrical circuit and a transducer where charge transport (always electronic) takes place, whereas charge transport in the analyte sample can be electronic, ionic, or mixed (Power et al., 2018; Karimi-Maleh et al., 2021a). Electroanalytical methods are a class of analytical chemistry techniques that measure the potential (volts) and/or current (amperes) in an electrochemical cell containing the analyte to investigate it (Naveen et al., 2017; Karimi-Maleh et al., 2021b; Karimi-Maleh et al., 2021c; Karimi-Maleh et al., 2021d; Mani et al., 2021; Karimi-Maleh et al., 2022). Two important electrochemical methods used in the parameter evaluation of tea products are voltammetry, which measures current as potential varies, and electronic sensing, which includes the electronic nose (E-nose), electronic tongue (E-tongue), and electronic eye (E-eye). ASV, CV, SWV, and staircase voltammetry, in particular, use various types of electrodes, such as inert carbon electrodes, glassy carbon electrodes, and paraffin-impregnated graphite electrodes, to help determine the quality of tea products. We searched the WOS core database using the keywords *electrochemistry* and *tea* and discovered a total of 865 papers dedicated to the study of tea using electrochemistry from 1994 to 2021, as shown in **Figure 1**. It's worth noting that similar research has gotten a lot of attention in the recent decade.

We conducted a bibliometric analysis of these 865 studies to see if electrochemical biosensing may replace existing analytical approaches in tea quality control. For the bibliometric analysis and visual presentation, CiteSpace was employed (Chen, 2004; Chen, 2006; Chen and Song, 2019). The analysis report compares electrochemical techniques for tea analysis in terms of procedural and thematic changes. We also kept an eye on the cutting-edge of electrochemical sensing technologies in the field of tea analysis. We included both the analysis and data comparison of specific publications in traditional review writing, in addition to the overall pulse of bibliometric analysis. The use of a mix of electrochemical biosensing techniques and nanomaterial technologies was highlighted in particular. The most representative of these pieces was also thoroughly examined.

LITERATURE INFORMATION ANALYSIS

Changes in the Literature Category of Electrochemical Sensing Technology for Tea Analysis

Changes in electrical signals are the basis for signal production in electrochemical biosensing systems. Changes in current and resistance can be used as signals. Electrochemical oxidation-reduction is present in the bulk of these signal alterations. Under a result, the vast majority of tea research using electrochemical sensing technologies is classified as chemistry, electrochemistry, or analytical chemistry (**Figure 2A**). It's important to note that each published paper does not fall into a single category. As a result, understanding other categories can assist in determining the target themes and areas to cross. What was unexpected, as indicated in **Figure 2A**, was the importance of materials science in this area. This is because the production of

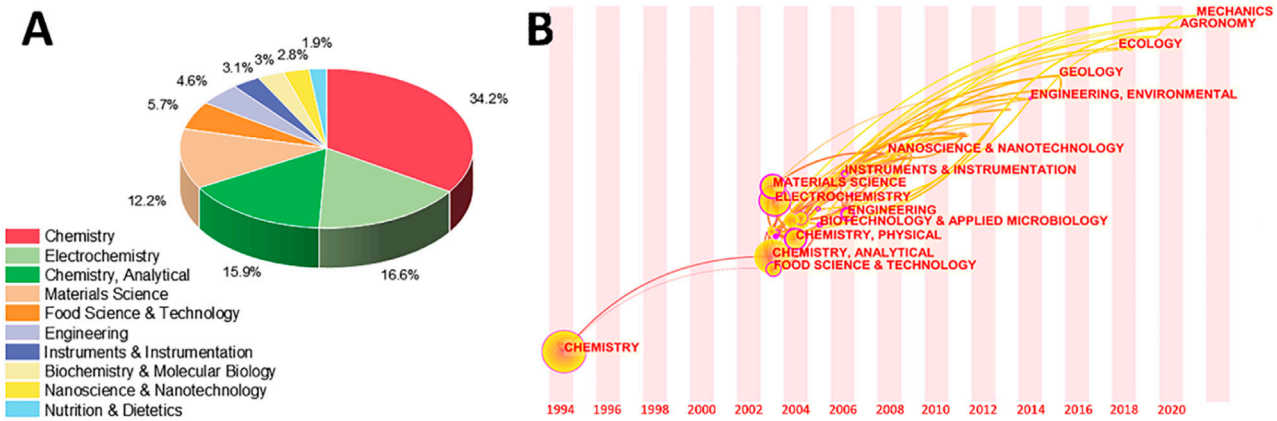

FIGURE 2 | (A) The amount of contribution of electrochemical sensing techniques for tea research in different categories. **(B)** Variation of category with year in the literature of electrochemical sensing techniques for tea analysis.

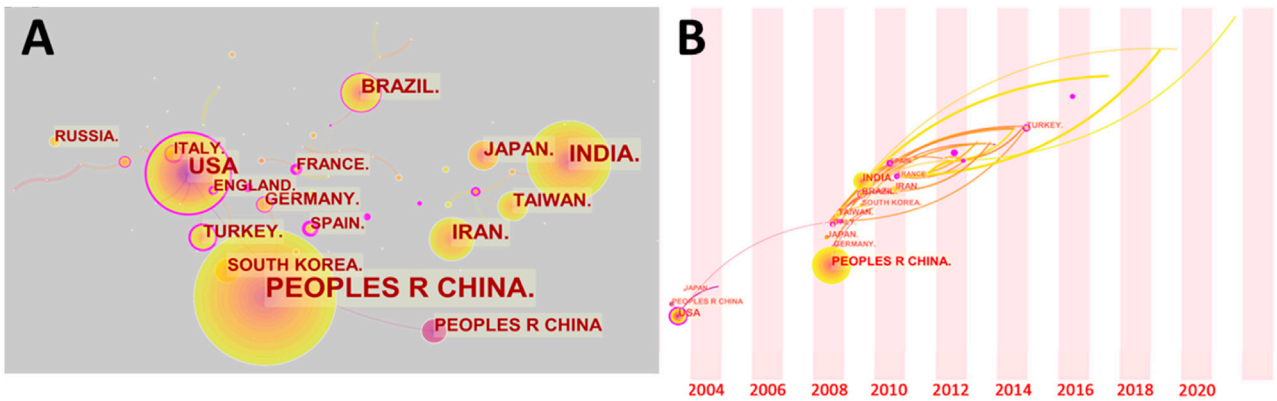

FIGURE 3 | (A) Author's country collaboration network map. **(B)** Time of involvement of different countries in electrochemical sensing research on tea.

very sensitive electrochemical biosensors frequently necessitates the use of novel materials, according to a later study for the literature (Guth et al., 2009; Wang et al., 2015a; Alqarni et al., 2020; Ranga et al., 2021). This also helps to explain why instruments and instrumentation account for 4.6 percent of all category statistics. Similarly, nanotechnology and nanotechnologies accounted for 2.8 percent of the total. Traditional fields of tea research include food science and technology, biochemistry and molecular biology, and nutrition and dietetics. The development of electrochemical sensing technology has clearly been employed for tea research in various areas, as evidenced by this pie chart.

Figure 2B depicts a history of when these categories first appeared on this subject and how they interacted. The most significant categories arrived between 2003 and 2008, with the exception of chemistry, which debuted in 1994. This shows that the electrochemical sensing analysis approaches for tea research surge began in 2003, which is consistent with the results in **Figure 1**. The illustration also highlights the current multidisciplinary interaction of electrochemical analytical tools for tea research with ecology, geology, and agronomy. This indicates that the technology discussed in this article is well-established and has application potential, allowing it to be expanded into other areas for research. Abhradip and Chandan (Pal and Das, 2020a), for example, used electrochemical sensors to assess the solid tea waste extract's ability to inhibit boiler quality steel under acidic circumstances.

Author Country Distribution and Cooperation

Despite the fact that the tea tree is a widely produced plant, tea consumption in various countries has been influenced by cultural factors. As a result, in the scientific study of tea, there is relative independence between countries. Although many nations have actively investigated the development of electrochemical sensing technology in tea, as illustrated in **Figure 3A**, international

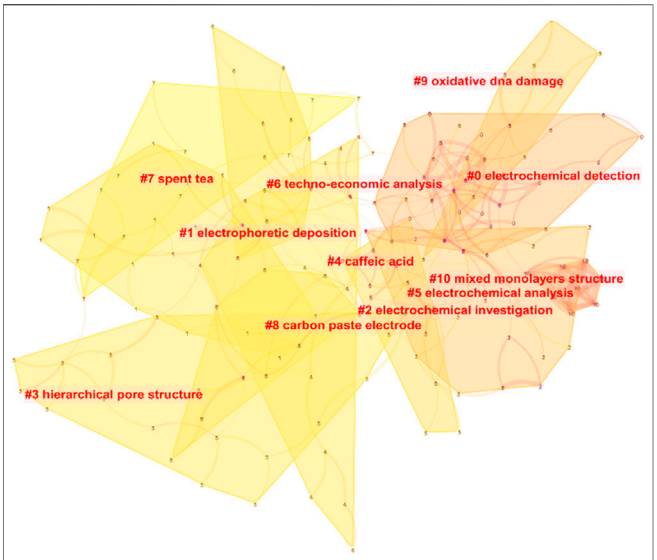

FIGURE 4 | Cluster analysis of research content of electrochemical sensing for tea.

cooperation is rare. Only a small percentage of articles have authors who are from different nations. The United States and the United Kingdom play a significant role in international cooperation on this issue. China and India, on the other hand, the two countries with the highest number of publications on this subject, have focused their research mostly on their own countries. Despite the fact that certain countries are geographically close, such as France and Germany, they continue to perform independent studies on this topic. Surprisingly, despite tea's reputation as an Eastern beverage, the most influential publications in early studies were from the United States. The radical chemistry of epigallocatechin gallate and epigallocatechin was reported by Hagerman et al. (Hagerman et al., 2003). To validate the generation of hydroxyl radicals, the electrochemical redox potentials of both molecules were measured. For phenolic component identification, Luo et al. (Luo et al., 2003) employed liquid chromatography with coulometric electrochemical detection. In this investigation, tea blends were employed as a true sample. **Figure 3B** depicts the timelines of the various countries participated in this research. Many countries got actively involved in research on this topic between 2008 and 2010. Until recently, this topic continues to draw scholars from a variety of countries, who began to participate in the research. Cameroon, the Netherlands, and Finland, for example, have all published studies on this topic in the last 2 years. Dongmo et al. (Dongmo et al., 2020) from the University of Dschang developed an electrochemical biosensor for catechol detection in tea samples utilizing amino-grafting of montmorillonite. University of Yaoundé researchers Deutchoua et al. (Djitieu Deutchoua et al., 2019) devised two electrochemical techniques for determining antioxidant properties. As authentic samples, tea extracts were employed in this study. Overall, research on electrochemical sensing technology in tea is dominated by China, the United States, India, Brazil, Iran, and Japan. They were responsible for more than 70% of the academic papers.

RESEARCH CONTENT ANALYSIS

Cluster Analysis of Research Content

Cluster analysis of the content reveals some of the most important study avenues for this subject. Electrochemical detection, electrophoretic deposition, electrochemical investigation, hierarchical pore structure, caffeic acid, electrochemical analysis, techo-economic analysis, spent tea, carbon paste electrode, oxidative DNA damage, and mixed monolayers structure were among the 11 top themes identified by bibliometric clustering of research on electrochemical sensing for tea (**Figure 4**). The results show that the information on electrochemical procedures, which includes analytical techniques and sensor preparation techniques, is the most important aspect of this topic. Malakootian et al. (Malakootian et al., 2020) used a carbon paste electrode modified with Eu^{3+}-doped NiO to detect Pb (II) and Cd (II) in black tea.

The clustering analysis results included caffeic acid, which is particularly important in tea, in addition to the development of electrochemical techniques and sensors. For the detection of caffeic acid in tea leaves, many of these studies propose an electrochemical sensing device. Chang and colleagues, for example, suggested a ratiometric electrochemical sensor for the detection of caffeic acid (Yin et al., 2021). Caffeic acid is electro-oxidized with two electrons in a diffusion-controlled method. Caffeic acid's hydroxyl groups undergo two-electron transfer and release two protons, resulting in the quick production of the matching quinone. The surface modification presented in this paper can help improve caffeic acid diffusion at the electrode. For caffeic acid detection, Arajo et al. (Araújo et al., 2020a) developed a screen-printed electrode modified with carbon nanotubes (**Figure 5**). However, limited research has been done to assess the antioxidant effects of caffeic acid in tea (Lima et al., 2020). Furthermore, because caffeic acid has substantial electrochemical activity, it has been utilized as a signal in various experiments to demonstrate successful caffeic acid production (Li et al., 2020).

Tea extracts are frequently utilized to examine the consequences of oxidative DNA damage due to their high antioxidant activity. Yury et al. (Kuzin et al., 2016), for example, described a GCE that had been electropolymerized with methylene blue. For analysis, the DNA solution was first combined with an oxidant before being immobilized on the modified GCE. By interrupting the DNA-methylene blue interactions, the voltammetric signal can be utilized to assess the degree of DNA damage (Kuzin et al., 2016). The presence of an antioxidant can help to slow down this process. They put this methodology to the test to see if it could determine the antioxidant capabilities of green tea extract. In addition to voltammetry, impedimetric technology (Kuzin et al., 2015). can be used for a similar purpose. Uliana and colleagues looked into whether tea may preserve DNA from dye-induced damage (Uliana et al., 2014). They proved that the tea solution could prevent adenine and guanine from reacting with the dye

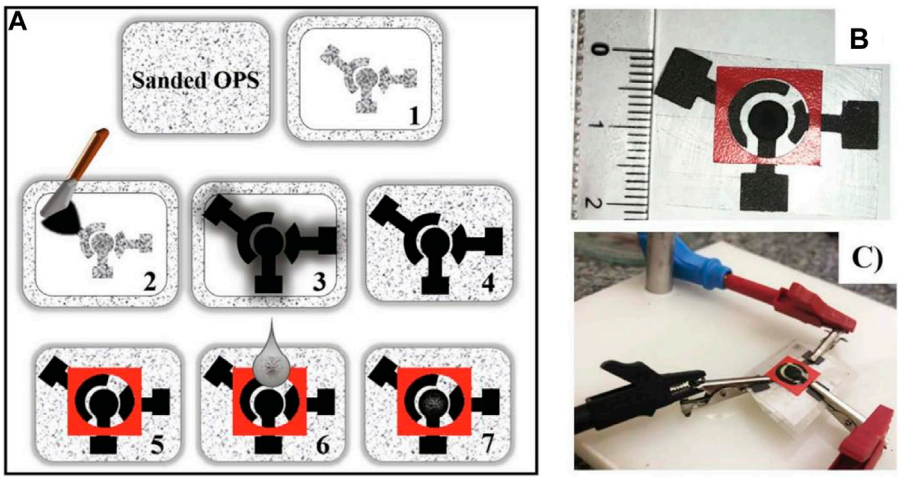

FIGURE 5 | (A) Schematic representation, **(B,C)** images of SPE based caffeic acid biosensor (Araújo et al., 2020a). Copyright: Elsevier B.V.

Top 27 Keywords with the Strongest Citation Bursts

Keywords	Strength	Begin	End	1994 - 2021
catechin	8.29	2003	2009	
hplc	4.48	2003	2012	
performance liquid chromatography	9.92	2004	2012	
liquid chromatography	7.58	2004	2009	
green tea	7.49	2004	2012	
electrochemical detection	7.17	2004	2012	
flavonoid	6.22	2004	2013	
human plasma	3.6	2004	2007	
capillary electrophoresis	3.92	2005	2015	
electrochemical oxidation	4.5	2006	2015	
tea	11.42	2009	2014	
cyclic voltammetry	4.59	2010	2016	
polyphenol	5.32	2011	2013	
adsorption	4.44	2013	2017	
film	6.26	2014	2017	
behavior	4.4	2014	2016	
antioxidant capacity	4.02	2014	2017	
electrochemical behavior	3.7	2014	2018	
electrochemical performance	3.4	2014	2015	
glassy carbon electrode	4.91	2016	2018	
electrochemical sensor	3.45	2016	2017	
graphene oxide	5.75	2017	2019	
uric acid	5.67	2018	2021	
ascorbic acid	4.24	2018	2019	
dopamine	3.39	2018	2021	
graphene	3.54	2019	2021	
gold nanoparticle	3.37	2019	2021	

FIGURE 6 | Top 27 keywords with the strongest bursts of electrochemical sensing techniques for tea analysis.

using electrochemical analysis. The current intensity of the adenine molecule had fallen by 60% of its initial value after an interaction duration of 180 s. Green tea appears to be able to minimize DNA molecule damage, according to the findings. Sumkova and Labuda (Šimková and Labuda, 2009) also suggested an electrochemical biosensor for detecting DNA

damage and integrated the sensor into a commercial flow-through cell. This apparatus has been used to evaluate the antioxidative effects of ta extracts with great effectiveness.

Tea that has been discarded can be recycled as a valuable biological resource. High-temperature carbonization of wasted tea has been used to make electrodes or electrocatalysts in several investigations (Choi et al., 2016; Deng et al., 2016; Ahsan et al., 2020; Gao et al., 2021). In these investigations, electrochemical techniques were utilized as a characterisation tool to assess the performance of biochar. Gao et al. (Gao et al., 2021), for example, produced a biomorphic carbon electrode from discarded black tea and then used it to store potassium ions. Ahsan and colleagues constructed a cobalt-based electrocatalyst using discarded tea leaves as a template and then employed it for hydrogen and oxygen evolution and oxygen reduction (Ahsan et al., 2020).

Keywords Analysis

Keyword analysis can also reveal which research hotspots are being followed. The use of citation burst to analyze keywords might reveal how the study topic's focus has shifted over time. The top 27 keywords with the greatest bursts of electrochemical sensing techniques for tea analysis are shown in **Figure 6**. The term bursts began in 2003, which was the year that additional research findings were published. Catechin, HPLC, performance liquid chromatography, green tea electrochemical detection, and flavonoid are some of the terms used in the beginning. The identification of catechin and flavonoids in tea using liquid chromatography is the main focus here. Electrochemical analysis was used as a technique for identifying samples after chromatographic separation of mixed samples at this point, rather than as a stand-alone sensing technique for tea detection. Long et al. (Hong et al., 2003), for example, suggested a method for detecting natural phenolic compounds in tea using liquid chromatography and multi-channel electrochemical detection. Similarly, Kotani et al. (Kotani et al., 2003) used an HPLC with electrochemical detection to identify catechins. Coulometric detection is frequently the most widely used electrochemical technology when paired with liquid chromatography (Chu et al., 2004; Kotani et al., 2007; Novak et al., 2010a; Shao et al., 2010; Narumi et al., 2014). Electrophoretic (Kartsova and Ganzha, 2006; Kartsova and Alekseeva, 2008) and voltammetric (Novak et al., 2010b; Hocker et al., 2017) approaches had previously been employed in conjunction with chromatographic methods to examine tea.

Around 2010, due to the development of electrochemical analysis techniques, a lot of work started to focus on the study of electrochemical behavior and antioxidant capacity measurement. It is worth noting that the studies of electrochemical behavior are not primarily an investigation of the electrochemical properties of tea components. According to the specific literature revealed, these works mainly focused on the use of extracts of tea leaves as inhibitors for the corrosion protection of metals. Changes in the electrochemical behavior of metals can be used as a characterization of the degree of corrosion. For example, Rauf et al. (Rauf and Mahdi, 2012) evaluated the effects of green tea on corrosion. Electrochemical frequency modulation and cyclic polarization scans were used for characterization. Unfortunately, the results showed that green tea did not show particularly excellent corrosion resistance. However, in a report by Tang et al. (Tang et al., 2018) green tea extract had good anti-corrosion efficacy on carbon steel. Pal and Das (Pal and Das, 2020b) also claimed that solid waste extract from tea factories is an excellent inhibitor.

The preparation of electrochemical sensors for tea detection has been the hottest topic since 2016. GCE has been chosen as the most commonly used commercial electrode. Different nanomaterials have been synthesized for the modification of GCE. Among them, graphene and its derivatives and gold nanoparticles have been studied the most.

Electrochemical Biosensor Performance Comparison

According to the results of the preceding two sub-sections' analyses, the most essential aspect of electrochemical biosensing in the research of tea is the detection of tea components. We used bibliometrics to summarize these specific works. The major characteristics (linear detection range, LDR, and limit of detection, LOD) of various electrochemical sensors for tea composition detection are summarized in **Table 1**.

From **Table 1**, it can be seen that catechol, catechin, caffeine, rutin, gallic acid, quercetin, and tea polyphenols were the most detected tea components by electrochemical sensing technique. Acetamiprid, theophylline, Pb (II) and Cd (II) are the most frequently detected hazardous substances. This is due to the fact that excessive levels of pesticides and heavy metals in tea can lead to food safety problems. Overall, with the development of electrochemical sensing analysis technology, the detection range of different analytes has been enhanced while the detection limits have been reduced. For example, catechol had a detection limit of 0.1 M in the report in 2009 (Lin et al., 2009), which was reduced to 7.34 nM in 2018 (Manavalan et al., 2018). Among these, the widespread use of carbon nanomaterials has proven to be a game-changer. Due to the synergistic role of matrix or composite, carbonaceous materials (Graphene, CNTs, Carbon Nanofibres, and Mesoporous carbon, etc.) and conducting polymer materials act as reliable catalysts with metal oxide nanoparticles for the production of nonenzymatic sensors. It is worth noting that electrochemical sensors are a very sensitive detection technology. The sensors described above already meet the needs of detection, so the pursuit of high sensitivity does not necessarily have practical value. The focus of future electrochemical sensor research should be on how to improve stability and repeatability. Also, miniaturization of electrochemical sensors to fit field detection is an important direction.

KEY AUTHORS AND PAPERS ANALYSIS

Author Co-citation Analysis

Figure 7 shows the relationship network of authors' co-cited information. From this figure, it can be seen that the research

TABLE 1 | Comparison of the performance of electrochemical biosensors for the detection of different tea components.

Biosensor	Analyte	LDR	LOD	Year	Reference
Al/SiO_2/CPE	Catechol	0.5–50 μM	0.1 μM	2009	Lin et al. (2009)
AgNPs/TiO_2/ITO	Catechol	0.1–500 μM	0.05 μM	2012	Wang et al. (2012a)
CNT/carbon paper	Catechol	1–100 μM	0.29 μM	2013	Yue et al. (2013)
Ag doped TiO_2/GCE	Catechol	1–15 μM	0.0249 μM	2016	Ravishankar et al. (2016)
Au@NG-PPy	Catechol	0.1–0.9 μM	0.0016 μM	2017	Vellaichamy et al. (2017)
Gr/GNRs/AgNPs/PPO	Catechol	2–2,300 μM	—	2018	Sandeep et al. (2018)
rGOSs@$SrWO_4$	Catechol	0.034–672.64 μM	7.34 nM	2018	Manavalan et al. (2018)
Banana tissue/CPE	Catechol	1.4–15.7 mg/L	0.1 mg/L	2019	Broli et al. (2019)
Biomimetic oxidase/GO	Catechol	50–1,600 μM	0.09 μM	2021	Jiaojiao et al. (2021)
f-SWCNTs/PEDOTM/GCE	Catechin	0.039–40.84 μM	0.013 μM	2015	Yao et al. (2015)
Pt/MnO_2/f-MWCNT/GCE	Catechin	2–950 μM	0.02 μM	2015	Ezhil Vilian et al. (2015)
(fMWCNT)/YHCF/GCE	Catechin	5–200 μM	0.28 μM	2015	Devadas and Chen, (2015)
N-doped carbon/GCE	Catechin	1–30 μM	0.088 μM	2017	Pang et al. (2017)
MIP	Catechin	5–100 μM	37 nM	2018	Chatterjee et al. (2018)
Cu@g-C_3N_4	Catechin	100–900 μM	15.12 μM	2021	Sanjay et al. (2021)
3DG/MWCNTs-Nc	Caffeic acid	0.2–174 μM	17.8 nM	2017	Sakthinathan et al. (2017)
Pt-PEDOT/rGO	Caffeic acid	5 nM–0.5 μM	2 nM	2018	Gao et al. (2018)
MWCNTs/SPE	Caffeic acid	2–50 μM	0.66 μM	2020	Araújo et al. (2020a)
MWCNT/SPEs	Caffeic acid	2–50 μM	0.2 μM	2020	Araújo et al. (2020b)
PMB@Ni–TPA/GCE	Caffeic acid	0.25–15.0 μM	0.2 μM	2021	Yin et al. (2021)
Poly-aspartic acid	Caffeine	0.25–30 μM	72 nM	2010	Wang et al. (2010)
Nafion/poly (safranine T)/GCE	Caffeine	0.3–100 μM	0.1 μM	2011	Guo et al. (2011)
MIPs/GNPs/MWNTs/GCE	Caffeine	0.5 nM-0.16 μM	90 pM	2012	Kan et al. (2012)
DNA-SWCNT/Nafion/GCE	Caffeine	0.02–1.5 μM	8 nM	2014	Wang et al. (2014)
PDDA-MWCNT	Caffeine	0.3–80 μM	0.05 μM	2017	Zhang et al. (2017a)
Polydopamine-gold	Caffeine	—	—	2017	Zhang et al. (2017b)
ZMWCNTMCPE/SDS/CPE	Caffeine	10–100 μM	75 nM	2019	Azab et al. (2019)
Nafion-NCNTs	Caffeine	0.08–6 μM	20 nM	2019	Wu et al. (2019)
SWCNT-SubPc	Caffeine	0.1–1.5 μM	13 nM	2019	Şenocak et al. (2019)
TiO_2/MIP	Caffeine	5–120 μM	0.6 μM	2020	Das et al. (2020)
Cu-MOF/graphene	Caffeine	5–450 mM	1.38 mM	2021	Venkadesh et al. (2021)
Plasma-triggered polydimethylsiloxane/ITO	Caffeine	50 nM-700 μM	20 nM	2021	Li et al. (2021)
MoO_3-GCNS	Caffeine	0.5–359 μM and 410–810 μM	21.24 nM	2021	Boopathy et al. (2021)
GC/Gr/SiC-NPs/[Cu(pydc) (apym)](2)	Caffeine	—	0.313 μM	2021	Hallaj et al. (2021)
Co_3O_4/GCE-Nafion	Caffeine	—	97 nM	2021	Kumar et al. (2021)
MIP(poly (o-phenylenediamine))	Epigallocatechin-3-gallate	0.5–10 μM	0.16 μM	2013	Duan et al. (2013)
MIP/GO/GC	Epigallocatechin-3-gallate	30 nM-10 μM	8.78 nM	2017	Liu et al. (2017)
$Ni(OH)_2$ NPs	Epigallocatechin-3-gallate	10–100 mM	7 nM	2019	Nandy Chatterjee et al. (2019)
SWCNTs/poly-EB/GCE	Rutin	0.16–20 μM	82 nM	2012	Wang et al. (2012b)
SMWCNT-PEDOT-IL	Rutin	—	77 nM	2016	Nagles and García-Beltrán, (2016)
G-MWCNTs/GCE	Rutin	0.01–1 μM	5 nM	2016	Yang et al. (2016)
PEDOT/M-EDTA	Rutin	—	1.67 nM	2018	Lu et al. (2018)
NiCo2S4/rGO@PANI	Rutin	0.01–200 μM	0.007 μM	2018	Wang et al. (2018)
Polyphenol oxidase-AuNPs-mesoporous carbon	Rutin	1.6–28 mM	0.51 mM	2019	Zhong et al. (2019)
Poly (safranine/nano NiO)CPE	Rutin	16.1–230 nM	5.4 nM	2019	Saritha et al. (2019)
GQDs/PEDOT/GCE	Rutin	0.05–10 μM	11 nM	2019	Meng et al. (2019)
Fe_3O_4@TAPB-DMTP-COFs	Luteolin	0.01–70 μM	7.2 nM	2020	Xie et al. (2020)
MoO_3-PPy NWs/MWCNTs	Luteolin	0.1 nM-10 μM	0.03 nM	2021	Zeng et al. (2021)
MIP	Morin	0.05–1.7 μM	0.01 μM	2016	Liu et al. (2016)
SiO_2/CPE	Pyrogallol	2–300 μM	0.7 μM	2014	Tashkhourian and Ghaderizadeh, (2014)
PEI-rGO/GCE	Gallic acid	0.1–10 mg/L	0.07 mg/L	2013	Luo et al. (2013)
Polyepinephrine/GCE	Gallic acid	1–20 μM	0.663 μM	2013	Abdel-Hamid and Newair, (2013)
SPCE/PME	Gallic acid	—	0.076 μM	2015	Su and Cheng, (2015)
APTS@GO/PPAH-SDS/GCE	Gallic acid	0.006–2000 μM	1.7 nM	2018	Baghayeri et al. (2018)
PLM/MWCNT/GCE	Gallic acid	0.004–1.1 μM and 1.7–20 μM	3.1 nM	2019	Koçak et al. (2019)
Graphene/GCE	Gallic acid	80 nM–2 μM	1.2 nM	2019	Chen et al. (2019)
3D IPCNT/CNS/GCE	Gallic acid	0.05–20 μM	53 nM	2020	Zhao et al. (2020)
NG-Au@Ag NPs	Gallic acid	1–16.2 μM	3.17 nM	2020	Feng et al. (2020)

(Continued on following page)

TABLE 1 | (*Continued*) Comparison of the performance of electrochemical biosensors for the detection of different tea components.

Biosensor	Analyte	LDR	LOD	Year	Reference
Silica gel/CPE	Quercetin	5–100 μg/L	3.53 μg/L	2012	Chen et al. (2012)
Porous alumina microfibers/CPE	Quercetin	0.025–1.5 μM	10 nM	2015	Li and Huang, (2015)
Platinum (II)-porphyrin/GCE	Quercetin	0.002–50 mg/L	0.8 μg/L	2015	Tian et al. (2015)
SWCNT/GCE	Quercetin	0.01–100 mM	7 mM	2019	Kuyumcu Savan, (2020)
GCE	Quercetin	7.9 nM-3.96 μM and 3.96–14.86 μM	2.2 nM	2020	Karaboduk and HASDEMİR, (2020)
Co_3O_4/GCE	Quercetin	0.01–3 mM	70 nM	2021	Khand et al. (2021)
MWCNTs-CS	Tea polyphenols	100–1,000 mg/L	10 mg/L	2009	Guo et al. (2009)
Diazonium-tyrosinase	Tea polyphenols	—	0.1 mM	2010	Cortina-Puig et al. (2010)
Pt NPs-rGO-laccase	Tea polyphenols	0.2–2 μM	2.75 μM	2013	Eremia et al. (2013)
Ferric chloride/GCE	Tea polyphenols	0.192–0.318 mg/L	—	2014	Chattopadhyay and Sarkar, (2014)
Iron phthalocyanine	Tea polyphenols	—	0.176 μM	2016	Maximino et al. (2016)
Chloramine-T/GCE	Tea polyphenols	—	0.674 mg/L	2016	Sen et al. (2015)
Tyrosinase- (Co-1.57 Al(OH) $_{(x)}$)SO_4	Tea polyphenols	Up to 10 μg/ml	0.33 pg/ml	2017	Soussou et al. (2017)
Cassava fiber-iron nanoparticles/spE	Tea polyphenols	3.5–31.5 μM	0.1 μM	2021	Shi et al. (2021)
Cetyltrimethyl ammonium bromide/CPE	Theophylline	0.8–200 μM	0.185 μM	2009	Hegde et al. (2009)
CdSe/GCE	Theophylline	1.0–40 μM and 40–700 μM	0.4 μM	2012	Yin et al. (2012)
ED-GO/GCE	Theophylline	0.8–60 μM	0.01 μM	2013	Cui and Zhang, (2013)
SWCNT-LMC/Nafion/GCE	Theophylline	0.3–38 μM	0.08 μM	2013	Gao and Guo, (2013)
MWNT|MnO_2/GCE	Theophylline	0.1–20 μM	0.01 μM	2015	Yang and Li, (2015)
WS_2/AgNP/GCE	Theophylline	0.05–150 μM	3 nM	2015	Wang et al. (2015b)
AuNP/MWCNT/GCE	Theophylline	0.5–20 μM	90 nM	2018	da Silva et al. (2018)
AFW/Nf/GCE	Theophylline	0.1–160 μM	0.0028 μM	2019	Karthika et al. (2019)
MIP/SL-MoS_2-BOMC/GCE	Theophylline	0.01–50 μM and 50–250 μM	5 nM	2019	Hu et al. (2019)
beta-NiS/Ppy	Theophylline	10 nM-900 μM	1 nM	2019	Muthukumaran et al. (2019)
DMN-AuNPs/GCE	Theophylline	0.05–2.0 μM	9.6 nM	2021	Zhang et al. (2021)
MoS_2/MWCNTs	Carbendazim	0.04–100 μM	7.4 nM	2020	Zhu et al. (2020)
V_2O_5/G-C_3N_4/PVA/GCE	Folic acid	0.01–60 μM	1.74 nM	2020	Karthika et al. (2020)
Polyacrylamide (MIP)/graphite	Flavins	20–100 μM	14 μM	2017	Nandy Chatterjee et al. (2017)
MWNT/GCE	Tannins	0.4–200 μM	0.1 μM	2004	Lü, (2004)
3D-CS/rGO/GCE	Acetamiprid	0.1 pM-0.1 μM	71.2 fM	2020	Yi et al. (2020)
Ag/His-GQD/G	Acetamiprid	0.1 fM-5 pM	0.04 fM	2020	Dan et al. (2020)
SPE-Gr	Sibutramine	2–120 μM	0.3 μM	2019	Lima et al. (2019)
Diamond paste electrode	Pb (II)	10–100 pM	—	2004	(Raluca-Ioana Stefan (2004, 2004)
BioExt/MWCNTs/GCE	Cd (II)	0.05–5 μM	1.01 nM	2020	Incebay et al. (2020)
rGO/Sb/GCE	Pb (II); Cd (II)	0.1–3 μM; 0.1–3 μM	45.5 nM; 70 nM	2020	Nunes et al. (2020)
Eu^{3+} doped NiO/CPE	Pb (II); Cd (II)	0.8–165 μg/L; 0.8–165 μg/L	0.1 μg/L; 0.4 μg/L	2020	Malakootian et al. (2020)
Mn-TiO_2 NTAs	Cd (II)	—	0.01 μM	2020	Jiang et al. (2020)

work of those authors has had an impact on the field. It is worth noting that the work here is not necessarily limited to the study of electrochemical biosensing for tea, but rather exemplifies the type of work that has had a greater impact on the topic. Lee et al.'s (Lee et al., 2002) work on the pharmacokinetics of catechins and epigal-locatechin-3-gallate in human-consumed tea gave early insights into electrochemical studies on tea. Yang et al. (Yang et al., 2001) explored the antioxidant activity of catechins in microsomal lipid peroxidation at an early stage and also influenced the study of electrochemical techniques for tea detection. Studies on antioxidants in tea have also been mainly influenced by Kilmartin et al. (Kilmartin et al., 2001) because their work suggested for the first time that cyclic voltammetry is an excellent technique for the evaluation of antioxidant activity.

In the design and fabrication of electrochemical biosensors, the electrochemical methodology established by Bard and Faulkner (Bard and Faulkner, 2002) is the most important foundation. Wang's textbook on analytical chemistry is also the basis for electrochemical sensor design (Wang, 2000). The study of the electron transfer process of glucose oxi-dase in glucose biosensors has laid the foundation for many subsequent mechanistic studies of biosensors (Liu et al., 2005). Meanwhile, the technique for rutin and quercetin detection in plants proposed by Chen et al. (Chen et al., 2000) was applied to the operation of a biosensor for tea analysis. Similarly, the technique for gallic acid detection proposed by Abdel-Hamid and Newair (Abdel-Hamid and Newair, 2013) also affects many of the later tests for substances in tea. The work conducted by Ziyatdinova et al. (Ziyatdinova et al., 2012) is also instructive for the detection of flavonoids. The discovery of graphene has become a very important material in the assembly of electrochemical sensors. The graphene-based bioenzyme sensors proposed by Yang et al. (Yang et al., 2011) have influenced the study of sensors targeting the detection of tea components. Electrochemical techniques are used in this work not only for the detection of analytes, but also as a method for the synthesis of nanomaterials.

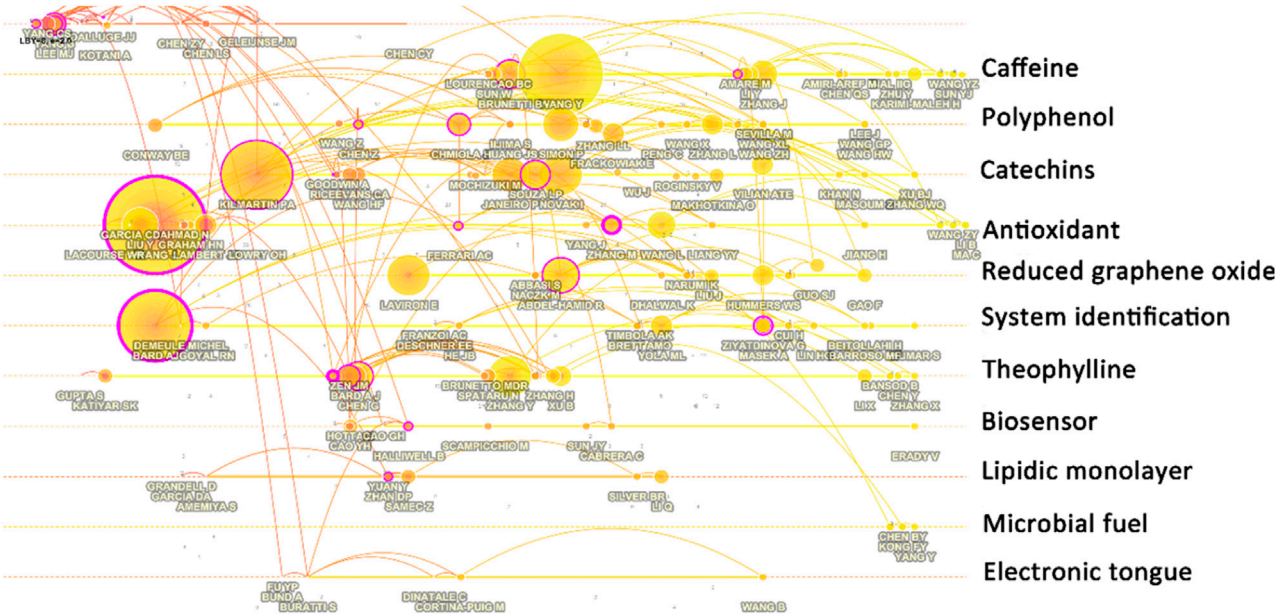

FIGURE 7 | Author co-citation analysis with different research content clusters.

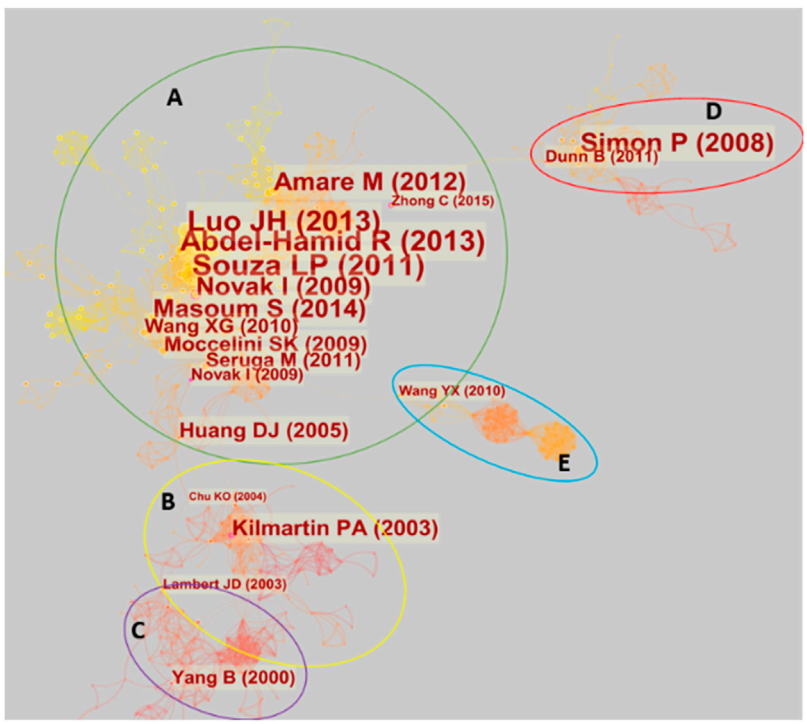

FIGURE 8 | Reference co-citation analysis with five clusters.

Reference Co-citation Analysis

Figure 8 illustrates the literature co-citation analysis relationship graph for electrochemical biosensors in tea analysis. As can be seen from the figure, the relationship between all the literature can be divided into five clusters, one of which contains a very large number of co-cited articles.

Cluster A has a dense network of pairs, representing work that is largely within a broad theme and is very closely linked.

Important papers in this cluster include the work on gallic acid by Abdel-Hamind and Newair (Abdel-Hamid and Newair, 2013), previously mentioned in the authors' co-citation analysis. Luo et al. (Luo et al., 2013) also proposed a method for gallic acid detection. Novak et al. (Novak et al., 2009) reported the electrochemical determination of epicatechin gallate using GCE. Masoum et al. (Masoum et al., 2014), Moccelini et al. (Moccelini et al., 2009) and Wang et al. (Wang et al., 2010) proposed electrochemical methods for catechin detection. Amare and Admassie (Amare and Admassie, 2012) proposed an electrochemical method for caffeine detection. Šeruga et al. (Šeruga et al., 2011) reported the method for polyphenols detection. All these techniques mentioned above are for the analysis of important components in tea.

Behind the detection of these substances is the passion of scientists for antioxidant substances. The explanation of the principle for antioxidant substances reported by Huang et al. (Huang et al., 2005) links cluster A and cluster B. In cluster B, the evaluation of the capacity of antioxidants using cyclic voltammetry reported by Kilmartin et al. (Kilmartin et al., 2001) is one of the most important works. A review of tea and tea polyphenols for cancer chemoprevention (Lambert and Yang, 2003) connects Cluster B and Cluster C. This links electrochemical sensing analysis to the specific health uses of tea components. Cluster C focuses on studies on the electrochemical behavior of tea components and human health, such as the relationship between the electrochemical oxidation of catechins and their antioxidant activity in microsomal lipid peroxidation (Yang et al., 2001).

Clusters D and E are two relatively independent groups. Cluster D focuses on tea as a source of carbon materials and has been used for energy storage. This cluster was accidentally included in the scope of this review because of the many electrochemical characterizations required in energy storage studies. This is a frequent occurrence in bibliometrics in sample statistics due to the sharing of similar keywords across different research directions. Cluster E is about the kinetic study of the electrochemistry of the tetraethylammonium/water interface. Since the abbreviation for tetraethylammonium is TEA, this literature was also accidentally included in the sample for this review.

CONCLUSION AND PERSPECTIVES

This bibliometrics-based review summarizes the progress of electrochemical analysis for tea component sensing. The following conclusions can be drawn based on the focus of research in different time periods:

1) Between 1994 and 2010, electrochemical techniques were often used as a detection step after the separation of samples by chromatographic techniques.
2) After the study of tea components gradually came to our attention, the electrochemical behavior of those components that have electrochemical activity began to be investigated.
3) Since some important components in tea have very pronounced electrochemical redox behavior, electrochemical sensors are starting to become a technique to detect the concentration of these components.
4) As materials science, especially nanomaterials, has become a hot topic, the use of nanomaterials to improve the performance of electrochemical sensors has become the focus of this field. A large number of papers have been published from 2010 onwards.
5) Electrochemical techniques allow not only the detection of specific components in tea but also the evaluation of their antioxidant properties. Therefore, different methodologies based on electrochemical biosensing have been established for measuring the antioxidant properties of tea.

Based on the bibliometric survey of trends in this area, we believe that future direction is likely to focus on the following areas:

1) The use of novel nanomaterial composites, particularly carbon materials and noble metal nanoparticles, will continue to be popular in the design and fabrication of biosensors.
2) Antioxidant property detection biosensors based on DNA ligand technology may become the norm for evaluating tea's antioxidant properties.
3) Because electrochemical sensors allow for rapid assessment of antioxidant properties, this technique can be applied to a wide range of *in vitro* biological experiments.
4) Miniaturization of electrochemical biosensors is an important step toward applying this technology in the field for food detection and quality control.

AUTHOR CONTRIBUTIONS

Conceptualization, JS; writing—original draft preparation, CW, YS, JS, and DD; writing—review and editing, JS; supervision, JS. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Abdel-Hamid, R., and Newair, E. F. (2013). Adsorptive Stripping Voltammetric Determination of Gallic Acid Using an Electrochemical Sensor Based on Polyepinephrine/Glassy Carbon Electrode and its Determination in Black Tea Sample. *J. Electroanalytical Chem.* 704, 32–37. doi:10.1016/j.jelechem.2013.06.006

Ahsan, M. A., Imam, M. A., Santiago, A. R. P., Rodriguez, A., Alvarado-Tenorio, B., Bernal, R., et al. (2020). Spent Tea Leaves Templated Synthesis of Highly Active and Durable Cobalt-Based Trifunctional Versatile Electrocatalysts for Hydrogen and Oxygen Evolution and Oxygen Reduction Reactions. *Green. Chem.* 22, 6967–6980. doi:10.1039/d0gc02155e

Alqarni, S. A., Hussein, M. A., Ganash, A. A., and Khan, A. (2020). Composite Material–Based Conducting Polymers for Electrochemical Sensor Applications: A Mini Review. *BioNanoScience* 10, 351–364. doi:10.1007/s12668-019-00708-x

Amare, M., and Admassie, S. (2012). Polymer Modified Glassy Carbon Electrode for the Electrochemical Determination of Caffeine in Coffee. *Talanta* 93, 122–128. doi:10.1016/j.talanta.2012.01.058

Araújo, D. A. G., Camargo, J. R., Pradela-Filho, L. A., Lima, A. P., Muñoz, R. A. A., Takeuchi, R. M., et al. (2020). A Lab-Made Screen-Printed Electrode as a Platform to Study the Effect of the Size and Functionalization of Carbon Nanotubes on the Voltammetric Determination of Caffeic Acid. *Microchemical J.* 158, 105297. doi:10.1016/j.microc.2020.105297

Araújo, D. A. G., Camargo, J. R., Pradela-Filho, L. A., Lima, A. P., Muñoz, R. A. A., Takeuchi, R. M., et al. (2020). A Lab-Made Screen-Printed Electrode as a Platform to Study the Effect of the Size and Functionalization of Carbon Nanotubes on the Voltammetric Determination of Caffeic Acid. *Microchemical J.* 158, 105297. doi:10.1016/j.microc.2020.105297

Azab, S. M., Shehata, M., and Fekry, A. M. (2019). A Novel Electrochemical Analysis of the Legal Psychoactive Drug Caffeine Using a Zeolite/MWCNT Modified Carbon Paste Sensor. *New J. Chem.* 43, 15359–15367. doi:10.1039/c9nj04070f

Baghayeri, M., Amiri, A., Hasheminejad, E., and Mahdavi, B. (2018). Poly(Aminohippuric Acid)–Sodium Dodecyl Sulfate/Functionalized Graphene Oxide Nanocomposite for Amplified Electrochemical Sensing of Gallic Acid. *J. Iranian Chem. Soc.* 15, 1931–1938. doi:10.1007/s13738-018-1390-3

Bard, A. J., and Faulkner, L. R. (2002). *Student Solutions Manual to Accompany Electrochemical Methods: Fundamentals and Applicaitons*. 2e. John Wiley & Sons. 0-471-40521-3.

Beringer, T., Kulak, M., Müller, C., Schaphoff, S., and Jans, Y. (2020). First Process-Based Simulations of Climate Change Impacts on Global Tea Production Indicate Large Effects in the World's Major Producer Countries. *Environ. Res. Lett.* 15, 034023. doi:10.1088/1748-9326/ab649b

Boopathy, G., Keerthi, M., Chen, S.-M., Meenakshi, S., and Umapathy, M. J. (2021). Molybdenum Trioxide Embedded Graphitic Carbon Nitride Sheets Modified Electrode for Caffeine Sensing in Green Tea and Coffee Powder. *Mater. Chem. Phys.* 269, 124735. doi:10.1016/j.matchemphys.2021.124735

Broli, N., Vallja, L., Shehu, A., and Vasjari, M. (2019). Determination of Catechol in Extract of Tea Using Carbon Paste Electrode Modified with Banana Tissue. *J. Food Process. Preservation* 43, e13838. doi:10.1111/jfpp.13838

Chapagain, A. K., and Hoekstra, A. Y. (2007). The Water Footprint of Coffee and Tea Consumption in the Netherlands. *Ecol. Econ.* 64, 109–118. doi:10.1016/j.ecolecon.2007.02.022

Chatterjee, T. N., Das, D., Roy, R. B., Tudu, B., Sabhapondit, S., Tamuly, P., et al. (2018). Molecular Imprinted Polymer Based Electrode for Sensing Catechin (+ C) in Green Tea. *IEEE Sensors J.* 18, 2236–2244. doi:10.1109/jsen.2018.2791661

Chattopadhyay, S., and Sarkar, P. (2014). Estimation of Tea Polyphenols by Electrochemical Sensors via Ferric Chloride Modified Electrodes. *JOURNAL INDIAN CHEMICAL SOCIETY* 91, 2291–2298. doi:10.5281/zenodo.5746521

Chen, C. (2006). CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *J. Am. Soc. Inf. Sci. Tech.* 57, 359–377. doi:10.1002/asi.20317

Chen, C. (2004). Searching for Intellectual Turning Points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci.* 101, 5303–5310. doi:10.1073/pnas.0307513100

Chen, C., and Song, M. (2019). Visualizing a Field of Research: A Methodology of Systematic Scientometric Reviews. *PloS one* 14, e0223994. doi:10.1371/journal.pone.0223994

Chen, G., Zhang, H., and Ye, J. (2000). Determination of Rutin and Quercetin in Plants by Capillary Electrophoresis with Electrochemical Detection. *Analytica Chim. Acta* 423, 69–76. doi:10.1016/S0003-2670(00)01099-0

Chen, M., Lv, H., Li, X., Tian, Z., and Ma, X. (2019). Determination of Gallic Acid in Tea by a Graphene Modified Glassy Carbon Electrode. *Int. J. Electrochem. Sci.* 14, 4852–4860. doi:10.20964/2019.05.23

Chen, X., Li, Q., Yu, S., Lin, B., and Wu, K. (2012). Activated Silica Gel Based Carbon Paste Electrodes Exhibit Signal Enhancement for Quercetin. *Electrochimica Acta* 81, 106–111. doi:10.1016/j.electacta.2012.07.063

Choi, C., Seo, S.-D., Kim, B.-K., and Kim, D.-W. (2016). Enhanced Lithium Storage in Hierarchically Porous Carbon Derived from Waste Tea Leaves. *Scientific Rep.* 6, 39099. doi:10.1038/srep39099

Chu, K. O., Wang, C. C., Rogers, M. S., Choy, K. W., and Pang, C. P. (2004). Determination of Catechins and Catechin Gallates in Biological Fluids by HPLC with Coulometric Array Detection and Solid Phase Extraction. *Analytica Chim. Acta* 510, 69–76. doi:10.1016/j.aca.2003.12.060

Cortina-Puig, M., Muñoz-Berbel, X., Calas-Blanchard, C., and Marty, J.-L. (2010). Diazonium-Functionalized Tyrosinase-Based Biosensor for the Detection of Tea Polyphenols. *Microchimica Acta* 171, 187–193. doi:10.1007/s00604-010-0425-y

Cui, F., and Zhang, X. (2013). A Method Based on Electrodeposition of Reduced Graphene Oxide on Glassy Carbon Electrode for Sensitive Detection of Theophylline. *J. Solid State. Electrochemistry* 17, 167–173. doi:10.1007/s10008-012-1867-4

da Silva, W., Ghica, M. E., and Brett, C. M. A. (2018). Gold Nanoparticle Decorated Multiwalled Carbon Nanotube Modified Electrodes for the Electrochemical Determination of Theophylline. *Anal. Methods* 10, 5634–5642. doi:10.1039/C8AY02150C

Dan, X., Ruiyi, L., Zaijun, L., Haiyan, Z., Zhiguo, G., and Guangli, W. (2020). Facile Strategy for Synthesis of Silver-Graphene Hybrid with Controllable Size and Excellent Dispersion for Ultrasensitive Electrochemical Detection of Acetamiprid. *Appl. Surf. Sci.* 512, 145628. doi:10.1016/j.apsusc.2020.145628

Das, D., Chatterjee, T. N., Roy, R. B., Tudu, B., Hazarika, A. K., Sabhapondit, S., et al. (2020). Titanium Oxide Nanocubes Embedded Molecularly Imprinted Polymer-Based Electrode for Selective Detection of Caffeine in Green Tea. *IEEE Sensors J.* 20, 6240–6247. doi:10.1109/jsen.2020.2972773

de Carvalho Couto, C., dos Santos, D. G., Oliveira, E. M. M., and Freitas-Silva, O. (2021). Global Situation of Reference Materials to Assure Coffee, Cocoa, and Tea Quality and Safety. *Trac Trends Anal. Chem.*, 116381. doi:10.1016/j.trac.2021.116381

Deng, X., Chan, C. K., and Tüysüz, H. (2016). Spent Tea Leaf Templating of Cobalt-Based Mixed Oxide Nanocrystals for Water Oxidation. *ACS Appl. Mater. Inter.* 8, 32488–32495. doi:10.1021/acsami.6b12005

Devadas, B., and Chen, S.-M. (2015). Controlled Electrochemical Synthesis of Yttrium (III) Hexacyanoferrate Micro Flowers and Their Composite with Multiwalled Carbon Nanotubes, and its Application for Sensing Catechin in Tea Samples. *J. Solid State. Electrochemistry* 19, 1103–1112. doi:10.1007/s10008-014-2715-5

Djitieu Deutchoua, A. D., Ngueumaleu, Y., Kenne Dedzo, G., Kenfack Tonle, I., and Ngameni, E. (2019). Electrochemical Study of DPPH Incorporated in Carbon Paste Electrode as Potential Tool for Antioxidant Properties Determination. *Electroanalysis* 31, 335–342.

Dongmo, L. M., Jiokeng, S. L., Pecheu, C. N., Walcarius, A., and Tonle, I. K. (2020). Amino-Grafting of Montmorillonite Improved by Acid Activation and Application to the Electroanalysis of Catechol. *Appl. Clay Sci.* 191, 105602. doi:10.1016/j.clay.2020.105602

Duan, Y., Luo, X., Qin, Y., Zhang, H., Sun, G., Sun, X., et al. (2013). Determination of Epigallocatechin-3-Gallate with a High-Efficiency Electrochemical Sensor Based on a Molecularly Imprinted Poly(o-Phenylenediamine) Film. *J. Appl. Polym. Sci.* 129, 2882–2890. doi:10.1002/app.39002

Eremia, S. A. V., Vasilescu, I., Radoi, A., Litescu, S.-C., and Radu, G.-L. (2013). Disposable Biosensor Based on Platinum Nanoparticles-Reduced Graphene Oxide-Laccase Biocomposite for the Determination of Total Polyphenolic Content. *Talanta* 110, 164–170. doi:10.1016/j.talanta.2013.02.029

Ezhil Vilian, A. T., Madhu, R., Chen, S.-M., Veeramani, V., Sivakumar, M., Huh, Y. S., et al. (2015). Facile Synthesis of MnO2/Carbon Nanotubes Decorated with a Nanocomposite of Pt Nanoparticles as a New Platform for the Electrochemical Detection of Catechin in Red Wine and Green Tea Samples. *J. Mater. Chem. B* 3, 6285–6292. doi:10.1039/C5TB00508F

Feng, L., Gwee, X., Kua, E.-H., and Ng, T.-P. (2010). Cognitive Function an Tea Consumption in Community Dwelling Older Chinese in Singapore. *J. Nutr. Health Aging* 14, 433–438. doi:10.1007/s12603-010-0095-9

Feng, S., Zhou, X., Chen, X., Zhang, G., Liu, G., Wu, N., et al. (2020). Au@ Ag Core-Shell Nanomaterials Embedded in N-Doped Graphene: A Novel Electrochemical Sensor for Determination of Gallic Acid. *Int. J. Electrochem. Sci.* 15, 6908–6919. doi:10.20964/2020.07.28

Gao, L., Yue, R., Xu, J., Liu, Z., and Chai, J. (2018). Pt-PEDOT/RGO Nanocomposites: One-Pot Preparation and Superior Electrochemical Sensing Performance for Caffeic Acid in Tea. *J. Electroanalytical Chem.* 816, 14–20. doi:10.1016/j.jelechem.2018.03.024

Gao, Y., and Guo, L. (2013). A Sensitive Theophylline Sensor Based on a Single Walled Carbon Nanotube–Large Mesoporous Carbon/Nafion/Glassy Carbon Electrode. *Anal. Methods* 5, 5785–5791. doi:10.1039/C3AY41236A

Gao, Y., Ru, Q., Zheng, M., Pan, Z., Lei, T., Zhang, J., et al. (2021). Recovery of Kitchen Bio-Waste from Spent Black Tea as Hierarchical Biomorphic Carbon Electrodes for Ultra-long Lifespan Potassium-Ion Storage. *Appl. Surf. Sci.* 555, 149675. doi:10.1016/j.apsusc.2021.149675

Guo, D., Zheng, D., Mo, G., and Ye, J. (2009). Adsorptive Stripping Voltammetric Detection of Tea Polyphenols at Multiwalled Carbon Nanotubes-chitosan Composite Electrode. *Electroanalysis: Int. J. Devoted Fundam. Pract. Aspects Electroanalysis* 21, 762–766. doi:10.1002/elan.200804475

Guo, S., Zhu, Q., Yang, B., Wang, J., and Ye, B. (2011). Determination of Caffeine Content in Tea Based on Poly(Safranine T) Electroactive Film Modified Electrode. *Food Chem.* 129, 1311–1314. doi:10.1016/j.foodchem.2011.05.095

Guth, U., Vonau, W., and Zosel, J. (2009). Recent Developments in Electrochemical Sensor Application and Technology—A Review. *Meas. Sci. Tech.* 20, 042002. doi:10.1088/0957-0233/20/4/042002

Hagerman, A. E., Dean, R. T., and Davies, M. J. (2003). Radical Chemistry of Epigallocatechin Gallate and its Relevance to Protein Damage. *Arch. Biochem. Biophys.* 414, 115–120. doi:10.1016/s0003-9861(03)00158-9

Hallaj, R., Soltani, E., Mafakheri, S., and Ghadermazi, M. (2021). A Surface-Modified Silicon Carbide Nanoparticles Based Electrochemical Sensor for Free Interferences Determination of Caffeine in Tea and Coffee. *Mater. Sci. Eng. B* 274, 115473. doi:10.1016/j.mseb.2021.115473

Hegde, R. N., Hosamani, R. R., and Nandibewoor, S. T. (2009). Electrochemical Oxidation and Determination of Theophylline at a Carbon Paste Electrode Using Cetyltrimethyl Ammonium Bromide as Enhancing Agent. *null* 42, 2665–2682. doi:10.1080/00032710903243620

Hocker, N., Wang, C., Prochotsky, J., Eppurath, A., Rudd, L., and Perera, M. (2017). Quantification of Antioxidant Properties in Popular Leaf and Bottled Tea by High-Performance Liquid Chromatography (HPLC), Spectrophotometry, and Voltammetry. *Anal. Lett.* 50, 1640–1656. doi:10.1080/00032719.2016.1242008

Hong, L., Yongxin, Z., and Kissinger, P. T. (2003). Liquid Chromatography with Multi-Channel Electrochemical Detection for the Determination of Natural Phenolic Compounds. *Chin. J. Anal. Chem.* 31, 631–634. doi:10.1016/S1570-0232(02)00087-9

Hu, X., Xi, J., Xia, Y., Zhao, F., and Zeng, B. (2019). Space-Confined Synthesis of Ordered Mesoporous Carbon Doped with Single-Layer MoS2–Boron for the Voltammetric Determination of Theophylline. *Microchimica Acta* 186, 694. doi:10.1007/s00604-019-3824-8

Huang, D., Ou, B., and Prior, R. L. (2005). The Chemistry behind Antioxidant Capacity Assays. *J. Agric. Food Chem.* 53, 1841–1856. doi:10.1021/jf030723c

Incebay, H., Aktepe, L., and Leblebici, Z. (2020). An Electrochemical Sensor Based on Green Tea Extract for Detection of Cd(II) Ions by Differential Pulse Anodic Stripping Voltammetry. *Surf. Inter.* 21, 100726. doi:10.1016/j.surfin.2020.100726

Jiang, M., Ma, M.-J., Lin, C.-H., Yang, M., Fang, L., Liu, J.-H., et al. (2020). Uniform Manganese-Loaded Titanium Dioxide Nanotube Arrays for Accurate Detection of Trace Cd2+ in Water, Soil and Tea: Enhanced Stability and Sensitivity. *Chem. Eng. J.* 400, 125972. doi:10.1016/j.cej.2020.125972

Jiaojiao, X., Pengyun, W., Bin, Z., and Onyinye, A. I. (2021). Enhancing Electrochemical Sensing for Catechol by Biomimetic Oxidase Covalently Functionalized Graphene Oxide. *Bioproc. Biosyst. Eng.* 44, 343–353. doi:10.1007/s00449-020-02446-x

Kan, X., Liu, T., Li, C., Zhou, H., Xing, Z., and Zhu, A. (2012). A Novel Electrochemical Sensor Based on Molecularly Imprinted Polymers for Caffeine Recognition and Detection. *J. Solid State. Electrochemistry* 16, 3207–3213. doi:10.1007/s10008-012-1760-1

Karaboduk, K., and Hasdemīr, E. (2020). Simultaneous Determination of Quercetin and Luteolin in Mate and White Tea Samples by Voltammetry. *REVUE ROUMAINE DE CHIMIE* 65, 375–385. doi:10.33224/rrch/2020.65.4.07

Karimi-Maleh, H., Alizadeh, M., Orooji, Y., Karimi, F., Baghayeri, M., Rouhi, J., et al. (2021). Guanine-Based DNA Biosensor Amplified with Pt/SWCNTs Nanocomposite as Analytical Tool for Nanomolar Determination of Daunorubicin as an Anticancer Drug: A Docking/Experimental Investigation. *Ind. Eng. Chem. Res.* 60, 816–823. doi:10.1021/acs.iecr.0c04698

Karimi-Maleh, H., Ayati, A., Davoodi, R., Tanhaei, B., Karimi, F., Malekmohammadi, S., et al. (2021). Recent Advances in Using of Chitosan-Based Adsorbents for Removal of Pharmaceutical Contaminants: A Review. *J. Clean. Prod.* 291, 125880. doi:10.1016/j.jclepro.2021.125880

Karimi-Maleh, H., Karimi, F., Fu, L., Sanati, A. L., Alizadeh, M., Karaman, C., et al. (2022). Cyanazine Herbicide Monitoring as a Hazardous Substance by a DNA Nanostructure Biosensor. *J. Hazard. Mater.* 423, 127058. doi:10.1016/j.jhazmat.2021.127058

Karimi-Maleh, H., Khataee, A., Karimi, F., Baghayeri, M., Fu, L., Rouhi, J., et al. A Green and Sensitive Guanine-Based DNA Biosensor for Idarubicin Anticancer Monitoring in Biological Samples: A Simple and Fast Strategy for Control of Health Quality in Chemotherapy Procedure Confirmed by Docking Investigation. *Chemosphere* 2021, 132928. doi:10.1016/j.chemosphere.2021.132928

Karimi-Maleh, H., Orooji, Y., Karimi, F., Alizadeh, M., Baghayeri, M., Rouhi, J., et al. (2021). A Critical Review on the Use of Potentiometric Based Biosensors for Biomarkers Detection. *Biosens. Bioelectron.*, 113252. doi:10.1016/j.bios.2021.113252

Karthika, A., Sudhakar, C., Suganthi, A., and Rajarajan, M. (2019). Eco-Friendly Synthesis of Aloe Vera Plant Extract Decorated Iron Tungstate Nanorods Immobilized Nafion for Selective and Sensitive Determination of Theophylline in Blood Serum, Black Tea and Urine Samples. *J. Sci. Adv. Mater. Devices* 4, 554–560. doi:10.1016/j.jsamd.2019.09.004

Karthika, A., Suganthi, A., and Rajarajan, M. (2020). An In-Situ Synthesis of Novel V2O5/G-C3n4/PVA Nanocomposite for Enhanced Electrocatalytic Activity toward Sensitive and Selective Sensing of Folic Acid in Natural Samples. *Arabian J. Chem.* 13, 3639–3652. doi:10.1016/j.arabjc.2019.12.009

Kartsova, L. A., and Ganzha, O. V. (2006). Electrophoretic Separation of Tea Flavanoids in the Modes of Capillary (Zone) Electrophoresis and Micellar Electrokinetic Chromatography. *Russ. J. Appl. Chem.* 79, 1110–1114. doi:10.1134/S1070427206070135

Kartsova, L., and Alekseeva, A. (2008). Chromatographic and Electrophoretic Methods for Determining Polyphenol Compounds. *J. Anal. Chem.* 63, 1024–1033. doi:10.1134/s1061934808110026

Khand, N. H., Solangi, A. R., Ameen, S., Fatima, A., Buledi, J. A., Mallah, A., et al. (2021). A New Electrochemical Method for the Detection of Quercetin in Onion, Honey and Green Tea Using Co3O4 Modified GCE. *J. Food Meas. Characterization* 15, 3720–3730. doi:10.1007/s11694-021-00956-0

Kilmartin, P. A., Zou, H., and Waterhouse, A. L. (2001). A Cyclic Voltammetry Method Suitable for Characterizing Antioxidant Properties of Wine and Wine Phenolics. *J. Agric. Food Chem.* 49, 1957–1965. doi:10.1021/jf001044u

Koch, W., Kukula-Koch, W., Komsta, Ł., Marzec, Z., Szwerc, W., and Głowniak, K. (2018). Green Tea Quality Evaluation Based on its Catechins and Metals Composition in Combination with Chemometric Analysis. *Molecules* 23, 1689. doi:10.3390/molecules23071689

Kochman, J., Jakubczyk, K., Antoniewicz, J., Mruk, H., and Janda, K. H. (2021). Benefits and Chemical Composition of Matcha Green Tea: A Review. *Molecules* 26, 85.

Kotani, A., Miyashita, N., and Kusu, F. (2003). Determination of Catechins in Human Plasma after Commercial Canned Green Tea Ingestion by High-Performance Liquid Chromatography with Electrochemical Detection Using a Microbore Column. *J. Chromatogr. B* 788, 269–275. doi:10.1016/s1570-0232(02)01036-x

Kotani, A., Takahashi, K., Hakamata, H., Kojima, S., and Kusu, F. (2007). Attomole Catechins Determination by Capillary Liquid Chromatography with Electrochemical Detection. *Anal. Sci.* 23, 157–163. doi:10.2116/analsci.23.157

Kottawa-Arachchi, J. D., Gunasekare, M. K., and Ranatunga, M. A. (2019). Biochemical Diversity of Global Tea [Camellia Sinensis (L.) O. Kuntze] Germplasm and its Exploitation: A Review. *Genet. Resour. Crop Evol.* 66, 259–273. doi:10.1007/s10722-018-0698-2

Koçak, Ç. C., Karabiberoğlu, Ş. U., and Dursun, Z. (2019). Highly Sensitive Determination of Gallic Acid on Poly (L-Methionine)-Carbon Nanotube Composite Electrode. *J. Electroanalytical Chem.* 853, 113552. doi:10.1016/j.jelechem.2019.113552

Kumar, R., Qadir, G., Rajar, K., Balouch, A., Ibupoto, Z. H., and Parkash, A. (2021). Voltammetric Detection of Caffeine Content in Different Tea Stuffs by Using Co3O4/GCE-Nafion Electrode. *J. Iranian Chem. Soc.* 18, 701–708. doi:10.1007/s13738-020-02059-x

Kuyumcu Savan, E. (2020). Square Wave Voltammetric (SWV) Determination of Quercetin in Tea Samples at a Single-Walled Carbon Nanotube (SWCNT) Modified Glassy Carbon Electrode (GCE). *Anal. Lett.* 53, 858–872. doi:10.1080/00032719.2019.1684514

Kuzin, Y., Ivanov, A., Evtugyn, G., and Hianik, T. (2016). Voltammetric Detection of Oxidative DNA Damage Based on Interactions between Polymeric Dyes and DNA. *Electroanalysis* 28, 2956–2964. doi:10.1002/elan.201600297

Kuzin, Y., Porfireva, A., Stepanova, V., Evtugyn, V., Stoikov, I., Evtugyn, G., et al. (2015). Impedimetric Detection of DNA Damage with the Sensor Based on Silver Nanoparticles and Neutral Red. *Electroanalysis* 27, 2800–2808. doi:10.1002/elan.201500312

Lambert, J. D., and Yang, C. S. (2003). Cancer Chemopreventive Activity and Bioavailability of Tea and Tea Polyphenols. *Mutat. Research/Fundamental Mol. Mech. Mutagenesis* 523–524, 201–208. doi:10.1016/S0027-5107(02)00336-6

Lee, M.-J., Maliakal, P., Chen, L., Meng, X., Bondoc, F. Y., Prabhu, S., et al. (2002). Pharmacokinetics of Tea Catechins after Ingestion of Green Tea and (−)-Epigallocatechin-3-Gallate by Humans: Formation of Different Metabolites and Individual Variability. *Cancer Epidemiol. Prev. Biomarkers* 11, 1025–1032.

Li, Y., and Huang, W. (2015). Electrode Modified with Porous Alumina Microfibers as a Highly Sensitive Electrochemical Sensor for Quercetin. *Anal. Methods* 7, 2537–2541. doi:10.1039/C5AY00206K

Li, Y., Luo, Z., Li, G., Belwal, T., Li, L., Xu, Y., et al. (2021). Interference-Free Detection of Caffeine in Complex Matrices Using a Nanochannel Electrode Modified with Binary Hydrophilic–Hydrophobic PDMS. *ACS sensors* 6, 1604–1612. doi:10.1021/acssensors.1c00004

Li, Y., Qi, H., Fan, M., Zhu, Z., Zhan, S., Li, L., et al. (2020). Quantifying the Efficiency of O-Benzoquinones Reaction with Amino Acids and Related Nucleophiles by Cyclic Voltammetry. *Food Chem.* 317, 126454. doi:10.1016/j.foodchem.2020.126454

Lima, A. B., dos Santos, W. T., and Compton, R. G. (2019). Simple and Sensitive Determination of Sibutramine in Slimming Tea Beverages Using a Carbon Screen-printed Electrode with Adsorptive Stripping Voltammetry. *Electroanalysis* 31, 975–980. doi:10.1002/elan.201800888

Lima, A. P., dos Santos, W. T. P., Nossol, E., Richter, E. M., and Munoz, R. A. A. (2020). Critical Evaluation of Voltammetric Techniques for Antioxidant Capacity and Activity: Presence of Alumina on Glassy-Carbon Electrodes Alters the Results. *Electrochimica Acta* 358, 136925. doi:10.1016/j.electacta.2020.136925

Lin, H., Gan, T., and Wu, K. (2009). Sensitive and Rapid Determination of Catechol in Tea Samples Using Mesoporous Al-Doped Silica Modified Electrode. *Food Chem.* 113, 701–704. doi:10.1016/j.foodchem.2008.07.073

Liu, R., Long, L., Lei, C., Zhaoyang, W., Liu, Y., and Sijia, L. (2016). Fabrication and Application of a Molecularly Imprinted Sensor for Morin Detection. *Chin. J. Anal. Chem.*, 385–390.

Liu, Y., Wang, D.-Z., Zhang, S.-Z., and Zhao, H.-M. (2015). Global Expansion Strategy of Chinese Herbal Tea Beverage. *Adv. J. Food Sci. Tech.* 7, 739–745. doi:10.19026/ajfst.7.1731

Liu, Y., Wang, M., Zhao, F., Xu, Z., and Dong, S. (2005). The Direct Electron Transfer of Glucose Oxidase and Glucose Biosensor Based on Carbon Nanotubes/Chitosan Matrix. *Biosens. Bioelectron.* 21, 984–988. doi:10.1016/j.bios.2005.03.003

Liu, Y., Zhu, L., Hu, Y., Peng, X., and Du, J. (2017). A Novel Electrochemical Sensor Based on a Molecularly Imprinted Polymer for the Determination of Epigallocatechin Gallate. *Food Chem.* 221, 1128–1134. doi:10.1016/j.foodchem.2016.11.047

Lu, L., Wu, L., Wang, W., Long, X., Xu, J., and He, H. (2018). Electrochemical Sensor Based on Poly (3, 4-Ethylenedioxy-Thiophene) Doped with Transition Metals for Detecting Rutin in Buck Wheat Tea. *Int. J. Electrochem. Sci.* 13, 2126–2135. doi:10.20964/2018.02.66

Lü, S. (2004). Electrochemical Determination of Tannins Using Multiwall Carbon Nanotubes Modified Glassy Carbon Electrode. *Russ. J. Electrochemistry* 40, 750–754. doi:10.1023/B:RUEL.0000035260.35980.f7

Luo, J. H., Li, B. L., Li, N. B., and Luo, H. Q. (2013). Sensitive Detection of Gallic Acid Based on Polyethyleneimine-Functionalized Graphene Modified Glassy Carbon Electrode. *Sensors Actuators B: Chem.* 186, 84–89. doi:10.1016/j.snb.2013.05.074

Luo, W., Ang, C. Y., Gehring, T. A., Heinze, T. M., Lin, L. J., and Mattia, A. (2003). Determination of Phenolic Compounds in Dietary Supplements and Tea Blends Containing Echinacea by Liquid Chromatography with Coulometric Electrochemical Detection. *J. AOAC Int.* 86, 202–208. doi:10.1093/jaoac/86.2.202

Malakootian, M., Abolghasemi, H., and Mahmoudi-Moghaddam, H. (2020). A Novel Electrochemical Sensor Based on the Modified Carbon Paste Using Eu3+− Doped NiO for Simultaneous Determination of Pb (II) and Cd (II) in Food Samples. *J. Electroanalytical Chem.* 876, 114474. doi:10.1016/j.jelechem.2020.114474

Manavalan, S., Govindasamy, M., Chen, S.-M., Rajaji, U., Chen, T.-W., Ajmal Ali, M., et al. (2018). Reduced Graphene Oxide Supported Raspberry-like SrWO4 for Sensitive Detection of Catechol in Green Tea and Drinking Water Samples. *J. Taiwan Inst. Chem. Eng.* 89, 215–223. doi:10.1016/j.jtice.2018.05.001

Mani, V., Beduk, T., Khushaim, W., Ceylan, A. E., Timur, S., Wolfbeis, O. S., et al. (2021). Electrochemical Sensors Targeting Salivary Biomarkers: A Comprehensive Review. *Trac Trends Anal. Chem.* 135, 116164. doi:10.1016/j.trac.2020.116164

Marx, W., Haunschild, R., and Bornmann, L. (2017). Global Warming and Tea Production—The Bibliometric View on a Newly Emerging Research Topic. *Climate* 5, 46. doi:10.3390/cli5030046

Masoum, S., Behpour, M., Azimi, F., and Motaghedifard, M. H. (2014). Potentiality of Chemometric Approaches for the Determination of (+)-Catechin in Green Tea Leaves at the Surface of Multiwalled Carbon Nanotube Paste Electrode. *Sensors Actuators B: Chem.* 193, 582–591. doi:10.1016/j.snb.2013.12.022

Maximino, M. D., Martin, C. S., Paulovich, F. V., and Alessio, P. (2016). Layer-by-layer Thin Film of Iron Phthalocyanine as a Simple and Fast Sensor for Polyphenol Determination in Tea Samples. *J. Food Sci.* 81, C2344–C2351. doi:10.1111/1750-3841.13394

McCants, A. E. (2008). Poor Consumers as Global Consumers: The Diffusion of Tea and Coffee Drinking in the Eighteenth Century 1. *Econ. Hist. Rev.* 61, 172–200. doi:10.1111/j.1468-0289.2008.00429.x

Meng, R., Li, Q., Zhang, S., Tang, J., Ma, C., and Jin, R. (2019). GQDs/PEDOT Bilayer Films Modified Electrode as a Novel Electrochemical Sensing Platform for Rutin Detection. *Int. J. Electrochem. Sc* 14, 11000–11011. doi:10.20964/2019.12.40

Moccelini, S. K., Fernandes, S. C., de Camargo, T. P., Neves, A., and Vieira, I. C. (2009). Self-Assembled Monolayer of Nickel (II) Complex and Thiol on Gold Electrode for the Determination of Catechin. *Talanta* 78, 1063–1068. doi:10.1016/j.talanta.2009.01.038

Muthukumaran, P., Ramya, R., Thivya, P., Wilson, J., and Ravi, G. (2019). Nanocomposite Based on Restacked Crystallites of β-NiS and Ppy for the Determination of Theophylline and Uric Acid on Screen-Printed Electrodes. *New J. Chem.* 43, 19397–19407. doi:10.1039/c9nj04246f

Mzembe, A. N., Lindgreen, A., Maon, F., and Vanhamme, J. (2016). Investigating the Drivers of Corporate Social Responsibility in the Global Tea Supply Chain: A Case Study of Eastern Produce Limited in Malawi. *Corporate Soc. Responsibility Environ. Manag.* 23, 165–178. doi:10.1002/csr.1370

Nagles, E., and García-Beltrán, O. (2016). Determination of Rutin in Black Tea by Adsorption Voltammetry (AdV) in the Presence of Morin and Quercetin. *Food Anal. Methods* 9, 3420–3427. doi:10.1007/s12161-016-0538-y

Nandy Chatterjee, T., Banerjee Roy, R., Tudu, B., Pramanik, P., Deka, H., Tamuly, P., et al. (2017). Detection of Theaflavins in Black Tea Using a Molecular Imprinted Polyacrylamide-Graphite Nanocomposite Electrode. *Sensors Actuators B: Chem.* 246, 840–847. doi:10.1016/j.snb.2017.02.139

Nandy Chatterjee, T., Das, D., Banerjee Roy, R., Tudu, B., Hazarika, A. K., Sabhapondit, S., et al. (2019). Development of a Nickel Hydroxide Nanopetal Decorated Molecular Imprinted Polymer Based Electrode for Sensitive Detection of Epigallocatechin-3-Gallate in Green Tea. *Sensors Actuators B: Chem.* 283, 69–78. doi:10.1016/j.snb.2018.11.159

Narumi, K., Sonoda, J.-I., Shiotani, K., Shigeru, M., Shibata, M., Kawachi, A., et al. (2014). Simultaneous Detection of Green Tea Catechins and Gallic Acid in Human Serum after Ingestion of Green Tea Tablets Using Ion-Pair High-Performance Liquid Chromatography with Electrochemical Detection. *J. Chromatogr. B* 945–946, 147–153. doi:10.1016/j.jchromb.2013.11.007

Naveen, M. H., Gurudatt, N. G., and Shim, Y.-B. (2017). Applications of Conducting Polymer Composites to Electrochemical Sensors: A Review. *Appl. Mater. Today* 9, 419–433. doi:10.1016/j.apmt.2017.09.001

Ng, K.-W., Cao, Z.-J., Chen, H.-B., Zhao, Z.-Z., Zhu, L., and Yi, T. (2018). Oolong Tea: A Critical Review of Processing Methods, Chemical Composition, Health Effects, and Risk. *Crit. Rev. Food Sci. Nutr.* 58, 2957–2980. doi:10.1080/10408398.2017.1347556

Novak, I., Šeruga, M., and Komorsky-Lovrić, Š. (2010). Characterisation of Catechins in Green and Black Teas Using Square-Wave Voltammetry and RP-HPLC-ECD. *Food Chem.* 122, 1283–1289. doi:10.1016/j.foodchem.2010.03.084

Novak, I., Šeruga, M., and Komorsky-Lovrić, Š. (2010). Characterisation of Catechins in Green and Black Teas Using Square-Wave Voltammetry and RP-HPLC-ECD. *Food Chem.* 122, 1283–1289. doi:10.1016/j.foodchem.2010.03.084

Novak, I., Šeruga, M., and Komorsky-Lovrić, Š. (2009). Square-Wave and Cyclic Voltammetry of Epicatechin Gallate on Glassy Carbon Electrode. *J. Electroanalytical Chem.* 631, 71–75. doi:10.1016/j.jelechem.2009.03.005

Nunes, E. W., Silva, M. K. L., and Cesarino, I. (2020). Evaluation of a Reduced Graphene Oxide-Sb Nanoparticles Electrochemical Sensor for the Detection of Cadmium and Lead in Chamomile Tea. *Chemosensors* 8. doi:10.3390/chemosensors8030053

Pal, A., and Das, C. (2020). A Novel Use of Solid Waste Extract from Tea Factory as Corrosion Inhibitor in Acidic Media on Boiler Quality Steel. *Ind. Crops Prod.* 151, 112468. doi:10.1016/j.indcrop.2020.112468

Pal, A., and Das, C. (2020). A Novel Use of Solid Waste Extract from Tea Factory as Corrosion Inhibitor in Acidic Media on Boiler Quality Steel. *Ind. Crops Prod.* 151, 112468. doi:10.1016/j.indcrop.2020.112468

Pang, J., Wu, X., Li, A., Liu, X., and Li, M. (2017). Detection of Catechin in Chinese Green Teas at N-Doped Carbon-Modified Electrode. *Ionics* 23, 1889–1895. doi:10.1007/s11581-017-2006-0

Power, A. C., Gorey, B., Chandra, S., and Chapman, J. (2018). Carbon Nanomaterials and Their Application to Electrochemical Sensors: A Review. *Nanotechnology Rev.* 7, 19–41. doi:10.1515/ntrev-2017-0160

Raluca-Ioana Stefan (2004). Semere Ghebru Bairu Diamond Paste Based Electrodes for the Determination of Pb(II) at Trace Concentration Levels. *Talanta* 63, 605–608. doi:10.1016/j.talanta.2003.12.023

Ranga, R., Kumar, A., Kumari, P., Singh, P., Madaan, V., and Kumar, K. (2021). Ferrite Application as an Electrochemical Sensor: A Review. *Mater. Characterization*, 111269. doi:10.1016/j.matchar.2021.111269

Rauf, A., and Mahdi, E. (2012). Evaluating Corrosion Inhibitors with the Help of Electrochemical Measurements Including Electrochemical Frequency Modulation. *Int. J. Electrochem. Sci.* 7, 4673–4685.

Ravishankar, T. N., Suresh Kumar, K., Teixeira, S. R., Fernandez, C., and Ramakrishnappa, T. (2016). Ag Doped Titanium Dioxide Nanocomposite-Modified Glassy Carbon Electrode as Electrochemical Interface for Catechol Sensing. *Electroanalysis* 28, 452–461. doi:10.1002/elan.201500238

Sakthinathan, S., Kubendhiran, S., and Chen, S.-M. (2017). Hydrothermal Synthesis of Three Dimensional Graphene-Multiwalled Carbon Nanotube Nanocomposite for Enhanced Electro Catalytic Oxidation of Caffeic Acid. *Electroanalysis* 29, 1103–1112. doi:10.1002/elan.201600687

Sandeep, S., Santhosh, A. S., Swamy, N. K., Suresh, G. S., Melo, J. S., and Chamaraja, N. A. (2018). A Biosensor Based on a Graphene Nanoribbon/Silver Nanoparticle/Polyphenol Oxidase Composite Matrix on a Graphite Electrode: Application in the Analysis of Catechol in Green Tea Samples. *New J. Chem.* 42, 16620–16629. doi:10.1039/C8NJ02325E

Sanjay, B. P., Kumara Swamy, N., Yashas, S. R., and Sandeep, S. (2021). Design of an Electrochemical Sensor Using 2D Sheet-like Cu@g-C3n4 Transducer Matrix for Electroanalysis of Catechol. *J. Electrochem. Soc.* 168, 076511. doi:10.1149/1945-7111/ac1495

Saritha, D., Gupta, V. K., Reddy, A. V. B., Agarwal, S., Moniruzzaman, M., Anitha, K., et al. (2019). Development of a Simple, Selective, Stable and Ultrasensitive Poly (Safranine/Nano NiO) Modified Carbon Paste Electrode for Selective Detection of Rutin in Buckwheat and Green Tea Samples. *Int. J. Electrochem. Sci.* 14, 10093–10110. doi:10.20964/2019.11.48

Sen, S., Chattopadhyay, S., and Sarkar, P. (2015). Electrochemical Sensing of Tea Polyphenols by Chloramine-T Modified Electrodes: A New Approach. *J. Electrochem. Soc.* 163, B49–B55. doi:10.1149/2.0491603jes

Şenocak, A., Basova, T., Demirbas, E., and Durmuş, M. (2019). Direct and Fast Electrochemical Determination of Catechin in Tea Extracts Using SWCNT-Subphthalocyanine Hybrid Material. *Electroanalysis* 31, 1697–1707. doi:10.1002/elan.201900214

Šeruga, M., Novak, I., and Jakobek, L. (2011). Determination of Polyphenols Content and Antioxidant Activity of Some Red Wines by Differential Pulse Voltammetry, HPLC and Spectrophotometric Methods. *Food Chem.* 124, 1208–1216. doi:10.1016/j.foodchem.2010.07.047

Seth, R., Bhandawat, A., Parmar, R., Singh, P., Kumar, S., and Sharma, R. K. (2019). Global Transcriptional Insights of Pollen-Pistil Interactions Commencing Self-Incompatibility and Fertilization in Tea [Camellia Sinensis (L.) O. Kuntze]. *Int. J. Mol. Sci.* 20, 539. doi:10.3390/ijms20030539

Shao, X., Lv, L., Parks, T., Wu, H., Ho, C.-T., and Sang, S. (2010). Quantitative Analysis of Ginger Components in Commercial Products Using Liquid Chromatography with Electrochemical Array Detection. *J. Agric. Food Chem.* 58, 12608–12614. doi:10.1021/jf1029256

Shi, H., Chen, F., Zhao, S., Ye, C., Lin, C.-T., Zhu, J., et al. (2021). Preparation of Cassava Fiber-Iron Nanoparticles Composite for Electrochemical Determination of Tea Polyphenol. *J. Food Meas. Characterization* 15, 4711–4717. doi:10.1007/s11694-021-01030-5

Šimková, D., and Labuda, J. "Electrochemical Flow-Through System with DNA Biosensor for Biomedical and Food Technology Applications," in Proceedings of the 2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, November 24 2009, 181–186. doi:10.1109/isabel.2009.5373651

Soussou, A., Gammoudi, I., Moroté, F., Kalboussi, A., Cohen-Bouhacina, T., Grauby-Heywang, C., et al. (2017). Efficient Immobilization of Tyrosinase Enzyme on Layered Double Hydroxide Hybrid Nanomaterials for Electrochemical Detection of Polyphenols. *IEEE Sensors J.* 17, 4340–4348. doi:10.1109/jsen.2017.2709342

Su, Y.-L., and Cheng, S.-H. (2015). Sensitive and Selective Determination of Gallic Acid in Green Tea Samples Based on an Electrochemical Platform of Poly(Melamine) Film. *Analytica Chim. Acta* 901, 41–50. doi:10.1016/j.aca.2015.10.026

Tang, J., Wang, H., Jiang, X., Zhu, Z., Xie, J., Tang, J., et al. (2018). Electrochemical Behavior of Jasmine Tea Extract as Corrosion Inhibitor for Carbon Steel in Hydrochloric Acid Solution. *Int. J. Electrochem. Sci.* 13, 3625–3642. doi:10.20964/2018.04.41

Tanui, J. K., Fang, W., Feng, W., Zhuang, P., and Li, X. (2012). World Black Tea Markets: Relationships and Implications for the Global Tea Industry. *J. Int. Food Agribusiness Marketing* 24, 148–168. doi:10.1080/08974438.2012.665791

Tashkhourian, J., and Ghaderizadeh, S. M. (2014). SiO2-Modified Carbon Paste Electrode for Electrochemical Determination of Pyrogallol. *Russ. J. Electrochemistry* 50, 959–966. doi:10.1134/S1023193514100139

Tian, L., Wang, B., Chen, R., Gao, Y., Chen, Y., and Li, T. (2015). Determination of Quercetin Using a Photo-Electrochemical Sensor Modified with Titanium Dioxide and a Platinum(II)-Porphyrin Complex. *Microchimica Acta* 182, 687–693. doi:10.1007/s00604-014-1374-7

Uliana, C. V., Garbellini, G. S., and Yamanaka, H. (2014). Electrochemical Investigations on the Capacity of Flavonoids to Protect DNA against Damage Caused by Textile Disperse Dyes. *Sensors Actuators B: Chem.* 192, 188–195. doi:10.1016/j.snb.2013.10.091

Vellaichamy, B., Ponniah, S. K., and Prakash, P. (2017). An In-Situ Synthesis of Novel Au@NG-PPy Nanocomposite for Enhanced Electrocatalytic Activity toward Selective and Sensitive Sensing of Catechol in Natural Samples. *Sensors Actuators B: Chem.* 253, 392–399. doi:10.1016/j.snb.2017.06.147

Venkadesh, A., Mathiyarasu, J., and Radhakrishnan, S. (2021). Voltammetric Sensing of Caffeine in Food Sample Using Cu-MOF and Graphene. *Electroanalysis* 33, 1007–1013. doi:10.1002/elan.202060488

Wambu, E. W., Fu, H., and Ho, Y. (2017). Characteristics and Trends in Global Tea Research: A Science Citation Index Expanded-based Analysis. *Int. J. Food Sci. Tech.* 52, 644–651. doi:10.1111/ijfs.13317

Wang, G., He, X., Zhou, F., Li, Z., Fang, B., Zhang, X., et al. (2012). Application of Gold Nanoparticles/TiO2 Modified Electrode for the Electrooxidative Determination of Catechol in Tea Samples. *Food Chem.* 135, 446–451. doi:10.1016/j.foodchem.2012.04.139

Wang, H.-B., Zhang, H.-D., Zhang, Y.-H., Chen, H., Xu, L.-L., Huang, K.-J., et al. (2015). Tungsten Disulfide Nano-Flowers/Silver Nanoparticles Composites Based Electrochemical Sensor for Theophylline Determination. *J. Electrochem. Soc.* 162, B173–B179. doi:10.1149/2.0941507jes

Wang, H., Yuan, X., Zeng, G., Wu, Y., Liu, Y., Jiang, Q., et al. (2015). Three Dimensional Graphene Based Materials: Synthesis and Applications from Energy Storage and Conversion to Electrochemical Sensor and Environmental Remediation. *Adv. Colloid Interf. Sci.* 221, 41–59. doi:10.1016/j.cis.2015.04.005

Wang, J. (2000). *Analytical Electochemistry*. Second Edition. Wiley VCH. 978-0-471-28272-3.

Wang, J., Mu, J., Ma, J., Yang, Y., Wang, M., Zhu, L., et al. (2012). Determination of Rutin and Puerarin in Teas and Pharmaceutical Preparations Using Poly (Evans Blue) Film-Modified Electrodes. *J. Food Drug Anal.* 20, 611–616.

Wang, J., Zareef, M., He, P., Sun, H., Chen, Q., Li, H., et al. (2019). Evaluation of Matcha Tea Quality Index Using Portable NIR Spectroscopy Coupled with Chemometric Algorithms. *J. Sci. Food Agric.* 99, 5019–5027. doi:10.1002/jsfa.9743

Wang, X.-G., Li, J., and Fan, Y.-J. (2010). Fast Detection of Catechin in Tea Beverage Using a Poly-Aspartic Acid Film Based Sensor. *Microchimica Acta* 169, 173–179. doi:10.1007/s00604-010-0335-z

Wang, Y., Wei, X., Wang, F., and Li, M. (2014). Sensitive Voltammetric Detection of Caffeine in Tea and Other Beverages Based on a DNA-Functionalized Single-Walled Carbon Nanotube Modified Glassy Carbon Electrode. *Anal. Methods* 6, 7525–7531. doi:10.1039/c4ay00837e

Wang, Y., Yao, Z., Pan, Z., Wang, R., Yan, G., Liu, C., et al. (2020). Tea-planted Soils as Global Hotspots for N2O Emissions from Croplands. *Environ. Res. Lett.* 15, 104018. doi:10.1088/1748-9326/aba5b2

Wang, Y., Zhong, J., Ding, F., Zhao, Q., Zhang, Z., Liu, X., et al. (2018). A Bifunctional NiCo2S4/Reduced Graphene Oxide@polyaniline Nanocomposite as a Highly-Efficient Electrode for Glucose and Rutin Detection. *New J. Chem.* 42, 9398–9409. doi:10.1039/C8NJ00663F

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., et al. (2018). Draft Genome Sequence of Camellia Sinensis Var. Sinensis Provides Insights into the Evolution of the Tea Genome and Tea Quality. *Proc. Natl. Acad. Sci.* 115, E4151–E4158. doi:10.1073/pnas.1719622115

Wu, Y., Kou, J., Wang, L., Cheng, L., and Lu, K. (2019). Langmuir-Blodgett Films of Nafion-Nitrogen Doped Carbon Nanotubes as New Sensing Materials for the Determination of Caffeine in Tea. *Int. J. Electrochem. Sci.* 14, 11166–11177. doi:10.20964/2019.12.50

Xie, Y., Zhang, T., Chen, Y., Wang, Y., and Wang, L. (2020). Fabrication of Core-Shell Magnetic Covalent Organic Frameworks Composites and Their Application for Highly Sensitive Detection of Luteolin. *Talanta* 213, 120843. doi:10.1016/j.talanta.2020.120843

Xu, M., Wang, J., and Gu, S. (2019). Rapid Identification of Tea Quality by E-Nose and Computer Vision Combining with a Synergetic Data Fusion Strategy. *J. Food Eng.* 241, 10–17. doi:10.1016/j.jfoodeng.2018.07.020

Xu, M., Wang, J., and Zhu, L. (2019). The Qualitative and Quantitative Assessment of Tea Quality Based on E-Nose, E-Tongue and E-Eye Combined with Chemometrics. *Food Chem.* 289, 482–489. doi:10.1016/j.foodchem.2019.03.080

Yang, B., Kotani, A., Arai, K., and Kusu, F. (2001). Relationship of Electrochemical Oxidation of Catechins on Their Antioxidant Activity in Microsomal Lipid Peroxidation. *Chem. Pharm. Bull.* 49, 747–751. doi:10.1248/cpb.49.747

Yang, J., Deng, S., Lei, J., Ju, H., and Gunasekaran, S. (2011). Electrochemical Synthesis of Reduced Graphene Sheet–AuPd Alloy Nanoparticle Composites for Enzymatic Biosensing. *Biosens. Bioelectron.* 29, 159–166. doi:10.1016/j.bios.2011.08.011

Yang, L., Yang, J., Xu, B., Zhao, F., and Zeng, B. (2016). Facile Preparation of Molecularly Imprinted Polypyrrole-Graphene-Multiwalled Carbon Nanotubes Composite Film Modified Electrode for Rutin Sensing. *Talanta* 161, 413–418. doi:10.1016/j.talanta.2016.08.080

Yang, Y. J., and Li, W. (2015). High Sensitive Determination of Theophylline Based on Manganese Oxide Nanoparticles/Multiwalled Carbon Nanotube Nanocomposite Modified Electrode. *Ionics* 21, 1121–1128. doi:10.1007/s11581-014-1264-3

Yao, Y., Zhang, L., Wen, Y., Wang, Z., Zhang, H., Hu, D., et al. (2015). Voltammetric Determination of Catechin Using Single-Walled Carbon Nanotubes/Poly(Hydroxymethylated-3,4-Ethylenedioxythiophene) Composite Modified Electrode. *Ionics* 21, 2927–2936. doi:10.1007/s11581-015-1494-z

Yi, J., Liu, Z., Liu, J., Liu, H., Xia, F., Tian, D., et al. (2020). A Label-free Electrochemical Aptasensor Based on 3D Porous CS/RGO/GCE for Acetamiprid Residue Detection. *Biosens. Bioelectron.* 148, 111827. doi:10.1016/j.bios.2019.111827

Yin, C., Zhuang, Q., Xiao, Q., Wang, Y., and Xie, J. (2021). Electropolymerization of Poly(Methylene Blue) on Flower-like Nickel-Based MOFs Used for Ratiometric Electrochemical Sensing of Total Polyphenolic Content in Chrysanthemum Tea. *Anal. Methods* 13, 1154–1163. doi:10.1039/D1AY00028D

Yin, H., Meng, X., Su, H., Xu, M., and Ai, S. (2012). Electrochemical Determination of Theophylline in Foodstuff, Tea and Soft Drinks Based on Urchin-like CdSe Microparticles Modified Glassy Carbon Electrode. *Food Chem.* 134, 1225–1230. doi:10.1016/j.foodchem.2012.02.197

Ying, J., Zheng, Y., Zhang, H., and Fu, L. (2020). Room Temperature Biosynthesis of Gold Nanoparticles with Lycoris Aurea Leaf Extract for the Electrochemical Determination of Aspirin. *Revista Mexicana de Ingeniería Química* 19, 585–592. doi:10.24275/rmiq/mat741

Yue, X., Pang, S., Han, P., Zhang, C., Wang, J., and Zhang, L. (2013). Carbon Nanotubes/Carbon Paper Composite Electrode for Sensitive Detection of Catechol in the Presence of Hydroquinone. *Electrochemistry Commun.* 34, 356–359. doi:10.1016/j.elecom.2013.07.016

Zeng, Q., Chen, J., Gao, F., Tu, X., Qian, Y., Yu, Y., et al. (2021). Development of a New Electrochemical Sensing Platform Based on MoO3-Polypyrrole Nanowires/MWCNTs Composite and its Application to Luteolin Detection. *Synth. Met.* 271, 116620. doi:10.1016/j.synthmet.2020.116620

Zhang, G., Fu, H., Zou, D., Xiao, R., Liu, J., and Li, S. (2017). Electrochemical Determination of Caffeine in Tea Using a Polydopamine-Gold Nanocomposite. *Int. J. Electrochem. Sci.* 12, 11465–11472. doi:10.20964/2017.12.76

Zhang, H., Wu, S., Xing, Z., Wang, H.-B., and Liu, Y.-M. (2021). A Highly Sensitive Electrochemical Sensor for Theophylline Based on Dopamine-Melanin Nanosphere (DMN)–Gold Nanoparticles (AuNPs)-Modified Electrode. *Appl. Phys. A* 127, 844. doi:10.1007/s00339-021-04968-x

Zhang, M., Pan, B., Wang, Y., Du, X., Fu, L., Zheng, Y., et al. (2020). Recording the Electrochemical Profile of Pueraria Leaves for Polyphyly Analysis. *ChemistrySelect* 5, 5035–5040. doi:10.1002/slct.202001100

Zhang, Y., Shang, J., Jiang, B., Zhou, X., and Wang, J. (2017). Electrochemical Determination of Caffeine in Oolong Tea Based on Polyelectrolyte Functionalized Multi-Walled Carbon Nanotube. *Int. J. Electrochem. Sci.* 12, 2552–2562. doi:10.20964/2017.03.02

Zhao, H., Ran, Q., Li, Y., Li, B., Liu, B., Ma, H., et al. (2020). Highly Sensitive Detection of Gallic Acid Based on 3D Interconnected Porous Carbon Nanotubes/Carbon Nanosheets Modified Glassy Carbon Electrode. *J. Mater. Res. Tech.* 9, 9422–9433. doi:10.1016/j.jmrt.2020.05.102

Zheng, C., Wang, Y., Ding, Z., and Zhao, L. (2016). Global Transcriptional Analysis Reveals the Complex Relationship between Tea Quality, Leaf Senescence and the Responses to Cold-Drought Combined Stress in Camellia Sinensis. *Front. Plant Sci.* 7, 1858. doi:10.3389/fpls.2016.01858

Zhi, R., Zhao, L., and Zhang, D. (2017). A Framework for the Multi-Level Fusion of Electronic Nose and Electronic Tongue for Tea Quality Assessment. *Sensors* 17, 1007. doi:10.3390/s17051007

Zhong, T., Guo, Q., Yin, Z., Zhu, X., Liu, R., Liu, A., et al. (2019). Polyphenol Oxidase/Gold Nanoparticles/Mesoporous Carbon-Modified Electrode as an Electrochemical Sensing Platform for Rutin in Dark Teas. *RSC Adv.* 9, 2152–2155. doi:10.1039/C8RA08199A

Zhou, J., Zheng, Y., Zhang, J., Karimi-Maleh, H., Xu, Y., Zhou, Q., et al. (2020). Characterization of the Electrochemical Profiles of Lycoris Seeds for Species Identification and Infrageneric Relationships. *Anal. Lett.* 53, 2517–2528. doi:10.1080/00032719.2020.1746327

Zhu, X., Liu, P., Ge, Y., Wu, R., Xue, T., Sheng, Y., et al. (2020). MoS2/MWCNTs Porous Nanohybrid Network with Oxidase-like Characteristic as Electrochemical Nanozyme Sensor Coupled with Machine Learning for Intelligent Analysis of Carbendazim. *J. Electroanalytical Chem.* 862, 113940. doi:10.1016/j.jelechem.2020.113940

Ziyatdinova, G., Aytuganova, I., Nizamova, A., Morozov, M., and Budnikov, H. (2012). Cyclic Voltammetry of Natural Flavonoids on MWNT-Modified Electrode and Their Determination in Pharmaceuticals. *Collection Czechoslovak Chem. Commun.* 76, 1619–1631. doi:10.1135/cccc2011115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Analytical Science

ORIGINAL RESEARCH
published: 14 January 2022
doi: 10.3389/frans.2021.772844

Synchronization-Free Multivariate Statistical Process Control for Online Monitoring of Batch Process Evolution

Rodrigo Rocha de Oliveira and Anna de Juan**

Chemometrics Group, Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Barcelona, Spain

Synchronization of variable trajectories from batch process data is a delicate operation that can induce artifacts in the definition of multivariate statistical process control (MSPC) models for real-time monitoring of batch processes. The current paper introduces a new synchronization-free approach for online batch MSPC. This approach is based on the use of local MSPC models that cover a normal operating conditions (NOC) trajectory defined from principal component analysis (PCA) modeling of non-synchronized historical batches. The rationale behind is that, although non-synchronized NOC batches are used, an overall NOC trajectory with a consistent evolution pattern can be described, even if batch-to-batch natural delays and differences between process starting and end points exist. Afterwards, the local MSPC models are used to monitor the evolution of new batches and derive the related MSPC chart. During the real-time monitoring of a new batch, this strategy allows testing whether every new observation is following or not the NOC trajectory. For a NOC observation, an additional indication of the batch process progress is provided based on the identification of the local MSPC model that provides the lowest residuals. When an observation deviates from the NOC behavior, contribution plots based on the projection of the observation to the best local MSPC model identified in the last NOC observation are used to diagnose the variables related to the fault. This methodology is illustrated using two real examples of NIR-monitored batch processes: a fluidized bed drying process and a batch distillation of gasoline blends with ethanol.

Keywords: batch process, online process monitoring, statistical process control, synchronization-free MSPC, local MSPC modeling

OPEN ACCESS

Edited by:
Federico Marini,
Sapienza University of Rome, Italy

Reviewed by:
Marina Cocchi,
University of Modena and Reggio Emilia, Italy
Jahan B. Ghasemi,
University of Tehran, Iran

****Correspondence:***
Rodrigo Rocha de Oliveira
rodrigo.rocha@ub.edu
Anna de Juan
anna.dejuan@ub.edu

Specialty section:
This article was submitted to Chemometrics, a section of the journal Frontiers in Analytical Science

Received: *09 September 2021*
Accepted: *27 December 2021*
Published: *14 January 2022*

Citation:
Rocha de Oliveira R and de Juan A (2022) Synchronization-Free Multivariate Statistical Process Control for Online Monitoring of Batch Process Evolution. Front. Anal. Sci. 1:772844. doi: 10.3389/frans.2021.772844

INTRODUCTION

Industrial sectors often rely on batch processes to produce their intermediate or final products. Batch processes consist of cyclic repetitions of an established recipe aiming at the production of products meeting specific quality specifications. They are also characterized by complex, dynamic and nonstationary behavior. Thus, monitoring a batch evolution in real-time is a challenging, but essential action to obtain end products with desired quality, reducing costs and increasing process understanding. (van Sprang et al., 2002; Rendall et al., 2019; Rato and Reis, 2020).

Abbreviations: MCR-ALS, multivariate curve resolution—alternating least squares; MSPC, multivariate statistical process control; NIR, near infrared; NOC, normal operating conditions; PCA, principal component analysis; PC, principal component; PLS, partial least squares regression.

Nowadays, with the emergence of Industry 4.0, batch processes are monitored not only with typical process sensors, e.g., temperature, pressure, flow, etc, but also with advanced sensors probes based on spectroscopic techniques such as near-infrared (NIR), mid-infrared, and Raman (Cimander and Mandenius, 2004; Pöllänen et al., 2006; Ávila et al., 2012; Besenhard et al., 2018; Grassi et al., 2019; Avila et al., 2021). The collection and use of process sensor measurements from historical batches that followed the normal operating conditions (NOC) and reached the targeted product specifications is the basis for the development of multivariate statistical process control (MSPC) models and related charts, ready to be used to test the evolution of new batches (Kourti, 2005; Ferrer-Riquelme, 2009; Wold et al., 2009; Colucci et al., 2019; Vidal-Puig et al., 2019; França et al., 2021). Offline MSPC charts can be used to diagnose the root cause of a disturbance from a finished faulty batch. However, it is even more important the online use of MSPC charts for real-time monitoring of batch evolution to enable taking quick action in case of detection of process disturbances.

Process data measurements from a single batch consist of the collection of several variables, J, (process data and/or spectroscopic measurements) at different process points throughout the batch, K_i. These measurements are usually organized in a data matrix, $\mathbf{X_i}$, with dimension $(K_i \times J)$ to be used for process monitoring and/or control purposes. Most data-driven modeling strategies aiming at building online MSPC charts to monitor process evolution require that data from several NOC batches, I, that have the same batch length, i.e. batch data matrices with the same numbers of rows K, and follow the same and synchronized process dynamics. When this happens, the data can be arranged in a three-dimensional data array, $\underline{\mathbf{X}}$, with dimensions $I \times K \times J$. Most of the MSPC models are built based on data-driven multivariate analysis methods, such as principal component analysis (PCA) and partial least squares (PLS); for this purpose, different unfolding strategies of the $\underline{\mathbf{X}}$ array can be used according to the modeling approach used as originally introduced elsewhere (Nomikos and MacGregor, 1995; Wold et al., 1998). However, because of the inherent batch process complexity and nonstationary behavior, the batch duration, K_i, is not always the same and equally relevant, key process events do not occur at the same time point when comparing different NOC batch runs of the same process. This uneven and not synchronized batch data cannot be represented in this perfect three-dimensional data array, $\underline{\mathbf{X}}$, unless adjusted using different batch synchronization tools to cope with this problem (González-Martínez et al., 2014b).

Great progress has been made to develop strategies for batch alignment based on a maturity index or indicator variable coming directly from a process variable or estimated by PLS models or using more advanced algorithms, such as correlation optimized warping or dynamic time warping (Kassidas et al., 1998; Ramaker et al., 2004; González-Martínez et al., 2014a; Liu et al., 2017; Spooner and Kulahci, 2018; Zhao et al., 2020). Most of these methods were designed for the monitoring of finished batches using offline MSPC models and only an attempt proposed by (González-Martínez et al., 2011) described a method based on time warping that allows batch alignment for online MSPC.

Despite the methodologies mentioned above, having naturally non-synchronized batches is the most common situation in practice and batch alignment is a delicate operation that can induce artifacts in the definition of MSPC models when scarce information is available or when is not properly applied. Hence, the need for MSPC approaches that can circumvent the synchronization step for online process monitoring and control. Very few attempts have been carried out in this direction. (Rato et al., 2017) used the translation-invariant wavelet decomposition and PCA for the monitoring of the semiconductor manufacturing process. Another method based on a search grid capturing the batch trajectory in the PCA score space was proposed by (Westad et al., 2015) and was used for the monitoring of two industrial processes.

In this paper, a new synchronization-free approach of multivariate statistical process control (MSPC) for online monitoring and diagnostics of batch processes is introduced. It is based on the modeling of an overall NOC historical batch trajectory, defined by individual non-synchronized NOC batches, and the subsequent construction of derived PCA-based local MSPC models covering the complete process, i.e., the complete overall NOC batch trajectory. These local models are used to identify whether new batch observations are inside the NOC trajectory and, when this is the case, to provide an estimate of the process progress. The approach is illustrated using two real examples of NIR-monitored batch processes but is readily applicable for the online monitoring of batch processes of different typologies monitored by one or more diverse sensors.

PROCESS CASE STUDIES AND DATA SETS

Two case studies from previous works are used to illustrate and test the online batch MSPC models for tracking process trajectories. A brief experimental description of these NIR-monitored processes with the related spectral preprocessing implemented is presented below.

Process 1: Fluidized Bed Drying of Pharmaceutical Granules

Batches of 500-g pharmaceutical wet granules (dry mass fraction of mannitol > 50% and excipients) were dried in a 4-L fluidized bed (4M8-Trix Formatrix, ProCepT, Belgium). The fluidized bed air inlet flow was controlled at 0.6 or 0.85 m^3/min and a temperature range from 22 to 30°C. In-line NIR measurements were collected approximately every second using a spectrophotometer with a MEMS Fabry-Perot interferometer (N-Series 2.2, Spectral Engines, Finland) coupled to a diffuse reflectance immersion probe (OFS-6S- 100HO/080704/1, Solvias, Switzerland). The spectra covered a wavelength range from 1750 to 2150 nm at 1-nm intervals. For each batch, off-line reference moisture content analysis was carried out using a thermogravimetric moisture analyzer (MB120, Ohaus, Germany) from samples retrieved at 6-min intervals to detect

drying endpoint (moisture < 2%). Because of different process conditions at the beginning and during each batch run, such as inlet air temperature and flow, different batch durations were required for each trial to reach the defined <2% moisture level, therefore, providing data matrices with uneven lengths. Faulty batches used in the testing of the proposed approach did not reach this moisture level. Suitable preprocessing was employed to filter out noise and baseline fluctuations on the NIR raw data observations before data analysis. The preprocessing steps included the application of a moving average of consecutive NIR observations followed by standard normal variate (SNV) normalization. For a detailed description of the experimental procedure and the visualization of the spectral data, the reader is referred to (Avila et al., 2020; de Oliveira et al., 2020). Some batches were selected from the previous work and additional faulty batches were used for model validation. Ten NOC batches, NOC1 to NOC10, were used for MSPC model building, and three for validation (one NOC, Batch NOC1, and two faulty batches, Batch Fault1 and Batch Fault2). This is an example of a batch process where the evolution of drying in time is not synchronized among batches since the initial and final material in every batch does not necessarily have the same moisture level.

Process 2: Automated Benchtop Batch Gasoline Distillation

Batches of 100-ml gasoline blends (mixture of pure gasoline and ethanol) were distilled in an automated batch distillation device designed for the in-line monitoring of distilled product with NIR spectroscopy. For every batch, vapor temperature readings and in-line NIR absorption spectra (900–2600 nm with 4 cm^{-1} resolution; Rocket, ARCoptix ANIR, Switzerland) were recorded for every unit of percentage distilled mass fraction of initial sample weight, in the 5–90% range. Therefore, the data matrices obtained had the same number of NIR observations per batch (86 NIR spectra) and every observation was related to the same distillation process stage, as defined by the percentage (w/w) of distilled sample mass. The gasoline batches were prepared by mixing ethanol AR (99% Sigma-Aldrich) and pure gasoline (from Petrobras refinery, Brazil) at different volume ratios from 10 to 40%. Distillation batches of gasoline blends with 27% ethanol were defined as NOC batches and all batches with a different ratio as faulty, or out of specification according to Brazilian legislation. The preprocessing steps used in this data set were Savitzky-Golay derivative (1^{st}-order derivative, 2^{nd}-order polynomial function and 9-point window) for baseline correction followed by spectral normalization to mitigate signal intensity fluctuations of the NIR spectra. More detailed information related to the experiments and spectra preprocessing can be found elsewhere (de Oliveira et al., 2017). In this work, nine NOC distillation batches were used to build the MSPC control charts for tracking process trajectory (B1 to B3, B5 to B9 and B11), and three for validation, where one was NOC (B4) and two were faulty batches (B13, B19). In this case, batch process trajectories were synchronized because the percentage of distillation weight gives a direct reference for batch progress evolution.

DATA TREATMENT

The online batch MSPC model building procedure for tracking process evolution in synchronized or non-synchronized batch processes is described below. The complete methodology involves the following steps:

a) Modeling of NOC batch process trajectories.
b) Construction of local MSPC models based on NOC batch process trajectories.
c) Use of an MSPC chart based on local MSPC models to track the evolution of new batches.

The first two steps are involved in the generation of the MSPC models, whereas the last step involves the use of the local MSPC models on new batches to test whether they follow the NOC trajectory or to detect faults. A detailed description of each step is presented below together with a visual description of the approach in **Figure 1**.

Modeling of NOC Batch Process Trajectories

The evolution of NOC batches, a.k.a "golden batches", can be defined using different multivariate analysis modeling strategies, such as PCA, independent component analysis, multivariate curve resolution, parallel factor analysis, etc. (Haack et al., 2004; Mortensen and Bro, 2006; Skibsted et al., 2006; Bogomolov, 2011; de Oliveira et al., 2017; Gomes et al., 2019). In this work, we use PCA as the basis to define the general NOC batch process trajectory.

The NIR spectra obtained in a NOC batch i are structured in a data matrix $\mathbf{X}_i\,(K_i \times J)$, where K_i are the number of spectra collected (related to time points for *Process* 1 and to % of distillation for *Process 2*) and J are the NIR channels per spectrum.

When several NOC batches are used to define the general process trajectory, the data matrices from the different NOC batches, $\mathbf{X}_i\,(K_i \times J)$, are placed one on top of each other to build an augmented multiset structure $\mathbf{X}\,(N \times J)$, where N is the number of rows related to the total number of observations from the I NOC batches, that is, $N = \sum_i^I K_i$. Note that this strategy does not require resizing or synchronization of uneven batch lengths, since the only requirement is that all batches share a common spectral dimension, J (Wold et al., 1998). The next step is to column mean-center this multi-batch structure and analyze it with PCA. This centering operation is not oriented to remove the mean trajectory of the batches in time, just to center the data and remove the average spectral shape in order to see the spectral process variation already from the first PC.

Principal component analysis (PCA) is used to obtain a global model of batch trajectories explaining the overall NOC process evolution. PCA is used to reduce the dimensionality of the preprocessed spectral data into a low-dimensional subspace of principal components (PC's), orthogonal among them, that preserve the relevant information of the original data and explain the maximum non-random variance (Jolliffe, 2002).

The PCA model for the augmented process data matrix $\mathbf{X}(N \times J)$ is expressed as in **Eq. 1**,

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} \quad (1)$$

where $\mathbf{T}(N \times A)$ is formed by the scores matrix, related to the observations of the batch process data, $\mathbf{P^T}(A \times J)$ is the loadings matrix, related to the importance of the NIR variables in the description of the A PC's and $\mathbf{E}(N \times J)$ is the residual matrix after modeling. The number of principal components of the model, A, can be found using a suitable cross-validation method. The loading matrix, $\mathbf{P^T}$, is common to all batches and the augmented score matrix, $\mathbf{T}$, accommodates $\mathbf{T_i}$ blocks, related to every batch, that can be formed by a different number of observations, K_i. The multiset structure for three NOC batches and the related PCA model is illustrated in **Figure 1A** (top left), where λ represents the J spectral channels of the NIR spectra.

Construction of Local MSPC Models Based on NOC Batch Process Trajectories

From the augmented score matrix of all NOC batches, individual batch score trajectories can be overlapped on a scatter score plot, as shown in **Figure 1A** (bottom left). The dots represent the scores for each observation and are colored according to the NOC batches used in the PCA model. Note that the overall trajectory evolution is the same for all NOC batches, but in a general non-synchronized case, the starting and endpoint of every batch do not need to coincide. The overlapped individual batch process trajectories define a global description of the variability of the NOC process evolution, helpful to observe whether a new batch process evolves as NOC batches or not, independently from the batch length and dynamics.

The evolution described by the overlapped NOC trajectories can be divided into a sufficient number of C local regions using a cluster analysis methodology, such as k-means and fuzzy c-means clustering algorithms. In general, any algorithm allowing an even distribution of observations in the different clusters would be potentially valid in this step. The number of clusters used to set the local MSPC models will be closely related to the process progress resolution desired to study the batch evolution and will be limited by the number of available NOC observations. Hence, the higher the number of clusters, the higher process progress resolution will be obtained; however, care must be taken to avoid building local MSPC models with an insufficient number of observations that could lead to a non-representative description of the process stage to be controlled. **Figure 1A** (bottom right) illustrates these local regions for $C = 11$, as indicated by the outer circle color of the neighbor observations inside each cluster. The seeding information for the local MSPC models is formed by the observations in two consecutive clusters. Therefore, the first local MSPC model contains the observations in the first two clusters of the process trajectory, the second local MSPC model uses the observations in clusters two and three and so forth until all the NOC process trajectory is covered. The observations used in consecutive local MSPC models overlap with each other so that all process trajectory regions are covered. As can be seen in

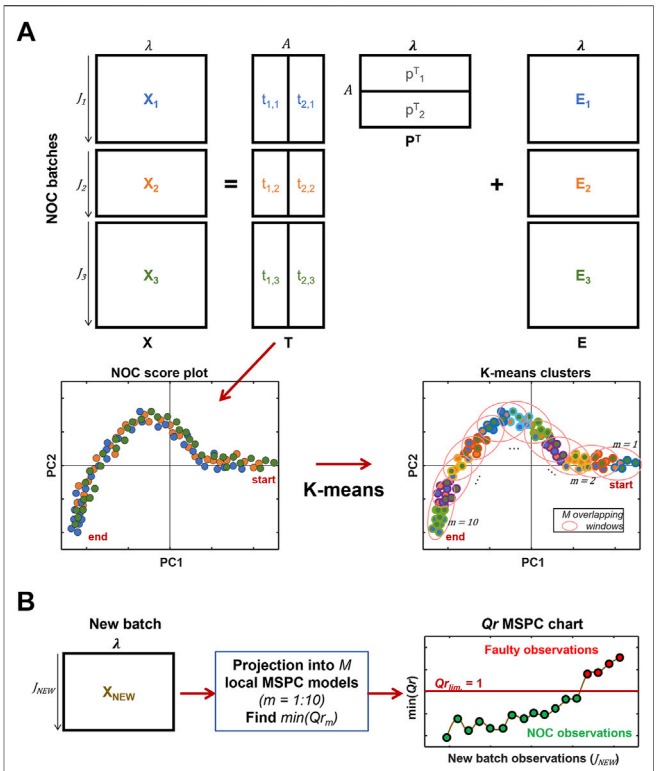

FIGURE 1 | Illustrative description of the different steps involved in the implementation of the local MSPC models for online monitoring of batch evolution. **(A)** PCA modeling of original batch process data for several NOC batches, visualization of process trajectories in the scatter score plot and definition of local regions in the NOC trajectory using k-means. **(B)** Monitoring of the evolution of a new batch using the projection of each observation onto the local MSPC models. The related reduced Q-statistics control chart is obtained plotting the minimum Qr value obtained in all M model projections per each observation.

Figure 1A, for a k-means analysis providing 11 clusters, 10 local MSPC models with overlapping information as defined by the red ellipses can be built.

The local MSPC models are built based on PCA and control chart limits are defined using the suitable local model statistics. The operational procedure to build each local MSPC model can be described as follows. First, the original observations, i.e. NIR spectra, for each local model are placed into a data matrix $\mathbf{X_m}(K_m \times J)$, where m indicates the index of the local model (from 1 to M) and K_m is the number of observations used to build the model. Then, this matrix is mean-centered and modeled with PCA, as in **Eq. 1**, generating the matrices of scores $\mathbf{T_m}(K_m \times \mathrm{A_m})$, loadings $\mathbf{P_m^T}(A_m \times J)$, and residuals $\mathbf{E_m}(K_m \times J)$. Note that the mean-center is performed using the mean of the matrix $\mathbf{X_m}$ not the global mean of the multi-bach structure. By doing so, since the local observations inside the matrix $\mathbf{X_m}$ should have similar spectral shape, the mean trajectory of the batch at that particular process stage is removed. Enough PC's, A_m, are included in each local model to provide the best fit using cross-validation (Wold, 1978). Finally, the control limits of the local control charts can be derived using the residuals and the

scores from the local PCA model (Rännar et al., 1998; Wold et al., 1998; Aguado et al., 2007). In this work, the controls charts are based only on the residual matrix, $\mathbf{E}_m$, deriving the Q-statistic control chart limit, Q_{lim}; however, other statistical parameters can readily be used to track the process evolution. The Q_{lim} is calculated according to the equation proposed by (Jackson and Mudholkar, 1979). Thus, once the local MSPC models and their related multivariate control charts limits are set, the online process evolution of new batches can be tracked based on the local models defined.

Use of an MSPC Chart Based on Local MSPC Models to Track New Batch Evolution

Calculation of squared residuals statistics (Q)

For online batch monitoring of new batch observations ($\mathbf{X_{NEW}}$ in **Figure 1B**), every new observation is projected onto all local MSPC models and a set of the related sum of squared residuals statistics, , are obtained as shown in **Figure 1B**. Thus, for every new online observation, $\mathbf{x}_k$ (a NIR spectrum in $\mathbf{X_{NEW}}$), its scores values, $\mathbf{t}_{k,m}$, are obtained for each local MSPC model using its related PCA loadings, $\mathbf{P}_m$, as follows,

$$\mathbf{t}_{k,m} = \mathbf{x}_k \mathbf{P}_m \quad (2)$$

Then, the residuals for the new observation in each local model are obtained as,

$$\mathbf{e}_{k,m} = \mathbf{x}_k - \mathbf{t}_{k,m} \mathbf{P}_m^T \quad (3)$$

And the related $Q_{k,m}$ as:

$$Q_{k,m} = \mathbf{e}_{k,m} \mathbf{e}_{k,m}^T \quad (4)$$

For an easier interpretation of the global multivariate control chart obtained from the outputs of the local MSPC models, reduced Q-statistics, $Qr_{k,m}$, are calculated by dividing the obtained $Q_{k,m}$ values by the related local model Q_{lim}. Thus, the control limits for all local MSPC models become equal to one, $Qr_{lim} = 1$. The reduced Q values for every new observation, $Qr_{k,m}$, are checked to see whether they are above or below the Qr_{lim}. If all $Qr_{k,m}$ values for the observation k are large and above one, this observation is diagnosed as faulty, and it is an indicator that the process is deviating from the NOC trajectory. Conversely, if one or more $Qr_{k,m}$ values are below the control limit, the observation follows the NOC trajectory. An easy way to visualize the diagnostic of every new observation by using a single Q chart is shown in **Figure 1B** (bottom right), where only the minimum Qr parameter after the projection in all local models is displayed for every new observation. Observations that follow the NOC trajectory are depicted by the green dots below the $Qr_{lim} = 1$, and the eventual deviations from it, with $\min(Qr_{k,m}) > 1$, in red. To assess the spectral variables making the greatest contributions to the deviation in Q we can display the Q-statistics contribution plots for the sought observation by plotting the elements of the residual vector, $\boldsymbol{e_{k,m}}$. The residuals used for the contribution plots are calculated using the best local MSPC model related to the last NOC observation.

For NOC observations, it is also possible to estimate the process stage of every observation by identifying the local MSPC model providing the lowest $Qr_{k,m}$ value. This visualization approach will be provided for the real process applications studied in this work in the next section.

RESULTS AND DISCUSSION

In this section, the results related to the construction of NOC trajectories and local MSPC models for each process case study are shown. Afterwards, the resulting MSPC charts for the online monitoring of new NOC and faulty batches are shown for each process. Complementary visualization of MSPC charts and fault diagnostics based on contribution plots are also presented.

Construction of NOC Trajectories and Local MSPC Models

The construction of PCA-based NOC trajectories for each process was calculated as explained in *step a* of the Data Treatment section using the training dataset, i.e. all NIR observations from selected complete NOC batches. This step was followed by k-means analysis on the overlapped individual NOC batch trajectories to define the clusters used to build the local MSPC models covering the overall NOC process trajectory

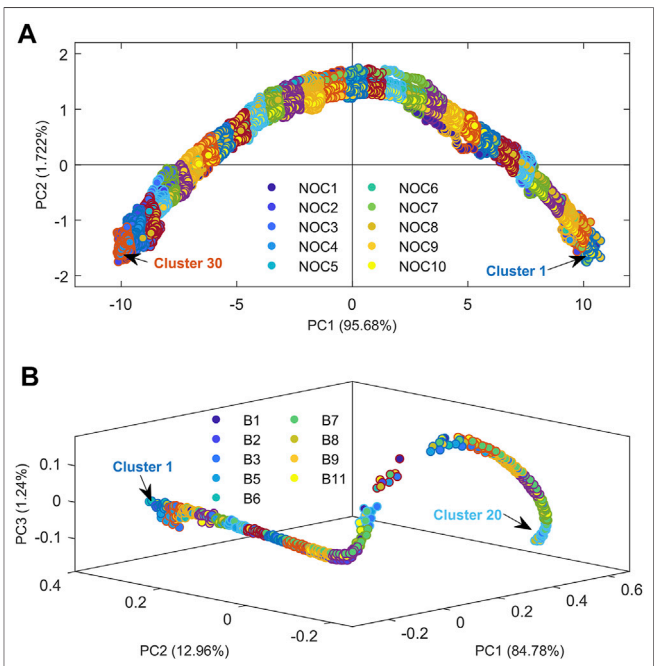

FIGURE 2 | PCA score plot for the online NIR observations showing the NOC batch process trajectories and local clusters found by k-means for **(A)** *Process 1*, fluidized bed drying, and **(B)** *Process 2*, gasoline blend distillation. The inner part of the circles is colored according to the related NOC batch, whereas the outer part reflects the observations included in every cluster and, hence, in the related local MSPC model.

(Data Treatment section *step b*). **Figure 2** shows the PCA score scatter plot and the k-means clusters used to build the local MSPC models describing the overall NOC batch process trajectories for the drying (*Process 1*) and the distillation processes (*Process 2*).

Principal Component Analysis of the NOC batches from *Process 1* (Fluidized bed drying) allowed description of the process evolution using only two PC's explaining a total of 97.61% of the data variance, as shown in the score plot of **Figure 2A**. The score plot described mostly the variation of the moisture content with the drying evolution from beginning to end of every NOC batch. Note that, because each batch had different initial and final moisture conditions, they started and finished at different points of the overall NOC trajectory; however, all individual batch trajectories followed the same evolution pattern, as shown in the PCA score plot. Once the overall NOC trajectory was defined, 30 clusters were defined using the k-means analysis along this trajectory, as displayed by the different outer circle colors associated with the observations inside each cluster in **Figure 2A**. For this example, 30 clusters and, hence, 29 local MSPC models, were considered sufficient to track in detail the process evolution. After that, a number indicating the process stage evolution was automatically assigned to each cluster according to the position in the overall NOC trajectory.

For *Process 2* (Distillation), three components were required by PCA to explain 98.99% of NOC batches variance because of the complex gasoline sample and the continuous variation of the distilled material composition. The complex overall NOC trajectory associated with the distillation process is shown in the 3 PC score scatter plot in **Figure 2B**. Despite the higher complexity of the overall NOC trajectory linked to the distillation process, all individual batches trajectories followed the same evolution pattern with good reproducibility. In contrast to the drying process, the NIR observations of the distillation process were acquired at specific percentages of distillation weight; therefore, the observations were naturally synchronized according to the process evolution. Note that all batches started and finished at the same point of the overall NOC batch trajectory in the score plot. The k-means algorithm applied on the PCA scores of **Figure 2B** was used to set 20 clusters along the overall NOC batch trajectory, as displayed in **Figure 2B**. The number of clusters is lower than in the previous example because of the limited number of available observations per batch run (only 86) and the need to avoid having clusters with a very low number of observations to build the local MSPC models.

Once the overall NOC batch process trajectories were defined for each process case, the original NIR observations inside the suitable two consecutive k-means clusters were used as seeding information to build local MSPC models for each step of the batch trajectory, as described in the Data Treatment section (*step b*). Thus, a total of 29 and 19 local PCA-based MSPC models were built for *Processes* 1 and 2, respectively. Local MSPC control chart limits based on the *Q*-statistics with a 99% confidence interval were calculated for each local MSPC model to be used for the online tracking of new batches evolution, as shown in the next subsection.

Online Tracking of New Batch Evolution with Local MSPC Models

The results of the use of local MSPC models for the online tracking of new batch evolution are described separately for each process case, as shown below. The new batches used were identified in previous studies as NOC or faulty; therefore, they will be useful to demonstrate and validate the proposed methodology.

Application to Process 1 (Fluidized Bed Drying)

The tracking of every observation in new fluidized bed drying batches was performed as described in the Data treatment section (*step c*), using the 29 local MSPC models built as explained above (**Supplementary Figure S1** and a related animation **Supplementary Figure S2** of the help to display how the *Qr* values issued from every MSPC local model are obtained for every observation in a batch).

The *Qr*-based MSPC control charts for the online tracking of observations in two drying batches are shown in **Figure 3**. **Figure 3A**; **Figure 3C** are contour plots related to validation Batch NOC1 and Batch Fault1, respectively, that show all the *Qr* values calculated after the projection of each online NIR observation of the batch onto all local MSPC models. A log-scale colormap has been used to highlight the differences at low *Qr* values. The horizontal axis of the contour plot represents the batch time at which every observation was collected and the right vertical axis the indices related to the local MSPC model used to describe the *Process 1* NOC batch trajectory, i.e. from 1 to 29. Additionally, in the left vertical axis, each local MSPC model index is associated with a percentage of the process progress from 0–100%, defined making a linear scaling that links the initial local model to 0% process progress and the final local model to 100% process progress. The process progress in this approach plays the same role as the process maturity concept proposed by other authors (Wold et al., 1998; Westad et al., 2015).

Thus, to track the behavior of an observation of a new batch, their related *Qr* values (associated with a specific process time) are examined. In the contour plots in **Figure 3A**; **Figure 3C**, the *Qr* values below the control limit, i.e. $Qr < 1$, are depicted as blue dots and the min ($Qr < 1$) for every observation in green. If an observation shows a NOC behavior (as all do in **Figure 3A** related to Batch NOC1), there will always be one or more *Qr* values below 1; i.e., all observations will show one or more blue dots and a green dot. Instead, when an observation deviates from the NOC trajectory, as in Batch Fault1 (**Figure 3C**), all *Qr* values related to that observation are above the control limit of 1 and neither blue nor green dots are observed.

To facilitate the interpretation and summarize the relevant information of the results in the contour plots, graphics displaying the min (*Qr*) value and the related process progress for every batch observation are proposed (see **Figure 3B** and **Figure 3D** for batches NOC1 and Fault1, respectively). **Figure 3B** shows that all observations for batch NOC1 followed the NOC

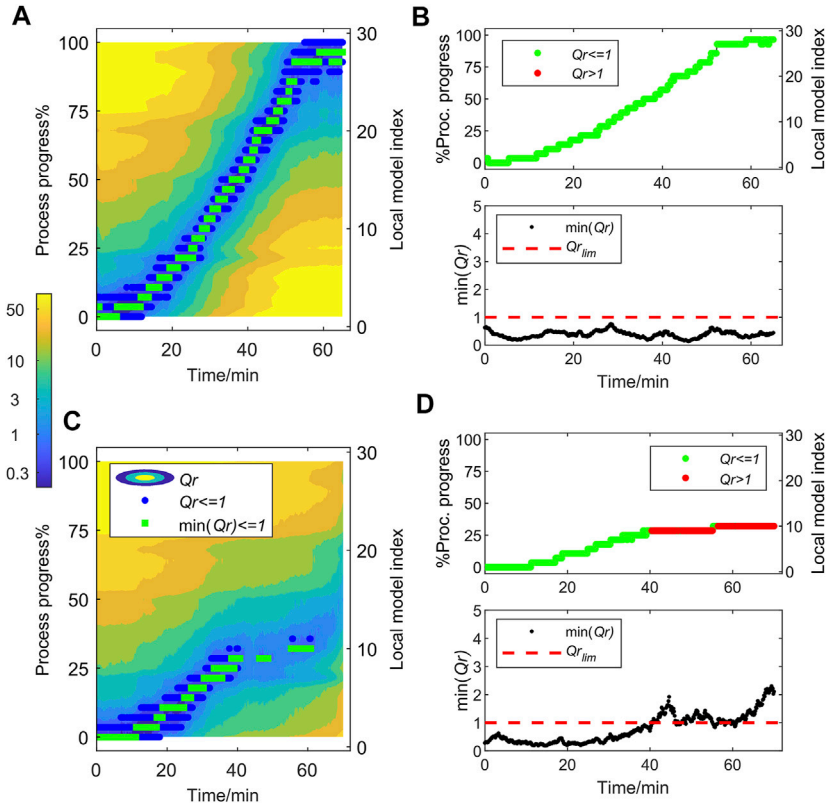

FIGURE 3 | *Qr*-based MSPC charts for fluidized bed drying (Process 1) batch NOC1 **(A and B)** and batch Fault1**(C and D)**. **(A and C)** Contour plots of the *Qr* values calculated after the projection of each NIR observation onto the local MSPC models. Blue dots show values of $Qr < 1$ (control limit), green squares the min ($Qr < 1$). **(B and D)** Charts show the min (*Qr*) value (bottom panel) and the related process progress associated with it (top panel) for every batch observation. In the process progress plot, NOC observations are displayed in green and faulty observations in red.

batch trajectory, seen because all min (*Qr*) values were below the control limit of 1 (bottom panel) and that the process progress covered the complete range (0–100%) (top panel). **Figure 3D** shows that batch Fault1 deviated from the NOC trajectory after approximately 40 min of batch time as flagged by the *Qr* above the local MSPC control limits (min (*Qr*) > 1) (bottom plot). When a fault happens, the related observations are displayed in red in the process progress plot to indicate that the evolution of the process is abnormal (top plot).

Detailed results and interpretation of the abnormal behavior for the online tracking of two faulty batches, Fault1 and Fault2, are shown in the **Supplementary Figure S3**; **Figure 4** (left plots), respectively. **Supplementary Figure S3A**; **Figure 4A** show the deviations of the two batches by displaying the score plot projections of NIR observations of these new batches onto the global PCA model used to describe the NOC batch trajectory. The score plot shows all training NOC batch trajectories as gray dots whereas the NOC observations from the new batches are overlayed as green dots when identified as NOC and as red dots when faulty. **Supplementary Figures S3B, S3C**; **Figure 4B** show the batch process progress and min (*Qr*) MSPC chart for the tracking of the online observations, where the abnormal observations are associated with min (*Qr*) values higher than 1 and flagged in red color in the process progress plot. Moreover, *Q* contribution plots from two faulty observations selected from each batch are shown in **Supplementary Figure S3D**; **Figure 4C**. The contribution plots were used to understand the reasons for the deviations from the NOC batch trajectory, as described below for each batch.

The deviation of drying batch Fault1 from the NOC trajectory was detected after approximately 40 min of batch time, see **Supplementary Figures S3B, S3C**. Although in **Supplementary Figure S3A** the faulty observations (red dots) right after 40 min were still close to the NOC trajectory, the related min (*Qr*) after projection onto local MSPC models was above the control limit indicating a deviation, which became even larger after ca. 65 min of batch time, see **Supplementary Figure S3C**. To help to diagnose this deviation, contribution plots are shown in **Supplementary Figure S3D** for two faulty observations selected at 64 and 69 min of batch time. These observations are marked in blue and orange squares in the score plot and MSPC charts. The *Q* contribution plots show that the absorption bands that gave higher contributions to *Q* were around 1750 and 1900 nm related to the 1st overtone of CH and OH bonds. No clear trend was observed when comparing the contribution plots of the two observations suggesting that this deviation may have been caused by changes of heterogeneity or particle comminution of the pharmaceutical granules.

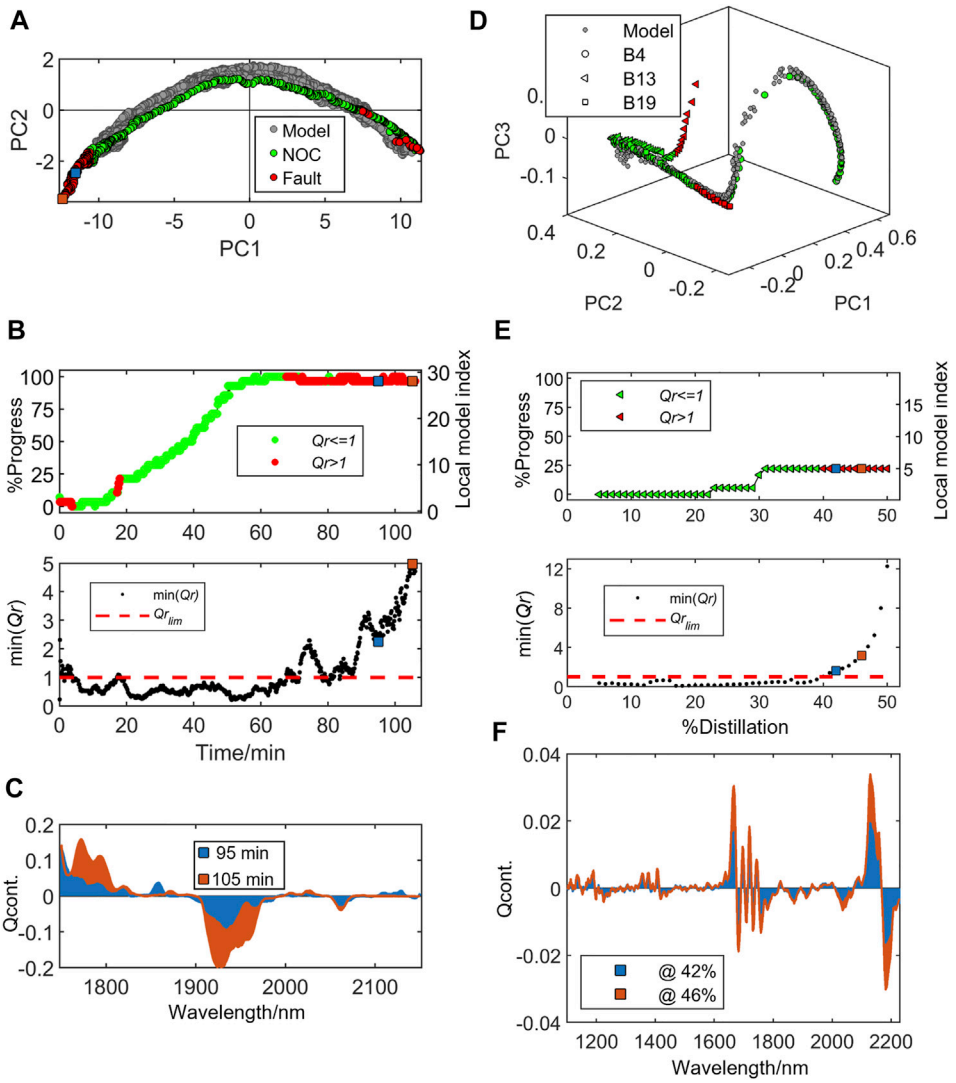

FIGURE 4 | Online tracking of new batch evolution using the local MSPC models for process 1 (fluidized bed drying) left plots **(A to C)** and process2 (gasoline blend distillation) right plots **(D to F)**. **(A)** Process 1 PCA score plot showing the training NOC batches trajectories (gray dots) and validation batch Fault2 trajectory in green (NOC observations) and red dots (faulty observations). **(B)** MSPC chart showing the process progress (top panel) and min (*Qr*)-based MSPC chart. **(C)** *Q* contribution (*Qcont*.) plots for two faulty observations at 95 and 105 min of batch Fault2 drying time represented by the blue and orange squares in the MSPC control chart. **(D)** Process 2 PCA score plot showing the training NOC batches trajectories (gray dots) and three validation batches (B4 circles, B13 triangles and B19 squares) trajectories in green (NOC observations) and red dots (faulty observations). **(E)** MSPC chart showing faulty batch B13 process progress (top panel) and min (*Qr*)-based MSPC chart. **(F)** *Q* contribution (*Qcont*.) plots for two faulty observations at 42 and 46% of distillation represented by the blue and orange squares in the MSPC control chart. Green and red marker face color in process progress chart indicate that the observation is inside or outside the trajectory confidence limits, respectively.

During the tracking of the additional batch, Fault2, three clusters of faulty observations were detected, see **Figure 4B**. The first faulty observations were detected during the first few minutes of the batch process. This deviation was related to the initial moisture content higher than the common starting point for the NOC batches used to build the MSPC models at the beginning of the process trajectory. However, after a few minutes of drying, the online observations fell inside the confidence interval. The second faulty situation occurred after ca. 18 min of batch time during just four consecutive observations, but it quickly returned inside the control limit. This probably was related to a fast change of moisture content sensed by the NIR probe due to granule heterogeneity. This can be noticed by the fast change in process progress just before minute 20 in **Figure 4B** (top panel). From this point until approximately 60 min of batch time, the batch followed the NOC trajectory reaching 100% of batch progress, that is, reaching the minimum moisture level of the NOC batches used to train the local MSPC models at the end of the process trajectory, see **Figure 4** (top panel). However, this batch was left to overdry reaching moisture levels lower than the endpoint of the historical NOC batches used for model training. The consequence of this action was successfully detected after

approximately 70 min of the batch time by the MSPC chart, **Figure 4** (bottom panel), where almost all consecutive observations were above the control limit. Looking at the bottom left of **Figure 4A** it can be observed how the PCA projections of these faulty observations were outside the NOC trajectory, but still following the drying process trend. Finally, two faulty observations at the end of this validation batch (at 100 and 105 min) were selected to check the contribution plots. These observations are marked in blue and orange squares in the score plot and MSPC charts. The *Q* contribution plots in **Figure 4C** show that the absorption bands that contributed more to *Q* were around 1750 and 1950 nm related to the 1st overtone of CH and OH bonds, respectively, being the band at 1950 nm identified generally as the most dominant water band. The *Q* contribution positive and negative sign for the bands at 1750 and 1950 nm, respectively, indicates that the moisture level for these two observations was lower than the endpoint of the historical batches used in the model building. Also, when comparing the two faulty contribution plots, the systematic growth of the *Q* contributions at 1750 and 1950 nm bands, indicates the continuing moisture content decrease. It is important to note that this overdrying batch was used in this work to demonstrate the ability of the local MSPC models to detect such situations. In real-time monitoring, this batch would have been terminated once reached 100% of process progress, thus, avoiding energy waste and possible detrimental effects due to the excessive granules processing time.

Application to Process 2 (Gasoline Distillation)

The local MSPC models built to track the batch gasoline distillation were tested. Three validation batches were used: one batch of on-specification gasoline blend with 27% of ethanol (batch B4) and two off-specification gasoline distillation batches, B13 and B19, with 15 and 30% ethanol blends, respectively. The results for all testing batches are shown in **Figure 4** (right plots) and **Supplementary Figure S4**.

The scatter score plot projections of the NIR observations for all three validation batches in the global PCA model used to build the *Process 2* NOC batch trajectory are represented in **Figure 4D** (same as **Supplementary Figure S4A**). In the score plot, gray dots identify the observations from the training batches describing the NOC batch trajectory, while the circles, triangles and squares are the projected observations from testing batches B4, B13 and B19, respectively. For the testing batches, the symbol face color indicates whether the observation was detected by the MSPC charts as faulty (red) or not (green). Process progress and min (*Qr*) MSPC charts for the testing batches are shown in **Figure 4E** for batch B13 and **Supplementary Figures S4B, S4C** for batches B4 and B19, respectively. Additionally, *Q* contribution plots for two selected faulty observations are shown in **Figure 4F**; **Supplementary Figure S4D** for batches B13 and B19, respectively.

The projections of the validation batch B4 in the global PCA model (**Supplementary Figure S4A**) followed the NOC batch trajectory described by the cloud of gray dots. Indeed, when looking at the MSPC charts in **Supplementary Figure S4B**, all observations are below the *Qr* control limit and the batch process progressed accordingly to the on-specification gasoline batches. On the other hand, when looking at the projections of batch B13 observations to the global PCA model, an obvious deviation of the NOC batch trajectory was observed, see the red triangles in **Figure 4D**. This deviation was detected by the min (*Qr*) local MSPC charts (**Figure 4E** bottom panel) after 40% of the initial batch weight was distilled. Note the interruption of the process progress after this point and all consecutive observations. The off-specification batch B19 deviation from the NOC batch trajectory was lightly noticed by the PCA score plot projections in **Supplementary Figure S4A** (red squares). However, this batch deviation was still detected by the local MSPC charts in **Supplementary Figure S4C** (bottom panel). Note that this sensitivity is important since batch B19 contains 30% alcohol (v/v), only a 3% more than the NOC batches. Similarly, the fault was first detected after ca. 40% of the distillation batch and all consecutive observations since then were detected outside the confidence interval for all local MSPC models.

The contribution plots (**Figure 4F**) for the selected fault observations at 42% (in blue) and 46% (in orange) fraction of distilled material of the B13 batch show that the two bands covering the 1650–1700 nm and 2100–2200 nm NIR contributed the most to the *Q*. The absolute increase of *Q* contributions at 1665, 2130 and 2180 nm indicated a possible increment of mid and high-density hydrocarbon fractions at these distillation points. Additionally, the negative contribution at 1685 nm indicated a lower content of ethanol and light hydrocarbon compounds. This agrees with the expected distillation behavior for off-specification gasoline blends with low ethanol content. This is confirmed when looking at the distillation profiles obtained by Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) for these compounds presented in our previous work for this specific batch (de Oliveira et al., 2017). For batch B19, **Supplementary Figure S4D** shows the contribution plots for the faulty observation at 44% (in blue) and 50% (in orange) of the batch distillation. The high negative contribution between 1680 and 1700 nm suggested the presence of a lower content of mid and heavy hydrocarbons fraction than expected for NOC batches at this point of distillation. These ethanol-rich fractions were related to the fact that this off-specification gasoline batch had a slightly higher ethanol content (30%) than NOC gasolines (27%).

CONCLUSION

The present work introduces a new approach for online monitoring of spectroscopic-monitored batch process evolution through the design of local MSPC models covering an overall NOC batch process trajectory, defined from the PCA modeling of non-synchronized NOC batches. The key element in this approach is that the different NOC batches follow a similar NOC trajectory in the PCA score map and this fact is clearly visible and can be used to build derived MSPC models without the need of batch synchronization. The tracking of the evolution of new batches does not require synchronization either. The methodology has been demonstrated with the building and

validation of online MSPC charts for the monitoring of two real batch process data of different nature using *in-situ* NIR measurements. In both process examples, the implementation of local MSPC charts has been successfully validated for the tracking of well-known new batches that followed or deviated from the overall NOC batch trajectory. The use of *Q* contribution plots was helpful to identify the sources of process abnormalities based on the chemical information provided by the NIR signal.

The fact that the proposed methodology does not require batch synchronization makes the data analysis pipeline simpler and flexible and offers many advantages for real-time process monitoring, from the building of the reference MSPC models to the test of new batches. Thus, the designed methodology allows the model building with historical NOC process data acquired with different online sampling rates and spanning evolution in different time (or process variable) ranges. The monitoring of new batches is also independent of the sampling rate used in the model building, which allows for changes in the sampling interval if required. Furthermore, the fact that the exam of the quality of new batch observations provides additionally a good indication of the process progress enables the potential use of this online tracking methodology for end-point detection, providing a single tool to control both the evolution and the end of the process. The presented methodology has been applied to NIR monitored processes but could be readily adapted to deal simultaneously with the output from several sensor outputs in a sensor fusion scenario, since a common trajectory for NOC batches would be seen. That would allow an integral control of the process evolution by combining the output from advanced sensors with other process data (temperature, flow, pressure, etc.).

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Distillation process data are available upon request to the authors. Drying data are not available due to confidential reasons. Requests to access these datasets should be directed to RR, rodrigo.rocha@ub.edu.

AUTHOR CONTRIBUTIONS

RR: Conceptualization, Methodology, Investigation, Data Curation, Visualization, Software, Formal analysis, Writing—original draft, Writing—review and editing. AJ: Conceptualization, Methodology, Supervision, Formal analysis, Writing—original draft, Writing—review and editing, Funding acquisition.

FUNDING

Spanish government. PID 2019-1071586B-IOO. Catalan Government: Excellence research group (2017 SGR 753).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frans.2021.772844/full#supplementary-material

REFERENCES

Aguado, D., Ferrer, A., Ferrer, J., and Seco, A. (2007). Multivariate SPC of a Sequencing Batch Reactor for Wastewater Treatment. *Chemom. Intell. Lab. Syst.* 85, 82–93. doi:10.1016/j.chemolab.2006.05.003

Ávila, T. C., Poppi, R. J., Lunardi, I., Tizei, P. A. G., and Pereira, G. A. G. (2012). Raman Spectroscopy and Chemometrics Foron-Linecontrol of Glucose Fermentation bySaccharomyces Cerevisiae. *Biotechnol. Prog.* 28, 1598–1604. doi:10.1002/btpr.1615

Avila, C. R., Ferré, J., de Oliveira, R. R., de Juan, A., Sinclair, W. E., Mahdi, F. M., et al. (2020). Process Monitoring of Moisture Content and Mass Transfer Rate in a Fluidised Bed with a Low Cost Inline MEMS NIR Sensor. *Pharm. Res.* 37, 84. doi:10.1007/s11095-020-02787-y

Avila, C., Mantzaridis, C., Ferré, J., Rocha de Oliveira, R., Kantojärvi, U., Rissanen, A., et al. (2021). Acid Number, Viscosity and End-point Detection in a Multiphase High Temperature Polymerisation Process Using an Online Miniaturised MEMS Fabry-Pérot Interferometer. *Talanta* 224, 121735. doi:10.1016/j.talanta.2020.121735

Besenhard, M. O., Scheibelhofer, O., François, K., Joksch, M., and Kavsek, B. (2018). A Multivariate Process Monitoring Strategy and Control Concept for a Small-Scale Fermenter in a PAT Environment. *J. Intell. Manuf.* 29, 1501–1514. doi:10.1007/s10845-015-1192-8

Bogomolov, A. (2011). Multivariate Process Trajectories: Capture, Resolution and Analysis. *Chemom. Intell. Lab. Syst.* 108, 49–63. doi:10.1016/j.chemolab.2011.02.005

Cimander, C., and Mandenius, C.-F. (2004). Bioprocess Control from a Multivariate Process Trajectory. *Bioproc. Biosyst. Eng.* 26, 401–411. doi:10.1007/s00449-003-0327-z

Colucci, D., Prats-Montalbán, J. M., Fissore, D., and Ferrer, A. (2019). Application of Multivariate Image Analysis for On-Line Monitoring of a Freeze-Drying Process for Pharmaceutical Products in Vials. *Chemom. Intell. Lab. Syst.* 187, 19–27. doi:10.1016/j.chemolab.2019.02.004

de Oliveira, R. R., Pedroza, R. H. P., Sousa, A. O., Lima, K. M. G., and de Juan, A. (2017). Process Modeling and Control Applied to Real-Time Monitoring of Distillation Processes by Near-Infrared Spectroscopy. *Anal. Chim. Acta* 985, 41–53. doi:10.1016/j.aca.2017.07.038

de Oliveira, R. R., Avila, C., Bourne, R., Muller, F., and de Juan, A. (2020). Data Fusion Strategies to Combine Sensor and Multivariate Model Outputs for Multivariate Statistical Process Control. *Anal. Bioanal. Chem.* 412, 2151–2163. doi:10.1007/s00216-020-02404-2

Ferrer-Riquelme, A. J. (2009). "Statistical Control of Measures and Processes," in *Comprehensive Chemometrics*, 97–126. doi:10.1016/B978-044452701-1.00096-X

França, L., Grassi, S., Pimentel, M. F., and Amigo, J. M. (2021). A Single Model to Monitor Multistep Craft Beer Manufacturing Using Near Infrared Spectroscopy and Chemometrics. *Food Bioprod. Process.* 126, 95–103. doi:10.1016/j.fbp.2020.12.011

Gomes, F. P. C., Garg, A., Mhaskar, P., and Thompson, M. R. (2019). Data-Driven Advances in Manufacturing for Batch Polymer Processing Using Multivariate Nondestructive Monitoring. *Ind. Eng. Chem. Res.* 58, 9940–9951. doi:10.1021/acs.iecr.8b05675

González-Martínez, J. M., Ferrer, A., and Westerhuis, J. A. (2011). Real-time Synchronization of Batch Trajectories for On-Line Multivariate Statistical Process Control Using Dynamic Time Warping. *Chemom. Intell. Lab. Syst.* 105, 195–206. doi:10.1016/j.chemolab.2011.01.003

González-Martínez, J. M., de Noord, O. E., and Ferrer, A. (2014a). Multisynchro: A Novel Approach for Batch Synchronization in Scenarios of Multiple Asynchronisms. *J. Chemom.* 28, 462–475. doi:10.1002/cem.2620

González-Martínez, J. M., Vitale, R., De Noord, O. E., and Ferrer, A. (2014b). Effect of Synchronization on Bilinear Batch Process Modeling. *Ind. Eng. Chem. Res.* 53, 4339–4351. doi:10.1021/ie402052v

Grassi, S., Strani, L., Casiraghi, E., and Alamprese, C. (2019). Control and Monitoring of Milk Renneting Using FT-NIR Spectroscopy as a Process Analytical Technology Tool. *Foods* 8, 405. doi:10.3390/foods8090405

Haack, M. B., Eliasson, A., and Olsson, L. (2004). On-line Cell Mass Monitoring of *Saccharomyces cerevisiae* Cultivations by Multi-Wavelength Fluorescence. *J. Biotechnol.* 114, 199–208. doi:10.1016/j.jbiotec.2004.05.009

Jackson, J. E., and Mudholkar, G. S. (1979). Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics* 21, 341–349. doi:10.1080/00401706.1979.10489779

Jolliffe, I. T. (2002). *Principal Components Analysis*. 2nd ed. New York: Springer.

Kassidas, A., Macgregor, J. F., and Taylor, P. A. (1998). Synchronization of Batch Trajectories Using Dynamic Time Warping. *Aiche J.* 44, 864–875. doi:10.1002/aic.690440412

Kourti, T. (2005). Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry. *Int. J. Adapt. Control. Signal. Process.* 19, 213–246. doi:10.1002/acs.859

Liu, Y.-J., André, S., Saint Cristau, L., Lagresle, S., Hannas, Z., Calvosa, É., et al. (2017). Multivariate Statistical Process Control (MSPC) Using Raman Spectroscopy for In-Line Culture Cell Monitoring Considering Time-Varying Batches Synchronized with Correlation Optimized Warping (COW). *Anal. Chim. Acta* 952, 9–17. doi:10.1016/j.aca.2016.11.064

Mortensen, P. P., and Bro, R. (2006). Real-time Monitoring and Chemical Profiling of a Cultivation Process. *Chemom. Intell. Lab. Syst.* 84, 106–113. doi:10.1016/j.chemolab.2006.04.022

Nomikos, P., and MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37, 41–59. doi:10.1080/00401706.1995.10485888

Pöllänen, K., Häkkinen, A., Reinikainen, S.-P., Rantanen, J., and Minkkinen, P. (2006). Dynamic PCA-Based MSPC Charts for Nucleation Prediction in Batch Cooling Crystallization Processes. *Chemom. Intell. Lab. Syst.* 84, 126–133. doi:10.1016/j.chemolab.2006.04.016

Rännar, S., MacGregor, J. F., and Wold, S. (1998). Adaptive Batch Monitoring Using Hierarchical PCA. *Chemom. Intell. Lab. Syst.* 41, 73–81. doi:10.1016/S0169-7439(98)00024-0

Ramaker, H.-J., Van Sprang, E. N. M., Westerhuis, J. A., and Smilde, A. K. (2004). The Effect of the Size of the Training Set and Number of Principal Components on the False Alarm Rate in Statistical Process Monitoring. *Chemom. Intell. Lab. Syst.* 73, 181–187. doi:10.1016/j.chemolab.2003.12.015

Rato, T. J., and Reis, M. S. (2020). An Integrated Multiresolution Framework for Quality Prediction and Process Monitoring in Batch Processes. *J. Manuf. Syst.* 57, 198–216. doi:10.1016/j.jmsy.2020.09.007

Rato, T. J., Blue, J., Pinaton, J., and Reis, M. S. (2017). Translation-Invariant Multiscale Energy-Based PCA for Monitoring Batch Processes in Semiconductor Manufacturing. *IEEE Trans. Automat. Sci. Eng.* 14, 894–904. doi:10.1109/TASE.2016.2545744

Rendall, R., Chiang, L. H., and Reis, M. S. (2019). Data-driven Methods for Batch Data Analysis - A Critical Overview and Mapping on the Complexity Scale. *Comput. Chem. Eng.* 124, 1–13. doi:10.1016/j.compchemeng.2019.01.014

Skibsted, E. T. S., Boelens, H. F. M., Westerhuis, J. A., Witte, D. T., and Smilde, A. K. (2006). Simple Assessment of Homogeneity in Pharmaceutical Mixing Processes Using a Near-Infrared Reflectance Probe and Control Charts. *J. Pharm. Biomed. Anal.* 41, 26–35. doi:10.1016/j.jpba.2005.10.009

Spooner, M., and Kulahci, M. (2018). Monitoring Batch Processes with Dynamic Time Warping and K-Nearest Neighbours. *Chemom. Intell. Lab. Syst.* 183, 102–112. doi:10.1016/j.chemolab.2018.10.011

van Sprang, E. N. M., Ramaker, H.-J., Westerhuis, J. A., Gurden, S. P., and Smilde, A. K. (2002). Critical Evaluation of Approaches for On-Line Batch Process Monitoring. *Chem. Eng. Sci.* 57, 3979–3991. doi:10.1016/S0009-2509(02)00338-X

Vidal-Puig, S., Vitale, R., and Ferrer, A. (2019). Data-driven Supervised Fault Diagnosis Methods Based on Latent Variable Models: a Comparative Study. *Chemom. Intell. Lab. Syst.* 187, 41–52. doi:10.1016/j.chemolab.2019.02.006

Westad, F., Gidskehaug, L., Swarbrick, B., and Flåten, G. R. (2015). Assumption Free Modeling and Monitoring of Batch Processes. *Chemom. Intell. Lab. Syst.* 149, 66–72. doi:10.1016/j.chemolab.2015.08.022

Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A. (1998). Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments. *Chemom. Intell. Lab. Syst.* 44, 331–340. doi:10.1016/S0169-7439(98)00162-2

Wold, S., Kettaneh-Wold, N., MacGregor, J. F., and Dunn, K. G. (2009). "Batch Process Modeling and MSPC," in *Comprehensive Chemometrics* (Elsevier), 163–197. doi:10.1016/B978-044452701-1.00108-3

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* 20, 397–405. doi:10.1080/00401706.1978.10489693

Zhao, J., Li, W., Qu, H., Tian, G., and Wei, Y. (2020). Real-time Monitoring and Fault Detection of Pulsed-spray Fluid-Bed Granulation Using Near-Infrared Spectroscopy and Multivariate Process Trajectories. *Particuology* 53, 112–123. doi:10.1016/j.partic.2020.02.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 15 March 2022
doi: 10.3389/fchem.2022.818974

Hyperspectral Video Analysis by Motion and Intensity Preprocessing and Subspace Autoencoding

Raffaele Vitale[1]*, *Cyril Ruckebusch*[1], *Ingunn Burud*[2] and *Harald Martens*[3,4]

[1]Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Lille, France, [2]Faculty of Science and Technology, Norwegian University of Life Sciences, Oslo, Norway, [3]Idletechs AS, Trondheim, Norway, [4]Department of Engineering Cybernetics, Norwegian University of Science and Technology, Trondheim, Norway

Hyperspectral imaging has recently gained increasing attention from academic and industrial world due to its capability of providing both spatial and physico-chemical information about the investigated objects. While this analytical approach is experiencing a substantial success and diffusion in very disparate scenarios, far less exploited is the possibility of collecting sequences of hyperspectral images over time for monitoring dynamic scenes. This trend is mainly justified by the fact that these so-called hyperspectral *videos* usually result in BIG DATA sets, requiring TBs of computer memory to be both stored and processed. Clearly, standard chemometric techniques do need to be somehow adapted or expanded to be capable of dealing with such massive amounts of information. In addition, hyperspectral video data are often affected by many different sources of variations in sample chemistry (for example, light absorption effects) and sample physics (light scattering effects) as well as by systematic errors (associated, *e.g.*, to fluctuations in the behaviour of the light source and/or of the camera). Therefore, identifying, disentangling and interpreting all these distinct sources of information represents undoubtedly a challenging task. In view of all these aspects, the present work describes a multivariate hybrid modelling framework for the analysis of hyperspectral videos, which involves spatial, spectral and temporal parametrisations of both known and unknown chemical and physical phenomena underlying complex real-world systems. Such a framework encompasses three different computational steps: 1) motions ongoing within the inspected scene are estimated by optical flow analysis and compensated through IDLE modelling; 2) chemical variations are quantified and separated from physical variations by means of Extended Multiplicative Signal Correction (EMSC); 3) the resulting light scattering and light absorption data are subjected to the On-The-Fly Processing and summarised spectrally, spatially and over time. The developed methodology was here tested on a near-infrared hyperspectral video of a piece of wood undergoing drying. It led to a significant reduction of the size of the original measurements recorded and, at the same time, provided valuable information about systematic variations generated by the phenomena behind the monitored process.

Keywords: hyperspectral videos, motion compensation, IDLE modelling, light scattering, light absorption, extended multiplicative signal correction, on-the-fly processing, BIG measurement DATA

OPEN ACCESS

Edited by:
Paolo Oliveri,
University of Genoa, Italy

Reviewed by:
Qifeng Li,
Tianjin University, China
Cristina Malegori,
University of Genoa, Italy

****Correspondence:***
Raffaele Vitale
raffaele.vitale@univ-lille.fr

Specialty section:
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

Received: *20 November 2021*
Accepted: *16 February 2022*
Published: *15 March 2022*

Citation:
Vitale R, Ruckebusch C, Burud I and Martens H (2022) Hyperspectral Video Analysis by Motion and Intensity Preprocessing and Subspace Autoencoding. Front. Chem. 10:818974. doi: 10.3389/fchem.2022.818974

1 INTRODUCTION

1.1 Hyperspectral Videos

In the last decade, hyperspectral imaging has experienced a significant diffusion mainly because of its capability of providing spatial and physico-chemical information about the systems under study - Hugelier et al. (2020). By returning whole spectra for all scanned pixels, in fact, a hyperspectral image permits to map the distribution of the constituents of the investigated samples all over the inspected field of view. For this reason, the applications of this analytical approach have lately dramatically increased in many domains of interest, like medicine, forensics, geoscience, urban and environmental surveillance and fire detection—Fischer and Kakoulli (2006); Chuvieco and Kasischke (2007); Hay et al. (2011); Matikainen and Karila (2011); Elmasry et al. (2012); Lu and Fei (2014); Silva et al. (2017); Khan et al. (2018); Vitale et al. (2020a).

Nonetheless, although hyperspectral imaging devices have become rather common tools in both academic and industrial chemistry laboratories, they are rarely configured so as to collect series of hyperspectral images over time for dynamic scene monitoring. There are two reasons behind this tendency: first of all, finding a reasonable compromise between spatial and spectral resolution and recording rate is not an easy and straightforward task; second, these so-called hyperspectral *videos* often translate into BIG DATA sets that can hardly be coped with by methodologies commonly resorted to for the analysis of individual hyperspectral images—for instance, Principal Component Analysis (PCA), Pearson (1901); Hotelling (1933), Partial Least Squares regression (PLS), Wold et al. (1983); Martens and Næs (1989), Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), Tauler et al. (1995), and Non-Negative Matrix Factorisation (NNMF), Lawton and Sylvestre (1971); Martens (1979). As an example, one can consider that, when storing hundreds of hyperspectral data arrays, the computer memory load is likely to increase up to the order of magnitude of the TBs. Modern workstations cannot readily handle such massive amounts of information and, therefore, standard chemometric techniques do need to be somehow adapted or extended to be possibly utilised in similar scenarios. Furthermore, hyperspectral video data typically account for various phenomena related to sample physics (*e.g.*, light scattering) and sample chemistry (light absorbance) and can be significantly affected by many different types of systematic errors (like those associated to nuisance fluctuations of the light source and/or the camera). Thus, identifying, disentangling, modelling and interpreting all these distinct sources of variations remains undoubtedly a challenging task.

1.2 Hyperspectral Video Analysis

Many known causal phenomena influence how light interacts with matter. The most important ones—light absorbance and light scattering—can even be approximated by relatively simple models, *e.g.*, the Beer-Lambert's law or the Kubelka-Munck theory—Bouguer, (1729); Lambert, (1760); Beer (1852); Kubelka and Munk (1931). Yet, albeit some of the aforementioned error factors affecting spectroscopic measurements—like illumination changes—can also be easily foreseen, a detailed mathematical characterisation of the spectral effects they might generate would be prohibitive from a computational point of view. For this reason, modelling and analysing hyperspectral videos constitutes a problematic challenge. Hyperspectral video data, in fact, yield information about the four main ontological aspects of reality: space, time, properties/attributes (for instance, a light intensity profile in the near-infrared—NIR—spectral range) and their interactions. Thus, a comprehensive description of a hyperspectral video would require the identification and the quantification of *factors* or *components* (both known and unknown) accounting for spatial, temporal and spectral variation patterns in such data. This would allow practitioners and users to gain new insights into complex systems of high relevance and into the interplay between the known and unknown phenomena driving their behaviour and evolution. As an immediate example of such an interplay, consider wood drying, a process exhibiting a deep economic and technical impact—McMillen (1964): water absorption properties allow, in principle, the moisture content of wood samples to be accurately determined. However, these properties may substantially change along with the thermodynamic state of water molecules (*i.e.*, free or bound) and might even mimic spectral contributions from other wood constituents, like cellulose, hemicellulose and lignin. In this as well as in numerous other real-world scenarios, disentangling and characterising the two aforementioned types of phenomena becomes, therefore, crucial from the perspective of understanding. In this article, a novel hybrid approach to achieve this objective is presented. It combines three multivariate approximation strategies for the compression and rational handling of hyperspectral videos: IDLE modelling—Westad and Martens (1999); Martens (2015) — Extended Multiplicative Signal Correction (EMSC)—Martens and Stark (1991); Martens et al. (2003)—and the On-The-Fly Processing (OTFP)—Vitale et al. (2017b). If, on the one hand, EMSC is a well-established tool in the chemometric community, on the other hand IDLE modelling and the OTFP have only recently been conceived, although they have already demonstrated their potential for fast processing of BIG DATA streams—Martens (2015); Vitale et al. (2017b); Stefansson et al. (2020); Vitale et al. (2020b). For the joint analysis of spatial and intensity changes in video recordings, IDLE splits the data variation into two domains as expressed in the following mathematical relation:

$$I = D(L) + E \tag{1}$$

by which a generic measured image (I) can be described as a function of the displacement (D) of a local intensity image (L) plus error (E). Imagine, for instance, that two different objects (*i.e.*, a black triangle and a black square—see **Figure 1A**) were photographed on a white table. After 1 minute, someone moves the first object along *y*, rotates it 90° and collects another picture (see **Figure 1B**). After 2 minutes, a third picture is taken after the square was painted grey (see **Figure 1C**). Assume now that D and L explain simultaneously vertical displacements and clockwise

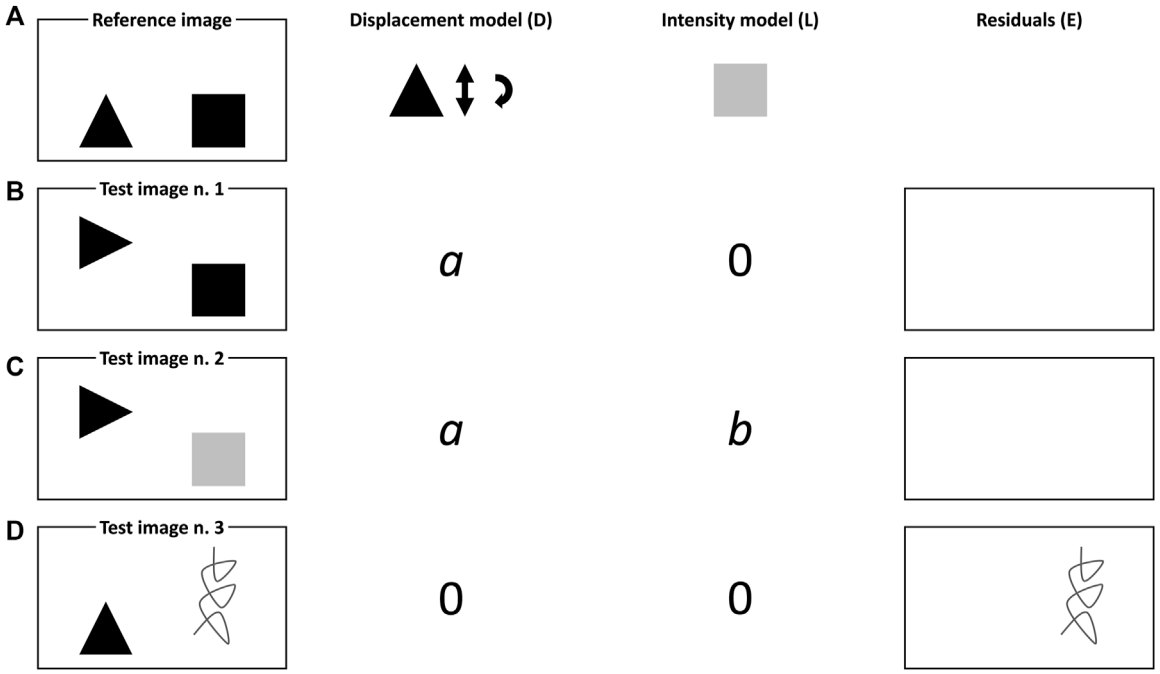

FIGURE 1 | Schematic representation of the basic principles of IDLE modelling. **(A)** Reference image, **(B)** Test image n.1, **(C)** Test image n.2 and **(D)** Test image n.3.

rotations of the black triangle and variations in the black square pixel intensities, respectively. In this simple illustration, D would encode the triangle movement as an individual coefficient (say, *a*), proportional to the dissimilarity from the object's original location and positioning and exhibiting a sign that depends on the sense of its motion. Analogously, L would quantify the change in the intensity of the pixels of the square as a positive parameter (say, *b*), since the transition from black to light grey implies an increase of the image lightness. Unexpected motions and colours as well as the appearance of unexpected items, like the dark grey spiral-like structure in **Figure 1D**, would be accounted for by the residuals E.

Conversely, the OTFP gradually constructs reduced-rank bilinear models that summarise virtually *ever-lasting* streams of multivariate responses and capture the evolving covariation patterns among their (spectral) variables in space and time. In other words, it represents an extension of classical PCA designed for processing such multivariate responses as soon as they are collected and, most importantly, without requiring entire raw datasets to be kept in memory. More specifically, the OTFP rests on a flexible bilinear subspace model structure which is automatically expanded when a new variation pattern is discovered—as for classical moving-window PCA implementations, Makeig et al. (2000); Wang et al. (2005), even if, here, relevance for *old* or *past* observations is never lost—or refined when the same variation patterns are repeatedly observed, while statistical redundancies are filtered out guaranteeing high rates of information compression. In contrast to black-box *deep learning* solutions, this PCA-like model-based approximation is graphically interpretable in its compressed state and allows at any time the original input to be reconstructed with a better signal-to-noise ratio (as measurement noise is eliminated).

Here, the sequential utilisation of IDLE, EMSC and the OTFP for the investigation of hyperspectral videos will be tested in a context similar to that envisioned before: the monitoring of the drying process of a wood specimen. The results of this study will highlight how this combination can enable an accurate estimation of the dynamic evolution of wood properties and how relatively simple quantitative spatial and temporal information can be extracted from a seemingly overwhelming stream of hyperspectral video data by coupling different mathematical modelling techniques.

1.3 Hyperspectral Video Data Structure

Hyperspectral videos can be regarded as time series of three-dimensional data arrays (hyperspectral *frames* or *snapshots*) with dimensions $N_x \times N_y \times J$, where N_x and N_y denote the number of pixels scanned along the horizontal and vertical direction, respectively, and J the number of wavelength channels sampled by the equipment employed. More broadly speaking, they can be thought of as the product of the concatenation of these arrays along a fourth time-related measurement mode. In spite of their multidimensional structure, hyperspectral video data are usually analysed in their unfolded form, *i.e.* as matrices of size $N_xN_yK \times J$ with K representing the amount of time points at which the aforementioned frames are collected. Each row of such matrices carries a single spectral profile recorded for an individual pixel at a given time point.

2 METHODS

In this article, EMSC and the OTFP are applied in a sequential fashion to assess/discover and quantify known and unknown

sources of data variability in hyperspectral videos. This strategy combines mechanistic and empirical multivariate modelling for describing all physical, chemical and instrumental variation patterns behind hyperspectral video recordings. In order to account for and compensate possible motions and pixel intensity changes which could originate complex non-linearities distorting the measured spatiospectral response, optical flow analysis—Horn and Schunck (1981, 1993)—and IDLE are applied in a preliminary preprocessing step.

The next sections will describe in detail the basics of the three different methodologies exploited here.

2.1 IDLE Modelling

Broadly speaking, the IDLE model is a mathematical description of real-world objects or scenes (characterised by spatiotemporal measurements like videos) in terms of their intensity and spatial variations. Here, IDLE is utilised as an empirical compression approach for sets of consecutive video frames, yielding high compression rates and, at the same time, enabling qualitative and quantitative data interpretation. IDLE is based on a three-step methodological procedure:

1. first of all, it segments out each of the relevant, independent objects (so-called *holons*) within a particular scene;
2. then, for each holon it estimates both D (accounting for motions and shape changes) and L, relative to a fixed, user-defined reference frame;
3. finally, it *morphs back* the holons in the investigated image to their spatial shape and location in the reference image. This facilitates a compact subspace modelling of both displacements and intensity changes.

2.1.1 Motion Estimation and Motion Compensation

IDLE modelling concerns how to reduce the complexities that arise when modelling objects that both move and change intensity (or spectral profile) at the same time. Imagine, for instance, a video composed of K grey-scale images ($\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_k, \ldots, \mathbf{I}_K$) of size $N_x \times N_y$ depicting certain objects whose shape and brightness varies over time. Let $\mathbf{I}_{\text{ref}}$ be one of these images, chosen to define a common reference for all the other ones. Analogously, let $\underline{\mathbf{O}}_{\text{ref}}$ ($N_x \times N_y \times 2$) define the horizontal and vertical pixel coordinates (or pixel *addresses*) at which these objects are visible in $\mathbf{I}_{\text{ref}}$. The reference intensity image at pixel adresses $\underline{\mathbf{O}}_{\text{ref}}$ is then $\mathbf{I}_{\text{ref},\underline{\mathbf{O}}_{\text{ref}}}$. At this point, the objects in the scene setting captured by each video frame, $\mathbf{I}_k$, can be described with respect to how they look in $\mathbf{I}_{\text{ref}}$. Neglecting motions, at time k, the local intensity-corrected version of $\mathbf{I}_{\text{ref},\underline{\mathbf{O}}_{\text{ref}}}$ can be expressed as:

$$\mathbf{L}_{k,\underline{\mathbf{O}}_{\text{ref}}} = \mathbf{I}_{\text{ref},\underline{\mathbf{O}}_{\text{ref}}} + \Delta\mathbf{I}_{k,\underline{\mathbf{O}}_{\text{ref}}} \tag{2}$$

with $\Delta\mathbf{I}_{k,\underline{\mathbf{O}}_{\text{ref}}}$ ($N_x \times N_y$) carrying the image intensity deviations from $\mathbf{I}_{\text{ref},\underline{\mathbf{O}}_{\text{ref}}}$.

Likewise, the pixel adresses where the objects from $\mathbf{I}_{\text{ref}}$ are observable in $\mathbf{I}_k$ become:

$$\underline{\mathbf{O}}_k = \underline{\mathbf{O}}_{\text{ref}} + \Delta\underline{\mathbf{O}}_k \tag{3}$$

where $\Delta\underline{\mathbf{O}}_k$ ($N_x \times N_y \times 2$) contain the so-called horizontal and vertical motion fields indicating how every pixel in $\mathbf{I}_k$ should be displaced so that the objects in $\mathbf{I}_k$ mimic their shape and location in $\mathbf{I}_{\text{ref}}$. Hence, merging **Eqs (2)**, **(3)**, the IDLE model for the kth frame can be written compactly as a function of how its intensity has changed ($\Delta\mathbf{I}_{k,\underline{\mathbf{O}}_{\text{ref}}}$) and how it has moved ($\Delta\underline{\mathbf{O}}_k$) compared to the reference one:

$$\mathbf{I}_{k,\underline{\mathbf{O}}_k} = \mathbf{I}_{k,\underline{\mathbf{O}}_{\text{ref}}+\Delta\underline{\mathbf{O}}_k} = \mathbf{I}_{\text{ref},\underline{\mathbf{O}}_{\text{ref}}} + \Delta\mathbf{I}_{k,\underline{\mathbf{O}}_{\text{ref}}} = \mathbf{L}_{k,\underline{\mathbf{O}}_{\text{ref}}} \tag{4}$$

According to this notation, the terms I, D and L in **Eq. (1)** would correspond to $\mathbf{I}_{k,\underline{\mathbf{O}}_k}$, $\Delta\underline{\mathbf{O}}_k$ and $\mathbf{L}_{k,\underline{\mathbf{O}}_{\text{ref}}}$, respectively.

From a practical perspective, $\Delta\underline{\mathbf{O}}_k$ can be obtained by motion estimation—Horn and Schunck (1981, 1993)—comparing $\mathbf{I}_k$ and $\mathbf{I}_{\text{ref}}$. This allows one to morph the objects from where they were located in $\mathbf{I}_k$ back to their pixel addresses in $\underline{\mathbf{O}}_{\text{ref}}$ and to their intensity at time k relative to $\mathbf{I}_{\text{ref}}$ ($\mathbf{L}_{k,\underline{\mathbf{O}}_{\text{ref}}}$). The intensity changes, $\Delta\mathbf{I}_{k,\underline{\mathbf{O}}_{\text{ref}}}$, as well as the motion fields $\Delta\underline{\mathbf{O}}_k$ are all given in the coordinate system of $\mathbf{I}_{\text{ref}}$, *i.e.*, $\underline{\mathbf{O}}_{\text{ref}}$.

2.1.2 Dual-Domain Bilinear Modelling of a Hyperspectral Video

Even when a hyperspectral video is handled, all the wavelength channels must follow the same spatial displacement at each time k. For this purpose, the unfolded vertical and horizontal motion fields, $\Delta\mathbf{o}_k^{\text{T}}$ ($1 \times 2N_xN_y$), can be estimated from an optimised combination of such channels, gathered column-wise into the matrix $\Delta\mathbf{O}$ ($K \times 2N_xN_y$), modelled bilinearly as

$$\Delta\mathbf{O} = \mathbf{T}_{\Delta\mathbf{O}}\mathbf{P}_{\Delta\mathbf{O}}^{\text{T}} + \mathbf{E}_{\Delta\mathbf{O}}^{\text{T}} \tag{5}$$

and applied to each entire hyperspectral frame. Here, $\mathbf{T}_{\Delta\mathbf{O}}$ ($K \times A_{\text{IDLE}}$) contains the projection coordinates of $\Delta\mathbf{O}$ on the directions defined by the columns of $\mathbf{P}_{\Delta\mathbf{O}}$ ($2N_xN_y \times A_{\text{IDLE}}$) and $\mathbf{E}_{\Delta\mathbf{O}}$ ($K \times 2N_xN_y$) carries the corresponding residuals not explained at the chosen rank, $A_{\text{IDLE}} < 2N_xN_y$.

Compact, low-dimensional bilinear models often summarize quite well the motions in $\Delta\underline{\mathbf{O}}$ when they are defined in the same reference coordinate system. Also the unfolded intensity images $\Delta\mathbf{i}_{k,\underline{\mathbf{O}}_{\text{ref}}}^{\text{T}}$ ($1 \times N_xN_y$) may be well approximated in a similar fashion if expressed in a common coordinate system:

$$\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}} = \mathbf{T}_{\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}\mathbf{P}_{\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}^{\text{T}} + \mathbf{E}_{\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}} \tag{6}$$

with $\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}$ ($K \times N_xN_y$) being the 2D array resulting from the column-wise concatenation of each $\Delta\mathbf{i}_{k,\underline{\mathbf{O}}_{\text{ref}}}^{\text{T}}$ vector.

Rewriting **Eq. (4)** in vectorial form, the aforementioned morphing operation can be therefore expressed for the kth video frame as:

$$\begin{aligned}\mathbf{i}_{k,\underline{\mathbf{O}}_k}^{\text{T}} &= \mathbf{i}_{k,\underline{\mathbf{O}}_{\text{ref}}+\Delta\underline{\mathbf{O}}_k}^{\text{T}} = \mathbf{i}_{k,\mathbf{o}_{\text{ref}}^{\text{T}}+\Delta\mathbf{o}_k^{\text{T}}}^{\text{T}} = \mathbf{i}_{k,\mathbf{o}_{\text{ref}}^{\text{T}}+\left(\mathbf{t}_{k,\Delta\mathbf{O}}^{\text{T}}\mathbf{P}_{\Delta\mathbf{O}}^{\text{T}}+\mathbf{e}_{k,\Delta\mathbf{O}}^{\text{T}}\right)}^{\text{T}} \\ &= \mathbf{i}_{\text{ref},\underline{\mathbf{O}}_{\text{ref}}}^{\text{T}} + \mathbf{t}_{k,\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}^{\text{T}}\mathbf{P}_{\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}^{\text{T}} + \mathbf{e}_{k,\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}^{\text{T}}\end{aligned} \tag{7}$$

where $\mathbf{t}_{k,\Delta\mathbf{O}}^{\text{T}}$, $\mathbf{e}_{k,\Delta\mathbf{O}}^{\text{T}}$, $\mathbf{t}_{k,\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}^{\text{T}}$ and $\mathbf{e}_{k,\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}^{\text{T}}$ denote the kth row vectors of $\mathbf{T}_{\Delta\mathbf{O}}$, $\mathbf{E}_{\Delta\mathbf{O}}^{\text{T}}$, $\mathbf{T}_{\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}$ and $\mathbf{E}_{\Delta\mathbf{I}_{\underline{\mathbf{O}}_{\text{ref}}}}$, respectively. Refolding is finally required for the sake of representation.

2.2 Extended Multiplicative Signal Correction

EMSC is a bilinear modelling approach that permits to separate, quantify and correct for distinct types of known chemical and physical data variation sources in the acquired signal profiles. As applied in this article, EMSC assumes that a generic spectrum, $\mathbf{x}$ (of dimensions $J \times 1$) can be mathematically described as:

$$\mathbf{x} = b\left(\mathbf{r} + \sum_i \Delta c_i \mathbf{s}_i\right) + a\mathbf{1} + d\mathbf{f} + g\mathbf{f}^2 + \mathbf{e} \tag{8}$$

where b is the effective relative pathlength; $\mathbf{r}$ ($J \times 1$) is a predetermined reference spectrum; Δc_i and $\mathbf{s}_i$ ($J \times 1$) denote the presumed concentration/abundance contribution and the spectral fingerprint of the ith main constituent of the system under study, respectively; $\mathbf{1}$ ($J \times 1$) is a column vector of ones; $\mathbf{f}$ ($J \times 1$) contains values monotonically increasing from −1 to 1; a, d and g constitute a set of coefficients; and $\mathbf{e}$ ($J \times 1$) carries the unmodelled residuals (*i.e.*, unmodelled chemical and/or physical variations as well as random measurement noise) resulting from this approximation.

Altogether, $\mathbf{1}$, $\mathbf{f}$ and $\mathbf{f}^2$ connote polynomial model dimensions accounting for smoothly wavelength-dependent phenomena (baseline level, slope and curvature, respectively).

Given $h_i = b\Delta c_i\ (\forall i)$, the unknowns in **Eq. (8)** can be retrieved by Ordinary or Weighted Least Squares (OLS/WLS) as:

$$\left[b\ h_1\ \ldots\ h_I\ a\ d\ g\right] = \mathbf{x}^{\mathrm{T}}\mathbf{W}_{\mathrm{EMSC}}\mathbf{W}_{\mathrm{EMSC}}\mathbf{M}\left(\mathbf{M}^{\mathrm{T}}\mathbf{W}_{\mathrm{EMSC}}\mathbf{W}_{\mathrm{EMSC}}\mathbf{M}\right)^{-1} \tag{9}$$

where $\mathbf{M} = [\mathbf{r}\ \mathbf{s}_1\ \ldots\ \mathbf{s}_I\ \mathbf{1}\ \mathbf{f}\ \mathbf{f}^2]$ and $\mathbf{W}_{\mathrm{EMSC}}$ ($J \times J$) is a diagonal matrix of weights associated to the different sampled spectral channels[1].

Since the constituent profiles, $\mathbf{s}_i$, are a required input for EMSC processing, this methodology has been chosen for describing expected variation patterns evolving all over the duration of a hyperspectral video.

Once the EMSC coefficients have been calculated as in **Eq. (9)**, they can be exploited for pretreating the input spectrum, $\mathbf{x}$, in order to filter varying light scattering effects as:

$$\mathbf{x}^{\mathrm{p}} = \frac{\left(\mathbf{x} - a\mathbf{1} - d\mathbf{f} - g\mathbf{f}^2\right)}{b} \tag{10}$$

with $^{\mathrm{p}}$ standing for *preprocessed*. In the present application of EMSC, the estimated chemical variations will also be subtracted from $\mathbf{x}$ as:

$$\begin{aligned}\mathbf{x}^{\mathrm{p}} &= \frac{\left(\mathbf{x} - a\mathbf{1} - d\mathbf{f} - g\mathbf{f}^2 - \sum_i h_i\mathbf{s}_i\right)}{b} \\ &= \frac{\left(\mathbf{x} - a\mathbf{1} - d\mathbf{f} - g\mathbf{f}^2\right)}{b} - \sum_i \Delta c_i\mathbf{s}_i\end{aligned} \tag{11}$$

Finally, if EMSC residuals are deemed to be affected by the effective optical pathlength of the sample, they can be computed as:

$$\begin{aligned}\mathbf{e} &= \mathbf{x} - b\mathbf{r} - \sum_i h_i\mathbf{s}_i - a\mathbf{1} - d\mathbf{f} - g\mathbf{f}^2 \\ &= \mathbf{x} - b\left(\mathbf{r} - \sum_i \Delta c_i\mathbf{s}_i\right) - a\mathbf{1} - d\mathbf{f} - g\mathbf{f}^2\end{aligned} \tag{12}$$

Pathlength-corrected residuals are subsequently estimated as:

$$\tilde{\mathbf{e}} = b^{-1}\mathbf{e} \tag{13}$$

2.3 The On-The-Fly Processing

After the IDLE-based motion estimation-compensation and the quantification-correction of known physical and chemical variations by EMSC preprocessing, the resulting unmodelled residuals are analysed by the OTFP in the attempt of looking for unknown, yet systematic variability patterns in data. The OTFP relies on a self-learning adaptive modelling principle which allows massive amounts of measurement recordings collected over time to be compressed with a minimal loss of meaningful information according to a PCA-like bilinear decomposition. Its global computational procedure encompasses five different steps:

1. the raw data stream, $\mathbf{X}$ (of dimensions, *e.g.*, $N_xN_yK \times J$), divided into a sequence of blocks, say $\mathbf{X}_g$ ($N_g \times J$, $g = 1, 2, \ldots, G$), is submitted to an optional lossless knowledge-based preprocessing stage including a linearisation—which can be conducted by means of approaches like Standard Normal Variate (SNV), Barnes et al. (1989), Multiplicative Scatter Correction (MSC), Martens et al. (1983), Fast Fourier Transform (FFT), Cooley and Tukey (1965), and wavelet decomposition, Walczak (2000)—and a signal-conditioning step;
2. the preprocessed data are projected onto a bilinear subspace already established at the previous point in time as:

$$\mathbf{X}_g^{\mathrm{p}} = \mathbf{T}_g^{\mathrm{p}}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_g^{\mathrm{p}} \tag{14}$$

with $\mathbf{T}_g^{\mathrm{p}}$ ($N_g \times A_{\mathrm{OTFP}}$) defining the projection coordinates or scores of all the N_g observations on the basis vectors or components defined by the columns of $\mathbf{P}$ ($J \times A_{\mathrm{OTFP}}$) and $\mathbf{E}_g^{\mathrm{p}}$ ($N_g \times J$) carrying unmodelled residuals, *i.e.*, the fraction of $\mathbf{X}_g^{\mathrm{p}}$ not explained by the model at the chosen rank, $A_{\mathrm{OTFP}} < J$;

3. the projection residuals are thereafter input to a second bilinear modelling stage aimed at detecting new components and isolating outliers. New components are encoded as additional subspace dimensions, whose final number is usually selected based on the total amount of the original data variance that is to be explained, although alternative criteria may also be exploited—Endrizzi et al. (2014); Vitale et al. (2017a); Vitale and Saccenti (2018). In other words, the OTFP algorithm learns to identify and quantify all the systematic types of covariation in the data as they stream, while filtering out random measurement

[1]If $\mathrm{diag}(\mathbf{W}_{\mathrm{EMSC}}) = \mathbf{1}$, the parameter estimation is carried out by OLS.

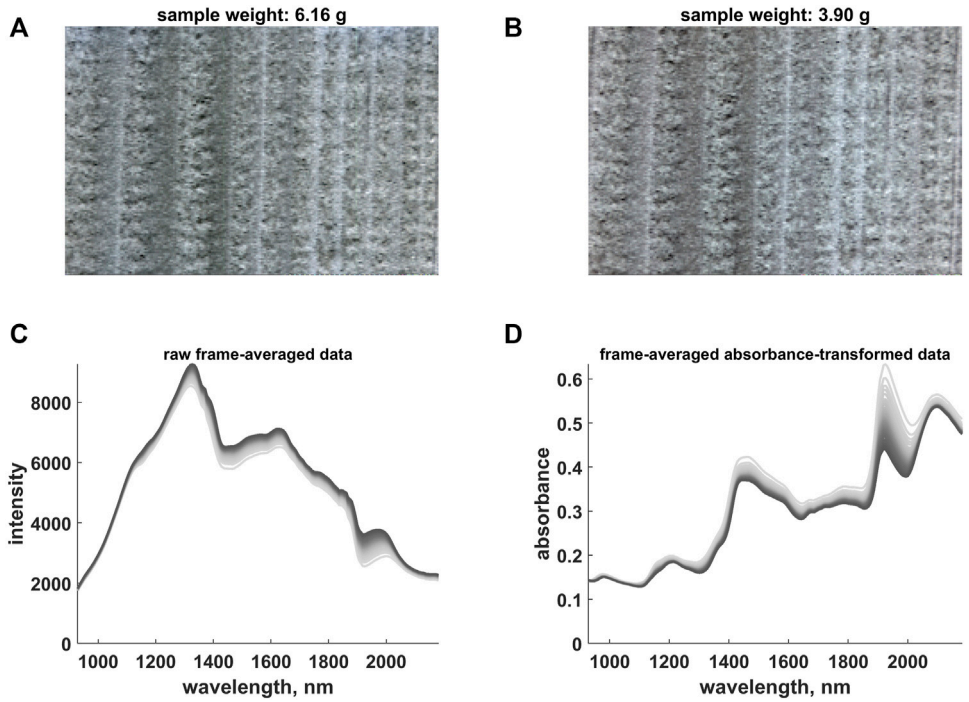

FIGURE 2 | (A) False Red Green Blue (RGB) representation of the first hyperspectral video frame. **(B)** False RGB representation of the last hyperspectral video frame. **(C)** Raw frame-averaged intensity data. **(D)** Frame-averaged absorbance-transformed data. The colour gradient (from light to dark grey) follows the time evolution of the hyperspectral video. Notice that the absorbance values measured at 980, 1,138 and 1,302 nm were used to generate **(A)** and **(B)**.

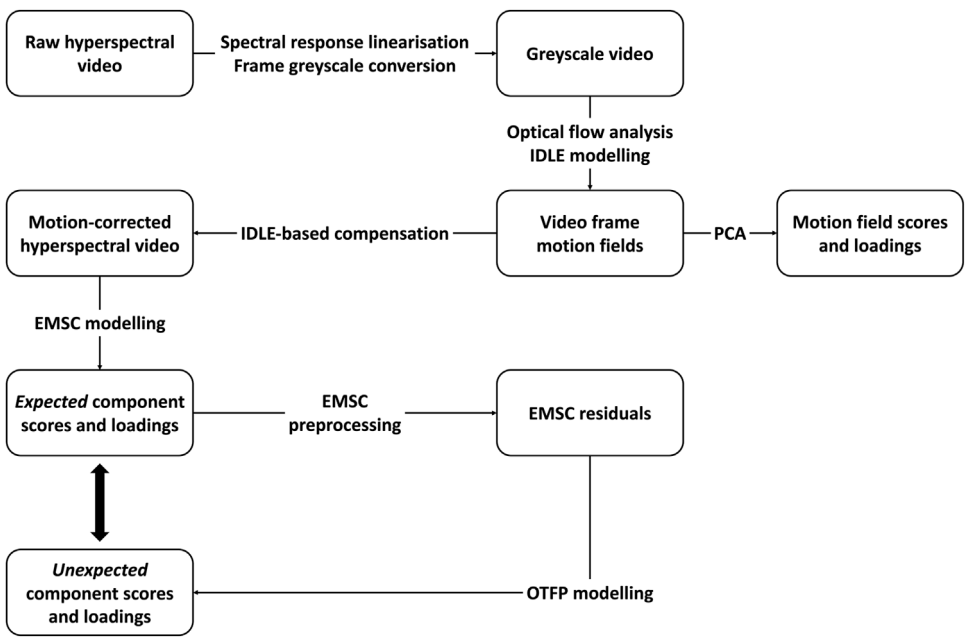

FIGURE 3 | Schematic flowchart of the hyperspectral video processing and analysis framework proposed in this article.

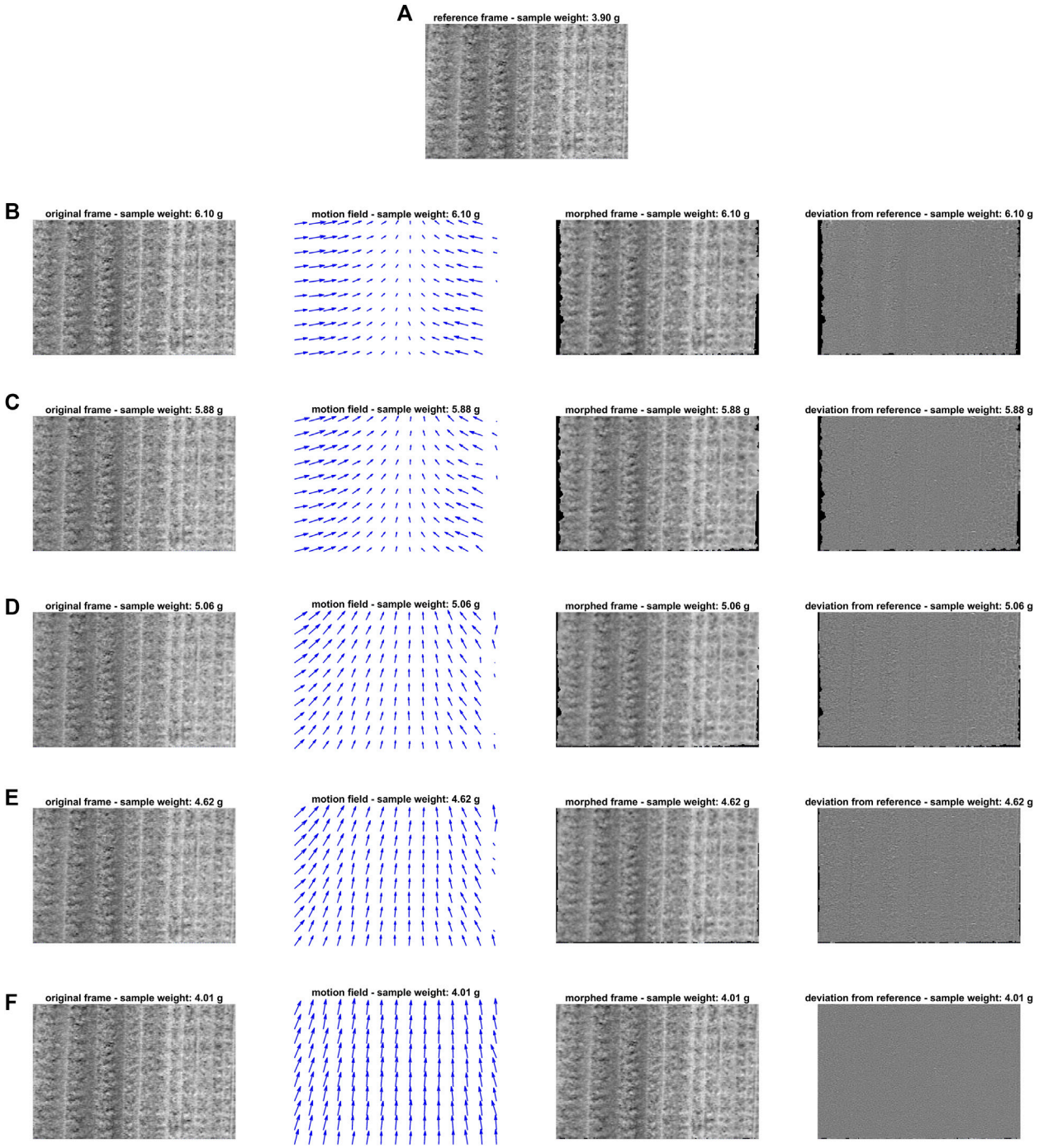

FIGURE 4 | IDLE modelling: **(A)** displays the reference video frame; **(B-F)** contain (from left to right) the representation of five different snapshots collected over the entire duration of the monitoring experiment (n. 2—sample weight: 6.10 g; n. 6—sample weight: 5.88 g; n. 21—sample weight: 5.06 g; n. 29—sample weight: 4.62 g; n. 40—sample weight: 4.01 g), of the motion fields yielded by their optical flow analysis highlighting how individual pixels shifted compared to the reference image, of the motion-compensated frames morphed in order to mimic the target one and of the intensity deviations between the motion-compensated and the reference snapshots. Notice that IDLE was applied to grey-scale images, obtained by averaging the preprocessed absorbance values (see **Section 4.1**) at 1,024, 1,195 and 1,309 nm, respectively.

errors and irrelevant outliers (if they do not contribute to the definition of a new pattern of variation);

4. at regular intervals, the OTFP model is refined and updated;
5. pretreatment as well as model parameters (*i.e.*, OTFP scores and loadings) are stored as output. At any time, they can be either used to reconstruct the original data, *e.g.*, for

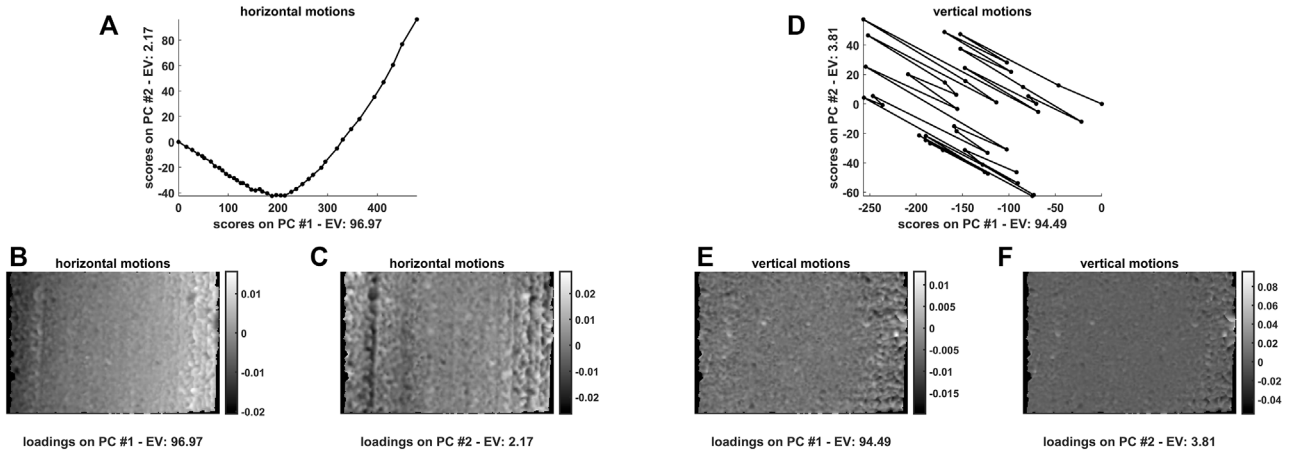

FIGURE 5 | First and second principal component scores and loadings resulting from a bilinear decomposition of the **(A-C)** horizontal and **(D-F)** vertical motions quantified for the 42 hyperspectral video frames at hand. The black solid line follows the temporal evolution of the experiment. The reference snapshot (n. 42—the last one of the sequence) is easily recognisable as it exhibits scores coordinates equal to [0, 0] (*i.e.*, no motions were estimated for it). The black areas around the loadings images contain pixels excluded from the computational procedure as they underwent an excessively large displacement with respect to frame n. 42. *EV* stands for *Explained Variance*.

visualisation, or exploited in their compressed form for efficient storage and transmission, human graphical interpretation and quantification.

A survey of the operational principles of the OTFP is provided in Vitale et al. (2017b).

3 DATASET

As model system, a piece of wood of the species Norway Spruce (*Pincea abies*) was submerged in water and soaked for approximately 24 h. Thereafter, it was placed on a digital scale for tracking in real time the variation of its weight and its drying process was monitored by means of a hyperspectral line scan camera (Specim, Oulu, Finland) automatically capturing reflectance images between 930 and 2,200 nm. More specifically, the sample was scanned at regular time intervals, *i.e.*, each time a decrease of around 0.05 g was observed (initial weight: 6.16 g—see **Figure 2A**; final weight: 3.90 g—see **Figure 2B**; total number of hyperspectral images: 42). The sample was illuminated by two halogen lamps positioned on the two sides of the hyperspectral device and never moved during the whole duration of the experiment. A region of interest of 150 × 225 pixels was segmented within each frame, which finally resulted in the generation of a four-dimensional dataset of size 150 × 225 × 42 × 200 (see also **Section 1.3**) and in a memory load of roughly 2.3 GB (double-precision floating-point format).

Although these data were already investigated before—Vitale et al. (2020b)—here, the key role of the linearisation of the instrumental response across space provided by the IDLE approach and its fundamental impact on the assessment and interpretation of the temporal variations of the water signal contributions will be explored.

4 RESULTS AND DISCUSSION

A flowchart schematising the general hyperspectral video analysis framework proposed in this work is provided in **Figure 3**.

4.1 Spectral Response Linearisation and Frame Greyscale Conversion

In order to compensate the wavelength-dependent variations associated to the light source, the intensity values registered at each jth wavelength and at each $n_x \times n_y$-th pixel of the kth video frame, $I_{n_x,n_y,k,j}$, were first converted into reflectance units as in :

$$R_{n_x,n_y,k,j} = \frac{\left(I_{n_x,n_y,k,j} - I_{n_x,n_y,j,\mathrm{d}}\right)}{\left(I_{n_x,n_y,j,\mathrm{w}} - I_{n_x,n_y,j,\mathrm{d}}\right)} \tag{15}$$

with $I_{n_x,n_y,j,\mathrm{d}}$ and $I_{n_x,n_y,j,\mathrm{w}}$ the intensity recorded at the jth wavelength and at $n_x \times n_y$-th pixel for a dark reference and a white reference (a Spectralon sample), respectively. Thereafter, they were transformed into apparent absorbance (that is to say, linearised with respect to the chemical response) according to the following relation:

$$A_{n_x,n_y,k,j} = \log\left(\frac{1}{R_{n_x,n_y,k,j}}\right) = x_{n_x,n_y,k,j} \tag{16}$$

An example of raw and absorbance-converted spectral profiles is provided in **Figure 2C,D**, which highlight the presence of strong baseline variations probably caused by fluctuations in the illumination conditions or in the angular distribution of the reflected light. In order to minimise the bias that such fluctuations (unrelated to sample motions[2]) may induce in the IDLE-based quantification and compensation, an additional

[2]For example, those due to water diffusion.

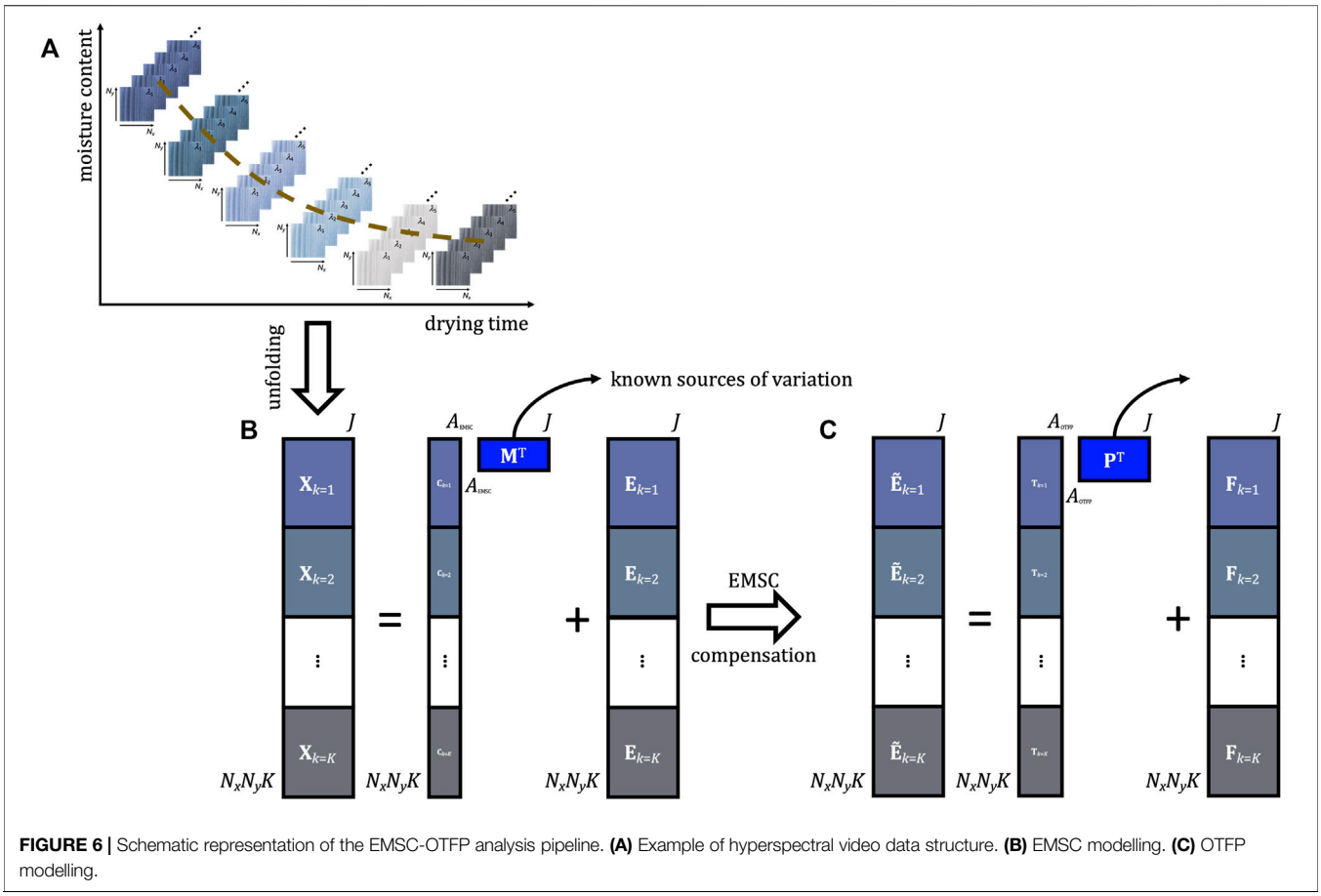

FIGURE 6 | Schematic representation of the EMSC-OTFP analysis pipeline. **(A)** Example of hyperspectral video data structure. **(B)** EMSC modelling. **(C)** OTFP modelling.

two-step pretreatment procedure was executed prior to the successive data processing stage:

1. the spectra associated to the pixels of each video frame were pretreated according to an EMSC model similar to the one in **Eq. 8** and encompassing the profiles of two known components: dry wood[3] (reference) and pure water[4]. $\mathbf{W}_{\mathrm{EMSC}}$ was set equal to the identity matrix. More specifically, the correction performed for the $n_x \times n_y$-th pixel of the kth video frame can be expressed as:

$$\mathbf{x}^{\mathrm{p}}_{n_x,n_y,k} = \frac{\left(\mathbf{x}_{n_x,n_y,k} - a_k\mathbf{1} - d_k\mathbf{f} - g_k\mathbf{f}^2\right)}{b_k} - h_{k,\mathrm{water}}\mathbf{s}_{\mathrm{water}} \qquad (17)$$

with a_k, d_k, g_k, b_k and $h_{k,\mathrm{water}}$ being estimated as in **Eq. (9)** from the kth frame mean spectrum;

2. at each time point, a grey-scale image, $\mathbf{I}_k$, was then obtained by averaging, for every pixel, the resulting absorbance values at 1,024, 1,195 and 1,309 nm (at these wavelengths, the frame-averaged spectra in **Figure 2D** exhibited the lowest standard deviation). In order to compensate dissimilarities among the intensity cumulative histograms of the various snapshots, these final estimates were ultimately level- and range-adjusted as:

$$\mathbf{I}^{\mathrm{p}}_k = \left(\mathbf{I}_k - \tilde{I}_k\right)\frac{RMS_{\mathrm{ref}}}{RMS_k} + \tilde{I}_{\mathrm{ref}} \qquad (18)$$

where $\tilde{I}_k$ and $\tilde{I}_{\mathrm{ref}}$ are the median intensity values within the kth and the reference frame (n. 42—sample weight: 3.90 g), respectively, while RMS_{ref} and RMS_k represent the root-median-squared deviation of the pixel intensities in $\mathbf{I}_{\mathrm{ref}}$ and $\mathbf{I}_k$ from their corresponding median values.

4.2 IDLE Modelling

The level- and range-corrected grey-scale images output by the algorithmic procedure outlined in **Section 4.1** were then subjected to IDLE modelling. **Figure 4** summarises the outcomes of the motion estimation-compensation step: **Figure 4A** displays the reference video frame, while, for the sake of illustration, **Figure 4B–F** contain (from left to right) the representation of five other snapshots collected over the entire duration of the monitoring experiment, of the motion fields

[3]Calculated as the average profile of the last video frame.

[4]Measured in reflectance mode by a Nicolet 6700 FT-NIR instrument (Thermo Scientific Inc., Madison, WI, United States) at the same nominal resolution and within the same spectral range as for the hyperspectral video data dealt with in this study and, then, converted into absorbance units.

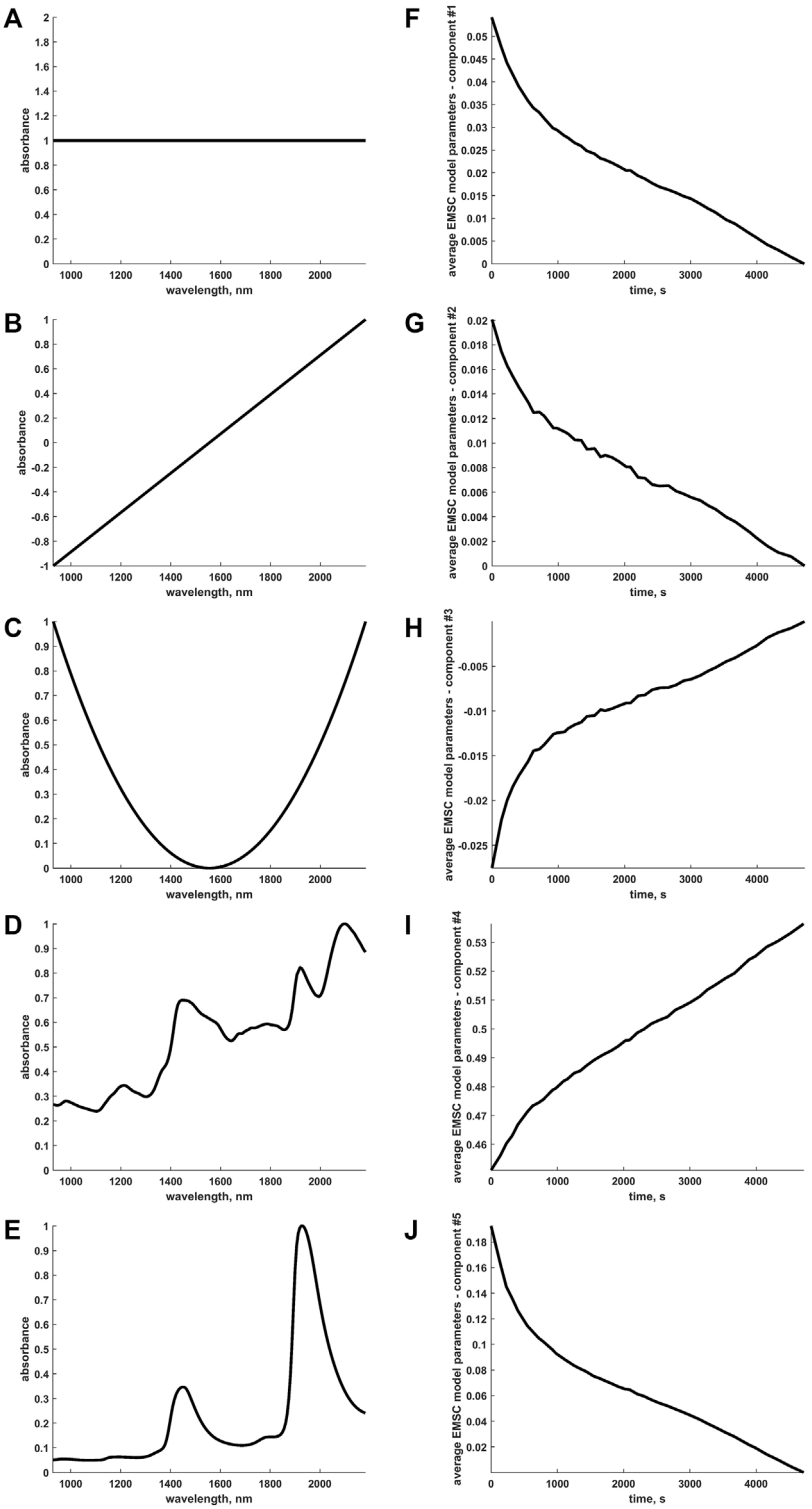

FIGURE 7 | (A-E) Characteristic (absolute max-normalised) spectral profiles submitted to the EMSC computational procedure. The first three represent a typical choice for EMSC modelling as they allow baseline offset, slope and quadratic curvature to be estimated for all the spectra or pixels of the hyperspectral video and compensated before the subsequent On-The-Fly Processing. The last two correspond to the spectra of dry wood and pure water, the two main constituents underlying the specific scene at hand. **(F-J)** Time evolution of the frame-averaged EMSC coefficients (rescaled to compensate the aforementioned absolute max-normalisation) associated to the sources of variation explained by the spectral profiles in **(A-E)**.

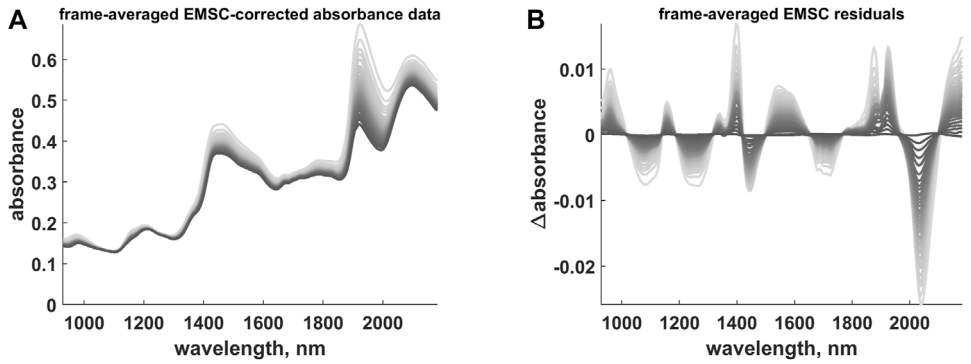

FIGURE 8 | (A) Frame-averaged EMSC-corrected absorbance data. **(B)** Frame-averaged EMSC residual profiles. The colour gradient (from light to dark grey) follows the time evolution of the hyperspectral video.

yielded by their optical flow analysis highlighting how individual pixels shifted compared to the reference image, of the motion-compensated frames morphed in order to mimic the target one and of the intensity deviations between the motion-compensated and the reference snapshots. As one can clearly see, except for minor edge artefacts, the aforementioned motion fields show how the wood sample horizontally squeezed as it dried and how such horizontal movements significantly decreased at the latest stage of the video recording (*i.e.*, when low amounts of water were present in the pores between wood fibres and compression finally slowed down or stopped). This is also corroborated by the gradual reduction of the number of pixels whose motions could not be properly estimated by IDLE (see the black areas surrounding the motion-compensated frames) because of their relatively large displacement with respect to snapshot n. 42[5]. Notice that these pixels did not undergo EMSC-OTFP processing. Moreover, minimal intensity-deviation-from-target values were observed after image compensation, confirming that wood spatial variations were successfully corrected for.

In order to get additional insights into the nature of such spatial variations along time, the quantified horizontal and vertical motions—retrieved from all the calculated motion fields and concatenated as detailed in Martens (2015)—were analysed by PCA. The resulting temporal scores and spatial loadings are graphed in **Figure 5**.

While vertical shifts appear to follow a random trend (see **Figure 5D**) and might be looked at as mainly due to sideways camera or measurement stage bumps (loadings values are also more or less homogeneously distributed all over the inspected field of view—see **Figures 5A,E,F**) a smoother and more structured evolution was found for the horizontal ones, which further substantiates what stated before about wood squeezing. Horizontal motion scores (see **Figure 5A**) seem to point out the occurrence of a two-phase process during which compression initially proceeds faster and finally decelerates. Horizontal motion loadings along the first principal component (see **Figure 5B**) emphasise the differences between the movements of the pixels of the left and the right side of the image, while those along the second principal component (see **Figure 5C**) permit to distinguish the distinct behaviour of lateral and central pixels.

4.3 EMSC Modelling

If on the one hand the IDLE approach is capable of quantifying and compensating the movements of a sample observed throughout a hyperspectral video (thus, enhancing the spatial linearity of the instrumental response), on the other hand the combined use of EMSC and the OTFP can enable the identification and retrieval of the most meaningful sources of information from the time series of resulting motion-free hyperspectral images.

The EMSC-OTFP analysis pipeline is schematically outlined in **Figure 6**.

Both EMSC and the OTFP are bilinear modelling techniques that can be utilised in an adaptive- or recursive-like way without requiring entire raw datasets to be kept in memory. The main difference between them regards their respective subspace definition. The matrix **M** (see **Eq. 9** and **Figure 6B**), in fact, is manually constructed by the user based on *a priori* knowledge about the system or the sample under study, which renders EMSC an ideal methodology for extracting and describing expected variation patterns evolving during the progression of a hyperspectral video. On the other hand, **P** (see **Eq. 14** and **Figure 6C**) is automatically learnt by the OTFP algorithm which gradually discovers (in real time) all the sources of systematic variation underlying the data at hand. Consequently, applying sequentially 1) EMSC to the (unfolded) motion-corrected data and 2) the OTFP to the resulting EMSC residuals yields two additive models accounting for both known and unknown phenomena driving the generation mechanism of hyperspectral videos and providing a detailed global overview of the captured dynamic scene.

Here, in a first step, the five profiles in **Figure 7A–E** were input to the EMSC algorithmic procedure: as also briefly outlined before, the first three constitute a standard choice for EMSC modelling as they permit to estimate and compensate baseline offset, slope and quadratic curvature for all the pixels of the hyperspectral video before the subsequent application of the OTFP. The last two profiles, instead, correspond to the spectra of dry wood (reference) and pure water, the two major constituents of the specific scene at hand. Representing the

[5]In fact, the higher the time difference between frames, the larger the distance that the pixels at the borders of these frames covered due to wood squeezing.

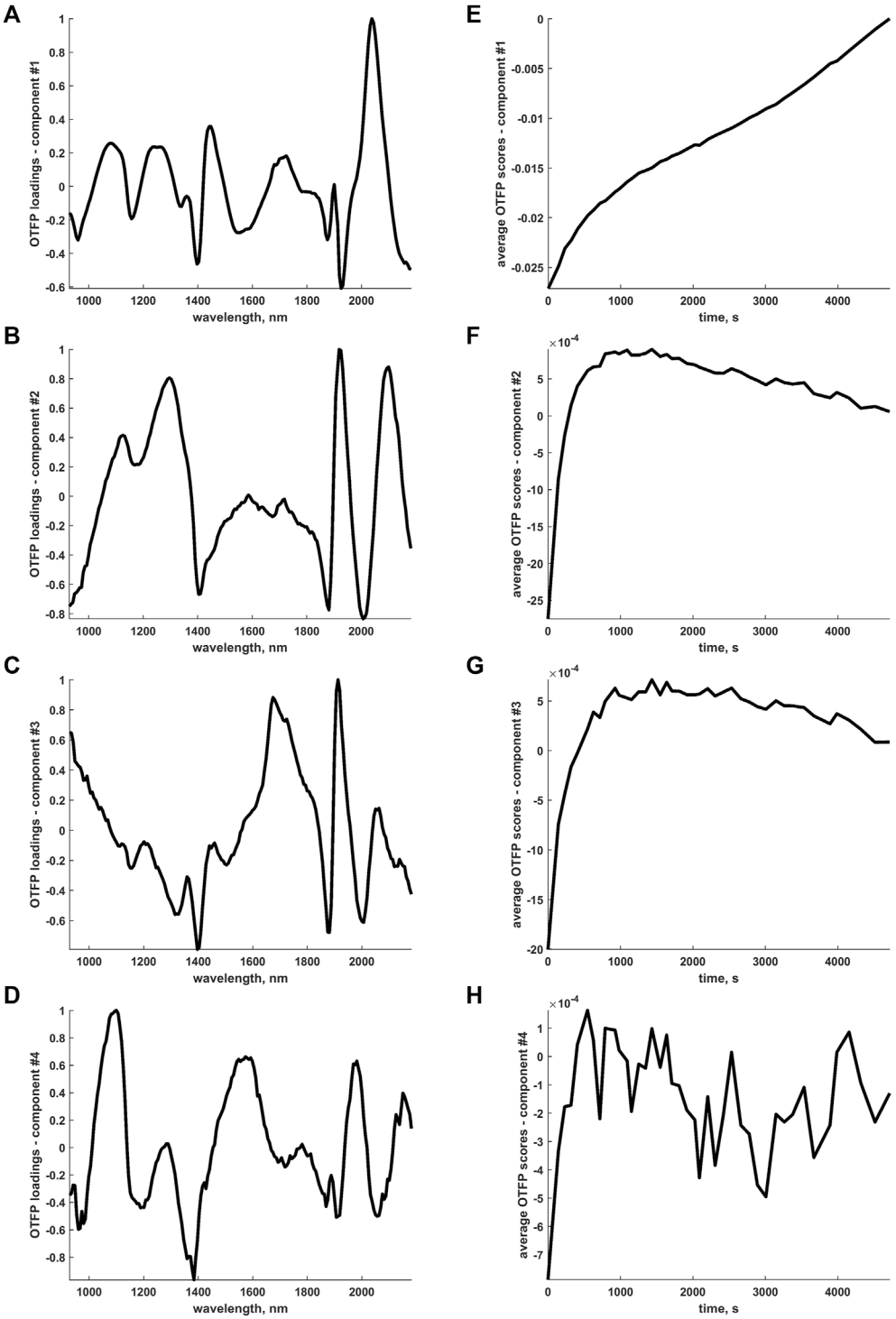

FIGURE 9 | (A-D) Pseudo-spectral (absolute max-normalised) loadings profiles retrieved by the OTFP computational procedure. **(E-H)** Time evolution of the frame-averaged OTFP scores (rescaled to compensate the aforementioned absolute max-normalisation) associated to the sources of variation explained by the loadings profiles in **(A-D)**.

time trend of the coefficients yielded for each one of these expected sources of data variability (averaged across all the pixels within every original video frame after motions were compensated, see **Figures 7F–J**) is a simple and immediate way to visualise and assess the information returned by EMSC and somehow characterise the dynamic evolution of known variability patterns during wood drying. From such graphs, one can easily observe that most of the modelled wood features change quite rapidly within the first stage of

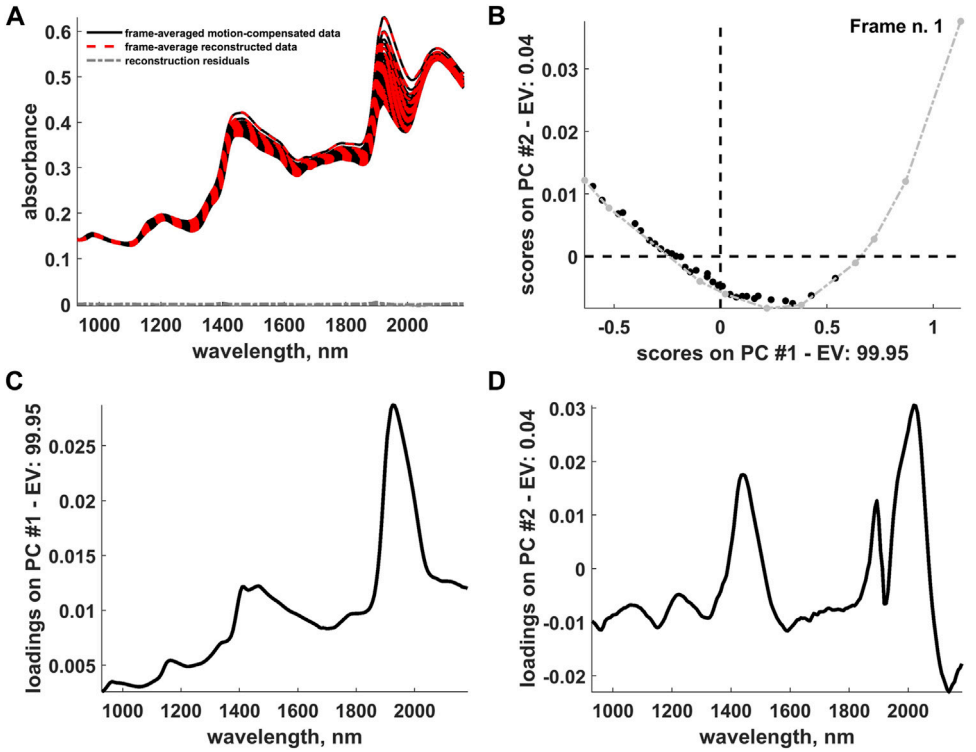

FIGURE 10 | (A) Representation of the frame-averaged motion-compensated data, frame-averaged data reconstructed after the IDLE, EMSC and OTFP analysis and reconstruction residuals. **(B)** Two-dimensional scores plot resulting from a PCA decomposition of the (pathlength-corrected) frame-averaged reconstructed data. Archetypal frames are highlighted in light grey and connected by a dashed-dotted grey line. The evolution of the scores from right to left follows the hyperspectral video progression from its beginning to its end. **(C)** First and **(D)** second component loadings yielded by the aforementioned PCA decomposition. *PC* and *EV* stand for *Principal Component* and *Explained Variance*, respectively.

the drying process. This may be due to the residual presence of a thin liquid water film on the surface of the wood sample at the beginning of the hyperspectral monitoring, which could have cloaked its spectral properties. **Figure 7J** also highlights that moisture loss was still proceeding when the experiment was interrupted. Conversely, regarding the wood contribution itself, an approximately constant increasing trend over time was observed. This behaviour accurately reflects the chemical nature of the sample drying which might have been clearly unveiled here because its continuous physical contractions were directly and explicitly accounted for, significantly reducing the spatial complexity of the considered video data. It goes without saying, then, that exploiting simultaneously both spectral and spatial information encoded in hyperspectral videos can significantly enhance the comprehension and understanding of the physico-chemical phenomena behind complex real-world systems.

4.4 OTFP Modelling

After the EMSC compensation (see **Figure 8A**), the resulting residual profiles (see **Figure 8B**) were submitted to the OTFP for automatically retrieving all the systematic sources of variation left unmodelled by the first data analysis steps[6].

[6]4 OTFP components were required to explain around 80% of the EMSC residual variance.

Even if the interpretation of the OTFP output may seem more complicated due to the fact that the OTFP subspace features PCA-like orthogonal bases, smooth and rather well-defined time trends were found for the frame-averaged OTFP coefficients or scores (see **Figure 9E–H**). Such time trends highlight the existence of at least two structured phases in the process of wood drying. Consider, for example, **Figure 9F**: an initial fast transition from negative to positive scores values can be observed followed by a smoother descendant evolution approximately plateauing at around 0. Given also that most of the OTFP loadings profiles in **Figure 9A–D** show large contributions associated to the main water absorption regions, one can reasonably envision the occurrence of more complex phenomena directly related to the thermodynamic state of water itself (*i.e.*, free or bound).

4.5 Data Reconstruction and Postprocessing

For a tentative exploration of the thermodynamic phenomena mentioned in **Section 4.4**, the pathlength-corrected absorbance spectra, obtained by reconstructing and averaging the 42 motion-compensated hyperspectral video frames after EMSC and OTFP processing (see **Figure 10A**), were decomposed by standard PCA and graphed in the scores plot in **Figure 10B**. This plot clearly highlights the occurrence of a two-phase transition process

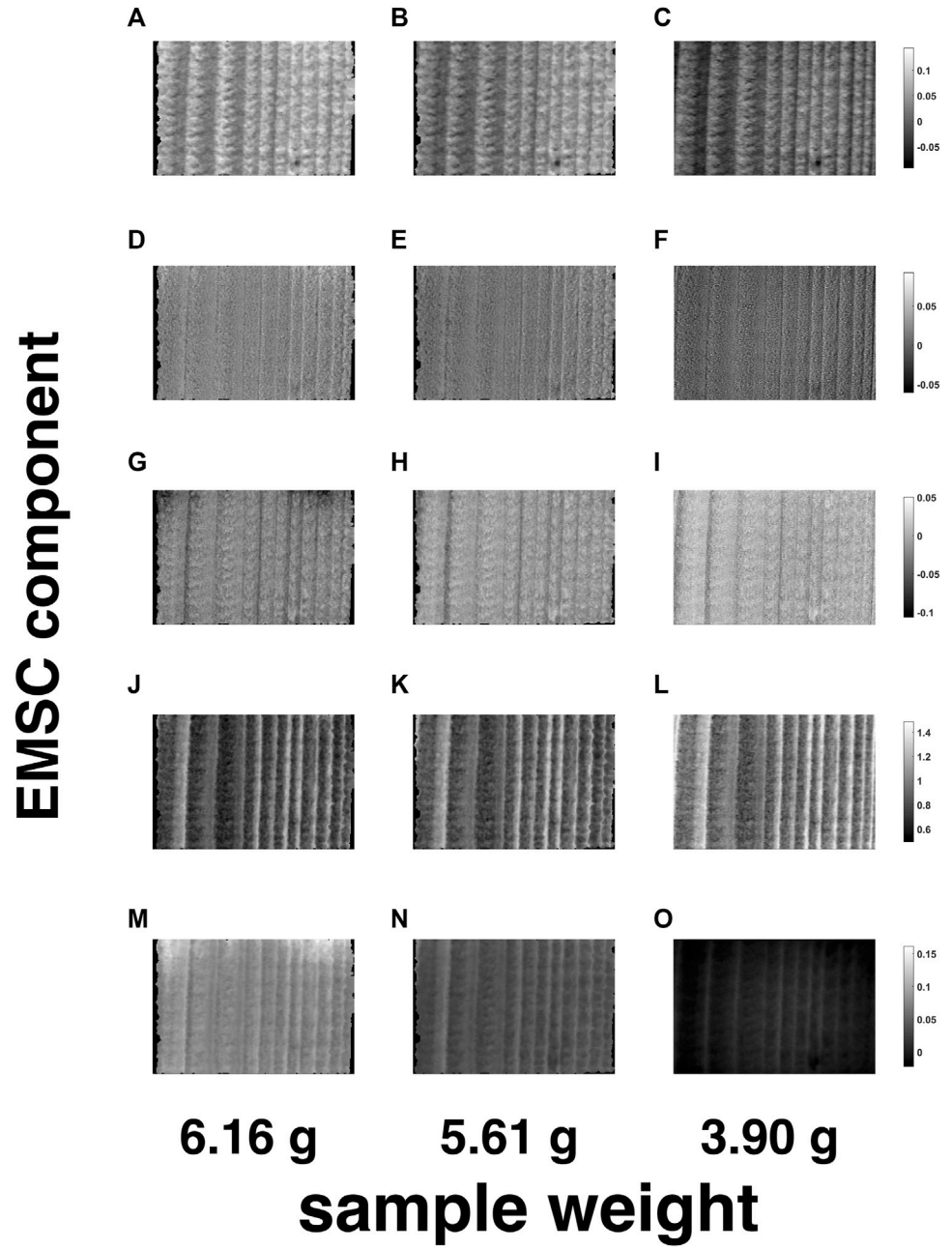

FIGURE 11 | Spatial representation of the EMSC coefficients related to the EMSC components n. 1—**(A-C)**—n. 2—**(D-F)**—n. 3—**(G-I)**—n. 4—**(J-L)**—and n. 5—**(M-O)**—for three of the 10 archetypal frames highlighted in **Figure 10B**. The black areas around the loadings images contain pixels excluded from the computational procedure.

during wood drying affecting mainly the water bands of such NIR spectra (see the loadings in **Figures 10C,D**) and characterised by 10 archetypal time instants (see the grey dots in **Figure 10B**)—Ruckebusch et al. (2020). **Figures 11**, **12** provide an illustration of the distribution of the EMSC coefficients and the OTFP scores over the surface of the wood sample at three of these time instants. This representation allows assessing the aforementioned transition process at a spatial level: overall, the coefficient spatial distribution seems to get smoother as the experiment evolves, which might be explained in the light of the continuous migration/diffusion of water molecules through the pores of the wood specimen (see, *e.g.*, **Figure 12D–F**). However, all these aspects will be investigated in future research also by means of more rational subspace axis rotations—performed, for instance,

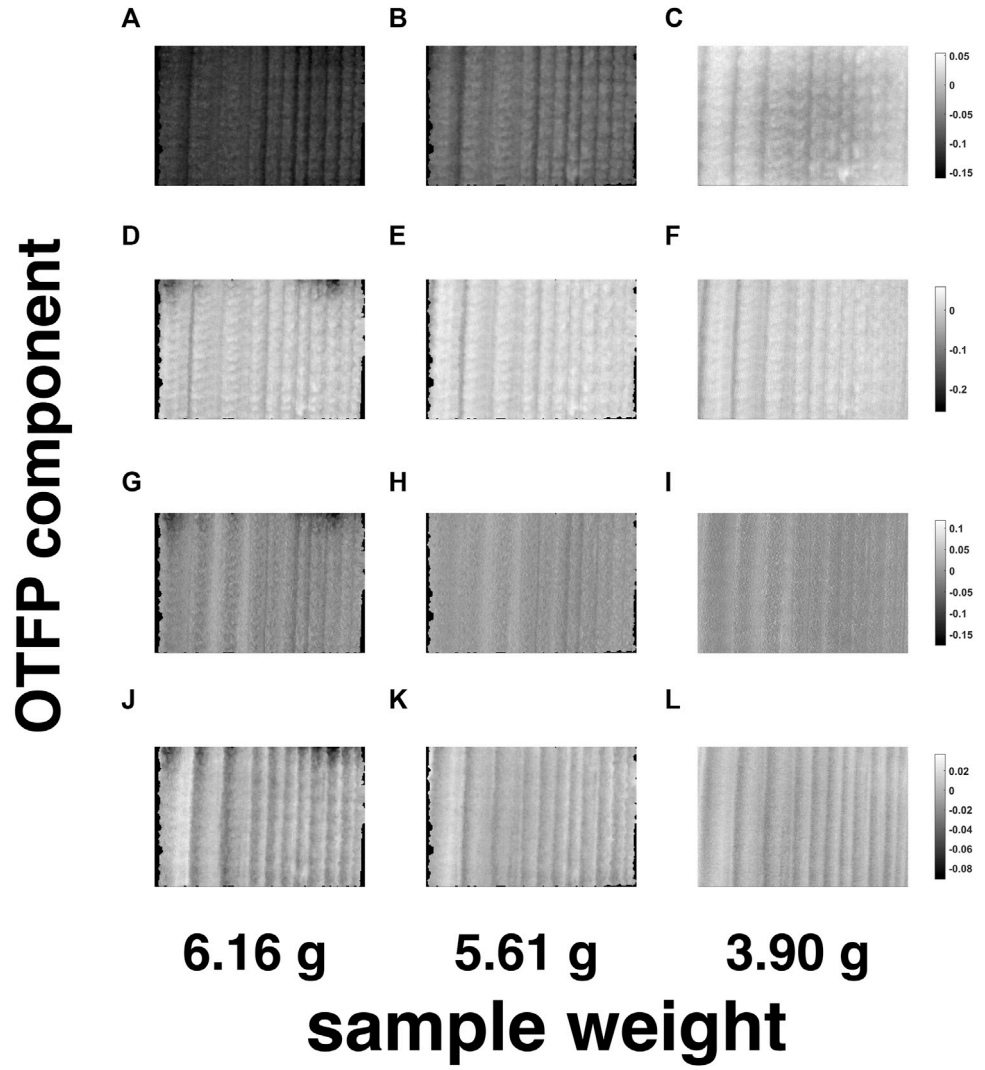

FIGURE 12 | Spatial representation of the OTFP scores related to the OTFP factors n. 1—**(A-C)**—n. 2—**(D-F)**—n. 3—**(G-I)**—n. 4—**(J-L)**—for three of the 10 archetypal frames highlighted in **Figure 10B**. The black areas around the loadings images contain pixels excluded from the computational procedure.

by varimax, Kaiser (1958), Independent Component Analysis (ICA), Comon (1994); Hyvärinen et al. (2001), or MCR-ALS—aimed at optimising the meaningfulness of the OTFP factors from a physico-chemical perspective.

5 CONCLUSION

Hyperspectral videos generate a lot of informative data. However, these data require efficient mathematical modelling for being reliable, understandable and quantitatively interpretable. Here, a general framework by which hyperspectral videos can be analysed was proposed. The three computational steps of this framework result in a compact multi-domain hybrid subspace modelling approach, involving spatial, spectral and temporal parametrisation of both known and unknown chemical and physical phenomena underlying the studied systems. IDLE permits to characterise and compensate the complex motions that the investigated objects may undergo over the measurement time. EMSC is capable of providing a simple mathematical description of a range of phenomena (and of their temporal evolution) that operators expect or presume *a priori* to be occurring over the duration of the hyperspectral video recording. Finally, the OTFP compresses and summarises all the information related to unknown or unexpected events which may happen during the progression of the data collection. In other words, one can look at the combination of these three different methodologies as an algorithmic extension of how human beings observe reality: the eyes capture spatial changes in the external environment and submit particular signals to the brain that afterwards processes them distinguishing between

what was somehow forecastable in advance (based on past experiences) and what is completely new and unforeseen. In this regard, rather than the individual application of the aforementioned techniques (some of which are already well-established in the field of chemometrics), it is their fusion into a comprehensive algorithmic architecture for the global assessment and interpretation of time-series of high-dimensional hyperspectral images to be innovative and unprecedented.

The sequential IDLE-EMSC-OTFP hybrid framework presented here rests on a combination of targeted and non-targeted data modelling of both known and unknown variation sources. In contrast to classical subspace decomposition strategies (*e.g.*, PCA, PLS, MCR-ALS, NNMF and ICA), it enables the description not only of additive spectral response variations, but also of multiplicative ones (like physical structure effects on the optical pathlength) and hard and soft shape changes (due, for example, to sample repositioning and/or shrinkage).

Moreover, differently from machine learning methods based on Artificial Neural Networks (ANN)—Gasteiger and Zupan (1993)—and Convolutional Neural Networks (CNN)—Gu et al. (2018)—the IDLE-EMSC-OTFP modelling approach yields a strong dimensionality reduction of torrents of input data and results graphically interpretable in their compressed state, revealing how spectral properties, spatial patterns and temporal dynamics are strictly intertwined into unified variation components, whose assessment and interpretation might provide fundamental insights into underlying chemical, physical and instrumental causalities. In the future, relying on a trilinear rather than bilinear OTFP model structure—exploiting, for instance, the principles of Parallel Factor Analysis (PARAFAC), Harshman (1970); Carroll and Chang (1970); Bro (1997)—may enhance this process further.

These conclusions are substantiated and thoroughly corroborated by the outcomes reported in this article. In fact:

1. motion estimation-compensation by spatiotemporal IDLE modelling allowed shrinkage induced by wood drying to be modelled and corrected for, reducing the spatial complexity of the hyperspectral imaging data;
2. EMSC preprocessing permitted a simpler spectral modelling by detecting and disentangling light absorption/light scattering-related variation patterns and their respective evolution over time;
3. the continuous data-driven bilinear subspace decomposition returned by the OTFP enabled the study of the dynamics of the various physical and chemical variations left unmodelled in the stream of hyperspectral residuals after the previous two steps.

In the light of all this and considering its computational efficiency when massive (potentially ever-lasting) flows of multi-channel measurements are handled, the developed approach could have an enormous impact also within the more general context of BIG DATA.

DATA AVAILABILITY STATEMENT

Data are available under request. Inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RV and HM conceived and designed the study. IB collected and organised the database. RV and HM performed the statistical data analysis. CR and IB validated both outcomes and conclusions. RV wrote the first draft of the article. RV, CR, IB and HM wrote sections of the paper. All authors contributed to the revision of the manuscript and approved its submitted version.

REFERENCES

Barnes, R. J., Dhanoa, M. S., and Lister, S. J. (1989). Standard normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* 43, 772–777. doi:10.1366/0003702894202201

Beer, A. (1852). Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. *Ann. Phys. Chem.* 162, 78–88. doi:10.1002/andp.18521620505

Bouguer, P. (1729). *Essai d'optique sur la gradation de la lumière*. Editor C. A. Jombert. First edn (Paris, France).

Bro, R. (1997). PARAFAC. Tutorial and Applications. *Chemometrics Intell. Lab. Syst.* 38, 149–171. doi:10.1016/s0169-7439(97)00032-4

Carroll, J. D., and Chang, J.-J. (1970). Analysis of Individual Differences in Multidimensional Scaling via an *N*-Way Generalization of "Eckart-Young" Decomposition. *Psychometrika* 35, 283–319. doi:10.1007/bf02310791

Chuvieco, E., and Kasischke, E. (2007). Remote Sensing Information for Fire Management and Fire Effects Assessment. *J. Geophys. Res.* 112, article number G01S90. doi:10.1029/2006jg000230

Comon, P. (1994). Independent Component Analysis, a New Concept? *Signal. Process.* 36, 287–314. doi:10.1016/0165-1684(94)90029-9

Cooley, J. W., and Tukey, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Comp.* 19, 297–301. doi:10.1090/s0025-5718-1965-0178586-1

Elmasry, G., Kamruzzaman, M., Sun, D.-W., and Allen, P. (2012). Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-Food Products: a Review. *Crit. Rev. Food Sci. Nutr.* 52, 999–1023. doi:10.1080/10408398.2010.543495

Endrizzi, I., Gasperi, F., Rødbotten, M., and Næs, T. (2014). Interpretation, Validation and Segmentation of Preference Mapping Models. *Food Qual. Preference* 32, 198–209. doi:10.1016/j.foodqual.2013.10.002

Fischer, C., and Kakoulli, I. (2006). Multispectral and Hyperspectral Imaging Technologies in Conservation: Current Research and Potential Applications. *Stud. Conservation* 51, 3–16. doi:10.1179/sic.2006.51.supplement-1.3

Gasteiger, J., and Zupan, J. (1993). Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* 32, 503–527. doi:10.1002/anie.199305031

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent Advances in Convolutional Neural Networks. *Pattern Recognition* 77, 354–377. doi:10.1016/j.patcog.2017.10.013

Harshman, R. (1970). Foundations of the PARAFAC Procedure: Models and Conditions for an "explanatory" Multimodal Factor Analysis. *UCLA Working Pap. Phonetics* Vol. 16, 1–84. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.5652&rep=rep1&type=pdf

Hay, G. J., Kyle, C., Hemachandran, B., Chen, G., Rahman, M. M., Fung, T. S., et al. (2011). Geospatial Technologies to Improve Urban Energy Efficiency. *Remote Sensing* 3, 1380–1405. doi:10.3390/rs3071380

Horn, B. K. P., and Schunck, B. G. (1993). "Determining Optical Flow": a Retrospective. *Artif. Intelligence* 59, 81–87. doi:10.1016/0004-3702(93)90173-9

Horn, B. K. P., and Schunck, B. G. (1981). Determining Optical Flow. *Artif. Intelligence* 17, 185–203. doi:10.1016/0004-3702(81)90024-2

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* 24, 417–441. doi:10.1037/h0071325

Hugelier, S., Vitale, R., and Ruckebusch, C. (2020). *Image Processing in Chemometrics. Comprehensive Chemometrics*. Second edn., Vol. 4. Amsterdam, Netherlands: Elsevier, B.V., 411–436. doi:10.1016/b978-0-12-409547-2.14597-4:

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. First edn. Hoboken: United States of America: John Wiley & Sons, Ltd.

Kaiser, H. F. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika* 23, 187–200. doi:10.1007/bf02289233

Khan, M. J., Khan, H. S., Yousaf, A., Khurshid, K., and Abbas, A. (2018). Modern Trends in Hyperspectral Image Analysis: a Review. *IEEE Access* 6, 14118–14129. doi:10.1109/access.2018.2812999

Kubelka, P., and Munk, F. (1931). An Article on Optics of Paint Layers. *Z. Tech. Phys.* 12, 593–601. Available at: https://www.graphics.cornell.edu/westin/pubs/kubelka.pdf

Lambert, J. (1760). in *Photometria sive de mensura et gradibus luminis, colorum et umbrae*. Editor C. P. Detleffsen. First edn (Augsburg, Gerrmany.

Lawton, W. H., and Sylvestre, E. A. (1971). Self Modeling Curve Resolution. *Technometrics* 13, 617–633. doi:10.1080/00401706.1971.10488823

Lu, G., and Fei, B. (2014). Medical Hyperspectral Imaging: a Review. *J. Biomed. Opt.* 19, article number 010901. doi:10.1117/1.jbo.19.1.010901

Makeig, S., Enghoff, S., Jung, T., and Sejnowski, T. (2000). "Moving-window ICA Decomposition of EEG Data Reveals Event-Related Changes in Oscillatory Brain Activity," in Proceedings of the Second International Workshop on Independent Component Analysis and Signal Separation, 627–632.

Martens, H. (1979). Factor Analysis of Chemical Mixtures. *Analytica Chim. Acta* 112, 423–442. doi:10.1016/s0003-2670(01)85040-6

Martens, H., Jensen, S., and Geladi, P. (1983). "Multivariate Linearity Transformation for Near-Infrared Reflectance Spectrometry," in Proceedings of the Nordic Symposium on Applied Statistics, 205–234.

Martens, H., and Næs, T. (1989). *Multivariate Calibration*. First edn. Hoboken: United States of America: John Wiley & Sons, Ltd.

Martens, H., Nielsen, J. P., and Engelsen, S. B. (2003). Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures. *Anal. Chem.* 75, 394–404. doi:10.1021/ac020194w

Martens, H. (2015). Quantitative Big Data: where Chemometrics Can Contribute. *J. Chemometrics* 29, 563–581. doi:10.1002/cem.2740

Martens, H., and Stark, E. (1991). Extended Multiplicative Signal Correction and Spectral Interference Subtraction: New Preprocessing Methods for Near Infrared Spectroscopy. *J. Pharm. Biomed. Anal.* 9, 625–635. doi:10.1016/0731-7085(91)80188-f

Matikainen, L., and Karila, K. (2011). Segment-Based Land Cover Mapping of a Suburban Area-Comparison of High-Resolution Remotely Sensed Datasets Using Classification Trees and Test Field Points. *Remote Sensing* 3, 1777–1804. doi:10.3390/rs3081777

McMillen, J. (1964). Wood Drying-Techniques and Economics. Approved Technical Article, Food Products Laboratory, Forest Service, U.S. Department of Agriculture. Available at: https://www.fpl.fs.fed.us/documnts/pdf1964/mcmil64a.pdf

Pearson, K. (1901). LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Phil. Mag. J. Sci.* 2, 559–572. doi:10.1080/14786440109462720

Ruckebusch, C., Vitale, R., Ghaffari, M., Hugelier, S., and Omidikia, N. (2020). Perspective on Essential Information in Multivariate Curve Resolution. *Trac Trends Anal. Chem.* 132, article number 116044. doi:10.1016/j.trac.2020.116044

Silva, C. S., Pimentel, M. F., Amigo, J. M., Honorato, R. S., and Pasquini, C. (2017). Detecting Semen Stains on Fabrics Using Near Infrared Hyperspectral Images and Multivariate Models. *Trac Trends Anal. Chem.* 95, 23–35. doi:10.1016/j.trac.2017.07.026

Stefansson, P., Fortuna, J., Rahmati, H., Burud, I., Konevskikh, T., and Martens, H. (2020). *Hyperspectral Time Series Analysis: Hyperspectral Image Data Streams Interpreted by Modeling Known and Unknown Variations*. First edn., Vol. 32. Amsterdam, Netherlands: Elsevier, B.V., 305–331. doi:10.1016/b978-0-444-63977-6.00014-6:

Tauler, R., Smilde, A., and Kowalski, B. (1995). Selectivity, Local Rank, Three-Way Data Analysis and Ambiguity in Multivariate Curve Resolution. *J. Chemometrics* 9, 31–58. doi:10.1002/cem.1180090105

Vitale, R., Hugelier, S., Cevoli, D., and Ruckebusch, C. (2020a). A Spatial Constraint to Model and Extract Texture Components in Multivariate Curve Resolution of Near-Infrared Hyperspectral Images. *Analytica Chim. Acta* 1095, 30–37. doi:10.1016/j.aca.2019.10.028

Vitale, R., and Saccenti, E. (2018). Comparison of Dimensionality Assessment Methods in Principal Component Analysis Based on Permutation Tests. *Chemometrics Intell. Lab. Syst.* 181, 79–94. doi:10.1016/j.chemolab.2018.08.008

Vitale, R., Stefansson, P., Marini, F., Ruckebusch, C., Burud, I., and Martens, H. (2020b). *Fast Analysis, Processing and Modeling of Hyperspectral Videos: Challenges and Possible Solutions. Comprehensive Chemometrics*. Second edn., Vol. 4. Amsterdam, Netherlands: Elsevier, B.V., 395–409. doi:10.1016/b978-0-12-409547-2.14605-0:

Vitale, R., Westerhuis, J., Næs, T., Smilde, A., de Noord, O., and Ferrer, A. (2017a). Selecting the Number of Factors in Principal Component Analysis by Permutation Testing - Numerical and Practical Aspects. *J. Chemometr* 31, article number e2937. doi:10.1002/cem.2937

Vitale, R., Zhyrova, A., Fortuna, J. F., de Noord, O. E., Ferrer, A., and Martens, H. (2017b). On-The-Fly Processing of Continuous High-Dimensional Data Streams. *Chemometrics Intell. Lab. Syst.* 161, 118–129. doi:10.1016/j.chemolab.2016.11.003

Walczak, B. (2000). *Wavelets in Chemistry, Data Handling in Science and Technology*. First edn. Amsterdam, Netherlands: Elsevier, B.V.

Wang, X., Kruger, U., and Irwin, G. W. (2005). Process Monitoring Approach Using Fast Moving Window PCA. *Ind. Eng. Chem. Res.* 44, 5691–5702. doi:10.1021/ie048873f

Westad, F., and Martens, H. (1999). Shift and Intensity Modeling in Spectroscopy-General Concept and Applications. *Chemometrics Intell. Lab. Syst.* 45, 361–370. doi:10.1016/s0169-7439(98)00144-0

Wold, S., Martens, H., and Wold, H. (1983). *Matrix Pencils. Lecture Notes in Mathematics*. in *Chap. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method*. First edn., Vol. 973. Berlin/Heidelberg, Germany: Springer-Verlag, 286–293. doi:10.1007/bfb0062108

Conflict of Interest: HM is a co-founder of Idletechs AS which develops and commercialises the On-The-Fly Processing (OTFP) tool.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership