Multi-block data analysis for online monitoring of anaerobic co-digestion process

Lorraine-Fifame Awhangbo, R. Bendoula, Jean-Michel Roger, Fabrice Béline

Multi-block data analysis for online monitoring of anaerobic co-digestion process

L. Awhangbo[1, 4], R. Bendoula[2], J.M. Roger[2, 3] & F. Béline[1]

[1]Irstea, UR OPAALE, 17 av. de Cucillé, CS 64427, F-35044, Rennes, France
[2] ITAP, Univ Montpellier, Irstea, Montpellier SupAgro, 361, rue J.F. Breton, BP 5095, F-34196, Montpellier, France
[3] Chemhouse Research group, Montpellier, France
[4] Univ. Bretagne Loire, France

Awhangbo Lorraine
Bendoula Ryad (Corresponding author: ryad.bendoula@irstea.fr)
Roger Jean-Michel jean-michel.roger@irstea.fr
Fabrice Béline fabrice.beline@irstea.fr

Abstract

Anaerobic digestion is a chemical process whose purpose is to maximize biogas production whilst concomitantly treating organic waste mostly through co-digestion due to the variety of substrates. To avoid failures, the process requires the monitoring of several parameters and or inhibitors. The existing strategies and methods used in the process monitoring still lack sensitivity and robustness, when taken individually. The current study investigated the use of sequential and orthogonalized partial least squares (SO-PLS) regression to relate these parameters to several blocks of data coming for near infrared spectroscopy, chemical routine analysis and kinetics of biogas production. The models produced were able to extract relevant information from each block's data and discard redundancies. Moreover, to meet biogas plant operators' requirements, variable selection was performed on the infrared blocks using a recent method: SO-CovSel. SO-CovSel is a method resulting from coupling SO-PLS and Covariance Selection (CovSel) method. The method has been demonstrated to be suitable for multi-response calibration purposes with infrared calibration. It has provided good predictions and provided an interesting interpretation of wavelengths involved in the monitoring of parameters of stability in anaerobic co-digestion.

Keywords: Anaerobic digestion, online monitoring**,** Multi-block analysis, SO-PLS, SO-CovSel

ABBREVIATIONS

AD Anaerobic Digestion

AcoD Anaerobic co-Digestion

LCFA Long Chain Fatty Acids

NH_4^+ ammonium

NIRS Near InfraRed Spectroscopy

OLR Organic Load Rate

PoLiS Polarization Light Spectroscopy

R(λ) reflectance from the immersed probe

$R_{bs}(\lambda)$ total backscattered reflectance of the remote probe

$R_{ms}(\lambda)$ multiple scattered reflectance of the remote probe

$R_{ss}(\lambda)$ single scattered reflectance of the remote probe

TS Total Solids

VFA Volatile Fatty Acids

VS Volatile Solids

1. INTRODUCTION

Anaerobic digestion (AD) is a sensitive process involving a synergistic effort of a diverse group of microbial communities for metabolizing diverse organic substrates and then producing biogas. However, AD usually features process imbalances and disturbances due to environmental fluctuations and inhibitions, especially in anaerobic co-digestion (AcoD). While AcoD can enhance methane production due to several co-substrates [1], substrate diversity also increases environmental fluctuation risks. Co-substrates such as nitrogen-rich products, fats, oils and grease (FOG), highly biodegradable products… boost methane production due to their high methanogenic potential. However these co-substrates might induce the accumulation of volatile fatty acids (VFA), ammonia/ammonium (NH_3/NH_4^+) and long chain fatty acids (LCFA), known to severely affect anaerobic processes. Thus, control strategies and instrumentations must be up to the level of development of the process, including AcoD specificities. Several studies have explored AD process parameters and have found some parameters (pH, Alkalinity, VFA …) which are characteristic of the process state and used for its monitoring [2]. These parameters reflect, to a certain extent, the microbial behavior of the digester, but are not always accurate especially during AcoD where the quality and quantity of substrate vary over time. Furthermore, the recommended early warning indicators are different among digesters [3]. Moreover, these indicators were only effective for some specific substrates and operating conditions probably due to their different sensitivities to environmental fluctuations in different AD systems [3-4].

Online monitoring and process control techniques could achieve a high-efficiency and stabilized performance of the process [5]. Ideal monitoring methods should be fast, sensitive, non-destructive, robust, and give early indications of imbalance in the microbial status of the process [6]. In addition, they should generate data on several parameters without analyte consumption and interferences with the metabolism of the process and be resistant to the changing environments encountered in reactors [7]. In this context, spectroscopic techniques, often coupled to optical fibers, seem suitable for processes monitoring and have been investigated for different bioprocess applications, in wavelength ranges including ultraviolet–visible (UV–Vis), near infrared (NIR) and mid infrared (MIR). Near infrared spectroscopy (NIRS), has been usually chosen over the other techniques for AD monitoring because it is an ideal compromise between cost, technology and in-situ applicability. Information must be extracted from the acquired spectra by chemometric techniques, mainly Partial Least Squares regression (PLS), to produce calibrations and estimate the key parameters. Subsequently, several studies have been conducted based on NIRS to determine the stability of the process via VFA prediction [8-11]. Parameters such as Total and Volatile Solids (TS/VS), ammonium (NH_4^+) [12] and ammonia [13] were also predicted with some success. However, the validity of these models and their robustness depend on the concentration range of the predicted parameter [14], the feed substrates and the proportion of TS contents in the final digestate's samples [15]. Indeed, low TS content in the digestate suggests high water content. And, water has several absorbance bands in the NIR region which therefore upset PLS model calibration [16].

Thus, although a range of electrochemical, titrimetric, chromatographic and spectroscopic devices can be deployed for online monitoring and control of AD process, none seem to be ideal [6]. This may be attributed to all the possible interferences that could hinder the measures from these

different devices. However, it remains true that each technique has some part of valid information on the process state or behavior. A relevant hypothesis is that the combination of these measures could allow a better monitoring of the process. Combining information from many datasets can improve interpretation of the trends observed in the studied system. Indeed, modern monitoring technologies usually provide large amounts of datasets, from different devices, on processes. Rather than a separate analysis of these datasets, integrated approaches are necessary as they can improve the understanding of complex systems by (*i*) identifying common correlation trends; (*ii*) revealing hidden structures not known *a priori* and (*iii*) extracting more information from the studied datasets [17]. While single dataset analysis is widespread, integrated analysis of several different types of datasets, also called blocks, is challenging. Several methods for integrated analysis have been used in bioprocesses including Bayesian factor analysis [18]; network analysis [19] and multivariate linear projections such as multi-block analysis [20-21].

Recently, research in relation to the performance of multi-blocks algorithms have increased due to their usefulness in bioprocess monitoring. Multi-block methods are used to explore and model the relationships between several datasets. Most of these multivariate linear projection methods are based on PLS regression, for example: Hierarchical-PLS [22], Multi-Block-PLS (MB-PLS) [23], Sequential and Orthogonalized Partial Least Squares (SO-PLS) [24], Sequential and Orthogonalized multi-way version of PLS (SO-N-PLS) [25], Parallel Orthogonalized Partial Least Squares (PO-PLS) [26]. There is also Predictive-ComDim [27] which derived from Common Components and Specific Weights Analysis. Multi-block methods are generally used in biological systems such as metabolomics, industrial pharmaceutical process or quality control, and enable important biological conclusions on the monitored process [17].

AcoD is also a biological system that could use integrated analysis of multiple datasets. Digesters are generally equipped with basic sensors for their monitoring. A survey on 400 full-scale AD plants indicated that 95% of their in-line instrumentation was limited to pH, temperature, water flow, biogas flow (quantity or kinetics), level and pressure [28]. There is also an indirect level of monitoring through the digester's feed substrates. Given its development on digester monitoring, NIRS analysis on digestate could also become a routine analysis on biogas plants, representing another source of information. It is therefore possible to collect, on one hand, an easily acquirable dataset of physicochemical parameters on the digester and, on the other hand, a spectral dataset of digestates. As control strategies, multi-block analysis based on these inline instrumentations could be the answer to prevent biogas plants failures.

To make this monitoring strategy more productive, variable selection procedure can also be explored. Variable selection allows selection of a sub-set of variables to be used for the creation of a reduced regression model. This procedure is used to remove non informative and noisy variables and to improve model's prediction and interpretation ability. Moreover, reducing variables can simplify their acquisition. The system could become cheaper, especially for NIRS, and less time consuming [29]. The cost-effective point is primordial for biogas plant operators. There are several methods of variable selection in the context of PLS regression [30-31]. A recent method called covariance selection (CovSel) was also made for variable selection with multi-response calibration as usually encountered with infrared devices [31]. It is only recently that variable selection methods have been introduced in a multi-block context. For example, combinations of three variable selection methods (variable importance in projection (VIP), selectivity ratio (SR), and forward

selection) were thoroughly examined with MB-PLS and SO-PLS [32]. SO-PLS was also joined with a CovSel, resulting in a new method of sequential and orthogonalized covariance selection (SO-CovSel) [33]. SO-CovSel was found to be very parsimonious in selecting variables making the method suitable for a number of practical applications.

Therefore, the objective of this study was to explore the applicability of SO-PLS method to predict relevant parameters in AcoD using inline physicochemical parameters and NIRS measurements on digestates. AcoD experiments were conducted with highly biodegradable or fat-containing co-substrates, known to induce inhibitions in the digester but also to create interferences and hinder infrared measurements. In particular, two infrared probes were evaluated for the prediction of state indicators such as VFA, LCFA and NH_4^+. Focus was put on interpretation as well as prediction ability of these multi-block models and how to assess reliability of the interpretations. Variable selection allows focusing on relevant wavelengths from the used probes.

2. MATERIALS AND METHODS

2.1.Digester operating conditions

A continuously stirred tank reactor (CSTR) with a working volume of 35L was operated at 38± 1°C. A semi-continuous feeding mode was used. The substrate mixture (kept at 4°C) was added once every morning prior to digestate sampling. Different operating conditions were tested with organic load rate (OLR) varying between 1.5 and 5kgCOD.m^{-3}.d^{-1} (Table 1). Digestate samples were collected from these performed co-digestion experiments.

Table 1: co-digestion experiments performed with their characteristic

N°	Substrates Mixture	OLR
1	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270g.l^{-1}) + Food fats (20g.l^{-1})	4.9
2	Pig slurry + Horse feed residues (20g.l^{-1}) + Catering waste (200 g.l^{-1})	3.4
3	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270 g.l^{-1})	2.2
4	Pig slurry + Horse feed residues (40g.l^{-1}) + Fruit waste (540 g.l^{-1})	4.2
5	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270g.l^{-1})	1.4
	Pig slurry + Horse feed residues (40g.l^{-1}) + Fruit waste (540g.l^{-1})	3.0
6	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270g.l^{-1})	1.7
7	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270g.l^{-1})	1.6
	Pig slurry + Fruit waste (135g.l^{-1}) + Protein waste (20g.l^{-1})	3.8
8	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270 g.l^{-1})	1.7
	Pig slurry + Horse feed residues (20g.l^{-1}) + Fruit waste (270 g.l^{-1}) + Food fats (20g.l^{-1})	3.4

OLR: organic load rate

2.2.Chemical analysis

2.2.1. Potential online parameters

As stated, the in-line instrumentation of digesters is limited [28]. In this study, pH was measured, both in the digestate (pH_{out}) and in the feed substrates (pH_{in}), according to APHA Standard Methods [34]. The brief composition of the substrate, i.e. feed mass and their contents in Total Solids (TS) and Volatile Solids (VS), is also easily determined or estimated.

Biogas production was automatically determined with a wet tipping bucket flow meter connected to

the acquisition program. The operating of these gas meters is based on the "tipping bucket" principle in which liquid is displaced by gas in a specially-designed chamber [35]. Such a gas meter also allowed the measurement of periodic tipping of the container. Therefore, Biogas Production Rate (BPR) was calculated by dividing the volume of biogas (36.4 ml) by the time elapsed since the previous tipping, considering the volume and the temperature of the headspace of the digester constant. Each BPR obtained was called instantaneous BPR and all BPR obtained between two feedings, corresponding to a day in our case, constituted the evolution of the instantaneous BPR and was called daily BPR kinetics. These daily BPR kinetics were always linearly interpolated with a time step of 0.001 day to have the same number of points (1001) for each kinetic. Standard BPR kinetics obtained is described in figure 1 which showed the three steps which occurred daily on the digester. Changes in BPR kinetics were found to be sensitive of the digester's failures.

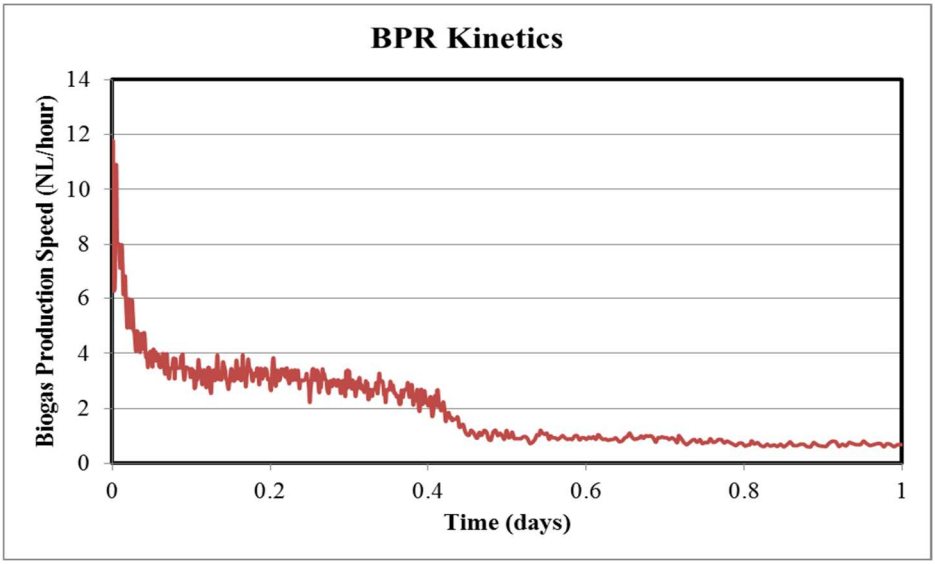


Figure 1: Example of BPR kinetics acquired daily on the digester.

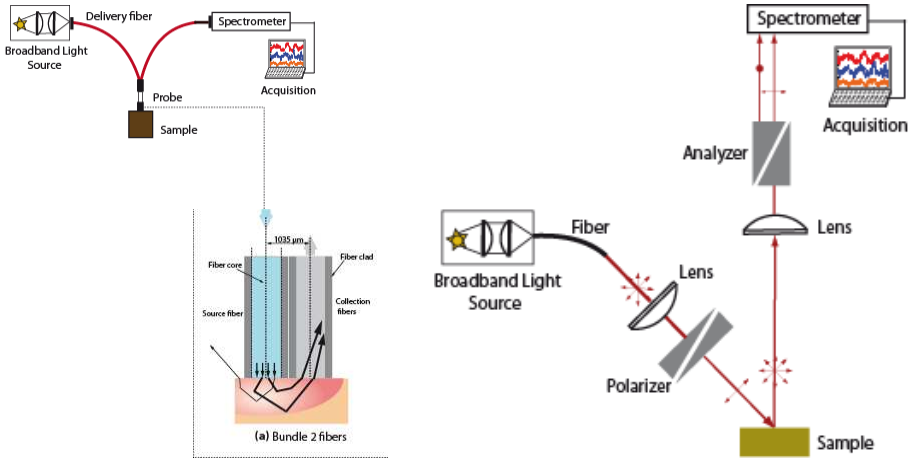
2.2.2. Reference analysis

Ammonium (NH_4^+), VFA and LCFA contents were determined on all digestates samples. VFA contents were determined, on the supernatant after centrifugation of the samples, with high performance liquid chromatography (HPLC, Varian©, U3000). Gas chromatography/mass selective (GC/MS, Agilent Technologies, 7890B/5977A) was used to determine LCFA. The procedure required LCFA extraction from raw samples with hexane/isopropanol (3/2) solvent using an accelerated solvent extractor (Dionex, ASE 350) followed by a concentration step before the chromatographic analysis. TS and VS were also determined on digestate samples.

2.3.NIRS Probes

Two spectroscopic systems (Figure 2) were tested off-line on the collected digestate samples, maintained at the reactor bath temperature and scanned whilst 'fresh'. Both systems use the same light source (Tungsten-Halogen source: Ocean Optics HL-200-FHSA) and the same spectrometer (LabSpec1: ASD, Boulder). The spectral range of measurement extends from 350nm to 2500nm with a step of 1nm, and a resolution of 3nm for the range 350nm-1000nm and 10nm for the range 1000nm - 2500nm. The first probe is an immersible probe, based on diffuse optical spectroscopy and a reflectance $R(\lambda)$ is computed from this measurement. The second spectroscopic system is a remote probe, based on polarization light spectroscopy (PoLiS) to reduce deformations of spectra

especially in biological media such as digestates. From this system measurements, the weakly scattered reflectance $R_{ss}(\lambda)$ and the multiple scattered reflectance $R_{ms}(\lambda)$ were computed for each sample and summed into the total backscattered reflectance $R_{bs}(\lambda)$ [36]. The specificities of these probes are fully described in Awhangbo et al., 2020 [37].



(a) **(b)**

Figure 2: Schematic (a) the immersed probe and (b) the remote or polarized probe

2.4.Data Sets

Four infrared predictor blocks were considered in this study: X_1, X_2, X_3 and X_4 respectively from the multiple scattered reflectance $R_{ms}(\lambda)$, the weakly scattered reflectance $R_{ss}(\lambda)$ and the total backscattered reflectance $R_{bs}(\lambda)$ of the polarized probe and the reflectance $R(\lambda)$ from the immersed probe. These infrared data were preprocessed using Savitsky–Golay (SG) smoothing or derivate algorithm [38], detrending [39] and Standard Normal Variate (SNV) [40] before computations as described in table 2 below. Noise presence and baseline deviation were considered when choosing the appropriate pretreatments. To these infrared blocks, were added blocks X_5 and X_6. X_5 consisted of a block of six chemical parameters (pH_{in}, pH_{out}, feed mass, TS, VS and biogas production) easily acquirable on the digesters cited above. X_6 corresponded to data from BPR kinetics acquired during the process. The response block Y consists of state indicators VFA, LCFA and ammonium.

166 samples were used in this study and all analyzed by each spectroscopic probe and listed chemical technique.

Table 2: Preprocessing performed on the spectra with respect to the parameters and on each probe.

Signals	Pre-processing
Total VFA	
$R(\lambda)$	Reflectance, 1st Derivative SG 71pt window, SNV, 450–1700nm
$R_{bs}(\lambda)$	Reflectance, Smoothing SG 71pt window, 450–1700nm
$R_{ms}(\lambda)$	Reflectance, Smoothing SG 71pt window, 450–1700nm
$R_{ss}(\lambda)$	Reflectance, Smoothing SG 71pt window, 450–1700nm
Total LCFA	
$R(\lambda)$	Reflectance, 2nd Derivative SG 71pt window, 450–1700nm
$R_{bs}(\lambda)$	Reflectance, 1st Derivative SG 71pt window, 2-order Detrend, 450–1700nm
$R_{ms}(\lambda)$	Reflectance, 1st Derivative SG 71pt window, 2-order Detrend, 450–1700nm
$R_{ss}(\lambda)$	Reflectance, 1st Derivative SG 71pt window, 2-order Detrend, 450–1700nm

NH_4^+	
$R(\lambda)$	Absorbance, Smoothing SG 71pt window, 450–1800nm
$R_{bs}(\lambda)$	Absorbance, Smoothing SG 71pt window, SNV, 450–1800nm
$R_{ms}(\lambda)$	Absorbance, Smoothing SG 71pt window, 2-order Detrend, 450–1800nm
$R_{ss}(\lambda)$	Reflectance, 1st Derivative SG 71pt window, 450–1800nm

2.5.Computations

Experiments N°1,2,3,4&7 (107 measures) were used in training and experiments N°5,6&8 (59 measures) were used in test. Cross-validation procedures were first performed on the training set to deduce the number of latent variables (LV). The same split (training/ test) and CV blocks were used in all models. These models feature performance parameters such as the coefficient of determination (R^2), the Root Mean Squared Error of Cross-Validation (RMSECV) and of Prediction (RMSEP). All computations and multivariate data analysis were performed with Matlab software v. R2013b (The Mathworks Inc., USA).

3. THEORY

3.1.Sequential and Orthogonalised Partial Least Square Regression (SO-PLS)

In SO-PLS, the general multi-block linear regression with N blocks of independent variables can be represented by the equation [41]:

$$Y=X_1B_1+X_2B_2+...+X_NB_N+E$$

where $X_1(A\times J)$, $X_2(A\times M)$ and $X_N(A\times N)$ are the N predictor blocks with the same number of observation A; $B_1(J\times R)$, $B_2(M\times R)$ and $B_N(N\times R)$ are the regression coefficients ; $E(A\times R)$ is the residual matrix and $Y(A\times R)$ the response block. SO-PLS algorithm consists in sequentially extracting information from different data structures. Orthogonalization performed in SO-PLS allows removing redundancies among blocks and help focusing on the incremental contributions of each new block. The optimal number of components in the model can be defined for each block by: incremental or global estimation of components giving the lowest RMSECV. A detail algorithm of the SO-PLS method is available in literature [24, 41].

3.2.Sequential and Orthogonalised Covariance Selection (SO-CovSel)

As mentioned above, SO-CovSel method results from the coupling of SO-PLS and CovSel [33]. SO-CovSel algorithm presents the same structure as SO-PLS, but the reduction features achieved by PLS are operated by CovSel. The scheme of the algorithm, considering a case of two predictor blocks $X_1(B\times J)$ and $X_2(B\times M)$ and a response Y (B×R), is summarized in the following steps:

- First CovSel is applied in order to select features from X_1, by selecting the most useful variable and projecting the data orthogonally to this selected variable, hence obtaining $X_{1_{Sel}}$,
- Next, an ordinary least squares fit of Y to $X_{1_{Sel}}$ as $Y = X_{1_{Sel}} B_{X_1} + E_Y$, B_{X_1} with the coefficients and E_Y the residuals,
- Similarly with the SO-PLS algorithm, X_2 is orthogonalized with respect to $X_{1_{Sel}}$ obtaining X_2^{orth},
- CovSel is next applied to X_2^{orth} obtaining the reduced matrix $X_{2_{Sel}}^{orth}$,

- The previously estimated residuals E_Y are fitted to $X_{2_{Sel}}^{orth}$ with another ordinary least squares regression as $E_Y = X_{2_{Sel}}^{orth} B_{X_2} + E_{Y\,final}$, with B_{X_2} the new coefficients and $E_{Y\,final}$ the final residuals,
- The full predictive model is then computed as the ordinary least squares fit of Y to $X_{1_{Sel}}$ and $X_{2_{Sel}}^{orth}$ used as independent variables and can be written as:

$$Y = X_{1_{Sel}} B_{X_1} + X_{2_{Sel}}^{orth} B_{X_2} + E_{Y\,final}$$

In SO-CovSel, the number of variables selected on one input block affects the selection made on the following ones. The optimal number of variables to be selected in each block is evaluated through a procedure similar to the global approach used in SO-PLS. Consequently, all the possible combinations of numbers of selected variables are tested and the combination providing the lowest RMSECV is selected as the optimal one.

4. RESULTS AND DISCUSSION

4.1.Digester operation

The operating conditions used in these experiments induced different states in the digester. VFA and LCFA accumulated in the digester, for several experiments. VFA varied between 0 and 13500 mg/L, while LCFA varied between 0 and 2200 mg/L. NH_4^+ ranged from 1200 to 4100 mg/L. These high concentration values highlight the failures that occurred during these experiments. TS and VS varied little during all experiments due to a constant and low TS of the pig slurry used and highly biodegradable co-substrates added. TS content was found varying between 1.4 and 2.6% in all experiments.

4.2.SO-PLS models on infrared data

In a previous study [37], PLS and SO-PLS models were performed on single and different combinations of the infrared signals X_1, X_2, X_3 and X_4, with the same training / test split and the same blocks (in CV procedure) as in this study. Both global and incremental approaches were tested in the study as well as parsimony testing. The best model results from this previous study are summarized in Table 3.

Table 3: SO-PLS models results based on infrared signals

		X_2 & X_3 & X_4			X_3			X_1 & X_2 & X_4		
Training	Range (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)
VFA	0 – 13548	3-4-2	0.12	3931						
LCFA	0 – 2269				7	0.31	391			
NH_4^+	1180 – 4090							2-2-10	0.45	464
Test	Range (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)
VFA	0 – 10096	3-4-2	0.78	1343						
LCFA	0 – 251				7	0.62	188			
NH_4^+	1420 – 2530							2-2-10	0.68	496

Parameter predictions based on infrared blocks were improved by the SO-PLS multi-block analysis. For VFA, the most interesting model is a three-block model performed with signals $R_{ss}(\lambda)$ and $R_{bs}(\lambda)$ of the polarized probe, and $R(\lambda)$ of the immersed probe with a R^2 of 0.78 and a RMSEP of 1343mg/L. For NH_4^+, a three-block model performed with the decomposed signal of the polarized probe ($R_{ms}(\lambda)$ and $R_{ss}(\lambda)$) and the immersed probe $R(\lambda)$ have increased the accuracy of the parameter prediction with an R^2 of 0.68 and an error of 496 mg/l. For LCFA, only $R_{bs}(\lambda)$ signals have provided the most interesting model with a mono-block analysis. No combination was able to increase the performance of this model due to the particularity of this parameter. Indeed, LCFA in AcoD processes results in flotation phenomena in the sludge [42]. Therefore, only the remote probe was able to capture the variations of this parameter. It is worth noting that each parameter reacted differently to these infrared measures and this is highlighted by the different blocks used in their prediction.

4.3.SO-PLS models including chemical data

In order to further improve these infrared models, blocks X_5 and X_6 were also added for multi-block analysis. Prior to SO-PLS computations, PLS models were first built on blocks X_5 and X_6 in order to separately analyze their ability to predict the key parameters. The results of these models are shown in Table 4 below.

Models from the chemical data block $\mathbf{X_5}$ were not very successful, with low R^2 and high RMSECV and RMSEP. However, looking at R^2 values in CV procedures, these models still displayed some potential, especially for VFA and NH_4^+. And with the test set, NH_4^+ prediction was able to produce a significant R^2 (0.3) and a lower RMSEP (417 mg/l) than the best infrared multi-block model (496 mg/l). This suggests that X_5 does contain latent information about the digester state. Models with BPR kinetics produced more interesting results than chemical data blocks. Indeed they were able to predict all parameters except NH_4^+, with significant R^2 and low RMSEP errors. These models also display some potentialities in the prediction of the key parameters. Therefore, a multi-block approach can reconcile these models by extracting, for each block, the most interesting and useful latent variables with respect to the anaerobic process. It would also compensate the loss of information related to interference from infrared measurements.

Table 4: PLS models results based on Chemical data and BPR kinetics

		Chemical data (X_5)			BPR Kinetics (X_6)		
Training	Range (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)
VFA	0 – 13548	3	0.46	2700	3	0.09	3583
LCFA	0 – 2269	3	0.3	498	3	0.01	533
NH_4^+	1180 – 4090	6	0.07	661	6	0.06	667
Test	Range (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)
VFA	0 – 10096	3	0.00	3203	3	0.21	2598
LCFA	0 – 251	3	0.01	169	3	0.14	157
NH_4^+	1420 – 2530	6	0.30	417	6	0.01	362

Based on the previous multi-blocks results with infrared blocks [37], SO-PLS was now performed including blocks X_5 and X_6. The other infrared combinations were not tested with these new chemical blocks as, based on the deflation step used in the SO-PLS algorithm, the same information will be added to all the previous blocks used in the same order. And, the potentially best model would still be derived from these infrared data combinations. Therefore, SO-PLS models were computed with X_2, X_3, X_4, X_5 and X_6 for VFA, with X_1, X_2, X_4, X_5 and X_6 for NH_4^+ and with X_3, X_5 and X_6 for LCFA. The results of these models are summarized in Table 5 below. It is worth noting that these results were computed with the incremental approach. Indeed, from the previous work on infrared blocks [37], it was highlighted that this approach is better suited for the data used in this study, based on all SO-PLS models tested. To avoid over-fitting, parsimony testing was also used especially with the increase of the number of data blocks.

Table 5: SO-PLS models results based on infrared signals and chemical data.

		X_2 & X_3 & X_4 & X_5 & X_6			X_3 & X_5 & X_6			X_1 & X_2 & X_4 & X_5 & X_6		
Training	Range (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)
VFA	0 – 13548	3-4-2-1-1	0.17	3721						
LCFA	0 – 2269				7-1-3	0.28	411			
NH_4^+	1180–4090							2-2-10-6-0	0.35	515
Test	Range (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)
VFA	0 – 10096	3-4-2-1-1	0.85	1115						
LCFA	0 – 251				7-1-3	0.69	204			
NH_4^+	1420–2530							2-2-10-6-0	0.75	493

Most models have improved with the multi-block analysis of these data including chemical blocks X_5 and X_6. For all parameters CV model results were similar to CV results obtained with infrared data blocks (Table 2) with lower RMSECV in the case of VFA and similar errors in the cases of LCFA and NH_4^+. In the incremental approach, the blocks are optimized one after the other. The final multi-block model results are always influenced by each incremental step performed. High CV errors can also derive from the first incremental step taken in these models. The most noticeable improvements were found in the prediction of these parameters.

For VFA, R^2 improved from 0.78 to 0.85 with RMSEP lower by 17% than the multi-block model with only infrared data. Each chemical block only added one latent variable to achieve this improvement. In the incremental approach, the blocks are optimized one after the other. In the first incremental step with X_5 only, the model with 10 (3-4-2-1) LVs provided a RMSECV of 3921mg/l, a R^2 (test) of 0.78 and a RMSEP of 1388 mg/l which is similar to SO-PLS model with infrared data blocks. The final multi-block model results are always influenced by each incremental step performed. It can therefore be deduced that most additional information came from block X_6 of BPR kinetics. This is also confirmed by mono-block models of X_5 and X_6 which showed that block X_6 was better predictors of VFA than block X_6.

For LCFA, which was only well predicted by the polarized signal $R_{bs}(\lambda)$, the results were mitigated. While the multi-block model gained in accuracy with a R^2 improved from 0.62 to 0.69, it loss in

precision with a slightly higher RMSEP (204 mg/l) than the mono-block model (188 mg/l). This is due to first incremental step, (i.e. with only block X_5). Indeed, at this step, the model obtained with X_3 and X_6 with 8 (7-1) LVs provided a RMSECV of 391 mg/l, a R^2 (test) of 0.62 and a RMSEP of 184 mg/l. It should be noted that the RMSEP at this step was the smallest obtained. At the second incremental step, two outcomes were possible: a parsimonious model with no LV from block X_6 (7-1-0) because it has the lowest RMSECV or the second possibility which took information in block X_6 as shown in Table 5.

For NH_4^+, the model produced similar CV results as a multi-block model with only infrared data blocks. In test, the model provided a slightly lower RMSEP of 493 mg/l (previously 496 mg/L) with a R^2 which improved from 0.68 to 0.75. As expected, the model produced for NH_4^+ was parsimonious as no LV was selected in BPR kinetics (block X_6). As a reminder, the mono-block model of NH_4^+ prediction with BPR kinetics was not successful.

To better identify the improvement achieved by these multi-blocks model, scatter plots of predicted versus reference values of each model was made and analyzed for each parameter (Figure 3). These plots show that the predictions were accurate on the global behavior of the digester during these experiments, especially in the case of VFA accumulation. There was remarkable improvement was with VFA prediction where higher values were more precise when using all blocks (including chemical blocks). For LCFA and NH_4^+, with these new multi-block models, improvements were made on the prediction of lower values.

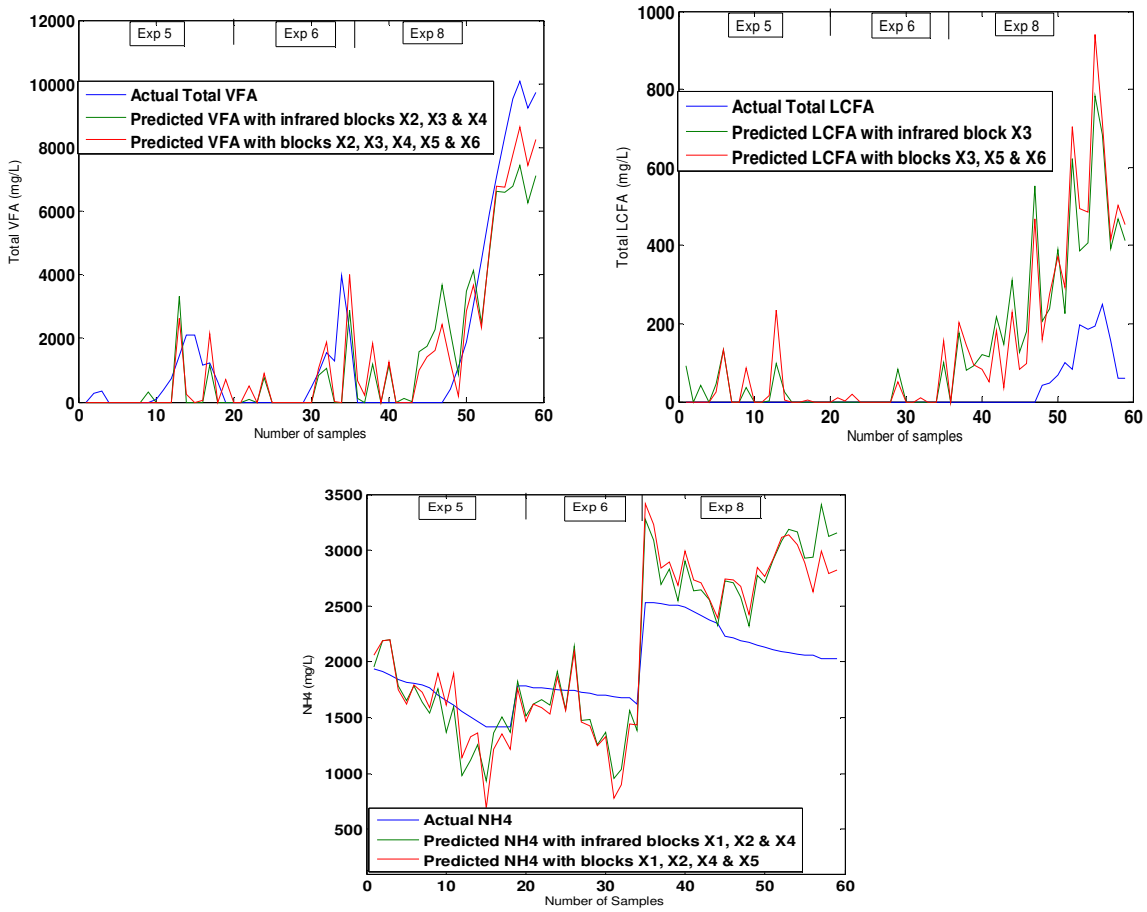


Figure 3: Scatter plots of measured versus predicted values for VFA, LCFA and NH_4^+ with each

multi-block model.

4.4.Variable selection with SO-CovSel on infrared blocks

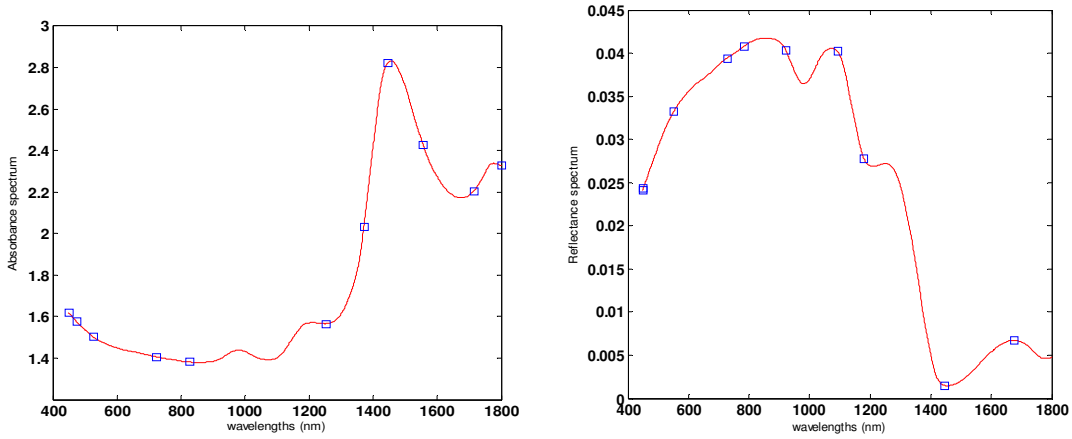
While for SO-PLS data are thoroughly preprocessed in order to build the best model, it is inappropriate to do so in variable selection as it might distort the selection process. Moreover, the infrared spectra were sometimes preprocessed differently for each parameter. Therefore, before performing SO-CovSel on these infrared data, spectra were only reduced to wavelengths used in SO-PLS (450 – 1850 nm) and smoothed using Savitsky–Golay, first with reflectance and then with absorbance. The same conditions of split (training / test) and CV-blocks as in SO-PLS were used for SO-CovSel models. To avoid over-fitting, the initial variable number was set based on LVs selected in PLS and SO-PLS regression. The results of the variable selection models performed with SO-CovSel are shown in Table 6.

Table 6: SO-CovSel models results based on infrared signals

		Reflectance			Reflectance			Absorbance		
		X_1, X_2, X_3 and X_4			X_3 and X_4			X_1, X_2, X_3 and X_4		
Training	Range (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)	LV	R^2	RMSECV (mg/l)
VFA	0 – 13548	5-3-0-2	0.11	4060	8-0	0.43	2771	0-0-0-11	0.44	2780
LCFA	0 – 2269		0.47	360		0.25	412		0.13	445
NH_4^+	1180–4090		0.06	729		0.1	678		0.12	610
Test	Range (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)	LV	R^2	RMSEP (mg/l)
VFA	0 – 10096	5-3-0-2	0.75	1420	8-0	0.66	1642	0-0-0-11	0.50	3390
LCFA	0 – 251		0.09	193		0.36	327		0.17	131
NH_4^+	1420–2530		0.22	452		0.18	370		0.58	512

In general, models produced by SO-CovSel are slightly less accurate than SO-PLS predictions [33], and this was observed in the results of the present study. Moreover, spectra were preprocessed differently for the two algorithms. Specifically, with reflectance spectra the model was parsimonious as no variable was selected in block X_3. This can be explained by the fact that X_3 is the global signal from decomposed signals X_1 and X_2. Selected wavelengths are resumed in figure 4 below. These wavelengths corresponded mostly to C-H third overtone assimilated to oils (450, 923nm), N-H overtone or amino groups (552, 785nm) and aromatic compounds (1093, 1446, 1677nm). With these wavelengths, VFA was the best predicted parameter with a R^2 of 0.75 and RMSEP of 1420 mg/l. LCFA prediction was not successful as no LV was selected in block X_3 which was previously identified as the most appropriate block for the prediction of this parameter. A SO-CovSel model with X_3 and X_4 resulted in a parsimonious model with 8 LVs from only X_3 and a better model result for LCFA with R^2 of 0.36. Similar wavelengths as in the previous model were selected (451, 1093, 1661nm) probably for VFA prediction. Wavelengths related to methyl groups (744, 841, 1800, 1353nm) frequently encountered in LCFA compounds were also selected in the

models. While VFA and LCFA were better predicted with reflectance spectra, NH_4^+ was better predicted with absorbance spectra. This was also the case in SO-CovSel models with NH_4^+ predicted with R^2 of 0.60 and RMSEP error of 420 mg/l. There were also similar wavelengths as with reflectance spectra (450, 1446, 1372, 1800 nm). Some of these selected wavelengths corresponded to acetone absorbance (475, 1254) and amino group bands (528, 828, 1557nm) (Figure 4).



(a) **(b)**

Figure 4: Selected variables by SO-CovSel the X-blocks in absorbance (a) and reflectance (b). The red line represents the average X_4 spectrum. Blue squares: variables selected in the models.

5. CONCLUSION

This study demonstrates the potential of the combination of several analytical methods for the monitoring of anaerobic digestion. The application of SO-PLS methods improved prediction of the process state indicator parameters such as VFA, LCFA and ammonium. These on-line routine analyses allow the gain of substantial information in the prediction of these parameters. Moreover SO-CovSel provided a pertinent and interpretable variable selection however leading to slightly worse predictions than SO-PLS. Nevertheless, it enables correlated and important conclusions in the infrared data blocks used and the relevant wavelengths of the used probes.

References

[1] Hagos, K., Zong, J., Li, D., Liu, C., & Lu, X. 2017. Anaerobic co-digestion process for biogas production: Progress, challenges and perspectives. *Renewable and Sustainable Energy Reviews*, 76, 1485-1496. https://doi.org/10.1016/j.rser.2016.11.184

[2] Boe, K., Batstone, D. J., Steyer, J.-P., & Angelidaki, I. 2010. State indicators for monitoring the anaerobic digestion process. *Water Research, 44*(20), 5973-5980. https://doi.org/10.1016/j.watres.2010.07.043

[3] Li, D., Chen, L., Liu, X., Mei, Z., Ren, H., Cao, Q. Yan, Z., 2017. Instability mechanisms and early warning indicators for mesophilic anaerobic digestion of vegetable waste. *Bioresour. Technol*., 245 (Part A), 90-97.https://doi.org/10.1016/j.biortech.2017.07.098

[4] Wu, D., Li, L., Zhao, X., Peng, Y., Yang, P., Peng, X. 2019. Anaerobic digestion: A review on process monitoring. *Renewable and Sustainable Energy Reviews,* 103, 1-12. https://doi.org/10.1016/j.rser.2018.12.039

[5] Björnsson, L., Murto, M., Jantsch, T. G., & Mattiasson, B. 2001. Evaluation of new methods for the monitoring of alkalinity, dissolved hydrogen and the microbial community in anaerobic digestion. *Water Research*, *35*(12), 2833-2840. https://doi.org/10.1016/S0043-1354(00)00585-6

[6] Madsen, M., Holm-Nielsen, J. B., & Esbensen, K. H. 2011. Monitoring of anaerobic digestion processes: A review perspective. *Renewable and Sustainable Energy Reviews, 15*(6), 3141-3155. https://doi.org/10.1016/j.rser.2011.04.026

[7] Vojinovic, V., Cabral, J. M. S. & Fonseca, L. P. 2006. Real-time bioprocess monitoring Part I: in situ sensors. *Sens Actuator B-Chem* 114:1083–1091. https://doi.org/10.1016/j.snb.2005.07.059

[8] Jacobi, H.F., Moschner, C.R. & Hartung, E. 2009. Use of near infrared spectroscopy in monitoring of volatile fatty acids in anaerobic digestion. *Water Sci. Technol*, 60(2): 339-346. doi: https://doi.org/10.2166/wst.2009.345

[9] Krapf, L. C., Nast, D., Gronauer, A., Schmidhalter, U. & Heuwinkel, H. 2013. Transfer of a near infrared spectroscopy laboratory application to an online process analyser for in situ monitoring of anaerobic digestion. *Bioresour. Technol, 129*, 39-50. https://doi.org/10.1016/j.biortech.2012.11.027

[10] Lomborg, C.J., Holm-Nielsen, J.B., Oleskowicz-Popiel, P. & Esbensen, K.H. 2009. Near infrared and acoustic chemometrics monitoring of volatile fatty acids and dry matter during co-digestion of manure and maize silage. *Bioresour Technol*, 100(5), 1711-1719. https://doi.org/10.1016/j.biortech.2008.09.043

[11] Stockl, A., & Oechsner, H. 2012. Near-infrared spectroscopic online monitoring of process stability in biogas plants. *Engineering in Life Sciences, 12*(3), 295-305. https://doi.org/10.1002/elsc.201100065

[12] Krapf, L. C., Gronauer, A., Schmidhalter, U. & Heuwinkel, H. 2011. Near Infrared Spectroscopy Calibrations for the Estimation of Process Parameters of Anaerobic Digestion of Energy Crops and Livestock Residues. *Journal of Near Infrared Spectroscopy, 19*(6), 479-493. https://doi.org/10.1255/jnirs.960.

[13] Finzi, A., Oberti, R., Negri, A.S., Perazzolo, F., Cocolo, G., Tambone, F., Cabassi, G. & Provolo, G. 2015. Effects of measurement technique and sample preparation on NIR spectroscopy analysis of livestock slurry and digestates. *Biosystems Engineering* 134, 42-54 https://doi.org/10.1016/j.biosystemseng.2015.03.015

[14] Saeys, W., Darius, P. & Ramon, H. 2004. Rapid on site analysis of hog manure using a visual and near-infrared diode array reflectance spectrometer. *Journal of Near Infrared Spectroscopy*, 12 (5), 299-310. https://doi.org/10.1255/jnirs.438

[15] Reed, J.P., Devlin, D., Esteves, S.R.R., Dinsdale, R. & Guwy, A.J. 2011. Performance parameter prediction for sewage sludge digesters using reflectance FT-NIR spectroscopy. *Water Research*, 45(8), 2463-2472. https://doi.org/10.1016/j.watres.2011.01.027

[16] Xie, L., Xingqian, Y., Liu, D. & Ying, Y. 2009. Quantification of glucose, fructose and sucrose in bayberry juice by NIR and PLS. *Food Chemistry*, 114, 1135-1140. https://doi.org/10.1016/j.foodchem.2008.10.076

[17] Surowiec, I., Skotare, T., Sjögren, R., Gouveia-Figueira, S., Orikiiriza, J., Bergström, S. & Trygg, J. 2019. Joint and unique multiblock analysis of biological data – multiomics malaria study. Faraday Discussions, 218(0), 268-283. https://doi.ord/10.1039/C8FD00243F

[18] Li, D., Yang, H. Z., & Liang, X. F. 2013. Prediction analysis of a wastewater treatment system using a Bayesian network. *Environmental Modelling & Software, 40*, 140-150. https://doi.org/10.1016/j.envsoft.2012.08.011

[19] Steyer, J.-P., Rolland, D., Bouvier, J.-C., & Moletta, R. 1997. Hybrid fuzzy neural network for diagnosis - application to the anaerobic treatment of wine distillery wastewater in a fluidized bed reactor. *Water Science and Technology, 36*(6), 209-217. https://doi.org/10.1016/S0273-1223(97)00525-8

[20] Lee, D. S. & Vanrolleghem, P. A. 2003. Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. *Biotechnology and Bioengineering, 82*(4), 489-497. https://doi.org/10.1002/bit.10589

[21] Hong J.J. & Zhang J. 2010. Quality Prediction for a Fed-Batch Fermentation Process Using Multi-Block PLS. In: Lee J.H., Lee H., Kim JS. (eds) EKC 2009 Proceedings of the EU-Korea Conference on Science and Technology. *Springer Proceedings* in Physics, vol 135. https://doi.org/10.1007/978-3-642-13624-5_15

[22] Wold, S., Kettaneh, N. & Tjessem, K. 1996. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics, 10*(5-6), 463-482. https://doi.org/10.1002/(SICI)1099-128X(199609)10:5/6<463::AID-CEM445>3.0.CO;2-L

[23] Westerhuis, J. A., Kourti, T. & MacGregor, J.F. 1998. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics: A Journal of the Chemometrics Society* 12.5, 301-321. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S

[24] Næs, T., Tomic, O., Mevik, B.H. & Martens, H. 2011. Path modelling by sequential PLS regression. *Journal of Chemometrics*, *25*(1), 28-40. https://doi.org/10.1002/cem.1357

[25] Biancolillo, A., Næs, T. Bro, R. Måge, I. Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemometr. Intell. Lab. Syst. 164 (2017) 113–126. https://doi.org/10.1002/cem.3120 https://doi.org/10.1016/j.chemolab.2017.03.002

[26] Måge, I., Menichelli, E. & Næs, T. 2012. Preference mapping by PO-PLS: Separating common and unique information in several data blocks, *Food Qual. Pref.* 24, 8–16. https://doi.org/10.1016/j.foodqual.2011.08.003.

[27] El Ghaziri, A., Cariou, V., Rutledge, D.N. & Qannari, E.M. 2016. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K+1) datasets. *Journal of Chemometrics*, *30*(8), 420-429. https://doi.org/10.1002/cem.2810

[28] Spanjers, H. & van Lier, J.B. 2006 Instrumentation in anaerobic treatment – research and practice. *Water Sci Technol.* 53(4-5), 63–76. https://doi.org/10.2166/wst.2006.111

[29] Mehmood, T., Liland, K. H., Snipen, L. & Sæbø, S. 2012. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62-69. https://doi.org/10.1016/j.chemolab.2012.07.010

[30] Gauchi, J. P. & Chagnon, P. 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 171-193. https://doi.org/10.1016/S0169-7439(01)00158-7

[31] Roger, J. M., Palagos, B., Bertrand, D. & Fernandez-Ahumada, E. 2011. CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *106*(2), 216-223. https://doi.org/10.1016/j.chemolab.2010.10.003

[32] Biancolillo, A., Liland, K. H., Måge, I., Næs, T. & Bro, R. 2016. Variable selection in multi-block regression. *Chemometrics and Intelligent Laboratory Systems*, *156*, 89-101. https://doi.org/10.1016/j.chemolab.2016.05.016

[33] Biancolillo, A., Marini, F. & Roger, J. M. 2019. SO-CovSel: A novel method for variable selection in a multiblock framework. *Journal of Chemometrics*, e3120. https://doi.org/10.1002/cem.3120

[34] APHA, 2012. Standard Methods for the Examination of Water and Wastewater (22nded.), American Public Health Association, American Water Works Association, Water Environment Federation.

[35] Walker, M., Zhang, Y., Heaven, S., Banks, C.J. 2009. "Potential errors in the quantitative evaluation of biogas production in anaerobic digestion processes". *Bioresour. Technol.,* 100(24), 6339-6346. http://dx.doi.org/10.1016/j.biortech.2009.07.018

[36] Bendoula, R., Gobrecht, A., Moulin, B., Roger, J.M., Bellon-Maurel, V. 2015. Improvement of the chemical content prediction of a model powder system by reducing multiple scattering using polarized light spectroscopy. *Applied Spectroscopy,* 69(1), 95-102 https://doi.org/10.1366/14-07539

[37] Awhangbo, L., Bendoula, R., Roger, J.M., Béline, F. 2020. Multi-block SO-PLS approach based on infrared spectroscopy for anaerobic digestion process monitoring. *Chemometrics and*

Intelligent Laboratory Systems, 196, 103905. https://doi.org/10.1016/j.chemolab.2019.103905

[38] Savitzky, A. & Golay, M. J. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8), 1627–1639. https://doi.org/10.1021/ac60214a047

[39] Zeaiter, M., Roger, J.M. & Bellon-Maurel, V., 2005. Robustness of models developed by multivariate calibration. Part II: the influence of pre-processing methods. *TrAC,Trends Anal. Chem*. 24, 437–445. https://doi.org/10.1016/j.trac.2004.11.023

[40] Barnes, R., Dhanoa, M. & Lister, J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc*. 43, 772–777. https://doi.org/10.1366/0003702894202201

[41] Biancolillo, A. Naes, T. Chapter 6 - The sequential and orthogonalised PLS regression (SO-PLS) for multi-block regression: theory, examples and extensions, in: M. Cocchi (Ed.), Data Handling in Science and Technology, Vol 31, Elsevier, Amsterdam, (2019), pp. 157-177. https://doi.org/10.1016/B978-0-444-63984-4.00006-5

[42] Palatsi, J., Affes, R., Fernandez, B., Pereira, M.A., Alves, M.M. & Flotats, X. 2012. Influence of adsorption and anaerobic granular sludge characteristics on long chain fatty acids inhibition process. *Water Res*, 46(16), 5268-5278. https://doi.org/10.1016/j.watres.2012.07.008