Multivariate Analysis of Industrial Biorefinery

Processes: Strategy for Improved Process

Understanding with Case Studies in Fatty Acid

Production

Pieter Nachtergaele^{1,2}, Joris Thybaut³, Steven De Meester⁴, David Drijvers², Wouter Saey⁵,
Jo Dewulf¹*

¹ Ghent University, Research Group STEN, Coupure Links 653, B-9000 Ghent, Belgium

² Olean NV, Assenedestraat 2, B-9940 Evergem, Belgium

³ Ghent University, Laboratory for Chemical Technology, Technologiepark 125, B-9052
Ghent, Belgium

⁴ Ghent University, Department of Green Chemistry and Technology, Graaf Karel de
Goedelaan 5, B-8500 Kortrijk, Belgium

⁵ KU Leuven, Department of Biosystems, Kasteelpark Arenberg 30, B-3001, Leuven,
Belgium

* Corresponding author. Tel.: +32 92 64 59 18. E-mail: Pieter.Nachtergaele@ugent.be

ABSTRACT: A major difficulty in operating biorefinery processes is the large feedstock variability. A systematic multivariate analysis (sMVA) strategy for improved process

understanding of industrial biorefinery processes is proposed to support identification of effects of feedstock and process variability on product quality. This strategy comprises nine steps categorized in dataset organization, exploratory analysis and regression. Different MVA techniques are used, such as principal component analysis (PCA) and Partial Least Squares Regression (PLS). As a case study, two main operations in fatty acid production are investigated: oil hydrolysis and fatty acid distillation. Key feedstock properties and process parameters affecting the product properties were identified for both operations. For fatty acid production, product quality largely depends on the type of fat or oil used, such as canola or palm oil, due to the large difference in composition and quality between the oil types. However, if a single oil type is used, the variability in product quality does not always critically depend on the variability in feedstock properties. For both operations, flow rate variations, mainly caused by planning issues, were identified as a main cause. In oil hydrolysis, the feed flow rate influences the residence time and thereby directly influences the hydrolysis and side reactions. In fatty acid distillation, a better control is required of the middle reflux ratio in relation to this changing flow rate. The case study showed that applying our proposed sMVA strategy improves the understanding of a biorefinery process by identifying critical sources of variability, which allows more targeted decisions for optimization and control.

1. Introduction

Biorefineries provide an important alternative for petroleum-based processes to improve the sustainability of chemicals, plastics and fuels.¹ A key bottleneck in the development of biorefineries is the pronounced feedstock variability.^{2, 3} Operating a continuous process with a variable feed composition and quality poses significant challenges for process control to ensure resource efficiency, high product yields and excellent product quality.⁴ Petroleum refineries already use the power of mechanistic and statistical modelling to tackle this challenge of feedstock diversity and adapt process conditions for compositional variability of the incoming crude oil.⁵ Applying these techniques to biorefineries would aid in utilizing variable feedstock streams more efficiently, enabling the use of lower quality-grade feedstocks and growing the biofuel/biochemical sectors to the billion-ton scale.⁶ The complexity of biomass and biorefinery processes, however, makes mechanistic modelling highly challenging, as it would be nearly impossible to take the effect of all feedstock components and process variables into account. Alternatively, statistical process modelling can be used to identify the critical sources of feedstock and process variability that have an effect on product quality and to gain insight in how control strategies can be used to manage this variability. Thereby, statistical modelling could serve as an important basis to steer further research and modelling efforts for the optimization, monitoring and control of biorefinery processes. A lack of officially endorsed practices and methods, however, is often seen as one of the main disadvantages of this statistical modelling.⁷ This motivates the development of general strategies for applying statistical modelling in industrial biorefinery processes.

There are different economic and environmental drivers to use oils and fats as a renewable feedstock in biorefineries.⁸⁻¹⁰ The oleochemical industry uses fats and oils as feedstock for the

production of renewables-based fatty acids (FA), glycerol, biofuels and derivatives. Oleochemical feedstocks range from low quality animal fat to higher quality canola and other vegetable oils. Fatty acid compositions, minor components and quality vary greatly among these oils and fats.¹¹ Massive amounts of data are being collected from industrial biorefinery processes via numerous sensors, monitoring sheets filled out by operating personnel and routine quality control checks of feedstocks and products.¹² The range in which feedstock and operating variables vary during normal production may be limited and correlations with product quality are often difficult to interpret correctly due to existing control schemes.¹³ In comparison to an ideal lab setting, however, industrial processes are not always adequately controlled and flow rates often fluctuate due to planning. These process variations may cause significant variations in final product quality in addition to those caused by feedstock variations.¹⁴ Statistical modelling of industrial data can be a very interesting tool for gaining process understanding if it allows determining these feedstock and process variations that have a significant effect on final product quality during normal production. Hydrolysis, a chemical reaction, and distillation, a physical separation, are the two key operations in the production of FA. However, the effect of feedstock quality and operating parameters on the stability of the final FA has not been studied comprehensively.

Multivariate analysis (MVA) tools have been applied successfully in a vast amount of applications to investigate industrial datasets. The basic principle behind MVA techniques is a dimensionality reduction of the data to simplify the visualization and interpretation. Industrial datasets typically involve a large number of variables, which may be highly correlated. The effective dimensionality of the dataset is typically considerably smaller than the number of variables, as they actually only represent a small number of underlying events.¹³ MVA techniques try to capture the variability in the high-dimensional dataset and project it onto a lower number of

linear combinations, called principal components (PC) in the case of Principal Component Analysis (PCA) or latent variables (LV) in the case of partial least squares (PLS) regression.⁷ Non-linear techniques, e.g. artificial neural networks, may offer a better predictive ability but are typically harder to interpret and therefore less suited if the goal is gaining process understanding.¹⁵ ¹⁶ Many reviews have been performed on MVA techniques and their applications in industry.^{7, 17} Lourenco et al. (2012) reviewed the use of MVA techniques specifically for bioreactor monitoring, while Tomba et al. (2013) describe a general procedure for applying these techniques to support the development of continuous pharmaceutical processes on an industrial scale.^{15, 18}

In this article, a systematic multivariate analysis (sMVA) strategy is proposed to guide process engineers and data scientists in this multivariate modelling of continuous industrial scale biorefinery processes. The goal of following this MVA strategy is to improve process understanding, so more targeted decisions for process optimisation and control can be made. The strategy attempts to maximize the knowledge that can be gained from industrial datasets, while simultaneously taking into account its limits. This work does not propose a new multivariate method for the analysis of datasets, but rather provides a guideline for applying existing techniques in the complex setting of renewables-based processes considering the available data resources and the limitations faced at industrial scale. As a case study, the developed strategy for MVA of industrial biorefinery processes is applied to the operations fat/oil hydrolysis and FA distillation. For both operations, the critical sources of variability in product quality are identified and suggestions for further process optimization and control are made.

2. Development of a systematic MVA strategy (sMVA) for improved process understanding of industrial biorefinery processes

A 9-step systematic MVA strategy (sMVA) is proposed for applying MVA to gain process understanding of industrial biorefinery processes, which is schematically illustrated in **Figure 1**. The goal of the sMVA strategy is to identify the critical sources of feedstock and process variability that have an effect on product quality during normal production. The sMVA strategy covers dataset organization, exploratory analysis and regression, with each divided into three steps. The sMVA strategy is developed for continuous processes only, as batch operations require taking into account their nonlinear dynamic nature. In section 1 of the supporting information, a step-by-step guide can be found explaining these steps in detail. The attached step-by-step guide includes recommendations on how to perform each step, an explanation of the methodologies used and applications to the case study.

DATASET ORGANIZATION	
Build a structured and clean dataset that can easily be uploaded and analysed in statistical software.
STEP 1: COLLECT DATA Collect data from different sources, ensuring a good representation of the variability in the system.
STEP 2: STRUCTURE DATA Integrate data into one repository and structure data into blocks.
STEP 3: CLEAN DATA Optimize dataset by looking at missing values, correcting mistakes, removing duplicate data,...
EXPLORATORY ANALYSIS	
Explore and prepare dataset for further analysis.
STEP 4: DETERMINE GROUPING Detect grouping of samples and split the dataset into groups when grouping can be explained.
STEP 5: DETECT OUTLIERS Detect outlying samples and remove if linked to experimental error.
STEP 6: CONTROL BACKGROUND Control the effect of variables that cannot be controlled.
REGRESSION ANALYSIS	
Estimate the relationships between product variables and process/feedstock variables.
STEP 7: PERFORM REGRESSION Choose regression technique and perform regression analysis.
STEP 8: SELECT VARIABLES Select variables with highest importance in the model.
STEP 9: INTERPRET RESULTS Make reliable conclusions based on a comprehensive interpretation of all results.

Figure 1. Systematic MVA Strategy for improved process understanding of industrial biorefinery processes. The goal of applying this 9-step strategy is to identify the critical sources of variability that have an effect on product quality and finding strategies that can be used to manage this variability.

3. Materials and Methods

3.1 Description of the Case Study

Animal fat (tallow) or vegetable oil hydrolysis and fatty acid (FA) distillation are the two main operations in the production of fatty acids at the oleochemical plant of Oleon NV in Ertvelde, Belgium. The first step in fatty acid production is fat/oil hydrolysis. In this step, a reaction takes place between water and triglycerides to form glycerol and fatty acids. The process is performed in a counter-current splitting tower, at temperatures up to 260°C and pressures up to 60 bar. There are three different hydrolysis units (Tower 1, 2, 3), and the capacity per installation varies between 5-15 ton FA/hour. The second step is refining the crude FA by distillation, forming a bottom, middle (distilled) and top FA fraction. The process takes place at temperatures up to 250°C and under vacuum. There are two distillation units (Distillation 1, 2), which are identical by design and each have a capacity of 12-17 ton crude FA/hour. A simplified flowsheet of the two operations is given in **Figure 2**, together with a list of measured variables. There are five types of feedstocks used: unrefined tallow (Tallow), refined tallow (Tallow+), palm oil (Palm), palm fatty acid distillate (PFAD) and rapeseed/canola oil (Canola).

Figure 2. Simplified flowsheet of the two operations in fatty acid production with a list of the measured variables a. Fat/oil Hydrolysis and b. FA Distillation.

The properties that define the quality of fatty acids depend on the intended application. For fatty acids used in technical applications, the acid value and iodine value are most important, while for cosmetics also colour and heat stability should be taken into account. A list of the quality measurements used in this study can be found in **Table S3** in the supporting information, together with a short explanation of the parameter. Mostly quality parameters, such as p-anisidine value, are used both to describe oleochemical feedstocks and products. For oil hydrolysis, important considerations are a high degree of hydrolysis, a high concentration of glycerol (%Glycerol) and a minimal colour degradation of the fatty acids. For fatty acid distillation, low colour values and a

good stability (storage & heat) of the distilled fatty acids (middle fraction) is desired. For both processes, many different process variables are measured. In oil hydrolysis, different flows, temperatures and pressures are taken into account. The water to oil ratio is the ratio of the combined input flows of water and steam on the flow of fat/oil. For FA distillation, also different flows, temperatures and pressures are measured. Middle reflux ratio is the ratio of the recycle stream at the middle of the column on the flow of distilled fatty acids. %Top and %bottom fatty acids is the ratio of the flows of respectively top and bottom fatty acids on the flow of crude fatty acids.

3.2 Application of the sMVA strategy to Fatty Acid Production

In this section, the methodologies used for every step of the strategy are given as applied for the fatty acid case study. All MVA analyses in this study were performed using the Eigenvector PLS_Toolbox 8.6.2 for Matlab® (R2018a).

3.2.1 Dataset Organization

Step 1: Collect Data

MVA is a data-driven analysis and the data used needs to properly represent the variability in the system.⁷ For process parameters, the representability of the dataset was controlled by comparing the distributions of process indicators in the dataset with the distribution based on the total historical dataset. For feedstock parameters, the representability was ensured by collecting a dataset spanning a whole year and distributed over the different feedstock types. The data of both processes was collected for a 12-month period in 2017-2018. The actual datasets are of a confidential nature and are therefore not included in this article.

Step 2 Structure Data

Typically, a continuous industrial biorefinery process can be seen as a series of operations that convert a feedstock into products. Therefore, the variables in the dataset were divided into the

following blocks: feedstock properties (Z), process parameters (X) and product properties (Y).¹³

In this work, two other types of variables are also considered: Class information (C) and Background Variables (B). Class information variables, such as feedstock type, are discrete and divide the total dataset into certain groups (e.g. by feedstock type used). Background variables cannot be manipulated, but may still affect the variables in Z, X and Y (e.g. day of sampling). The number of campaigns (n) and variables (c, z, x, y, b) per block are given in **Table 1**.

Table 1. Number of observations (campaigns) and variables for each process.

	Fat/Oil Hydrolysis	Fatty Acid Distillation
# Campaigns (<i>n</i>)	102	158
# Class Variables (<i>c</i>)	2	2
# Feedstock Variables (<i>z</i>)	33	35
# Process Variables (<i>x</i>)	14	25
# Product Variables (<i>y</i>)	13	20
# Background Variables (<i>b</i>)	4	4

Step 3: Clean Data

The dataset was cleaned from clear errors a priori by checking for spelling mistakes, duplicate data or other clear errors that can be identified and corrected before starting the analysis. A cell-wise algorithms, Detecting Deviating Cells (DDC), was used to make this inspection more efficient.¹⁹

3.2.2 Exploratory Analysis

Step 4: Determine Grouping

Principal Component Analysis (PCA) is a useful tool to determine grouping in multivariate data.²⁰ The principal components (PCs) are sequentially extracted to maximize the amount of

variance in the total dataset captured by each component. The score of a campaign for a PC is the projection of the original data of the campaign for that PC. Plotting the scores of campaigns in a scatter plot can therefore be used to inspect the relation between them. If the relative positions of campaigns in the scatter plot are close, it indicates similarity.⁷ If clear grouping is observed, it may be recommended to build separate regression models for the different groups instead of combining the data into a single analysis, because the assumptions at the basis of MVA methods such as multivariate normality may not be fulfilled for the entire dataset.²¹

For both hydrolysis and FA distillation, a PCA analysis was performed on the blocks feedstock properties (Z), process parameters (X) and product properties (Y) separately to inspect for grouping in the data and to better understand the relations between variables in one block. Because PCA and PLS models are not scale invariant, the data was mean-centred and all variables were scaled to unit variance before analysis.²² Class variables (C) were used to help explain the grouping observed in the data. The grouping of campaigns was inspected by analysing the scores of campaigns and colouring the campaigns by the different class variables, e.g. feedstock type. If grouping was observed and could be explained by a class variable, the campaigns were split into groups accordingly. .

The loading of an original variable for a principal component measures how much a variable contributes to that PC. A very positive or negative loading means the variable has a strong contribution to the scores for that PC. A loadings plot for the different PCs was used to investigate how the original variables correlate to one another. In a biplot, the scores and loadings are superimposed in one figure. Biplots were used to allow for a better understanding of which variables determine the grouping found in the scores plot.⁷

Step 5: Detect Outliers

Outliers are campaigns that are extreme in the model space compared to the other campaigns (high Hotelling T^2) or are badly explained by the model (high Q residuals).⁷ Outlying campaigns can have a great influence on the model, and if this outlying behaviour is due to experimental error, the corresponding campaigns were removed before regression analysis. To detect outliers, a PCA on the Z, X and Y blocks was performed separately for every group. The Q-residuals & Hotelling T^2 of all campaigns were plotted together and used for identifying outlying samples. After identifying outliers in the separate blocks, a multiblock-PCA (MB-PCA) was performed on the model combining the Z, X and Y blocks of every group. MB-PCA allows the analysis of multiple blocks of data simultaneously. For MB-PCA, block scaling was performed, in which the scores of variables in a specific block are divided by the square root of the number of variables in that block. By doing so, each block starts with the same variance irrespective of the number of variables in each block.²⁰ The MB-PCA models were used to detect outlying campaigns for the combined Z, X and Y dataset.

Step 6: Control Background

The PCA and MB-PCA models from step 5 were also used to control the importance of background variables. This was done by observing the PC scores of campaigns (y-axis) combined with the values of a background variable (x-axis). If background variables clearly influenced the scores of campaigns, a further investigation of the cause of this trend was performed before regression. Step 5 and step 6 were performed in unison, as an outlier may also be explained by the effect of a background variable. For example, if the scores show a clear trend following the background variable 'day of sampling', this may be due to the seasonal variability of feedstocks, which is an aspect that should be captured by the model and therefore does not inhibit performing a regression

analysis. However, if the trend is due to an experimental error being introduced, enhanced or removed over time, it might cause serious misinterpretation of the model results.

3.2.3 Regression Analysis

Step 7: Perform Regression

In this step, the correlations between the independent variables feedstock properties (Z) and process parameters (X), and the dependent variables product properties (Y) were sought. PLS focusses on the covariance between the independent variables and the dependent variables and is the most commonly used MVA technique for regression.²³ Different adaptations of the original PLS technique exist to improve the predictive ability or to simplify interpretation, such as multi-block PLS (MB-PLS).²⁴ MB-PLS splits the independent variables into a number of meaningful blocks, in this case feedstock properties (Z) and process parameters (X). MB-PLS is not a PLS between each block and the product properties (Y) separately, but one PLS where the blocks are weighted in such a way that their combination is most predictive of Y.²⁰ As the standard MB-PLS method combines the information in the original Z and X variables in the same latent variables (LVs), it explores the correlation between these variables. However, this may complicate the interpretation of the effects of the individual blocks. Two separate PLS models, one between Z and Y and one between X and Y, solve this problem, but may lead to a difficult and possibly false interpretation of the results when the Z and X blocks are (highly) correlated. Variation in the feedstock properties (Z) may introduce variation in the process parameters (X), either directly via physical responses or indirectly due to feedback and feedforward control loops. A new algorithm, sequential multi-block PLS (SMB-PLS), was recently proposed that deals with these issues. The correlated information in Z and X is combined into a common latent variable (LV), whereas the

orthogonal information is captured by a different LV. This allows for an easier interpretation of the variance in Z and X that is related with the variance in Y.²⁵

For every group found in step 5, a separate PLS model was built with Z and with X as independent variables and Y as dependent variables. For both oil hydrolysis and fatty acid distillation, no direct correlation was expected or found between X and Z within one group. The variability in fatty acid composition is low within one feedstock type and there are currently no feedback or feedforward controls for product quality. Therefore, PLS and MB-PLS regression analysis were used for the different groups identified during the exploratory analysis. In addition, a regression analysis of the total dataset (all groups combined) of every process was performed to quantify the variability explained by the differences between groups, and to investigate if the correlations found on group level hold for the total dataset. For the total dataset, correlations between X and Z were expected and found because slightly different process parameters were used between feedstocks and the differences in composition may have influenced certain process parameters. Therefore, for the analysis of the total dataset, SMB-PLS was used as this method differentiates between the variability in X which is correlated with the variability in Z, and the variability in X which is orthogonal to the variability in Z.

The number of valuable LVs in a PLS model depends on the complexity of the relation between the dependent and independent variables and the signal to noise ratio. As industrial data was used, it can be expected that this signal to noise ratio is low.¹³ If too many LVs are used in the model, noise or irrelevant sources of variation may be included in the model, leading to overfitting. Cross-validation (CV, venetian blinds) was used to detect this overfitting. CV is a model validation technique which gives an indication how well the model would perform on new data by evaluating the model performance for different calibration-validation splits.⁷ Both an absolute value, the root

mean square error of cross-validation (RMSECV), and the cross-validated coefficient of determination (R^2_{CV}) were used for selecting the model complexity (the number of LV's used in the model).²⁶ A high RMSECV and low R^2_{CV} indicate that the model does not explain the variability in product quality well. The models for which the explained variability is high were further analysed in step 8. Whether the explained variability can be considered high or not depends on the system and the modelled Y variable.

Step 8: Select Variables

The selected models contain all independent variables, even though some variables may not actually explain any of the variability in the product properties (Y). The feedstock properties (Z) and process parameters (X) that are most important for modelling the variability in Y were determined by variable selection. The Z and X parameters which are the most important for explaining the variability in the models were selected based on their Variable Importance in Projection (VIP) scores. The VIP score of a variable indicates the importance of an independent variable in the projections used in a PLS model. Variables were selected if they have a VIP score close to or greater than 1.²⁶ Other variable selection methodologies have been developed and could be used. For example, the 'lasso' technique minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant.²⁷ Alternatively, stepwise regression can be used, in which independent variables are sequentially introduced into the model one at a time. Chong and Jun (2005) compared lasso, stepwise and PLS-VIP and found that PLS-VIP outperformed the other methods when there is collinearity between independent variables.²⁸

Step 9: Interpret Results

Finally, the PLS model was rebuilt using only the selected variables, and interpreted by analysing different plots such as a scatter plot where the predicted values are plotted as a function

of the measured values, the biplot and the regression vector. In addition, the results of previous steps, such as the effect of background variables, were taken into account when doing the interpretation.

4. Case Study Results

4.1 Dataset Organization

The results of steps 1-3 of the strategy are structured datasets of the fat/oil hydrolysis and FA distillation that properly represent the variability in the system and that can easily be uploaded into statistical software. Feedstock properties and process parameters vary between different campaigns. The variability of feedstock properties within one campaign is limited due to good mixing of the feedstock tanks. The variability in feedstock properties between and within feedstock types is given in **Table S4** (Fat/Oil Hydrolysis) and **Table S5** (FA Distillation) of the supporting information. For iodine value and acid value of oils and fats, the variability is large between feedstocks, while for colour and oxidation related parameters the variability is large both between and within an oil type. The variability of process parameters is given in **Table S7** (Fat/Oil Hydrolysis) and **Table S8** (FA Distillation) of the supporting information. The variability of the feed rates is high, both between and within a feedstock type, for both processes. For temperature variables, the variability is much lower. However, the absolute standard deviation (σ) of Temp Tower Bottom for tallow campaigns in FA Distillation is still 8°C, a difference that could significantly affect the relative volatility of components in the distillation tower. The pressure of the distillation tower also varies significantly between campaigns, while the pressure of the High Pressure (HP) steam used is relatively more stable. Finally, for FA distillation, middle reflux ratio shows the highest relative variability both between and within feedstocks. The variability of a selection of product properties of the middle fraction of the FA distillation is given in **Table S9**

(Fat/Oil Hydrolysis) and **Table S10** (FA Distillation) of the supporting information. There is a large difference in fatty acid quality between fatty acids produced from different oil types. For colour, distilled fatty acids from unrefined tallow and palm fatty acids are respectively the most and the least coloured. For all product properties, the variability is also large within one feedstock type.

4.2 Exploratory Analysis

The first 3 principal components (PCs) were retained in the model for the PCA analysis of product properties (Y) of the process FA distillation. **Figure 3** shows the PC1/PC2 biplot of this model. A clear grouping of the canola, palm, tallow and tallow+ samples can be observed. In this PC1/PC2 scores plot, the grouping is less clear for PFAD campaigns, due to their very variable PC1 scores. Grouping was also found for the class variable installation, even though the distillation towers are identical by design. However, this grouping was only observed in the process parameters (X) and not in the product properties (Y). This creates five groups for FA distillation: Tallow on Distillation 1 (D1), Tallow+ on D1, Palm on D2, Canola on D2 and PFAD on D2. For fat/oil hydrolysis, grouping was also observed for feedstock type and installation, creating five groups: Tallow on Tower 2 (T2), Tallow+ on T3, Palm on T3, Canola on T1 and PFAD on T2.

Figure 3. PCA biplot visualizing the campaigns and variables in the PC1-PC2 space for the product properties (Y) of the fatty acid distillation to detect grouping in the data (step 4).

The biplot in **Figure 3** also shows the multivariate nature of the product properties. There are two clear clusters of Y variables. In the bottom right corner, colour, heat stability and storage stability of the distillate fraction and the colour of the top fraction are clustered, indicating that they are correlated. As tallow campaigns are found in the same corner, tallow fatty acids likely have high values for these product properties. A second cluster of Y variables is found in the top right corner, clustering colour, heat stability and storage stability of the top fraction and the soap colour and p-anisidine value of the distillate fraction. This split between product properties was unexpected, as the properties of the top and distillate fractions were not clustered separately, but mixed. A correlation between the p-anisidine value and the soap colour of the distillate may indicate that secondary oxidation products play a role in colour formation during the saponification reaction, while this is not directly observed with initial colour, heat stability and storage stability.

In **Figure 4a**, Q-residuals and Hotelling T^2 of the campaigns of Tallow on D1 (MB-PCA of the combined [Z X Y] dataset) are given to detect outliers (step 5). For this group, three campaigns were identified as candidate outliers and inspected in more detail (D002, D042, and D113). As experimental errors were identified as the cause for the outlying behaviour of campaigns D002 and D042, these were removed. D113 however showed no indications of experimental error, but rather extreme conditions for the process parameter Flow Cooling Water. D113 is therefore not removed, as it may contain valuable information regarding this parameter. The effect of the background variable ‘date of sampling’ on the PC1 score for the group Tallow on D1 is shown in **Figure 4b** (step 6). There appears to be a trend. However, as there is no data available for campaigns between April and July, it is difficult to draw conclusions with respect to this.

Figure 4. MB-PCA results for Tallow on D1 a. Q-residuals vs Hotelling T^2 plot for detecting outliers (step 5). b. PC1 scores as a function of the effect of background variable ‘Date of Sampling’ (step 6).

The biplot of this MB-PCA analysis can also be investigated to obtain an initial overview of relations of variables between blocks. As an example, the biplot for the group Tallow on T2 is shown in **Figure 5**. Storage stability and other product properties linked to colour are found in the centre of the biplot, indicating their variability is not well explained by the model. Heat stability is found far from the centre, and close to flow parameters. Surprisingly, peroxide value is found in the opposite quadrant of heat stability, possibly indicating a negative correlation between the two product properties. Degree of hydrolysis and %C18:3 breakdown are found close together and in the proximity of process parameters linked to tower temperature and opposite of the flow parameters and water/oil ratio.

Figure 5. Biplot of the MB-PCA analysis of the combined ZXY dataset for the group tallow on T2 of the hydrolysis process.

4.3 Regression Analysis

4.3.1 Regression within group

Fat/Oil Hydrolysis

For the group Tallow+ on T3, the model outputs (R^2_{CV} and RMSECV) for a selected set of product properties are shown in **Figure 6**. For all parameters, the combined MB-PLS does not achieve a higher R^2_{CV} . In the MB-PLS models, Z and X variables were combined that probably represent different effects on the system, such as p-Anisidine Value (Z variable) and the pressure HP steam (X variable) in LV1 and LV2 of the MB-PLS model with dependent variable colour stability. This made the models more difficult to interpret, but also lowered the robustness of the models. Overall, product parameters linked to fatty acid quality, i.e., colour and oxidation, are better explained by the variability in feedstock properties. The R^2_{CV} for peroxide value is however very low. This was observed for all groups and can be attributed to the fact that oxidation products like peroxides break down at the high temperatures of the hydrolysis tower.

Figure 6. Model outputs for PLS models using only feedstock properties (Z) or process parameters (X) or a MB-PLS using both Z and X to predict selected product properties (x-axis) for the fat/oil hydrolysis process for the group Tallow+ on T3.

Of all parameters, the largest part of the variability is explained for the breakdown of C18:3 fatty acids, with over 80% explained by the variability in process parameters. C18:3, also called

linolenic acid, is unstable at the temperatures of the tower due to its three double bonds. A high breakdown of C18:3 fatty acids should be avoided as the breakdown products are removed during distillation and result in a loss of product. **Figure 7a** shows the plot of the VIP scores for the %C18:3 breakdown of Tallow on T2. Eight variables have a VIP score above 1 and are therefore selected. A new model was built using only these parameters and the observed correlations were interpreted (step 9). As can be seen in the regression vector shown in **Figure 7b**, five of the selected variables negatively contribute to the prediction of the %C18:3 Breakdown, being the feedstock (fat/oil and water) and product (fatty acids and glycerol) flows and the pressure of the tower. Three variables positively contribute to the %C18:3 Breakdown, namely the tower temperatures. The correlation map of these variables, shown in **Figure 7c**, indicates that the flows and pressure are highly correlated and so are the tower temperatures, each probably representing only one major effect on the system. Higher flows correspond to a lower residence time in the tower, thereby reducing the exposure of the fatty acids to the high temperatures of the tower. Secondly, higher temperatures increase the kinetics of the breakdown reactions. To limit the %C18:3 Breakdown, a different process control is required. If the feed flow cannot be fixed due to planning, an option is to slightly lower temperatures for lower flow rates. The higher residence time is in that case compensated by lower temperatures. However, before this strategy is implemented, these hypotheses derived from the observational data should first be verified with a designed experiment.

Figure 7. Results of the PLS model of the % C18:3 Breakdown for Tallow on T2 predicted using only process parameters (X) for variable selection (step 8) and interpretation (step 9). a. VIP scores for the different process parameters for variable selection. b. Regression vector for the selected variables. c. Correlation map for the selected variables (variables regrouped by similarity).

Over 75% of the variability in storage stability of crude tallow+ fatty acids is explained by the variability in feedstock properties. The variability in storage stability of crude fatty acids was linked to the fatty acid composition of the feedstock. A higher concentration of odd-chain (C15/C17/C21) and unsaturated (C16:1/C18:1/C18:3) fatty acids showed a correlation with a worse storage stability. Further investigation indicated that there was a difference in fatty acid

composition between the two suppliers of tallow+, with the tallow of one supplier resulting in fatty acids with a better storage stability.

From an economic perspective, %Glycerol and degree of hydrolysis are important parameters for the hydrolysis process, as they are linked to the yield of fatty acids and glycerol. The variability in these parameters explained by the variability in feedstock properties is below 20%. For %Glycerol, the variability explained by process parameters is 60%, while for degree of hydrolysis it is below 20%. This is probably because, in the past, optimization and control efforts have focused primarily on increasing the degree of hydrolysis, and the process is therefore operated to always reach a maximal degree of hydrolysis. This is done for example by using a large excess of water. The variability in %Glycerol is explained primarily by a difference in water to oil ratio between campaigns. A higher water to oil ratio results in a lower %Glycerol. The water/oil ratio is currently corrected by changing the water input at the top of the tower. The steam inlets, however, are not taken into account, while this stream also influences the actual water/oil ratio after condensation inside the tower. It is therefore recommended to change the control by including the steam flow.

Fatty Acid Distillation

The model outputs (R^2_{CV} and RMSECV) for the PLS models of selected product properties for FA distillation are shown in **Figure 8**. For Palm on D2, the variability in product properties explained by both feedstock properties and process parameters is very low. However, as the RMSECV values are also low, this primarily shows the stable product quality of the distilled palm fatty acids. For Tallow on D1, the variability in product properties explained by process parameters (5-50%) is higher than the variability explained by feedstock properties (2-10%). The same is observed for Tallow+ on T1 and for PFAD on D2. Only for Canola on D2, the variability explained by feedstock properties is slightly higher (3-50%) than by process parameters (2-40%). The

importance of feedstock variability for canola FA can be attributed to the seasonal variability. Canola oil quality is high directly after harvesting and pressing, and gradually deteriorates during storing. The oil quality is therefore best in autumn, just after harvesting, and worse in the summer when the final stock of the previous harvest is processed. In **Table 2**, the selected variables are given for the models with a total explained variability above 50%. The variability in heat stability explained by the variability in process parameters is high for both tallow and tallow+. Variable selection showed in both cases that the key process parameter influencing this property is middle reflux ratio. Currently the middle reflux is controlled with a fixed valve%, causing a fluctuating reflux ratio depending on the feed flow, which is determined by planning. This type of control can clearly be improved, and a designed experiment was set up to investigate the possibility to define a better control scheme. As an example, **Figure 9** shows the results of a designed experiment investigating the effect of middle reflux% on the colour and heat stability of distilled tallow fatty acids. The results show that fatty acid colour and heat stability are indeed improved significantly if a higher reflux ratio is used. Based on the findings of this study the control was changed from a fixed valve% to a higher and fixed middle reflux ratio, leading to a better and more stable product quality.

Tallow on Distillation 1

Tallow+ on Distillation 1

Palm on Distillation 2

PFAD on distillation 2

Canola on Distillation 2

■ Z ■ X

Figure 8. Model outputs for PLS models using only feedstock properties (Z) or process parameters (X) to predict selected product properties (x-axis) of the distilled fatty acids fraction of the fatty acid distillation process.

Table 2. List of product properties, the selected variables and their impact on this property (+ indicates a positive correlation and – a negative correlation) for the FA Distillation process.

Group	Product Property (Y)	Selected Feedstock and Process Variables (Z, X)	
		+	-
Tallow on D1	Heat Stability	Temp Tower Bottom	Middle Reflux Ratio,
			Flow Middle Reflux
Tallow+ on D1	Heat Stability		Middle Reflux Ratio
			Flow Middle Reflux
PFAD on D2	Colour	Temp Tower Bottom	Middle Reflux Ratio
		Temp Tower below Packing	
PFAD on D2	P-Anisidine Value	Temp Tower Bottom	Middle Reflux Ratio
		Temp Tower Middle	
		Flow Crude FA	
Canola on D2	Colour	Crude Soap Colour	
		Crude P-Anisidine Value	
		Crude Heat Stability	
		Crude Storage Stability	
		Crude Acid Value	
Canola on D2	Soap Colour	Crude Soap Colour	
		Crude P-Anisidine Value	
		Crude Heat Stability	
		Crude Storage Stability

Figure 9. Experimental results of a designed experiment, within one campaign of Tallow on D1, showing the effect of the process parameter middle reflux ratio on colour and heat stability of tallow fatty acids in the fatty acid distillation process.

4.3.2 Regression for total dataset

Figure 10a shows the relative importance of each block in the LV's for the product property Heat Stability for the process FA distillation for the total dataset (all groups together). The first three LV's contain the variability in Z important for product properties (Y) and the variability in X that is correlated with Z. An analysis of loadings of the first three LV's showed the different fatty acid compositions and the different process settings that are used between feedstocks. LV4 and LV5 contain the orthogonal information in X explaining Y, or in other words the variability in X important for Y that is not correlated with the variability in Z. The loadings of the X variables in LV4 and LV5 are shown in **Figure 10b**. Middle reflux ratio and flow of middle reflux both have negative loadings for both LV's, while the temperature at the bottom of the tower has positive

loadings for both LV's. This confirms the results found for tallow and tallow+, as it again indicates that the variability in the middle reflux ratio and bottom tower temperature are correlated with the variability in fatty acid heat stability. Thanks to the use of SMB-PLS, the need to determine grouping in step 4 can be bypassed. However, based on the results of only this analysis it is difficult to quantify the actual effects for the different feedstocks. Moreover, effects only visible in one of the feedstocks, like the effect of process parameters on p-Anasidine value for distilled PFAD fatty acids, could not be identified from the total dataset.

Figure 10. Outcome of using SMB-PLS on the total FA Distillation dataset for the product property Heat Stability. a. Relative Importance of each block in every LV. b. Loadings of the X variables in LV4 and LV5.

5. Discussion

The case studies in fatty acid production have shown how the sMVA strategy can be applied to gain process understanding of industrial-scale biorefinery processes. The critical sources of variability in the feedstock properties and process parameters that affect product quality could be identified. Moreover, based on the results, a first set of suggestions were made for further control and optimization studies. If only historical data is used, the possibility for direct optimization after applying the strategy is however limited. This strategy does not completely eliminate the need for experiments, but the number of variables and the number of experiments to be included in the designed experiments can be drastically reduced. By focusing on the key feedstock and process parameters, improvements can be achieved more rapidly. Applying this strategy can thus reduce the actual time and resources spent on optimization and control efforts. In this study, the strategy was tested for a chemical reaction and a physical separation process. It is recommended to test the strategy for bioprocesses as well, such as fermentation, as these may have additional challenges. This would allow further generalizing the strategy to be applicable to all biorefinery processes.

Some recommendations can be made towards future application of the strategy. Due to a lack of data for certain months, the effect of seasonal variability of feedstocks could not be explored fully. Therefore, it is recommended to collect a dataset over a longer period (e.g. 2 years). During exploratory analysis, grouping is best performed on the data blocks separately and class variables can be used to explain the observed groups. Grouping in feedstock properties (Z) or process parameters (X) was generally observed in product properties (Y) as well. As multiple installations were used for the same process, we could also investigate grouping by installation. This was only observed on the X dataset. Even though the chosen installation might not visibly affect product properties, it may drastically influence process parameters, even if the installations are identical

by design. This can be due to different reasons, such as sensor calibration or equipment maintenance. Therefore, if multiple installations are used, this grouping should also be investigated for the X dataset instead of only on the Y dataset. Although not formally included in the strategy, an investigation of the biplot of the combined ZXY dataset of every group using MB-PCA is useful to get an initial overview of the relations of variables between blocks. At this stage, hypotheses can be formulated regarding possible relations between dependent (Y) and independent variables (Z X).

During regression analysis, the expected correlation between feedstock properties (Z) and process parameters (X) was a critical factor in determining the most suited PLS methodology. As the correlation between Z and X, either directly via physical responses or indirectly due to feedback and feedforward control loops, was limited, two separate PLS models, one between Z and Y and one between X and Y, gave the best and easiest to interpret results. Using one MB-PLS model rarely improved the R^2_{CV} and the results were harder to interpret due to mixing of Z and X variables in LVs. However, if a significant correlation is found, one sequential multi-block PLS (SMB-PLS) is recommended as the model captures the correlated and orthogonal information between X and Y in separate LV's. This algorithm is, however, more complicated, and a commercial package or toolbox of this method is not yet available. Performing one regression analysis with all Y variables simultaneously resulted in models underestimating the variability that can be explained and that were harder to interpret. Therefore, a separate analysis for every cluster of correlated Y variables or every individual Y variable is preferred over a single analysis with all Y variables together. In general, product properties for which an adequate control strategy is implemented showed low R^2_{CV} values, while the product properties for which no control exists could more often be explained by the variability in the feedstock and process variables. If the process is well controlled, the

variation in quality will be limited because the larger effects are handled and removed by the process control and only a low percentage of the variation in Y remains, which can be explained by the variability in Z and X. However, if the explained variability is high, this indicates that the process control is not adequate. In case of collinearity between independent variables, an estimated negative regression coefficient was sometimes obtained for a variable that is positively correlated to the response. It is therefore recommended to be careful in interpreting the regression coefficient values and to investigate the correlation between the selected independent variables, for example via a correlation matrix. A different approach in the suggested sMVA strategy that should be tested in future research is the possibility of using SMB-PLS for exploratory analysis instead of PCA. In that case, the total dataset can be used, and insights would be gained both regarding correlations between independent variables and between independent and dependent variables. In that case, grouping in data could be controlled directly in the scores plot of the SMB-PLS analysis. After SMB-PLS analysis, targeted PLS analyses can be performed for specific groups or variables.

For both fatty acid production processes, clear groups were found by feedstock type. This shows that a large part of the variability in product quality is due to the difference between feedstock types. For both processes, the optimal process parameters will differ between feedstocks and optimization studies are required to determine the optimal settings of parameters per fat/oil type. Nevertheless, large variability still exists between campaigns using the same fat/oil as feedstock. For fat/oil hydrolysis, feed flow rate was found to be the most important process parameter, while for fatty acid distillation, it was found to be the variability in the middle reflux ratio causing variability in product quality. However, both are linked to the variability in the feedstock flow rate. If a highly variable flow rate cannot be avoided, a proper control strategy should be implemented so other process variables are adequately controlled to compensate for the changing flow rate.

Specifically for fat/oil hydrolysis, past optimization efforts have focused on increasing the degree of hydrolysis. In the future, also %glycerol and side reactions such as %C18:3 breakdown should be taken into account. Overall, for the processes in fatty acid production, variability in feedstock properties was not always the main cause for variability in product quality. This is contrary to existing belief, where out-of-spec campaigns are often attributed to the quality of the feedstock instead of to different process parameters.

The performed MVA analysis can be an important basis to steer further research and modelling efforts for the optimization, monitoring and control of oleochemical processes. Existing mechanistic models for fat/oil hydrolysis don't include side reactions such as the breakdown of C18:3 and use simplified compositions, such as one fatty acid and glyceride component.^{29, 30} This study has shown the importance of differentiating between feedstock types, which underlines the importance of using a more specific composition if modelling is used to optimize process parameters. Existing models of fatty acid distillation focus on the vapour-liquid equilibrium (VLE) of fatty acids and don't include quality parameters such as colour.³¹ Including quality parameters such as colour in a mechanistic model is highly challenging due to the many components involved in colour formation. The development of a hybrid model, combining the mechanistic VLE models and statistical quality models, could provide the most holistic view of the process. Finally, it would be beneficial to apply the sMVA strategy to other oleochemical processes, especially those more challenging to model mechanistically such as crystallization.

6. Conclusions

A systematic MVA strategy for improved process understanding of industrial biorefinery processes was proposed. The 9-step strategy allows determining the critical sources of variability in product quality of an industrial biorefinery process. The strategy is limited to gaining process

understanding, but it can be an important starting point to propose hypotheses to be tested through designed experiments or modelling studies focused on optimization, monitoring and control. The strategy was successfully applied to two processes in fatty acid production. In future research, the sMVA strategy can be applied to other (bio)-processes to further show its general applicability.

Different MVA techniques were tested and recommendations were given for every step of the strategy. Single and multi-block PCA analysis of product properties, feedstock properties and process parameters allow identification of grouping, outliers and background effects. The optimal PLS regression methodology, resulting in the best and easiest to interpret results, can be chosen based on the correlation between feedstock properties and process parameters.

For the case studies in fatty acid production, product quality largely depends on the type of fat or oil used, such as canola or palm oil. However, if a single oil type is used, the variability in product quality does not always critically depend on the variability in feedstock properties. It was found that process parameters, such as the middle reflux ratio in fatty acid distillation, need a better control in relation to a changing feed flow rate. Based on these findings, control improvements were proposed, such as a higher and fixed middle reflux ratio instead of a fixed valve% controlling the reflux rate in fatty acid distillation. The findings of these case studies can be used to develop or improve mechanistic or hybrid models of both processes to further steer the optimization and control of both operations.

Glossary

%Variance Explained	The amount of variance in the dataset that is explained by the model or by a specific PC/LV. The total variance explained is a measure of how well the model fits the data.
Biplot	In a biplot, the scores and loadings are superimposed in one figure. This allows for a better understanding of which variables determine the grouping found in the scores plot.

Coefficient of Determination (R^2)	Relative value for the total variance included in Y that is explained by a regression model. R^2_{Cal} is a measure of the %variance in Y in the calibration dataset that is captured by the model. R^2_{CV} is an indication of the %variance in Y captured for new data according to internal validation and is therefore a more realistic assessment of the predictive ability of the model.
Hotelling T^2	The Hotelling's T^2 measures the distance from a sample to the centre of the model. A sample with a high Hotelling's T^2 is different from the average sample and may therefore have a large influence on the model.
Loadings	The loading of an original variable for a PC/LV measures how much the variable contributes to that PC/LV. A very positive or negative loading means the variable has a strong contribution for that PC/LV, while a loading close to zero means it contributes less to that PC/LV. Plotting the loadings of the original variables on two PC/LV in a scatter plot indicates how the original variables relate to one another.
Multi-Block PLS (MB-PLS)	MB-PLS splits the X variables into meaningful blocks, with each block usually corresponding to a feedstock, a process unit or a section of a unit. MB-PLS is not a PLS between each X block and Y separately, but one PLS where the blocks are weighted in such a way that their combination is most predictive of Y. ²⁰
Partial Least Squares Regression (PLS)	PLS regression projects a large number of variables into a lower dimensional set of new variables called latent variables (LV). The PLS algorithm searches the correlation between the independent variables (X) and the dependent variables (Y). The LV are the projections that maximize the amount of variance in X that is related with the variance in Y. This technique is the most commonly used MVA technique for regression. ²³
Principal Component Analysis (PCA)	PCA projects a dataset with a large number of variables into a lower dimensional set of new variables called principal components (PC). The PC are sequentially extracted to maximize the amount of variance in the total dataset captured by each component. This technique can be used for the exploratory analysis of a dataset to find grouping of samples, clustering of variables or to detect outlying samples. It is a good starting point before regression analysis as it provides the practitioner with an overview of the data. ³²
Q residuals	The Q residuals measures the distance between a sample and its projection on the model. A sample with high Q residuals is therefore badly explained by the model.
Regression coefficient	By representing a PLS model as a multiple regression equation ($Y = b_0 + b_1x_1 + b_2x_2 + \dots$), the regression coefficients $b_1, b_2, \dots$ for each independent variable over all model components can be calculated.

	These values indicate the importance of the independent variables to the model and the relations with the dependent variables (Y).
Root Mean Square Error (RMSE)	The root mean square error (RMSE) for a variable indicates the difference between the predicted values compared to measured values. It is a measure of how well the model explains/predicts a variable. It is an absolute value and can be calculated for calibration (RMSEC), cross-validation/internal validation (RMSECV) or external prediction (RMSEP).
Scores	The score of a sample for a PC/LV is the projection of the data of the sample for that PC/LV. For PCA, plotting the scores of samples on two PC is used to inspect the relations between samples. If the relative positions of samples in the scatter plot are close, it indicates similarity. This plot therefore allows identifying grouping of samples or detecting outlying samples.
Sequential MB-PLS (SMB-PLS)	When splitting X variables into different blocks, variables between these blocks may be correlated. The key feature of this algorithm is that correlated information between two X blocks are combined into a common LV whereas the orthogonal information is captured by a different LV. This allows for an easier interpretation of the variance in different X blocks that is related with the variance in Y. ²⁴
Variable Importance in Projection (VIP)	The Variable Importance in Projection (VIP) of a variable indicates the importance of each original independent variable in the regression model. VIP scores can therefore be used to select the most important variables of a model, e.g. all variables with a VIP score above 1. ³³

Supporting Information

The following supporting information is provided:

1. Systematic MVA strategy for improved process understanding of industrial biorefinery processes: Step-by-Step Guide
2. Information on Oleochemical Properties
3. Information on the Variability in the Dataset

This information is available free of charge via the Internet at <http://pubs.acs.org/>

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. These authors contributed equally.

Funding Sources

This work was supported by Oleon NV and the Flanders agency for Innovation and Entrepreneurship (VLAIO), acting on behalf of the Hermes fund [grant number HBC.2017.0217]. Oleon NV was not involved in the modelling or development of the framework but supported the authors during the interpretation of the results of the case studies. Oleon NV approved to submit this manuscript for publication. VLAIO supports independent research and did not have a role in the preparation, review and approval of the manuscript and decision to submit the manuscript for publication.

Acknowledgments

The authors wish to thank Oleon NV and the Flanders agency for Innovation and Entrepreneurship for their financial support. The authors also wish to thank the process department of Oleon NV Ertvelde and the department of chemical engineering of Université Laval, Québec.

Abbreviations

B, Background Variables; *b*, #Background Variables; C, Class Information; *c*, #Class Variables; Ca:b, Fatty Acid species with a carbon atoms and b double bounds; corr, Correlated; CV, Cross Validation; D1/D2, Distillation unit #; E, Residual matrix; F, Residual matrix of Y after step 1 in SMB-PLS; FA, Fatty Acids; HP, High Pressure (Steam); LV, Latent Variable; MB-PLS, Multi-Block Partial Least Squares Regression; MVA, Multivariate Analysis; *n*, #Campaigns; orth, Orthogonal; PC, Principal Component; PCA, Principal Component Analysis; PFAD, Palm Fatty

Acid Distillate; PLS, Partial Least Squares Regression; R^2_{cv} , Coefficient of determination for Cross Validation; RMSE, Root Mean Square Error; RMSECV, Root Mean Square Error of Cross Validation; RMSEP, Root Mean Square Error of Prediction; SMB-PLS, Sequential Multi-Block Partial Least Squares Regression; sMVA, Systematic Multivariate Analysis Strategy; T, Super Block or transposed ; t, Scores ; T1/T2/T3, Hydrolysis Tower # ; TEMP, Temperature; VIP, Variable Importance in Projection; VLE, Vapour-Liquid Equilibrium; w^T , Loadings; X, Process Parameters; x, #Process Variables; $\bar{x}$, Average; Y, Product Properties; y, #Product Variables; Z, Feedstock Properties; z, #Feedstock Variables; σ , Standard Deviation

References

1. Langholtz, M. H.; Stokes, B. J.; Eaton, L. M.; Brandt, C. C.; Davis, M. R.; Theiss, T. J.; Turhollow Jr, A. F.; Webb, E.; Coleman, A.; Wigmosta, M. *2016 Billion-ton report: advancing domestic resources for a thriving bioeconomy, volume 1: economic availability of feedstocks*; Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States): 2016.
2. Maity, S. K., Opportunities, recent trends and challenges of integrated biorefinery: Part I. *Renewable and Sustainable Energy Reviews* **2015**, 43, 1427-1445.
3. Lee, R. A.; Lavoie, J.-M., From first-to third-generation biofuels: Challenges of producing a commodity from a biomass of increasing complexity. *Animal Frontiers* **2013**, 3, (2), 6-11.
4. Pinzi, S.; Lopez-Gimenez, F.; Ruiz, J.; Dorado, M., Response surface modeling to predict biodiesel yield in a multi-feedstock biodiesel production plant. *Bioresource technology* **2010**, 101, (24), 9587-9593.
5. Hsu, C. S.; Robinson, P. R., *Practical advances in petroleum processing*. Springer Science & Business Media: New York, 2007; Vol. 1, p 410.
6. Narani, A.; Coffman, P.; Gardner, J.; Li, C.; Ray, A. E.; Hartley, D. S.; Stettler, A.; Konda, N. M.; Simmons, B.; Pray, T. R., Predictive modeling to de-risk bio-based manufacturing by adapting to variability in lignocellulosic biomass supply. *Bioresource technology* **2017**, 243, 676-685.
7. Ferreira, A. P.; Tobyn, M., Multivariate analysis in the pharmaceutical industry: enabling process understanding and improvement in the PAT and QbD era. *Pharmaceutical development and technology* **2015**, 20, (5), 513-527.
8. Rincón, L. E.; Hernández, V.; Cardona, C. A., Analysis of technological schemes for the efficient production of added-value products from Colombian oleochemical feedstocks. *Process Biochemistry* **2014**, 49, (3), 474-489.
9. dos Santos, M. T.; Gerbaud, V.; Le Roux, G. A., Beyond biofuels: economic opportunities, recent advances and challenges in property modeling for vegetable oils. *Green Processing and Synthesis* **2014**, 3, (6), 401-410.

10. Nachtergaele, P.; De Meester, S.; Dewulf, J., Environmental sustainability assessment of renewables-based propylene glycol at full industrial scale production. *Journal of Chemical Technology & Biotechnology* **2019**, 94, (6), 1808-1815.
11. Gunstone, F., *Vegetable oils in food technology: composition, properties and uses*. 2 ed.; Blackwell Publishing Ltd.: West Sussex, 2011.
12. Rathore, A. S.; Chopda, V. R.; Gomes, J., Knowledge management in a waste based biorefinery in the QbD paradigm. *Bioresource technology* **2016**, 215, 63-75.
13. Kourti, T., Pharmaceutical manufacturing: the role of multivariate analysis in design space, control strategy, process understanding, troubleshooting, and optimization. *Chemical Engineering in the Pharmaceutical Industry: Drug Product Design, Development, and Modeling* **2019**, 601-629.
14. Duchesne, C.; MacGregor, J. F., Establishing multivariate specification regions for incoming materials. *Journal of quality technology* **2004**, 36, (1), 78-94.
15. Tomba, E.; De Martin, M.; Facco, P.; Robertson, J.; Zomer, S.; Bezzo, F.; Barolo, M., General procedure to aid the development of continuous pharmaceutical processes using multivariate statistical modeling—an industrial case study. *International journal of pharmaceutics* **2013**, 444, (1-2), 25-39.
16. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J., OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2006**, 20, (8-10), 341-351.
17. Kourti, T., Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of adaptive control and signal processing* **2005**, 19, (4), 213-246.
18. Lourenço, N.; Lopes, J.; Almeida, C.; Sarraguça, M.; Pinheiro, H., Bioreactor monitoring with spectroscopy and chemometrics: a review. *Analytical and bioanalytical chemistry* **2012**, 404, (4), 1211-1237.
19. Rousseeuw, P. J.; Bossche, W. V. D., Detecting deviating data cells. *Technometrics* **2018**, 60, (2), 135-145.
20. Westerhuis, J. A.; Kourti, T.; MacGregor, J. F., Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics: A Journal of the Chemometrics Society* **1998**, 12, (5), 301-321.
21. Eriksson, L.; Andersson, P. L.; Johansson, E.; Tysklind, M., Megavariate analysis of environmental QSAR data. Part I—A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Molecular diversity* **2006**, 10, (2), 169-186.
22. Eriksson, L.; Byrne, T.; Johansson, E.; Trygg, J.; Vikström, C., *Multi-and megavariate data analysis basic principles and applications*. Umetrics Academy: Malmö, 2013; Vol. 1, p 500.
23. Abdi, H., Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley interdisciplinary reviews: computational statistics* **2010**, 2, (1), 97-106.
24. Lauzon-Gauthier, J.; Manolescu, P.; Duchesne, C., The Sequential Multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability. *Chemometrics and Intelligent Laboratory Systems* **2018**, 180, 72-83.
25. Azari, K.; Lauzon-Gauthier, J.; Tessier, J.; Duchesne, C., Establishing multivariate specification regions for raw materials using SMB-PLS. *IFAC-PapersOnLine* **2015**, 48, (8), 1132-1137.

26. Kjeldahl, K.; Bro, R., Some common misunderstandings in chemometrics. *Journal of Chemometrics* **2010**, 24, (7-8), 558-564.
27. Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, 58, (1), 267-288.
28. Chong, I.-G.; Jun, C.-H., Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems* **2005**, 78, (1-2), 103-112.
29. Jones, M. N.; Forero-Hernandez, H.; Zubov, A.; Sarup, B.; Sin, G., Splitting Triglycerides with a Counter-Current Liquid–Liquid Spray Column: Modeling, Global Sensitivity Analysis, Parameter Estimation and Optimization. *Processes* **2019**, 7, (12), 881.
30. Forero-Hernandez, H.; Jones, M. N.; Sarup, B.; Jensen, A. D.; Abildskov, J.; Sin, G., Comprehensive development, uncertainty and sensitivity analysis of a model for the hydrolysis of rapeseed oil. *Computers & Chemical Engineering* **2020**, 133, 106631.
31. Perederic, O. A.; Cunico, L. P.; Kalakul, S.; Sarup, B.; Woodley, J. M.; Kontogeorgis, G. M.; Gani, R., Systematic identification method for data analysis and phase equilibria modelling for lipids systems. *The Journal of Chemical Thermodynamics* **2018**, 121, 153-169.
32. Jolliffe, I. T.; Cadima, J., Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, 374, (2065).
33. Jun, C.-H.; Lee, S.-H.; Park, H.-S.; Lee, J.-H. In *Use of partial least squares regression for variable selection and quality prediction*, 2009 International Conference on Computers & Industrial Engineering, 2009; IEEE: 2009; pp 1302-1307.

Table of Contents and Abstract Graphics