CHEMICAL ENGINEERING TRANSACTIONS
VOL. 43, 2015
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Chief Editors: Sauro Pierucci, Jiří J. Klemeš

ISBN 978-88-95608-34-1; **ISSN** 2283-9216
DOI: 10.3303/CET1543259

Quality Control of Industrial Detergents through Infra-Red Spectroscopy Measurements Coupled with Partial Least Square Regression

Alessandra Taris[a], Massimiliano Grosso*[a], Fabio Zonfrilli[b], Vincenzo Guida[b]
[a]Dipartimento di Ingegneria Meccanica, Chimica e dei Materiali, Università degli Studi di Cagliari, Via Marengo 2, 09123, Cagliari, Italy
[b]Procter & Gamble, Pomezia R&D Research Center, Via Ardeatina 100, 00040 Pomezia, Italy
massimiliano.grosso@dimcm.unica.it

Quality control of industrial products has become essential in modern industry as it aims to satisfy customer demands. Therefore it requires fast and simple procedures in order to ensure efficient on-line process monitoring and detect abnormal deviation from certain product specifications. In this work, commercial detergent quality control was performed by means of (i) Fourier Transform Infra-Red (FT-IR) spectroscopy and (ii) Partial Least Square Regression (PLS-R) that allows for prediction from multivariate spectra. Sodium hydroxide and non-ionic surfactant concentrations were considered for the calibration PLS-R model. Results demonstrated excellent predictive performance of the PLS-R model. In addition, its robustness was evaluated by mimicking a fault in the process, in this case a deviation of anionic surfactant concentration. It was found that a Q_x statistic can be introduced with the purpose to assess whether sodium hydroxide and non-ionic surfactant concentration are correctly estimated in presence of external interference.

1. Introduction

Commercial hard surfaces detergents are a complex blend composed by different chemical species, whose quality strongly depends on the relative proportions. Quality control is usually performed on the end-product at the completion of the process with off-line conventional analytical techniques (e.g. chromatography) that introduce time delay and consequently reduce the effectiveness of quality control. Indeed, in case the product does not meet certain specifications, it has to be reprocessed, implying an increase of cost and time waste. Thus, for the analysis and control of critical quality variables (i.e. the compounds proportions) during the manufacturing process, real time analyzers are required. For the case at hand, a proper experimental tool might be the attenuated total reflectance (ATR) coupled with Fourier transform infrared (FTIR) spectrometer (Stuart, 2004). This is an innovative, non-destructive analytical technique, capable of measuring in very fast times aqueous samples, characterizing materials in a really efficient way and that are well suited for on-line measurements. Nevertheless, because of the large amount of spectral information, interpretation and correlation of the collected spectra with quality variables is a challenging task. Therefore, multivariate approaches represent a powerful tool for data analysis and compression.
In this work, we tackle the issue of the on-line monitoring of detergent mass production through multivariate statistical process control approach applied on FTIR measurements of detergent samples. In particular, Partial Least Squares Regression (PLS-R) is considered as the best multivariate technique for the quantitative analysis of spectroscopic data because it enables to overcome common problems such as collinearity, band overlaps and interactions. It is here implemented for the determination of the concentration of some selected compounds (sodium hydroxide and non-ionic surfactant). For this purpose, two different sets of detergent samples are designed by jointly varying these two compounds: (i) a training set used to calibrate the PLS-R model and (ii) a validation set to test the PLS-R model. Moreover, small variations of the other components in the blend were introduced in order to mimic typical fluctuations unavoidably present in the standard mass

Please cite this article as: Taris A., Grosso M., Zonfrilli F., Guida V., 2015, Quality control of industrial detergents through infra-red spectroscopy measurements coupled with partial least square regression, Chemical Engineering Transactions, 43, 1549-1554
DOI: 10.3303/CET1543259

production. These two samples sets respect the standard of the end-product and hereafter they will be referred as normal operating conditions (NOC) and defined as in-control samples. Different works in literature usually determine the best calibration PLS-R model for compounds concentration in detergents formulations (Rohman et al., 2011) without considering the possible presence of external interferences that could worsen its prediction ability. In fact, since the PLS-R model is built only on a limited number of compounds, it could be no longer consistent when the system is *out-of-control*, that is in presence of large deviations of other compounds not taken into account in the PLS-R model calibration.
Here, a Q_x statistic was also proposed to assess the concentration prediction reliability when a fault occurs, i.e. a perturbation from the NOC. To this aim the fault was simulated considering (iii) an out-of-control samples set where anionic surfactant concentration is higher than the reference value defined for the NOC. Eventually, the estimation of the Q_x statistic greatly improves the effectiveness of quality monitoring during detergents manufacturing process because it will be capable of classifying samples as in-control or out-of-control.

2. Experimental

2.1 Samples sets

The detergents used for the preparation were commercial formulations used for the mass production. Three different samples sets were considered and their Infrared spectra are reported in Figure 1: a training set (34 samples) used to develop the PLS model (Figure 1.a) and a validation set (6 samples) to test prediction ability (Figure 1.b). In both these sets, sodium hydroxide and non-ionic surfactant concentration vary (± 10 % of the average nominal values), while anionic surfactant concentration was kept at low level. In addition, a further set was generated which is an out-of-control test set (12 samples), characterized by anionic surfactant concentration 22 % higher than the NOC value (Figure 1.c), while variations of sodium hydroxide and non-ionic surfactant concentration for these samples were the same designed for the in-control samples. Concentrations of other compounds (sodium carbonate, fatty acid, pH buffer, chelating agents, anphoteric surfactant, ethanol, perfume, polymer additive) were slightly varied in all the samples in order to simulate industrial process fluctuations.

2.2 Infrared measurements

The infrared measurements were performed at the Procter & Gamble Brussels Innovation Center (BIC) in on a Thermo Scientific Nicolet™iS™10 FT-IR Spectrometer with a deuterated triglycine sulfate (DTGS) detector and a KBr/Ge mid-infrared optimized beamsplitter. The spectra cover the range from 3000 to 800 cm^{-1} with a wavenumber resolution equal to 1.928 cm^{-1}.

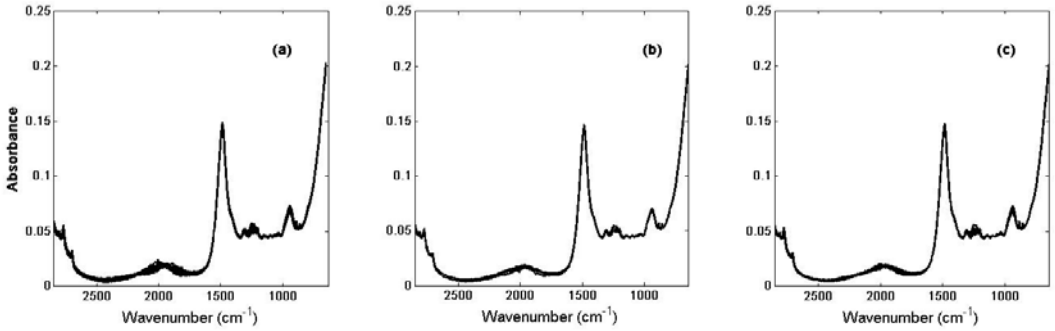

Figure 1. Infrared spectra for (a) training (34 samples), (b) validation (6 samples) and (c) out-of-control (12 samples)

Experiments were carried out by jointly varying the non-ionic surfactant and the sodium hydroxide according to an *I*-optimal design (Montgomery, 2008). The experimental conditions are arranged into a response matrix $\mathbf{Y}_{(I\times M)}$ where *M* and *I* refer to the number of compounds (for the case at hand *M*=2) and the number of samples, respectively. The corresponding experimental spectra are assembled into a matrix $\mathbf{X}_{(I\times J)}$ consisting of the *I* experimental infrared spectra collected at the different *J* wavenumbers.

3. Methods

3.1 Model calibration

Given a predictor matrix $\mathbf{X}_{(I\times J)}$ and a response matrix $\mathbf{Y}_{(I\times M)}$ the PLS algorithm projects **X** and **Y** onto a low-dimensional space defined by a small number of latent variables *A* (Li et al., 2010) as expressed in equations (1) and (2)

$$\underset{I\times J}{\mathbf{X}} = \underset{I\times A}{\mathbf{T}} \cdot \underset{A\times J}{\mathbf{P}^{\mathbf{T}}} + \underset{I\times J}{\mathbf{E}} \tag{1}$$

$$\underset{I\times M}{\mathbf{Y}} = \underset{I\times A}{\mathbf{T}} \cdot \underset{A\times M}{\boldsymbol{Q}^{\boldsymbol{T}}} + \underset{I\times M}{\mathbf{F}} \tag{2}$$

Where *A* « *J* is the number of the latent variables that allow for describing adequately the experimental data, **T** is the orthonormal score matrix, **P** and **Q** are the loading matrices for **X** and **Y** respectively. **E** and **F** are the residuals matrices of **X** and **Y**. In general, **X** and **Y** are usually pre-processed and scaled to unity variance and mean centred. The basic idea in PLS-R is that the covariance between **X** and **Y** should be maximized and there are several ways to solve the maximum optimization problem and compute PLS model matrices **P** and **Q**. In this work, the SIMPLS algorithm developed by De Jong (1993) was used since it appears faster and easier to interpret than nonlinear iterative partial least-squares one (NIPALS). PLS can be implemented to infer a single response variable (PLS1) or multiple response variables (PLS2). Here, PLS2 was adopted as it seemed more appropriate for process monitoring. This sounds reasonable since the joint regression of multiple response variables should provide more information than the ones collected by building *M* different independent PLS models (Li et al., 2010).

3.2 Quality variables prediction

PLS method can determine the i-th sample concentration $\mathbf{y_i}$ from the corresponding spectrum $\mathbf{x_i}$ (MacGregor et al., 1994) as expressed in (3) and (4).

$$\underset{1\times M}{\mathbf{y_i}} = \underset{1\times J}{\boldsymbol{x_i}} \cdot \underset{J\times M}{\mathbf{B}} \tag{3}$$

where

$$\underset{J\times M}{\mathbf{B}} = \underset{J\times A}{\mathbf{R}} \cdot \underset{A\times M}{\mathbf{Q}^{\mathbf{T}}} \tag{4}$$

In equations (3) and (4) **R** is the pseudo-inverse of the **P** matrix and **B** is the regression coefficients matrix estimated through the matrices **R** and $\mathbf{Q}^{\mathbf{T}}$.

3.3 PLS-based statistical control

Besides the quantitative estimation of the compounds concentration, here a PLS-based monitoring was implemented (Kourti, 2005). In particular, the detection of deviations from nominal conditions and, as a consequence, the accuracy of the concentration prediction were performed by resorting to the Q_x statistic (Li et al., 2010). Such statistic is calculated for the i-th sample spectrum according to the Equation 5.

$$Q_{x,i} = \|\boldsymbol{x_i} - \widehat{\boldsymbol{x_i}}\|^2 = \left\|\boldsymbol{x_i} \cdot (\mathbf{I} - \mathbf{P} \cdot \mathbf{R}^{\mathbf{T}})\right\|^2 \tag{5}$$

$$Q_{x,lim} = g \cdot \chi^2_{h,\alpha} \tag{6}$$

Where $\widehat{\boldsymbol{x_i}}$ is the the i-th spectrum as predicted by the PLS model considering the first *A* latent variables. Under NOC, the Q_x statistic follows a chi-square distribution with *h* degrees freedom χ^2_h, and its threshold value is calculated through equation (6), where α is the significance level (usually 5 %), *g* and *h* depend on the mean and variance of the Q_x calculated for the calibration samples (Nomikos et al., 1995). When $Q_{x,i} > Q_{x,lim}$ the process behavior is supposed to be out-of-control.

4. Results and discussion

PLS model was built on the training set samples represented by the experimental matrix $\mathbf{X}_{(34\times1142)}$ and concentration matrix $\mathbf{Y}_{(34\times2)}$. Model matrices (**P**, **Q** and **B**) are evaluated using SIMPLS algorithm. For the case at hand, six latent variables were chosen as the variance explained for both **X** and **Y** achieves 92 and 97 % respectively. The sodium hydroxide and non-ionic surfactant concentration ($\mathbf{y}_i$) for the three different samples sets (training, validation and out-of-control test sets) were calculated according to equation (3).

The PLS2 model here developed demonstrates high predictive performance achieving R^2 values of 97.4 and 97.3 % for sodium hydroxide and non ionic surfactant concentration estimation, respectively. Similarly, the Root Mean Squared Error of Calibration (RMSEC) values are quite low and equal to 0.01 and 0.106. For the validation set the Root Mean Squared Error of Prediction (RMSEP) is equal to 0.0187 and 0.11.

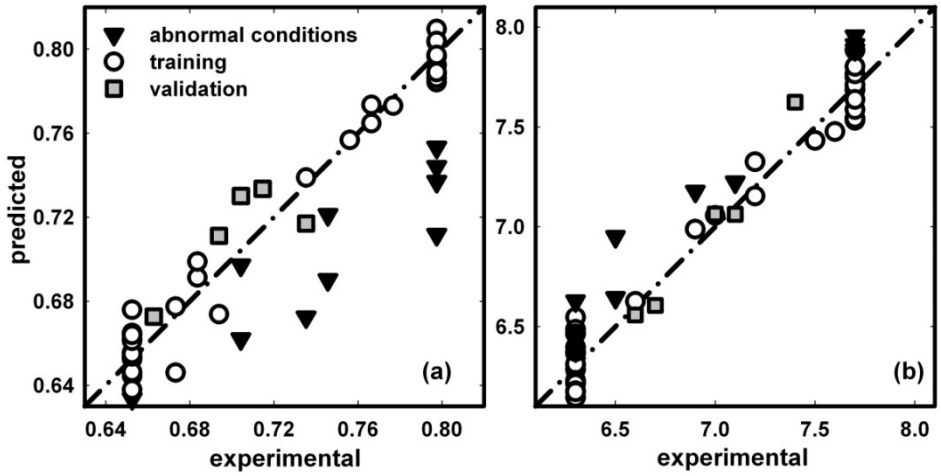

Figure 2. Experimental vs predicted concentration of sodium hydroxide (a) and non-ionic surfactant (b). White circles, gray squares and black triangles represent training, validation and out-of-control samples, respectively

The experimental vs predicted concentrations (arbitrary units) are reported in Figure 2 for the sodium hydroxide concentration (Figure 2.a) and non ionic surfactant concentration (Figure 2.b). It can be seen that the training samples (white circles) and the validation data (gray squares) are well predicted for both sodium hydroxide and non ionic surfactant. On the other hand, it was observed that the out-of-control samples, when projected in the PLS model (black triangles in Figures 2), cannot be accurately predicted. In particular, the sodium hydroxide concentration was underestimated, whereas the non-ionic surfactant was slightly overestimated. The explanation of such lack of fit seems obvious: possible variations of anionic surfactant concentration were not included into the PLS-R model calibration. As a consequence, the model is not suited to predict concentrations corresponding to out-of-control samples.

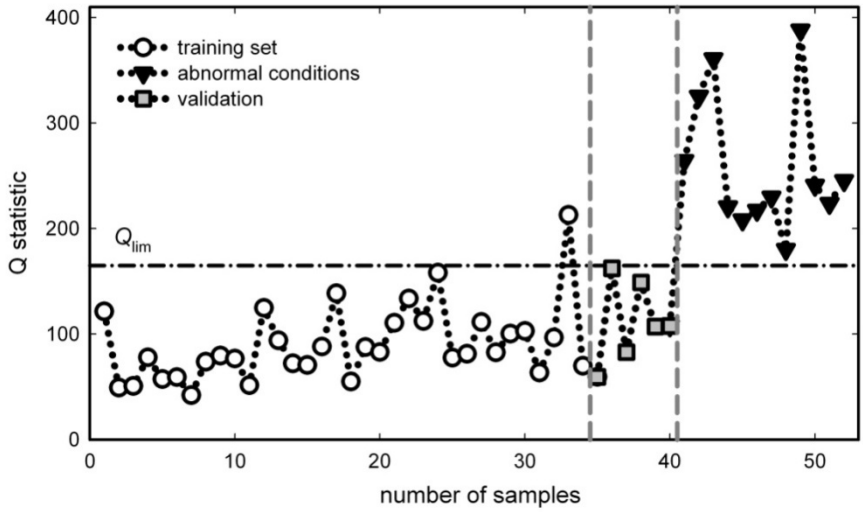

Figure 3. Q statistic evaluated for samples belonging to training (white circles), validation (gray squares) and out-of-control sets (black triangles), respectively. The dashed-dotted line represents the Q limit

For these reasons, a PLS-based monitoring is crucial. It can be mentioned that multivariate statistical control for detergent production has already been performed by means of PCA (Taris et al., 2014): they proposed an elliptical operating region in the T^2-Q control chart to detect out-of-control samples. Nevertheless, the goal of this work is the further quantification of quality variables and the assessment of the model robustness and

precision. Therefore, the Q_x statistic for each sample and the $Q_{x,lim}$ were calculated according to Equation (5) and (6) and they are reported in Figure 3. It can be observed that samples belonging to training and validation sets assume Q_x values smaller than the threshold. Thus, they are correctly classified as in-control, that is the PLS-R model is supposed to correctly predict the quality variables. On the other hand, the out-of-control samples exceed the limit and anomalous conditions are detected. In this case, the PLS-R model cannot be used to infer the compound concentrations.

The efficiency of the Q_x statistic is further evaluated by means of the Receiver Operating Characteristic (ROC) curves (Scheipers et al., 2005). These are two-dimensional graphs of the true positive rates (TPs; i.e., successes) versus the false positive rates (FPs; i.e., false alarms). The area under the ROC curve is the so-called AUC index, which is a scalar measure of the overall performance of a classifier, averaging across different thresholds that can be used to generate a classifier. In general, a model with a larger AUC is preferred to a model with a smaller one. The AUC of a random classifier is 0.5, whereas AUC=1 corresponds to perfect classification. Here, two ROC curves were determined: (i) training set was compared with out-of-control set and (ii) training with validation set and depicted in Figure 4. The ideal scenario would be: (i) an AUC_1 value as close as possible to 1, when comparing the training set with out-of-control set (thus meaning a perfect separation between the two classes) and (ii) an AUC_2 value close to 0.5 when comparing the training set with the validation set. The obtained AUC values for cases (i) and (ii) were AUC_1=0.989 and AUC_2=0.77. This confirms the capability of the Q_x statistic to distinguish the in-control from the out-of-control samples.

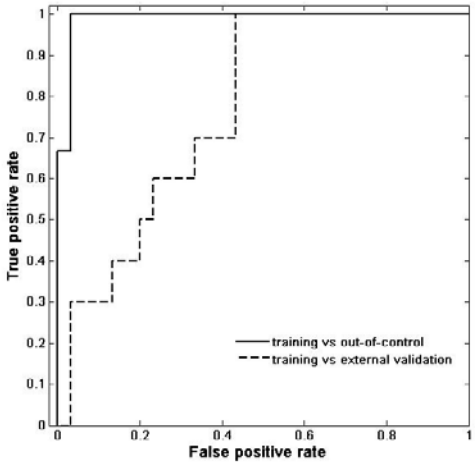

Figure 4. Roc curves resulting from the comparison of the training with out-of-control data (solid line) and the training with validation data (dashed line) for the Q_x statistic.

The proposed strategy for the on-line quality monitoring of commercial detergents is summarized in Figure 5: new spectra collected during the manufacturing process are projected onto the PLS subspace (previously calibrated using in-control samples), Q_x statistic is evaluated for each spectrum and compared to the threshold value. If the new sample complies with in-control samples, then prediction of quality variable can be carried out.

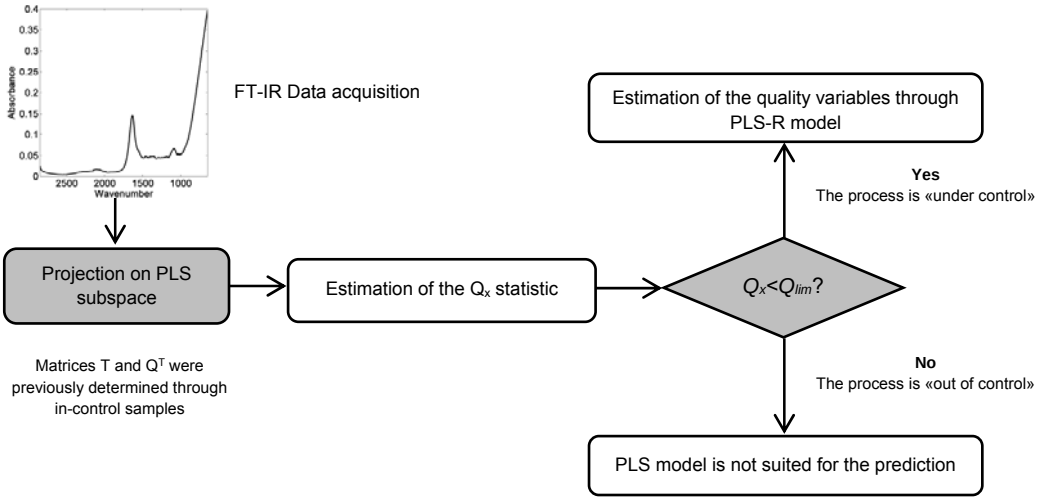

Figure 5. Flow diagram to be implemented during commercial detergent manufacturing process for quality monitoring.

5. Conclusions

Fourier Transform Infrared Spectroscopy combined with multivariate statistical analysis techniques were exploited for real time monitoring of the quality of hard surface detergents. The proposed strategy involves the development of a data-driven PLS regression model for the quantification of some quality variables and the employment of the Q_x statistic as a diagnostic tool able to find out the samples deviating from normal process behavior. The goal is twofold: the detection of anomalous conditions and then prediction of the quality variables. For this purpose concentration of sodium hydroxide and non-ionic surfactant in the blend was estimated and two sets of experimental data were taken into account: (i) a set of data following normal operating conditions and used for the calibration and validation model and (ii) a set of data collected far from the NOC that mimic the occurrence of a process fault during the manufacturing process.
In this work, Q_x statistic was demonstrated as an effective tool to clearly detect the out-of-control samples whose quality variables cannot be determined through PLS model. Furthermore, it was found out that PLS model correctly estimated the compounds concentration under nominal operating conditions.

References

De Jong S., 1993. Simpls: An Alternative Approach to Partial Least Squares Regression. Chemometrics and intelligent laboratory systems,18(3), 251-253.

Kourti T., 2005. Application of latent variable methods to process control and multivariate statistical process control in industry. International journal of adaptive control and signal processing, 19(4), 213-246.

Li G., Qin S. J., Zhoua D., 2010. Geometric properties of partial least squares for process monitoring. Automatica, 46(1), 204-210.

MacGregor J.F., Jaeckle C., Kiparissides C., Koutoudi M., 1994. Process monitoring and diagnosis by multiblock PLS methods. AIChE Journal, 40(5), 826-838.

Montgomery D., 2008. Design and analysis of experiments. Wiley, Hoboken, NJ, USA.

Nomikos P, MacGregor J.F., 1995. Multivariate SPC charts for monitoring batch processes. Technometrics, 37, 41–59.

Rohman A., Che Man Y.B, 2011. Determination of sodium fatty acid in soap Formulation Using Fourier Transform Infrared (FTIR) spectroscopy and multivariate calibrations. Journal of Surfactants and Detergents, 14(1), 9-14.

Scheipers U., Perrey C., Siebers S., Hansen C., Ermert H., 2005. A tutorial on the use of ROC analysis for computer-aided diagnostic systems. Ultrasonic Imaging, 27(3), 181-198.

Stuart B., 2004. Infrared spectroscopy: fundamentals and applications. Wiley, Chichester, UK.

Taris A., Grosso M., Zonfrilli F., Guida V., 2014. Statistical Control of Commercial Detergents Production through Fourier Transform Infra-Red Spectroscopy. Computer Aided Chemical Engineering, 33, 601–606.