The following full text is a publisher's version.

For additional information about this publication click this link.
https://hdl.handle.net/2066/225074

Please be advised that this information was generated on 2021-11-03 and may be subject to change.
Synchronizing process variables in time for industrial process monitoring and control

Tim Offermans a, Ewa Szymańska b, Lutgarde M.C. Buydens a, Jeroen J. Jansen a,∗

a Radboud University, Institute for Molecules and Materials, Heyendaalseweg 135, 6525AJ Nijmegen, The Netherlands
b FrieslandCampina, Amersfoort, The Netherlands

A R T I C L E  I N F O

Article history:
Received 4 February 2020
Revised 23 April 2020
Accepted 20 May 2020
Available online 29 May 2020

Keywords:
Soft-sensor
Process control
Dynamic synchronization
Regression
Industry
Statistical modelling

A B S T R A C T

The use of soft-sensors in industry is becoming more popular, as they allow for the prediction of critical product qualities from process variables in real-time. The requirement for this that all process variables are dynamically synchronized is often not met. Although different methods for dynamically synchronizing process variables are reported, no critical comparison of these methods is available. In this study we show that the choice in synchronization method significantly influences a soft-sensor’s accuracy. From the methods studied, median filtering using a moving window with a width of 168 minutes placed before the target times led to the highest sensor accuracy for the production plant studied, a method not reported in literature. This optimal width is remarkable, as the total processing time of the plant is 30 minutes. This suggests that changes in the physical state of the plant can affect the production quality than one might expect.

© 2020 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

(Bio)chemical production data contains many sources of variation, both known and unknown. These need to be understood, to define appropriate process control operations to retain or regain Normal Operating Conditions (NOC). Multivariate projection methods such as Principal Component Analysis and Partial Least Squares regression may offer clear insight in the variations of large numbers of variables in the data. They can be used to describe and understand the relations between process variables such as temperatures, pressures and flow rates throughout the production plant, and are able to define multivariate control limits (MacGregor and Kourti, 1995, Camacho, Picó et al., 2008).

Multivariate regression methods, such as Partial Least Squares (PLS), can be used to predict product quality variables that are costly or impossible to measure from process variables that can be readily measured. Prediction models applied as such are often referred to as soft-sensors, and can greatly improve process monitoring and control of a production plant (Fortuna, Graziani et al., 2007, Lin, Recke et al., 2007). They have, for instance, been used to successfully predict the sediment values of milk powder from operating data measured routinely from the production plant, superseding off-line analysis of this property (Rimpiläinen, Kaipio et al., 2015).

Robust and predictive soft-sensors require high-quality historical data (Kourti and MacGregor, 1995). For such a dataset, the dependent variables should be synchronized with the independent variables in time. However, in practice those variables may not all be measured at the same sampling frequency, and misalignments with considerable residence time may exist between the values measured. This is especially relevant for production facilities where the variables are measured using different control units throughout the plant, and is applicable to both continuous and batch productions. In such cases, the process variables should be dynamically synchronized prior to the development of a soft-sensor, or any bilinear model.

The problem of time missynchronization is addressed in literature involving soft-sensor development, and different methods to cope with it have been used. For the aforementioned study on predicting sediment values of milk powder, the variables were synchronized by calculating their averages over a 10 minute window spanning around the sampling times of the sediment values (Rimpiläinen, Kaipio et al., 2015). This method is also referred to as Boxcar smoothing or averaging (Holcomb and Norberg, 1955). In a study on industrial fluidized catalytic cracking, Slama et al. also dealt with time missynchronization by calculating averages over a window, but used hourly averages (Slama, 1991). An alternative method is to use the values of the variables nearest to the time
points of interest, and is used by Gabrielson et al. to synchronize process and spectroscopic data in time (Gabrielson, Jonsson et al., 2006).

Many other studies that describe the multivariate analysis of process variable do however not discuss synchronization of those variables (Zamprogna, Barolo et al., 2005, Ferrer, 2007, Måge, Mekv et al., 2008, Cuentas, Peñabaena-Niebles et al., 2016). Whether the problem did not exist for the plant in question or the researchers did not deem the method used for synchronization relevant to discuss is unclear. Furthermore, the studies that do report the synchronization method do not motivate the choice for that method, or compare different methods. This illustrates that there is a lack of consensus regarding the method that is most fitting for dynamic synchronization of process variables prior to multivariate analysis.

The problem of time missynchronization is to a certain extend comparable to that of missing data. This topic is covered in literature, also in context of multivariate statistical process control, and several methods to cope with it have been postulated and compared (Arteaga and Ferrer, 2002). However, such studies most often refer to cases where measurements are missing at random. In case of time missynchronization, data is not only missing by design but the fraction of data that is missing is generally larger than is considered with missing value imputation.

Another related problem encountered with multivariate modelling of process data that is covered in literature and should be noted is that of batch missynchronization (Kassidas, MacGregor et al., 1998, Neogi and Schlags, 1998, González-Martínez, Vitale et al., 2014). This problem arises when data from different production batches with different production lengths have to be compared, in which case the time axis of the batches have to be synchronized. This is a fundamentally different problem than that of process variable missynchronization. Batch synchronization aligns datasets of which the variables are already dynamically synchronized, while variable synchronization deals with aligning the variables of those individual datasets in the first place. However, one method used for batch synchronization can also be used for variable synchronization. For their study on an emulsion process, Neogi et al. used linear interpolation to match the time axes of different production batches (Neogi and Schlags, 1998). Linear interpolation could also be used to synchronize individual process variables to a universal time axis, and should be considered a viable method for dealing with variable missynchronization.

The objective of this study is to systematically and critically compare different methods for dynamically synchronizing process variables, and to identify which method leads to the most robust and informative statistical model. Different methods were compared and included the ones found in literature (window-based mean filtering, nearest value interpolation and linear interpolation), as well as three additional ones (window-based median filtering, cubic spline and previous value interpolation). The soft-sensors, which are PLS calibration models, were calibrated and validated on differently synchronized datasets using two scenarios: one where the same sample selection is used for each synchronized dataset, and one where the sample selection is optimized for each dataset individually. The performances of all soft-sensors as well as the actual values in each dataset were compared, and the most effective synchronization method was identified.

A dataset obtained from a dairy processing plant that produces milk protein powder from skim milk through precipitation has been synchronized and used in this study. The dried milk protein powder is obtained after consecutively heating, precipitation, washing and drying of the skim milk, as schematically shown in Fig. 1. The critical quality attribute is the mineral content of the protein powder product, which is measured off-line only a few times per day. The availability of a soft-sensor predicting the mineral content from the process variables measured on-line would enable more frequent monitoring of product quality. As the process variables for this production process are belonging to consecutive unit operations (the total processing time is about 30 minutes), synchronizing them in time is crucial for the development of a reliable soft-sensor. Dynamic synchronization using linear interpolation is currently offered by the data storage system of this production plant for inspection of the process data, and can thus be considered as the benchmark method.

2. Materials and methods

2.1. Dataset

A total of 175 mineral content values were collected from the milk protein powder plant over a period of three months. For the same time period, values for 32 process variables were collected to obtain a raw dataset. These variables include temperatures, flow rates, pressures and power consumptions, among others, and were selected using expert knowledge from plant operators. Setpoints variables are not taken into account. The average sampling interval differs per process variable, and ranges between 20 seconds and five minutes, as shown in Fig. 2. The sampling interval of the mineral content is around 8 hours.

2.2. Synchronization methods

The raw process variables were synchronized to the mineral content values using all of the synchronization methods found...
in literature (window-based filtering/averaging using means, nearest value interpolation and linear interpolation), as well as three additional ones. The principle of each method used is exemplified in Fig. 3A-F. The first additional method is window-based filtering/averaging using medians instead of means. As medians are more robust population estimators than means, using them as an alternative to synchronize the process variables may lead to more robust models. For both median and mean filtering, placement of the window around or before the target times as well as the width of the window will be studied. Secondly, cubic spline interpolation is tested as alternative to linear interpolation. Using cubic splines instead of a linear function for interpolation results in a smoother interpolant and incurs a smaller error, which can increase the accuracy of the synchronization process (Hall and Meyer, 1976). The final method tested matches the last registered value of each process variable to the time points of interest. This interpolation method might be the least accurate one, but it only uses data measured before the time points of interest and therefore corresponds closest to the on-line application of a process monitoring strategy.

For linear interpolation, the value for a process variable \( X \) at a desired target time is found by fitting the linear function given in Eq. 1 through the two available data points surrounding the time point. The coefficients for this function, \( a \) and \( b \), are fitted piecewise for each target time \( t \).

When cubic splines are used to interpolate the value for a process variable \( X \) at a desired target time, a third-order polynomial function is fitted through the data points around that target time, following Eq. 2. Two data points before and two data points after the target time \( t \) are used to fit the four coefficients \( a, b, c \) and \( d \), as exemplified in Fig. 3B. For a more elaborate explanation on cubic spline interpolation, the reader is referred to a comprehensive tutorial by McKinley and Levine (McKinley and Levine, 1998). For this study, the built-in interpolation code of MATLAB (interp1.m) was used for both linear and cubic spline interpolation.

\[
X(t) = at + b
\]  
(1)

\[
X(t) = a(t)^3 + b(t)^2 + ct + d
\]  
(2)

For nearest-value interpolation of a process variable, the measurement closest to the target time is selected as the representative value at that target time, as illustrated in Fig. 3C. If the distance in time of the two surrounding values is equal, which is the case for \( X_2 \) in Fig. 3C, the previous value is chosen. For previous

Fig. 3. A-F: The different data synchronization methods explored in this study, exemplified using dummy data. Process variables \( X_1 \) and \( X_2 \) are synchronized to \( Y \) at a target time of 5 hours, using linear, cubic spline, nearest value and previous value interpolation (A-D, respectively), and using window filtering with window placement before the target time (E) or around the target time (F). Dots represent measured values, red dots represent measured values selected as or used to calculate the synchronized value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
value interpolation, the last measured value for a process value is chosen at any target time, even when the next measurement is closer in time (Fig. 3D).

Mean and median filtering/averaging can essentially be seen as a single method with three parameters: the population estimator used to calculate the average value over the moving window (mean or median), the placement of the window with respect to the target times and the width of the window. Besides using both means and medians, placements of the window before and around the target times were applied (Fig. 3E and F, respectively) and the window width was varied from three minutes to five hours with linear increments of three minutes.

A full-factorial experimental design is used to test the effects of all three parameters. The total number of settings for synchronization using window-filtering is therefore 400: 2 population estimator options, 2 window placement options and 100 window width options. Together with linear, cubic spline, nearest value and previous value interpolation, this brings the total number of synchronizations performed and investigated to 404.

2.3. Soft-sensor approach

Soft-sensors were developed on each synchronized dataset according to the approach schematically shown in Fig. 4. The remainder of this section will explain this approach step-by-step.

Missing values are present mostly for mean and median filtering. They occur when no process values are registered for a certain target window, and have to be cleared or imputed from the data before bilinear modelling (Walczak and Massart, 2001). Each sample containing a missing value for at least one process variable was therefore removed. The datasets were individually processed, as the location for the missing values can be different for each of them. Because of this, the sample selection is no longer unified for each dataset after missing value removal (Fig. 5).

Outliers are mostly the result of system errors or the incorrect registration of the production status of the plant. They were identified in each dataset based on the Hotelling’s T² and the Q-statistic from principal component models (Varmuz and Filzmoser, 2009). Samples for which either one of these statistics was more than two standard deviations removed from the median value in the respective model were marked as outlier and were removed from the dataset. This step was also applied to each synchronized dataset individually, as the manifestation of the outliers in the data may be different for each synchronization method. All principal component models were built on autoscaled data, as process variables are measured in different units (Gurden, Westerhuis et al., 2001). For each model, as much principal components were used as were required to describe at least seventy percent of the variance in the data.

After these sample filtering steps, soft-sensors were calibrated for each synchronized dataset by regressing the mineral content on the process variables, using partial least squares (PLS) (Geladi and Kowalski, 1986). All datasets were autoscaled prior to modelling. The prediction performance of each PLS model was validated using double cross-validation, in which the inner loop was used to optimize the number of latent variables and the outer loop was used for validating the models’ prediction performances, as proposed in (Szymańska, Saccenti et al., 2012). Both loops used random 5-fold cross-validation, and the entire validation was repeated 100 times.

The synchronized datasets were cleared from missing values and outliers individually. This ensures that the full potential of the synchronization methods is exploited, and simulates the direct application of each method on the plant. However, to enrich the comparison of the synchronization methods, soft-sensors were calibrated using two different scenarios. In one scenario,
datasets are modelled as is, with each dataset having a different sample selection that was individually optimized. In the other scenario, soft-sensors are calibrated on each dataset, but after the sample selection has been unified. For this, samples that suffered from missing values or that are identified as outlier in at least one of the synchronized datasets were removed from all datasets. This scenario allows for a fundamental comparison of the synchronized values and the soft-sensors, as each soft-sensor is representing the exact same data. Only datasets synchronized using mean or median filtering with a window shorter than 15 minutes were kept out of this comparison, as including them would limit the unified sample selection to only 9 samples. The samples that are represented in each synchronized dataset are referred to as common samples. The samples that each synchronized dataset with different sample selections may have besides those common samples are referred to as additional samples.

2.4. Evaluation and comparison

The performance of the models are interpreted in terms of the Pearson correlation coefficient $r$ between predicted and reference mineral content (Figs. 7A-B, 8A-B and 9A-E). These coefficients are invariant to the scale of the dependent variable, which makes the results better comparable to those of related studies. As the validation is repeated, $r$ was averaged over all repeats and the 95% confidence interval was calculated for each of those reported average values.

To directly compare the values synchronized by each of the methods, the datasets were clustered based on the Euclidian distance between them, using average linkage. The data matrices have to have equal dimensions for this, and clustering is therefore only done datasets with unified sample selection. From the datasets corresponding to mean or median filtering, only window widths of 15, 30, 60, 120 and 300 minutes were used to ensure clarity of the results. Each dataset was autoscaled before clustering, and the clustering results are represented as a dendrogram (Fig. 6).

As mentioned in the introduction, mean and median filtering can be regarded as one method with three factors: population estimator, window placement and window width. To quantify the effect of each of these factors on the prediction performance of the eventual soft-sensor, a three-way ANOVA was performed on the validated performances of the soft-sensors calibrated on datasets synchronized using a window-based method and with unified sample selection (Table 1) (Mickey, Dunn et al., 2004). The sample size was limited in order to prevent ANOVA from identifying a factor to be of significant importance, even if the true differences between the levels of that factor are trivially small (Anderson, Burnham et al., 2000). For the window width factor, only settings of 15, 30, 60, 120 and 300 minute were included. Furthermore, the results of only 5 of the 100 validation repeats are used as replicate measurements.

3. Results and discussion

3.1. Data dimensions

The number of samples present after removal of missing values and outliers is for each synchronized data matrix is given in Fig. 5A-B. From these figures, it follows that the choice of synchronization method affects the number of mineral content samples that can eventually be used to develop a reliable soft-sensor. Most methods could successfully synchronize data for more than 80% of the mineral content samples. Methods that use a window more narrow than fifteen minutes could extract only a limited number of samples. Using these relatively narrow moving windows leads to many missing values, as some process variables might not be measured for up to fifteen minutes due to their high and inconsistent sampling intervals. For this plant, it would therefore not be advisable to use these narrow windows, as the limited number of samples exported would restrict the development of a reliable soft-sensor.

For windows wider than 15 minutes, the fraction of synchronized samples slowly decreases as the windows are further widened. This is because samples of which the process data (partly) represents non-effective production time are not synchronized. The chance of including such data increases with increasing window width. For instance, if a mineral content is measured in the first hour of operation, a window of four hours placed before that value will not yield a value while a window of half an hour placed before that value would.

In total, 95 out of 175 samples (54%) could be synchronized by all methods, except for mean and median filtering using windows more narrow than 15 minutes. This sample selection was applied to obtain the datasets with the unified sample selection scenario.

3.2. Clustering

The clustering results on the synchronized datasets with unified sample selection are represented as a dendrogram in Fig. 6. All these datasets contain 95 samples, and are thus subsets of the datasets for which the dimensions are shown in Fig. 5A-B. The results show that spline interpolation leads to the most distinct values. Linear, nearest value and previous value interpolation form a cluster, implying that these methods are more comparable to each other than to the other methods.

For the window-based methods, the choice between means and medians as population estimator seems to have the least influence on the values synchronized. Furthermore, it seems that wider windows lead to more distant datasets. This is in correspondence with the effect that wider windows incorporate less relevant data, and implies that the window width is an important parameter to set for both mean and median filtering.

For comparable window widths, data matrices generated using either the mean or median are more similar than data matrices that
used either window placement before or around the target value. This suggests that, after window width, window placement is more influential on the synchronized value than that the choice between calculating means or medians is. Remarkable is that using five-hour windows placed before the target time gives data more similar to that of spline interpolation than that of any other method.

3.3. Soft-sensor performance on unified sample selection

The performances of the soft-sensors calibrated on the synchronized datasets with unified sample selection are given in Fig. 7A-B. The performances are given in terms of Pearson correlation coefficient between predicted and measured mineral content value. The highest performance, r=0.68, is obtained when synchronizing the data using median filtering with windows of 168 minutes placed before the target times. Synchronizing using cubic spline interpolation leads to the lowest prediction performance: r=0.50. Of the synchronization methods that do not use window filtering, nearest value interpolation leads to the best prediction performance. A likely reason for this is that the synchronized values are actually measured values and not calculated ones, and that they are sampled closest to the target times as possible. Remarkable is that linear interpolation, the current benchmark method for this production plant, does not yield the highest prediction performance.

For the window-based methods, there are some relations between the model performance and the choice for population estimator, window placement and window width. For windows smaller than 180 minutes, Median filtering nearly always outperforms mean filtering, regardless of the window placement. An explanation for this is that medians are more robust population estimators than means, and are thus less sensitive to outliers. Major outliers have been removed, but minor outlying values might still be present as outlier removal was based on PCA models that did not explain all variance in the data.

Increasing the window width from 15 to 180 minutes increases performance when the window is placed before the target times. This likely caused by the fact that more of the raw data is used to calculate the synchronized values when the windows are widened. This leads to a more robust description of the state of the process in terms of process values, increasing the performance of the models.

However, widening the windows that are placed around the target value does not seem to affect the performance. A likely reason for this is that the aforementioned effect of increased robustness is countered by the fact that for these windows, process data measured after the target time is incorporated in the synchronized data. This data is less representative of that time point: changes happening in the process after a mineral content value is measured will not be related to that value. This (second) effect decreases the modelling performance.

When the windows are widened beyond 180 minutes, the performance remains roughly the same or gradually decreases. For such wide windows, data sampled long before or after the target times is incorporated. These values are less representative and will decrease the performance.

The use of median-filtering with 168-minutes placed before the target times is found optimal for this production plant. The use of medians and the placements of windows before the target time is likely to be optimal for other production plants as well, but the optimal window width will be more case-specific as it depends on the total process time of the plant and on the sampling frequencies of the variables used. To optimize the window width for a new soft-sensor, it can be incorporated as a modelling parameter in the soft-sensor calibration by testing a number of window widths and validating the performance of the subsequent sensors. In this study, 100 window widths were tested, spanning from 0.1 to 10 times the total processing time of the plant. However, as this study revealed a clear global optimum for the window width, testing only 10 windows would also suffice and would reduce the optimization time ten-fold.
3.4. ANOVA

The results of the ANOVA performed on the validated performance of models calibrated on datasets synchronized with mean or median filtering and with unified sample selection are given in Table 1. These results show that all three factors have a significant influence on the modelling result, as their p-values are lower than 0.05. The p-value for choice in population estimator is lower than that for the choice in window placement and width, indicating that the choice in population estimator has more of an influence on the soft-sensor performance than the choice in window placement and width. The only interaction factor that is found to be significant is that of window placement and window width. This is in correspondence with the results shown in Fig. 7B: increasing the width of windows placed before the target time from 15 to 180 improves the soft-sensor performance, regardless of the choice in population estimator. Increasing the width of windows placed around the target time has little effect on the soft-sensor performance. The third-level interaction between all three factors is not found to be significant.

3.5. Soft-sensor performance on different sample selections

The performances of the soft-sensors calibrated on synchronized datasets with different sample selections are given in Fig. 8A-B. Note that these datasets are the same ones for which Fig. 5A-B show the sample sizes. As for the modelling scenario using unified sample selection, synchronizing using cubic spline interpolation leads to the lowest performance (r=0.50). The best performance is obtained for the datasets synchronized using median filtering with windows of 171 minutes placed before the target times (r=0.68). This corresponds to the results found for the models on unified data shown in Fig. 7B, although the optimal window width for the models with different sample selection is slightly larger.

For the datasets obtained using mean or median filtering, the spread in performance over increasing window width seems to be larger in comparison to the results for the unified datasets (Fig. 7B). This is a direct effect of each dataset representing a different selection of samples. Increasing or decreasing the window width changes the synchronized values, but also the selection of samples which can be modelled. This also causes the large spread in results for datasets obtained using window widths below 15 minutes. Far less samples could be modelled for these datasets than for datasets obtained using wider windows, as can be seen in Fig. 5B. Based on these validated performances, it follows that the change in sample selection can have a larger impact on the modelling results than the change in values has. This underlines that the selection of samples that can eventually be modelled is an important consideration when selecting a window width.

Remarkable is that for mean filtering with windows placed before the target times, the prediction performance drops steeply when the windows are widened beyond 219 minutes. This is caused by an error in the annotation of the production status of the plant, which was only discovered after performing the dynamic synchronization and subsequent soft-sensor development. This drop in performance is not observed when medians are used instead of means, showing that median filtering is also more robust against these types of errors.

Overall, the results for the datasets with different sample selections are comparable to or lower than those for the datasets with unified sample selection. It was expected that these performances would be better, as each dataset is individually optimized in terms of sample selection and their potential for predictive modelling is fully exploited. However, the datasets with unified sample selection contain only samples that were not considered to be outlying in (nearly) all of the synchronized datasets. These samples will therefore represent the relations in the process well, and modelling them will lead to a good prediction performance. The samples that a dataset with different sample selection contains additional to the 95 common samples might not have been detected as an outlier in that specific dataset, but they have been detected as outliers in other datasets. These additional samples will represent the relations in the plant less well than the common samples, and includ-
ing them in the calibration will then not improve the performance of that model.

To confirm this, the prediction performances for the soft-sensors on datasets with different sample selection were re-calculated for the 95 common samples that are represented in all of them, and for the samples that are additional in each model separately. These results are shown in Fig. 9A-E. Note that the models were not re-validated, and that these figures represent the same validation results as Fig. 8A-B. From Fig. 9A-E, it follows that nearly all models indeed perform generally better on the common samples than on the additional samples.

4. Conclusion

In our study, different dataset synchronization methods were compared while developing a soft-sensor. The best modelling results are obtained for synchronization using median filtering with windows placed before the target times, a method that is not reported in literature related to soft-sensor development. Window width is of considerable influence on the quality of the synchronized data, and should be optimized per case. For this production facility, median filtering with windows of 168 minutes placed before the target times gave the best results. This is significantly wider than the average throughput time of the plant of 30 minutes, and indicates that changes in the process can affect the performance of the plant considerably longer than the process throughout time. In cases where sampling frequencies of the process variables are very different, the window width should be optimized per process variable individually.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Tim Offermans: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. Ewa Szymańska: Conceptualization, Methodology, Resources, Writing - review & editing, Visualization, Supervision, Project administration. Lutgarde M.C. Buydens: Supervision, Project administration, Funding acquisition. Jeroen J. Jansen: Conceptualization, Methodology, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2020.106938.

References

Anderson, D.R., Burnham, K.P., Thompson, W.L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. J. Wildl. Manag. 64 (4), 912–923.
Arteaga, F., Ferrer, A., 2002. Dealing with missing data in MSPC: several methods, different interpretations, some examples. J. Chemom. 16, 408–418.
Camacho, J., Picó, J., Ferrer, A., 2008. Bilinear modelling of batch processes. Part I: theoretical discussion. J. Chemom. 22, 299–308.
Cuetas, S., Peñaba-Niebles, R., García, E., 2016. Support vector machine in statistical process monitoring: a methodological and analytical review. Int. J. Adv. Manuf. Technol. 91, 485–500.
Ferrer, A., 2007. Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autopsy assembly process. Qual. Eng. 19 (4), 311–325.
Fortuna, L., Graziani, S., Rizzo, A., Xibilia, M.G., 2007. Soft Sensors for Monitoring and Control of Industrial Processes. Springer Science & Business Media, London.
Gabrielson, J., Jonsson, H., Trygg, J., 2006. Combining process and spectroscopic data to improve batch modelling. Am. Inst. Chem. Eng. 52 (9), 3164–6172.
Geladi, P., Kowalski, B.R., 1986. Partial least squares regression: a tutorial. Anal. Chim. Acta 185, 1–17.
González-Martínez, J.M., Vitale, R., d. Noord, O.E., Ferrer, A., 2014. Effect of synchronization on bilinear batch process modelling. Ind. Eng. Chem. Res. 53, 4339–4351.
Garden, S.P., Westerhuis, J.A., Bro, R., Smilde, A.K., 2001. A comparison of multway regression and scaling methods. Chemom. Intell. Lab. Syst. 59, 121–136.
Hall, C.A., Meyer, W.W., 1976. Optimal error bounds for cubic spline interpolation. J. Approx. Theory 16 (2), 105–122.
Holcomb, D.F., Norberg, R.E., 1955. Nuclear spin relaxation in alakl metals. Phys. Rev. 94 (4), 1074–1091.
Kassidas, A., MacGregor, J.F., Taylor, P.A., 1998. Synchronization of batch trajectories using dynamic time warping. Am. Inst. Chem. Eng. 44 (4), 864–875.
Kourt, T., MacGregor, J.F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemom. Intell. Lab. Syst. 28, 3–21.
Lin, B., Recke, B., Knudsen, J.K.H., Jørgensen, S.B., 2007. A systematic approach for soft sensor development. Comput. Chem. Eng. 31, 419–425.
MacGregor, J.F., Kourt, T., 1995. Statistical process control of multivariate processes. Control. Eng. Practice 3 (3), 403–414.
Mäge, I., Mevik, B.H., Næs, T., 2008. Regression models with process variables and parallel blocks of raw material measurements. J. Chemom. 22, 443–456.
McKinley, S. and M. Levine (1998). “Cubic spline interpolation.” Coll. Redwoods.
Mickey, R.M., Dunn, O.J., Clark, V.A., 2004. Applied Statistics: Analysis of Variance and Regression. Wiley, New York.
Neogi, D., Schlags, C.E., 1998. Multivariate statistical analysis of an emulsion batch process. Ind. Eng. Chem. Res. 37, 3971–3979.
Rimpiläinen, V., Kaipio, J.P., Depree, N., Young, B.R., Wilson, D.I., 2015. Predicting functional properties of milk powder based on manufacturing data in an industrial-scale powder plant. J. Food Eng. 153, 12–19.
Slama, F.C., 1991. Multivariate statistical analysis of data from an industrial fluidized catalytic cracking process using PCA and PLS. MSc, McMaster University.
Szymańska, E., Saccetti, E., Smilde, A.K., Westerhuis, J.A., 2012. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics 8, 3–16.
Varmuza, K., Filzmoser, P., 2009. Introduction to Multivariate Statistic Analysis in Chemometrics. CRC press, New York.
Walczak, B., Massart, D.L., 2001. Tutorial: dealing with missing data part I. Chemom. Intell. Lab. Syst. 58, 15–27.
Zamprogna, E., Barolo, M., Seborg, D.E., 2005. Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis. J. Process Control 15, 39–52.