Journal of Food Engineering 153 (2015) 12–19

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: www.elsevier.com/locate/jfoodeng

# Predicting functional properties of milk powder based on manufacturing data in an industrial-scale powder plant

Ville Rimpiläinen [a,*], Jari P. Kaipio [a], Nick Depree [b], Brent R. Young [c], David I. Wilson [d]

[a] Department of Mathematics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[b] Light Metals Research Centre, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[c] Department of Chemical and Materials Engineering, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[d] Electrical and Electronic Engineering, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand

ARTICLE INFO

*Article history:*
Received 20 October 2014
Received in revised form 5 December 2014
Accepted 8 December 2014
Available online 16 December 2014

*Keywords:*
Dairy
Milk processing
Powders
Powder functional properties
Multivariate

ABSTRACT

The fundamental science relating key physical and functional properties of milk powder to plant operating conditions is complex and largely unknown. Consequently this paper takes a data-driven approach to relate the routinely measured plant conditions to one vital function property known as sediment in an industrial-scale powder plant. Data from four consecutive production seasons was examined, and linear regression models based on a chosen set of processing variables were used to predict the sediment values. The average prediction error was well within the range of the uncertainty of the laboratory test. The models could be used to predict the effect of each individual plant variable on the sediment values which could be beneficial in quality optimisation. In addition the choice of the training data set used to compute regression coefficients was studied and the resultant regression models were compared to alternative PLS models built on the same data.



## 1. Introduction

Milk powders are widely used in the food industry as in bakery products, soups and sauces, ready meals, milk based beverages, confectioneries, milk chocolates, yoghurts and cheeses (Oldfield and Singh, 2005; Sharma et al., 2012). A key motivation to transform liquid milk into powder is to increase shelf-life and reduce transportation costs. Milk powders also possess attractive physical and functional properties. The physical properties include powder structure, particle size distribution, flowability and bulk density while the functional properties describe how the powder behaves for the customer and include reconstitution properties, heat stability and foaming properties.

The functional properties depend on the raw milk composition, the degree of standardization, the processing and subsequent storage conditions, and how the powder is used in the particular food system (Oldfield and Singh, 2005). Since some of the functional properties can be built-into the powder, there are economical interests in manufacturing such tailor-made powders due to the added value (Sharma et al., 2012).

The functional properties are usually tested by sampling the final product some time after production. However this *a posteriori* testing strategy runs the risk that an out-of-specification campaign results in considerable material being downgraded or disposed leading to significant economical losses. Furthermore, because the science currently available to explain the relations between processing conditions and functional properties of powder is immature, simply knowing that a campaign is out of spec does not help to make changes in the production.

As a solution, a data-driven approach can be used to (1) establish relations between the real-time measurable processing conditions and offline tested functional properties, (2) to predict the functional properties (in some cases in real-time) based on plant data, and (3) estimate the variance caused by each processing condition. This is sometimes known as process analytical technology (PAT) or real-time quality control (RTQ) (Bakeev, 2005; FDA, 2004; Munir et al., 2014; Munson et al., 2006; Swarbrick, 2007; van den Berg et al., 2013).

The main objective of this study was to investigate strategies to be able to predict a key functional property using only operating data measured routinely from the plant thereby making the real-time quality control possible and avoiding the necessity of a time consuming, and somewhat subjective offline laboratory analysis. The functional test is one of the many sediment tests which quantify the volume of the undissolved milk (Anon, 2014). For instant whole milk powder, the less sediment, the better, although the upper acceptable limit depends on the specific product and the customer's requirements (Sharma et al., 2012).

* Corresponding author. Tel.: +64 9 923 8818; fax: +64 9 373 7457.
*E-mail address:* ville@math.auckland.ac.nz (V. Rimpiläinen).

http://dx.doi.org/10.1016/j.jfoodeng.2014.12.010
0260-8774/

Previously, it has been found that high evaporator preheat temperatures, long holding times and high values of total solids increase the sediment values (Oldfield et al., 2000). Moreover, it has been found that the milk homogenization settings have only marginal effect and that lecithin addition decreases sediments (Oldfield et al., 2000). Furthermore, there seems to be an optimum value for the concentrate temperature which, however, may be plant-dependent (Oldfield et al., 2000; Oldfield and Singh, 2005). Regarding the composition, milk with a low fat content (1.5–26%) is preferable than high (>26%) (Kelly, 1998; Oldfield and Singh, 2005) while varying amounts of protein (24.9–30.8%) have only a marginal influence on the sediment results (Kelly, 1998).

Generally speaking, the main contribution of this paper is to describe a data-driven approach that can predict end-point properties of industrial products based on real-time manufacturing data and estimate how much variation in the properties is caused by each plant variable. More specifically, this approach is here tested and evaluated by predicting sediment properties of milk powder with the help of real-time manufacturing data from an industrial-scale milk powder plant. The predictions were regressed using typical operating variables that are routinely logged, and approximating the joint distribution as Gaussian. Moreover, it is suggested how the nominal values of the operating variables could be adjusted in order to improve (lower) the sediment values.

## 2. Theory of conditional probability distributions

The aim of this work is to predict the scalar offline measured sediment values, $s$, given a vector of possibly correlated $m$ routine plant observations, $\mathbf{D} \stackrel{\text{def}}{=} [d_1, d_2, \ldots, d_m]^{\mathrm{T}}$. In this paper, the joint distribution of the plant observations and laboratory measurements is approximated as Gaussian (normal). Thus, the model predictions are linear functions of the plant observations, and easily implemented in practice.

The mean and standard deviation of probability distribution of the sediments can be predicted if correlated information on processing variables is available. This information can be incorporated by conditioning the sediment distribution with the process data. Here, the Gaussian approximations and the theory how to calculate conditional probability distributions (CPDs) are briefly described.

The following notation is used in this section: vectors and matrices are denoted upright bold. The mean is denoted by a bar, $\bar{\mathbf{z}}$, and the model prediction by a hat, $\hat{s}$.

To start, the observation $s$ and the vector of plant data $\mathbf{D}$ are concatenated to get the augmented vector

$$\mathbf{z} = \begin{bmatrix} s \\ \mathbf{D} \end{bmatrix}. \tag{1}$$

The joint-normal approximation of $s$ and $\mathbf{D}$ has mean value $\bar{\mathbf{z}}$ and covariance matrix $\mathbf{P}$

$$\bar{\mathbf{z}} = \begin{bmatrix} \bar{s} \\ \bar{\mathbf{D}}, \end{bmatrix} \tag{2}$$

$$\mathbf{P} = \begin{bmatrix} P_s & \mathbf{P}_{sD} \\ \mathbf{P}_{Ds} & \mathbf{P}_D \end{bmatrix}. \tag{3}$$

The joint-probability distribution is of the form

$$\pi(s, \mathbf{D}) = \pi(\mathbf{z}) \propto \exp\left(-\frac{1}{2}(\mathbf{z} - \bar{\mathbf{z}})^{\mathrm{T}} \mathbf{P}^{-1} (\mathbf{z} - \bar{\mathbf{z}})\right), \tag{4}$$

and the inverse of the joint-covariance matrix can be partitioned as

$$\mathbf{P}^{-1} \stackrel{\text{def}}{=} \mathbf{B} = \begin{bmatrix} B_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{5}$$

with dimensions $B_{11} \in \Re^{1\times 1}, \mathbf{B}_{12} = \mathbf{B}_{21}^{\mathrm{T}} \in \Re^{1\times m}$ and $\mathbf{B}_{22} \in \Re^{m\times m}$.

The joint-probability distribution can be written as

$$\pi(s, \mathbf{D}) = \pi(s|\mathbf{D})\pi(\mathbf{D}), \tag{6}$$

where $\pi(s|\mathbf{D})$ is the conditional probability distribution of $s$ given $\mathbf{D}$ and $\pi(\mathbf{D})$ is the probability distribution of $\mathbf{D}$. The prediction of sediments is based on utilizing the above mentioned conditional probability distribution. It can be written in the form (Eaton, 2007),

$$\pi(s|\mathbf{D}) \propto \exp\left(-\frac{1}{2}(s - \bar{s}_{\cdot|\mathbf{D}})^{\mathrm{T}} \mathbf{P}_{\cdot|\mathbf{D}}^{-1} (s - \bar{s}_{\cdot|\mathbf{D}})\right), \tag{7}$$

where

$$\bar{s}_{\cdot|\mathbf{D}} = \bar{s} - B_{11}^{-1}\mathbf{B}_{12}(\mathbf{D} - \bar{\mathbf{D}}), \tag{8}$$

$$\mathbf{P}_{\cdot|\mathbf{D}} = B_{11}^{-1} = \frac{1}{B_{11}}. \tag{9}$$

This means that the most probable outcome of the sediment test when the processing data is known is given by (8) with variance (standard deviation squared) given by (9). The notation for the predicted sediment is $\hat{s} = \bar{s}_{\cdot|\mathbf{D}}$ and the corresponding standard deviation $\sigma = 1/\sqrt{B_{11}}$. For example, if the plant is operated with the nominal operating conditions $\bar{\mathbf{D}}$, it would yield sediment result $\hat{s} = \bar{s}$. Outside these conditions, the predicted sediment value is corrected by a value proportional to the difference $(\mathbf{D} - \bar{\mathbf{D}})$.

## 3. Materials and methods

### 3.1. Standard powder sediment tests

In this study, the stability of instant whole milk powder was the primary focus and it was quantified by a standard offline laboratory sediment measurement test. The testing procedure follows (Anon, 2014) which in turn is derived from Anon (1977). The exact nature of the test depends on the customer's requirements and for this specific industrial application are proprietary. However all the sediment tests follow the same basic procedure where a measured powder sample is mixed in controlled conditions in water (or other customer-specific solvents), and the resultant undissolved material is quantified. While the test is reasonably free from any operator subjectiveness, compared to other powder functional tests, the results tend to be severely quantified (see Fig. 5 for an example).

Due to confidentiality restrictions, the sediment values presented in this work have been normalised. Such normalisation does not have an effect on the method itself.

Sediment values from four consecutive production seasons consisting of 339, 300, 273 and 284 measured samples formed the basis for this work. Results that were further than 3.5 standard deviations away from the mean value were considered outliers and removed from the data set prior to processing.

Fig. 1 shows the histograms of the normalised sediment values for each production season and overlayed is the approximating Gaussian distributions. In addition, the average and standard deviation are noted for each season. It is immediately evident that seasons 1 and 2 are similar, as are seasons 3 and 4. The first two seasons have slightly lower sediment values than the latter, and the Gaussian approximation fit is better for seasons 3 and 4. This indicates that there has been operational changes between seasons 2 and 3.

### 3.2. Plant overview and manufacturing data

An overview of the plant layout and the different manufacturing stages are shown in Fig. 2. The raw milk from the farms is stored in mixing tanks (or silos) before passing through the preheating stage

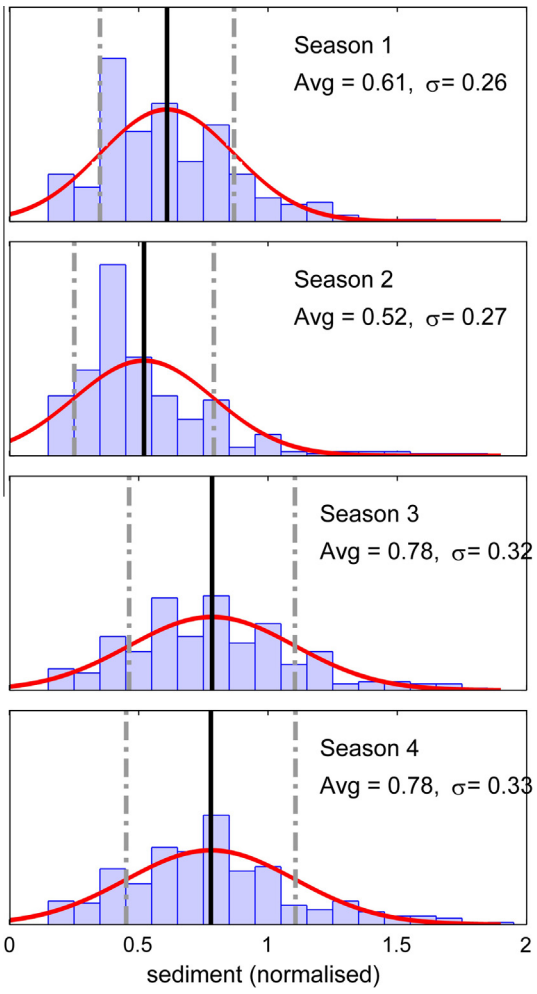


**Fig. 1.** The bars depict the histograms of the normalised sediment values for each of the studied seasons and the red lines show their Gaussian approximations. The black line shows the average value and the dashed gray lines show the one standard deviation intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and then to the evaporator. Next, the water content of the milk is reduced in a sequence of falling film evaporators. The concentrated milk is heated in order to reduce the viscosity and then homogenised. In the spray dryer, the concentrated milk is transformed into milk powder particles with the help of pressurised hot air. The remaining moisture in the powder is removed in the vibrating fluidised-bed dryers. The now powder product goes to the packing line through blending bins. The plant throughput is significant; in the tonnes/hour range. More detailed descriptions of milk powder manufacturing can be found in Baldwin and Pearce (2005).

The plant variables that were examined (see, Table 1) were chosen because they were expected to have the most influence on the key functional property of interest, *s*. These variables were chosen based on expert domain knowledge of the plant engineers, or because they have been explicitly mentioned in the literature.

The direct contact heater and steam injector temperatures are related to the preheating stage. The preheating temperatures, total solids (measured prior to the concentrate tanks) and concentrate temperature are considered important by Oldfield et al. (2000) and Oldfield and Singh (2005). The flow and density of milk correspond to the material that is fed to the evaporator chain. Separator temperature and total solids are related to the material that comes out from the evaporators. The homogenization stage was characterised with the help of pressure, and the spray drying stage with the help of the dryer pressure and two dryer temperatures. The last operating conditions that were considered here were the vibrating fluidised-bed pressure and temperature.

In this article, only the real-time measured processing conditions were used as predictors for the sediment values. The effects of varying composition and standardization of the milk were considered negligible. The storage conditions were not taken into account since the tests were carried out immediately after manufacturing. Moreover, the testing conditions were standardized, therefore those were not considered as variables in the analysis.

The way the sediment values were synchronised with the corresponding values of the plant variables is illustrated in Fig. 3. First, the time lags that occur during the start-up and shut-down of the plant were omitted. Then, the first and the last production unit that entered the packing line were placed at the ends of the remaining period of time. Production data was selected based on the unit labels that were tested for the powder functionals. The values used for **D** in Section 2 were the averages of the selected plant variables over 10 min spanning around the time point that corresponded the approximated manufacturing time of the lab tested milk powder unit. It is important to note that the laboratory measurements are not regularly sampled, nor is the production continuous throughout the season given that the evaporators and dryers need to be regularly cleaned. However given the coarseness of the sampling, and the length of the season, the sampling is adequately approximated as uniform.

## 4. Results and discussion

### 4.1. Predicted sediments

A prediction model based on conditional probability distributions (CPDs) was constructed following the scheme given in Section 2. In each case, the first 75 sediment samples were used as a training data set. Then the model was used to predict only the next sediment value after which the model was updated. In other words, the model was updated after every new sediment test result.

Fig. 4 shows the prediction results that were calculated each season separately while Fig. 5 concatenates all the seasons. For each case, the average prediction error, $\bar{\epsilon}$, was calculated from the absolute differences between the predicted, $\hat{s}(t_i)$, and the actual measured sediment values, $s(t_i)$ at time $t_i$

$$\bar{\epsilon} = \frac{1}{N}\sum_{i=1}^{N}|\hat{s}(t_i) - s(t_i)|, \tag{10}$$

where $N$ is the number of predicted sediment samples. Table 2 summarises the statistics of the predictions. The fourth column gives the smallest and the highest standard deviation that were estimated with the help of Eq. (9). The last column gives the average prediction error based on Eq. (10).

For seasons 1, 3 and 4, the proportion of the sediment predictions lying within $1\sigma$ is consistent with the Gaussian assumption given that one would then expect 68%. The high percentage of season 2 (77.3%) can be explained by the fact that the Gaussian approximation for the sediments did not fit very well for this season (as indicated in Fig. 1). For instance, the mode of the distribution is located at 0.4 which is somewhat lower than the average value (0.52) of the approximated Gaussian distribution.

The values of average prediction errors were close to each other; the values for seasons 3–4 were slightly higher. The estimated standard deviations were the smallest for season 1 and highest for season 4.

Fig. 5 shows the result when all the seasons were considered together. The black vertical lines show points where the previous season ends and new begins. Here the average prediction error

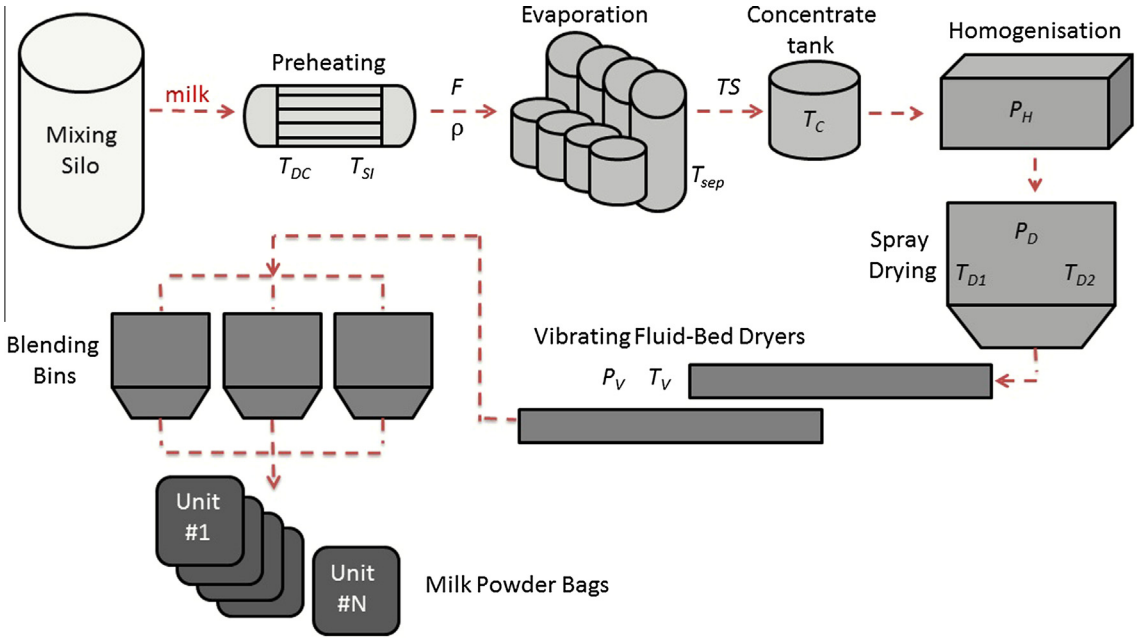


**Fig. 2.** Overview of the milk powder production showing the position of the plant measurements given in Table 1.

**Table 1**
The plant variables that were used for the prediction of *s*.

| Variable | Description |
|---|---|
| $T_{DC}$ | Direct contact heater temperature |
| $T_{SI}$ | Steam injector temperature |
| $F$ | Milk flow |
| $\rho$ | Milk density |
| $T_{sep}$ | Separator temperature |
| TS | Milk total solids |
| $T_C$ | Milk concentrate temperature |
| $P_H$ | Homogenisation pressure |
| $P_D$ | Dryer pressure |
| $T_{D1}$ | Dryer temperature 1 |
| $T_{D2}$ | Dryer temperature 2 |
| $P_V$ | Vibrating fluidised-bed pressure |
| $T_V$ | Vibrating fluidised-bed temperature |

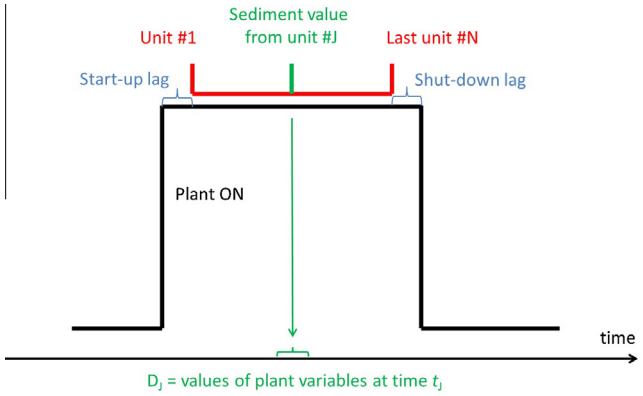


**Fig. 3.** The diagram shows how the sediment values were synchronised with the values of the processing variables.

was 0.22 units. It can be seen that there is a sudden jump when the season 3 begins. The predictions seem to follow this jump even though the highest sediment values were not captured by the model.

### 4.2. Controlling sediment with plant variables

In Section 3.1, it was highlighted that the seasonal averages and the shapes of the distributions of the sediments are different between seasons 1–2 and 3–4. In Fig. 5, it was noticed that there is a jump in both the predicted and measured sediment values when season 3 begins. This can be explained by the CPD approach.

In Section 2, it was stated that the expected sediment value based on the known production data is given by

$$\hat{s} = \bar{s} - B_{11}^{-1}\mathbf{B}_{12}(\mathbf{D} - \bar{\mathbf{D}}), \tag{11}$$

which can be re-written as a deviation

$$\hat{s} - \bar{s} = -B_{11}^{-1}\mathbf{B}_{12}(\mathbf{D} - \bar{\mathbf{D}}), \tag{12}$$

$$\Delta\hat{s} = \mathbf{C}\Delta\mathbf{D}. \tag{13}$$

Eq. (13) gives the sensitivity of the sediment as a function of the deviations of the plant variables from the nominal case. In order to estimate how much normal variation in the operating conditions can change the sediment values, Table 3 shows the predicted $\Delta\hat{s}$ when each plant variable one by one is increased with $+1 \times \mathrm{STD}(D_i)$ from $\bar{D}_i$. The values in **C** were taken from the calculation of the last predicted point.

Based on Table 3, one could control the sediment values in the following way. In season 1, the easiest way to decrease the sediments would be to increase both the dryer temperatures ($T_{D1}$ and $T_{D2}$). (These are denoted with a ⊞ symbol in Table 3.) In season 2, increasing the direct contact temperature ($T_{DC}$) would decrease the sediments the most. In season 3, one should decrease steam injector temperature, $T_{SI}$, (denoted with a ⊟), and increase $T_{DC}$. In season 4, decreasing $T_{SI}$ would work the best. When all the seasons are considered together (the last column of Table 3), $T_{SI}$ and the concentrate temperature $T_C$ would have the highest influence on sediments.

Increase in spray dryer temperatures $T_{D1}$ and $T_{D2}$ means hotter air and thus faster drying and faster particle formation. According to (Sharma et al., 2012) this can result in the hardening of the powder particles and lower bulk density. An increase in $T_{DC}$ could

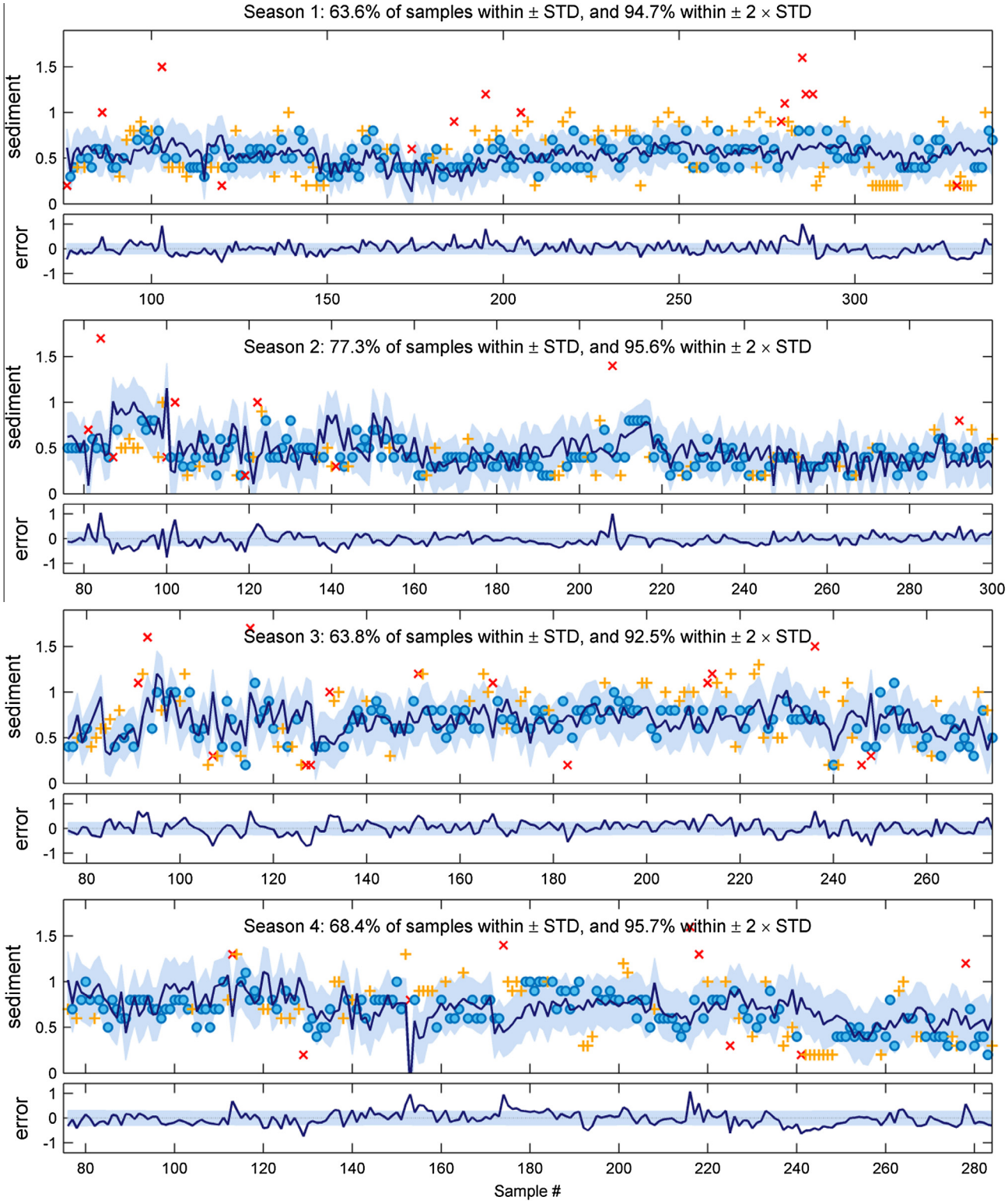


**Fig. 4.** Prediction results for seasons 1–4. The upper trend pictures show the predicted sediment values with a solid blue line and the predicted one standard deviation ($\sigma$) limits with a light blue color. The symbols ●, +, × are for the actual measured values that lie within $\pm\sigma, \pm 2\sigma$ and outside $2\sigma$ prediction intervals, respectively. The lower trend pictures show the error between the predicted and measured value with a solid blue line and the corresponding one standard deviation limit with a light blue color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reduce the temperature gradient within the preheating stage (since the material goes from the direct contact heater to the steam injector) and thus reduce the amount in sediments. Decreasing $T_{SI}$ in the preheating should be beneficial since it is known that high temperatures in the preheating stage can worsen the sediments (Oldfield et al., 2000).

Interestingly, $T_C$ had only small seasonal values in Table 3 even though it should have significance based on literature. This was

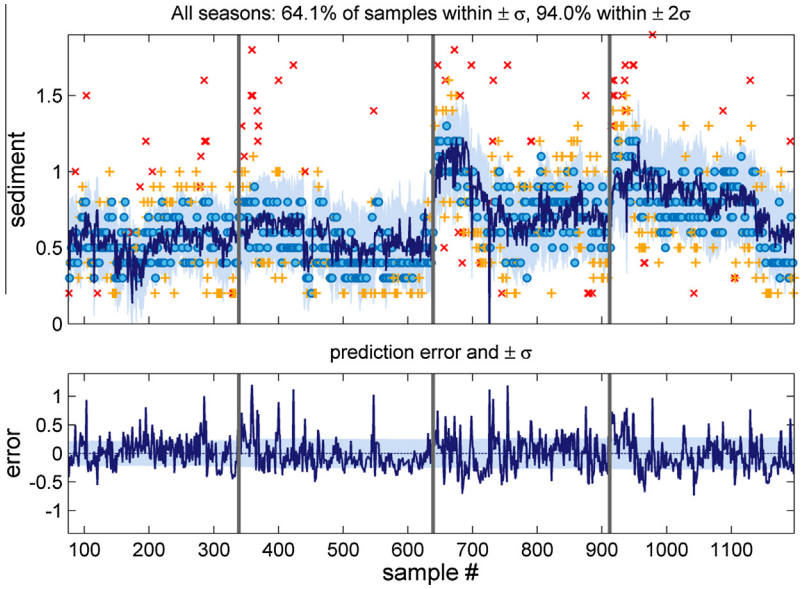


**Fig. 5.** Prediction results (solid blue line) and the actual measured values (●, +, ×) when all the seasons were considered together. Symbols are the same as for Fig. 4. The black vertical lines show when the previous season ends and new begins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Statistics related to the predictions. The first four rows correspond to the prediction results shown in Fig. 4 and the last row to the results shown in Fig. 5.

| Season | Values within $1\sigma$ (%) | Values within $2\sigma$ (%) | Range of $\sigma$ | $\bar{\epsilon}$ Eq. (10) |
|---|---|---|---|---|
| 1 | 63.6 | 94.7 | 0.20–0.23 | 0.19 |
| 2 | 77.3 | 95.6 | 0.23–0.28 | 0.18 |
| 3 | 63.8 | 92.5 | 0.23–0.27 | 0.22 |
| 4 | 68.4 | 95.7 | 0.27–0.31 | 0.22 |
| 1–4 | 64.1 | 94.0 | 0.20–0.27 | 0.22 |

**Table 3**
The predicted change of sediments $\Delta\hat{s}$ (in normalised units) if the value of each plant variable one by one was increased by the amount of one standard deviation from its nominal value. Variables that have the most influence on decreasing the sediment level are denoted ⊞ (variable to be increased) and ⊟ (to be decreased).

| Variable | Season 1 | Season 2 | Season 3 | Season 4 | All seasons |
|---|---|---|---|---|---|
| $T_{DC}$ | 0.003 | −0.107 ⊞ | −0.105 ⊞ | −0.003 | −0.019 |
| $T_{SI}$ | 0.022 | 0.059 | 0.139 ⊟ | 0.093 ⊟ | 0.081 ⊟ |
| $F$ | −0.037 | 0.017 | 0.038 | 0.044 | 0.019 |
| $\rho$ | −0.007 | −0.018 | −0.002 | 0.012 | 0.009 |
| $T_{sep}$ | −0.010 | 0.013 | 0.076 | −0.064 | −0.014 |
| TS | 0.012 | −0.014 | −0.011 | −0.023 | −0.010 |
| $T_C$ | 0.006 | 0.003 | 0.028 | 0.007 | −0.085 ⊞ |
| $P_H$ | −0.008 | 0.034 | 0.018 | −0.048 | −0.019 |
| $P_D$ | −0.015 | 0.039 | 0.001 | −0.057 | 0.017 |
| $T_{D1}$ | −0.081 ⊞ | −0.011 | 0.073 | 0.011 | −0.020 |
| $T_{D2}$ | −0.060 ⊞ | −0.053 | 0.021 | 0.000 | −0.069 |
| $P_V$ | −0.010 | 0.008 | 0.033 | 0.057 | 0.029 |
| $T_V$ | −0.031 | 0.056 | −0.061 | −0.016 | −0.003 |

due to the fact that $T_C$ was kept almost constant during every season: the predictions were based on the utilization of the joint-covariance matrix and the joint covariance matrix can capture correlations between variables only if there is sufficient statistical variation in the variables. However, $T_C$ shows up when all the seasons are considered together and this is because the set-point was decreased by 3 °C between seasons 2 and 3. So based on Table 3, the higher $T_C$ values resulted in better sediment test results. It was mentioned by Oldfield et al. (2000) and Oldfield and Singh (2005) that there should be a plant-specific optimum nominal value for this temperature.

It was expected that total solids, TS, should have a strong effect on the sediment values based on Oldfield et al. (2000), but in fact Table 3 shows that the effect is minimal. This was because the set-point of TS was tightly controlled at the plant by the operators, and there was not enough variation in its values to create correlations to sediments. In fact, it was noticed that the average error was slightly smaller when TS was removed from the prediction model. However, it is not recommended to exclude this variable in case TS happens to show more variation in the future.

Furthermore, it should be noted, that one needs to be careful when extrapolating the results in Table 3, since those are linear approximations that are the most accurate when the plant variables are close to their nominal operating values.

### 4.3. Comparison between updated and fixed training data sets

In this section, it is studied how the selection of the training data set of the prediction model influences the average error. The training set refers to the data that is used to calculate the average values and the covariances in Eqs. (8) and (9). The usual problem when building a prediction model is to choose the training set in such way that it is representative for the *future* observations. This can be challenging for a number of reasons. For example, there is often maintenance work that is carried out during the off-season which can change the correlations between plant variables and sediments. Therefore, historical data from previous production seasons may not be representative for the future observations. These correlation changes can be due to new processing equipment, improvements in the production line, new measurement devices or adjustments in the controlling devices, for example. Finally set-point changes may stimulate possible nonlinearities that can alter the nature of the correlations.

Since data from different production seasons may not be comparable, it makes sense to choose the training set from the data of the current production season; however, the nominal operation setting of the plant may also change or vary within the production-season which means that even the data of the on-going production season may not be representative. To illustrate this, Fig. 6 shows the average prediction error, $\bar{\epsilon}$, when different numbers of data

points from the beginning of the seasons were used as training sets and the rest of the sediments were predicted with this model. The number of the data points in these fixed training sets varied from 75 to 175.

In our case, the issue with the varying plant operation within the production season was solved by updating the training set after every new data point. Fig. 6 also shows the average error when the training set was updated continuously (black solid line). As can be seen, the average prediction error for the updated training set is lower or equal for every case.

Fig. 7 shows the average prediction error when the training set is formed with data from previous seasons. It can be seen that when the training set consists of three consecutive seasons, the average prediction error for the fourth season is close to the error of the updated training set scheme (black line). These results show that the updated training set scheme always gives errors equal to, or lower than the average error of the fixed training set.

### 4.4. Comparison between CPD and PLS predictions

Partial least squares (PLS) (sometimes also known as projection to latent structures) has become popular as a prediction method in both academy and industry (Vinzi et al., 2010; Wold et al., 2001). Due to this popularity, and because PLS can be considered as a distribution-free method (Dijkstra, 2010), it is here compared with the CPD approach which gives the best linear estimate with respect to the mean square criterion (Melsa and Cohn, 1978). Here, the training set of both CPD and PLS were updated after every new sediment value and in the PLS prediction, different numbers of PLS components were used.

Fig. 8 shows how the average prediction error changes as a function of the number of PLS components used in the prediction for each season. As a reference, the average error of the CPD approach is shown with solid black line. It is immediately evident that the error levels are comparable between the two schemes, and that at least 4 PLS components are required to match the CPD accuracy. With the exception of season 2, the CPD approach matches, or betters, the PLS prediction accuracy. The final trend in Fig. 8 shows the case when all the seasons were considered together. In this case, the CPD approach works better than PLS.

It is worth noting that the difference between the average error values of the methods was zero (up to numerical accuracy) when all 13 PLS components were used. (The dimension of **D** is 13.) This is because both PLS and CPD approach are based on the utilization of the joint-covariance matrix (Dijkstra, 2010; Wehrens, 2011). Therefore, when exactly the same data is used, the predicted sediments are also the same.

In some cases when less than full amount of PLS components are used, PLS delivers more accurate results. This is because PLS is designed to maximize the covariance between the predictor and response variables and to omit the less-correlated (or uncorrelated) data (Dijkstra, 2010; Wehrens, 2011). In other words, adding more PLS components does not necessarily mean that one adds more useful data for the prediction. Based on Fig. 8, the average prediction error is the lowest for seasons 1–4 when 4, 6, 5 and 7 PLS components respectively, are used for prediction.

The reasons why PLS was not finally adopted here were that, first, there was not a significant improvement in the prediction accuracy and, second, the CPD approach offered a convenient way to evaluate the influence of each and every predictor variable on the sediments as outlined in Section 4.2.

### 4.5. Discussion and future work

**Gaussian approximation:** It can be argued how well the Gaussian approximation fits for both the sediment data and the processing data. Nevertheless, the used method can be simply considered as the best linear estimate of the sediments that corre-

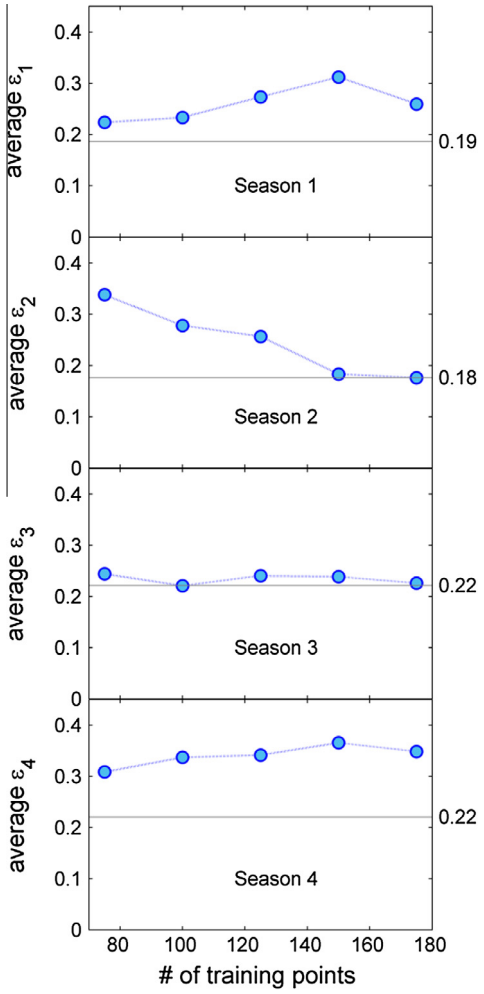


**Fig. 6.** The average of the prediction error is shown with blue circles for each season when different amounts of data points were used in the training set of the prediction model. For reference, the average prediction errors when the updated training set scheme was used are shown with solid black lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

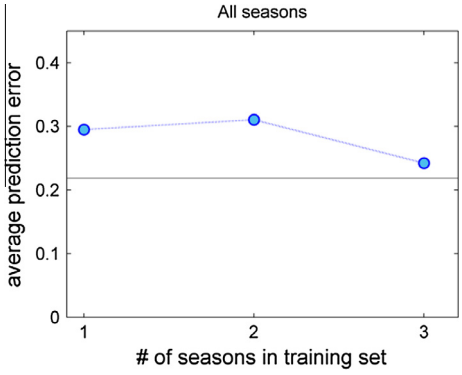


**Fig. 7.** The average of the prediction error is shown with blue circles when all the four seasons were considered one after the other. Here the training set was constructed using all the sediments of different amounts of previous seasons. For reference, the average prediction error when the updated training set scheme was used is shown with solid black line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

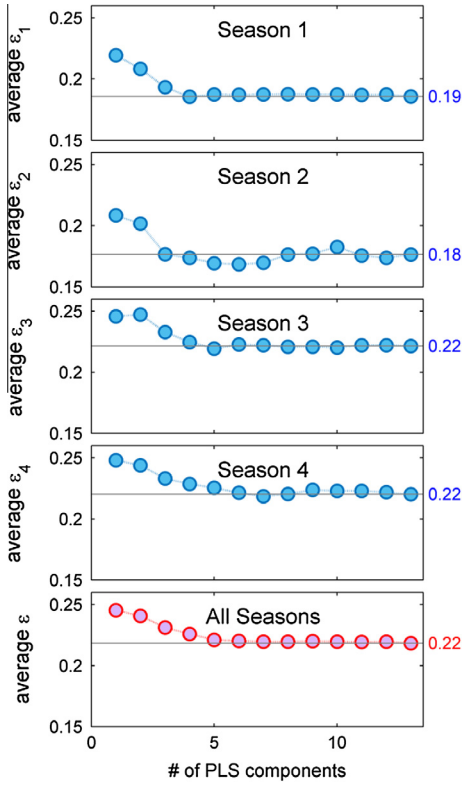


**Fig. 8.** The average of the prediction errors as a function of the number of PLS components used in the prediction model are shown with blue circles for seasons 1–4. The red circles show the case when all seasons were considered together. For reference, the average prediction error when CPD was used is shown with solid black lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sponds to Gaussian approximations. Other types of distributions might work better in some cases; however, it was shown that PLS, which can be considered as a distribution-free method (Dijkstra, 2010), gave similar accuracy as the CPD approach when average prediction errors were compared.

**Plant data:** In this work a relatively simple method was used to synchronise in time the sediment measurements to the plant measurements. Conceivably a more accurate method based on the flow times of the material through the plant could be used to exactly match the sample with the plant conditions. However due to the large flowrates, and the relatively small powder sample that is tested, this synchronisation error is likely to be negligible.

**Excluded variables:** The composition of milk and the concentrations of additives such as lecithin were not considered in the computations because of their marginal variations. However, if these variables are expected to show more noticeable variations, the real-time measurable milk content data can be easily included in the prediction model in the future.

## 5. Conclusions

The ease at which instant whole milk powder dissolves is quantified by a sediment test and is an important functional property. This time-consuming offline laboratory sediment test can be circumvented by employing data-driven models that utilize measurements that are routinely measured from the plant.

The approach developed in this paper gives linear models that work the best when the plant operates close to the nominal operating values. The approach was tested with data from four consecutive production seasons and validated against the laboratory results, and it was found that the average prediction errors were around 0.2 normalised units which is deemed adequate by the Dairy Industry. Moreover, it was shown that the prediction models can be used to find suggestions how to adjust plant variables in order to improve the sediment results.

The presented data-driven approach makes real-time quality control of the functional properties of milk powder possible. Since this can significantly reduce downgrades and losses, the economical benefits can be substantial. Finally, it is worth emphasizing that the described prediction approach is suitable also in other industrial cases, not just in milk powder manufacturing.

## Acknowledgements

The authors would like to thank Fonterra's Advanced Process Control group for the collaboration. This research was supported by University of Auckland, Faculty of Science (FRDF Project 9844/3700389) and the Primary Growth Partnership (PGP) scheme sponsored by Fonterra Co-operative Group and the New Zealand Government.

## References

Anon, 1977. ADMI Solubility Index. Technical Report 16.1. Ministry of Agriculture and Fisheries Dairy Division. Wellington, New Zealand.
Anon, 2014. GEA Niro analytical methods. Technical Report. GEA Niro. Gladsaxevej 305, PO Box 45, DK 2860 Soeborg, Denmark. <www.niro.com/methods>.
Bakeev, K.A. (Ed.), 2005. Process Analytical Technology. Blackwell Publishing.
Baldwin, A., Pearce, D., 2005. Milk powder. In: Onwulata, C. (Ed.), Encapsulated and Powdered Foods. CRC Press.
Dijkstra, T.K., 2010. Latent variables and indices: Herman Wold's basic design and partial least squares. In: Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications. Springer.
Eaton, M.L., 2007. Multivariate Statistics: A Vector Space Approach. Institute of Mathematical Statistics.
FDA, 2004. PAT – a framework for innovative pharmaceutical development, manufacturing, and quality assurance. Technical Report. U.S. Department of Health and Human Services, Food and Drug Administration.
Kelly, P.M., 1998. Coffee-stability of dried creamers. Technical Report. Dairy Products Research Centre (Moorepark, Fermoy, Co. Cork, Ireland). <http://www.teagasc.ie/research/reports/dairyproduction/4341/eopr-4341.pdf>.
Melsa, J., Cohn, D., 1978. Decision and Estimation Theory. McGraw-Hill.
Munir, M.T., Yu, W., Young, B.R., Wilson, D.I., 2014. The potential for process analytical technologies in the dairy industry. Trends Food Sci. Technol., submitted for publication.
Munson, J.C., Freeman Stanfield, C., Gujral, B., 2006. A review of process analytical technology (PAT) in the U.S. pharmaceutical industry. Curr. Pharm. Anal. 2, 405–414.
Oldfield, D., Singh, H., 2005. Functional properties of milk powder. In: Onwulata, C. (Ed.), Encapsulated and Powdered Foods. CRC Press.
Oldfield, D.J., Teehan, C.M., Kelly, P.M., 2000. The effect of preheat treatment and other process parameters on the coffee stability of instant whole milk powder. Int. Dairy J. 10, 659–667.
Sharma, A., Jana, A.H., Chavan, R.S., 2012. Functionality of milk powders and milk-based powders for end use applications – a review. Compr. Rev. Food Sci. Food Safe. 11, 518–528.
Swarbrick, B., 2007. Process analytical technology: a strategy for keeping manufacturing viable in Australia. Vib. Spectrosc. 44, 171–178.
van den Berg, F., Lyndgaard, C.B., Sørensen, K.M., Engelsen, S.B., 2013. Process analytical technology in the food industry. Trends Food Sci. Technol. 31, 27–35.
Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), 2010. Handbook of Partial Least Squares: Concepts, Methods and Applications. Springer.
Wehrens, R., 2011. Multivariate regression. In: Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Science. Springer.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58, 109–130.