Statistical Data Validation Methods for Large Cheese Plant Database

S. A. Jimenez-Marquez,* C. Lacroix,* and J. Thibault†

*Dairy Research Centre STELA, Pavillon Paul-Comtois, Université Laval, Québec, QC, Canada G1K 7P4

†Department of Chemical Engineering, 161 Louis Pasteur, University of Ottawa, Ottawa, ON, Canada K1N 6N5

ABSTRACT

Production data of the cheesemaking process are used to monitor milk fat and protein recoveries in cheese, cheese yield, and composition and eventually to predict these parameters. Due to the large impact of these factors on cheese quality and plant profitability, it is very important to use reliable data for analysis, modeling, and control of the process. This paper tested six methods for detecting erroneous data in industrial cheesemaking databases. The data analyzed came from 4 yr of stirred-curd Cheddar cheese production in an industrial cheesemaking facility, comprising over 10,000 vats. Single vat outliers were detected using a simple statistical criterion of $\bar{x} \pm 3.6$ SD on single variable distributions, Fourier series modeling of seasonal variables (fat, protein, lactose, and total solids in milk, and protein in whey), and the multivariate Mahalanobis outlier analysis. Detection of outlier productions (corresponding to several vats) was done by applying the $\bar{x} \pm 3.6$ SD criterion to variables obtained through calculating the fat mass balance, fat retention coefficient, and yield efficiency. Data treatment enabled the detection of outlier data, but also pinpointed variables with a low reliability (manually registered times). Single variable and multivariable methods proved complementary, and the use of both types of methods is recommended when validating an existing database.

(Key words: statistical cheese data validation, seasonal variation, fat mass balance, yield efficiency)

Abbreviation key: $E_{\text{act-th}}$ = yield efficiency based on actual yield and theoretical yield, $E_{\text{adj-th}}$ = yield efficiency based on adjusted actual yield and theoretical yield, **FMB** = fat mass balance, K_f = fat retention coefficient, **MAE** = mean absolute error, **P1** or **P2** = first or second principal component of transformed mass of fat in milk and that in cheese and whey, ΔR = percent

decrease of the range of the data, **S-CM** = starter and its culture medium, ΔSD = percent decrease of the standard deviation of the data, **SD/R** = ratio of the standard deviation to the range expressed in percentage, **SSC** = simple statistical criterion, Y_{act} = actual yield, Y_{VSPmod} = modified Van Slyke & Price yield formula, Y_{adj} = adjusted actual yield, Y_{th} = theoretical yield.

INTRODUCTION

Cheesemaking factories record a large quantity of production data that is used to monitor process efficiency. Among these parameters are milk, whey, and cheese compositions; milk volumes and cheese weights; starter quality and activity; processing times; temperatures; and various pH. The registered values can then be used to monitor fat and protein recoveries in cheese, as well as cheese yield, and, in some cases, to predict these production parameters based on past performance (Lawrence, 1993; Lolkema, 1993; Emmons and Lacroix, 2000). However, despite the expense and effort, most of the times, many of these data are only collected, but not used, for process analysis and control.

When records of cheese are studied, for example, to identify the relationship between final cheese quality or cheese yields and process conditions, it is necessary to have a representative, valid, and complete database for important parameters for cheese manufacture (Lacroix et al., 1993; Hunter et al., 1997; Paquet et al., 2000). However, it is well known that industrial registers usually contain many erroneous data. These errors have many sources, such as human error in registering data, nonrepresentative samples for analysis, and faulty equipment calibration (Lacroix et al., 1993, 2000). Obtaining reliable data can be even more challenging when the cheesemaking operation is semi-continuous versus strictly batch, such as in the use of a continuous draining step after batch processing of incoming milk. In the continuous whey drainage and cheddaring process, curd from one cheese vat is sometimes mixed with that from other vats of the same

Received March 30, 2001.

Accepted March 22, 2002.

Corresponding author: C. Lacroix; e-mail: christophe.lacroix@al.n.ulaval.ca.

production, or even sometimes of other productions. It is therefore difficult to accurately estimate the weight of cheese, cheese yield, or efficiency for a specific vat, or to use mass balance calculations to verify the validity of recorded data for a cheese vat (Paquet et al., 2000).

Several studies have used industrial data or samples to analyze the performance and level of control of cheese manufacture. Various cheese yield formulas from milk composition have been developed with different approaches and reviewed by Emmons et al. (1990). Measuring and modeling of seasonal variations of milk constituents (Ng-Kwai-Hang et al., 1987; Barbano, 1990; Lacroix et al., 1996), control of industrial production (Lawrence and Johnston, 1993; Lolkema, 1993), and modeling of cheese attributes with artificial neural networks (Bos et al., 1992; Ni and Gunasekaran, 1998; Paquet et al., 2000) using industrial databases have also been reported. However, in many of these studies, which used data from large databases, there was no explanation of how data were filtered or reconciled to ensure the reliability of the subsequent analyses. For example, Bos et al. (1992) used 145 data points from actual plant productions described by 24 composition and process variables to model and predict cheese moisture using neural networks. However, no information was provided on how those specific data points were selected over all other productions in the plant where the data originated. Recently, Paquet et al. (2000) used an extensive industrial database to develop a neural network model for the pH of cheese curd at various stages during the cheesemaking process. A total of 6258 cheese vats from over 3 yr of Cheddar cheese production in a large cheese plant were reduced to just over 1800 vats by applying a confidence interval defined by the mean and standard deviation to subsets of single variables. However, no multivariable methods, such as mass balances, were carried out for validating the data.

The objective of this investigation was to compare several techniques for outlier detection in a large cheese plant database. Single and multivariable approaches were used to analyze single cheese vats, as well as total productions (groups of contiguous vats). The techniques tested were meant to find outlier values that would normally not be found by the simple inspection of the data and were intended to detect both erroneous, as well as nonrepresentative data points. This would render the data more reliable for subsequent multivariable modeling of the cheese process, in addition to usual performance analyses such as yield efficiency and fat retention.

MATERIALS AND METHODS

The Process and Data

The cheesemaking data were obtained from a large capacity Cheddar cheese plant with a continuous drain-

ing system, processing about 1 million L of milk per day (Agropur, Granby, PQ, Canada). The coagulation was carried out sequentially in eight large-capacity (22,000 L) milk vats, in which milk was ripened, coagulated, and the resulting coagulum was cut, stirred, and cooked. The curd was then pumped onto a continuous conveyor system where it was drained, stirred, and dry-salted. The curd was transferred in 220-kg hoops and vacuum-pressed.

Data used in this study correspond to 4 yr of commercial production of stirred-curd Cheddar cheese and include over 10,600 vats grouped in 504 productions for three different recipes. Fat in milk was standardized for only one recipe. All composition, pH, and acidity measures were made after sampling, using methods commonly used in industry (i.e., infrared analysis for milk components and cheese moisture and fat, electrode probe for pH measurements, and phenolphthalein titration for acidities), and good laboratory practices. Most weights and volumes, as well as all flow rates and temperatures, were measured by in-line sensors and registered automatically, while time variables were recorded either automatically or manually. Over 150 parameters recorded for each vat were reduced to 61 useful process parameters via selection and transformation. Elimination of a given variable from the database was based on knowledge of the registered values based on: 1) sampling procedures, 2) measuring equipment and its usage in the plant, 3) recording method (manual vs. automated), and 4) proportion of missing data. Parameters having a constant value, presenting very little variation, or missing more than 50% data, were not retained.

Preliminary Variable Classification and Data Inspection

To later subdivide the data into smaller subsets for analysis, all quantitative variables were classified into four groups according to their variation (or lack of) with respect to three selected variables: 1) cheese recipe (e.g., cooking time, protein in whey, and moisture in cheese), 2) use of standardized or nonstandardized milk (e.g., fat, protein, lactose, and total solids in milk), 3) starter and its culture medium (**S-CM**) (e.g., starter activity measures, and starter production times), and 4) no variation (e.g., calcium chloride quantity, rennet, and milk volumes, and temperature at rennet addition). Classification of a process variable into a given group was decided based on knowledge about the variations of the variable at the plant, and complemented with the results of an ANOVA on the means of the subdivisions of data (subsets) of the process variable, corresponding to each value (level) of the classification variable. When the ANOVA was not valid or gave a level of significance

lower than 70% ($P > 0.3$), the classification decision was based entirely on knowledge about the variable. Some variables were classified into more than one group.

An initial visual inspection of the data for each variable was carried out to eliminate abnormal or extreme data (e.g., negative time spans, abnormal milk volume per vat, composition values of zero), to minimize their impact on subsequent calculations and give the validation approaches greater outlier detection sensitivity. After this initial inspection, the filtered database was then validated by the outlier detection approaches tested.

Validation Methods

Simple statistical criterion (SSC). This method consisted of establishing upper and lower limits for data acceptance defined by a probability confidence interval based on the mean ($\bar{x}$) and SD of a given dataset. The application of this method consisted in calculating the mean and SD of the data subset, and eliminating any values outside of the confidence interval. This procedure will be referred to as validation pass and was repeated until no more data points were found outside the confidence interval. The confidence interval was arbitrarily defined as the mean plus or minus 3.6 times SD of the set of values ($\bar{x} \pm 3.6 \text{ SD}$) after observing that the number of eliminated values increased considerably when a narrower confidence interval was tested on some of the variables. The selected criterion excludes 1 in every 3143 data points, or 0.032% of data, in a set of normally distributed values. Although data may not be normally distributed, this method was selected because of its common usage and simplicity.

This method was the basis for most validation approaches in this research, and in its most simple form, it was applied to data subsets of single variables. For the other approaches (described below), it was applied to new variables obtained by transforming the original ones or by combining several of them.

Fourier series modeling and SSC (Fourier-SSC). This method consisted in modeling the natural seasonal variations of variables using Fourier series in order to calculate the seasonal means and then to apply the SSC outlier criterion on the model residuals instead of the actual variable values. This approach was mainly aimed at milk constituents, but Fourier models were also tested on other variables to detect possible seasonal variations.

Data for each variable were divided into two subsets. One corresponded to vats that used nonstandardized milk, and the other to those that used standardized milk. For each subset, the mean for every production (a group contiguous of vats of the same recipe produced

without interruption) was calculated. These production means were then used to fit different Fourier series models to test for significant seasonal effects, using the general form of a Fourier series equation (Lacroix et al., 1996):

$$y(t) = a_0 + \sum_{i=1}^n a_i \cos(i\omega t) + b_i \sin(i\omega t) \quad [1]$$

where y is the modeled variable, n the model order, a_0 the model constant, a_i and b_i the model coefficients, t the time unit, and ω the angular velocity. For a full cycle of 2π radians and t expressed in days, $\omega = 2\pi/365$. The parameter ωt represents the position in the cycle (position of the day of production in the yearly cycle) of the data point. The production date was transformed to the day of the year starting with January 1 as d 0.

The least squares fitting method was used to adjust the different Fourier models for each variable. Starting with a first-order ($n = 1$) Fourier series, the model order (n) was progressively increased, to a maximum of 4 or until one or both terms (sine and cosine) were not significant ($P > 0.05$). Only models with determination coefficients (r^2) above 0.25 were retained. Model residuals were calculated by subtracting the seasonal means estimated by the corresponding Fourier model from the measured values. Outlier data were then found by applying the SSC on the model residuals.

Principal component analysis and SSC. This method consisted in transforming a pair of variables into principal components and then applying the SSC on the values of the second principal component (P2). This method was used with the mass of incoming and outgoing fat (in milk and in cheese and whey, respectively), because both variables contained an unknown and equivalent amount of error (errors of sampling, analysis, and measurement of weight and volume), and it was judged that neither one could be expressed in terms of the other. Principal component transformation gave both variables the same importance. The first principal component (P1) gave the major source of variation (tendency) and the deviation from this main tendency was represented by the second principal component (P2), which was used for outlier detection.

Mahalanobis outlier analysis. This method consisted in finding outlier vats within a data subset of a group of variables (hereafter referred to as variable group subset, Figure 1). For this method, a data point consisted of all the values of the variables selected to describe a single cheese vat. For each variable group subset of v variables and n data points (vats), a centroid was calculated in a v -dimensional space. The v coordinates (variable values) for each data point were then transformed to a single absolute value, equivalent to

Figure 1. Schematic example of data subdivision for single variables and groups of variables. The variables exhibited variations with respect to the recipe, their values were therefore subdivided per recipe for each variable and as a group, yielding single variable subsets and variable group subsets. Several subsets of single variables constitute a variable group subset. A data point is information for an individual vat; for a single variable it is a single value and for a group of variables it is the corresponding group of values.

the distance between that point and the centroid. This distance, called Mahalanobis distance, was calculated using the estimated covariance matrix of the variable group subset (Help menu, JMP IN version 3.0. 1996; SAS Inst., Cary, NC). Data transformation yielded a vector of n Mahalanobis distances. Outliers were identified as the data points (vats) with a Mahalanobis distance greater than the Mahalanobis threshold distance ($D_{\text{threshold}}$) for the variable group subset, calculated by the following equation:

$$D_{\text{threshold}} = \sqrt{v F_{\text{Quantile}}(\text{Probability}, \frac{v}{n - v - 1} \text{degrees of freedom})} \quad [2]$$

$$\sqrt{v F_{\text{Quantile}}(\text{Probability}, \frac{v}{n - v - 1} \text{degrees of freedom})}$$

where F_{Quantile} is the zero-centered Fisher quantile function (inverse of the Fisher probability function). In this study, a probability value of 0.9999 was used. The Mahalanobis distances and F_{Quantile} calculations were carried out with the statistical software.

Validation of Data from Individual Vats

Data from individual vats were validated using the SSC and Fourier-SSC applied on single variables, and the Mahalanobis outlier analysis on groups of variables. For the single variable SSC, data for each variable were divided into subsets (Figure 1). This division was based

on the previous classification of each variable into variation groups (variation according to the recipe, the use or not of standardized milk, the S-CM, or neither of these). For this method, the subdivision of the data for each variable was done according to only one classification variable, where the number of subdivisions corresponded to the number of levels of the qualitative variable selected (i.e., three levels for the recipe, two for milk standardization, and six for S-CM combinations). For variables exhibiting no variations with respect to the three classification variables, no data subdivision was done. A total of 176 variable subsets were analyzed.

For the Fourier-SSC approach, data from all variables tested for seasonal variations was divided into vats that used standardized milk and vats that did not. For nonstandardized milk data, Fourier modeling revealed that only fat, protein, lactose, and total solids in milk, and protein in whey exhibited highly significant seasonal effects along with model determination coefficients (r^2) above 0.25. Other variables also exhibited significant seasonal effects but yielded r^2 values less than 0.25 and were thus not validated with Fourier-SSC.

For the Mahalanobis analysis, the data were divided into variable group subsets (Figure 1). Variable groups were composed of variables previously classified in the same variation group (i.e., that presented variations according to recipe, the use or not of standardized milk, the S-CM, or neither of these). However, not all variables classified in the same variation group were placed in the same variable group for this analysis. The number of subdivisions of the data of the variable group corresponded to the number of levels of the respective classifications variable (i.e., three recipes, two levels of milk standardization, and six S-CM combinations). For variables presenting no variations with respect to the classification variables, no data subdivision was done. Because this method can only analyze complete data points, some variables were not used because their inclusion added too many missing values, which resulted in too many data points excluded from the analysis. The method was applied to 31 variable group subsets, corresponding to 46 variables.

Validation of Data from Productions

The SSC method was applied to variables derived from the calculations of fat mass balance, fat retention coefficient, and yield efficiency, in order to detect outlier productions. A production is a series of contiguous cheese vats of the same recipe. A production basis instead of an individual vat basis was necessary because the continuous drainage step in the process did not allow assigning an accurate cheese weight to a particu-

lar milk vat. Indeed, the correlation between cheese weight and milk weight recorded per vat was very low ($r = 0.121$), but very high ($r = 0.998$) for the corresponding sums of these weights per production.

For a production with less than one third of vats with missing values for any of the variables required for calculation, the missing data were filled in with mean values from adjacent vats of the same production. Productions with over one third of vats with missing values were not considered in the analysis. The mean values added were only used for these calculations and were not retained in the original database.

Calculation of fat mass balance (FMB). The error of the fat mass balance for every production equals the mass of fat in milk (m_{FM}) minus the sum of the masses of fat in cheese and whey ($m_{FC} + m_{FW}$):

$$\text{Error} = m_{FM} - (m_{FC} + m_{FW}) = \sum_{i=1}^{nVats} x_{FM-i} m_{M-i} - \left(\sum_{i=1}^{nVats} x_{FC-i} m_{C-i} + \sum_{i=1}^{nVats} x_{FW-i} (m_{M-i} - m_{C-i}) \right) \quad [3]$$

where x_{FM} , x_{FC} , and x_{FW} are the mass fractions of fat in milk, cheese and whey, respectively; m_M and m_C are the masses (weights) of milk and cheese, respectively, $nVats$ is the number of vats for the production and the subscript $-i$ is the vat number within the production.

Milk volume was converted to mass by multiplying by milk density at 32°C (milk temperature at vat filling). The following formula, adapted from a published formula (USDA, 1965), was used to estimate milk density for a given composition and temperature:

$$\rho_M = \rho_{H_2O} \frac{1}{1 + C_{FM} x_{FM} - C_{SNFM} x_{SNFM}} \quad [4]$$

where ρ_M and ρ_{H_2O} are milk and water densities, respectively; x_{FM} and x_{SNFM} are mass fractions of fat and SNF in milk, respectively; and C_{FM} and C_{SNFM} are the coefficients for fat and SNF in milk, respectively. Polynomial equations for C_{FM} and C_{SNFM} as a function of temperature, in the range from 10 to 38.9°C, were fitted using C_{FM} and C_{SNFM} data determined from a large number of milk and cream samples for 10, 20, and 38.9°C (USDA, 1965):

$$C_{FM} = -3.967 \times 10^{-5} T^2 + 3.560 \times 10^{-3} T + 0.01648 \quad [5]$$

$$C_{SNFM} = -1.091 \times 10^{-6} T^2 - 3.773 \times 10^{-4} T + 0.3894 \quad [6]$$

where T is milk temperature (°C). C_{FM} and C_{SNFM} for 32°C were equal to 0.0898 and 0.3758, respectively.

The mass of fat in milk (m_{FM}) and that for cheese and whey (m_{FC+FW}) per production were transformed into principal components per recipe. The SSC ($\bar{x} \pm 3.6$ SD) was then repeatedly applied to the values of the P2 per recipe, until no new outlier productions were found.

Calculation of the fat retention coefficient (K_f). The fat retention coefficient corresponds to the fraction of fat in milk that is recovered in cheese. It was calculated per production by:

$$K_f = \frac{\sum_{i=1}^{nVats} x_{FC-i} m_{C-i}}{\sum_{i=1}^{nVats} x_{FM-i} m_{M-i}} \quad [7]$$

The SSC outlier criterion ($\bar{x} \pm 3.6$ SD) was then repeatedly applied to the K_f values for each recipe, until no new outlier productions were found.

Calculation of yield efficiency. Yield efficiency was calculated as the ratio of either actual yield or adjusted actual yield to theoretical yield. Actual yield (Y_{act}) corresponds to the quantity of cheese per 100 kg of milk, and for a production it was calculated as:

$$Y_{act} = 100 \frac{\sum_{i=1}^{nVats} m_{C-i}}{\sum_{i=1}^{nVats} m_{M-i}} \quad [8]$$

Actual adjusted yield for a production (Y_{adj}) was calculated by adjusting Y_{act} for target cheese moisture, salt, and an estimated value of whey solids using the following formula (Emmons, 1993):

$$Y_{adj} = Y_{act} \times \frac{1 - \overline{x_{H_2Oact}} - \overline{x_{NaClact}} - \overline{x_{WSCact}}}{1 - x_{H_2Otarget} - x_{NaCltarget} - 0.0695 x_{H_2Otarget}} \quad [9]$$

where $x_{H_2Otarget}$, $x_{NaCltarget}$ are the target mass fractions for moisture and salt in cheese, respectively; and $\overline{x_{H_2Oact}}$, $\overline{x_{NaClact}}$, and $\overline{x_{WSCact}}$ are the production weighted averages of the mass fractions of actual cheese moisture, salt, and whey solids in cheese, respectively. For a production, the weighted averages were based on milk mass per vat and were calculated as the sum of the products, for all vats in the production, of the value of the variable and the corresponding milk mass per vat divided by the sum of the milk mass for all vats in the production. The mass fraction of whey solids in cheese (x_{WSC}) was estimated from the mass fraction of moisture in cheese (x_{H_2O}) and a constant of 0.065 for the mass

fraction of solids in whey (x_{WS}), as suggested by Emmons (1993):

$$x_{WSC} = \frac{x_{WS}}{1 - x_{WS}} x_{H_2O} = 0.0695 x_{H_2O} \quad [10]$$

A modified Van Slyke and Price (1927) cheese yield equation (Y_{VSPmod}) was used to calculate the theoretical yield for each vat. The equation was adapted to use protein and fat in milk mass fractions (x_{PM} , and x_{FM} , respectively) and the casein to protein in milk ratio (λ), giving the following formula for Cheddar cheese containing 37% moisture and 1.7% salt:

$$Y_{VSPmod} = 1.09 \frac{93x_{FM} + 100\lambda x_{PM} - 0.1}{1 - 0.37} \quad [11]$$

The λ values were estimated from a second-order Fourier model, calculated for the region that supplied milk to the cheese plant of this study (Lacroix et al., 1996). In a similar manner to Y_{act} , Y_{VSPmod} was adjusted for the targets of moisture, salt and whey solids in cheese for each recipe, in order to calculate theoretical yield per vat (Y_{th}):

$$Y_{th} = \quad [12]$$

$$Y_{VSPmod} \times \frac{1 - 0.37 - 0.017 - 0.0257}{1 - x_{H_2Otarget} - x_{NaCltarget} - 0.0695x_{H_2Otarget}}$$

Theoretical yield per production ($\bar{Y}_{th}$) was then obtained by calculating the weighted average of Y_{th} based on milk mass per vat. Yield efficiency based on actual yield (E_{act-th}) was then calculated as:

$$E_{act-th} = \frac{Y_{act}}{\bar{Y}_{th}} \quad [13]$$

while that based on adjusted actual yield (E_{adj-th}) was calculated as:

$$E_{adj-th} = \frac{Y_{adj}}{\bar{Y}_{th}} \quad [14]$$

The SSC outlier criterion ($\bar{x} \pm 3.6$ SD) was then repeatedly applied to the values of E_{act-th} , and E_{adj-th} per recipe, until no new outlier productions were found.

Software Tools

The statistical software used for principal component analysis, Fourier model fitting, Mahalanobis outlier analysis, and general statistical descriptor calculation

was JMP IN version 3.0 (1996; SAS Inst., Cary, NC). The detection of outliers for single variable subsets with the SSC was carried out with an in-house program.

RESULTS

Validation of Data from Individual Vats

Single-variable SSC. The application of the simple statistical criterion ($\bar{x} \pm 3.6$ SD) led to the detection of outliers in 94 (53%) subsets (outlier-positive subsets) of the total 176 single variable subsets analyzed. Of these 94 outlier-positive subsets, 64% had fewer than 3000 data points (SSC would eliminate 1 in 3143 normally distributed values, or 0.032% of data points). Table 1 summarizes the results obtained for this validation approach. Although the percentage of outliers detected in a subset ranged from 0.02 to 8.44%, it was below 1% for 61% of outlier-positive subsets, and below 3% for 89% of them. Noticeably, of the 10 subsets with over 3% outliers, six corresponded to manually registered time variables.

Outlier removal had more impact on the range than on the SD of subsets (Table 1). Outlier removal always caused a greater decrease in the range than in the SD of subsets ($\Delta R > \Delta SD$), and thus systematically led to an increase in the ratio of the SD to the range of values (**SD/R**) (Table 1). Presenting the SD as a percentage of the range of values was used to compare data dispersion of variables with different units and magnitudes. The SD/R ratio for outlier-positive subsets was always greater than 13.9% after outlier removal, and it ranged from 14 to 24% for 90% of these subsets. The SD/R value of 13.9% ($= 1/7.2$) corresponds to the definition of the outlier criterion ($\bar{x} - 3.6$ SD < acceptable data < $\bar{x} + 3.6$ SD), which allows a maximum range equal to 7.2 SD. Outliers were found in all subsets that had SD/R ratios below 13.9% before validation. This was the case for 67% of outlier-positive subsets; but for the remaining 33%, SD/R before validation was greater than 13.9%. All outlier-negative subsets exhibited SD/R ratios greater than 16.8%.

Most of the 94 outlier-positive subsets (83%) required a maximum of two validation passes for all outliers to be detected (one pass for 51%, and two passes for 32% of outlier-positive subsets). For the remaining 17% outlier-positive subsets that needed three or more passes, more than half corresponded to manually registered variables. Also, of the eight outlier-positive subsets that presented the greatest changes in mean value ($> 2\%$ with respect to the subset range), seven were manual time registers.

Fourier-SSC. Table 2 presents model coefficients obtained for nonstandardized milk productions. The highest r^2 values of Fourier models were obtained for fat

Table 1. Summary of data validations using the simple statistical criterion (SSC) and the Mahalanobis (Mahal) outlier analyses for subsets of variables grouped according to variation with respect to the cheese recipe (Recipe), milk composition due to standardization (Milk Std), starter and starter culture medium (S-CM) for starter production times (–Times) and starter activity measures (–Activity), and a group of variables that did not present variations according to neither of those parameters (No variation). ΔR is the percent decrease in the range of subset values and ΔSD is the percent decrease in SD, after outlier removal. SD/R is the ratio of the SD to the range of the subset.

Method	Group of variables, variation by	Number of variables in group	Number of data subsets ¹ obtained from group	Number of subsets with outliers	Range of number of data points ² in a subset	Range of data points ² detected as outliers in a subset (%)	Number of validation passes ³	Descriptors of single variable subsets ⁴			
								ΔR (%)	ΔSD (%)	SD/R (%)	
										Before ⁵	After ⁵
SSC	Recipe	22	66	43	1212–5921	0–8.44	0–9	0–67	0–45	6.2–47.2	14.0–48.2
	Milk Std	5	10	10	2937–7570	0.14–2.82	1–4	18–56	1.3–30	8.7–13.2	14.0–15.7
	S-CM–Times	3	15	0	1018–1698	0	0	0	0	18.3–23.1	18.3–23.1
	–Activity	15	75	20	55–114	0–3.51	0–2	0–70	0–39	11.1–24.6	16.7–24.7
	No variation	10	10	10	10,415–10,606	0.09–0.85	1–2	13–98	0.8–87	2.9–20.0	14.4–28.5
Mahal	Recipe	18	3	3	1404–5031	0.71–2.82	1	0–63	–0.3–19	6.3–46.3	8.6–46.8
	Milk Std	4	2	2	2935–7450	0.79–1.70	1	6–25	0.3–20	8.7–13.1	9.0–14.1
	S-CM –Times	3	5	1	977–1275	0–2.46	1	0	–0.6–0	17.4–22.7	17.4–22.7
	–Activity	12	20	6	59–114	0–2.53	1	0–46	–0.5–19	11.1–24.5	12.9–24.5
	No variation	9	1	1	10,125	2.21	1	0–85	0.6–73	2.6–24.4	4.8–24.2

¹For SSC a subset is a subgroup of data for a single variable, while for Mahal it is a data subgroup with values from all variables comprising the group. One data subset was generated for each level of the variable or combination of variables under which the validated variables were grouped.

²A data point corresponds to a single cheese vat (starter vat in the case of starter activity variables). For SSC, it is a single value of a variable; while for Mahalanobis, it is the complete set of values, for the vat in question, for each of the variables that comprise the variable group.

³Maximum number of subset analyses for which new outliers were detected. Zero indicates no outlier was detected upon the first validation pass.

⁴Does not include values of 100% for ΔR and ΔSD (final range and SD values equal to zero, respectively), nor undetermined SD/R values (both range and SD values equal to zero).

⁵Before and after outlier removal.

Table 2. Fourier series models for fat, protein, lactose and total solids in milk, and protein in whey calculated for production (group of vats) means and nonstandardized milk for 4 yr of production.

Variable ¹	Fourier model coefficients ^{2,3}									r ²	Standard deviations	
	a ₀	a ₁	b ₁	a ₂	b ₂	a ₃	b ₃	a ₄	b ₄		Prod. means ⁴	Residuals ⁵
Fat	3.6675	0.0676	0.0184	-0.0458	-0.0415	NS	0.0218			0.65	0.086	0.051
Protein	3.1420	0.0539	-0.0186	-0.0138	-0.0141	0.0143	0.0210	0.0123	-0.0110	0.63	0.060	0.037
Lactose	4.5377	NS	0.0668							0.32	0.084	0.069
Total solids	11.9581	0.1147	0.0654	-0.0632	-0.0653	NS	0.0429			0.57	0.155	0.102
Protein in whey	0.8598	0.0189	-0.0243	-0.0106	-0.0065*					0.32	0.043	0.035

¹Milk variables unless otherwise stated.

²Model equation: $y = a_0 + a_1\cos(\omega t) + b_1\sin(\omega t) + a_2\cos(2\omega t) + b_2\sin(2\omega t) + \dots + a_4\cos(4\omega t) + b_4\sin(4\omega t)$, where ωt represents the position in the cycle expressed in radians.

³All coefficients were highly significant ($P \leq 0.001$), except for * ($P \leq 0.05$) and NS for nonsignificant ($P > 0.05$).

⁴A production mean is the mean of the variable values for all vats in the production.

⁵Residuals are calculated by subtracting the Fourier model estimate from the corresponding production mean.

and protein in milk, while the lowest corresponded to lactose in milk and protein in whey. The SD for model residuals were always lower than those for variable values. Residual SD were 18 to 41% lower than production mean SD for the variables presented in Table 2. Fourier models for fat, protein, and lactose in milk, and protein in whey data from productions using nonstandardized milk are shown in Figure 2. Only protein, lactose, and total solids in milk, and protein in whey retained highly significant seasonal effects for standardized milk. However, only the model for total solids of standardized milk yielded a much lower r^2 value (0.26 vs. 0.57).

Table 3 presents a summary of the validation carried out with the Fourier-SSC analysis. The number of outliers detected by applying the SSC on model residuals was in all cases above 0.032% (confidence interval threshold) and ranged from 0.46 to 1.00% for nonstandardized milk subsets, and from 0.17 to 1.77% for standardized milk subsets. For most variables, the method detected a lower percentage of outliers in the standardized milk subsets (11 to 63% lower). The exception was total solids in milk, where the percentage of outliers was 77% greater for the standardized milk subsets.

A minimum of two validation passes were necessary for all subsets analyzed, but most (seven out of the nine subsets) required three or more passes. The SD/R ratios for model residuals of all nine subsets were above 13.9% after removing outliers, whereas for measured data, this was verified for seven of the subsets. The percent decrease of SD (Δ SD) for model residuals after outlier removal was always greater than that for measured data. Likewise, for 8 out of the 9 subsets, the percentage decrease of subset range (Δ R) for model residuals was greater than that for the corresponding measured data.

Mahalanobis outlier analysis. Table 1 summarizes validation results obtained with this method. Only one validation pass was used because the percentage

of outliers detected in outlier-positive variable group subsets was relatively high, ranging from 0.71 to 2.82%. Among the 31 variable group subsets analyzed, a total of 13 (42%) were outlier-positive. The method was able to detect outliers even in variable group subsets where the SD/R ratio of all single variable subsets was greater than 20%.

It was observed that this method detected outliers in the center, not only in the extremes of the distributions of the single variable subsets constituting the variable group subsets analyzed (Figure 1). Indeed, for over half (48 of 90) of the single variable subsets contained in the 13 outlier-positive variable group subsets, neither the minimum nor the maximum values of the subsets were removed as outliers; consequently, their range did not change ($\Delta R = 0$), and for 12 of these subsets, the removal of outliers actually caused an increase, instead of a decrease, in the SD of the subset ($-0.6 < \Delta SD < 0$). As a result, the SD/R ratio increased for these 12 subsets, but actually decreased for the other 36. For the other 42 single variable subsets where the range did decrease ($\Delta R > 0$) after outlier removal, the SD also decreased ($\Delta SD > 0$), but to a lesser degree for 40 of them, such that SD/R increased (but decreased for the other two subsets). Whether SD/R increased or decreased, the final SD/R values for 86% of single variable subsets ranged from 10 to 25%. This range was 12.9 to 24.5% for all single variable subsets contained in variable group subsets where no outliers were found.

Validation of Data from Productions

Fat mass balance-SSC. Figure 3 shows the high correlation ($r = 0.999$) between the mass of fat in cheese and whey and that in milk for recipe 1 per production. The mass balance mean absolute error (MAE) increased as the amount of fat in the production increased (number of vats in a production). Indeed, the correlation

Figure 2. Fourier models calculated for fat, protein, and lactose in nonstandardized milk, and protein in whey for the 4-yr production database. The range represents the difference between the highest and lowest value of the model.

between the number of vats per production and the FMB MAE was highly significant, with correlation coefficients (r) of 0.68, 0.77, and 0.67, for recipes 1, 2, and 3, respectively. Recipe 1, which had the highest mean number of vats per production, exhibited the largest MAE (Table 4). For this reason, the FMB MAE could not be used as an outlier indicator.

Figure 4 illustrates outlier detection by applying SSC to the P2 of the transformed m_{FM} and m_{FC+FW} data for recipe 1. For all three recipes, the application of SSC to P2 yielded 2% outlier productions, which corresponded to 2.6% of vats. For each recipe, the range and SD of P2 decreased, while the SD/R ratio increased after outlier removal (Table 4).

Fat retention coefficient-SSC. Figure 5 illustrates the change in K_f for recipe 1 productions (expressed as the difference with respect to the mean value to protect data confidentiality). For data from all three recipes, this validation method eliminated 3.5% of productions as outliers, corresponding to 2.6% of vats. The decrease in SD for K_f after validation ranged from 30 to 37%, and the range of K_f decreased between 48 and 63% for all three recipes after outlier removal. The SD/R ratio for all three recipes largely increased upon outlier removal, from 10.2 to 12.7% before validation to 17.2 to 17.8% after outlier removal.

Yield efficiency-SSC. No differences ($P > 0.05$) were found between the means of E_{act-th} and E_{adj-th} . The greatest difference between them for any single recipe before or after outlier production removal was 0.0008. However, both variables were not highly correlated (correlation coefficients after outlier removal of 0.567, 0.572, and 0.745 for recipes 1, 2 and 3, respectively) and exhibited appreciably different behaviors (Figure 6).

Figure 6 presents calculated values of E_{act-th} and E_{adj-th} for recipe 1 productions. Validation with E_{adj-th} detected two additional outliers than with E_{act-th} for recipe 1, but the opposite was observed for recipe 2 (Table 4). For all three recipes, SSC validation of E_{act-th} and E_{adj-th} eliminated 1.2 and 1.4% of productions, corresponding to 0.6 and 0.7% of vats, respectively. For recipes 1 and 3 the SD/R ratio, as well as the decreases in the range and SD, after outlier removal, were greater for E_{adj-th} than for E_{act-th} , while the opposite was true for recipe 2 (Table 4).

DISCUSSION

The extensive cheesemaking information recorded on paper or computer files is underutilized for process analysis and control, partly because of the common occurrence of many erroneous data. In addition, there are currently no procedures for detecting erroneous data in large industrial cheese manufacturing databases. In this study, different validation approaches were tested for data both from single vats and total productions (groups of vats).

Validation of Single Vat Data

For methods that used the SSC, it was observed that the subset SD/R ratio always increased after outlier removal, and was higher than 13.9%. The SD/R ratio was useful in determining whether a data subset needed SSC validation. This was always the case when $SD/R < 1/n_s$, where n_s is the number of SD in the confidence interval used (equal to 7.2 in this study). However, if $SD/R > 1/n_s$, validation may still be required, as it was noted for

Table 3. Validation summary for seasonal variables [milk fat, protein, lactose, and total solids (TS); and protein in whey (PW)] by Fourier-SSC (simple statistical criterion), SSC and Mahalanobis (Mahal) methods, obtained for 7644 nonstandardized (N-Std) milk vats and 2964 standardized (Std) milk vats. ΔR is the percent decrease in the range of the subset, and ΔSD the percent decrease in the SD of the subset after outlier removal; SD/R is the ratio of SD to the range of the data subset.

Method	Milk ¹	Variable	Number of detected outliers	Number ¹ of validation passes	Descriptors of individual subsets of variables							
					Measured values				Fourier model residuals ²			
					ΔR (%)	ΔSD (%)	SD/R (%)		ΔR (%)	ΔSD (%)	SD/R (%)	
							Before ³	After ³			Before ³	After ³
Fourier-SSC	N-Std	Fat	38	3	30	2.9	10.5	14.6	55	5.7	6.7	14.1
		Protein	75	4	35	5.2	10.7	15.7	52	9.1	7.5	14.2
		Lactose	55	3	34	6.7	9.5	13.4	50	8.4	7.6	14.0
		TS	75	3	37	6.7	9.2	13.6	50	10.2	7.9	14.0
		PW	34	3	41	3.9	8.7	14.1	48	4.8	7.8	14.1
	Std	Protein	26	2	24	3.8	13.2	16.6	20	6.6	12.7	14.9
		Lactose	15	2	21	3.8	12.3	15.1	34	4.4	10.1	14.7
		TS	52	6	24	9.9	11.9	14.1	29	11.7	11.4	14.1
		PW	5	3	25	1.6	11.1	14.6	32	2.0	10.8	15.5
SSC	N-Std	Fat	20	2	33	2.4	10.5	15.3				
		Protein	28	2	28	3.2	10.7	14.3				
		Lactose	57	3	38	7.0	9.5	14.2				
		TS	48	2	39	5.7	9.2	14.3				
		PW ⁴	28	2, 1 ⁵	39	3.6	8.7	13.7				
Mahal	N-Std	Fat	59	1	12	2.1	10.5	11.7				
		Protein	59	1	3	1.7	10.7	10.8				
		Lactose	59	1	21	4.7	9.5	11.5				
		TS	59	1	3	2.0	9.2	9.3				
		PW ⁴	107	1, 1 ⁵	3	1.1	8.7	8.9				

¹Maximum number data set analyses for which new outliers were detected.

²A model residual is the measured variable value minus the value predicted by the corresponding model.

³Before and after removal of outlier data points.

⁴For these two methods, protein in whey (PW) was validated per recipe, for the two recipes that use nonstandardized milk. Thus, for comparison purposes, descriptor values reported were obtained by combining the data from two subsets.

⁵Numbers correspond to the validation passes needed for the first and second recipe subsets, respectively.

one third of the outlier-positive data subsets detected by the SSC, which had SD/R ratios above 13.9%.

The number of validation passes and of outliers detected can be indicators of the reliability of data. Single variable subsets validated with SSC showed that two validation passes were sufficient for detecting outliers in most subsets and that less than 3% of the data were discarded as outliers. However, for time variables registered manually, the method required up to nine validation passes and detected as much as 5.29% of the data as outliers for some subsets. This suggested the low reliability of these data.

The subdivision of data into smaller subsets was carried out for two reasons: 1) to generate sets of data with similar process conditions, and 2) to increase sensitivity of the validation method for outlier detection by removing sources of variation. This last point was particularly observed during the analysis of curd pH data. Although all three recipes had the same curd pH target, different numbers of outliers were detected by the SSC, with and without data subdivision (Table 5). Applying SSC to all pH data detected a total of 46 outliers for all three recipes versus 54 detected by SSC when applied to data for

each recipe. Also, different data points were detected as outliers because the upper and lower limits for each recipe were different than those for the entire pool of curd pH data. This is an example of an increase in outlier detection sensitivity obtained by subdividing data according to known or observed sources of variation. However, subdivision needs to be carried out with caution, because excessive subdivision can lead to subsets that are too small, with data that are too disperse for outlier detection by a statistical criterion. In this study, however, despite the wide confidence interval given by $\bar{x} \pm 3.6$ SD (eliminating 1 in 3143 points with a normal distribution), the SSC detected outliers in data sets with as few as 55 values. This suggests the presence of aberrant data points and confirms the need for at least this simple level of data verification.

The Fourier-SSC method was clearly more adequate for validating milk and whey compositional variables than SSC. The seasonal variations for milk and whey components observed in this study (Table 2 and Figure 2) have also been reported for fat, protein, and lactose in milk for Northeastern America (Barbano, 1990; Ng-Kuwai-Hang et al., 1987) and the Netherlands (Lolkema,

Figure 3. Correlation between fat in milk and fat in cheese and whey for production schedules for recipe 1 and 4 yr of production, using the fat mass balance (FMB) vat-based calculation approach. The line corresponds to a perfect correlation.

1993), and for total protein and noncasein nitrogen in milk for different regions of the Province of Québec (Lacroix et al., 1996). However, one difficulty in using Fourier-SSC is finding the most adequate seasonal model for the variable in question. Seasonal effects on bulk milk composition can change from one year to another as a result of temperature and weather differences or other factors (Lawrence, 1993). Consequently, a model adjustment could be required for different years. The

Figure 4. Detection of outlier productions for recipe 1 and 4 yr of production by the simple statistical criterion (SSC) applied to the second principal component (P2) of the transformed fat in milk and fat in cheese and whey data, using the FMB vat-based calculation approach. Dotted lines (---) indicate the limits for the confidence interval given by $\bar{x} \pm 3.6$ SD.

use of fourth-order terms in the Fourier models is probably not needed to explain a variable's general seasonal variation. However, in this study the objective was to best fit the data for all 4 yr and use the resulting model for data validation. Table 3 shows that for vats that used nonstandardized milk, except for lactose in milk, Fourier-SSC detected more outliers than SSC alone. However, it required an equal or greater number of validation passes.

Different outliers for fat in nonstandardized milk for the 4 yr of production were detected by SSC, Fourier-SSC, and Mahalanobis (Figure 7), and the methods showed different outlier detection sensitivities (Table 3). Fourier-SSC generally detected a greater number of outliers than SSC due to a narrower confidence interval calculated with model residuals. Also, the lack of sensitivity to seasonal variations of SSC resulted in the elimination of possibly accurate data with seasonally low or high values, and the retention of potentially erroneous data near the yearly average during those high and low periods. This was clearly observed during the summer period (Figure 7). Most outlier vats for SSC, Fourier-SSC, and Mahalanobis were the same at the extremes of the distributions of single variable subsets. However, unlike SSC and Fourier-SSC, Mahalanobis also detected outliers at the center of the single variable distributions. Many, but not all outlier vats detected by Mahalanobis in the center of the distribution of fat in nonstandardized milk, were also detected as outliers in protein, lactose, or total solids in milk (the other variables composing the nonstandardized milk variable group subset), by either SSC or Fourier-SSC.

Fourier-SSC and Mahalanobis better accounted for seasonal variations than SSC. However, an advantage of Mahalanobis over Fourier-SSC is that an a priori model was not needed, as seasonality was expressed by the dispersion of data in a multidimensional (multivariable) space. However, two disadvantages to the use of Mahalanobis are that only complete vat records can be analyzed, and that the choice of variables for the variable group is not always obvious. One is often inclined to include as many variables as possible in the variable group in order to reduce the amount of work involved in the manipulation and analysis of data. Although including an additional variable in a variable group may add more information about the other variables, it will contribute additional error to the data and will cause the Mahalanobis threshold distance to increase (Figure 8). Therefore, if the new variable does not bring pertinent information for describing the data, its addition may decrease the outlier detection sensitivity of the method.

When analyzing one variable at a time, for nonstandardized milk composition variables, Mahalanobis detected more single variable values as outliers than SSC,

Table 4. Summary for validation of data from groups of vats (productions) by applying the simple statistical criterion (SSC) to variables obtained in the calculations of the fat mass balance (P2), fat retention coefficient (K_f), and yield efficiency ($E_{\text{act-th}}$ and $E_{\text{adj-th}}$) of the operation for recipes 1, 2 and 3, corresponding to 239, 152, and 113 productions, respectively. ΔR is the percent decrease in the range of the subset, and ΔSD the percent decrease in the SD of the subset after outlier removal; SD/R is the ratio of SD to the range of the data subset.

		Descriptors of subset of variables								
Calculation	Recipe	MAE ¹ (kg fat/vat)	Variable ²	ΔR (%)	ΔSD (%)	SD/R (%)		Outlier		Validation passes ⁴
						Before ³	After ³	Productions	Vats	
Fat mass balance	1	728	P2	50	21	8.9	14.0	4	142	2
	2	513		53	29	12.0	18.1	5	124	5
	3	506		19	9	14.5	16.3	1	13	1
Fat retention coefficient	1		K _f	61	37	10.5	17.7	11	198	4
	2			48	30	12.7	17.2	3	44	1
	3			63	34	10.2	17.8	4	44	3
Yield efficiency	1		E _{act-th}	53	19	9.5	16.5	3	44	1
	2			51	18	10.4	17.4	2	10	2
	3			21	6	14.2	17.0	1	11	1
	1		E _{adj-th}	63	26	8.6	16.9	5	64	2
	2			43	16	11.4	16.8	1	4	1
	3			44	13	11.9	18.6	1	11	1

¹Mean absolute error per vat in fat mass balance (FMB), corresponding to the difference between fat in milk and fat in cheese and whey.

²P2 = Second principal component coordinates for transformed fat in milk and fat in cheese and whey from the FMB; K_f = fat retention coefficient; $E_{\text{act-th}}$ = yield efficiency based on actual yield; $E_{\text{adj-th}}$ = yield efficiency based on adjusted actual yield.

³Before and after removal of outlier data points.

⁴Maximum number of data analyses for which new outliers were detected.

and more than Fourier-SSC for three out of the five variables (Tables 3 and 6). However, when considering data for all four variables without missing values, the total number of outlier vats detected by SSC and Fourier-SSC were greater than for Mahalanobis (102 and 154, respectively vs. 59). Of those outlier vats, only 42 outlier vats were common to all three methods, where SSC and Fourier-SSC had more common outlier vats between them (89), than Mahalanobis had with either method (46 common outlier vats with each one). The higher prob-

ability value used to calculate Mahalanobis distances (equation 2) compared to that used for the SSC confidence interval (0.9999 vs. 0.99968, or 1 outlier in 10,000 vs. 3143 normally distributed data points, respectively) may explain the lower number of outlier vats for Mahalanobis versus the other methods.

Validation of Data from Productions

Despite nearly identical means and SDs for $E_{\text{act-th}}$ and $E_{\text{adj-th}}$ for all recipes before and after validation, $E_{\text{adj-th}}$

Figure 5. Detection of outlier productions for recipe 1 and 4 yr of production by the simple statistical criterion (SSC) applied to the fat retention coefficient K_f (kg of fat in cheese/kg of fat in milk) estimated by the vat-based calculation approach. The dotted lines (---) indicate the limits for the confidence interval given by $\bar{x} \pm 3.6$ SD. Change in K_f is presented instead of the actual value to maintain confidentiality.

Figure 6. Detection of outlier productions for recipe 1 and 4 yr of production by the simple statistical criterion (SSC) applied to yield efficiency variables $E_{\text{act-th}}$ (—) and $E_{\text{adj-th}}$ (○). Yield efficiency is dimensionless. $E_{\text{act-th}}$ is the quotient of actual yield divided by theoretical yield, while $E_{\text{adj-th}}$ is the quotient of adjusted actual yield divided by theoretical yield. The dotted lines (---) indicate the limits for the confidence intervals for $E_{\text{act-th}}$ and $E_{\text{adj-th}}$, given by $\bar{x} \pm 3.6$ SD.

Table 5. Effect of data subdivision on validation of curd pH data using the simple statistical criterion (SSC). ΔR and Δs represent the percent decreases, after outlier removal, of the range and of the SD, respectively. s/R is the ratio of the standard deviation to the range of the data subset.

Recipe	Before outlier removal	After outlier removal							
		Data subdivision per recipe				No data subdivision			
		ΔR (%)	Δs (%)	s/R (%)	Outliers	ΔR (%)	Δs (%)	s/R (%)	Outliers
1	9.3	38	3.6	14.3	22	32	3.1	13.2	17
2	10.5	30	3.0	14.5	10	25	4.0	13.4	16
3	9.9	41	10.7	14.9	22	34	7.2	13.9	13
All	8.8	38	4.1	13.5	54	41	3.8	14.2	46

led to the detection of one additional outlier production compared to E_{act-th} . For recipe 1, SSC found five outlier productions with E_{adj-th} versus three with E_{act-th} (Table 4 and Figure 6). However, for recipe 2, E_{act-th} -SSC detected one additional outlier production compared to E_{adj-th} -SSC. These data suggest that both calculations for yield efficiency may lead to the same level of outlier detection sensitivity.

Yield efficiency eliminated most seasonal variations naturally present in actual yield (Figures 6 and 9, re-

spectively). This resulted in yield efficiency values with a much lower dispersion than those of Y_{act} (SD/R ratios for recipe 1 before outlier removal of 8.6 and 17.6%, respectively), and gave greater outlier sensitivity to SSC. Figures 6 and 9 show that production 91 was detected as an outlier with either E_{act-th} or E_{adj-th} and SSC, but was masked by seasonal variations of Y_{act} . These data suggest that unusually low winter or unusually high summer yield values are more likely to be detected using yield efficiency. Although Y_{act} could be

Figure 7. Comparison of outlier detection for fat in nonstandardized milk by the simple statistical criterion (SSC), Fourier-SSC and Mahalanobis methods, for 4 yr of production. Dotted (---) and solid lines (—) indicate the limits for the confidence interval limits ($\bar{x} \pm 3.6$ SD) calculated by SSC with actual data, and with Fourier model residuals, respectively. Open circles (O) show the outliers detected by the Mahalanobis method.

Figure 8. Change of Mahalanobis threshold distance for a confidence interval of 99.99% as a function of the number of data points for different numbers of variables in the variable group data subset.

validated with Fourier-SSC, the development and selection of Fourier models renders the procedure more complex than simply calculating yield efficiency.

Of the validation approaches tested on production data, K_f -SSC detected the greatest number of outlier productions (18 vs. 10 and 7 for FMB-SSC and E_{adj-th} -SSC, respectively). This higher outlier sensitivity may in part be explained by the fact that K_f calculation uses the lowest number of variables. All three calculations (FMB, K_f and yield efficiency) use the same four mea-

Figure 9. Seasonal change of actual yield (Y_{act}) for recipe 1 and 4 yr of production. The dotted line (---) indicates the lower limit for the confidence interval given by $\bar{x} \pm 3.6$ SD. Yield units are kilograms of cheese per 100 kg of milk. The change of Y_{act} is presented instead of the actual value to maintain confidentiality.

sured variables: milk mass, cheese mass, fat in milk, and fat in cheese. The FMB also includes the fraction of fat in whey, while yield efficiency adds protein in milk, moisture and salt in cheese, for a total of five and seven variables, respectively. The inclusion of more measured variables into the analysis increases the sources of error on the estimated parameter and may decrease outlier sensitivity. However, the use of additional pertinent variables may actually help explain normal low or high data values that could otherwise be eliminated by a method that uses fewer variables (i.e., SSC).

Among the FMB-SSC, K_f -SSC, and E_{adj-th} -SSC validation approaches, FMB-SSC and E_{adj-th} -SSC were the most complementary for outlier detection. Of the 16 outlier productions detected by the two approaches, only one was common. In contrast, all the E_{adj-th} -SSC outliers were also detected by K_f -SSC. The correlation coefficients between the FMB second principal component (P2) and E_{adj-th} were low (0.33, 0.47, and 0.52, for recipes 1, 2, and 3, respectively), while those between E_{adj-th} and K_f were much greater (0.58, 0.76, and 0.79, for recipes 1, 2, and 3, respectively).

Comparison of Vat and Production Data Validation Approaches

Among the validation approaches tested, the SSC is the easiest to apply. For all others, except for Mahalanobis, data were first transformed and then applied the SSC. Despite this similarity, the approaches applied to production data yielded complementary outliers to those found by SSC alone. Table 6 shows that FMB-SSC, K_f -SSC, and E_{adj-th} -SSC had very few outliers in common (in parenthesis) with SSC for milk composition variables and protein in whey. These three approaches, along with Mahalanobis, identified outliers both at the extremes, as well as the center of the data distributions of the single variable subsets. One advantage of the approaches for production data over Mahalanobis was that variables were assigned different weights in the calculations of FMB, K_f and yield efficiency (e.g., salt is not assigned the same weight as fat in yield efficiency calculations), whereas for Mahalanobis, equal weight is assigned to each variable in the group regardless of their importance in describing the data variation. On the other hand, a disadvantage of validating production data versus individual vat data is that all vats in an outlier production are eliminated, including vats that may contain reliable data.

In terms of effectiveness for detection of outlier vats, no clear advantage was observed when comparing the three validation approaches tested for individual vat data (SSC, Fourier-SSC, and Mahalanobis) versus the

Table 6. Outlier vats for milk and whey composition variables detected by the simple statistical criterion (SSC), Fourier-SSC, Mahalanobis, fat mass balance-SSC (FMB-SSC), fat retention coefficient-SSC (K_F -SSC) and yield efficiency-SSC (E_{adj-th} -SSC) methods, for data from vats using non-standardized milk. The number of identical outlier vats found by the methods and SSC are given in parenthesis.

Validated values from	Validation approach	Number outlier vats ¹				
		Fat in milk n = 7543	Protein in milk n = 7544	Lactose in milk n = 7458	TS in milk ² n = 7471	Protein in whey n = 7428
Individual vats	SSC	20	28	57	48	28
	Fourier-SSC	38 (17)	75 (26)	55 (51)	75 (40)	34 (26)
	Mahalanobis ³	59 (12)	59 (9)	59 (30)	59 (12)	107 (1)
Productions ⁴	FMB-SSC	134 (1)	134 (0)	133 (0)	133 (0)	155 (0)
	K_F -SSC	234 (2)	233 (0)	231 (1)	230 (0)	238 (1)
	E_{adj-th} -SSC	74 (1)	73 (0)	74 (1)	73 (0)	74 (1)

¹n corresponds to the number of vat data used for the analysis.

²Total solids in milk.

³Analysis was carried out on groups of variables using complete vat data only. The four milk variables were analyzed together as a group, and protein in whey was grouped with 17 other process variables for analysis.

⁴A production is a group of contiguous vats of the same cheese recipe.

three tested for production data (FMB-SSC, K_F -SSC, and yield efficiency-SSC). This is exemplified for 7450 complete vats (with no missing values) for the four non-standardized milk variables, in which K_F -SSC detected the greatest number of outlier vats (230) followed by Fourier-SSC (154), FMB-SSC (133), SSC (102), E_{adj-th} -SSC (73), and Mahalanobis (59) (Table 6). However, it is important to note that the actual number of vats eliminated by a given validation approach, actually depends on the group of variables selected.

Effect of Outlier Detection and Grouping of Variables on Number of Validated Data

Upon validation, the number of complete data points decreased as the number of grouped variables in-

creased. Outlier removal eliminates vat data, but missing data also leads to the elimination of incomplete vat registers when variables were grouped. Table 7 shows how, for six common cheesemaking variables obtained from different points in the process, the number of complete vat registers decreases due to the grouping of variables (missing data in the variables). The number of vats with incomplete data before validation increased from 3.9 to 13.9%, as the number of grouped variables increased from 2 to 6. The removal of outliers by SSC slightly increased the loss of vats (range from 5.3 to 15.9%). Considering the group of six variables, missing data by grouping alone accounted for most (87%) vat records removed. Even when the outliers from SSC and all five validation approaches (Fourier-SSC, Mahala-

Table 7. Effect of missing values upon grouping of variables and of outlier removal by the single variable simple statistical criterion (SSC) and by the combined analyses of the single variable SSC, Fourier-SSC, Mahalanobis, fat mass balance-SSC (FMB-SSC), fat retention coefficient-SSC (K_F -SSC), and yield efficiency-SSC (E_{adj-th} -SSC), on the final number of validated complete vat records (data points) for selected variables from a database of 8868 cheese vat records for 3 starters and 4 yr of production.

Group of variables ¹	After outlier removal by							
	Before outlier removal		Single variable SSC ($\bar{x} \pm 3.6$ SD)			All validation approaches		
	Number of complete vat records for variable group ²	Complete vat records lost upon grouping of variables ^{3,4} (%)	Loss of complete vat records due only to outliers ³ (%)	Remaining complete vat records for the variable group ²	Total vats lost due to missing and outlier values ³ (%)	Loss of complete vat records due only to outliers ¹ (%)	Remaining complete vat records for the variable group ²	Total vats lost due to missing and outlier values ³ (%)
FM	8784		0.9	8707	1.8	4.9	8347	5.9
FM + SA	8523	3.9	1.4	8400	5.3	5.8	8006	9.7
FM + SA + PW	8341	6.0	1.7	8188	7.7	18.1	6735	24.1
FM + SA + PW + CM	8136	8.3	1.7	7987	9.9	15.9	6730	24.1
FM + SA + PW + CM + CA	7646	13.8	1.8	7484	15.6	10.4	6728	24.1
FM + SA + PW + CM + CA + CpH	7632	13.9	2.0	7456	15.9	10.2	6725	24.2

¹FM: fat in milk, SA: starter activity pH, PW: protein in whey, CM: cheese moisture, CA: acidity at cutting time, and CpH: curd pH.

²A complete vat record is one in which there are no missing values for any of the variables constituting the group.

³With respect to the original number of production vats (8868).

⁴A vat record is lost due to the presence of missing values for at least one of the variables constituting the variable group.

nobis, FMB-SSC, K_f -SSC, and $E_{\text{adj-th}}$ -SSC) were combined, and the total loss of vats increased up to 24.4% for the group of six variables, missing data still accounted for the greatest number (57%) of lost vats. This illustrates the large impact that missing data have on the final number of complete vat data available for analysis, and suggests the need for careful selection of the variables used simultaneously for a given application. Also, the loss of over 20% of vats for as few as three variables, when using a combination of several validation approaches, suggests that the selection of a large confidence interval for the validation methods is important in order to prevent excessive removal of data.

CONCLUSIONS

Results from this study show that cheese manufacturing data should ideally be validated using a combination of complementary validation methods, each selected according to the type of variable or variables analyzed. Validation strategies for individual vat and production (group of vats) data yielded complementary results (i.e., different outlier vats) and validation of both types of data can be recommended for industrial cheesemaking data. The greatest contrast in terms of outlier detection was observed between the single variable (SSC and Fourier-SSC) and multivariable methods (Mahalanobis, FMB-SSC, K_f -SSC, and yield efficiency-SSC). The latter can detect outliers in the center of single variable data distributions, as well as in the extremes. Validation of data from productions could be used prior to validation of data from individual vats, in order to eliminate entire productions containing data that could bias outlier detection for individual vat data.

Although convenient in terms of effort and time, the sole use of the traditional probability confidence interval (SSC) has considerable limitations and could keep many abnormal values in the data or eliminate values that are within normal variable variations. The SSC could be advantageously substituted by the Fourier-SSC method when dealing with variables such as milk, whey, and cheese compositional or process parameters exhibiting seasonal effects, for more efficient outlier detection. Among the production data validation approaches, K_f -SSC was simpler to carry out and detected a larger number of outliers than FMB-SSC and yield efficiency-SSC. However, the latter two approaches may be better suited for detecting unusual variations in fat in whey or cheese moisture, respectively, because the K_f calculation does not consider these variables. Data used in this study showed that grouping of variables has great impact on the number of complete vat data available for analysis, due to the presence of missing values, and that in some cases it contributed to a

greater loss of vat data than outlier removal by all data validation methods. For this reason, when using multivariable methods that analyze single vat data (i.e., Mahalanobis), the authors recommend using groups with small numbers of variables.

After validation, cheese manufacturing data can be used to develop representative models for selected variables, such as process outputs (e.g., pH, moisture, fat, or salt in cheese) as a function of measured process inputs (e.g., milk fat and protein, starter activity) or setpoints (e.g., curd cutting speeds, cooking time, starter quantity). Because process variables and models can be used for process control, it is essential, for cheese quality, process efficiency and economics, that validation be carried out on data that will be used for model development and process analysis.

ACKNOWLEDGMENTS

This research project was supported by Agropur (Granby, QC, Canada), the Conseil des recherches en pêche et en agroalimentaire du Québec (Partnership program, project number 4065), and the Natural Sciences and Engineering Research Council of Canada (NSERC, Strategic program).

REFERENCES

- Barbano, D. M. 1990. Seasonal and regional variations in milk composition in the US. Cornell Nutrition Conference for Feed Manufacturers, Oct. 23–25, Rochester, NY.
- Bos, A., M. Bos, and W. E. van der Linden. 1992. Artificial neural networks as a tool for soft-modelling in quantitative analytical chemistry: The prediction of the water content of cheese. *Anal. Chim. Acta* 256:133–144.
- Emmons, D. B. 1993. Definition and expression of cheese yield. Pages 12–20 in *Factors Affecting the Yield of Cheese*. Special Issue no. 9301. Int. Dairy Fed., Brussels, Belgium.
- Emmons, D. B., C. A. Ernstrom, C. Lacroix, and P. Verret. 1990. Predictive formulas for yield of cheese from composition of milk: A review. *J. Dairy Sci.* 73:1365–1394.
- Emmons, D. B., and C. Lacroix. 2000. Management information systems for control of yield. Pages 55–59 in *Practical Guide for Control of Cheese Yield*. Int. Dairy Fed., Brussels, Belgium.
- Hunter, E. A., D. A. McNulty, and J. M. Banks. 1997. Statistical design and analysis of experiments in cheese technology. *Lebens. Wissen. Technol.* 30:121–128.
- Lacroix, C., P. Verret, and D. B. Emmons. 1993. Design of experiments and statistical treatment of yield data. Pages 128–150 in *Factors Affecting the Yield of Cheese*. Special Issue no. 9301. Int. Dairy Fed., Brussels, Belgium.
- Lacroix, C., P. Verret, and P. Paquin. 1996. Regional and seasonal variations of nitrogen fractions in commingled milk. *Int. Dairy J.* 4:1–15.
- Lacroix, S., S. A. Jimenez-Marquez, and D. B. Emmons. 2000. Factors affecting the accuracy of plant cheese yield estimates. Pages 99–113 in *Practical Guide for Control of Cheese Yield*. Int. Dairy Fed., Brussels, Belgium.
- Lawrence, R. C. 1993. Cheese yield potential of milk. Pages 109–120 in *Factors Affecting the Yield of Cheese*. Special Issue no. 9301. Int. Dairy Fed., Brussels, Belgium.
- Lawrence, R. C., and K. A. Johnston. 1993. Estimation of processing efficiency in commercial plants. Pages 53–57 in *Factors Affecting*

- the Yield of Cheese. Special Issue no. 9301. Int. Dairy Fed., Brussels, Belgium.
- Lolkema, H. 1993. Cheese yield used as an instrument for process control—Experience in Friesland, The Netherlands. Pages 156–197 in *Factors Affecting the Yield of Cheese*. Special Issue no. 9301. Int. Dairy Fed., Brussels, Belgium.
- Ng-Kwai-Hang, K. F., J. E. Moxely, and A. S. Marziali. 1987. Gross component of milk and yield of Cheddar cheese in some Quebec cheese factories. *Can. Inst. Food Sci. Technol. J.* 20:372–377.
- Ni, H., and S. Gunasekaran. 1998. Food quality prediction with neural networks. *Food Technol.* 52(10):60–65.
- Paquet, J., C. Lacroix, and J. Thibault. 2000. Modelling of pH and acidity for industrial cheese production. *J. Dairy Sci.* 83:2393–2409.
- USDA. 1965. Full committee report of a study conducted in 13 federal milk order markets on volume-weight conversion factors for milk. Supplement to Marketing Research Report 701, Washington, DC.
- Van Slyke, L. L., and V. W. Price. 1927. *Cheese*. Orange Judd Publishing Company Inc., New York, NY.