XIX IMEKO World Congress
Fundamental and Applied Metrology
September 6–11, 2009, Lisbon, Portugal

COMPARISON OF PRINCIPAL COMPONENT REGRESSION (PCR) AND PARTIAL LEAST SQUARE (PLS) METHODS IN PREDICTION OF RAW MILK COMPOSITION BY VIS-NIR SPECTROMETRY. APPLICATION TO DEVELOPMENT OF ON-LINE SENSORS FOR FAT, PROTEIN AND LACTOSE CONTENTS

Rocío Muñiz [1], *Miguel A. Pérez* [1], *Cristina de la Torre*[1], *Carlos E. Carleos*[2], *Norberto Corral*[2], *Jesús A. Baro*[3]

[1]Dpt. de Ingeniería Eléctrica (University of Oviedo). Gijón, Spain, rociomuve@gmail.com
[2] Dept. de Estadística e Investigación Operativa (University of Oviedo), Gijón, Spain, carleos@uniovi.es
[3]Dpto, de Ciencias Agroforestales (University of Valladolid), Palencia, Spain, jabaro@arrakis.es

Abstract – Visible and Near InfraRed (VIS-NIR) spectrometry from 400 to 1100 nm in addition to Partial Least Squares (PLS) regression or Principal Component Regression (PCR) is a very interesting method to measure several important parameters of non-homogenised fresh milk such as fat, lactose and total protein content. These parameters can be used to analyze the nutritional properties of milk and, consequently they are very important to determine the economic value of produced milk.

This paper studies and compares the potential use of PCR and PLS statistical methods to obtain the values of milk nutrients composition in milk, and present the application to the development of on-line sensors for those nutrients.

The potential of VIS-NIR spectrometry in a spectral region below 1100 nm has been studied in this paper due to working in this region, a low-cost system would be obtain.

Several fresh milk samples taken during milking process were analyzed by means of standard measurement procedures and VIS-NIR spectrometry in order to verify the capabilities and precision of proposed method.

As will be seen in next sections, this method is very interesting for fat content estimation, but it present some problems for total protein and lactose measurement, probably due to the low value of protein and lactose spans.

Keywords: Milk composition, on-line sensors, spectrometry, PCR, PLS

1. INTRODUCTION

Daily measurement of nutritional milk parameters such as total protein content, lactose concentration and fat content could be used for:

a) Cow selection and genetics improvement.

b) Cow feed tuning in order to increase economic efficiency.

c) Milk differentiation to obtain predefined values of fat content, total protein or lactose in the farm outlet.

Modern dairy farms include several control and automation systems, able to provide interesting data for farm management and improving the economical results of exploitation [6].

NIR spectrometry has been used to estimate milk composition, but previous works are referred to dry milk, homogenised milk or high cost spectrometry equipment [5, 6] or requires sampling or previous treatment of milk samples [7, 8], avoiding a cow-side final implementation.

The purpose of this work is to investigate the potential of VIS-NIR spectrometry below 1100 nm, in addition to statistical analysis by Partial Least Squares (PLS) regression or Principal Component Regression (PCR) to estimate the values of main components of fresh raw milk. Additional objective of this work is the comparison of results of PCR and PLS application to spectrometry data.

All spectrometry equipment consists of an excitation light source able to produce a continuous spectrum for all wavelengths and a photo-detection system for measuring the received light in the same light spectrum. The reduction of range of interesting light wavelengths simplifies the design of complete system and decreases the final cost because low-cost LEDs and photodiodes can be used for excitation and light detection. Moreover, photodiodes can be used without cooling systems or temperature controllers, keeping an enough Signal-to-Noise ratio.

2. MATERIALS AND METHODS

To investigate the potentiality of VIS-NIR spectrometry, several milk samples has been taken from a farm during milking (along milking and from different cows). Each milk sample is divided into two similar sub-samples and preserved using refrigeration and bronopol (2-Bromo-2-nitro-1,3-propanediol). First sub-sample is sent to a certified laboratory for composition analysis, using standard procedures, obtaining reference values for fat (TG), total protein (TP) and lactose (TL) content; second sub-sample is analyzed in our laboratory by spectrometry. Finally, results

ISBN 978-963-88410-0-1

of both analyses are compared in order to determine the capability of VIS-NIR spectrometry to estimate the milk composition. Fig. 1 shows this general procedure.

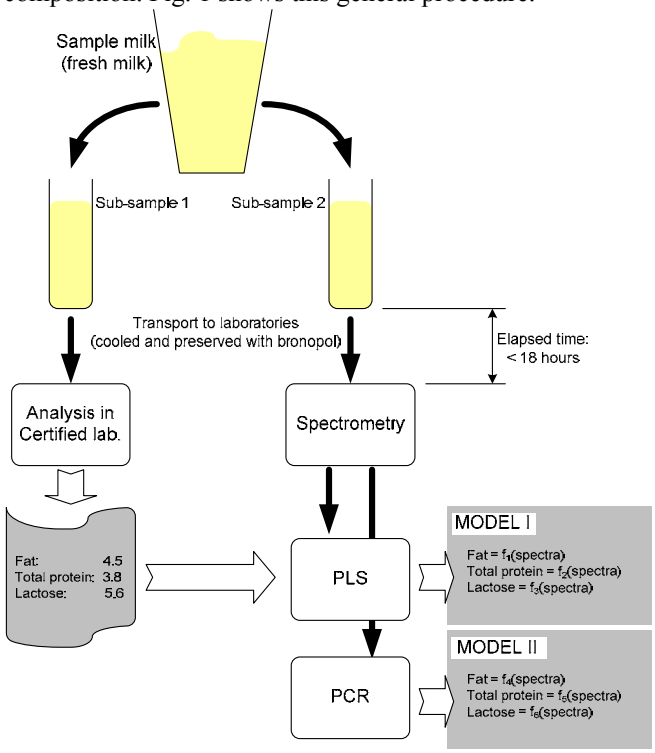

Fig. 1. Sampling and analysis procedure for fresh milk samples: each sample is divided into two sub-samples for analysis by spectrometry and PLS or PCR, and for analysis in a certified laboratory. This reference analysis has been carried out in LILA laboratory.

The analysis of each milk sample by spectrometry is carried out using a low-cost VIS-NIR spectrophotometer from Ocean Optics, able to provide 1236 values in the 400.33 to 949.59 nm, resulting in a resolution of 0.444 nm. Three different spectra are obtained by means of custom-designed analyzing cell connected to spectrophotometer and light source using several optical fibres as we can see in Fig. 2. When an appropriate excitation lamp is used, this system is able to provide orthogonal spectrum (M90) caused by scattered light, transmittance spectrum (TR) and reflectance spectrum (RE).

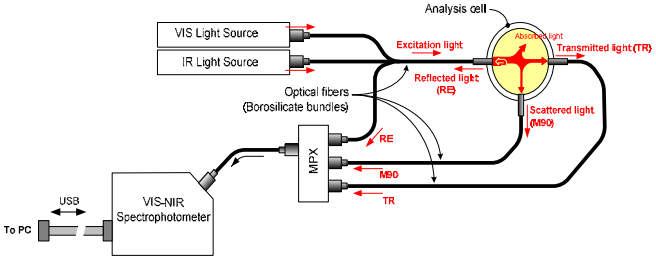

Fig. 2. Spectrum analyzer with optical fibres for obtaining three spectra (transmitted, reflected and scattered light).

All spectral data can include attenuation and disturbances introduce by light transmission path or changes in emission of light source. In order to avoid their effects, all spectra are divided by ultra-pure water spectrum, resulting in ratiometric spectra, independent on attenuation and disturbances. Fig. 3 shows these spectra.

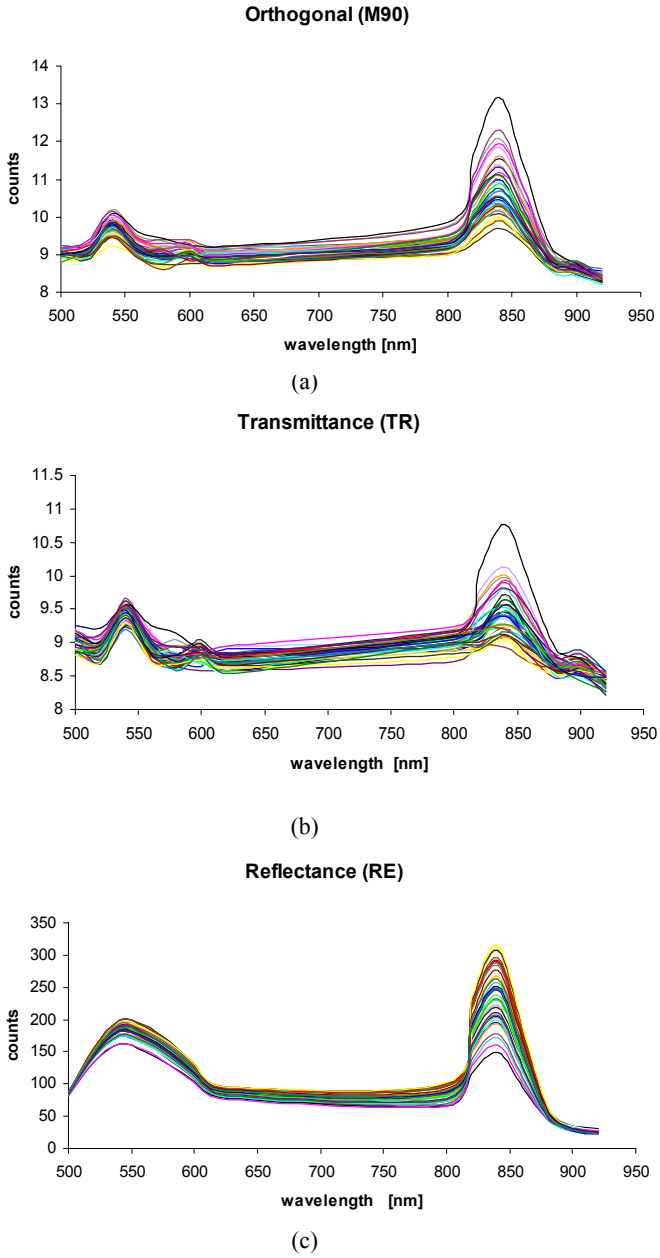

Fig. 3. Orthogonal (a), transmittance (b) and reflectance (c) spectra from fresh milk spectrum analyzer in Fig. 2.

Spectral data has been smoothed by applying iterative local linear polynomial fit with tricubic weighting [1] to redraw smooth spectra with a low resolution of 20 nm. Thus, the total number of input variables for statistical treatment is reduced and, the problem simplified, without significant data lost.

Regression-based methods are used for prediction, using TG, TP and TL as dependent variables and smoothed spectra M90, TR y RE, with 20nm of resolution as independent variables. For each value of three smoothed spectra, square and cubic terms are generated such as additional input variables to include non-linear behaviour of

model. Thus, model includes 504 input variables (56 × 3 ×3), 56 values of each spectrum, its square and cubic terms and three spectra).

Total number of input variables is lower than number of observations. So, a multivariate technique for dimensional reduction must be applied. In this work, we use two different techniques: first, a traditional Principal Component Regression (PCR) and, second, the useful PLS (Partial Least Squares). PLS was used in univariate response, that is, PLS-1 [2].

Both, PCR and PLS-1 methods are based on calculation of orthogonal components from a linear combination of original variables to reduce the total number of variables. The objective of PLS-1 is to extract the components from correlations between original independent variables and dependent variable. In our case, to choice the final components number, the average squared error of predicted values is calculated for all cases, by means of leave-one-out cross-validation. The use of R statistical environment simplifies these calculations and procedures [3].

3. EXPERIMENTAL RESULTS AND DISCUSSION

VIS-NIR spectrometry and PLS-1 is applied to quantify three of most important components of fresh raw milk. To verify the results of proposed method several experiments must be carried out. Several samples (35 samples) are taken and analyzed by following the procedure shown in Fig. 1; in all cases, they are un-homogenized fresh raw milk obtained along milking from Holstein-Friesian cows.

Three VIS-NIR spectra (RE, TL and M90) is obtained and smoothed to reduce the total number of input variables. Figs. 4 and 5 show, respectively, the average squared error of predicted value for dependent variables (TG, TL and TP) in function of used components number for PCR and PLS methods. Moreover, Table 1 shows the optimum number of used components for both methods and the percentage of explained variance.

Table 1 Comparison of PCR and PLS-1 results in prediction of milk composition. An overall interpretation could establish an excellent behaviour for prediction of fat content (it uses only one component and can explain a high percentage of variance); results are interesting for lactose content, although using many components.

Variable	*Number of components (PCR)*	*Number of components (PLS-1)*	*Explained variance (%)*
Fat content (TG)	1 ✓✓	1 ✓✓	82 ✓✓
Lactose content (TL)	11	8 ✓	62 ✓
Total protein (TP)	2	2	17

Used models allows us to predict fat and lactose content in raw milk with a high percentage of explained variance, but the results are not good enough for explaining total protein content. Fat content can be predicted with only one component, that is, it has a linear behaviour. Results of VIS-NIR spectrometry for lactose component are acceptable, but methods, PCR and PLS-1 need many input components.

The comparison conclusions of results from PCR and PLS-1 establish a better behaviour of PLS-1 in prediction of lactose content because it uses less input components, but are quite similar for other predicted variables.

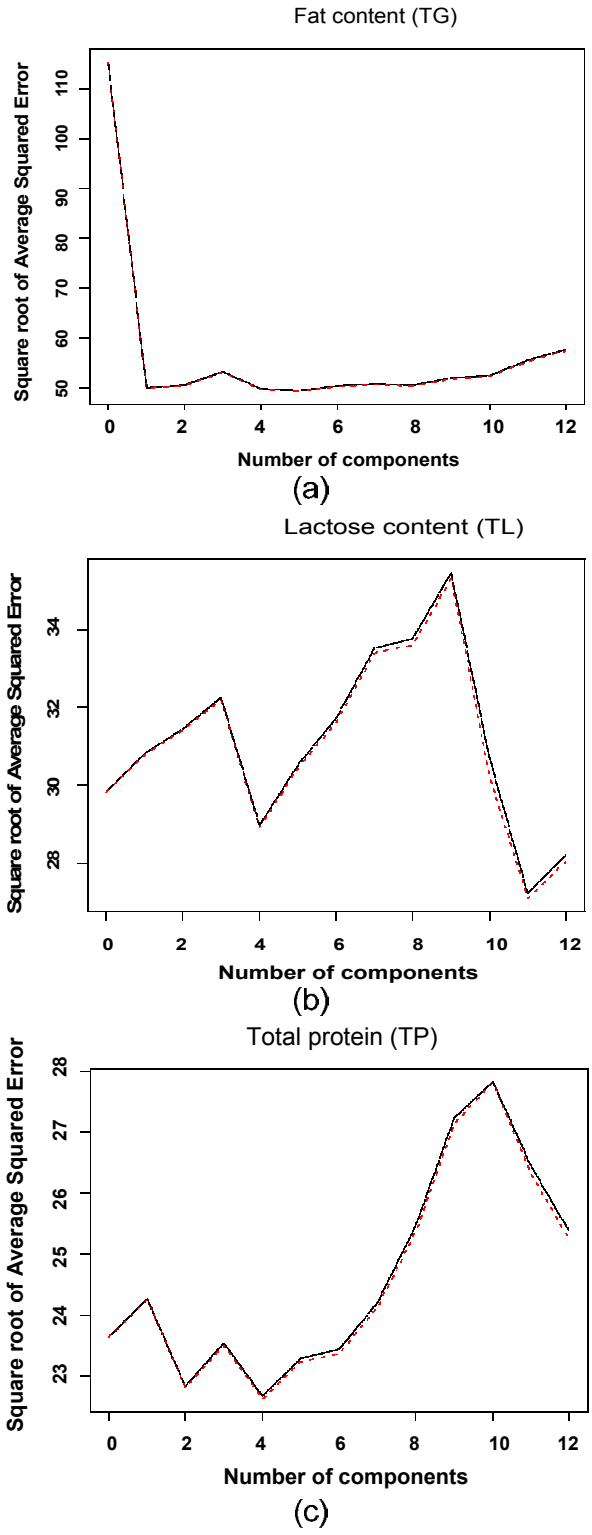

Fig. 4. Average squared error of predicted value for dependent variables, fat content, TG, lactose content, TL and total protein content, TP in function of used components number for PCR method..

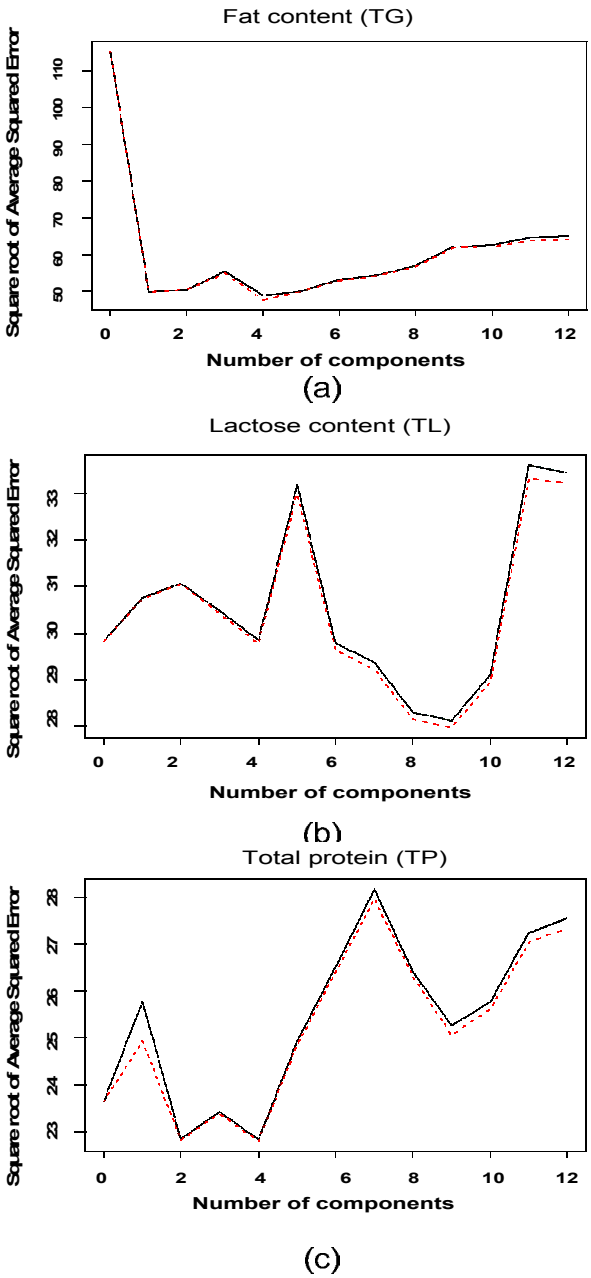

Fig. 5. Average squared error of predicted value for dependent variables, fat content, TG, lactose content, TL and total protein content, TP in function of used components number for PLS-1.

These results have been demonstrated the potentially of measurement of fat content in fresh raw milk with only one point of spectrum – a wavelength – with enough sensitivity in NIR region, where emitters and receivers have low-cost. An on-line fat sensor has been developed and full tested under laboratory and real conditions (now, it is placed in a milking parlour). This system is able to provide real-time measurement of fat control during milking. A picture of that sensor appears in Fig. 6.

This sensor has been tested in a farm from March, 2008 without any interference nor disturbance with the milking machine and other associated sub-system. In Fig. 7 we can see some examples of values provided by this sensor during milking for some cows.

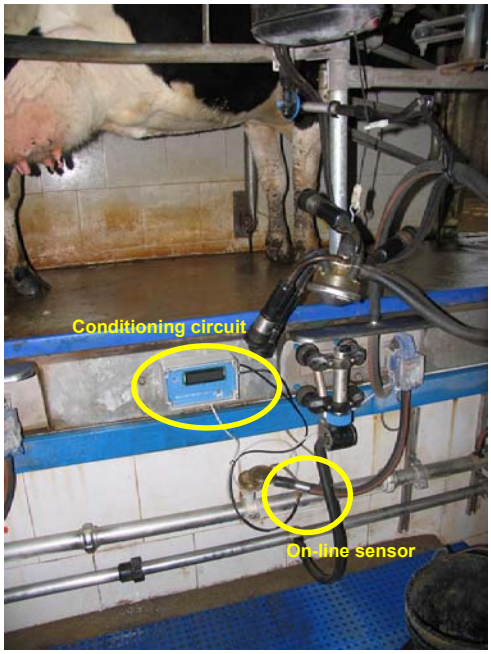

Fig. 6. On-line fat sensor placed in a parlour.

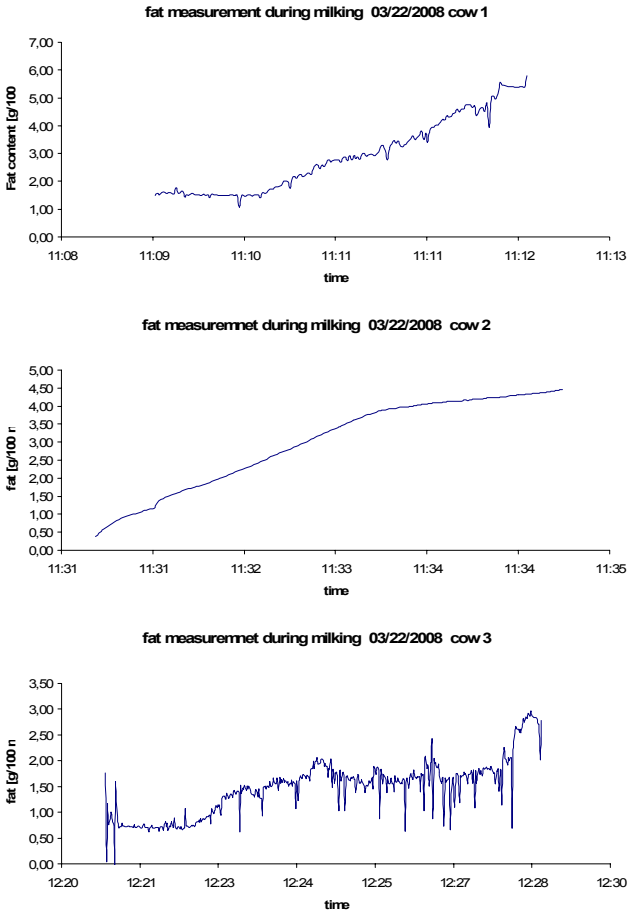

Fig. 7. Three examples of fat readout provided by the developed sensor. As we can see, output value may be noisy due to some flow milk fluctuations. In those cases, a exponential smoothing should be applied to obtain a clean signal.

4. CONCLUSIONS

This paper has investigated the potentially of VIS-NIR spectrometry and PLS method to predict the content of several important components of fresh raw milk (fat, lactose and total protein). This method results very interesting to future on-line applications in farm in relation to previous works [5, 6, 7, 8] because it can be used for un-homogenized fresh raw milk without any kind of previous treatment, homogenization or dilution, and it could be implemented with low-cost devices such as LEDs and photodiode-array in visible and near-infrared wavelengths (<1100 nm).

Experimental results have demonstrated the capability of method to predict fat and lactose content of milk with high explanation of variance. PLS-1 and PCR method produce similar results for three studied output variables (fat, lactose and total protein), although PLS-1 uses fewer input components to predict lactose content.

REFERENCES

[1] WS Cleveland Robust locally weighted regression and smoothing scatterplots. *J Amer Statist Assoc* 74:829-836, 1979

[2] H Martens, T. Naes Multivariate calibration. Wiley. 1989.

[3] R Development Core Team *R: A language and environment for statistical computing* R Foundation for Statistical Computing,

[4] BD Ripley *Pattern Recognition and neural networks*. Cambridge University Press. Cambridge. Chapter 7. 1996

[5] R. Tsenkova et al. Near-infrared spectroscopy for dairy management: measurement of unhomogenized milk composition, *J. Dairy Science* 82:2344-2351, 1999

[6] R. Tsenkova et al. Near-infrared spectroscopy for biomonitoring: cow milk composition measurement in a spectral region from 1100 to 2400 nm, *J. of Animal Sci*. 78: 515-522, 2000.

[7] I. Eshkenazi et al. A Three-Cascade Enzime Biosensor to Determine Lactose Concentration in Raw Milk, *J. Dairy Science* 83:1939-1945. 2000.

[8] Young-Ha Woo et al. Development of a New Measuring Unit for Rapid Determination of Fat, Lactose and Protein in Raw Milk Using Near Infra-Red Transmittance Spectroscopy, *Applied Spectroscopy*, No 56, Vol 5, 2002.

[9] Qi Xin et al. The rapid determination of fat and protein content in fresh raw milk using the laser light scatterinmg technology, *Optics and Lasers in Engineering,* 44, pp.858-869. 2006.