Near-Infrared Spectroscopy for Dairy Management: Measurement of Unhomogenized Milk Composition

R. TSENKOVA,* S. ATANASSOVA,† K. TOYODA,* Y. OZAKI,‡ K. ITOH,§ and T. FEARN||

*Department of Environmental Information and Bio-Production Engineering, Kobe University, Kobe 657, Japan
†Department of Mathematics and Physics, Thracian University, Stara Zagora 6000, Bulgaria
‡Department of Chemistry, Kwansei Gakuin University, Nishinomiya 662, Japan
§Laboratory of Agricultural Process Engineering, Hokkaido University, Sapporo 060, Japan
||Department of Statistical Science, University College London, London WC1E 6BT, United Kingdom

ABSTRACT

The potential of near-infrared spectroscopy to measure fat, total protein, and lactose contents of unhomogenized milk was studied for use in dairy management, as a new tool for on-line milk analysis in the process of milking. Influence of the spectral region, sample thickness, and spectral data treatment on the accuracy of determination was investigated.

Transmittance spectra of 258 milk samples, collected at different stages of the milking process, were obtained with a spectrophotometer (NIRSystems 6500; FOSS-NIRSystems, Silver Spring, MD) in the wavelength range from 400 to 2500 nm with sample thicknesses of 1 mm, 4 mm, and 10 mm.

The spectral region and sample thickness were found to be significant factors for milk fat and total protein determination but not the lactose determination. The best accuracy was obtained with the 1100 to 2400 nm region, 1-mm sample thickness, and the first derivative data transformation. For the spectral region from 700 to 1100 nm, close accuracy was obtained for fat with a 10-mm sample and for total protein with a 1-mm sample thickness. The sample thickness did not change significantly the accuracy of lactose determination. Different treatments of spectral data did not improve the calibrations for fat and protein.

For the region from 700 to 1100 nm, where inexpensive on-line sensors could be used, the highest positive coefficients for fat were at 930, 968, 990, 1026, 1076, and 1092 nm; for lactose were at 734, 750, 786, 812, 908, 974, 982, and 1064 nm; and for total protein were at 776, 880, 902, 952, and 1034 nm.

(Key words: unhomogenized milk composition, near-infrared spectroscopy)

INTRODUCTION

Milk composition is an important factor in dairy farming. Its measurement is essential for the efficient utilization of cows, cattle breeding, and dairy industry. Fat and total protein content of milk have economic importance because, in most countries, milk trade is based on these components and the somatic cell count. Daily measurement of milk composition is becoming a very important information source for dairy management decisions and could be a valuable diagnostic tool as well (15). Decrease in the milk fat content could be an indicator for imbalance in the forage to concentrates ratio in the diet (6). Lactose content of milk is usually very stable. Its deviation could be used as an additional indicator for mastitis diagnosis. Total protein content of milk is an individual parameter for each cow, which changes with and within the lactation stage and when somatic cells count increases. Protein content of milk could also be used as an energy supply indicator for each cow. Low protein content of milk indicates energy deficiency. With dairy management requirements for high milk quality and the rapid development of milking robots a clear necessity exists for a nondestructive, rapid method and instrumentation for on-line milk analysis. Near-infrared (NIR) technology is quite suitable for that purpose and especially the short wavelength (near NIR) range because of the inexpensive optical sensors available on the market.

Routine analytical methods used for milk composition measurement are destructive, expensive, time and labor consuming, and off-line by nature. A widely used instrument for milk analysis is Milko Scan (Foss-Elec-
MILK COMPOSITION MEASUREMENT

MATERIALS AND METHODS

Milk Samples

A total of 258 individual milk samples was collected from three Holstein cows twice a month for a 6-mo period. Equal portions of milk (3 L) were taken during the morning and evening milkings with a milk yield meter (True Test; Alfa Laval Agri, Tumba, Sweden). The experiment was carried out between the second and the eighth month of lactation for cows that were in their second lactation. All cows were kept on the same diet during the experiment.

All samples were analyzed simultaneously with the reference method (1) for fat, total protein, and lactose contents by Milko Scan (Foss-Electric A/S, Hillerød, Denmark).

NIR Spectra

Transmittance (T) milk spectra with path lengths of 1 mm, 4 mm, and 10 mm were obtained from all samples with a spectrophotometer (NIRSystems 6500; FOSS-NIRSystems, Silver Spring, MD) by using quartz cuvettes with 1 mm thick walls. Spectra were taken in the wavelength range from 400 to 2500 nm at 2-nm intervals and were recorded in the linked computer as absorbance [i.e., log (1/T)]. Prior to spectral analysis, each sample was warmed to 40°C in a water bath.

Spectral data were analyzed to obtain calibration models in two regions: the region from 700 to 1100 nm with 1-, 4-, and 10-mm path lengths and the region from 1100 to 2400 nm with a 1-mm path length. In the 1100 to 2500 nm region, the absorptions for 4- and 10-mm path lengths were very high, and the spectra were noisy; hence, no calibrations were performed. The spectral data for 1-mm sample thickness from 2400 to 2500 nm were also noisy and were not used.

Treatment of NIR Data

A commercial software program (Pirouette Version 2.0; Infometrics, Inc., Woodinville, WA) was used to process the data and to develop models for determination of fat, total protein, and lactose contents. Calibration models for each combination of spectral region, sample thickness, and data treatment were obtained.

The methods for data treatment included smoothing the spectral data over different intervals, from 5- to 25-point windows (10-50 nm), and first or second derivative transformation of log (1/T) data with different window sizes, from 5- to 25-point intervals. The smoothing and derivative transformations were based on the Savitzki-Golay second-order polynomial filter (13).

The calibration was performed using partial least square (PLS) regression. The PLS utilized both the spectra and the respective reference data for the examined samples to determine the latent variables (PLS factors). In the development of all calibration models, 15 PLS factors were set up as maximum. The optimum number of PLS factors used in the models was deter-
TABLE 1. Range, mean, and standard deviation (SD) of milk components in examined samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat, %</td>
<td>0.34</td>
<td>10.41</td>
<td>3.505</td>
<td>1.921</td>
</tr>
<tr>
<td>Protein, %</td>
<td>2.60</td>
<td>3.99</td>
<td>3.132</td>
<td>0.232</td>
</tr>
<tr>
<td>Lactose, %</td>
<td>3.94</td>
<td>4.74</td>
<td>4.399</td>
<td>0.163</td>
</tr>
</tbody>
</table>

mined by a cross-validation method. In cross-validation, five samples were temporarily removed from the calibration set to be used for validation. With the rest of the samples, a PLS model was developed and applied to predict the respective milk component content of the group of five samples. The results were compared with the respective reference values. This procedure was repeated several times until a prediction for all samples was obtained. Performance statistics were accumulated for each group of removed samples. The validation errors were combined into a standard error of cross validation (SECV), which was accepted as a measure of the accuracy of determination. The optimum number of PLS factors in each model was defined to be the one that corresponded to the lowest SECV.

RESULTS AND DISCUSSION

The experiment was designed to follow the changes in milk composition during lactation and during morning and evening milking processes. Results of standard chemical analysis for the examined milk components are presented in Table 1. Collected samples varied widely in milk composition and especially in fat content, which tested the ability of NIR spectroscopy to provide milk composition analysis at defined intervals in the milking process.

The spectra of the milk samples taken during morning milking of one of the examined cows are shown in Figure 1. The spectra were split into two regions at 1100 nm because of the change of detectors in NIRSystem 6500 spectrophotometer, which were silicon for the region from 700 to 1100 nm and lead sulfide (PbS) from 1100 to 2500 nm. Because of the water absorbance, strong absorption bands around 960, 1440, and 1950 nm (21) dominated the spectra. The characteristic absorption bands of fat and other milk components such as protein and lactose were very weak in comparison with the water bands and were hard to detect. The spectrum was shifting upwards with the fat content elevation.

Preliminary investigations showed that smoothing and pretreatment of milk spectral data were needed for measurement of milk composition by NIRS. The best calibration models for fat, total protein, and lactose content determination, in terms of smoothing intervals and derivative windows, based on log (1/T) data, first derivative, and second derivative transformation are given in Tables 2 to 4. The comparison of the calibration models was made on the basis of calibration and cross-validation statistics that included standard error of calibration, coefficient of multiple correlations, SECV, cross-validation correlation coefficient (CVr), and coefficient of variation (CV) = SECV/Mean value) \times 100.

We found that the accuracy of fat content determination of bovine milk depended strongly on the spectral regions and path lengths. The best results were obtained for the region from 1100 to 2400 nm with 1-mm sample thickness (Table 2). The SECV for the model based on the first derivative spectral data transformation with a 17-point derivative window was 0.110, the CVr was 0.998, and the CV was 3.15%. This accuracy of prediction was close to the accuracy of the reference method. Figure 2 illustrates the relationship between the chemical data for fat and the respective values predicted by the NIR calibration equation. The accuracy of the model, based on log (1/T) spectral data, was similar to the best first derivative model.

In the spectral region from 700 to 1100 nm, the lowest SECV and the highest CVr for fat content determination were obtained using spectral data for samples that were 10 mm thick. The model derived using log (1/T) data smoothed with windows with 13 data points had SECV = 0.191, CVr = 0.995, and CV = 5.46%. The lowest accuracy in fat determination was found for models derived using spectral data from 700 to 1100 nm of samples that were 1 mm thick.

These results were consistent with the investigation of Pierce and Wehling (8) for determination of fat con-
TABLE 2. Near-infrared spectra calibration and validation statistics for fat content determination in unhomogenized milk.

<table>
<thead>
<tr>
<th>Spectral region and sample thickness</th>
<th>Spectral data(^1) transformation</th>
<th>Smoothing interval or derivative windows</th>
<th>PLS(^2) factors</th>
<th>Calibration statistics(^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>700–1100 nm</td>
<td>Log(1/T)</td>
<td>9</td>
<td>10</td>
<td>SEC 0.310, R 0.986, SECV 0.376, CVr 0.998, CV% 10.72</td>
</tr>
<tr>
<td>1 mm</td>
<td>D1</td>
<td>9</td>
<td>9</td>
<td>0.284, 0.988, 0.347, 0.981, 9.91</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>10</td>
<td>0.236, 0.992, 0.262, 0.989, 7.48</td>
</tr>
<tr>
<td>700–1100 nm</td>
<td>Log(1/T)</td>
<td>9</td>
<td>10</td>
<td>0.172, 0.996, 0.213, 0.994, 6.09</td>
</tr>
<tr>
<td>4 mm</td>
<td>D1</td>
<td>17</td>
<td>10</td>
<td>0.178, 0.996, 0.203, 0.994, 5.78</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>8</td>
<td>0.180, 0.996, 0.203, 0.994, 5.78</td>
</tr>
<tr>
<td>700–1100 nm</td>
<td>Log(1/T)</td>
<td>13</td>
<td>10</td>
<td>0.171, 0.996, 0.191, 0.995, 5.46</td>
</tr>
<tr>
<td>10 mm</td>
<td>D1</td>
<td>13</td>
<td>10</td>
<td>0.167, 0.996, 0.193, 0.994, 5.52</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>10</td>
<td>0.221, 0.993, 0.246, 0.989, 7.03</td>
</tr>
<tr>
<td>1100–2400 nm</td>
<td>Log(1/T)</td>
<td>9</td>
<td>10</td>
<td>0.103, 0.999, 0.115, 0.998, 3.27</td>
</tr>
<tr>
<td>1 mm</td>
<td>D1</td>
<td>17</td>
<td>10</td>
<td>0.098, 0.999, 0.110, 0.998, 3.15</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>10</td>
<td>0.116, 0.998, 0.133, 0.997, 3.80</td>
</tr>
</tbody>
</table>

\(^1\)D1 = first derivative transformation, D2 = second derivative transformation, and T = transmittance.

\(^2\)Number of factors in the calibration model; PLS = partial least square.

\(^3\)SEC = Standard error of calibration, R = coefficient of multiple correlation, SECV = standard error of cross-validation, CVr = cross-validation correlation coefficient, and CV = variation coefficient = \(\frac{\text{SECV}}{\text{mean}} \times 100\).

Content in Cheddar cheese. They found that the spectral region from 1100 to 2500 nm performed better than did the spectral region from 700 to 1100 nm.

For the 700 to 2400 nm region, SECV for lactose content determination varied in a narrow range from 0.082 to 0.099, and the corresponding CV varied between 1.87 and 2.26% (Table 3). This accuracy of prediction was close to the accuracy of the reference method. The observed small total variation in the lactose content explained the relatively low CVr that ranged from 0.773 to 0.827. The best calibration model for lactose content determination was obtained when applied the first derivative transformation of log (1/T) data in the 1100 to 2400 nm region with a 1-mm sample thickness. The SECV for this model was 0.082, CVr was 0.827, and the variation coefficient was 1.87%. Figure 3 illustrates the relationship between chemical data for lactose and the respective values predicted by the NIR calibration equation. Accuracy of the model, based on the same spectral region and log (1/T) data with a smoothing interval of 17 data points, was similar. In the spectral region from 700 to 1100 nm, the best model for lactose content determination was obtained using log (1/T) data with a 4-mm sample thickness, which had the lowest SECV value of 0.089. This accuracy was very close to the results in the region from 1100 to 2400 nm.

Table 4 shows that the accuracy of total protein content determination for bovine milk depended strongly on the spectral region and the path length. The best results were obtained for the region from 1100 to 2400 nm with a 1-mm sample thickness. The SECV for the best model, based on the first derivative spectral data transformation with a 25-point derivative window, was 0.096, CVr was 0.848, and CV was 3.06%. This result showed that the accuracy of this method was close to the accuracy of the reference method. Figure 4 illustrates the relationship between the data for chemical analysis of total protein and the respective values predicted by the NIR calibration equation. The SECV values for the best models, based on log (1/T) and second derivative transformation of the spectral data, 0.115 and 0.118 respectively, were similar to the SECV of the first derivative model.

Satisfactory accuracy for determination of total protein content in the near NIR region, from 700 to 1100 nm, was found when a 1-mm sample thickness was used. The lowest SECV and the highest CVr were ob-
TABLE 3. Near-infrared spectra calibration and validation statistics for lactose content determination in unhomogenized milk.

<table>
<thead>
<tr>
<th>Spectral region and sample thickness</th>
<th>Spectral data transformation</th>
<th>Smoothing interval or derivative windows</th>
<th>PL Sp factors</th>
<th>Calibration statistics$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>700–1100 nm</td>
<td>Log(1/T)</td>
<td>13</td>
<td>7</td>
<td>SEC 0.086, R 0.813, SECV 0.092, CVr 0.778, CV% 2.09</td>
</tr>
<tr>
<td>1 mm</td>
<td>D1</td>
<td>13</td>
<td>5</td>
<td>SEC 0.087, R 0.808, SECV 0.092, CVr 0.779, CV% 2.08</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>6</td>
<td>SEC 0.085, R 0.816, SECV 0.092, CVr 0.800, CV% 2.08</td>
</tr>
<tr>
<td>700–1100 nm</td>
<td>Log(1/T)</td>
<td>21</td>
<td>10</td>
<td>SEC 0.082, R 0.829, SECV 0.089, CVr 0.784, CV% 2.02</td>
</tr>
<tr>
<td>4 mm</td>
<td>D1</td>
<td>13</td>
<td>8</td>
<td>SEC 0.083, R 0.822, SECV 0.090, CVr 0.776, CV% 2.05</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>9</td>
<td>SEC 0.084, R 0.818, SECV 0.089, CVr 0.781, CV% 2.03</td>
</tr>
<tr>
<td>700–1100 nm</td>
<td>Log(1/T)</td>
<td>5</td>
<td>6</td>
<td>SEC 0.090, R 0.840, SECV 0.096, CVr 0.809, CV% 2.18</td>
</tr>
<tr>
<td>10 mm</td>
<td>D1</td>
<td>25</td>
<td>6</td>
<td>SEC 0.091, R 0.836, SECV 0.094, CVr 0.820, CV% 2.13</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>6</td>
<td>SEC 0.091, R 0.838, SECV 0.099, CVr 0.795, CV% 2.26</td>
</tr>
<tr>
<td>1100–2400 nm</td>
<td>Log(1/T)</td>
<td>17</td>
<td>10</td>
<td>SEC 0.077, R 0.857, SECV 0.086, CVr 0.773, CV% 1.95</td>
</tr>
<tr>
<td>1 mm</td>
<td>D1</td>
<td>25</td>
<td>10</td>
<td>SEC 0.076, R 0.854, SECV 0.082, CVr 0.828, CV% 1.87</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>25</td>
<td>10</td>
<td>SEC 0.081, R 0.841, SECV 0.091, CVr 0.783, CV% 2.07</td>
</tr>
</tbody>
</table>

$^1$D1 = first derivative transformation, D2 = second derivative transformation, and T = transmittance.

$^2$Number of factors in the calibration model, PLS = partial least square.

$^3$SEC = Standard error of calibration, R = coefficient of multiple correlation, SECV = standard error of cross-validation, CVr = cross-validation correlation coefficient, CV = variation coefficient – (SECV/mean value) × 100.

The PLS modeling aided not only the development of quantitative models for milk composition determination but was also used as a tool for discerning the location of information related to fat, total protein, and lactose. Different smoothing intervals or different derivative windows gave different accuracies for fat, total protein, and lactose determinations. Models based on the second derivative data transformation tended to yield better results when using larger window sizes. The influence of the data transformation on the accuracy of fat and protein content determinations was smaller than the influence of the spectral region and sample thickness.

The CLM modeling aided not only the development of quantitative models for milk composition determination but was also used as a tool for discerning the location of information related to fat, total protein, and lactose. No information was found in the literature concerning determination of milk composition in the near NIR region. In this study, regression vectors of the calibration models were studied, based on log (1/T) spectral data in this region. High positive coefficients in the regression vector for the model of fat determination in the 700 to 1100 nm spectral region with a 10-mm sample thickness were found at 730, 776, 930, 968, 990, 1026, 1076, and 1092 nm (Figure 5). The highest positive coefficient was observed at 930 nm. The absorption at 930 nm could be associated with the C-O band of oil.

The regression vector for the best lactose determination model in the region from 700 to 1100 nm with a 4-mm sample thickness is presented in Figure 6. High positive coefficients in the regression vector were found at 734, 750, 786, 812, 908, 974, 982, and 1064 nm. The highest positive coefficient was observed at 1064 nm. High positive coefficients in the regression vector for the total protein determination model in the 700 to...
1100 nm region with a 1-mm sample thickness were found at 726, 736, 760, 776, 794, 880, 902, 952, and 1034 nm (Figure 7). The absorptions at 760, 880, and 902 nm could be associated with the N-H third overtone bands of protein and at 1034 nm with the N-H second overtone band. The peak at 952 nm was very close to the water band and probably was connected with the influences of water-soluble protein of milk on water absorption.

**CONCLUSIONS**

The NIR spectroscopy was an adequate method for determination of fat, total protein, and lactose of unhomogenized cow milk for wide range of milk composition changes, which were observed in the milking process and over one lactation.

Spectral region and sample thickness were found to be significant factors for determination of milk fat and total protein. The best accuracies for fat, total protein, and lactose content determinations were obtained with the 1100 to 2400 nm region, combined with a 1-mm sample thickness and first derivative data transformation. For the spectral region from 700 to 1100 nm, the best results for fat measurement were obtained for samples that were 10 mm thick; total protein required a 1-mm sample thickness. The different sample thicknesses in the region from 700 to 1100 nm did not have strong...
influence on the accuracy of lactose content determination. The influence of spectral data treatment on the accuracy of determining milk fat and total protein contents was less than the influence of the spectral region and the sample thickness. The best results for determination of fat, total protein, and lactose contents in the region from 700 to 1100 nm had accuracies very close to that obtained in the region from 1100 to 2400 nm. Our study showed that the short-wave NIR region could be successfully used for on-line measurement of milk composition in the process of milking for different cows and different stages of their lactation.

The wavelengths with the largest contribution for determination of constituents of milk in the short-wave NIR spectral region from 700 to 1100 nm were found to be 730, 776, 930, 968, 990, 1026, 1076, and 1092 nm for fat content; 734, 750, 786, 812, 908, 974, 982 and 1064 nm for lactose content; and 726, 736, 760, 776, 794, 880, 902, 952, and 1034 nm for total protein content.

ACKNOWLEDGMENTS

This work was supported by the Program for Promotion of Basic Research Activities for Innovative Bioscience (PROBRAIN), Ministry of Agriculture, Fisheries and Food, Japan. We thank the staff of the University Farm at Hokkaido University, Sapporo, Japan for providing cows for the experiment. The authors are grateful to GL Science Inc., Tokyo, Japan for providing the software Pirouette, Version 2.0.

REFERENCES