Multivariate characterization of chemical and physical descriptors in pulp using NIRR

Ralf J. O. Olsson, Per Tomani, Mikael Karlsson, Tomas Joseffson, Kjell Sjöberg, and Christer Björklund

ABSTRACT: The authors propose a method using diffuse-reflectance, near-infrared spectrometry (NIRR) and multivariate data evaluation to determine yield, kappa number, and content of lignin, glucose, xylose and uronic acid during kraft digestion of birch under different conditions. A factorial design established the optimal linearization algorithm for the spectrometric data. The method readily predicts yield, kappa number, and concentrations of lignin, glucose, xylose, and uronic acid from a single NIRR measurement with good correlation.

KEYWORDS: Algorithms, betula papyrifera, *bibliographies, chemical properties, data, evaluation, glucose, infrared spectroscopy, kappa number, kraft pulps, lignin, mathematical models, physical properties, uronic acids, xylose, yield.*

The characterization of pulp samples is difficult and tedious due to the complex nature of pulp. Traditional spectrometric methods such as Fourier transform infrared (FTIR) *(1–4),* near infrared spectrometry (NIR) *(2, 5–8),* and others such as nuclear magnetic resonance (NMR), cross polarized/magic angle spinning/nuclear magnetic resouance (NMR,CP/MAS-NMR) CP/MAS-NMR *(3, 9),* and (Py-GC/MS) *(10)* have had frequent application in the chemical analysis of pulp samples. These may have major disadvantages, however, including complex datastructures that make it difficult to extract the desired information and in most cases sample pretreatment before analysis. This is time consuming and adds additional error sources.

The spectral contribution from a material under analysis is in many applications less than one tenth of the total spectrum. This makes the method sensitive to matrix effects. In addition, individual constituent contributions can overlap, and intercorrelations and nonlinearity can appear. Despite these phenomena and difficulties, there are various spectrometric methods that exist in the literature.

Several authors have used NIR for the determination of chemical and physical properties of pulp samples. Their work has used either absorbance measurements at one wavelength (common linear regression) or absorbance measurements at several wavelengths (multiple linear regression [MLR]). Published results have shown the following analyses:

- Lignin, cellulose and hemicellulose in fresh Sweetgum and Lobolly pine using stepwise regression *(5)*
- Yield and cellulose content in 14 different pine species *(6)*
- Hardwood and softwood content in aspen and oak mixtures and lignin in unbleached kraft pulp mixed with cotton linters *(7)*
- Cellulose and lignin in mixed eucalypt species *(8)*.

Olsson, Karlsson, Sjöberg, and Björklund are with Department of Chemical Engineering and Technology, Chemical Technology, Royal Institute of Technology, S-100 44, Stockholm, Sweden; Tomani is with Swedish Pulp and Paper Research Institute, STFI, S-114 86, Stockholm, Sweden; and Joseffson is with Department of Pulp and Paper Chemistry and Technology, Wood Chemistry, Royal Institute of Technology, S-100 44, Stockholm, Sweden.

I. Experimental conditions

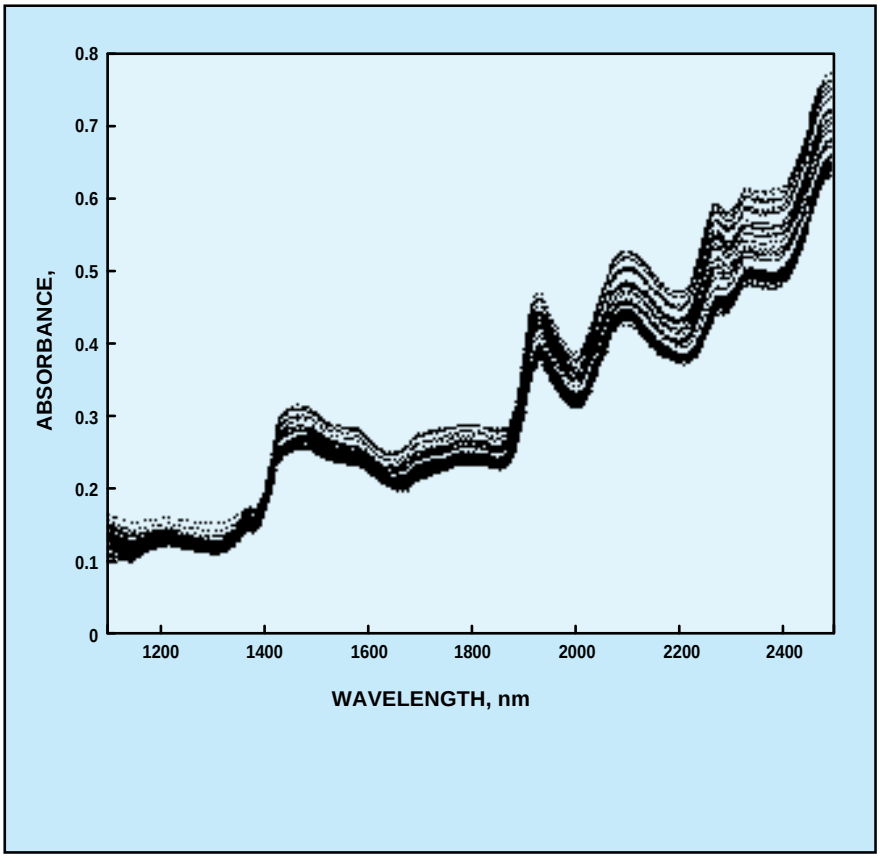
Experiment[a]	*Effective alkali, %*	*Sulfidity, %*	*Digestion temperature, T_s, °C*	*Time, t_s, from start to reach T_s, min*
A[b]	17	40	170	100
B	17	40	170	100
C	17	25	170	100
D	24	30	170	100
E	17	40	155	85
F[c]	17	40	170	53
G	17	40	170	53
R1	17	40	170	100
R2	17	40	170	100

[a] Experiments A-E started at 70°C with a temperature ramp of 1°C/min to simulate batch cooking, F-G started at 130°C with a ramp of approximately. 5°C/min to simulate continuous cooking.
[b] Digested with dried pulp (105°C)
[c] Digested with N_2 at 5 bar.

II. Mean values and ranges of the chemical composition and physical measurements

	Kappa number	*Yield, % w/w*	*Lignin, % w/w*	*Glucose, % w/w*	*Xylose, % w/w*	*Uronic acid, % w/w*
Mean value	51.5	63.8	10.6	60.0	22.6	2.2
Range	110–17.1	100–48.8	23.4–1.7	72.3–39.9	25.5–16.4	3.9–0.6

1. NIR spectra of 46 pulps with spectra containing 351 data points (wavelengths)



This method of modeling spectrometric data has drawbacks. It does not use the full information obtainable from accessible data. Modern spectrometers yield a redundancy of data—typically hundreds to thousands of variables (wavelengths or data points). The selection of wavelengths can be difficult, since the NIR spectral data can have overlapping peaks.

By using multivariate methods to evaluate spectroscopic data, it is possible to select useful data to interpret and model highly complex spectra with overlapping peaks. From the NIR domain of the spectrum, one can obtain information from the bulk of the sample due to the high penetration of the NIR beam.

In a recent paper on multivariate techniques, partial least-squares regression (PLS) assessed information from NIR spectra. The constituents modeled were lignin, glucose, and xylose in birch kraft pulp using ab-

sorbance values measured at 19 different wavelengths *(2)*.

The purpose of the present study was to develop a method to determine yield, kappa number, and content of lignin, glucose, xylose, and uronic acid content from a single NIR measurement on a pulp sample. Such a method would assume prior calibration and validation procedures on previously characterized samples. To evaluate the sensitivity of the method, this study used pulp digested under different conditions and removed after different digestion times. This pulp then underwent washing, drying, and milling before scanning with an NIR spectrophotometer. Measurement of absorbance values used 351 different wavelengths for each sample. A calibration model used the PLS algorithm. This algorithm can handle more variables (wavelengths) than samples. *A priori* selection of analytical wavelengths was not necessary. It was possible to model highly interrelated data and some nonlinearities occurring from wavelength shifts in the spectra *(11)*.

Experimental procedures

Samples of wood chips

Pure birch (*Betula papyrifera*) chips stored for approximately 1.5 years underwent drying to 94% on a dry basis followed by sieving. The accepted fraction passed 28 mm round holes and 6-mm slits but not 16-mm round holes. Bark and twigs were removed.

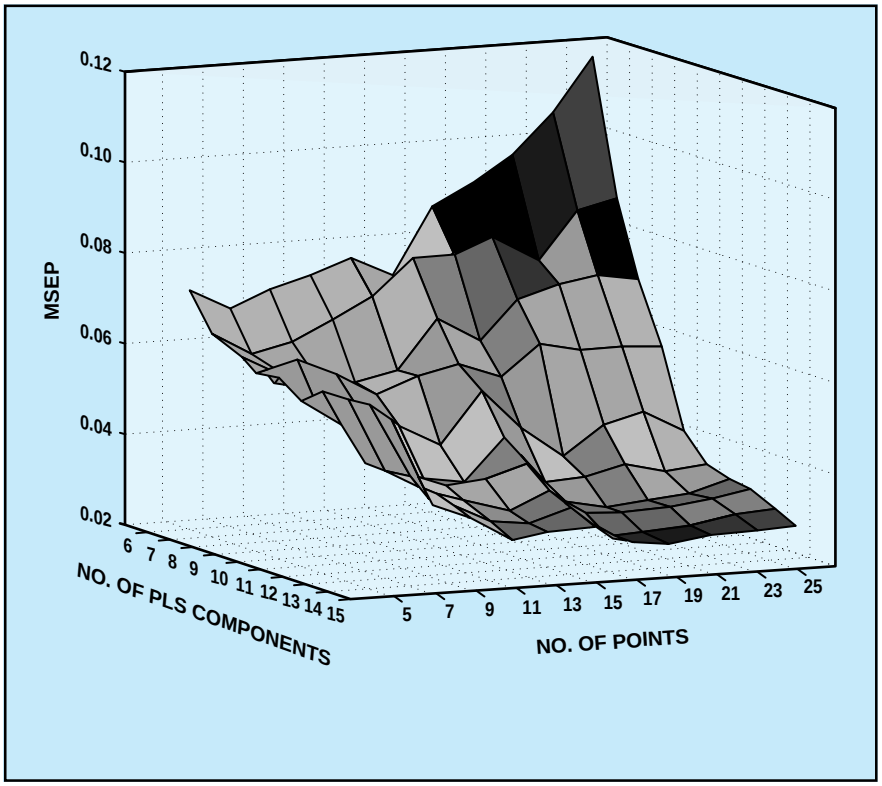
Digestion

Digestion of the chips used various cooking conditions to simulate the variability of an actual process digestion. **Table I** provides the values for the experimental conditions of the program. Two batches, R1 and R2, were identical to test reproducibility after full digestion, i.e. they were not sampled during digestion.

III. Sequence of pretreatments according to the factorial design

Model	*MSC*	*K-M*	*Derivative*
1	On	On	On
2	Off	On	On
3	On	Off	On
4	Off	Off	On
5	On	On	Off
6	Off	On	Off
7	On	Off	Off
8	Off	Off	Off

2. Error surface as a function of MSEP, number of PLS-components, and number of adjacent points in the derivative for the modeling of uronic acid



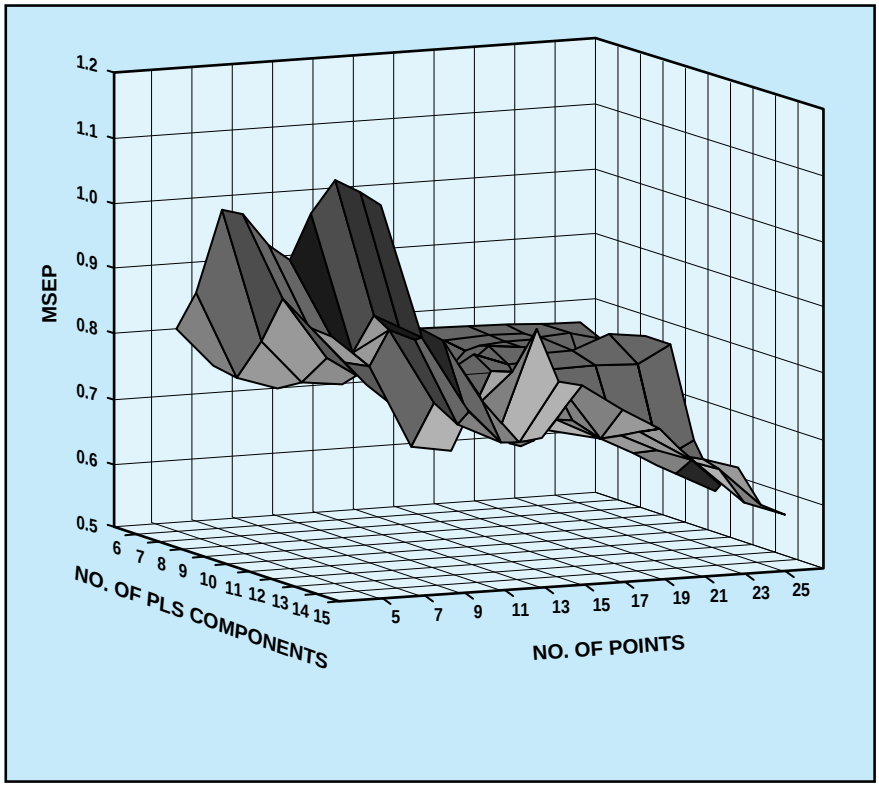
The digesters were laboratory autoclaves of 2.5-dm^3 volume each holding 350 g of dry chips. The ratio of wood to liquor was 1:4 in all experiments. A polyethylene glycol bath heated the digesters. Chip removal from the filled autoclaves occurred at consecutive times during the digestion. Analysis of the produced pulp provided values for yield, kappa number, and amounts of lignin, glucose, xylose, and uronic acid. **Table II** provides the mean values and ranges for the analyses of the 46 samples.

IV. Comparison of PLS models and previously reported MLR models

	Experimental	*Ref. 5*[a]	*Ref. 6*[b]	*Ref. 7*[c]	*Ref. 8*[d]
r^2:					
Yield	0.997	-	0.891	-	-
Kappa number	0.987	-	-	0.932	-
Lignin	0.995	0.815	-	0.973	0.704
MSEP:					
Yield	0.43	-	93.13	-	-
Kappa number	10.95	-	-	142.44	-
Lignin	0.34	86.08	-	12.47	93.54

[a]Using 1680, 1908, 2040, and 2268 nm (or the closest) determined as best R^2 for all 351 wavelengths.
[b]Using 1680, 1940, 1980, 2100, and 2208 nm (or the closest).
[c]Using second derivative, 11 points, second order polynomial. Using 1680 nm for kappa number and 2144 and 2180 nm for lignin.
[d]Using 1680, 1722, 1818, and 2208 nm.

3. Error surface as a function of MSEP, number of PLS-components, and number of adjacent points in the derivative for the modeling of lignin



Physical measurement

Yield

The samples of digested pulp underwent washing with hot water for approximately 12 h, then oven drying at 40°C to a constant dry-weight of 95%. Calculation of the dry content provided a measurement for yield. The low drying temperature prevented carbohydrate decomposition.

Chemical analysis

Kappa number

SCAN-C 1:77 was the method to determine kappa number using half the specified sample amount. Determination used the ground (40 mesh) samples.

Carbohydrates

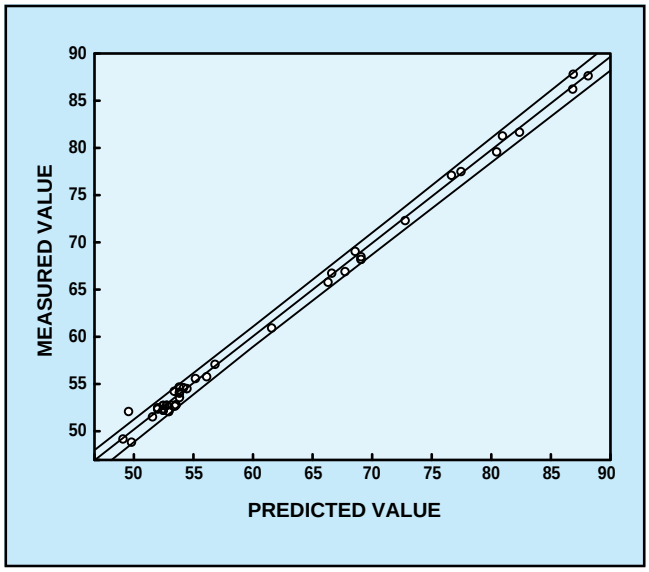
Carbohydrate analysis used the method of Theander and Westerlund *(12)*. This technique consists of hydrolysis of polysaccharides to monosaccharides followed by lignin determination, reduction of monosaccharides with potassium borohydride (KBH_4), and acetylation with acetic acid anhydride with subsequent gas chromatographic (GC) analysis.

The chromatographic determination used a 30 m x 0.32 mm (0.15 μm film thickness) fused silica capillary column in a gas chromatograph equipped with an autoinjector and a detector. All these items are commercially available. The analysis was isothermal at 210°C.

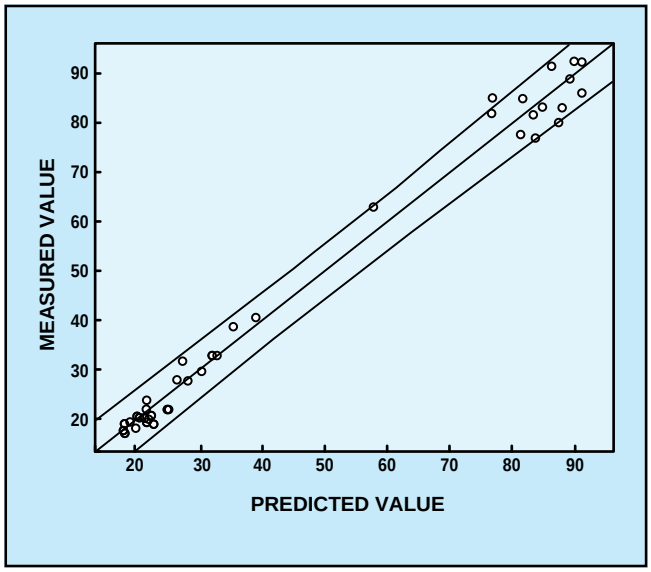
Lignin

Lignin determination was as dry solid residue after hydrolysis. This method provides a measurement of the nonacidic-soluble lignin content in the samples.

4. Plot of predicted vs. measured values for yield, $r^2 = 0.997$



5. Plot of predicted vs. measured values for kappa number, $r^2 = 0.987$



Uronic acid

Uronic acid analysis used the modified method of Bylund and Donetzhuber *(13)*. This consists of boiling the pulp in hydriodic acid (HI) and determining the difference in conductance as the resulting carbon dioxide reacts with sodium hydroxide.

NIR measurements

The NIR data came from a commercial infrared analyzer with a spectral operating range between 1100–2500 nm in even intervals of 4 nm giving 351 wavelength measurements. The samples underwent scanning after drying and grinding. **Figure 1** shows the plots for NIR measurements for 46 pulps. Each individual spectrum contains 351 data points or wavelengths. Measurement of diffuse reflectance data in the figure was as apparent absorbance where absorbance equals the logarithm of the reciprocal of the reflected light (=transmittance) Again the equipment used is comercially available.

V. Evaluation of the convolution derivative parameters showing the minimum MSEP and the corresponding number of PLS-components and number of adjacent spectral points used in the modeling

	Polynomial order: second or third		
	PLS-components in model	*No. of points in derivative*	*Minimum MSEP*
Yield	15	19	55.08
Kappa number	10	7	1175
Lignin	15	25	26.94
Glucose	15	11	17.52
Xylose	15	9	22.81
Uronic acid	15	19	1.25

Data analysis

Linearizing transform

Diffuse reflectance spectrometry does not fulfill the assumptions leading to Beer's law. For NIR data, the apparent absorbance does not correlate linearly with concentration. This is primarily due to light scatter variation that depends on the physical dimensions of the sample. The most common theories used to com-

VI. Evaluation of the factorial design with the model numbers corresponding to that of Table III

Model evaluation	*Model no.*	*PLS-components in model*	r^2	*RMSEP,%*	*SEP*	*MSEP*
Yield	4	12	0.997	0.66	1.06	0.43
Kappa number	3	13	0.987	3.35	4.85	10.95
Lignin	7	9	0.995	0.59	0.68	0.34
Glucose	2	15	0.997	0.50	0.62	0.25
Xylose	5	15	0.949	0.47	0.64	0.22
Uronic acid	3	14	0.960	0.13	0.21	0.02

pensate for this problem are the following:

- The Kubelka-Munk transform (K-M) that takes absorption and scatter into account *(14)*
- The multiplicative scatter correction (MSC) where each spectrum undergoes "correction" in both offset and slope by comparing it to the mean spectrum of the total data set *(15)*
- The second derivative of the spectra that results in a transformed spectrum consisting of only the relative changes between adjacent wavelengths where peak intensities of derivatized spectra depend more linearly on the concentration *(16)*.

Equation 1 shows the Kubelka-Munk transform:

$$A_{ik} = (K/S)_{ik} = (1 - R_{ik})^2/2R_{ik} \quad (1)$$

Equation 2 shows the MSC transform:

$$A_{ik} = (R_{ik} - \hat{a}_i)/\hat{b}_i \quad (2)$$

where

A_{ik} = transformed absorbance

R_{ik} = apparent absorbance

$\hat{a}_i$ = least squares estimate of the intercept parameter

$\hat{b}_i$ = least squares estimate of the slope parameter

i = the sample spectrum available (1, 2, . . ., 46)

k = the wavelengths available (1, 2, . . ., 351).

The second derivative is performed with the Savitzky-Golay convolution derivative *(17)*.

Since most software packages support these transforms and their implementation is optional, their use may be haphazard. The following study established the best pretreatment:

- A univariate study concerning how the model error (MSEP) arising from the derivative depends on three factors: the polynomial order (second or third) of the function fitted to the spectral data, the number of spectral points (5–25 points) used in the polynomial to calculate the derivative, and the number of "PLS-components" used in the predicting model.
- Using the best variable settings in the derivative, a factorial design model *(18)* established the optimal linearizing function or combination of functions (K-M, MSC, Derivative). This minimized the prediction errors from the modeling of the data. The factorial designs have modeled the different linearization transforms as variables at an on or off state. The design is a full 2^3 design seen in **Table III**. There was separate performance of this methodology for each pulp descriptor.

Calculation of the mean squared error of prediction (MSEP) as a function of the number of latent variables kept in the PLS model provided an evaluation of the influences of the different linearizing functions. The linearizing functions that yielded the smallest MSEP for the different descriptors were put in the subsequent PLS model. Equation 3 shows the calculation for MSEP:

$$MSEP = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{c}_i - c_i\right)^2 \quad (3)$$

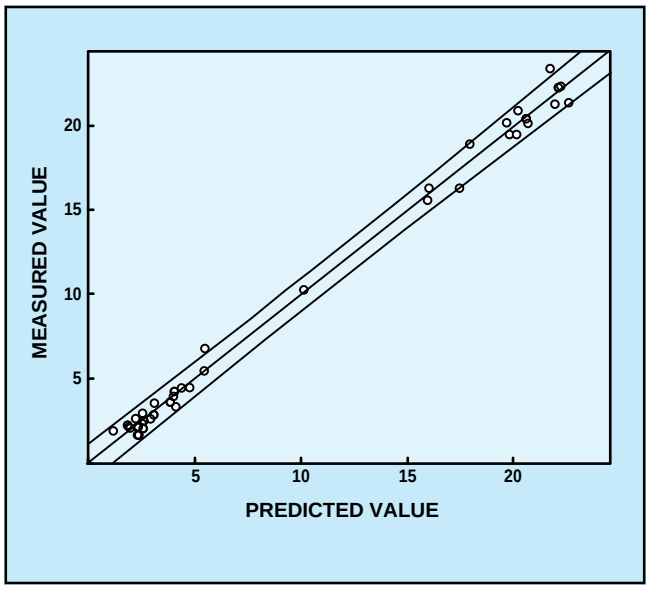
where

n = the number of samples (46)

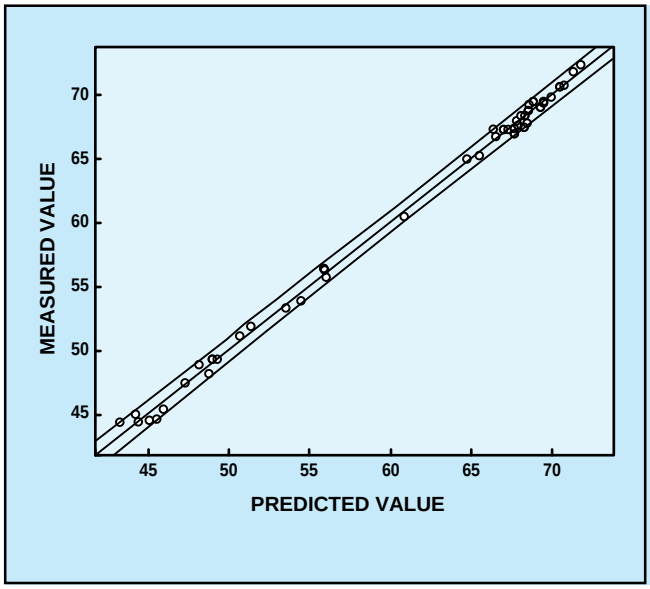
$\hat{c}_i$ = modeled descriptor value

c_i = the traditionally measured descriptor value

6. Plot of predicted vs. measured values for lignin, $r^2 = 0.995$



7. Plot of predicted vs. measured values for glucose, $r^2 = 0.997$



i = the sample number (1, 2, . . ., 46).

Mathematical modeling

This work has used commercial software throughout for the numerical calculations *(20)*. The PLS-algorithm used for modeling the relationships between the spectra and descriptors is a customized function based on the NIPALS-algorithm*(21)*. The PLS-algorithm falls under the principal components type of modeling described in detail by Geladi *et al. (22)*.

The following is one definition of principal components technique. "Principal component analysis (is) a mathematical procedure for resolving sets of data into orthogonal components whose linear combinations approximate the original data to *any* desired degree of accuracy. As successive components are calculated, each component accounts for the *maximum* amount of residual variance in the set of data. In spectroscopy, the data are usually spectra, and the number of components are smaller than or equal to the number of variables or the number of spectra, whichever is less" *(23)*.

All models have one descriptor correlated to the spectral data (PLS1). The convergence criterion for establishing the inner relations was 1×10^{-10} or 100 iterations, whichever occurred first. The method for establishing the significant number of PLS-components was cross-validation *(24)* (Jack-knifing) with one sample omitted. The predicted samples are not in the modeling set. The descriptor values were mean-centered and scaled to unit variance before modeling (autoscaling or z-transform).

There was a comparison of the different suggested mathematical models. The correlation was with PLS using the full data-set to MLR at the wavelengths suggested by Schultz *et al.*, Wright *et al.*, Easty *et al.*, and Garbutt *et al. (5–8)*. **Table IV** shows the results where applicable.

Results and discussion

The method does not seem sensitive to different digestion conditions, since no systematic prediction errors occur. The evaluation of the MSEP to establish the polynomial order and number of points resulted in the two types of error surfaces of **Figs. 2 and 3**. The error surface of Fig. 2 resembles that of the modeling for yield, glucose, and xylose. **Table V** lists the number of points for the minimum MSEP and the corresponding MSEP values. Note that second and third order polynomials yielded identical results.

Table VI shows the results from the modeling with the best linearizing function or functions for the different descriptors with the correlation coefficient (r^2) *(25)*, standard error of prediction (SEP), root mean squared error of prediction (RMSEP), and MSEP values. Equations 4 and 5 show the calculations for SEP and RMSEP respectively:

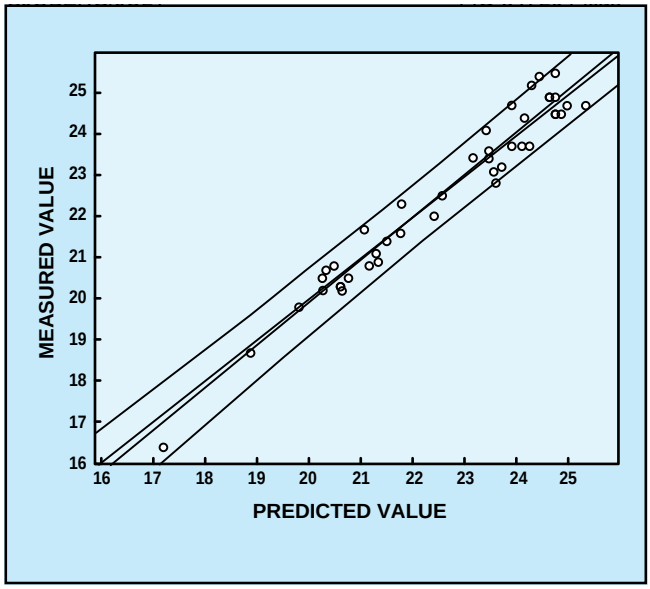
$$SEP = \{[(n-1)^{-1} \sum_{i=1}^{n} [\hat{c} - c_i - (\hat{c} - c)]^2\}^{1/2} \quad (4)$$

$$RMSEP = (MSEP)^{1/2} \quad (5)$$

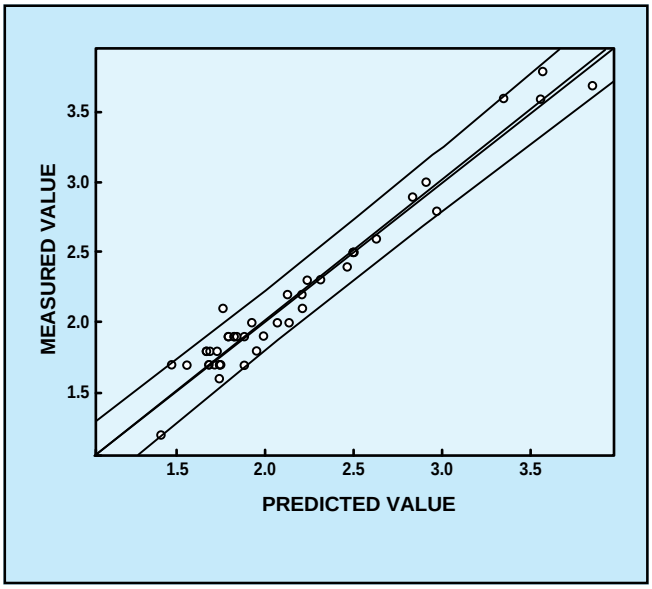
Figures 4–9 show plots of the measured vs. modeled values for the different descriptors with a 90% confidence interval for the prediction (dispersion bands).

The results of the (univariate) derivative linearization study shown in Table III indicate that the order of polynomial (second or third) has no apparent significance. This is due to

8. Plot of predicted vs. measured values for xylose, $r^2 = 0.949$



9. Plot of predicted vs. measured values for uronic acid, $r^2 = 0.960$



the very smooth and noise-free spectra obtained. The fitted mathematical function does not have to be of an order higher than two.

The number of spectral points used in the derivative necessary for an "optimal" model will generally increase with model complexity. A model with many (10–15) principal components gains from many points (13–25) as Figs. 2 and 3 indicate. This is probably related to noise reduction of the models' loadings that contain the critical information in the prediction step.

The evaluation of the different linearization sequences implies that there is no apparent "best" combination of linearizations in a general sense. The improvement of using the "optimal" pretreatment sequence shown in Figs. 4–9 for this data set ranges from 19–240% calculated for the MSEP compared to using raw data. One should also bear in mind that the "optimal" sequence will depend on the type of spectral data *(26)*. It therefore is possible to conclude that the linearization of the spectral data is crucial to model performance.

The comparison between the models suggests that the PLS-model has an increased predictability compared to MLR. The PLS-model also has built-in self-diagnostics to check whether an unknown spectrum belongs to the model. This is often important, since extrapolation of the models (PLS and MLR) due to unknown interferences, instrumental errors, or unusually high or low concentration can result in large erroneous results.

An interesting result is that it is possible to measure lignin accurately at high concentrations where the kappa number determination is less accurate. This suggests that the proposed spectrometric method would be a more accurate method to determine lignin content. Another interesting result is the prediction of yield. This is traditionally a physical descriptor measured by gravimetric measurement of samples.

Conclusions

The method proposed in this paper can replace the traditional methods of analysis for yield, kappa number, and levels of lignin, glucose, xylose, and uronic acid content in birch kraft pulp. The PLS-model is more rigid and accurate than the previously proposed MLR-model. The method does not seem sensitive to different digestion conditions. The linearization procedure of the spectral data affects the model performance. An obvious advantage to the method is that the time needed for the NIR analysis and subsequent prediction step is about one minute with a standard spectrophotometer. TP

Literature cited

1. Backa, S. and Brolin, A., *Tappi J.* 74(5): 218(1991).
2. Wallbäcks, L., Edlund, U., Nordén, *et al.*, *Tappi J.* 74(10): 201(1991).
3. Wallbäcks, L., "Pulp Characterization Using Spectroscopy and Multivariate Data Analysis," Ph.D. thesis, University of Umeå, Umeå, Sweden, 1991.
4. Friese, M. I. and Banerjee, S., *Appl. Spect.* 46(2): 246(1992).
5. Schultz, T. P. and Burns, D. A., *Tappi J.* 73(5): 209(1990).
6. Wright, J. A., Birkett, M. D., and Gambino, M.J.T., *Tappi J.* 73(8): 164(1990).
7. Easty, D. B., Berben, S. A., DeThomas, F. A., *et al.*, *Tappi J.* 73(10): 257(1990).
8. Garbutt, D.C.F., Donkin, M. J., and Meyer, J. H., *Paper Southern Africa* 12(2): 45(1992).
9. Wallbåcks, L., Edlund, U., Norden, B., *J. of Wood Chemistry and Technology* 9(2): 235(1989).
10. Kleen, M., "Characterization of Wood and Pulp Using Analytical Pyrolysis and Multivariate Data Analysis," Ph.D. thesis,

Royal Institute of Technology, Stockholm, Sweden, 1993.
11. Isaksson, T. and Naes, T., *Appl. Spect.* 42(7): 1273(1988).
12. Theander, O. and Westerlund, A. E., *J. Agric. Food Chem.* 34(2): 330(1986).
13. Bylund, M. and Donetzhuber, A., *Svensk Papperstidning* 71(15): 505(1968).
14. Kubelka, P. and Munk, F., *Z. Tech. Physik* 12: 593(1931).
15. Geladi, P., MacDougall, D., and Martens, H., *Appl. Spect.* 39(3): 491(1985).
16. O'Haver, T. C. and Begley, T., *Anal. Chem.* 53: 1876(1981).
17. Savitzky, A. and Golay, M.J.E., *Anal. Chem.* 36(8): 1627(1964).
18. Box, G.E.P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, Wiley, New York, 1978, pp. 306-352.
19. Martens, H. and Naes, T., *Multivariate Calibration*, Wiley, Chichester, 1989, p. 250.
20. *MATLAB*, The Mathworks Inc., Natick.
21. Wold, H., *Research Papers in Statistics*, (F. David, Ed.), Wiley, New York, 1966, p. 411-444.
22. Geladi, P. and Kowalski, B. R., *Anal. Chim. Acta*. 185: 1(1986).
23. ASTM E131-90, *Standard Definitions of Terms and Symbols Relating to Molecular Spectroscopy*.
24. Wold, S., *Technometrics* 20(4): 397(1978).
25. Draper, N. R. and Smith, H., *Applied Regression Analysis*, Wiley, New York, 1966, pp. 33-35,57.
26. Olsson, R.J.O., in *Near Infra-Red Spectroscopy* (K. I. Hildum, T. Isaksson, T. Naes, *et al*., Eds.), Ellis Horwood Ltd., Chichester, 1992, p. 103.

The authors thank Assi AB, Stora AB, MoDo AB, Södra AB, and SCA AB for their financial support of the delignification project. They also thank Nils Hartler and Göran Gellerstedt for the initial introduction to the field of pulping and Bo Karlberg for useful discussions and help.

Received for review July 16, 1994.

Accepted Dec. 16, 1994.