

# Determination of carbon fraction and nitrogen concentration in tree foliage by near infrared reflectance: a comparison of statistical methods

Katherine L. Bolster, Mary E. Martin, and John D. Aber

**Abstract:** Further evaluation of near infrared reflectance spectroscopy as a method for the determination of nitrogen, lignin, and cellulose concentrations in dry, ground, temperate forest woody foliage is presented. A comparison is made between two regression methods, stepwise multiple linear regression and partial least squares regression. The partial least squares method showed consistently lower standard error of calibration and higher  $R^2$  values with first and second difference equations. The first difference partial least squares regression equation resulted in standard errors of calibration of 0.106%, with an  $R^2$  of 0.97 for nitrogen, 1.613% with an  $R^2$  of 0.88 for lignin, and 2.103% with an  $R^2$  of 0.89 for cellulose. The four most highly correlated wavelengths in the near infrared region, and the chemical bonds represented, are shown for each constituent and both regression methods. Generalizability of both methods for prediction of protein, lignin, and cellulose concentrations on independent data sets is discussed. Prediction accuracy for independent data sets and species from other sites was increased using partial least squares regression, but was poor for sample sets containing tissue types or laboratory-measured concentration ranges beyond those of the calibration set.

**Résumé :** Les auteurs présentent une nouvelle évaluation de la spectroscopie de la réflectance dans le proche infrarouge comme méthode pour mesurer la concentration d'azote, de lignine et de cellulose dans le feuillage moulu et sec des plantes ligneuses de la forêt tempérée. Deux méthodes de régression sont comparées : la régression linéaire séquentielle et la régression basée sur les moindres carrés partiels. La méthode basée sur les moindres carrés partiels produisait toujours une erreur standard de calibration plus faible et une valeur de  $R^2$  plus élevée dans le cas des équations de première et de seconde différences. L'équation de première différence obtenue avec la régression basée sur les moindres carrés partiels produisait une erreur standard de calibration et une valeur de  $R^2$  de 0,106% et de 0,97 pour l'azote, de 1,613% et 0,88 pour la lignine et de 2,103% et 0,89 pour la cellulose. Les quatre longueurs d'onde les plus étroitement corrélées dans la région du proche infrarouge, ainsi que les liens chimiques auxquels elles sont associées, sont illustrés pour chaque constituants et chacune des deux méthodes de régression. La possibilité de généraliser les deux méthodes pour prédire la concentration d'azote, de lignine et de cellulose à partir d'ensembles de données indépendantes est discutée. La précision de la prédiction pour des ensembles indépendants de données et des espèces provenant d'autres sites a été améliorée par l'utilisation de la régression basée sur les moindres carrés partiels. Par contre, elle était faible pour des ensembles d'échantillons contenant des types de tissus ou des gammes de concentrations autres que ceux utilisés pour la calibration.

[Traduit par la Rédaction]

## Introduction

Near infrared reflectance spectroscopy (NIRS) uses reflectance signals resulting from bending and stretching vibrations in bonds between carbon, nitrogen, hydrogen, and oxygen to measure the concentration of major classes of chemical compounds in organic materials (Wetzel 1983; Wessman 1990). NIRS has been shown to be an accurate, precise, and rapid alternative method to wet chemistry

procedures for determining the protein, fiber, moisture, and oil concentrations of agricultural products (Norris et al. 1976; Shenk et al. 1981; Winch and Major 1981), and for determining protein, lignin, and cellulose concentrations in woody plant foliage (Wessman et al. 1988; McLellan et al. 1991). Calibration is required to correlate the spectral response of each sample at individual wavelengths with the known chemical concentrations from laboratory analyses. Varying by constituent, accuracy in NIRS analysis is dependent upon accuracy of the laboratory procedures, and upon measurement of the same chemical components by both NIRS and laboratory methods.

Stepwise multiple linear regression (MLR) of first or second difference spectra is the method most generally used for developing these calibration equations for native

Received November 3, 1994. Accepted October 19, 1995.

**K.L. Bolster, M.E. Martin, and J.D. Aber.** Complex Systems Research Center, Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, NH 03824, U.S.A.

woody plants (Wessman et al. 1988; McLellan et al. 1991). Although it has been used successfully, MLR has the following limitations: (i) the large number of wavelengths available for inclusion in calibration equations compared with the number of samples and the number of major constituents (Lindberg et al. 1983); (ii) the nonadditive behavior of spectra of pure constituents versus a mixture of the same constituents (Aber et al. 1994; Lindberg et al. 1983; Sjoström et al. 1983); (iii) the extensive spectral overlap of individual chemical constituents, or the multicollinearity problem (Lindberg et al. 1983; Lorber et al. 1987; Martens and Jensen 1983; Naes and Martens 1984; Otto and Wegscheider 1985; Wold et al. 1984); (iv) the loss of information and resulting increase in signal to noise ratio when reducing all the available data to a few selected wavelengths for the calibration equation (Martens and Jensen 1983); (v) the time required to evaluate the many possible combinations of wavelengths for each constituent to determine the best calibrations (Lindberg et al. 1983; Martens and Jensen 1983; Westerhaus 1989); and (vi) the potential that the best fitting wavelength combination fits the random errors as well as the model (Westerhaus 1989).

An alternative technique, partial least squares (PLS) regression, uses singular-value decomposition on the full spectrum of data to reduce the data to a smaller set of independent variables, or factors, as in principal component regression (PCR). Unlike PCR, however, PLS calculates for each principal component the loadings, or weights for each wavelength, using information on chemical concentrations, so that the variation relevant to modeling of the chemical variation in the data is described in the first PLS factors (Martens and Naes 1987). The advantage of PLS over PCR occurs if data noise, such as light scatter, causes greater variation in the data than chemical variation. By incorporating the chemical variability into the first factors and the data noise into the later, less important factors (Martens and Naes 1987), PLS often produces better results with fewer factors than PCR (Naes and Martens 1984). It also avoids the problems of wavelength selection and the apparent arbitrary nature of calibration development using MLR. PLS regressions result in one loading, or score, per wavelength for each principal component used in the calculations. The scores are regressed against the laboratory values to produce one coefficient value per wavelength for the final equation.

Few studies have compared various NIR calibration methods for predictability of chemical constituents, and none of these have been for foliage of native woody plants. Research with agricultural products shows consistently higher predictive power using equations developed with PLS rather than PCR or MLR (Martens and Jensen 1983; Naes and Martens 1985; Otto and Wegscheider 1985; Lindberg et al. 1983; Shenk and Westerhaus 1991b).

For any method there is a requirement that the range of constituent wet chemical values in the calibration samples represent all future unknown samples of the same type (Martens and Jensen 1983). The most important consideration (Card et al. 1988) is incorporation of a wide range of variation in the calibration data set in order to build that variation into the calibration equation. Accuracy

in prediction of unknown samples is dependent upon including the full range of variability in the factors affecting the chemical and physical properties of the unknown samples (Marten et al. 1983; Shenk 1989).

As part of the National Aeronautics and Space Administration (NASA) Accelerated Canopy Chemistry Program (ACCP), we have developed the largest data set available to date of foliage from native woody plants (Maine, New Hampshire, Massachusetts, and Wisconsin) analyzed for carbon fractions and nitrogen (Newman et al. 1994). The purposes of this paper are (i) to further evaluate the accuracy of NIRS predictions for dried, ground, green woody foliage; (ii) to compare two regression methods, MLR and PLS, and determine the best analytical procedure for green foliage of native woody plants; and (iii) to describe the generalizability of the method to predict protein, lignin, and cellulose concentrations for species with laboratory-measured values outside the range of the calibration samples.

## Materials and methods

### Sample collection

A total of 558 samples of fresh deciduous and conifer canopy foliage were collected in June 1992 at Harvard Forest, Petersham, Massachusetts (188 samples); and Blackhawk Island, Wisconsin Dells, Wisconsin (182 samples); and in September 1992 at Howland Experimental Forest, Howland, Maine (188 samples). These foliage samples, from a total of 16 deciduous and 10 conifer species, were collected in conjunction with remote sensing overflights of the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) (Vane et al. 1993) and are used as the calibration data set in this study, referred to as the 92AVIRIS data set (Table 1). All 558 samples were analyzed in the laboratory using standard wet chemical and combustion methods for nitrogen, lignin, and cellulose concentrations (Newman et al. 1994).

Two additional data sets were used for validation in this study. A total of 97 samples from 9 deciduous and 4 conifer species were collected at the same Harvard Forest site in 1993 and analyzed by the same wet chemical and combustion methods in the same laboratory (93HF data set, Table 2). The sampling site and all but one of the species in the 93HF data set were included in the calibration set. Chemical concentrations for all three constituents fall within the constituent ranges for the 92AVIRIS data set (Table 3).

The third data set consists of 31 samples of leaf, wood, and root tissues collected from sites throughout North America and the tropics as part of the National Science Foundation (NSF) Long-term Ecological Research (LTER) intersite decomposition study. All samples were again analyzed by the same wet chemistry and combustion procedures in the same laboratory (Newman et al. 1994). A subset of the LTER data set, LTERFOL, consisted of only the foliage samples from 14 woody plants. The samples in the LTER data set encompass variability in tissue types as well as sampling sites not represented in the calibration data set. The range of laboratory values for all three constituents exceeds that of the calibration data set (Table 3). The samples in the LTERFOL subset, while similar in tissue type, include only six of the species represented in the calibration data set. The minimum percent nitrogen laboratory value for LTERFOL is outside the range of the calibration data, while both the minimum and maximum values for lignin also fall outside the calibration range (Table 3).

**Table 1.** Laboratory reference value statistics by species for 1992 AVIRIS data set.

	n	% nitrogen				% lignin				% cellulose			
		Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
<i>Abies balsamea</i> (L.) Mill.	3	1.00	1.50	1.19	0.27	23.06	30.17	26.43	3.57	30.19	31.78	31.21	0.88
<i>Acer pensylvanicum</i> L.	1	1.66	1.66	1.66		20.46	20.46	20.46		36.88	36.88	36.88	
<i>Acer rubrum</i> L.	94	1.02	2.74	1.63	0.31	13.83	21.56	18.15	1.48	24.20	37.68	30.33	2.73
<i>Acer saccharum</i> L.	30	1.24	2.60	2.08	0.40	15.38	20.68	17.57	1.19	30.50	46.00	38.99	4.68
<i>Betula lenta</i> L.	1	2.20	2.20	2.20		18.31	18.31	18.31		40.98	40.98	40.98	
<i>Betula papyrifera</i> Marsh.	30	1.21	2.64	1.95	0.43	17.55	25.45	21.61	2.08	29.45	38.89	34.11	2.65
<i>Betula populifolia</i> Marsh.	1	2.98	2.98	2.98		25.66	25.66	25.66		34.40	34.40	34.40	
<i>Carya ovata</i> (Mill.) K. Koch	2	2.74	2.76	2.75	0.01	14.43	16.35	15.39	1.36	53.99	54.00	53.99	0.01
<i>Fagus grandifolia</i> Ehrh.	14	1.59	2.82	2.00	0.46	13.90	30.41	24.21	6.47	37.35	49.47	44.01	3.83
<i>Fraxinus americana</i> L.	10	1.81	2.77	2.15	0.26	15.04	20.24	17.62	1.54	51.96	67.57	55.73	4.51
<i>Larix laricina</i> (Du Roi) K. Koch	10	2.25	2.67	2.47	0.16	25.67	29.33	29.71	3.07	27.24	32.11	31.75	2.96
<i>Picea abies</i> (L.) Karst.	20	0.98	1.45	1.24	0.12	21.88	28.10	24.58	1.93	36.90	44.97	40.71	2.22
<i>Picea glauca</i> (Moench) Voss	1	1.09	1.09	1.09		24.33	24.33	24.33		40.82	40.82	40.82	
<i>Picea mariana</i> (Mill.) BSP	5	0.90	1.04	0.95	0.05	20.33	22.65	21.99	0.94	33.17	35.97	34.28	1.21
<i>Picea rubens</i> Sarg.	50	0.69	1.24	0.93	0.09	21.16	31.05	26.89	2.16	31.76	48.12	40.18	2.31
<i>Pinus resinosa</i> Ait.	31	0.93	1.41	1.14	0.18	23.97	29.64	26.48	1.38	36.75	41.16	38.91	1.47
<i>Pinus strobus</i> L.	40	1.08	2.09	1.54	0.29	23.25	31.00	26.43	1.80	33.38	49.89	38.50	3.04
<i>Populus grandidentata</i> Michx.	10	1.79	2.58	2.29	0.32	22.95	32.20	27.88	2.57	30.59	39.74	34.97	2.53
<i>Populus tremuloides</i> Michx.	7	1.76	2.40	2.09	0.25	23.10	28.55	26.55	1.75	27.53	36.03	32.46	2.91
<i>Prunus serotina</i> Ehrh.	15	2.40	3.19	2.63	0.26	13.81	20.95	17.79	2.08	35.86	52.12	41.07	3.95
<i>Quercus alba</i> L.	29	2.30	3.06	2.62	0.22	18.00	27.00	21.71	2.19	37.26	49.55	43.64	2.43
<i>Quercus rubra</i> L.	94	1.93	3.09	2.39	0.42	20.49	33.41	26.38	2.97	34.07	48.29	41.85	3.32
<i>Thuja occidentalis</i> L.	4	1.09	1.43	1.27	0.15	25.12	27.11	26.23	0.85	31.01	37.83	34.64	2.83
<i>Tilia americana</i> L.	22	2.76	3.51	3.11	0.22	14.07	24.28	19.42	2.56	41.63	54.41	47.38	3.41
<i>Tsuga canadensis</i> (L.) Carrière	46	0.83	1.39	1.08	0.26	13.75	20.69	16.64	4.13	23.69	38.87	29.11	7.24
<i>Ulmus rubra</i> Muhl.	5	2.43	3.34	2.97	0.34	12.42	16.54	14.58	1.48	41.20	52.26	46.52	4.35

**Table 2.** Laboratory reference value statistics by species for 1993 Harvard Forest data set.

	n	% nitrogen				% lignin				% cellulose			
		Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
<i>Acer pensylvanicum</i> L.	1	1.69	1.69	1.69		18.60	18.60	18.60		40.93	40.93	40.93	
<i>Acer rubrum</i> L.	14	1.35	2.03	1.69	0.18	16.43	22.85	19.17	1.83	24.91	35.66	30.92	2.67
<i>Betula alleghaniensis</i> Britt.	7	2.09	2.60	2.30	0.19	22.86	26.25	24.74	1.42	35.80	41.46	37.52	2.09
<i>Betula lenta</i> L.	7	2.13	2.50	2.31	0.15	23.63	28.62	26.47	1.75	33.88	45.57	39.97	3.60
<i>Betula papyrifera</i> Marsh.	4	2.11	2.21	2.18	0.05	19.09	24.43	21.50	2.48	34.44	39.84	36.74	2.37
<i>Fagus grandifolia</i> Ehrh.	4	1.85	2.39	2.09	0.25	26.26	30.04	28.20	1.70	43.99	53.15	47.24	4.08
<i>Fraxinus americana</i> L.	2	1.87	1.89	1.88	0.01	16.01	16.57	16.29	0.40	48.51	49.04	48.78	0.37
<i>Picea abies</i> (L.) Karst.	1	1.00	1.00	1.00		25.57	25.57	25.57		38.07	38.07	38.07	
<i>Pinus resinosa</i> Ait.	5	1.00	1.15	1.06	0.06	24.69	26.48	25.87	0.73	37.42	41.67	39.90	1.62
<i>Pinus strobus</i> L.	6	1.28	1.96	1.42	0.27	24.69	26.48	25.89	0.96	37.42	41.67	39.24	1.00
<i>Quercus rubra</i> L.	36	1.57	2.42	1.99	0.21	20.45	27.05	23.91	1.58	36.27	43.89	40.68	1.82
<i>Quercus velutina</i> Lam.	4	1.98	2.67	2.26	0.34	24.91	28.51	26.30	1.72	35.24	42.35	38.91	3.54
<i>Tsuga canadensis</i> (L.) Carrière	6	0.90	1.11	1.00	0.09	14.95	17.87	16.50	1.06	28.25	34.48	31.17	2.31

### Sample preparation

All fresh green leaf samples initially were air dried and then oven-dried at 70°C. Dried leaves were ground in a Wiley mill and passed through a 1-mm screen for optimal particle size (Windham et al. 1989). Prior to scanning, ground samples were redried overnight (approx. 15h) in a forced air convection oven at 70°C.

### Spectral data collection

Near infrared and visible diffuse reflectance spectral data were collected on a NIRSystems 6500 monochromator (NIRSystems Inc., Silver Spring, Md.) with a spinning cup module, which scanned at wavelengths from 400 to 2498 nm, at 2-nm intervals with a bandwidth of 10 nm. All reflectance values are relative to a ceramic standard. The software used both for data collection

**Table 3.** Laboratory reference value statistics for 1992 AVIRIS calibration and validation data sets.

Data set	n	% nitrogen				% lignin				% cellulose			
		Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
92AVIRC3	372	0.78	3.46	1.84	0.66	10.32	33.70	22.41	4.61	23.69	57.06	37.68	6.34
92AVIRV3	186	0.69	3.51	1.87	0.67	14.07	33.41	22.90	4.51	24.20	57.97	38.26	6.33
92AVIRIS	558	0.69	3.51	1.85	0.66	10.32	33.70	22.58	4.58	23.69	57.97	37.87	6.34
93HF	97 <sup>a</sup>	0.90	2.67	1.86	0.43	14.95	30.04	23.32	3.44	24.91	53.15	38.59	4.86
LTER	31	0.22	2.40	0.88	0.58	7.60	44.62	23.10	8.40	24.56	74.73	46.36	13.87
LTERFOL	14	0.22	2.40	0.90	0.61	7.60	41.40	25.49	8.73	26.72	49.20	38.04	5.29

<sup>a</sup>n = 94 for lignin and cellulose.

and for data analysis was NIRS2, version 3.00, from Infrasoftware International (Silver Spring, Md.). Because spectrophotometric signals are not linearly related to the chemical composition of a sample, a response linearization pretreatment for reflectance data to apparent absorbance,  $A = \log(1/R)$  (Hruschka 1987; Martens and Naes 1987; Naes and Martens 1984), is done prior to storing the spectral values.

Using the spinning cup module on the instrument, the sample is automatically rotated and scanned 32 times in 60 s. The reference ceramic is scanned 16 times before and 16 times after the sample scans to adjust for any possible drift that may have occurred during scanning. The resulting sample spectrum is the average of the 32 scans. To minimize any possible packing or subsampling error, two subsamples were scanned for each sample. All material from the first subsample was replaced in the jar, the sample was again thoroughly mixed, and another subsample was scanned. The sample spectrum used in the data analysis was the average of the two subsample spectra.

## Calibration

### Regression methods

Two types of regression analyses were compared: stepwise forward MLR, proposed by Norris et al. (1976) for this application; and PLS regression, developed by Wold (1982). Stepwise regression (referred to as step-up in the software) selects the first wavelength that gives the best single-term calibration; then selects the best second wavelength for the second variable for a two-term equation, and so on. The final equation consists of the statistically best wavelength combination. One problem with MLR is the probability of overfitting the data. Therefore, the general rule in stepwise regression is one term in the equation, and 10 samples in the calibration data set, for each regression constant and each parameter that varies, such as wavelength (Hruschka 1987). Four terms, or wavelengths, were used in this study.

PLS was run with a maximum of 15 factors, or principal components, set by the software based on the number of samples in the data set. Final equation statistics resulted from the average of four cross-validation groups, or internal subsets of calibration samples, which during the regression calculations are alternately excluded from the calibration to be used for internal prediction, and then are included in the regression calculations. In other words, using four cross-validation groups produces four iterations of calibrations and predictions so that each sample is predicted once during equation development. According to Martens and Naes (1987), the best internal validation criterion is cross validation, repeated until all the calibration samples have been treated as unknowns. The equations with the lowest SECV were used to predict against the validation data sets. Final calibration equations were developed using the entire

92AVIRIS data set. Robustness and generality of the calibration equations were then tested against two independent data sets: 93HF and LTER, including the LTERFOL subset.

### Regression statistics

To compare regression equations for the calibration data set, the coefficient of determination ( $R^2$  in MLR, or one minus the variance ratio ( $1 - VR$ ) in PLS) and the standard error of calibration statistic (SEC in MLR, or SECV in PLS, the standard error of cross validation) are evaluated independently for each constituent. To compare equation performance on external validation data sets, the standard error of prediction (SEP) for both MLR and PLS is evaluated for each constituent.

The PLS statistic for total explained variation,  $1 - VR$ , or the ratio of cross-validation error squared to the standard deviation of the data set squared, is a conservative estimate of  $R^2$ , where  $R^2$  equals one minus the ratio of SEC to the standard deviation. SEC, the standard deviation of the residuals, or differences between the actual laboratory, wet chemistry values and the spectrometrically predicted values for samples in the calibration data set, is calculated as the square root of the mean square for the residuals. SEC is an estimate of the best accuracy obtainable using the specific wavelengths of the calibration equation (Mark and Workman 1991).

SECV, however, is a true estimate of the prediction accuracy of the equation and is always larger than the calibration error (Infrasoft International 1992). SECV is the square root of the mean square of the residual for  $N - 1$  degrees of freedom. Based on an iterative calibration-validation algorithm, which selects sets of samples from the calibration set for calibration and for validation, SECV is used as the statistic for determining the best number of independent variables for the calibration equation. The equation with the lowest SECV is often selected as the best calibration (Mark and Workman 1991).

Usually larger than the standard error of calibration, SEP is calculated as the root mean square for residuals for  $N$  degrees of freedom, using the difference between laboratory and predicted values on independent validation data sets (Mark and Workman 1991). SEP is a true measure of the performance of the equation on unknown samples and is the preferred statistic to use for comparison of regression equations (Westerhaus 1989).

### Math treatments

Several math treatments (i.e., data transformations using finite-difference approximations to derivatives and smoothing) were tested to compare first and second difference transformations, sampling intervals (also called segment or gap) for the finite-difference calculation, and smoothing intervals. Derivative spectroscopy enhances small peaks or shoulders that may be masked by broad band features or background noise, thereby

**Table 4.** Calibration and validation statistics for partial least squares (PLS) regression and stepwise multiple linear regression (MLR) methods.

Data set	PLS regression <sup>a</sup>			Data set	Stepwise MLR <sup>a</sup>		
	(1,5,5,1)	(2,5,5,1)	(2,10,10,1)		(1,5,5,1)	(2,5,5,1)	(2,10,10,1)
<b>% nitrogen</b>							
92AVIRIS	0.106	0.105	0.107	92AVIRIS	0.157	0.127	0.129
1 - VR	0.974	0.975	0.974	R <sup>2</sup>	0.943	0.963	0.962
No. factors	13	13	13	No. terms	4	4	4
92AVIRV3	0.97, 0.11	0.97, 0.11	0.97, 0.11	92AVIRV3	0.85, 0.15	0.97, 0.12	0.97, 0.12
93HF	0.93, 0.14	0.93, 0.14	0.93, 0.14	93HF	0.87, 0.17	0.90, 0.15	0.91, 0.14
LTER	0.81, 0.29	0.82, 0.31	0.82, 0.31	LTER	0.75, 0.38	0.82, 0.32	0.78, 0.28
LTERFOL	0.86, 0.25	0.88, 0.26	0.84, 0.27	LTERFOL	0.72, 0.41	0.88, 0.35	0.78, 0.29
<b>% lignin</b>							
92AVIRIS	1.613	1.617	1.641	92AVIRIS	2.603	2.260	2.491
1 - VR	0.876	0.875	0.871	R <sup>2</sup>	0.677	0.756	0.704
No. factors	13	12	13	No. terms	4	4	4
92AVIRV3	0.88, 1.60	0.89, 1.53	0.87, 1.65	92AVIRV3	0.62, 2.77	0.75, 2.25	0.66, 2.64
93HF	0.78, 1.69	0.75, 1.74	0.74, 1.82	93HF	0.27, 3.11	0.62, 2.13	0.45, 2.59
LTER	0.56, 6.00	0.60, 5.80	0.68, 5.24	LTER	0.21, 8.26	0.36, 7.34	0.48, 6.09
LTERFOL	0.73, 5.16	0.69, 5.67	0.72, 5.29	LTERFOL	0.38, 8.29	0.44, 7.18	0.54, 6.11
<b>% cellulose</b>							
92AVIRIS	2.103	2.135	2.148	92AVIRIS	2.720	3.002	2.786
1 - VR	0.890	0.886	0.885	R <sup>2</sup>	0.816	0.776	0.807
No. factors	11	11	13	No. terms	4	4	4
92AVIRV3	0.86, 2.34	0.86, 2.35	0.86, 2.34	92AVIRV3	0.78, 2.95	0.73, 3.26	0.77, 3.02
93HF	0.87, 1.76	0.84, 1.98	0.83, 2.04	93HF	0.82, 2.07	0.69, 2.75	0.74, 2.56
LTER	0.85, 5.57	0.87, 7.51	0.85, 9.01	LTER	0.84, 5.96	0.61, 8.77	0.81, 7.51
LTERFOL	0.69, 3.94	0.69, 3.08	0.64, 3.21	LTERFOL	0.48, 3.85	0.34, 6.14	0.48, 4.96

**Note:** Goodness of fit statistics are  $R^2$  (MLR equations) and 1 - VR (variance ratio, PLS). For calibration data set, 92AVIRIS value is the standard error of calibration (SECV or SEC). Values for validation data sets are  $R^2$  and standard error of prediction (SEP).

<sup>a</sup>Numbers in parentheses are sets that represent the math treatment and are described in the text.

isolating weaker signals and separating overlapping bands, and also removes base-line shifts resulting from differences in particle size (Dixit and Ram 1985). A math treatment is described by four numbers in parentheses, which are, respectively, (i) the order of the difference; (ii) the interval, or number of data points (at 2 nm each) over which an average is taken; the number of data points used for (iii) a first running-average smooth and (iv) a second running-average smooth (a value of 1 indicates no smooth). In this software smoothing was done after the difference calculation. There is no single math treatment or calibration procedure that gives the best results for all constituents and for all products (Shenk and Westerhaus 1993). These parameters may be found experimentally through regression results.

#### Data set selection

To develop the initial calibration equations, two-thirds of the 92AVIRIS samples were used for calibration, while one-third of the samples were reserved for validation. To create an even distribution of constituent concentrations for both the calibration and validation data sets, two methods were used. Previous studies in this laboratory (McLellan et al. 1991) and others (Abrams 1989; Windham et al. 1989) have shown that random selection of calibration samples gave the best results. However, an alternative method available using the Center program in the NIRS2 software package (Infrasoft International 1992) is

sample selection based on the spectral variability within the data set (Abrams 1989; Windham et al. 1989; Shenk and Westerhaus 1991a, 1991b). The Mahalanobis distances ( $H$ ), or  $R^2$  distances, from a sample to the average spectrum in the file were computed. The spectra in the data set were then reordered according to ascending  $H$  value, to assure an even spectral distribution for a split between calibration-validation samples within a data set. With both random and centering methods, every third sample was removed from the total data set for validation, while the remaining two-thirds were used to develop the initial calibration equations. The centered files were produced using the same math treatments that were used for regression analyses.

Math treatments of (1, 5, 5, 1), (2, 5, 5, 1), and (2, 10, 10, 1) were chosen both for the preliminary trials of calibration data set splits into calibration and validation files based on spectral centering, and for the regression analyses. A math treatment of (2, 10, 10, 1), second difference, calculated over ten 2-nm data points, and smoothed once over the same range, most closely corresponds to that used in previous studies in our laboratory (McLellan et al. 1991). Results of the random calibration-validation data-set split versus the centered splits, using all four math treatments, showed that equations with the best SECV and SEP were developed using the randomly split file.

## Results and discussion

### Calibration statistics

Overall, comparisons between MLR and PLS regression showed consistently lower SEC and higher  $R^2$  values for the PLS equations (Table 4). According to SEC values, the best overall math treatment using PLS was (1, 5, 5, 1), although differences between treatments were small. Second derivative calibration equations work best for MLR (cf. Marten et al. 1983; Shenk and Westerhaus 1991b). Comparing equations with the (1, 5, 5, 1) math treatment, PLS showed an improvement in SECV over the SEC for MLR, from 0.157 to 0.106% for nitrogen, from 2.603 to 1.613% for lignin, and from 2.720 to 2.103% for cellulose. Likewise, the PLS (1, 5, 5, 1) equations showed an improvement in  $1 - VR$  over  $R^2$  for MLR, from 0.943 to 0.974 for nitrogen, from 0.667 to 0.876 for lignin, and from 0.816 to 0.890 for cellulose. Differences between regression statistics by constituent are due in part to differences in the precision and accuracy of the laboratory reference methods used. Equations for those constituents for which laboratory methods have higher precision and accuracy are expected to explain a greater proportion of variation in the data set.

### Validation statistics

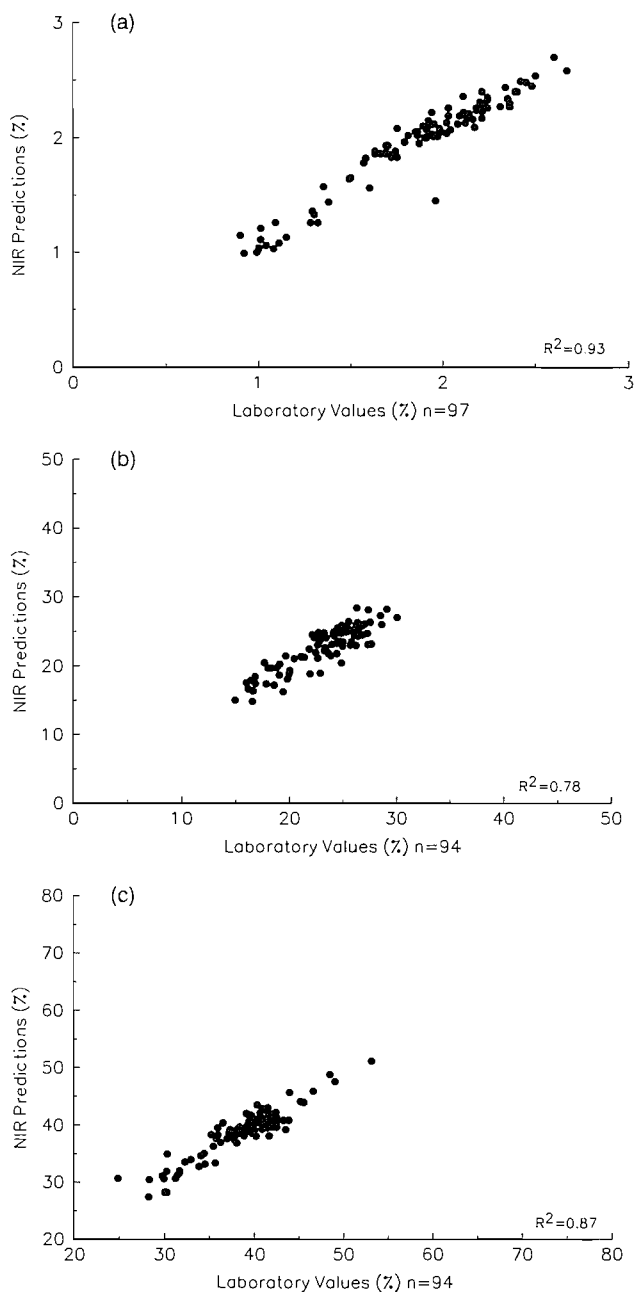
There was a close relationship between those math treatments that calibrated well and those that validated well against both the one-third of the 92AVIRIS data set held back from the calibration (92AVIRV3) and the independent data set collected at the Harvard Forest in 1993 (93HF, Table 4). Again, the first difference treatment was best for PLS, and the second difference treatments were generally best for MLR. As expected, SEP values were slightly above SEC values for each treatment, with the increases in SEP over SEC similar to those seen in previous studies (e.g., McLellan et al. 1991). All values approximate the analytical precision of wet chemical analyses (McLellan et al. 1991; Newman et al. 1994). This substantiates the applicability of the NIRS method for the determination of the concentration of nitrogen, lignin, and cellulose in foliar samples that fall within the same species groups and concentration ranges as the calibration data set (Fig. 1).

However, SEP values are 2–4 times higher for the LTER and LTERFOL validation data sets (Table 4). While validation was better for nitrogen than for lignin or cellulose, and was better for the more restricted LTERFOL data set (especially for cellulose for which the validation data set fell within the range of the calibration data set, Table 3), all values were substantially higher than the analytical precision of wet chemistry and combustion techniques (Fig. 2 and 3). In general, PLS produced lower SEP values for all constituents, data sets, and math treatments.

### Wavelength selection

The arbitrary and inconsistent nature of wavelength selection has been a constant criticism leveled against the MLR technique. PLS avoids this limitation by using information in all wavelengths and assigning a coefficient to each. The wavelengths with the highest coefficients contain the most information on the concentration of that constituent and

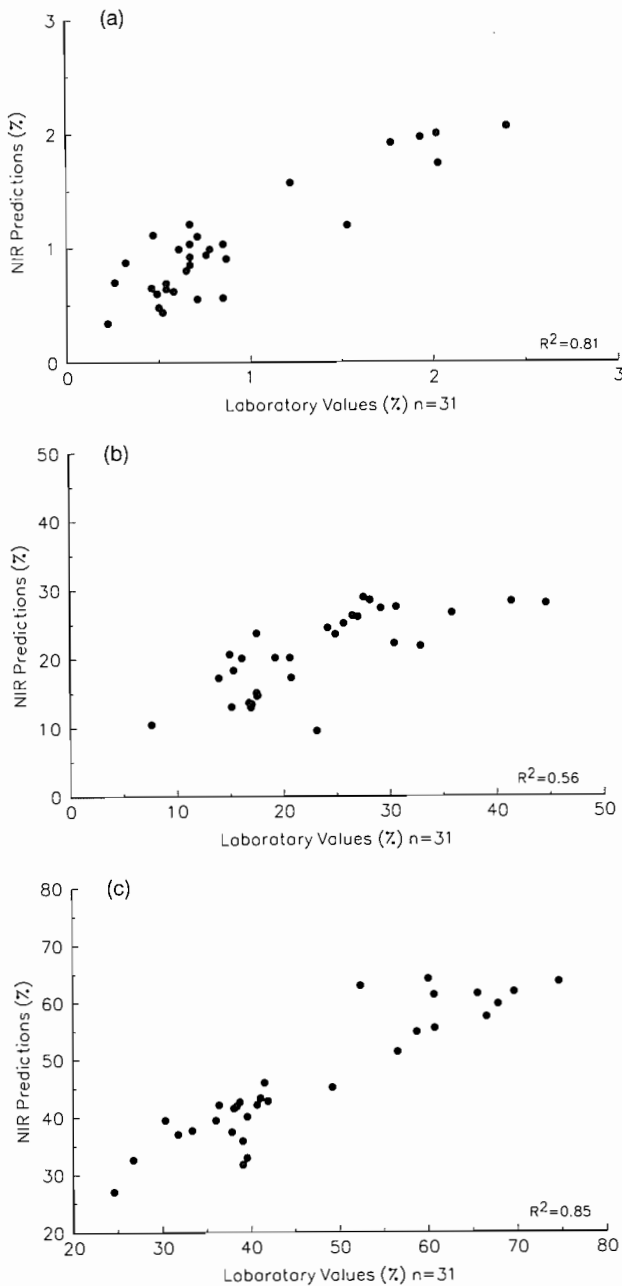
**Fig. 1.** Measured chemical constituent concentrations for the Harvard Forest 1993 validation data set compared with values predicted using the calibration equation derived from the 92AVIRIS data set by PLS using the (1, 5, 5, 1) math treatment: (a) nitrogen, (b) lignin, and (c) cellulose.



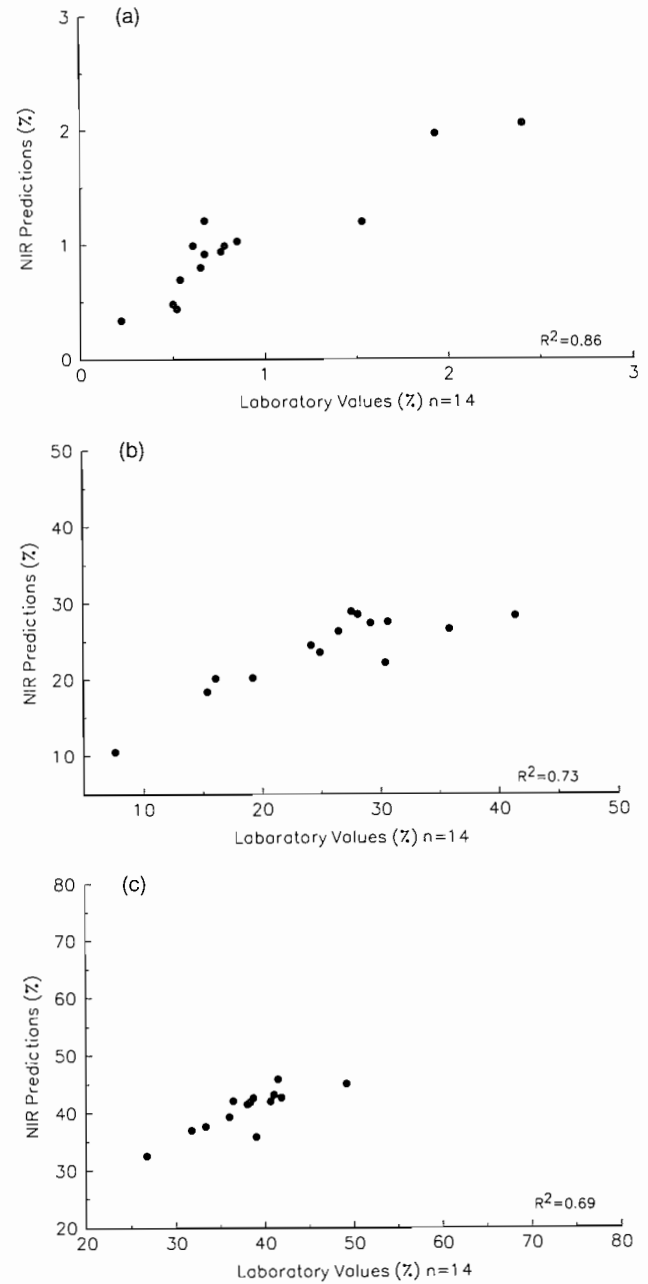
should relate to known absorption features of each constituent. Comparing results of statistical wavelength selection by MLR and maximum coefficients in PLS with published spectral relationships to chemical bond types allows an analysis of the consistency of the two techniques and the accuracy of each in identifying wavelengths with high information content.

With a spectral resolution of the instrument at 10 nm, and with a 5- or 10-data point (i.e., a 10- or 20-nm) interval used in the difference calculation, peaks may occur within

**Fig. 2.** Measured chemical constituent concentrations for the LTER validation data set compared with values predicted using the calibration equation derived from the 92AVIRIS data set by PLS using the (1, 5, 5, 1) math treatment: (a) nitrogen, (b) lignin, and (c) cellulose.



**Fig. 3.** Measured chemical constituent concentrations for the LTERFOL validation data set compared with values predicted using the calibration equation derived from the 92AVIRIS data set by PLS using the (1, 5, 5, 1) math treatment: (a) nitrogen, (b) lignin, and (c) cellulose.



a 10-nm range for first difference equations, and within a 30- or 60-nm range for second difference calculations. Data smoothing with a 5- or 10-data point interval further offsets the location of constituent absorption peaks. Thus, we consider wavelengths with maximum coefficient values occurring within 10-nm of a reported structure as representative of the structure (Table 5; chemical structures are according to Shenk et al. 1992; Osborne and Fearn 1986; Murray and Williams 1987; several wavelengths occurring

at greater than 10 nm but less than 20 nm from a given structure are indicated).

For the 92AVIRIS data set, the best statistical combination of four wavelengths (MLR) and the four wavelengths with highest regression coefficients (PLS) for predicting nitrogen concentration are generally similar between regression techniques and across math treatments (Table 5). They are also generally associated with known absorption features of nitrogen-containing compounds (see similar

**Table 5.** 1992 AVIRIS regression equations: wavelength (nm) and corresponding chemical structure.

	(1,5,5,1)		(2,5,5,1)		(2,10,10,1)	
<b>PLS regression<sup>a</sup></b>						
Nitrogen	1972	CONH <sub>2</sub>	1984	CONH <sub>2</sub>	1984	CONH <sub>2</sub>
	2044	CONH/H <sub>2</sub>	2056	Protein	2056	Protein
	2068	Urea	2076	Urea	1956	Starch
	2476	Protein	2344	Protein	2080	Urea
Lignin	1676	Aromatic	1688	Aromatic	1688	Aromatic
	1428	ArOH	2384	Oil	2384	Oil
	1716	·CH <sub>2</sub>	2324	Starch	1748	-SH
	2248	Amino acid	1748	-SH	1396	·CH <sub>2</sub>
Cellulose	2468	·CH <sub>2</sub>	2288	Starch	2368	Cellulose* <sup>†</sup>
	2224	Amino acid <sup>†</sup>	2368	Cellulose* <sup>†</sup>	2452	Starch*
	2160	·HC=CH	1696	·CH <sub>3</sub>	2396	Oil <sup>†</sup>
	2384	Oil	2152	CONH <sub>2</sub> *	1804	Cellulose <sup>†</sup>
<b>Stepwise MLR<sup>b</sup></b>						
Nitrogen	2140	·NC=CH	1556	CONH <sup>†</sup>	2168	·HC=CH
	1928	CONH	1980	CONH <sub>2</sub>	1980	CONH <sub>2</sub>
	2156	CONH <sub>2</sub> *	2044	CONH/H <sub>2</sub>	1192	·CH <sub>3</sub>
	2044	CONH/H <sub>2</sub>	2076	Urea	2056	Protein
Lignin	2392	Oil <sup>†</sup>	2384	Oil	1288	O—H Stretch**
	1908	P—OH	2452	Starch* <sup>†</sup>	2384	Oil
	2444	Starch* <sup>†</sup>	1436	·CH <sub>2</sub>	1484	·CONH <sub>2</sub>
	1868	C—Cl*	2068	Oil	1928	Starch
Cellulose	2084	·OH	2144	·NC=CH	2144	·NC=CH
	2472	·CH <sub>2</sub>	2260	Cellulose	2256	Cellulose <sup>†</sup>
	1528	RNH <sub>2</sub>	1420	ArOH	2364	Cellulose* <sup>†</sup>
	1744	-SH	2368	Cellulose* <sup>†</sup>	1560	CONH

**Note:** Numbers in parentheses are sets that represent the math treatment and are described in the text. Chemical structures are within 10 nm of specified wavelengths, according to Shenk et al. (1992), except when marked with one asterisk, according to Osborne and Fearn (1986), or with two asterisks, according to Murray and Williams (1987). Structures indicated by a dagger are 10- to 20-nm approximations to specified wavelengths.

<sup>a</sup>Data shown are the four wavelengths with the highest regression coefficient values.

<sup>b</sup>Data shown are the wavelengths used for the four terms of the equation.

results by Johnson and Billow 1996; Wessman et al. 1988; Peterson et al. 1988). Wavelengths with the highest regression coefficient values for nitrogen were consistent in PLS first and second difference equations. The highest value occurred near a N—H combination band feature at 1980 nm (Shenk et al. 1992; Wessman 1990). High coefficients also occurred near 2060 nm, another N—H combination band feature, and within the range of nitrogen-containing compounds between 2040 and 2080 nm (Shenk et al. 1992; Wessman 1990). Figure 4a is a nitrogen correlogram of the 92AVIRIS data set, and Fig. 5a is a plot of the PLS regression coefficients for nitrogen in the first difference equation. For this application, a correlogram is a plot of the correlation coefficient (*r*) for the chemical concentration of a constituent for all samples at each given wavelength. The correlograms are calculated for first difference absorbance data for comparison with the full-spectrum plot of regression coefficients in the PLS first

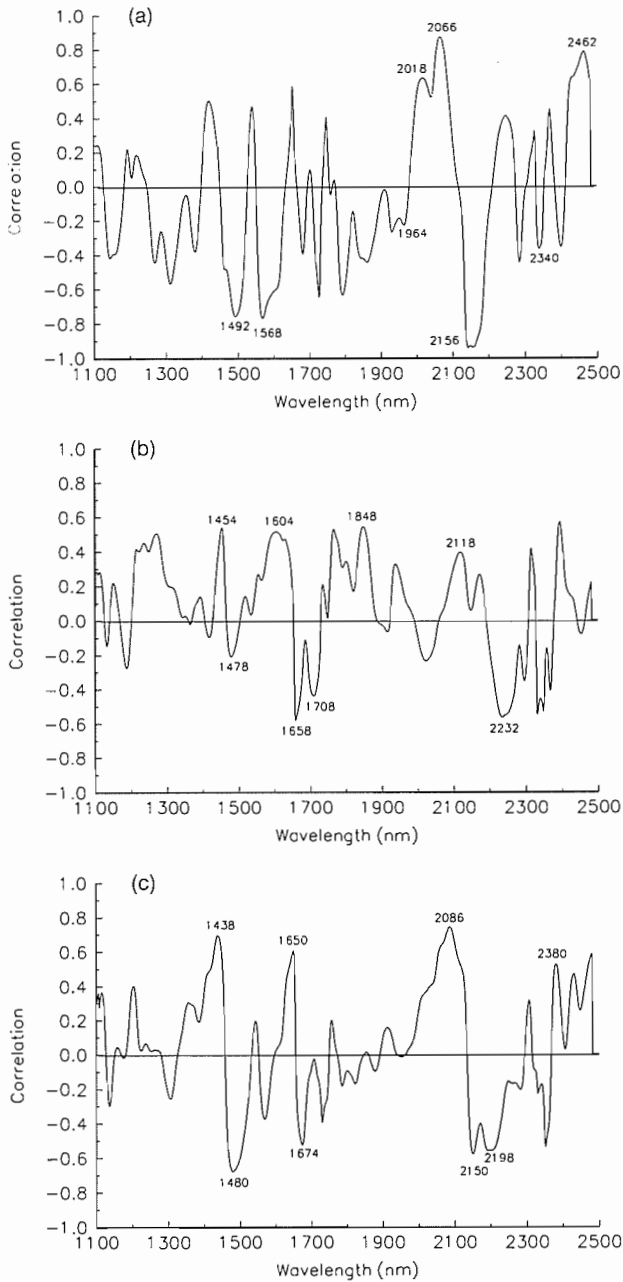
difference equation. First difference plots of correlation coefficients and PLS regression coefficients show similar prominent spectral features.

Lignin, a compound that is not well defined, has characteristic aromatic C—H overtone absorption features near 1685 and 1143 nm (Shenk et al. 1992; Wessman 1990). The most highly correlated wavelength for all PLS equations occurred close to the first overtone feature near 1685 nm. Overall, there is less wavelength consistency between equations for lignin than for nitrogen (Table 5). The first difference simple correlation plot (Fig. 4b) and the first difference PLS regression coefficient plot (Fig. 5b) for lignin show the most highly correlated wavelengths for both correlation and regression equation.

Wavelengths associated with cellulose include structurally similar chemical compounds such as starch (Table 5). The most highly correlated feature reported previously for this constituent occurred at 2140 nm (McLellan et al. 1991).



**Fig. 4.** Correlation coefficients from simple linear regressions of individual wavelengths against constituent concentrations for the 92AVIRIS data set: (a) nitrogen, (b) lignin, and (c) cellulose.

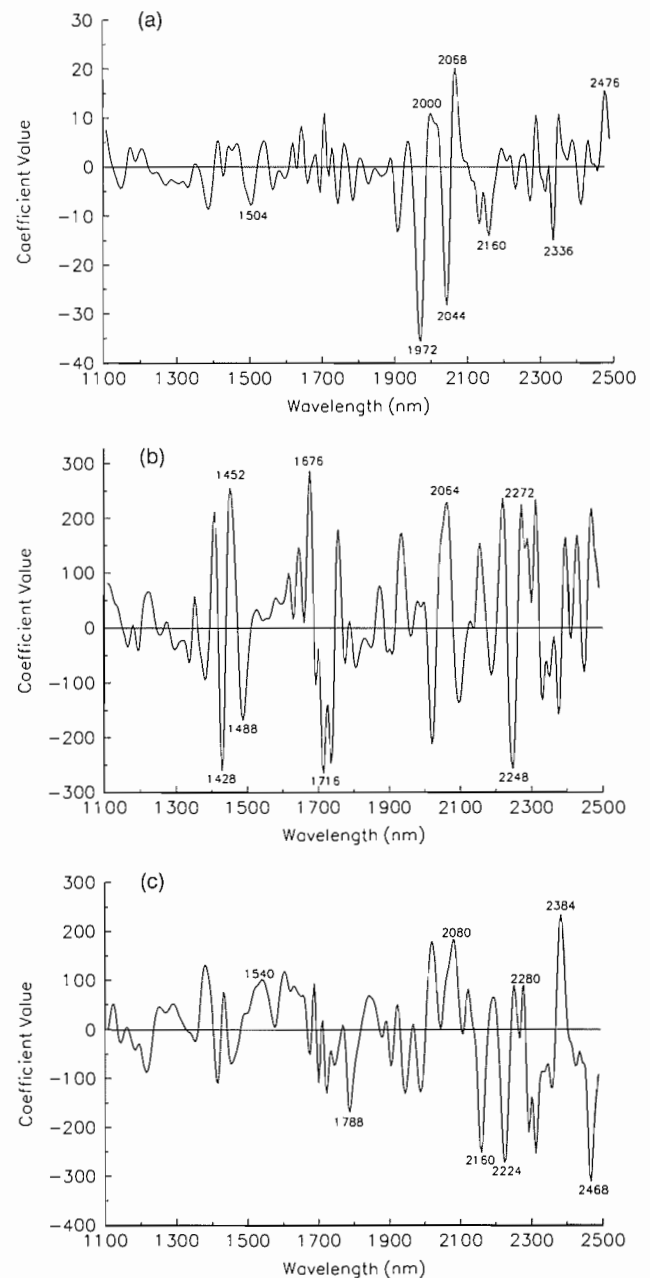


A strong absorption feature between approximately 2000 and 2100 nm seen in the correlation plot (Fig. 4c) represents a strong O-H combination band that occurs at 2100 nm (Wessman 1990). Two prominent wavelengths within this region are also seen on the regression coefficient plot for cellulose (Fig. 5c).

## Conclusions

Results show that the full-spectrum calibration method, PLS regression, performed better than stepwise MLR on dry, ground, green foliage samples of native woody plants.

**Fig. 5.** Regression coefficients derived for all wavelengths by PLS for the 92AVIRIS data set. Higher coefficients signify wavelengths with more information content for the constituent plotted: (a) nitrogen, (b) lignin, and (c) cellulose.



Problems due to multicollinearity between wavelengths and the reduction of relevant information into a few available wavelengths selected as MLR regression factors are overcome in the PLS method. Comparisons between correlograms and PLS regression coefficients showed similarities in prominent spectral features by constituent. Using all available wavelength information, and concentrating into the first factors the near infrared information with relevance to the chemical constituents, PLS regression increased prediction accuracy on independent data sets.

Validation of the equations derived from foliage of temperate woody species was acceptable when the validation data set contained samples from similar sites and species, and concentrations of constituents predicted fell within the range of those in the calibration data set. However, SEP on samples from different types of tissues or with measured concentrations of nitrogen, lignin, and cellulose beyond the range of variation in the calibration data set were 2–4 times higher than analytical precision of wet chemical and combustion techniques. The development of fully generalized NIRS calibration equations will require including in the calibration data set additional tissue types as well as ecologically diverse species with broad ranges of constituent concentrations.

## Acknowledgments

This research was supported by a grant from the National Aeronautics and Space Administration, HIRIS Science Team (NAS-5-31379), as part of the Accelerated Canopy Chemistry Program (NAGW 3194). We gratefully acknowledge the Long-Term Ecological Research Network for their samples from the Long-Term Intersite Decomposition Experiment Team Project (BSR-9108329) supported by the National Science Foundation. We thank Stephen Newman, Gloria Quigley, and Margaret Soulia for wet chemistry analyses, and Dr. Janet Campbell for helpful discussion of statistical analyses.

## References

- Aber, J.D., Bolster, K.L., Newman, S., Soulia, M., and Martin, M.E. 1994. Analyses of forest foliage II: Measurement of carbon fraction and nitrogen content by end-member analysis. *J. Near Infrared Spectrosc.* 2: 15–23.
- Abrams, S.M. 1989. Populations. *In* Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. *Edited by* G.C. Marten, J.S. Shenk, and F.E. Barton. U.S. Dep. Agric. Agric. Handb. 643. p. 37.
- Card, D.H., Peterson, D.L., Matson, P.A., and Aber, J.D. 1988. Prediction of leaf chemistry by the use of visible and near infrared reflectance spectroscopy. *Remote Sens. Environ.* 26: 123–147.
- Dixit, L., and Ram, S. 1985. Quantitative analysis by derivative electronic spectroscopy. *Appl. Spectrosc. Rev.* 21: 311–418.
- Hruschka, W.R. 1987. Data analysis: wavelength selection methods. *In* Near-infrared technology in the agricultural and food industries. *Edited by* P.C. Williams and K.H. Norris. American Association of Cereal Chemists, Inc., St. Paul, Minn. pp. 35–55.
- Infrasoft International. 1992. NIRS2, Version 3.00. Routine operation and calibration software for near infrared instruments. Perstorp Analytical, Inc., Silver Spring, Md.
- Johnson, L.F., and Billow, C.R. 1996. Spectrometric estimation of total nitrogen concentration in Douglas-fir foliage. *Int. J. Remote Sensing.* 17: 489–500.
- Lindberg, W., Persson, J.A., and Wold, S. 1983. Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. *Anal. Chem.* 55: 643–648.
- Lorber, A., Wangen, L.E., and Kowalski, B.R. 1987. A theoretical foundation for the PLS algorithm. *J. Chemometrics.* 1: 19–31.
- Mark, H., and Workman, J. 1991. *Statistics in spectroscopy.* Academic Press, Inc., New York.
- Marten, G.C., Halgerson, J.L., and Cherney, J.H. 1983. Quality prediction of small grain forages by near infrared reflectance spectroscopy. *Crop Sci.* 23: 94–96.
- Martens, H., and Jensen, S.A. 1983. Partial least squares regression: a new two-stage NIR calibration method. *In* Progress in cereal chemistry and technology. *Edited by* J. Holas and J. Kratochvil. Proceedings of the 7th World Cereal and Bread Congress, 28 June – 2 July 1982, Prague. Elsevier, New York. pp. 607–647.
- Martens, H., and Naes, T. 1987. Multivariate calibration by data compression. *In* Near-infrared technology in the agricultural and food industries. *Edited by* P.C. Williams and K.H. Norris. American Association of Cereal Chemists, Inc., St. Paul, Minn. pp. 57–87.
- McLellan, T.M., Martin, M.E., and Aber, J.D. 1991. Comparison of wet chemistry and near infrared reflectance measurements of carbon-fraction chemistry and nitrogen concentration of forest foliage. *Can. J. For. Res.* 21: 1689–1693.
- Murray, I., and Williams, P.C. 1987. Chemical principles of near-infrared technology. *In* Near-infrared technology in the agricultural and food industries. *Edited by* P.C. Williams and K.H. Norris. American Association of Cereal Chemists, Inc., St. Paul, Minn. pp. 17–34.
- Naes, T., and Martens, H. 1984. Multivariate calibration. II. Chemometric methods. *Trends Anal. Chem.* 3: 266–271.
- Naes, T., and Martens, H. 1985. Comparison of prediction methods for multicollinear data. *Commun. Stat. Simul. Comput.* 14: 545–576.
- Newman, S.D., Soulia, M.E., Aber, J.D., Dewey, B., and Ricca, A. 1994. Analyses of forest foliage I: Laboratory procedures for proximate carbon fractionation and nitrogen determination. *J. Near Infrared Spectrosc.* 2: 5–14.
- Norris, K.H., Barnes, R.F., Moore, J.E., and Shenk, J.S. 1976. Predicting forage quality by infrared reflectance spectroscopy. *J. Anim. Sci.* 43: 889–897.
- Osborne, B.G., and Fearn, T. 1986. *Near infrared spectroscopy in food analysis.* John Wiley & Sons Inc., New York.
- Otto, M., and Wegscheider, W. 1985. Spectrophotometric multi-component analysis applied to trace metal determinations. *Anal. Chem.* 57: 63–69.
- Peterson, D.L., Aber, J.D., Matson, P.M., Card, D.H., Swanberg, N., Wessman, C., and Spanner, M. 1988. Remote sensing of forest canopy and leaf biochemical contents. *Remote Sens. Environ.* 24: 85–108.
- Shenk, J.S. 1989. Equation selection. *In* Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. *Edited by* G.C. Marten, J.S. Shenk, and F.E. Barton. U.S. Dep. Agric. Agric. Handb. 643. pp. 26–27.
- Shenk, J.S., and Westerhaus, M.O. 1991a. Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. *Crop Sci.* 31: 469–474.
- Shenk, J.S., and Westerhaus, M.O. 1991b. Population structuring of near infrared spectra and modified partial least squares regression. *Crop Sci.* 31: 1548–1555.
- Shenk, J.S., and Westerhaus, M.O. 1993. Analysis of agriculture and food products by near infrared reflectance spectroscopy. *Infrasoft International, Port Matilda, Pa.*
- Shenk, J.S., Landa, I., Hoover, M.R., and Westerhaus, M.O. 1981. Description and evaluation of near infrared reflectance spectro-computer for forage and grain analysis. *Crop Sci.* 21: 355–358.
- Shenk, J.S., Workman, J.J., Jr. and Westerhaus, M.O. 1992. Application of NIR spectroscopy to agricultural products. *In* Handbook of near-infrared analysis. *Edited by* D.A. Burns and E.W. Ciurczak. Marcel Dekker, Inc., New York. pp. 383–431.
- Sjöstrom, M., Wold, S., Lindberg, W., Persson, J.A., and

- Martens, H. 1983. A multivariate calibration problem in analytical chemistry solved by partial least squares models in latent variables. *Anal. Chim. Acta*, **150**: 61–70.
- Vane, G., Green, R.O., Chrien, T.G., Enmark, H.T., Hansen, E.G., and Porter, W.M. 1993. The Airborne Visible / Infrared Imaging spectrometer (AVIRIS). *Remote Sens. Environ.* **44**: 127–143.
- Wessman, C.A. 1990. Evaluation of canopy biochemistry. *In Remote sensing of biosphere functioning. Edited by R.J. Hobbs and H.A. Mooney. Springer-Verlag, New York.* pp. 135–156.
- Wessman, C.A., Aber, J.D., Peterson, D.L. and Melillo, J.M. 1988. Foliar analysis using near infrared reflectance spectroscopy. *Can. J. For. Res.* **18**: 6–11.
- Westerhaus, M.O. 1989. Equation development. *In Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. Edited by G.C. Marten, J.S. Shenk, and F.E. Barton. U.S. Dep. Agric. Agric. Handb. 643.* pp. 38–39.
- Wetzel, D.L. 1983. Near-infrared reflectance analysis: sleeper among spectroscopic techniques. *Anal. Chem.* **55**: 1165A–1171A.
- Winch, J.E., and Major, H. 1981. Predicting nitrogen and digestibility of forages using near infrared reflectance photometry. *Can. J. Plant Sci.* **61**: 45–51.
- Windham, W.R., Mertens, D.R., and Barton, F.E., II. 1989. Protocol for NIRS calibration: sample selection and equation development and validation. *In Near infrared reflectance spectroscopy (NIRS): analysis of forage quality. Edited by G.C. Marten, J.S. Shenk, and F.E. Barton. U.S. Dep. Agric. Agric. Handb. 643.* pp. 96–103.
- Wold, H. 1982. Soft modeling, the basic design and some extensions. *In Systems under indirect observation. Part II. Edited by K.G. Joreskog and H. Wold. North Holland Publishing Co., Amsterdam.* pp. 1–54.
- Wold, S., Ruhe, A., Wold, H., and Dunn, W.J., III. 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Siam J. Sci. Stat. Comput.* **5**: 735–743.