Mathematical Treatment of Near-Infrared Reflectance Data for the Estimation of Protein[1]

F. S. LAI, Y. POMERANZ, D. TRAYLOR, and S. AFEWORK, U.S. Grain Marketing Research Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Manhattan, KS 66502

Cereal Chem. 61(4): 327-329

ABSTRACT

Protein was determined by Kjeldahl and near-infrared reflectance methods on corn from two crop years, and in two- and six-row barley and malt from numerous locations from one crop year. Kjeldahl protein values and reflectance values obtained at six wavelengths on calibration samples were used to prepare linear regression lines (calibrations) by including all significant terms from multiple regression equations (up to six values) with and without interaction terms up to the sixth degree. A slight but consistent improvement was seen in correlation coefficients, prediction ranges, and standard errors of estimate when interaction terms were included. The improvement was valid for predicting protein content in corn samples from the same or from different crop years, in barley or malt of two- or six-row types, and for predicting protein in malt from protein in barley.

Optimization of the treatment of near-infrared reflectance (NIR) spectroscopy data has been the subject of many publications. The two main approaches have focused either on selecting specific, sharp, and clearly defined bands (Norris 1978, Williams and Panford 1979, Shenk et al 1981) or on treating the data to compensate for or eliminate abberrant interferences from moisture, particle size, or other factors (Norris and Williams 1977).

The first approach depends on the availability of an expensive scanning instrument that allows for the possibility of selecting the best wavelengths. In addition, the process of selection may be time-consuming and complicated. Consequently, most studies have been directed toward treatment-transformation of data, which has been successful only to a limited extent. One reason is that bands in the near-infrared range are basically quite broad and result from the contributions of many components that are all transformed in varying degrees by the treatment of data at a selected wavelength. Calculating composition on the basis of reflectance data taken at several (up to six) wavelengths and including those data in a multiple linear correlation is a compromise, at best (Hymowitz et al 1974). Actually, the prediction power may be affected adversely by including too many measurements in the multiple linear regression equation.

The second approach—to compensate for interfering factors—was undertaken in this study. Data from six NIR readings for predicting protein content were obtained on a Technicon Infralyzer (model 2.5A) and treated in two ways: by solving multiple linear regression equations (up to six terms); and by including in a regression equation interaction terms of the log values. In both methods, only statistically significant (at the 5% level) terms were included. Selection of significant terms for inclusion was made by a computer program, several of which are available. We used the Stepwise procedure (SAS 1979) because it provided insight into the relationship between the independent and the dependent variables.

MATERIALS AND METHODS

Samples of corn, barley, and barley malt were included in the comparisons. We obtained 370 corn samples from the 1979 crop and 388 samples from the 1980 crop from the corn-breeding program of P. J. Loesch, Jr., Ames, IA. The 1979 corn samples covered a much wider range of protein (8.2–15.2%) than the 1980 samples (9.0–12.0%). We also obtained from the Barley and Malt Laboratory, Madison, WI, 198 barley and 194 malt samples of the 1979 crop grown in the Mississippi Valley Uniform Nursery, Central and Eastern Stations, and Rocky Mountain and Western Stations. No distinction was made between two- and six-row barleys. Malts were prepared from the barley on an experimental scale at the USDA Barley and Malt Laboratory, as described by Dickson et al (1968). Alternating samples were used for calibration. Every other sample served as an independent sample for prediction, using the reflectance values obtained at six wavelengths by an Infralyzer. All data were expressed on an as-is moisture basis.

Multiple Linear Regression Analysis without Interaction

Protein values obtained from the Kjeldahl method were utilized in the following equation: protein percentage = $K_0 + K_1 \log(1) + K_2 \log(2) + \ldots + K_6 \log(6)$. Here log (1), log (2),..., log (6) are the reflected energy levels from each of the six different wavelength bands found on the Grain Analyzer. $K_1, K_2, \ldots, K_6$ represent the calculated multiple regression coefficients, and K_0 represents the intercept. After $K_0, K_1, \ldots, K_6$ values were obtained, the equation was used to predict the protein content in the independent samples. The SAS program developed an optimized linear equation by first finding the one-variable model that produced the highest R^2, and then, for each of the other independent variables, calculating the F-statistics reflecting that variable's contribution to the model if it were to be included. Next, only those variables that produced an F-statistic significant at the 5% level were added to the model one by one. After a variable was added, all the variables in the model were examined, and any variable that did not produce an F-statistic significant at 5% level was deleted. Only after this check was made and the necessary deletions accomplished was another variable added to the model. The process ended when no additional variable had an F-statistic significant at the 5% level, or when the variable to be added to the model was just deleted from it.

Multiple Regression Analysis with Interaction

In addition to the linear terms without interaction, ie, log (i), i = 1,2,..., 6, terms of the second, third, fourth, fifth, and sixth order were included. The second order interaction term is defined as (log 1) × (log 2), (log 1) × (log 3), and all other possible two-log combinations. The third, fourth, and higher orders of interaction terms are similarly defined:

Second order: $(\log i) \times (\log j)$ $i \leqslant j$, i, j = 1, 2,..., 6
Third order: $(\log i) \times (\log j) \times (\log k)$ $i \leqslant j \leqslant k$, i,j,k = 1,2,...,6
Fourth order: $(\log i) \times (\log j) \times (\log k) \times (\log l)$ $i \leqslant j \leqslant k \leqslant l$, i,j,k,l = 1,2,..., 6
Fifth order: $(\log i) \times (\log j) \times (\log k) \times (\log l) \times (\log m)$ $i \leqslant j \leqslant k \leqslant l \leqslant m$, i,j,k,l,m = 1,2,..., 6
Sixth order: $(\log i) \times (\log j) \times (\log k) \times (\log l) \times (\log m) \times (\log n)$ $i \leqslant j \leqslant k \leqslant l \leqslant m \leqslant n$, i,j,k,l,m,n = 1,2,..., 6.

[1] Mention of firm names or trade products does not imply that they are endorsed or recommended by the U.S. Department of Agriculture over other firms or similar products not mentioned.

©1984 American Association of Cereal Chemists, Inc.

The maximum number of the possible interaction terms for the second, third, fourth, fifth, and sixth order were 21, 26, 21, 12, and 7, respectively. The theoretical number of terms for inclusion in the multiple regression equation was six without interaction terms and 93 with interaction terms up to the sixth degree.

RESULTS AND DISCUSSION

A statistical evaluation of the calibration samples is shown in Table I. Except for the 1979 barleys, which showed no improvement, the inclusion of the interaction terms consistently increased the correlation coefficients and substantially decreased the standard error of estimate. In this computation, optimized and nonoptimized (all six terms included) multiple regressions without interaction were identical.

Table II lists the number of calibration and prediction samples, the number of terms used to calculate multiple correlation coefficients, and the resultant coefficients. In computation design I, we used half of the 1979 corn samples to develop two equations, with and without interactions. Those equations were then used to predict the protein content of the other half of the 1979 corn samples. Similar treatment was applied to 1980 corn, as shown in computation design II. The optimization required 10 terms for the 1979 corn and 39 terms for the 1980 corn. However, for the combined 1979 and 1980 corn, only nine terms were required. This reduction in number of terms may be due to the wide range of protein content values in the combined samples from those two years. Note that the prediction for 1980 corn required 39 terms when based on the calibration of 1980 corn samples and only five terms when based on the 1979 corn samples. The resultant correlation coefficients in both cases were practically identical.

In computation design III, we used half of the 1979 and 1980 corn samples to develop two optimized calibration equations. The equations were then used to predict the protein content in the other

TABLE I
Linear Correlation Coefficients (r) and Standard Errors of Estimates of Calibration Samples for Kjeldahl versus Near-Infrared Reflectance Protein

Grain	Year	No. of Samples	Optimized Multiple Regression		Optimized Multiple Regression Including Interaction	
			r	s	r	s
Corn	1979	185	0.921	0.50	0.944	0.43
Corn	1980	194	0.841	0.31	0.856	0.30
Corn	1979,1980	379	0.917	0.45	0.935	0.40
Barley	1979	99	0.966	0.49	0.966	0.49
Malt	1979	97	0.974	0.45	0.983	0.37
Barley, Malt	1979	196	0.949	0.60	0.971	0.46

TABLE II
Correlations Coefficients for Optimized Linear Regressions between Kjeldahl and Near-Infrared Reflectance Predicated Protein

Computation Design	Grain	Calibration Samples		Prediction Samples		Optimized Multiple Regression		Optimized Multiple Regression Including Interaction	
		Year	No. of Samples	Year	No. of Samples	r	No. of Terms[a]	r	No. of Terms[b]
I	Corn	1979	185	1979	185	0.923	6	0.892	10
II	Corn	1980	194	1980	194	0.834	5	0.861	39
III	Corn	1979–1980	379	1979–1980	379	0.915	6	0.941	9
IV	Corn	1979	185	1980	388	0.770	6	0.862	5
V	Corn	1980	194	1979	370	0.852	5	0.903	33
VI	Corn	1979–1980	379	1979	185	0.901	6	0.940	9
VII	Corn	1979–1980	379	1980	194	0.830	6	0.847	9
VIII	Barley	1979	99	1979	98	0.970	6	0.969	31
IX	Malt	1979	97	1979	97	0.971	4	0.976	7
X	Barley,Malt	1979	196	1979	195	0.953	5	0.972	33

[a]Used to predict protein content, out of a total number of six terms.
[b]Used to predict protein content, out of a total number of 93 terms.

TABLE III
Ranges of Kjeldahl Protein of the Calibration and Prediction Samples, and the Near-Infrared Reflectance (NIR) Predicted Protein

Computation Design	Kjeldahl Protein (%)		NIR Prediction Protein (%)	
	Calibration Samples	Prediction Samples	Optimized Multiple Linear Regression	Optimized Multiple Regression Including Interaction
I Corn 79	8.17–15.22	7.19–14.71	8.23–14.85	6.86–15.33
II Corn 80	8.97–11.98	8.81–11.59	9.02–11.24	9.04–11.34
III Corn 79-80→ corn 79-80	8.17–15.22	7.19–14.71	8.44–14.57	8.30–14.81
IV Corn 79→ corn 80	8.17–15.22	8.81–11.98	8.87–18.43	9.03–11.68
V Corn 80→ corn 79	8.97–11.98	7.19–15.22	9.35–14.34	5.34–15.2
VI Corn 79-80→ corn 79	8.17–15.22	7.19–14.71	8.49–14.57	8.28–14.81
VII Corn 79-80→ corn 80	8.17–15.22	8.81–11.59	8.44–11.39	8.72–11.39
VIII Barley 79	8.1–15.3	8.30–18.80	8.14–7.30	7.83–18.88
IX Malt 79	8.3–18.8	8.10–15.30	8.36–15.87	8.11–15.32
X Barley/malt 79	8.3–15.7	8.30–18.80	7.81–17.62	8.02–18.42

TABLE IV
Slopes and Intercepts of the Linear Regression between Near-Infrared Reflectance Predicted (y) and Kjeldahl Protein (x)

Computation Design	Optimized Multiple Linear Regression		Optimized Multiple Regression Including Interaction	
	Slope	Intercept	Slope	Intercept
I	0.800	2.3	0.887	1.3
II	0.637	3.7	0.678	3.3
III	0.819	2.0	0.860	1.5
IV	1.135	0.47	0.715	2.9
V	0.609	4.7	0.898	1.2
VI	0.729	3.1	0.801	2.3
VII	0.794	2.1	0.811	1.9
VIII	0.887	1.4	0.922	0.94
IX	0.990	0.2	0.996	0.10
X	0.899	1.3	0.941	0.73

TABLE V
Standard Errors of Estimate of the Linear Regression between NIR Predicted (y) and Kjeldahl Protein (x)

Computation Design	Optimized Multiple Linear Regression	Optimized Multiple Regression Including Interaction
I	0.45	0.61
II	0.26	0.24
III	0.44	0.37
IV	0.55	0.25
V	0.49	0.56
VI	0.48	0.40
VII	0.32	0.31
VIII	0.44	0.46
IX	0.45	0.41
X	0.55	0.44

half of the 1979 and 1980 corn samples. In computation designs IV–VII, calibration equations were first prepared on data from corn samples from one or both years and then used to predict protein content in samples from either the same or different years. Similarly, in computation designs VIII–XI, a single combined calibration equation was prepared for barley and malt to predict the protein content of both. As indicated previously, the barleys (and corresponding malts) were from many locations throughout the United States and included cultivars and selections of the two- and six-row types.

One objective in developing these computation designs was to determine the extent to which a single calibration equation could be developed for corn that ranged widely in protein content (Table III) and had been harvested in different years. Another was to determine the need to produce a separate calibration equation for the combination of barley and malt. In every case, except for the 1979 corn, inclusion of the interaction term slightly improved the correlation coefficient (Table II). The improvement for corn was noticeable in prediction from one year to another and in prediction of each year from a combined calibration. The best improvement for barley and malt occurred in their combined calibration.

The ranges of the calibration and prediction samples given in Table III do not coincide. According to the Wilicoxon signed-ranks test (Conover 1971), the lower range predicted by the multiple regression for interaction is significantly ($P = 0.05$) closer to the lower range of the Kjeldahl protein than that predicted by multiple linear regression without interaction. The same is true for the upper range. Improvements are also seen in the nonparametric rankings of data in Tables IV and V.

Slopes and intercepts of the linear regression lines for predicted versus Kjeldahl protein are given in Table IV. The slopes of the multiple regression lines with interaction are closer to 1 than those slopes computed without interaction terms. Similarly, intercepts were closer to 0 when interaction terms were included in computations.

The standard errors of estimate of the linear regressions between the NIR-predicted (y) and Kjeldahl protein (x) are given in Table V. In general, the standard errors of estimate for multiple linear regressions with interaction were smaller than the errors without the interaction terms.

We believe that the small but consistent improvement in computation of NIR reflectance data from the inclusion of interaction terms warrants the modified calculation. The high correlation coefficients found in some instances need little improvement; yet, such improvement is possible and has been recorded. In other cases, inclusion of interaction terms increases the potential of computed results to convert marginal prediction values into meaningful ones. Furthermore, the use of a single, more universal and accurate calibration equation may have some value in reducing the need to calibrate instruments every year. Finally, in this age of versatile and industrial compact computers, the availability of canned programs that are easily adaptable to specific uses should reduce the complexity of added calculation.

In summary, the inclusion of interaction terms in developing linear regression lines used for the prediction of composition from NIR data provides several advantages, including improved agreement with ranges determined by the reference method, improved (to theoretical) slopes of regression lines, and reduced values of intercepts of linear regression lines.

LITERATURE CITED

CONOVER, W. J. 1971. Practical Nonparametric Statistics. John Wiley & Sons, Inc., New York.

DICKSON, A. D., STANDRIDGE, N. N., and BURKHART, B. A. 1968. The influence of rust infection in Larker barley on malting and brewing quality. Page 37 in: Am. Soc. Brew. Chem. Proc. The Society, St. Paul, MN.

HYMOWITZ, T., DUDLEY, J. W., COLLINS, F. I., and BROWN, C. M. 1974. Estimations of protein and oil concentration in corn, soybean, and oat seed by near-infrared light reflectance. Crop Sci. 14:713.

NORRIS, K. H. 1978. Near-infrared reflectance spectroscopy—The present and future. Page 245 in: Cereals '78: Better Nutrition for the World's Millions. Y. Pomeranz, ed. Am. Assoc. Cereal Chem., St. Paul, MN.

NORRIS, K. H., and WILLIAMS, P. C. 1977. Optimization of mathematical treatment of reflectance data for the estimation of protein. (Abstr.) Cereal Foods World 22:461.

SAS. 1979. SAS Users Guide, 1979 ed. SAS Institute, Inc., Casy, NC.

SHENK, J. S., LANDA, I., HOOVER, M. R., and WESTERHAUS, M. O. 1981. Description and evaluation of a near infrared reflectance spectro-computer for forage and grain analysis. Crop Sci. 21:355.

WILLIAMS, P. C., and PANFORD, J. A. 1979. Use of the Neotec Model 6350 research composition analyzer for optimization of wavelength and mathematic treatment for analysis of hard red spring wheat for protein. (Abstr.) Cereal Foods World 24:455.

[Received February 24, 1983. Revision received January 26, 1984. Accepted February 1, 1984]