Vis/NIR spectroscopy to estimate crude protein (CP) in Alfalfa crop: Feasibility study

Mohammadmehdi Maharlooei[1], Seyed Ahmad Mireei[2], Alimohammad Shirzadifar[3], Saravanan Sivarajan[3], Sreekala G. Bajwa[3*], Marisol Berti[4], John Nowatzki[3]

1- Department of Mechanical Engineering of Biosystems, Shahid Bahonar University of Kerman, Kerman, Iran.
2- Department of Biosystems Engineering, College of Agriculture, Isfahan University of Technology, Isfahan 84156-83111, Iran.
3- Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND, USA.
4- Department of Plant Sciences, North Dakota State University, Fargo, ND, USA.

A paper from the Proceedings of the
13th International Conference on Precision Agriculture
July 31 – August 4, 2016
St. Louis, Missouri, USA

Abstract. The fast and reliable quality determination of alfalfa crop is of interest for producers to make management decisions, the dealers to determine the price, and the dairy producers for livestock management. In this study, the crude protein (CP), one of the main quality indices of alfalfa, was estimated using the visible and near-infrared (Vis/NIR) spectroscopy. A total of 68 samples from various variety trials of alfalfa crop were collected under the irrigated and rainfed conditions. The diffuse reflectance spectral data of the undisturbed alfalfa stem and leaves were collected in the wavelength range of 400-2500 nm using a portable fiber optic spectroradiometer. Different spectral pretreatments were implemented to eliminate the irrelevant information from spectral data. Partial least squares regression (PLSR) method was then used to extract the CP predictive models using the spectral data as independent variables and the CP values obtained from the standard analytical method as dependent one. Primary results showed that the wavelength region of 400-900 nm had

the better predictive ability in comparison with the 900-2500 nm spectral region. In the 400-900 nm region, the best model was obtained from multiplicative scatter correction (MSC) preprocessing with a correlation coefficient in leave-one-out cross validation (r_{cv}) of 0.758 and a root mean square error in leave-one-out cross validation (RMSECV) of 1.096%. While, the best model in the spectral region of 900-2500 nm resulted from 2nd derivative pretreatment with an r_{cv} and RMSECV of 0.627 and 1.309%, respectively. When the whole spectral region (400-2500 nm) was used, no improvement in the predictive ability of PLS models was achieved in comparison with 400-900 nm region, where the best model resulted in r_{cv} and RMSECV values of 0.757 and 1.098%, respectively. Collecting the data during the subsequent years and increasing the harvesting times in each year can improve the robustness of CP predictive models by enhancing the variability in CP value.

Keywords. Crude protein, Spectroradiometer, Spectral pretreatment, Partial Least Squares Regression.

Introduction

Today, precision agriculture technologies help farmers in decision making using site-specific management system. There is a widely held belief among scientists that the most important part of precision agriculture cycle is in-field data acquisition stage (Morgan and Ess, 1997). Although the scientist primarily believed that the main result of precision agriculture leads to yield management, now they understand that crop quality is much more affected by site-specific management of production inputs than crop quantity (Blackmore, 1996). Many studies have been conducted in the recent decades to fulfill these necessary demands. The results of a number of these studies have led to some commercial yield mapping systems in the grain crops. Further investigations are still in progress to accomplish these needs in non-grain crops specially hays and forages.

Like yield monitoring systems, crop quality monitoring also needs on-the-go sensors to measure quality indices. Most of these efforts are based on employing near-infrared spectroscopy (NIRS) and remote sensing (RS) techniques.

Approaches based on NIRS have been applied to feces and/or forage and models were developed to analyze the diet quality of grazing animal intake (Stuth et al., 1989; Coleman et al., 1989; Coleman and Murray, 1993; Leite and Stuth, 1995; Lyons et al., 1995; Coates, 2000; Decruyenaere et al., 2002; Stuth et al., 2003). Lyons and Stuth (1992) found that monitoring forage diet quality and intake using NIRS scanning of fecal samples appeared to be promising. They demonstrated that grass in vivo digestibility can be estimated by NIRS applied to feces with the same accuracy as that obtained with conventional analysis methods. If there are appropriate calibration equations, NIRS is a rapid and non-destructive technology that could predict the digestibility and intake of a large set of similar samples.

The aim of this research was hence to evaluate the feasibility of using Vis/NIR spectroscopy for nondestructive measuring the Crude Protein (CP) of alfalfa.

Materials and methods

Sample preparation and treatments

In order to collect the alfalfa samples with a wide range of CP values, five different varieties of the crop were hand harvested from North Dakota State University (NDSU) research fields located in Fargo and Carrington. Data was collected from four different field conditions with respect to planting dates and irrigation availability. The Fargo site was rain-fed only and Carrington had both irrigated and rain-fed experiments. Experiments were planted in 2011 and 2014 in Fargo and in 2014 in Carrington. These fields were selected to include variability associated with irrigation and alfalfa's growth cycle on CP. Experiments in Fargo included five and six alfalfa varieties for 2014 and 2011 respectively and four replications for each. Experiments in Carrington included five alfalfa varieties and four replications for both irrigated and rain-fed conditions. A Total number of 86 samples (1 sample per each treatment and reflection) were hand harvested from middle rows at 10% bloom stage. In order to have enough leaves and stems for laboratory analysis, each sample consisted of 20 stems, cut from the stem collar. All samples were placed in the plastic bags with a trip cooler which guaranteed the stems freshness.

The reflectance spectral data of the undisturbed samples were collected using a portable Vis/NIR fiber optic spectrometer (OceanOptics, FL, USA). The whole unit consists of two spectrometers: USB 2000, covering the visible portion of the spectrum (200 nm to 1100 nm) and NIRQuest, covering the infrared portion of the spectrum (1100 nm to 2500 nm). Both these devices were integrated together to collect the spectral data in the wavelength range of 400-2500nm. To collect the reflectance spectral data, samples were lied on a flat surface and reflectance data was collected on three points of plant stems and averaged.

The CP values of the samples were determined using AOAC laboratory standard analytical methods as a reference (AOAC, 2000). The Nitrogen content which is directly related to CP was measured by using Kjeldahl method (Eq. 1).

$$CP\% = 6.25 \times N\% \quad (1)$$

Where CP= Crude protein content

N= Plant tissue Nitrogen Content

Spectral data analysis

In order to analyze the spectral data, all data were first reorganized in the Microsoft Excel V2007 software and then imported to the Unscrambler V9.7 (CAMO AS, Trondheim, Norway) software package for further analysis. In the first step, all the reflectance spectra were converted into the absorption ones using the equation of A=log(1/R). Partial least square regression (PLSR) was used to develop the CP predictive models. According to the ability of the used spectrometer, the PLSR models were developed using spectral regions of 400-900 nm, 900- 2500 nm, and the whole spectral region of 400-2500 nm. This was carried out to identify the best spectral region for estimating the CP. Moreover, different spectral pretreatment including Savitzky-Golay smoothing (3 nm smoothing points), area normalization, multiplicative scatter correction (MSC), 1st and 2nd derivatives, baseline correction, offset elimination, and standard normal variate (SNV) were applied on the raw spectra to remove any kind of irrelevant information. The leave-one-out cross validation (LOOCV) method was used to evaluate the predictability of the developed models.

Results and Discussion

Crude protein (CP) predictive models

Table 1 shows the performance of PLS models in predicting the average CP using different spectral regions and pretreatments procedures. Generally, the wavelength region of 400-900 nm resulted in better predictive ability in comparison with the 900-2500 nm spectral region. Among different pretreatment methods in this region, MSC led to the best model with a correlation coefficient in leave-one-out cross validation (r_{cv}) of 0.758 and a root mean square error in leave-one-out cross validation (RMSECV) of 1.0960. The best model in 900-2500 nm, however, obtained from 2nd derivative preprocessing in which the corresponding r_{cv} and RMSECV reached to relatively poor values of 0.627 and 1.3092, respectively. Moreover, by using the whole spectral region (400-2500 nm), the similar results with 400-900 nm region were achieved where the best model with MSC pretreatment resulted in the r_{cv} and RMSECV values of 0.757 and 1.0981, respectively. It should also be noted that by using the 400-900 nm spectral region, the PLS models generally were developed with the fewer

number of latent variables (LVs) which can be due to the fewer number of input wavelengths variables. Furthermore, the results presented in Table 1 were achieved using all of 68 samples in the data set and no outlier were detected and eliminated in the first step of the analysis.

Table 1. Effect of different spectral regions and preprocessing methods on the performance of PLS models in predicting the CP.

Wavelength region (nm)	Spectral pretreatment	LV	Calibration		Leave-out-out Cross Validation	
			r_c	RMSEC	r_{cv}	RMSECV
400-900	Raw	5	0.804	0.9851	0.736	1.1374
	Savitzky-Golay	5	0.803	0.9857	0.736	1.1376
	Area Normalization	4	0.802	0.9882	0.744	1.1234
	MSC	**3**	**0.798**	**0.9977**	**0.758**	**1.0960**
	1st Der.	1	0.741	1.1114	0.725	1.1571
	2nd Der.	1	0.711	1.1646	0.694	1.2092
	Baseline Offset	5	0.804	0.9845	0.734	1.1414
	Linear Baseline Correction	3	0.791	1.0125	0.745	1.1216
	SNV	4	0.800	0.9924	0.746	1.1187
900-2500	Raw	7	0.791	1.0124	0.572	1.3782
	Savitzky-Golay	8	0.787	1.0194	0.588	1.3593
	Area Normalization	6	0.742	1.1100	0.621	1.3173
	MSC	6	0.713	1.1618	0.557	1.3961
	1st Der.	5	0.715	1.1577	0.583	1.3651
	2nd Der.	**3**	**0.689**	**1.1996**	**0.627**	**1.3092**
	Baseline Offset	7	0.731	1.1299	0.580	1.3691
	Linear Baseline Correction	8	0.787	1.0194	0.588	0.3593
	SNV	5	0.692	1.1953	0.551	1.4012
400-2500	Raw	6	0.825	0.9357	0.723	1.1602
	Savitzky-Golay	6	0.825	0.9365	0.724	1.1592
	Vector Normalization	4	0.812	0.9657	0.744	1.1225
	MSC	**5**	**0.826**	**0.9335**	**0.757**	**1.0981**
	1st Der.	4	0.857	0.8545	0.733	1.1431
	2nd Der.	2	0.738	1.1185	0.707	1.1879
	Baseline Offset	6	0.824	0.9383	0.729	1.1510
	Linear Baseline Correction	4	0.797	1.0002	0.734	1.1408
	SNV	4	0.812	0.9651	0.744	1.1227

LV: latent variables; r_c: correlation coefficient in calibration; r_{cv}: correlation coefficient in validation

In order to detect the outliers, the ratio of NIR standard deviation of each individual sample (s(ei)) to the total residual standard deviation (s(e)) was investigated. For the best model obtained from 400-900 nm region, a total of four samples were detected as outlier and removed from the sample set. Table 2 summarizes the results of the best PLS models for predicting the CP values using different spectral regions. As shown, the PLS model based on 400-900 nm resulted in relatively good results with r_{cv} and RMSECV of 0.825 and 0.9334%, respectively. The predicted vs. true CP values obtained from this model is depicted in Fig. 1a. The regression coefficient of the model with respect to wavelength variables is illustrated in Fig. 2a.

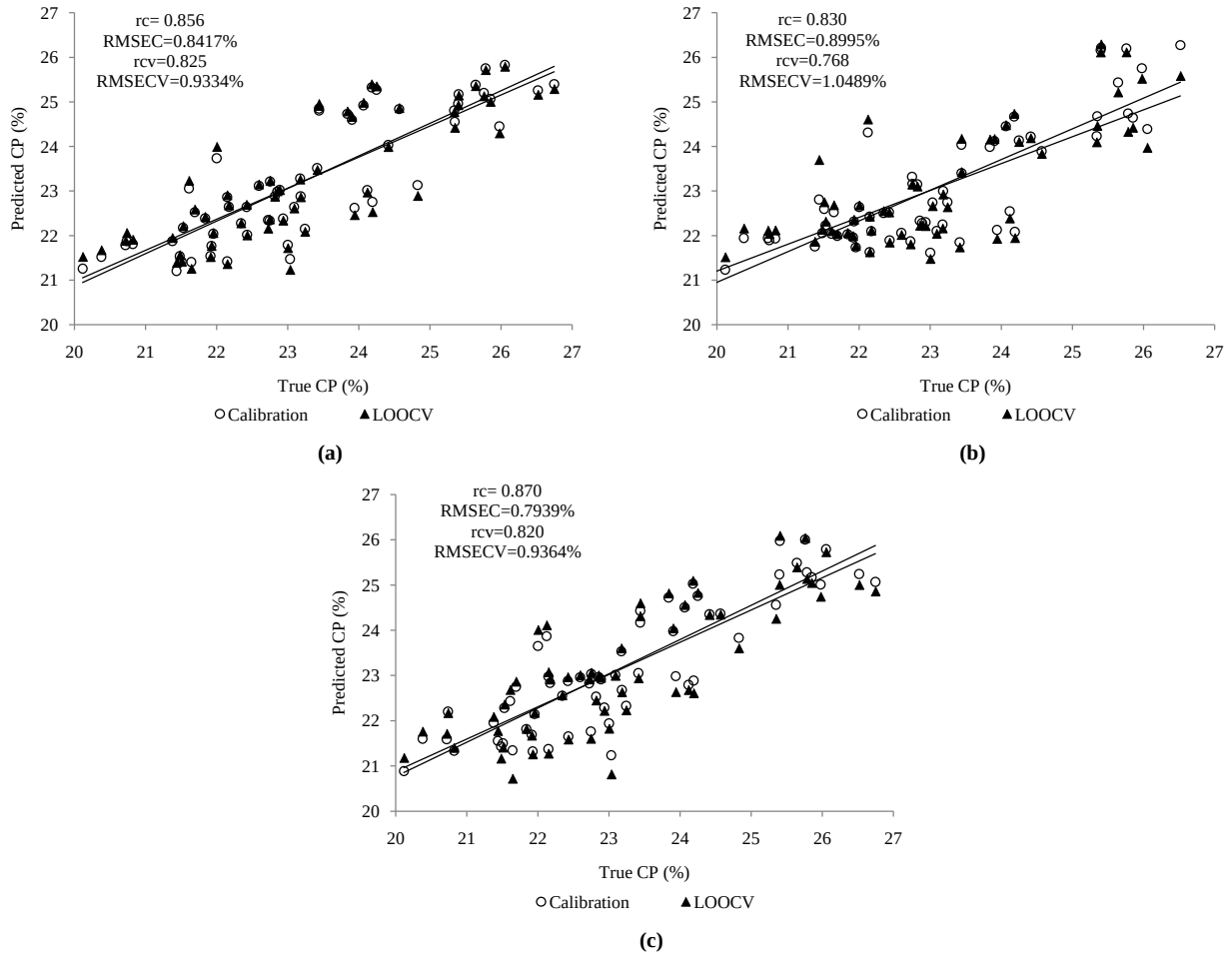

Fig. 1. Predicted vs. true CP values of the best PLS models obtained from (a) 400-900 nm, (b) 900-2500 nm, and (c) 400-2500 nm spectral regions.

In order to detect the outliers, the ratio of NIR standard deviation of each individual sample (s(ei)) to the total residual standard deviation (s(e)) was investigated. For the best model obtained from 400-900 nm region, a total of four samples were detected as outlier and removed from the sample set. Table 2 summarizes the results of the best PLS models for predicting the CP values using different spectral regions. As shown, the PLS model based on 400-900 nm resulted in relatively good results with r_{cv} and RMSECV of 0.825 and 0.9334%, respectively. The predicted vs. true CP values obtained from this model is depicted in Fig. 1a. The regression coefficient of the model with respect to wavelength variables is illustrated in Fig. 2a.

Table 2. The results of best PLS models for predicting the CP values in different spectral regions.

Wavelength region (nm)	Spectral pretreatment	LV	Calibration		Leave-out-out Cross Validation	
			r_c	RMSEC (%)	r_{cv}	RMSECV (%)
400-900	MSC	3	0.856	0.8417	0.825	0.9334
900-2500	2nd Der.	2	0.830	0.8995	0.768	1.0489
400-2500	MSC	5	0.870	0.7939	0.820	0.9364

LV: latent variables; r_c: correlation coefficient in calibration; r_{cv}: correlation coefficient in validation

Regarding 900-2500 spectral region, the outlier detection analysis showed that four samples should be removed from the sample set. After outlier removal, the r_{cv} and RMSECV of the developed model reached 0.768 and 1.0489%, respectively (Table 2). Figure 1b shows the predicted vs. true CP values obtained from this model and the respective regression coefficient values of the model are illustrated in Fig. 2b.

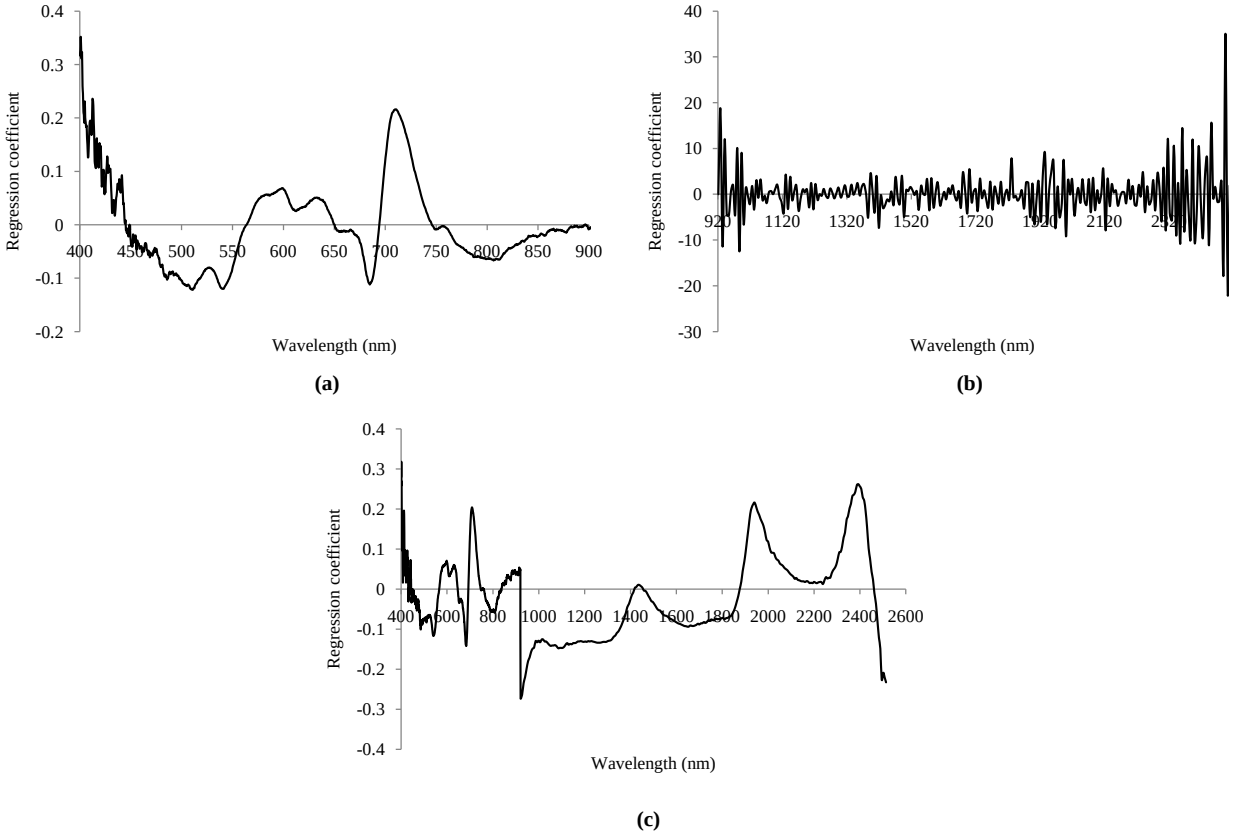

Fig. 2. Regression coefficients of the best PLS models obtained from (a) 400-900 nm, (b) 900-2500 nm, and (c) 400-2500 nm spectral regions.

Finally, in using the entire spectral data (400-2500 nm), four samples were detected as outlier and eliminated from the sample set. The developed PLS model resulted in the close predictive ability with the PLS model based on 400-900 nm region, and the r_{cv} and RMSECV were 0.820 and 0.9364%, respectively. The predicted vs. true CP values obtained from this model and the respective regression coefficients are depicted in Figs. 1c and 2c, respectively.

One thing that should be noted is that in the Figs. 2a and 2c, the wavelengths of 683 and 706 nm in the region of 400-900 nm, and the wavelengths of 1935 and 2387 in the region of 900-2500 nm had the higher regression coefficient values with respect to other wavelengths. As a rule-of-thumb, the wavelengths with a larger value of regression coefficient have a higher

degree of determination for describing the desired parameter (CP). In our case, the high regression coefficient value in the wavelengths of 683 is associated with plant pigments such as chlorophyll which indirectly can be related to protein content. Moreover, the peak in regression coefficient curve around the 706 nm can be related to C-H fourth overtone associated with the carbohydrate content of the crop. In the longer wavelengths of NIR region, the wavelength of 1935 is located in the strong water absorption waveband of 1900-1950 nm and thus is related to the water content of the crop. Finally, the wavelength of 2387 can be related to C-H stretch and CH_2 deformation combination associated with starch (Pojic et al 2010).

Summary

In this research, the feasibility of using spectroscopy in order to estimate the crude protein in alfalfa crop for five different varieties and four different field conditions (rain fed and irrigated fields in Carrington and Fargo) were investigated. Primary results showed a good potential of using this technique in estimating crude protein as a non-destructive test. The results of using the different preprocessing methods in three different wavelength regions including 400-900 nm, 900 -2500 nm, and 400 -2500 nm were compared. The results showed that multiplicative scatter correction (MSC) method in the wavelength range of 400-900 nm and 400-2500 nm and 2^{nd} derivative method in 900-2500 nm range led to the better CP predictability compared to other methods. Further investigations with different varieties need to be performed to develop more robust models to predict crude protein in alfalfa crop.

References

AOAC. (2000) Official methods of analysis of AOAC international (15th ed). Washington, USA, Association of Official Analytical Chemistry.

Coates, D.B., & Penning, P. (2000). Measuring animal performances. In 't Mannetje L. & Jones R.M., (Eds.) Field and laboratory methods for grassland and animal production research. (pp. 353-402) Wallingford, UK: CABI Publishing.

Coleman, S.W., Stuth, J.W., & Holloway, J.W. (1989). Monitoring the nutrition of grazing cattle with near-infrared analysis of feces. In Proceedings of the 16th International grassland congress. (pp. 881-882) Nice, France.

Decruyenaere V. et al., (2002). Improvement and validation of the NIRS analysis applied to feces to measure grass intake in pasture. In Durand J.L., Emile J.C., Huygue Ch. & Lemaire G., (Eds.) Proceedings of the 19th General meeting of the European Grassland Federation on multi-function grasslands, quality forages, animal products and landscapes, (pp. 196-197) La Rochelle, France. British Grassland Society.

Coleman, S.W., & Murray I. (1993). The use of near-infrared reflectance spectroscopy to define nutrient digestion of hay by cattle. *Animal Feed Science and Technology,* 44(3-4), 237-249.

Leite, E. R., & Stuth, J. W. (1995). Fecal NIRS equations to assess diet quality of free-ranging goats. *Small Ruminant Research*, *15*(3), 223-230.

Lyons, R. K., & Stuth, J. W. (1992). Fecal NIRS equations for predicting diet quality of free-ranging cattle. *Journal of Range Management*, 45, 238-244.

Lyons, R. K., Stuth, J. W., & Angerer, J. P., (1995). Technical note: Fecal NIRS question field validation. *Journal of Range Management*. 48, 380–382.

Pojić, M., Mastilović, J., Palić, D., & Pestorić, M. (2010). The development of near-infrared spectroscopy (NIRS) calibration for prediction of ash content in legumes on the basis of two different reference methods. Food chemistry, 123(3), 800-805.

Stuth, J. W., Kapes, E.D., & Lyons, R.K., (1989). Use of near infrared spectroscopy to assess nutritional status of cattle diets on rangeland. In Proceedings of the 16th International grassland congress. (pp. 889-890) Nice, France.

Stuth, J., Jama, A., & Tolleson, D. (2003). Direct and indirect means of predicting forage quality through near infrared reflectance spectroscopy. Field Crops Research, 84(1), 45-56.