Prediction of crude protein content in rice grain with canopy spectral reflectance

H. Zhang[1,2], T.Q. Song[1,2], K.L. Wang[1,2], G.X. Wang[3], H. Hu[4], F.P. Zeng[1,2]

[1]*Key Laboratory of Agro-Ecological Processes in Subtropical Region, Institute of Subtropical Agriculture, Chinese Academy of Sciences, Changsha, P.R. China*
[2]*Huanjiang Observation and Research Station for Karst Ecosystem, Chinese Academy of Sciences, Huanjiang, P.R. China*
[3]*Institute of Agricultural Ecological Research, College of Life Sciences, Zhejiang University, Hangzhou, P.R. China*
[4]*Key Laboratory of Digital Agriculture, Institute of Digital Agricultural Research, Zhejiang Academy of Agricultural Sciences, Hangzhou, P.R. China*

ABSTRACT

Non-destructive and rapid monitoring methods for crude protein content (CPC) in rice grain are of significance in nitrogen diagnosis and grain quality monitoring, and in enhancing nutritional management and use efficiency. In this study, CPC and canopy spectra in rice were measured based on rice field experiment. Key spectral bands were selected by principal component analysis (PCA) method, and the predicted models were built by multiple linear regressions (MLR), artificial neural network (ANN) and partial least squares regression (PLSR). The results showed that there is a significant correlation between CPC content and key spectral bands. The results of prediction for the three models were in order of PLSR > ANN > MLR with correlation values of 0.96, 0.92 and 0.90, respectively, for the validation data. Therefore, it is implied that CPC in rice (grain quality) could be estimated by canopy spectral data.

Keywords: nitrogen content; principal component analysis; partial least squares regression; artificial neural networks; key spectral bands

Crude protein content (CPC) in rice grain, as an important component of rice nutritional quality, is important for health of people whose main food in daily life is rice (Tang et al. 2004). Accurate and timely estimation of the CPC of rice grain can help farmers make appropriate decisions concerning fertilizer application, rice variety selection, and harvest classification (Diker and Bausch 2003, Wang et al. 2004). Determining CPC with Kjeldahl and combustion methods is a commonly used laboratory method. This laboratory method is accurate and reliable, but usually time consuming and costly (Li et al. 2006, Starks et al. 2006a,b). In contrast, remote sensing could provide spatial and temporal measurements of surface properties and was recognized as a reliable method for the estimation of various variables related to physiology and biochemistry (Hinzman et al. 1986, Diker and Bausch 2003).

Many studies described the capability of remote sensing technology to rapidly measure many crop nitrogen content (Rondeaux et al. 1996, Haboudane et al. 2004, 2008). At the leaf level, values derived from spectral measurements admittedly depend predominantly on the amount of chlorophyll. The best relationships between chlorophyll and

Supported by the Ministry of Science and Technology of China, Project No. 2011AA100503; by the National Natural Science Foundation of China, Projects No. 31070425, 31000224, 30970508 and U1033004; by the Chinese Academy of Sciences Action Plan for the Development of Western China, Project No. KZCX2-XB3-10; by the West Light Foundation of Chinese Academy of Sciences, Project No. 2012F046, and partially supported by the Prospective Study Program of Key Laboratory of Digital Agriculture from Zhejiang Academy of Agricultural Sciences (ZAAS).

reflectance measurements were obtained (Zhou and Wang 2003, Cartelat et al. 2005). At the canopy level, spectral measurements also appear to correlate well with total aerial N (Lukina et al. 2001, Zhao et al. 2004, Mistele and Schmidhalter 2008). At the landscape scales, the spectral reflectance of TM channel 5 derived from canopy spectra or image data at grain filling stage was all significantly correlated to grain protein content in wheat (Zhao et al. 2005). These studies suggest that remotely sensed data could be used to monitor plant nitrogen status from leaf level to landscape level.

Multiple linear regression (MLR) is widely used to regress reflectance measures against crop nitrogen concentrations. But, as well known, low estimate accuracy is its drawback (Martin and Aber 1997, Yi et al. 2007). Moreover, the selected wavelengths in stepwise regression are not always related to the biochemical of interest but to wavelengths that are related to biomass amount or other biochemicals (Yi et al. 2007). Compared with MLR, artificial neural networks (ANNs) have the ability to deal complicated spectral information with target attributes without any constraints for sample distribution making them ideal for describing the intricate and complex nonlinear relationships which exist between spectral signatures and various crop conditions (Gorr et al. 1994, Kimes et al. 1998). In addition, partial least squares regression (PLSR) is also an important statistical method that bears some relation to principal components regression (PCA), instead of finding hyperplanes of maximum variance between the response and independent variables. It can use fewer new variables than the original ones to figure out the difficult analysis such as the superposition of a spectral band and find a linear regression model by projecting the predicted variables and the observable variables to a new space (Rännar et al. 1994, Tenenhaus et al. 2005).

Previous studies implemented the regression model between reflectance and plant nitrogen by using MLR, ANNs and PLSR (He et al. 2006, Yi et al. 2007). However, the predictive ability of the three modelling methods using fresh canopy-level spectral reflectance has not been well compared. Our studies showed that the ANNs model provided better accuracy in retrieval of rice neck blasts compared with the results from the MLR model (Zhang et al. 2011). The objectives of the present investigation were to compare the predictive power of MLR, ANNs and PLSR methods using different models, i.e., (i) the R-MLR (stepwise multiple linear regression) model based on reflectance, (ii) the R-ANN model based on spectral reflectance, and (iii) R-PLSR model based on spectral reflectance, and to finally propose a suitable estimation model.

MATERIAL AND METHODS

Field experiments. The field experiments were conducted at State Monitoring Station of Rice Soil Fertility and Fertilizer Effect, Haining, Zhejiang Province, China, which is located at 120°25'E, 30°26'N. Rice (Zhegeng 22) was examined in a two-year field experiment (years 2009 and 2010). The average annual temperature was 15.3°C and the average annual precipitation was 1350 mm per year with the highest values occurring in the summer. The experiments were conducted on different fields each about 1 ha in size. Heterogeneous fields were chosen to obtain differences in both the N status and biomass. The experimental design consisted of five fertilization rates (0, 80, 120, 160 and 200 kg N/ha) each with 10 replicates for a total of 50 plots.

Measurements of spectral reflectance. AvaSpec-2048 spectrometer (Avantes inc., Apeldoorn, Netherlands) was used to get spectral of all canopy reflectance. This spectrometer is fitted with a 25 field of view fiber optics, operating in the 200–1100 nm spectral region with a sampling interval of 1.4 nm and spectral resolution of 1.2 nm. The measurements were carried out from a height of 1.0 m above the canopy and 0.44 m view diameter under clear sky conditions between 10:00 h and 14:00 h (Beijing local time). Measurements of vegetation radiance were made at 10 sample sites in each plot, with each sample averaging 20 scans at an optimized integration time. The saved spectrum file contained continuous spectral reflectance at 0.6 nm step over the band region of 200–1100 nm. A panel radiance measurement was taken before and after the vegetation measurement by two scans each time.

Measurements of crude protein content (CPC). In September 2009 and 2010, after the canopy spectral measurements, rice grain were put into an oven to dry at 105°C for half an hour and then at 70°C till the constant weight was acquired. The dry grains were then ground with a mortar and pestle for N measurement. Nitrogen concentration (N) was determined by Kjeldahl after acid digest and the results were expressed in mg N/g grain dry weight. The CPC in rice grain was computed by N × 5.95.

Data analysis. Spectral data were firstly exported from binary by using the manufacturer's

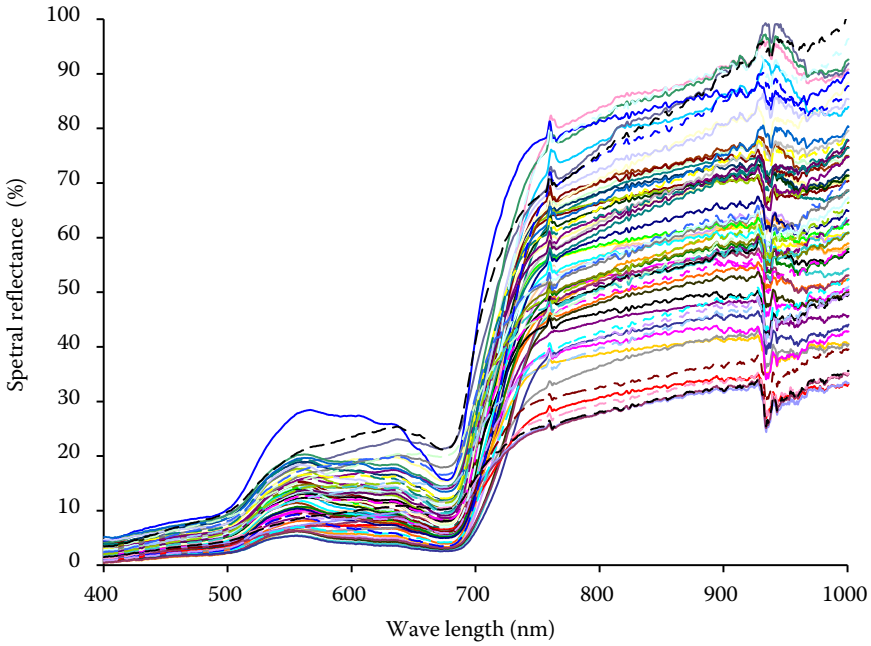

Figure 1. Spectral reflectance of canopy in rice changing with wavelength

program of AvaSoft 7.3 for USB 2 (Apeldoorn, the Netherlands). Reflectance data were smoothed with a five-point moving average to suppress instrumental and environmental noise in the data before the data were further analyzed.

Because spectral reflectance often contains large amounts of redundant information, the main purpose of principal component analysis (PCA) is to build the linear combinations of the original variables that represent the most original variations of the data set being investigated. After key spectral bands were selected by the PCA method, the predicted models were built by MLR, ANN and PLSR model. The precision of regression models was assessed by root mean square error of training (RMSEt), root mean square error of prediction (RMSEp), correlation coefficient of training (R_t^2), correlation coefficient of predication (R_p^2), and residual prediction deviation (RPD).

RESULTS

The canopy in rice under different nitrogen fertilization levels exhibited a different raw reflectance in the 520–680 nm and 720–1000 nm spectral regions (Figure 1). Because excessive spectral bands and noise interference may affect precision of regression model, principal component analysis (PCA) method was used to reduce original input variables dimension. In other words, the number of variables can be reduced by removing the lower-level components without any notable loss of information contained in the original data set by PCA. Based on PCA, the first principal component represents 71.20% of the spectral information, and the first four principal components contain information as high as 94.48%, while the other principal components contain less information (Table 1).

Table 1. Percentage of explained variance for the first four principal components (PCs)

	PC1	PC2	PC3	PC4
Explained variance (%)	71.20	16.49	5.57	1.22
Cumulative variance (%)	71.20	87.69	93.26	94.48

After principal components selected by PCA, the predicted models were built by multiple linear regressions (MLR), artificial neural network (ANN) and partial least squares regression (PLSR) method

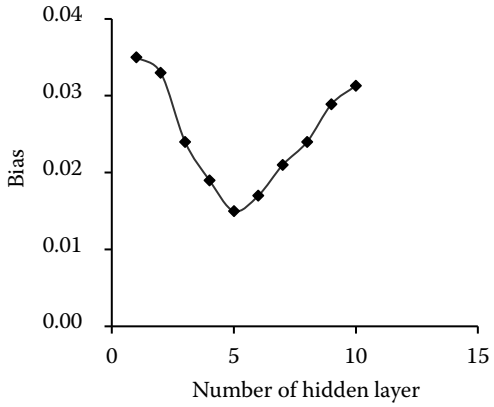

Figure 2. Bias changes with the point number of hidden layer

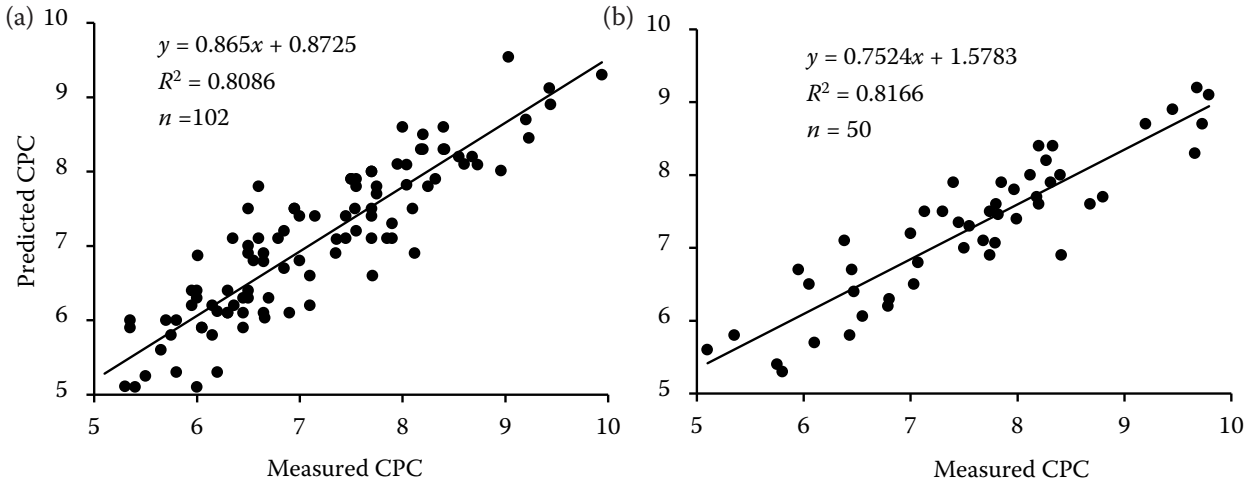

Figure 3. Scatter plots of measured versus predicted crude protein content (CPC) for linear regression based on multiple linear regressions (MLR). (a) calibration data set and (b) validation data set

respectively. For MLR model, factor scores of the first four principal components scores were used as independent variables to conduct regression. The linear model was expressed as equation:

$$CPC = 9.91 - 0.0026 \times PC1 - 0.0011 \times PC2 + + 0.0035 \times PC3 - 0.078 \times PC4$$

For ANN model, which included three layers network architecture, consisting of one input layer, one hidden layer, and one output layer, was established in MATLAB 6.5 (Stanford, USA). Then, first four principal components scores were put as input variables, and CPC as target variable. The number of neurons in the hidden layer is optimized by using the training, validation, and testing data sets. A total of 102 samples out of 152 samples were used for training, and the rest was equally divided for validation. The bias was used to select the optimum number of neurons in the hidden layer. The numbers of neurons in the hidden layers were determined when the minimum values of bias were found (Figure 2). After time-consuming trials, the ANN model with a 4-5-1 architecture was determined.

For PLSR model, spectra were imported into Unscramble V9.7 software (CAMO, Oslo, Norway). The region of 400–1000 nm was analyzed by PLSR by mean-centering, normalizing. Model was constructed using CPC standards in order to account for spectral differences based on matrix. Model was also cross-validated using a leave-one-out approach. Key bands were identified and PLSR model was calculated.

The (RMSEt), RMSEp, R_t^2, R_p^2 and RPD for calibration and validation data sets for all models are summarized in Table 2. It could be seen that the best result was obtained by the PLSR model, which was the case for both the calibration and the validation sets, and followed by the ANN, and MLR models (Figures 3–5, Table 2)

DISCUSSION

For a typical crop canopy, reflectance is low between the 480- and 680- nm region due to the strong absorption by chlorophylls and other pigments, but is high in the NIR region due to the

Table 2. Results of rice seed protein models in calibration and validation

Model	Calibration (n = 102)			Validation (n = 50)			
	r_c	$RMSE_c$	Bias	r_p	$RMSE_p$	Bias	RPD
MLR	0.8992	0.5701	0.0041	0.9037	0.4281	0.0031	3.99
ANN	0.9172	0.3845	0.0032	0.9201	0.2525	0.0024	5.06
PLSR	0.9525	0.2530	0.0025	0.9570	0.1817	0.0015	6.83

MLR – multi regression model; ANN – neural network model; PLSR – partial least squares regression; r_c – correlation coefficient of calibration; r_p – correlation coefficient of validation; $RMSE_c$ – mean square root of interactive calibration; $RMSE_p$ – mean square root of interactive validation; RPD – relative stand error of predication

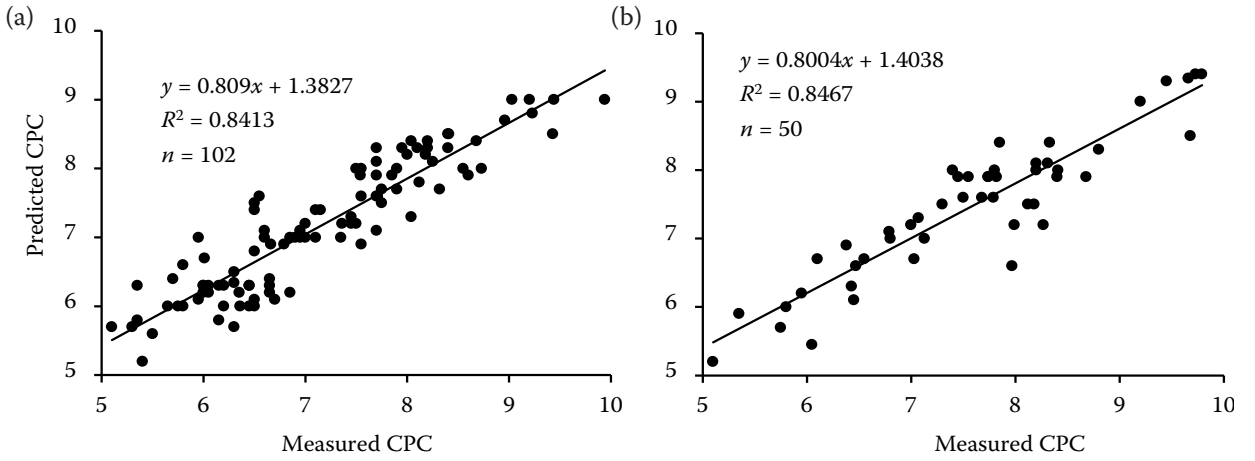

Figure 4. Scatter plots of measured versus predicted crude protein content (CPC) for linear regression based on artificial neural network (ANN). (a) calibration data set and (b) validation data set

microcellular structures in leaf material and canopy structures (Thomas and Oerther 1972, Feng et al. 2008). In this study, our results showed that there were two obvious band ranges between 520–680 nm and 720–1000 nm observed (Figure 1). Similar results were reported in previous studies (He et al. 2006, Li et al. 2006, Yi et al. 2007).

For the three models, the results of prediction of MLR, ANN and PLSR were in order of PLSR > ANN > MLR (Figures 3–5, Table 2). The prediction accuracy of the nonlinear model is better than the linear model. There was over fitting in MLR method because only some spectrum information was used, while other spectrum information was lost. Compared to MLR, the ANN has a strong advantage to fit the nonlinear problem. Some researchers implemented the ANN method for the analysis of spectral data, and for improving the inversion precision of crop biochemical parameters (He et al. 2006, Yi et al. 2007, 2010, Zhang et al. 2011). At the same time, we also found that the node number in hidden layer affected prediction bias (Figure 2). It is implied that the fitting accuracy of ANN model was less than that of PLSR model, which may be because of its over fitting and reducing generalization ability. The most important feature of PLSR model is that it can use all spectrum information, compress the sample quantity required, integrate highly related wavelength point into an independent variable, and establish regression model based on a few independent variables. PLSR model could avoid over fitting phenomenon through the internal inspection, and its fitting precision is higher than those of ANN and MLR.

In summary, this study showed the promising potential of CPC monitoring using canopy-level spectral reflectance and the three algorithms. Spectrometer (instrument equipment), chemical measurement software (data analysis) and model application (model inversion) were integrated. In addition, our study also improves the accuracy of

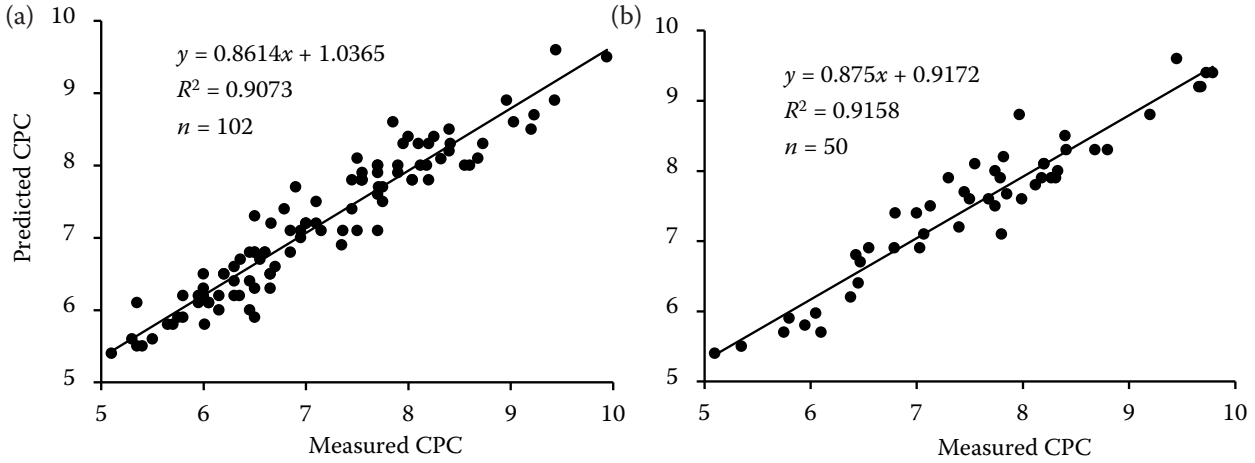

Figure 5. Scatter plots of measured versus predicted crude protein content (CPC) for linear regression based on partial least squares regression (PLSR). (a) calibration data set and (b) validation data set

spectral information acquisition and reliability of model building (Yi et al. 2010). As a method of data reducing and representation, PCA was very useful to analyze spectral reflectance data. However, the spectral response properties of vegetation canopy was also found to depend on atmospheric (e.g., illumination, cloudy shadow), edaphic (e.g., soil type, soil moisture), and biotic (e.g., crop variety, leaf area index) conditions (Sankaran et al. 2010). Further research is still needed to solve these questions.

Acknowledgements

The authors thank Jianna Li for help with data analysis, Dr. Wei Hu for improving language, Tingmao Wang for help with data collection in the experiment, and two anonymous reviewers for help improving the manuscript.

REFERENCES

Cartelat A., Cerovic Z.G., Goulas Y., Meyer S., Lelarge C., Prioul J.L., Barbottin A., Jeuffroy M.H., Gate P., Agati G., Moya I. (2005): Optically assessed contents of leaf polyphenolics and chlorophyll as indicators of nitrogen deficiency in wheat (*Triticum aestivum* L.). Field Crops Research, *91*: 35–49.

Diker K., Bausch W.C. (2003): Potential use of nitrogen reflectance index to estimate plant parameters and yield of maize. Biosystems Engineering, *85*: 437–447.

Feng W., Yao X., Zhu Y., Tian Y.C., Cao W.X. (2008): Monitoring leaf nitrogen status with hyperspectral reflectance in wheat. European Journal of Agronomy, *28*: 394–404.

Gorr W.L., Nagin D., Szczypula J. (1994): Comparative study of artificial neural network and statistical models for predicting student grade point averages. International Journal of Forecasting, *10*: 17–34.

Haboudane D., Miller J.R., Pattey E., Zarco-Tejada P.J., Strachan I.B. (2004): Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. Remote Sensing of Environment, *90*: 337–352.

Haboudane D., Tremblay N., Miller J.R., Vigneault P. (2008): Remote estimation of crop chlorophyll content using spectral indices derived from hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, *46*: 423–437.

He Y., Li X.L., Shao Y.N. (2006): Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model. Spectroscopy and Spectral Analysis, *26*: 850–853. (In Chinese)

Hinzman L.D., Bauer M.E., Daughtry C.S.T. (1986): Effects of nitrogen fertilization on growth and reflectance characteristics of winter wheat. Remote Sensing of Environment, *19*: 47–61.

Kimes D.S., Nelson R.F., Manry M.T., Fung A.K. (1998): Attributes of neural networks for extracting continuous vegetation variables from optical and radar measurements. International Journal of Remote Sensing, *19*: 2639–2663.

Li Y., Zhu Y., Tian Y., Yao X., Zhou C., Cao W. (2006): Quantitative relationship between leaf nitrogen accumulation and canopy reflectance spectral. Scientia Agricultura Sinica, *32*: 203–209. (In Chinese)

Lukina E.V., Freeman K.W., Wynn K.J., Thomason W.E., Mullen R.W., Stone M.L., Solie J.B., Klatt A.R., Johnson G.V., Elliott R.L., Raun W.R. (2001): Nitrogen fertilization optimization algorithm based on in-season estimates of yield and plant nitrogen uptake. Journal of Plant Nutrition, *24*: 885–898.

Martin M.E., Aber J.D. (1997): High spectral resolution remote sensing of forest canopy lignin, nitrogen and ecosystem processes. Ecological Applications, *7*: 431–443.

Mistele B., Schmidhalter U. (2008): Spectral measurements of the total aerial N and biomass dry weight in maize using a quadrilateral-view optic. Field Crops Research, *106*: 94–103.

Rännar S., Lindgren F., Geladi P., Wold S. (1994): A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. Journal of Chemometrics, *8*: 111–125.

Rondeaux G., Steven M., Baret F. (1996): Optimization of soil-adjusted vegetation indices. Remote Sensing of Environment, *55*: 95–107.

Sankaran S., Mishra A., Ehsani R., Davis C. (2010): A review of advanced techniques for detecting plant diseases. Computers and Electronics in Agriculture, *72*: 1–13.

Starks P.J., Zhao D., Philips W.A., Coleman S.W. (2006a): Herbage mass, nutritive value and canopy spectral reflectance of bermudagrass pastures. Grass and Forage Science, *61*: 101–111.

Starks P.J., Zhao D., Philips W.A., Coleman S.W. (2006b): Development of canopy reflectance algorithms for real-time prediction of bermudagrass pasture biomass and nutritive values. Crop Science, *46*: 927–934.

Tang Y.L., Huang J.F., Wang R.C. (2004): Study on estimating the contents of crude protein and crude starch in rice panicle and paddy by hyperspectra. Scientia Agricultura Sinica, *37*: 1282–1287. (In Chinese)

Tenenhaus M., Vinzi V.E., Chatelin Y.M., Lauro C. (2005): PLS path modeling. Computational Statistics and Data Analysis, *48*: 159–205.

Thomas J.R., Oerther G.F. (1972): Estimating nitrogen content of sweet pepper leaves by reflectance measurements. Agronomy Journal, *64*: 11–13.

Wang Z.J., Wang J.H., Liu L.Y., Huang W.J., Zhao C.J., Wang C.Z. (2004): Prediction of grain protein content in winter wheat (*Triticum aestivum* L.) using plant pigment ratio (PPR). Field Crops Research, *90*: 311–321.

Yi Q.X., Huang J.F., Wang F.M., Wang X.Z., Liu Z.Y. (2007): Monitoring rice nitrogen status using hyperspectral reflectance and artificial neural network. Environmental Science and Technology, *41*: 6770–6775.

Yi S.L., Deng L., He S.I., Zheng Y.Q., Mao S.S. (2010): A spectrum based models for monitoring leaf potassium content of *Citrus sinensis* (L.) cv. Jincheng Orange. Scientia Agricultura Sinica, *43*: 780–786. (In Chinese)

Zhang H., Hu H., Zhang X.B., Zhu L.F., Zheng K.F., Jin Q.Y., Zeng F.P. (2011): Estimation of rice neck blasts severity using spectral reflectance based on BP-neural network. Acta Physiologiae Plantarum, *33*: 2461–2466.

Zhao C.J., Liu L.Y., Wang J.H., Huang W.J., Song X.Y., Li C.J. (2005): Predicting grain protein content of winter wheat using remote sensing data based on nitrogen status and water stress. International Journal of Applied Earth Observation and Geoinformation, *7*: 1–9.

Zhao C.J., Zhou Q.F., Wang J.H., Huang W.J. (2004): Spectral indices redefined in detecting nitrogen availability for wheat canopy. Communications in Soil Science and Plant Analysis, *35*: 853–864.

Zhou Q.F., Wang J.H. (2003): Comparison of upper leaf and lower leaf of rice plants in response to supplemental nitrogen levels. Journal of Plant Nutrition, *26*: 607–617.

Received on September 6, 2012

Corresponding author:

Prof. Fuping Zeng, Chinese Academy of Sciences, Institute of Subtropical Agriculture, Key Laboratory of Agro-Ecological Processes in Subtropical Region, Changsha, P.R. China
phone (fax): + 86 731 8461 5233, e-mail: fpzeng@isa.ac.cn