ESTIMATION OF LEGUME CONTENT IN GRASS-LEGUME MIXTURES FROM HYPERSPECTRAL MEASURMENT

Kensuke Kawamura[*1], Nariyasu Watanabe[2], Seiichi Sakanoue[2], Hyo-Jin Lee[1], Jihyun Lim[1] and Rena Yoshitoshi[1]

[1] Associate Professor, Graduate Student, Graduate Student, Graduate Student,
Graduate School for International Development and Cooperation (IDEC), Hiroshima University,
1-5-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8529, Japan; Tel: +81-82-424-6929
E-mail: kamuken@hiroshima-u.ac.jp; tallwind25@hotmail.com; limjihyun7@gmail.com;
rena.yoshi1210@gmail.com

[2] Researcher, Researcher, National Agriculture and Food Research Organization (NARO), Hokkaido Agricultural Research Center, 1 Hitsujigaoka, Sapporo, Hokkaido 062-8555, Japan; Tel: +81-11-857-9313
E-mail: nariyasu@affrc.go.jp; saka@affrc.go.jp

KEY WORDS: Canopy Reflectance, Genetic Algorithm, Legume Content, Partial Least Squares Regression, Waveband Selection

ABSTRACT: Legume content in grass-legume mixtures is a key parameter for deciding the forage quality and the amount of fertilizer application to the pasture due to nitrogen (N) fixation. This study investigated the ability of field hyperspectral radiometer (400-2350 nm) with genetic algorithm partial least squares (GA-PLS) regression for estimating legume content in a mixed sown pasture of Hokkaido, Japan. Canopy reflectance data and plant samples were obtained from 50 selected sites in two seasons (n = 100); spring (May) and summer (July) 2007. The predictive accuracy of GA-PLS was compared with that of multiple linear regression (MLR), and standard full-spectrum PLS (FS-PLS) for spring, summer, and combined datasets. The predictive ability of the model was assessed by the coefficient of determination (R^2) and the root mean squared error of cross validation (RMSECV). Overall, the highest R^2 and the lowest RMSECV were obtained in the GA-PLS models for all datasets (R^2 = 0.62-0.86, RMSECV = 4.10-7.20%). In MLR, selected wavebands in the final models were blue (400-456 nm) and red bands (659-670 nm) in visible wavelength, red-edge region (704-724 nm), near infrared regions (813, 937, and 1121 nm), and shortwave infrared regions (2303-2344 nm) that are mainly linked to known biochemical components such as chlorophyll, nitrogen, lignin and cellulose. These results suggest that legume content in grass-legume mixtures can be predicted from field hyperspectral measurements, and that the predictive accuracy of the model can be improved by waveband selection using the GA-PLS.

1. INTRODUCTION

Legume content in grass-legume mixtures is a key parameter for deciding the forage quality and the amount of fertilizer application to the pasture due to nitrogen (N) fixation. Thus, timely assessment of legume content in grass-legume mixtures is required by land managers for efficient management of pastures. In laboratory, near-infrared spectroscopy (NIRS) is being a promising tool to determine clover content in grass-legume mixtures (Locher et al., 2005), which is based on diffuse reflectance of ground samples. The technology was used, at first, for predicting nutrients and feeding value in dried and fresh crop material (Berardo, 1997, Norris et al., 1976). However, the estimates of legume content involve clipping, bagging, oven-drying and weighing which are time consuming and resource intensive. Thus, their potential use for estimating legume content in the field and for application in large area is limited. In this context, remote sensing techniques offer a cost-effective solution for the quantitative estimation of pasture biochemical (e.g. chlorophyll and nitrogen contents) and biophysical variables (*e.g.* biomass, leaf area index) from local to regional scale.

Predicting herbage mass and forage quality status with field spectrometry is a complex undertaking, since plant reflectance is strongly influenced by multiple, overlapping absorption features (Curran et al., 1992, Ferwerda et al., 2006). Several studies applied stepwise multiple linear regression (MLR), which use several spectral wavelengths to estimate herbage mass and forage quality (Zhao et al., 2007). However, the drawbacks of extensive spectral overlap of individual biochemical properties, and multi-collinearity problems are well known. This usually occurs when the number of observations is smaller than the number of wavebands used in the analysis (Curran, 1989). In contrast, the partial least squares (PLS) regression has been recommended as an alternative approach to MLR to broaden the information contained in each model and thereby avoid over-fitting.

PLS regression is a bilinear calibration method (Wold et al., 2001), unlike MLR, which uses all available spectral wavebands. The PLS used firstly in a laboratory calibration such as chemometrics and near infrared spectroscopy (NIRS), then now increasingly employed in the field hyperspectral data analyses for wheat (Hansen and Schjoerring, 2003), paddy field (Nguyen and Lee, 2006, Takahashi et al., 2000) and heterogeneous grassland (Cho et al., 2007, Darvishzadeh et al., 2008). However, recent literatures indicated that waveband selection refine the predictive accuracy of the PLS model by optimizing important wavebands both in the laboratory NIRS (Jiang et al., 2002) and in the field hyperspectral measurements using iterative stepwise elimination (Kawamura et al., 2008) and genetic algorithm (GA) (Kawamura et al., 2010). Because GA has the ability to simulate a natural evolution of an individual, GA is well suited for solving variable subset selection problems (Ding et al., 1998).

In previous study (Kawamura et al., 2011), using multiple linear regression (MLR), we demonstrated the potential exist in a field hyperspectral measurements for estimating legume content in a mixed sown pasture at Hokkaido, Japan. In this study, we compared the performance of the GA-PLS with standard full-spectrum PLS (FS-PLS) and MLR models for predicting legume content using field hyperspectral data.

2. MATERIALS AND METHODS

2.1 Study site

The study was conducted in a 3.1 ha mixed sown pasture at the Hokkaido Agricultural Research Center (42°59' N, 141°24' E). This area is in a snowy, cold region with annual snowfall of 5 m. The annual mean temperature is 7.1°C and the annual precipitation is 960 mm. After over-seeding with perennial ryegrass (*Lolium perenne* L.) and fertilizer in 2002, the paddock had been used in a grazing trial without further fertilizer application. At the time of the measurements, the pasture was dominated by perennial ryegrass, orchardgrass (*Dactylis glomerata* L.) and white clover (*Trifolium repens* L.)(Watanabe et al., 2004). Ten Japanese Black cows (*Bos taurus* L.) and eight calves were stocked during the growing season from early May to late October in 2007.

2.2. Canopy reflectance measurement

Field hyperspectral measurements were made on 22-23 May and 24 July 2007 between 10:00-14:00 using a portable spectroradiometer (FieldSpec® Pro FR, ASD. Inc., USA), which measures spectral reflectance in the 350–2500 nm wavelength range. Canopy reflectance and forage biomass data were collected from 50 selected sites in each season. A spectralon (Labsphere, Inc., Sutton, NH, USA) reference panel (white reference) was used to optimize the ASD instrument prior to taking three canopy reflectance measurements at each plot. The sensor head with a 25° field-of-view (FOV) was held approximately 50 cm above the canopy at the nadir position, producing a view area with a 22 cm diameter at the canopy level. Spectral data in several wavelength regions were eliminated because of high levels of atmospheric noise (1341–1479, 1751–1999 and 2351–2500 nm); and low signal-noise ratio in the instrument (350–399 nm). Thus, a total of 1563 spectral bands between 400 nm and 2350 nm were used for analyses.

2.3. Pasture sampling and legume content

The forage samples were separated into white clover (WC), grasses, weeds and dead material (which include litter) and dried in a forced-air oven at 70ºC for 48 h to determine the total and individual aboveground component biomass (BM). Legume content was calculated using the dried green biomass (GBM) components only, using the following equation:

$$\text{Legume content} = \frac{\text{BM}_{\text{WC}}}{\text{BM}_{\text{WC}} + \text{BM}_{\text{grass}} + \text{BM}_{\text{weed}}} \times 100 \quad (1)$$

where BM_{WC}, BM_{grass} and BM_{weed} are BM of WC, grasses and weeds, respectively.

2.4. FS-PLS and GA-PLS regressions

All data handling and regression calculations were performed using PLS_Toolbox version 6.2 (Eigenvector Research Inc., Manson, WA, USA) in Matlab software version 7.12 (Mathworks Inc., Sherborn, MA, USA). The GA-PLS regression was analyzed using Leardi (2000). In the present study, five GA runs were performed because each GA-PLS gave a slightly different model. The parameters and their condition were taken from previous study (Kawamura et al., 2010). The predictive ability of the FS-PLS and GA-PLS were evaluated by the coefficient of determination (R^2) and the root mean squared error of cross validation (RMSECV) using a leave-one-out (LOO)

cross-validation, as similar to our previous studies (Kawamura et al., 2008, 2010).

To determine the significant wavelengths used in FS-PLS calibrations, the Variable Importance in the Projection (VIP) (Wold et al., 2001, Chong and Jun, 2005) was used. The VIP score gives a summary of the importance of an *x*-variable (waveband) for an observed *y*-variable (legume content), which was calculated as following equations,

$$\mathrm{VIP}_k(a) = K \sum_a w_{ak}^2 \left(\frac{\mathrm{SSY}_a}{\mathrm{SSY}_t} \right) \tag{2}$$

where $\mathrm{VIP}_k(a)$ is the importance of the *k*th predictor variables based on a model with a factors, w_{ak} is the corresponding loading weight of the *k*th variables in the *a*th PLS regression factor, SSY_a is the explained sum of squares of y by a PLS model with a factors, SSY_t is the total sum of squares of *y*, and *K* is the total number of predictor variables. A large VIP score, like PLS regression coefficient (*b*-coefficient), indicates an important *x*-variables (waveband) (Wold et al., 2001, Li et al., 2006).

3. RESULTS

3.1 Cross-validated calibration results

Cross-validated calibration results between canopy reflectance spectra and legume content using MLR, FS-PLS, and GA-PLS are shown in Table 1, with the selected number of wavebands (NW), the selected NW as a percentage of the full-spectrum (NW% = NW/whole waveband [n = 1563] × 100), and selected wavebands in our previous study using MLR (Kawamura et al., 2011). The NW (NW%) ranged between 80 and 188 (5.12–12.03%) in GA-PLS. The optimum NLV ranged between 5 and 9 in FS-PLS and GA-PLS, determined as the lowest RMSECV values calculated from LOO cross-validation, to avoid over-fitting of the model. Figure 1 shows the relationships between measured and cross-validated prediction values of legume content in each datasets using MLR, FS-PLS and GA-PLS models. Overall, the highest R^2 and the lowest RMSECV were obtained in the GA-PLS models for all datasets (R^2 = 0.62-0.86, RMSECV = 4.10-7.20%).

Table 1 Cross-validated optimum number of latent variables (NLV), coefficients o determination (R^2) and root mean squared errors of cross validation (RMSECV) from MLR (Kawamura et al., 2011), FS-PLS and GA-PLS models with selected number of wavebands (NW) and their percentage of the full-spectrum (1,563 bands)

	Regression	Cross-validation for whole samples in each data set					Selected wavebands in the MLR models‡				
Data set	method	NLV	R^2	RMSECV	NW	NW%	1	2	3	4	5
May	MLR		0.72	5.81	5	0.32	724	1324	2323	456	409
(n = 50)	FS-PLS	9	0.73	5.67	1563	100.00					
	GA-PLS	8	0.86	4.10	147	9.40					
July	MLR		0.39	8.47	5	0.32	400	408	2315	669	1131
(n = 50)	FS-PLS	5	0.30	9.14	1563	100.00					
	GA-PLS	7	0.72	5.73	80	5.12					
May+July	MLR		0.59	7.34	5	0.32	721	1132	707	2316	815
(n = 100)	FS-PLS	9	0.53	8.07	1563	100.00					
	GA-PLS	9	0.62	7.20	188	12.03					

NLV, number of latent variables (or PLS factors); RMSECV, root mean squared error of cross validation; NW, number of wavebands; NW%, number of wavebands percentage of all available bands.
‡Selected wavebands in the GA-PLS models were shown in figure 2.

3.2. Selected waveband regions from GA-PLS models

Selected wavelength regions from five GA-PLS (as for the lowest RMSECV) to estimate legume content are shown in Figure 2. GA-PLS selected a wider range of spectral wavelength regions from within the visible (400-700 nm), near- (NIR, 700-1300 nm) and shortwave-infrared (SWIR, 1300-2350 nm) spectra, with slightly different regions in each data set. In May data set, selected five wavebands in MLR (409, 456, 724, 1324, 2323 nm) were included in GA-PLS model. The other hands, in July and combined data sets, slightly different from MLR were selected in GA-PLS models. The wavelength regions selected for legume content prediction were generally different among the data sets. However, wavelengths in the red-edge (700-760 nm) and 2320-2325 nm regions were commonly selected in GA-PLS models.

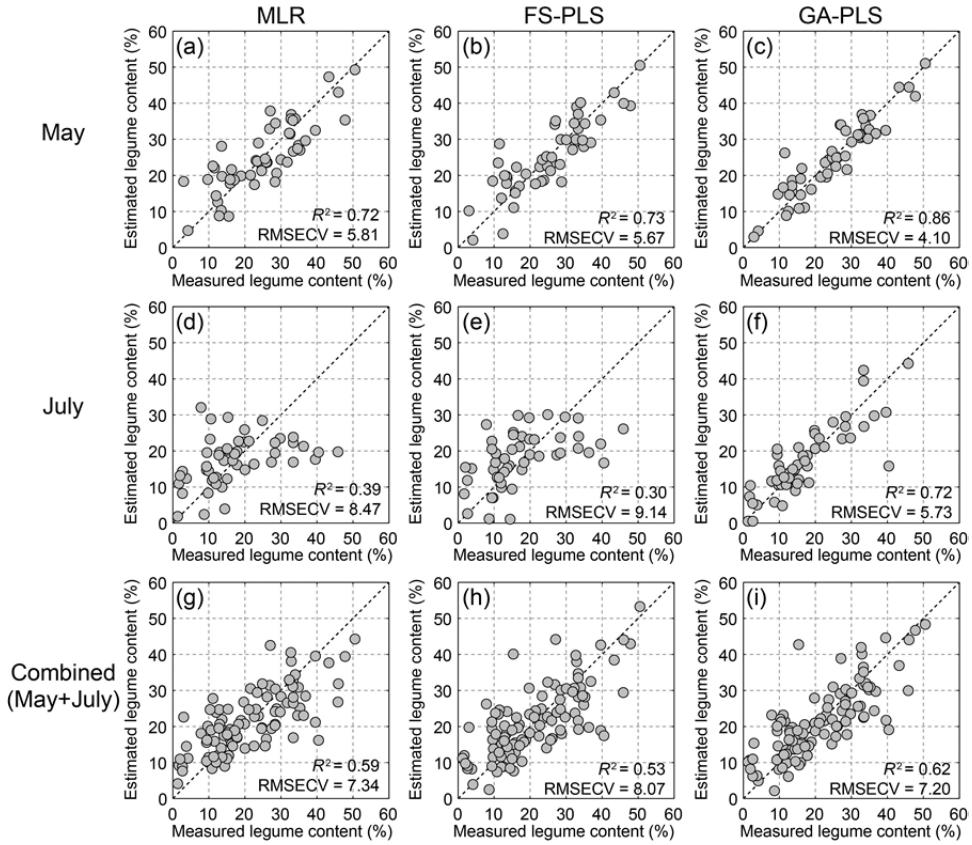

Figure 1. Comparison of laboratory-measured legume content in May (n = 50), July (n = 50), and the combined data set (n = 100), along with their estimated values using MLR (left), FS-PLS (center), and GA-PLS (right) regressions.

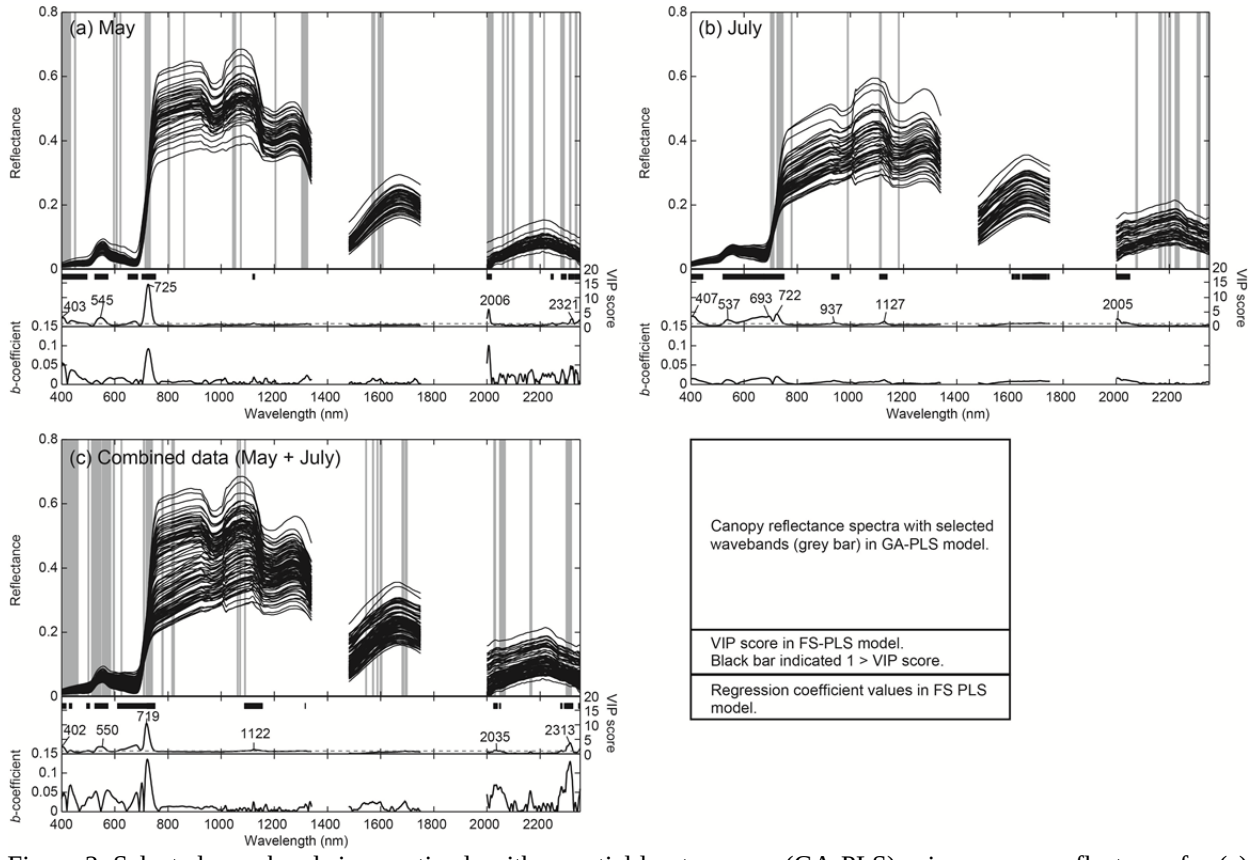

Figure 2. Selected wavebands in genetic algorithm partial least squares (GA-PLS) using canopy reflectance for (a) May (n = 50), (b) July (n = 50), and (c) combined data (May + July, n = 100) to estimate legume content, with VIP score (black bar; 1 > VIP score) and b-coefficient values in standard full-spectrum PLS (FS-PLS) model.

4. DISCUSSION

Our results indicate that legume content in grass-legume mixtures can be rapidly and nondestructively predicted from field hyperspectral measurement, and that the predictive accuracy of the PLS method was improved by selecting a subset of wavelengths related to pasture parameters and removing unrelated wavebands (Table 1). Particularly, GA based waveband selection greatly improved in July dataset (R^2 = 0.30-0.72). The NW (NW%) remaining after waveband selection ranged 80-188 (5.12–12.03%) in GA-PLS models, suggesting that over 88% of the waveband information from the canopy reflectance spectrum was redundant and did not contribute to or disturb the prediction. These results in agreements previous results that the most useful information for predicting forage parameters was provided by only less than 20% of the wavelength region (400-2350 nm) (Kawamura et al., 2008, 2010). These findings support previous results showing that the performance of PLS models can be improved through wavelength selection (Darvishzadeh et al., 2008, Cho et al., 2007, Kawamura et al., 2008). Moreover, spectral data efficiency would be improved by optimization of a wavelength subset in the PLS model.

The wavebands selected in GA-PLS models do not match wavelengths identified by previous study using MLR (Kawamura et al., 2011), but in most cases they are found within 20 nm of previously known wavebands. In the visible region, chlrophyll absorption related blue region (400-460 nm) was selected. Red-edge region (700-760 nm), which is strongly related to chlorophyll concentration and has been used to estimate N status of crop and grasses (Horler et al., 1983, Lamb et al., 2002), and 2320-2325 nm region were commonly selected in all data sets. They are considered to be potentially important spectral bands for predicting legume content using field hyperspectral data.

It is presumed that the predictive accuracy of remote sensing could not be better than that of laboratory NIRS methods because both plant materials and measurement configurations are far better controlled in NIRS. Nevertheless, remote sensing can provide quantitative information on legume content at geospatial and near real-time basis over wide areas, so that it has a great potential for efficient, productive and environment-friendly pasture management.

5. CONCLUSIONS

We investigated the performance of the GA-PLS in spectral assessment of legume content. Results indicated that legume content in grass-legume mixtures can be more accurately estimated by GA-PLS than MLR and FS-PLS using *in situ* hyperspectral reflectance data. GA based waveband selection in PLS calibration suggested that the important wavebands for estimating legume content are 5-12% of all 1,563 wavebands over the 400-2350 nm range. Using selected wavebands in the GA-PLS model, legume content were estimated with 62-86% accuracy using ground based canopy reflectance data. The use of PLS with GA waveband selection is promising for spectral assessment of legume content in grass-legume mixtures. Further, the waveband selection procedure refined the predictive ability, expected by optimization of the wavelength subset using GA-PLS. Such timely and accurate legume content prediction could provide some useful insights for the optimization of legume content management.

ACKNOWLEDGEMENT

We are grateful to Ms. Shizue Nakashima of the National Agriculture and Food Research Organization (NARO) Hokkaido Agricultural Research Center, Japan, for her assistance in field experiments. Funding from Research Fellowships of the Japan Society for the Promotion of Science (JSPS) for Young Scientists (No. 18/06934) supported this work.

REFERENCES

Berardo, N., 1997. Prediction of the chemical composition of white clover by near-infrared reflectance spectroscopy. Grass and Forage Science, 52, pp. 27-32.

Cho, M.A., Skidmore, A., Corsi, F., Van Wieren, S.E. and Sobhan, I., 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. International Journal of Applied Earth Observation and Geoinformation, 9, pp. 414-424.

Chong, I.G. and Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity i present. Chemometrics and Intelligent Laboratory Systems, 78, pp. 103-112.

Curran, J.P., Dungan, L.J., Macler, A.B., Plummer, E.S. and Peterson, L.D., 1992. Reflectance spectroscopy of

fresh whole leaves for the estimation of chemical concentration. Remote Sensing of Environment, 39, pp. 153-166.

Curran, P.J., 1989. Remote sensing of foliar chemistry. Remote Sensing of Environment, 30, pp. 271-278.

Darvishzadeh, R., Skidmore, A., Schlerf, M., Atzberger, C., Corsi, F. and Cho, M., 2008. LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements. ISPRS Journal of Photogrammetry and Remote Sensing, 63, pp. 409-426.

Ding, Q., Small, G.W. and Arnold, M.A., 1998. Genetic Algorithm-Based Wavelength Selection for the Near-Infrared Determination of Glucose in Biological Matrixes: Initialization Strategies and Effects of Spectral Resolution. Analytical Chemistry, 70, pp. 4472-4479.

Ferwerda, J.G., Skidmore, A.K. and Stein, A., 2006. A bootstrap procedure to select hyperspectral wavebands related to tannin content. International Journal of Remote Sensing, 27, pp. 1413-1424.

Hansen, P.M. and Schjoerring, J.K., 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sensing of Environment, 86, pp. 542-553.

Horler, H.N.D., Dockray, M. and Barber, J., 1983. The red edge of plant leaf reflectance. International Journal of Remote Sensing, 4, pp. 273-288.

Jiang, J.H., Berry, R.J., Siesler, H.W. and Ozaki, Y., 2002. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. Analytical Chemistry, 74, pp. 3555-3565.

Kawamura, K., Watanabe, N., Sakanoue, S. and Inoue, Y. 2008. Estimating forage biomass and quality in a mixed sown pasture based on PLS regression with waveband selection. Grassland Science, 54, pp. 131-146.

Kawamura, K., Watanabe, N., Sakanoue, S., Lee, H.J. and Inoue, Y., 2011. Waveband selection using a phased regression with a bootstrap procedure for estimating legume content in a mixed sown pasture. Grassland Science, 57, pp. 81-93.

Kawamura, K., Watanabe, N., Sakanoue, S., Lee, H.J., Inoue, Y. and Odagawa, S., 2010. Testing genetic algorithm as a tool to select relevant wavebands from field hyperspectral data for estimating pasture mass and quality in a mixed sown pasture using partial least squares regression. Grassland Science, 56, pp. 205-216.

Lamb, W.D., Steyn-Ross, M., Schaare, P., Hanna, M.M., Silvester, W. and Steyn-Ross, A., 2002. Estimating leaf nitrogen concentration in ryegrass (*Lolium* spp.) pasture using the chlorophyll red-edge: theoretical modelling and experimental observations. International Journal of Remote Sensing, 23, pp. 3619-3648.

Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. Journal of Chemometrics, 14, pp. 643-655.

Li, B., Liew, O.W. and Asundi, A.K., 2006. Pre-visual detection of iron and phosphorus deficiency by transformed reflectance spectra. Journal of Photochemistry and Photobiology B: Biology, 85, pp. 131-139.

Locher, F., Heuwinkel, H., Gutser, R. and Schmidhalter, U., 2005. Development of near Infrared reflectance spectroscopy calibrations to estimate legume content of multispecies legume-grass mixtures. Agronomy Journal, 97, pp. 11-17.

Nguyen, H.T., Lee, B.W., 2006. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. European Journal of Agronomy, 24, pp. 349.

Norris, K.H., Barnes, R.F., Moore, J.E. and Shenk, J.S., 1976. Predicting forage quality by infrared reflectance spectroscopy. Journal of Animal Science, 43, pp. 889-897.

Takahashi, W., Nguyen-Cong, V., Kawaguchi, S., Minamiyama, M. and Ninomiya, S. 2000. Statistical models for prediction of dry weight and nitrogen accumulation based on visible and near-infrared hyper-spectral reflectance of rice canopies. Plant Production Science, 3, pp. 377-386.

Watanabe, N., Umemura, K. and Sudo, K., 2004 Combination use of bite counter and GPS for monitoring foraging behaviour in grazing cow. Grassland Science, 50 (Ext.), pp. 404-405.

Wold, S., Sjöström, M. and Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58, pp. 109-130.

Zhao, D., Starks, P.J., Brown, M.A., Phillips, W.A. and Coleman, S.W., 2007. Assessment of forage biomass and quality parameters of bermudagrass using proximal sensing of pasture canopy reflectance. Grassland Science, 53, pp. 39-49.