SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

March 2011 Vol.54 No.3: 272–281
doi: 10.1007/s11427-011-4135-4

Estimating biophysical parameters of rice with remote sensing data using support vector machines

YANG XiaoHua[1,2*], HUANG JingFeng[1], WU YaoPing[2], WANG JianWen[2], WANG Pei[3], WANG XiaoMing[2] & Alfredo R. HUETE[4]

[1]*Institute of Remote Sensing & Information Application, Zhejiang University, Hangzhou 310029, China;*
[2]*Meteorological, Hydrographic, Spatial & Synoptic Central Station of General Staff Headquarters, Beijing 100081, China;*
[3]*State Key Laboratory of Earth Surface Processes & Resource Ecology, Beijing Normal University, Beijing 100875, China;*
[4]*Department of Soil, Water, and Environmental Science, University of Arizona, Tucson, AZ 85721, USA*

Received November 22, 2008; accepted September 8, 2009

Hyperspectral reflectance (350–2500 nm) measurements were made over two experimental rice fields containing two cultivars treated with three levels of nitrogen application. Four different transformations of the reflectance data were analyzed for their capability to predict rice biophysical parameters, comprising leaf area index (LAI; m^2 green leaf area m^{-2} soil) and green leaf chlorophyll density (GLCD; mg chlorophyll m^{-2} soil), using stepwise multiple regression (SMR) models and support vector machines (SVMs). Four transformations of the rice canopy data were made, comprising reflectances (*R*), first-order derivative reflectances (*D*1), second-order derivative reflectances (*D*2), and logarithm transformation of reflectances (*LOG*). The polynomial kernel (POLY) of the SVM using *R* was the best model to predict rice LAI, with a root mean square error (RMSE) of 1.0496 LAI units. The analysis of variance kernel of SVM using *LOG* was the best model to predict rice GLCD, with an RMSE of 523.0741 mg m^{-2}. The SVM approach was not only superior to SMR models for predicting the rice biophysical parameters, but also provided a useful exploratory and predictive tool for analyzing different transformations of reflectance data.

biophysical parameters; support vector machines; remote sensing

Citation: Yang X H, Huang J F, Wu Y P, *et al*. Estimating biophysical parameters of rice with remote sensing data using support vector machines. Sci China Life Sci, 2011, 54: 272−281, doi: 10.1007/s11427-011-4135-4

The assessment of biophysical vegetation properties, such as leaf area index (LAI) and green leaf chlorophyll density (GLCD), is a major goal of remote sensing in agriculture. Remote-sensing-based assessments of these variables are made possible as a result of the strong contrast between spectral reflectances of vegetation and the soil background and the dramatic reflectance changes associated with changing vegetative cover. Based on this contrast, numerous vegetation indices (VIs) have been developed during the past few decades, which are highly correlated with the amount of vegetation. The most common of these indices use the red and near-infrared (NIR) canopy reflectances in the form of ratios, such as the ratio VI [1] and the normalized difference vegetation index [2], and as linear combinations of red and NIR reflectances [3,4]. These indices generally use averaged spectral information over broad bandwidths [5], resulting in the loss of critical information available in specific narrow bands [6], and potentially limiting the accurate estimates of agricultural crop and natural vegetation biophysical and biochemical variables [7,8]. In addition, many of these vegetation indices are strongly influenced by the soil background, resulting in soil-dependent VI-biophysical relationships [9,10].

Further improvements in quantifying vegetation are possible using spectral data from distinct narrow bands, as indicated by numerous hyperspectral studies using field spec-

*Corresponding author (email: dr.xiaohuayang@gmail.com)

troradiometers [11–14]. These studies have shown narrow-band data to provide additional information over broadband data, enabling significant improvements in quantifying biophysical and biochemical variables of agricultural crops. A number of investigators have studied the relationship between canopy hyperspectral reflectance and canopy properties for major crops [15–17]. Hyperspectral studies have been successfully used in assessments of rice yield [13] and chlorophyll content of plants [6,7].

Among spectroscopic techniques, derivative analysis of reflectances is particularly promising for use with remote sensing data. Second-order derivatives and higher-order derivatives are relatively insensitive to variations in illumination intensity caused by changes in sun angle, cloud cover, or topography. Nonetheless, relatively few researchers have addressed applications of spectral derivatives in remote sensing [18,19]. Although some of these studies have used higher-order derivatives [20], first-order and second-order derivatives are most commonly used.

More recently, Filella and Peñuelas [21] and Mauser and Bach [22] concluded that derivative spectral indices are very sensitive to LAI and GLCD. Yoder and Pettigrew-Crosby [23] found first-order derivative spectra were the best predictors of nitrogen and chlorophyll for big-leaf maple grown under different fertilization treatments. The log(1/*R*) (also called pseudoabsorbance) is often used because it provides a curve similar to an absorption curve, with peaks occurring at the corresponding absorption wavelengths. Johnson and Billow [24] examined Douglas fir needles that were grown under various fertilization treatments and also found that the first-order derivative and log(1/*R*) of the fresh leaf spectra were strongly correlated with total nitrogen concentration.

The relationships found between biophysical parameters and specific narrow spectral bands have promoted the development of models to estimate biophysical parameters both at the leaf and canopy scales [25–28]. Most of the models, thus far, have been developed using regression analyses and assumptions of linearity between the dependent and independent variables. Some researchers have shown that stepwise multiple regression (SMR) models performed on discrete narrow bands provide flexibility in choosing the bands that provide maximum information at a given period of crop growth [29,30]. As crop conditions vary due to factors such as management conditions, soil characteristics, climatic conditions, and cultural practices, different band combinations can be used [13,30]. Recent studies have also successfully used SMR to select the optimal wavelengths that correlate best with leaf biochemistry [31].

Although linearity in a dataset may be achieved through mathematical transformations, data with complex non-linear properties are difficult or may never be approximated. In such situations, the use of non-linear regression requires a priori knowledge of the nature of the non-linear behavior, something that is not usually known and, furthermore, non-linear regressions are cumbersome to implement [32]. Support vector machines (SVMs) were first introduced for classification and non-linear function estimations [33,34]. A SVM for regression analysis is accomplished by solving a convex optimization problem, more specifically a quadratic programming (QP) problem. This is obtained by employing Vapnik's c-insensitive loss function [34], solving the approximation problem as an inequality constrained optimization problem, and exploiting the Mercer condition to relate the non-linear feature space mapping to the chosen kernel function. As for Mercer's condition, any kernel which can be expressed as $K(x,y)=\sum_{P=0}^{\infty} c_p (x\cdot y)^p$, where c_p are positive real coefficients and the series is uniformly convergent, satisfies Mercer's condition, a fact noted previously [35].

Moreover, the model complexity follows from solving this convex optimization problem. SVM models also scale to higher-dimensional input spaces very well [36]. However, less is known about the application of SVMs to estimate biophysical parameters using remote sensing data.

In this study, reflectances (*R*), first-order derivative reflectances (*D*1), second-order derivative reflectances (*D*2) and logarithm transformation of reflectances (log(1/*R*)) were selected as independent variables with field-measured LAI and GLCD as dependent variables to apply and test SMR and SVM prediction capabilities for these two variables. The main objective of this study was to assess and compare the predictive ability of the SVM models in estimating LAI and GLCD in rice with that of the more traditional SMR models.

1 Study area and methods

1.1 Study area

The study area was located at the Zhejiang University experimental field, Hangzhou, Zhejiang Province, China, located at 120°10′05″E, 30°14′03″N. To acquire a large dynamic range in rice LAI and GLCD, two experiments were designed in 2004 with different rates of nitrogen fertilization and two rice cultivars. The first experiment began 15 days earlier than the second experiment. Winter wheat and rice were included in the crop rotations of the two experimental fields with straw residue removed from the fields between plantings. The study area is characterized by a monsoon climate with a hot summer and a cool winter. The average annual rainfall is 1374.7 mm and the average annual temperature is 17.8°C. The soil is a sandy loam paddy soil with pH 5.7, organic matter content 16.5 g kg^{-1} and total N content of 1.02 g kg^{-1}.

1.2 Field experimental design

The experimental field was divided into 48 subplots of size

4.6 m×5.46 m. Half of the plots were used for the first experiment and the remaining plots were used for the second experiment. Each experiment involved four replicates of two rice cultivars ('Xiushui 110' and 'Xieyou 9308'), three nitrogen levels (0, 120, and 240 kg N hm^{-2}), and with a plant density of 45 plants m^{-2}. The first experiment was seeded on 30 May 2004 and the second experiment was seeded on 15 June 2004. Both sets of seedlings were transplanted to the field one month later. Nitrogen treatment levels included no nitrogen fertilizer, a normal application, and a superabundant dose of urea, applied as 45% base fertilizer, 35% tillering fertilizer, and 20% heading fertilizer. In addition, 225 kg hm^{-2} $Ca(H_2PO_4)_2$ was applied as a base fertilizer with 150 kg hm^{-2} KCl as a tillering fertilizer and 150 kg hm^{-2} KCl as a heading fertilizer.

1.3 Spectral measurements

Field canopy reflectance measurements were made with a full spectral range (350–2500 nm) Analytical Spectral Devices™ spectroradiometer. The spectroradiometer has a spectral resolution of 3 nm between 400 and 1000 nm, and approximately 10 nm between 1000 and 2500 nm. Due to severe noise in the water absorption spectrum from 1330 to 1480 nm and from 1780 to 1990 nm, only data from 350 to 1330 nm, 1480 to 1780 nm, and 1990 to 2300 nm were used in this study. The measurements of rice spectra were performed between 10:00 and 14:00 local time (GMT+8) on 20 July, 8 August, 28 August, 22 September, 5 October and 27 October, 2004 for the first experiment, and on 8 August, 28 August, 22 September, 5 October and 27 October, 2004 for the second experiment.

All rice fields were in flooded condition except on 27 October. At each plot, 10 reflectance measurements were consistently taken, with a nadir view from a height of 1 m above the canopy, using a 25° field of view lens. The measurement sites were selected randomly at each plot. The spectroradiometer data over rice were analyzed using PORTSPEC™ and VNIR™ software, supplied by the manufacturer of the instrument (Analytical Spectral Devices™), and SPSS version 11.5. The target reflectance was computed as the ratio of energy reflected off the rice canopy to the energy incident on a $BaSO_4$ white reference plate. Dark current values varied slightly with ambient temperature and were recorded for each integration time. The solar zenith angle was less than 45° for all measurements and no disturbing clouds were observed. Reflectances were then computed:

$$\text{Reflectance (\%)}=\frac{(\text{target}-\text{dark current})}{(\text{reference}-\text{dark current})}. \quad (1)$$

Figure 1 shows the reflectance curves of a few typical rice canopies.

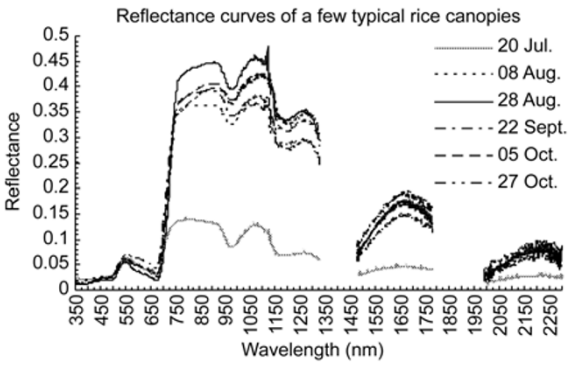

Figure 1 Reflectance curves of a few typical rice canopies.

1.4 Plant sampling and harvest procedure

Samples of green leaves were collected throughout the vegetative growth stages from early stem elongation until heading, coinciding with the same dates as the canopy reflectance measurements. A representative area of 0.088 m^2 (one hill of plants under nadir view) was cut and brought to the laboratory for measurement of rice biophysical parameters, comprising leaf length (LL; cm), leaf width (LW; cm), and green leaf fresh weight (GLFW) per square meter (g m^{-2}).

From these measurements, LAI (cm^2 cm^{-2}) was calculated:

$$LAI=\frac{\sum LL\times LW\times 0.83}{0.088\times 10000}, \quad (2)$$

where 0.83 is the rice leaf calibration coefficient.

Leaves and stems were separated by excision at the leaf base. One leaf was randomly selected among the youngest fully developed leaves for organic extraction of leaf chlorophyll. For leaf chlorophyll analysis, the leaf samples were chipped, weighed and then dipped in a 20 mL mixed solution of acetone, ethanol, and distilled water (4.5∶4.5∶1 proportions, respectively), for 24 h. The concentrations (mg L^{-1}) of Chl a, Chl b and total chlorophyll (Chl_t=Chl a+Chl b) in the extract were calculated using eqs. (3)–(5) and the contents (mg g^{-1}) of chlorophyll were calculated using eq. (6):

$$\text{Chl a (mg L}^{-1})=12.7A_{663}-2.69A_{645}, \quad (3)$$

$$\text{Chl b (mg L}^{-1})=22.9A_{645}-4.68A_{663}, \quad (4)$$

$$\text{Chl}_t\ \text{(mg L}^{-1})=8.02A_{663}+20.2A_{645}, \quad (5)$$

$$\text{GLCC}=Pc\times V\times 1000/Ma, \quad (6)$$

where A is the optical density, GLCC is the green leaf chlorophyll content (mg g^{-1}), Pc is the pigment concentration (mg L^{-1}), V is volume (mL) of the extracting solution, and Ma is the mass (g) of the sample.

The GLCD (mg chl_t m^{-2} soil) was calculated with eq. (7):

$$\mathrm{GLCD(mg\ m^{-2})} = \frac{GLCC \times GLFW}{0.088}. \quad (7)$$

1.5 Stepwise multiple regression model

Stepwise multivariable regression (SMR) models are most commonly used to predict crop biophysical variables in plants [37,38]. Using SPSS version 11.5, SMR selectively and stepwisely includes those most significant variables (various narrow bands in the forms of *R*, *D*1, *D*2 or *LOG*) into the model of dependents (LAI or GLCD). For LAI and GLCD, the first four narrow band variables explained 75% or above variability (Figure 2). This is nearly the same percentage of variability explained when the ratio of the number of independent variables or number of bands (*M*) to that of the total number of field samples (*N*, 100 in this case) for that variable is between 0.04 and 0.05 in different rice variables based on different spectral transformations. As *M* approaches *N*, the coefficient of determination value approaches 1. When *M*/*N* is higher than 0.05 (Figure 2), there were only small increases (often statistically insignificant) with the addition of another variable. Therefore, four-variable models were considered the best in this study.

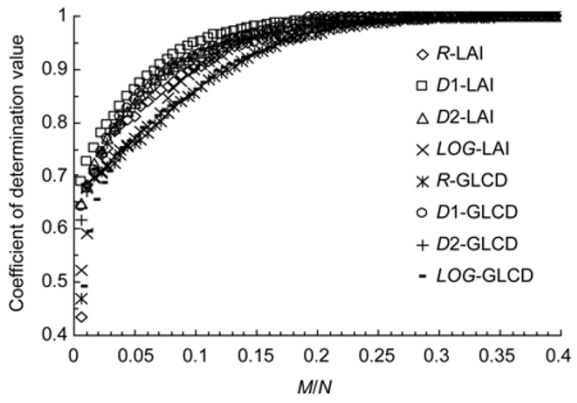

Figure 2 Plot of the ratio *M*/*N* versus the coefficient of determination value.

Various transformations of reflectance in 1590 discrete narrow bands are independent variables and biophysical parameters are dependent variables. The coefficients and bands of the regression equations are shown in Table 1.

1.6 Support vector machines

Support vector machines (SVMs) are based on the structural risk minimization principle from computational learning theory [34]. The idea of structural risk minimization is to find a hypothesis hyperplane to represent the data and for which the lowest true error of the data can be guaranteed. The true error is the probability that the hyperplane will make an error classification on an unseen and randomly selected test sample in the process of building the model. An upper bound can be used to relate the true error of a hypothesis hyperplane with the training data set error of the hyperplane. The complexity of the hypothesis space containing the hyperplane is measured by the Vapnik-Chervonenkis (VC) dimension that depicts the capacity of a hypothesis space. Capacity is a measure of complexity or expressive power, richness or flexibility of a set of functions [34]. SVMs find the hypothesis hyperplane which, approximately, minimizes this bound on the true error by effectively and efficiently controlling the VC-dimension of hypothesis space.

SVMs are universal learners, and in their basic form learn linear threshold functions. Nevertheless, by a simple 'plug-in' of an appropriate kernel function, they can be used to learn polynomial classifiers, radial basic function (RBF) networks, and three-layer sigmoid neural nets, for example. One remarkable property of SVMs is that their ability to learn is largely independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses space, based on the margin with which they separate the data, not the number of features. This means that we can generalize, even in the presence of many features, if our data is separable with a sufficiently wide margin using functions from the hypothesis space. The same margin ar-

Table 1 Stepwise multiple regression equations built in this study

Variable		Best four-variable models									
		C[g)]	Constant	Beta1[h)]	R_{Band1}	Beta2	R_{Band2}	Beta3	R_{Band3}	Beta4	R_{Band4}
LAI[a)]	*R*[c)]	0.738	1.966	31.729	R_{1124}[i)]	–17.073	R_{731}	–260.216	R_{1720}	221.385	R_{1710}
	*D*1[d)]	0.781	0.823	7098.553	R_{768}[j)]	831.342	R_{1757}	964.859	R_{1087}	1745.257	R_{1282}
	*D*2[e)]	0.746	2.431	17122.780	R_{718}[k)]	–5805.780	R_{1625}	2921.679	R_{2131}	–649.927	R_{1132}
	LOG[f)]	0.705	–0.507	5.386	R_{353}[l)]	–32.227	R_{807}	30.974	R_{751}	–4.810	R_{370}
GLCD[b)]	*R*	0.708	95.714	9140.299	R_{1125}	–49560.600	R_{2261}	14152.200	R_{2007}	14899.680	R_{2285}
	*D*1	0.745	–172.433	1910057.000	R_{768}	–447241.000	R_{1068}	1415155.000	R_{1620}	570679.000	R_{1625}
	*D*2	0.741	392.873	6159555.000	R_{718}	–515438.000	R_{2265}	-2.9×10^7	R_{861}	-3.6×10^7	R_{452}
	LOG	0.686	–176.370	3834.464	R_{354}	–8972.510	R_{807}	8309.175	R_{751}	–3531.270	R_{370}

a) Leaf area index; b) green leaf chlorophyll density; c) the reflectances; d) first derivative of reflectances; e) second derivative of reflectances; f) logarithm-transformed reflectances; g) determination coefficient of the SMR models; h) constants of the SMR models; i) reflectances of band 1124; j) first derivative of reflectances at band 768; k) second derivative of reflectances at band 718; l) logarithm-transformed reflectances at band 353.

guments also suggest a heuristic for selecting good parameter settings for the learner [34]. Jaakkola's heuristic uses the median separation of negative points to their nearest positive neighbor. The best parameter setting is the one that produces the hypothesis with the lowest VC-dimension. This allows fully automatic parameter tuning without expensive cross-validation.

In this study, three different kernel functions of SVMs, comprising a polynomial kernel (POLY), an RBF kernel, and an ANOVA kernel were employed. The POLY kernel was defined by eq. (8):

$$k(x,\ y) = (x \cdot y + 1)^d, \tag{8}$$

the RBF kernel was defined by eq. (9):

$$k(x, y) = \exp(-\gamma \parallel x - y \parallel^2), \tag{9}$$

and the ANOVA kernel was defined by eq. (10):

$$k(x, y) = \left(\sum_i \exp(-\gamma(x_i - y_i)) \right)^d. \tag{10}$$

In order to compare the predictive power of SVMs and SMR, the SVMs were developed and tested with the same data sets as those for the corresponding SMR. The four reflectance transformations of the four bands that were previously selected by the SMR analysis were used as net inputs to the SVM, with the biophysical parameters as net outputs in the SVM. Because the results from the basic SVMs with POLY kernel using different reflectance transformations for GLCD were too poor to accept (not all SVM kernel algorithms are appropriate for the same data set, as long as the best one can be selected), it was removed until later. The specific structures and training parameters of the SVMs used in this study are listed in Table 2.

1.7 Measurement of model performance

In most studies model performance analysis is usually conducted through comparisons of the correlation coefficient (r), average absolute error (ABSE), and root mean squared error (RMSE) between the predicted sets and the corresponding observed sets [39,40]. The calculated statistic r is defined as the proportion of variance of the response that can be explained by the regressing variable(s). However, the r statistic can be misleading when comparing results of experiments on the same variable but with different ranges [41], such as in the experiment reported herein. In such cases, one should calculate ABSE and RMSE rather than r, and for these reasons we decided to base the analysis of our results on RMSE and ABSE, defined as follows:

$$RMSE = \sqrt{\frac{\sum (P - P')^2}{N - 1}}, \tag{11}$$

$$ABSE = \frac{\sum (ABS(P - P'))}{N}, \tag{12}$$

Table 2 Specific structures and training parameters of different SVM models built in this study

Variables	Spectral transformations	Kernel[g)]	C[h)]	Epsilon[i)]	nu[j)]	γ[k)]	d[l)]
LAI[a)]	R[c)]	POLY[m)]	1	0.1	–	–	3
		RBF[n)]	1000	0.8	–	0.01	–
		ANOVA[o)]	1	0.0001	–	0.3	1
	$D1$[d)]	POLY	0.1	1.5	–	–	1
		RBF	1	0.1	–	0.1	–
		ANOVA	0.1	0.01	–	0.2	2
	$D2$[e)]	POLY	0.01	0.01	–	–	1
		RBF	0.1	0.00001	–	0.02	–
		ANOVA	100000	0.001	–	0.1	1
	LOG[f)]	POLY	100	–	0.8	–	1
		RBF	100	0.01	–	1.2	–
		ANOVA	100	–	0.5	0.5	1
GLCD[b)]	R	RBF	100	–	0.6	0.8	–
		ANOVA	0.01	0.001	–	0.1	4
	$D1$	RBF	1	0.0001	–	0.3	–
		ANOVA	0.1	0.01	–	0.7	1
	$D2$	RBF	1	0.01	–	0.1	–
		ANOVA	0.1	0.001	–	0.4	1
	LOG	RBF	1000	0.001	–	0.05	–
		ANOVA	1000	0.001	–	0.3	1

a) Leaf area index; b) green leaf chlorophyll density; c) reflectances; d) first derivative of reflectances; e) second-derivative of reflectances; f) logarithm-transformed reflectances; g) different kernel functions; h) SVM capacity parameters; i) and j) different SVM algorithms; k) and l) SVM kernel parameters; m) POLY kernel SVM; n) RBF kernel SVM; o) ANOVA kernel SVM.

where P' is the simulated value of LAI and GLCD of rice, P is the measured value of LAI and GLCD of rice, and N is the number of test samples.

2 Results

The 182 samples from the two experiments in our study were combined into one dataset, and then randomly divided into two subsets. The first subset (100 samples) was used to construct the models and the second subset (82 samples) was used to measure the effectiveness of the models.

2.1 Basic statistical properties of the measured data

The experimental treatments, comprising two rice cultivars and three nitrogen application strategies together with the temporal timing of plant sampling, resulted in a wide dynamic range of variation in the investigated crop variables. There was an almost 20-fold variation in LAI and 100-fold variation in GLCD (Table 3). The wide range in the investigated crop variables was planned in order to make the relationship between plant performance and reflectance measurements as realistic and universal as possible.

2.2 Training results of SVMs and SMR models for LAI

The transformed data from the four discrete narrow-band reflectances and LAI were trained by SVM and SMR models. The assessment results are presented in Figure 3 and Table 4.

From Figure 3, it can be seen that the predicted LAI values from the SVM-POLY model using *R* (Figure 3B) increases with increasing amounts of measured LAI, with data points more tightly located along the line $y=x$ than those from the SMR model (Figure 3A) or from the other two SVM models (Figure 3C and D). At measured LAI values greater than 5, the predicted LAI in the RBF model with *D*1 (Figure 3G) was slightly underestimated, but resulted in the lowest RMSE (1.2124, Table 4) among the other prediction models. The SVM-RBF model using *D*2 (Figure 3K) also underestimated LAI for measured LAI values greater than 5, while the other three models resulted in predicted LAI values (Figure 3I–L) more centralized along the line $y=x$. However, the SVM-ANOVA model using *D*2 (Figure 3L) performed the best in the LAI prediction of all the models and had the lowest RMSE (1.0858), lowest ABSE (0.8676), and the highest r (0.7941) (Table 4). All predictive LAI relationships in the four models using *LOG* were well centralized along the line $y=x$ (Figure 3M–P), with the SVM-POLY model the best for predicting LAI, based on the RMSE (1.1256), ABSE (0.8980), and r (0.7637) values (Table 4). This suggests that SVMs have a stronger potential to take into account non-linear characteristics in predicting LAI compared with SMR.

Comparing the RMSE between SVMs and SMR for predicting LAI (Table 4), the SVM-POLY model had the lowest RMSE (1.0496) compared with the other models when using *R*. In the case of *D*1, the SVM-RBF model had the lowest RMSE (1.2124) compared with the other models. With *D*2, the SVM-ANOVA had the lowest RMSE (1.0858), while the use of LOG resulted in the SVM-POLY model having the lowest RMSE (1.1256). From the ABSE results (Table 4), there was also much improvement in the relationship between predicted LAI and measured LAI, with the SVM-POLY model using *R* providing the lowest ABSE (0.8051) relative to the other models using *R*. The SVM-ANOVA model using *D*1 had the lowest ABSE (0.9194) compared with the other models using *D*1. The SVM-ANOVA model using *D*2 also showed the lowest ABSE (0.8676), and the SVM-POLY model using *LOG* had the lowest ABSE (0.8980). All results showed that SVMs have high potential for learning with overall good robustness, indicating that SVMs could further improve the relationships and accuracies between various transformations of reflectance and LAI.

Table 3 Selected properties of the investigated rice leaf area index (LAI) and green leaf chlorophyll density (GLCD)

Variable	Mean	Minimum[c)]	Maximum[d)]	Range[e)]
LAI[a)] (m^2 m^{-2})	3.5859	0.4652	8.2703	7.8051
GLCD[b)] (mg m^{-2})	1162.4108	31.8338	2970.0627	2938.2289

a) Leaf area index; b) green leaf chlorophyll density; c) minimum value in the preprocessed data set; d) maximum value in the preprocessed data set; e) difference between the maximum and minimum values.

2.3 Training results of SVM and SMR models for GLCD

The four transformed data sets using the four discrete narrow bands and GLCD were trained by SVMs and SMR, and the results of the regression and error analyses are presented in Figure 4 and summarized in Table 5.

The predictions of GLCD with the SVM-ANOVA model using *R* (Figure 4C) were well centralized along the line $y=x$, compared with results from the SMR and SVM-RBF models (Figure 4A and B). When measured GLCD values exceeded 1800 mg m^{-2}, the prediction of GLCD from the SVM-ANOVA model using *D*1 was underestimated (Figure 4F), and the predicted GLCD values from the SVM-RBF model using *D*1 (Figure 4E) were closer to the line $y=x$ than those from the SMR model (Figure 4D), hence the SVM-RBF model with *D*1 had the lowest RMSE (575.3 mg m^{-2}, Table 5).

There was also a trend for underestimation of GLCD values in the SVM-RBF and SVM-ANOVA model results using *D*2 (Figure 4H and I), with the SVM-RBF model having the lowest RMSE (553.2974 mg m^{-2}, Table 5). The points of the predicted GLCD values from the SVM-ANOVA model using *LOG* (Figure 4L) were well central-

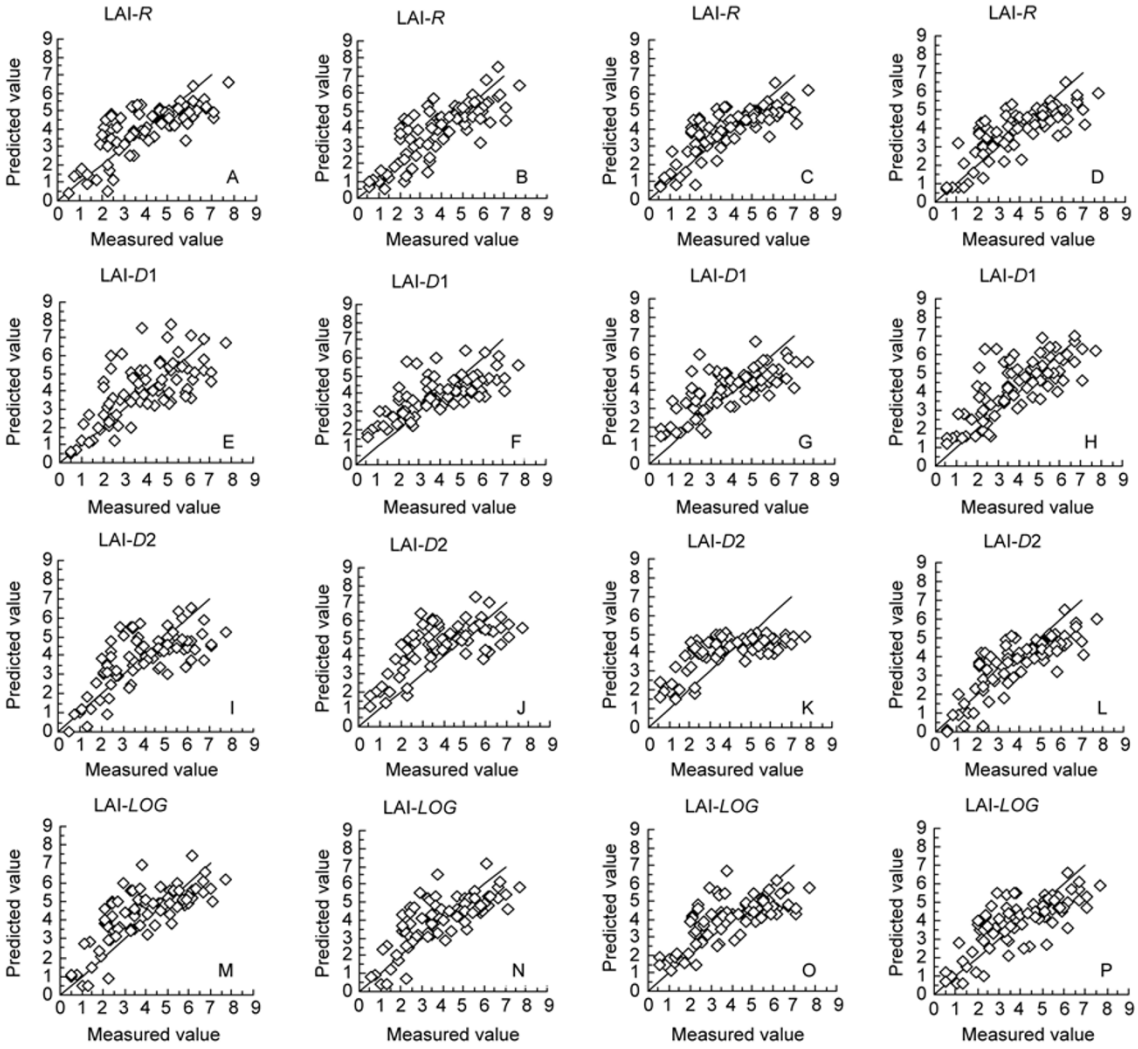

Figure 3 Scatter plots of predicted leaf area index (LAI) and measured LAI for the stepwise multivariable regression (SMR), the POLY kernel SVM (SVM-POLY), the RBF kernel SVM (SVM-RBF) and the ANOVA kernel SVM (SVM-ANOVA), using reflectance (*R*), the first derivative of reflectance (*D*1), the second derivative of reflectance (*D*2), and the logarithm-transformed reflectance (*LOG*). A–D, E–H, I–L and M–P represent SMR, SVM-POLY, SVM-RBF and SVM-ANOVA using *R*, *D*1, *D*2 and *LOG*, respectively. The line in each graph represents the line *y*=*x*.

Table 4 Test results of various support vector machines (SVMs) and stepwise multivariable regression models (SMRs) for the leaf area index (LAI)

Variable	Spectral transformation	SMR[f]			SVM-POLY[g]			SVM-RBF[h]			SVM-ANOVA[i]		
		RMSE[j] (m^2 m^{-2})	r[k]	ABSE[l] (m^2 m^{-2})	RMSE (m^2 m^{-2})	r	ABSE (m^2 m^{-2})	RMSE (m^2 m^{-2})	r	ABSE (m^2 m^{-2})	RMSE (m^2 m^{-2})	r	ABSE (m^2 m^{-2})
LAI[a]	*R*[b]	1.1198	0.7651	0.8905	**1.0496**[m]	**0.8024**	**0.8051**	1.0776	0.7844	0.8550	1.0797	0.7838	0.8465
	*D*1[c]	1.3349	0.6891	1.0066	1.2299	0.7090	0.9869	**1.2124**	**0.7185**	**0.9559**	1.2159	0.7534	**0.9194**
	*D*2[d]	1.2505	0.7084	0.9872	1.4713	0.6794	1.1968	1.3385	0.6633	1.1359	**1.0858**	**0.7941**	**0.8676**
	LOG[e]	1.2136	0.7572	0.9629	**1.1256**	**0.7637**	**0.8980**	1.2287	0.7067	0.9811	1.1582	0.7448	0.9406

a) Leaf area index; b) reflectances; c) first derivative of reflectances; d) second derivative of reflectances; e) logarithm-transformed reflectances; f) stepwise multivariable regression model; g) POLY kernel SVM; h) RBF kernel SVM; i) ANOVA kernel SVM; j) root mean squared error; k) correlation coefficients; l) average absolute error; m) data in bold show the best results for the prediction of LAI.

ized along the line *y*=*x* compared with the other two models using *LOG* (Figure 4J and K), and were also well correlated with the measured GLCD (0.7299, Table 5). This demonstrated that SVMs have better capability for non-linear mapping and estimation of GLCD compared with SMR approaches.

Comparing the overall SVM and SMR results for predicting GLCD, the SVM-ANOVA model based on *LOG* resulted in the lowest overall RMSE (523.0741 mg m^{-2}) of all models and data transformations. The SVM-ANOVA

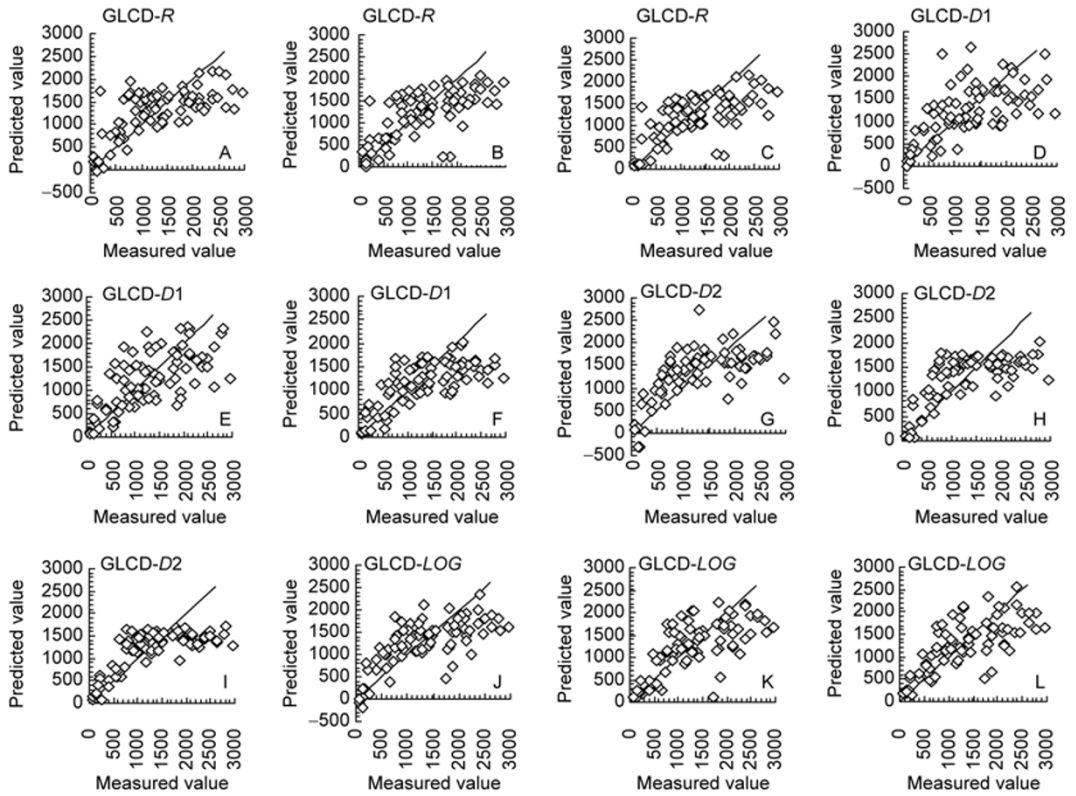

Figure 4 Scatter plots of predicted green leaf chlorophyll density (GLCD) and measured GLCD for the stepwise multivariable regression (SMR), the RBF kernel SVM (SVM-RBF) and the ANOVA kernel SVM (SVM-ANOVA), using reflectance (*R*), the first derivative of reflectance (*D*1), the second derivative of reflectance (*D*2), and the logarithm-transformed reflectance (*LOG*). A–C, D–F, G–I and J–L represent SMR, SVM-RBF and SVM-ANOVA using *R*, *D*1, *D*2 and *LOG*, respectively. The line represents the line of $y=x$.

Table 5 Test results of different SVM and SMR models for predicted GLCD

Variable	Spectral transformation	SMR[f]			SVM-POLY			SVM-RBF[g]			SVM-ANOVA[h]		
		RMSE[i] (mg m^{-2})	r[j]	ABSE[k] (mg m^{-2})	RMSE (mg m^{-2})	r	ABSE (mg m^{-2})	RMSE (mg m^{-2})	r	ABSE (mg m^{-2})	RMSE (mg m^{-2})	r	ABSE (mg m^{-2})
GLCD[a]	*R*[b]	574.5243	0.6550	454.6556	–	–	–	571.2719	0.6789	432.8516	570.7760[l]	0.6851	430.9946
	*D*1[c]	613.5800	0.6282	459.2451	–	–	–	575.2666	0.6755	444.4299	578.9491	0.7027	440.0834
	*D*2[d]	571.5815	0.6683	463.7467	–	–	–	553.2974	0.6866	442.2441	554.1760	0.7160	418.1450
	LOG[e]	551.3314	0.6930	436.6221	–	–	–	549.6904	0.7081	415.6934	523.0741	0.7299	394.7962

a) Green leaf chlorophyll density; b) reflectances; c) first derivative of reflectances; d) second derivative of reflectances; e) logarithm-transformed reflectances; f) stepwise multivariable regression model; g) RBF kernel SVM; h) ANOVA kernel SVM; i) root mean squared error; j) correlation coefficients; k)average absolute error; l) data in bold show the best results for the prediction of GLCD.

model based on *R* also had the lowest RMSE (570.7760) among all the other models (Table 5). For the case of *D*1 and *D*2, the SVM-RBF model had the lowest RMSE (575.2666 and 553.2794 mg m^{-2}, respectively), relative to the other models.

From the ABSE results (Table 5), significant improvements in the predictive GLCD relationships were achieved with the SVM models. The SVM-ANOVA models using *R*, *D*1, *D*2 and *LOG* datasets resulted in the lowest ABSE values (430.9946, 440.0834, 418.1450, and 394.7962 mg m^{-2}, respectively), relative to the other models using *R*, *D*1, *D*2 and *LOG*. All of the results showed that SVMs have higher accuracies for learning with good robustness, and further improve the relationships between different transformations of reflectance and GLCD.

2.4 Comparisons of SVM and SMR performance

The SVM and SMR methods for predicting LAI and GLCD resulted in significantly different *r*, RMSE, and ABSE values for the same datasets (Table 6). In the case of LAI, the SVM-POLY model using *R* improved the relationship between *R* and LAI with a lower RMSE 0.0702 and lower ABSE 0.0854. The SVM-RBF model using *D*1 improved the relationship between *D*1 and LAI with a lower RMSE 0.1225 and ABSE 0.0507, and the SVM-ANOVA model

Table 6 Improvement of different SVM models over the SMR models based on RMSE, *r*, and ABSE values

Variable		SVM-POLY[g]			SVM-RBF[h]			SVM-ANOVA[i]		
		ΔRMSE[j] (m^2 m^{-2})	Δr[k]	ΔABSE[l] (m^2 m^{-2})	ΔRMSE (m^2 m^{-2})	Δr	ΔABSE (m^2 m^{-2})	ΔRMSE (m^2 m^{-2})	Δr	ΔABSE (m^2 m^{-2})
LAI[a]	*R*[c]	**–0.0702**[m]	**0.0373**	**–0.0854**	–0.0422	0.0193	–0.0355	–0.0401	0.0187	–0.0440
	*D*1[d]	–0.1050	0.0199	–0.0197	**–0.1225**	**0.0294**	**–0.0507**	–0.1190	0.0643	–0.0872
	*D*2[e]	0.2208	–0.0290	0.2096	0.0880	–0.0451	0.1487	**–0.1647**	**0.0857**	**–0.1196**
	LOG[f]	**–0.0880**	**0.0065**	**–0.0649**	0.0151	–0.0505	0.0182	–0.0554	–0.0124	–0.0223
Variable		SVM-POLY			SVM-RBF			SVM-ANOVA		
		ΔRMSE (mg m^{-2})	Δr	ΔABSE (mg m^{-2})	ΔRMSE (mg m^{-2})	Δr	ΔABSE (mg m^{-2})	ΔRMSE (mg m^{-2})	Δr	ΔABSE (mg m^{-2})
GLCD[b]	*R*	–	–	–	–3.2524	0.0239	–21.8040	**–3.7483**	**0.0301**	**–23.6610**
	*D*1	–	–	–	**–38.3134**	**0.0473**	**–14.8152**	–34.6309	0.0745	–19.1617
	*D*2	–	–	–	**–18.2841**	**0.0183**	**–21.5026**	–17.4055	0.0477	–45.6017
	LOG	–	–	–	–1.6410	0.0151	–20.9287	**–28.2573**	**0.0369**	**–41.8259**

a) Leaf area index; b) green leaf chlorophyll density; c) reflectances; d) first derivative of reflectances; e) second derivative of reflectances; f) logarithm-transformed reflectances; g) POLY kernel SVM; h) RBF kernel SVM; i) ANOVA kernel SVM; j) increase of SVM over SMR on the root mean squared error; k) increase of SVM over SMR on the correlation coefficients; l) increase of SVM over SMR on the average absolute error; m) data in bold show the best results for the prediction of LAI and GLCD.

using *D*2 also improved the relationship between *D*2 and LAI with a lower RMSE 0.1647 and ABSE 0.1196. The SVM-POLY model using *LOG* also improved the relationship between *LOG* and LAI with lower RMSE 0.0880 and ABSE 0.0649.

Regarding GLCD, the relationship between *R* and GLCD was improved in the SVM-ANOVA model using *R* with a lower RMSE 3.7483 mg m^{-2} and ABSE 23.6610 mg m^{-2}. The relationship between *D*1 and GLCD was improved in the SVM-RBF model using *D*1 with a lower RMSE 38.3134 mg m^{-2} and ABSE 14.8152 mg m^{-2}. The relationship between *D*2 and GLCD was also improved by the SVM-RBF model with a lower RMSE 18.2841 mg m^{-2} and ABSE 21.5026 mg m^{-2}, as were the relationships by the SVM-ANOVA model using *LOG* with a lower RMSE 28.2573 mg m^{-2} and ABSE 41.8259 mg m^{-2}. Based on all these results, it can be concluded that SVMs improved the relationships between the four transformed reflectance datasets and the two biophysical parameters and, in all cases, the results using SVMs had higher accuracies than those using SMR.

3 Conclusion

In the present work, two different model approaches (SVM and SMR) using four different transformations of reflectance (*R*, *D*1, *D*2 and *LOG*), were used to compare their prediction capability to estimate rice LAI and rice GLCD (Tables 4 and 5). The SVM-POLY model using the four reflectance bands (R) was the best model to predict rice LAI, and the SVM-ANOVA model using LOG was the best model to predict rice GLCD.

At present, SVMs have only been used in a few agriculture studies with remote sensing. The present work represents an initial step in evaluating the merits of SVMs compared with the more traditional SMR models. We found SVMs performed better than SMR models, based on the four different reflectance transformations analyzed in this study, demonstrating that SVMs are a potentially useful method to understand and predict optical interactions over a wide range of rice canopy LAI and GLCD conditions. The large improvements observed in the SVM models over SMR also suggest that important non-linear processes exist in the relationships between remote sensing data and rice biophysical (e.g., LAI) and biochemical (e.g., chlorophyll) variables.

We found that selecting the appropriate kernel function and parameters of the kernel is critical to avoid 'over-fitting'. In the present work, three different kernel functions were selected, and many kernel parameters were tested for LAI and GLCD training. These two variables were optimized simultaneously when the highest improvements were observed. Consequently, SVMs provide a useful exploratory tool for improvement of the relationships between different transformations of reflectance and crop variables. Much work remains to be done to scale these greenness estimation relationships across a variety of canopies, so that this approach may become more robust and be applied in larger-scale remote sensing applications in agriculture.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 40571115 and 40271078), and the National Hi-Tech Research and Development Program of China (Grant No. 2006AA10Z203).

1 Pearson R L, Miller L D. Remote mapping of standing crop biomass for estimation of the productivity of the short-grass prairie, Pawnee National Grasslands, Colorado. In: Proceedings of the 8th international symposium on remote sensing of environment. ERIM International, Ann Arbor, MI, USA, 1972. 1357–1381

2 Tucker C J. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens Environ, 1979, 8: 127–150

3 Richardson A J, Wiegand C L. Distinguishing vegetation from soil background information. Photogramm Eng Rem Sens, 1977, 43: 1541–1552
4 Huete A R. A soil vegetation adjusted index (SAVI). Remote Sens Environ, 1988, 25: 295–309
5 Lyon J G, Yuan D, Lunetta R S, *et al*. A change detection experiment using vegetation indices. Photogramm Eng Rem Sens, 1998, 62: 143–150
6 Blackburn G A. Quantifying chlorophylls and caroteniods at leaf and canopy scales: An evaluation of some hyperspectral approaches. Remote Sens Environ, 1998, 66: 273–285
7 Thenkabail P S, Ward A D, Lyon J G. Landsat-5 thematic mapper models of soybean and corn crop characteristics. Int J Remote Sensing, 1995, 15: 49–61
8 Wiegand C J, Richardson A J, Escobar D E, *et al*. Vegetation indices in crop assessments. Remote Sens Environ, 1991, 35: 105–119
9 Elvidge C D, Lyon R J P. Influence of rock-soil spectral variation on the assessment of green biomass. Remote Sens Environ, 1985, 17: 265–269
10 Huete A R, Jackson R D, Post D F. Spectral response of a plant canopy with different soil backgrounds. Remote Sens Environ, 1985, 17: 37–53
11 Carter G A. Reflectance bands and indices for remote estimation of photosynthesis and stomatal conductance in pine canopies. Remote Sens Environ, 1998, 63: 61–72
12 Elvidge C D, Chen Z. Comparison of broad-band and narrowband red and near-infrared vegetation indices. Remote Sens Environ, 1995, 54: 38–48
13 Shibayama M, Akiyama T. Estimating grain yield by remote sensing of crop of rice canopies using high spectral resolution reflectance measurements. Remote Sens Environ, 1991, 36: 45–53
14 Yang X, Huang J, Wang F, *et al*. A modified chlorophyll absorption continuum index for chlorophyll estimation. J Zhejiang Univ Sci A, 2006, 7: 2002–2006
15 Baret F, Champion I, Guyot G, *et al*. Monitoring wheat canopies with a high spectral resolution radiometer. Remote Sens Environ, 1987, 22: 367–378
16 Gilabert M A, Gandia S, Melia J. Analyses of spectral-biophysical relationships for a corn canopy. Remote Sens Environ, 1996, 55: 17 11–20
17 Jackson R D, Pinter P J. Spectral response of architecturally different wheat canopies. Remote Sens Environ, 1986, 20: 43–56
18 Demetriades-Shah T H, Steven M D, Clark J A. High resolution derivative spectra in remote sensing. Remote Sens Environ, 1990, 33: 55–64
19 Peñuelas J, Gamon J A, Freeden A L, *et al*. Reflectance indices associated with physiological changes in nitrogen and water-limited sunflower leaves. Remote Sens Environ, 1994, 48: 135–146
20 Butler W L, Hopkins D W. Higher derivative analysis of complex absorption spectra. Photochem Photobiol, 1970, 12: 439–450
21 Filella I, Peñuelas J. The red edge position as indicators of plant chlorophyll content, biomass and hydric status. Int J Remote Sensing, 1994, 15: 1459–1470
22 Mauser W, Bach H. Imaging Spectroscopy in Hydrology and Agriculture-Determination of Model Parameters. In: Hill J, Megier J, eds. Imaging spectrometry—a tool for environmental observations. Dordrecht: Kluwer Academic Publishing, 1995. 261–283
23 Yoder B J, Pettigrew-Crosby R E. Predicting nitrogen and chlorophyll from reflectance spectra (400–2500 nm) at leaf and canopy scales. Remote Sens Environ, 1995, 53: 199–211
24 Johnson L F, Billow C R. Spectroscopic estimation of total nitrogen concentration in douglas—fir foliage. Int J Remote Sensing, 1996, 17: 489–500
25 Datt B. A new reflectance index for remote sensing of chlorophyll content in higher plants: Tests using eucalyptus leaves. J Plant Phy, 1999, 154: 30–36
26 Daughtry C S T, Walthall C L, Kim M S, *et al*. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sens Environ, 2000, 74: 229–239
27 Zarco-Tejada P J, Miller J R, Noland T L, *et al*. Scaling-up and model inversion methods with narrow-band optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. IEEE Trans Geosci Rem Sens, 2001, 39: 1491–1507
28 Cheng Q, Huang J, Wang X, *et al*. *In situ* hyperspectral data analysis for pigment content estimation of rice leaves. J Zhejiang Univ Sci A, 2003, 4: 727–733
29 Ripple W J. Determining coniferous forest cover and forest fragmentation with NOAA-9 advanced very high resolution radiometer rata. Photogramm Eng Rem Sens, 1994, 60: 553–540
30 Shibayama M, Takahashi W, Morinaga S, *et al*. Canopy water deficit detection in paddy rice using high resolution field spectroradiometer. Remote Sens Environ, 1993, 45: 117–126
31 Curran P J, Dungan J L, Macler B A, *et al*. Reflectance spectroscopy of fresh whole leaves for the estimation of chemical concentration. Remote Sens Environ, 1992, 39: 153–166
32 Keiner L E, Yan X H. A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. Remote Sens Environ, 1998, 66: 153–165
33 Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge: Cambridge University Press, 2000
34 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995
35 Smola A J, Scholkopf B, Muller K R. The connection between regularization operators and support vector kernels. Neural Network, 1998, 11: 637–649
36 Suykens J A K. Nonlinear modeling and support vector machines. In: Proceeding of IEEE instrumentation and measurement technology, Budapest, 2001
37 Thenkabail P S, Smith R B, Pauw E. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. Remote Sens Environ, 2000, 71: 158–182
38 Haboudane D, Miller J R, Tremblay N, *et al*. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. Remote Sens Environ, 2002, 81: 416–426
39 Monte R O, Bernard A, Engel Daniel R E, *et al*. Neural network prediction of maize yield using alternative data coding. Biosystems Engineering, 2002, 83: 31–45
40 Mutanga O, Skidmore A K. Integrating imaging spectroscopy and neural networks to map grass quality in the kruger national park, South Africa. Remote Sens Environ, 2004, 90: 104–115
41 Broge N H, Mortensen J V. Deriving green crop area index and canopy chlorophyll density of winter wheat from spectral reflectance data. Remote Sens Environ, 2002, 81: 45–57