International Journal of Remote Sensing
Vol. 32, No. 24, 20 December 2011, 9307–9325

Taylor & Francis
Taylor & Francis Group

Vegetation classification method with biochemical composition estimated from remote sensing data

KUN JIA, BINGFANG WU*, YICHEN TIAN, YUAN ZENG and QIANGZI LI
Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, PR China

(*Received 30 November 2009; in final form 29 November 2010*)

In this article, a vegetation classification hypothesis based on plant biochemical composition is presented. The basic idea of this hypothesis is that the vegetation species/crops have their own biochemical composition characteristics, which are separable from each other for those co-existing species at a specific region. Therefore, vegetation species can be classified based on the biochemical composition characteristics, which can be retrieved from hyperspectral remote-sensing data. In order to test this hypothesis, an experiment was conducted in north-western China. Field data on the biochemical compositions and spectral responses of different plants and an Earth-observing 1 (EO-1) Hyperion image were simultaneously collected. After analysing the relationship between biochemical composition and spectral data collected from Hyperion, the vegetation biochemical compositions were estimated using sample biochemical data and bands of Hyperion data. The vegetation classification was completed using the biochemical content classifier (BCC) and maximum-likelihood classifier (MLC) with all Hyperion bands (MLC_A) and selected bands (MLC_S), which were used for estimating considered biochemical contents (cellulose and carotenoid). The overall classification accuracy of the BCC (95.2%) was as good as MLC_S (95.2%) and better than MLC_A (91.1%), as was the kappa value (BCC 92.849%, MLC_S 92.845%, MLC_A 86.637%), suggesting that the BCC was a feasible classification method. The biochemical-based classification method has higher vegetation classification accuracy and execution speed, reduces data dimension and redundancy and needs only a few spectral bands to retrieve biochemical contents instead of using all of the spectral bands. It is an effective method to classify vegetation based on plant biochemical composition characteristics.

1. Introduction

Hyperspectral techniques have been developing rapidly in recent years. The ability of hyperspectral techniques to recognize different objects is a dramatic improvement over previous multispectral or single-wavelength techniques, because of their high spectral resolution. Many new hyperspectral instruments have been developed for use in remote sensing, such as the National Aeronautics and Space Administration (NASA)/Jet Propulsion Laboratory Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) (Green *et al*. 1998) and the Hyperion hyperspectral instrument carried by the Earth-observing 1 (EO-1) spacecraft.

*Corresponding author. Email: wubf@irsa.ac.cn

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online © 2011 Taylor & Francis
http://www.tandf.co.uk/journals
http://dx.doi.org/10.1080/01431161.2011.554454

Classification is one of the main usages of hyperspectral images. The maximum-likelihood classifier (MLC; Hoffbeck 1995) is a classical algorithm of image classification, whose origin can be traced back to electrical engineering (Nilsson 1965). And, currently, three main approaches based on intelligent data analysis are being applied in hyperspectral data classifications: expert systems (Skidmore 1989, De Jong and Riezebos 1991), artificial neural networks (ANNs; Heerman and Khazenie 1992, Binagli *et al*. 2005, Rogan *et al*. 2008, Pacifici *et al*. 2009) and decision trees (Friedl and Brodley 1997, Kandrika and Roy 2008, Tooke *et al*. 2009). However, no image classifier provides perfect results. For example, statistical methods such as the MLC are still in use for hyperspectral image classification (Binagli *et al*. 2005, Boschetti *et al*. 2007), which relies on the assumption that the probabilities of class membership can be modelled by a normal probability density function, and this assumption is not always valid. For ANNs, the training times can be lengthy, and designing the network architecture and choosing the values of the learning rate parameters are not straightforward (Foody and Arora 1997). The decision-tree algorithm has many advantages over the MLC and other intelligent classifiers: it is computationally fast, makes no statistical assumptions and can handle data that are represented using different measurement scales (Friedl and Brodley 1997); but most decision trees are fixed to binary splits for numeric attributes and do not allow backtracking in the tree construction phase (Ankerst *et al*. 1999).

Furthermore, a large number of image bands of hyperspectral images are too complex for parametric tools. The complexity of using such a large number of bands does not only reduce the precision of model estimation of these parametric tools, but also causes the singularity of covariance matrix inversion (Vaiphasa *et al*. 2007). This is even more serious for vegetation classification because of the comparability of spectra of different vegetation species/types. Many band selection methods have been developed to overcome this problem (Kavzoglu and Mather 2002, Ulfarsson *et al*. 2003, Vaiphasa *et al.* 2007), but they are always only suitable for specific studies.

Biochemical components of vegetation, such as the chlorophyll, protein, lignin, cellulose and nitrogen contents, are useful characteristics that can be measured in this manner because the contents and compositions of these compounds can directly or indirectly influence the vegetation's reflectance properties. Estimating biochemical content is thus an important application of remote sensing (Myneni *et al*. 1995, Verstraete *et al*. 1996). The earliest such study was applied to dry leaves. In the early 1960s, researchers at the United States Department of Agriculture (USDA) used near-infrared (NIR) spectroscopy to measure and analyse the spectra of dried and crushed plant leaves and successfully estimated the contents of cellulose, lignin, protein and starch (Curran 1989). Many researchers are now trying to find suitable bands to investigate the biochemical composition of fresh leaves. Fourty and Baret (1998) used the spectral reflectance of fresh leaves and a combination of several bands to estimate biochemical compositions. The estimation of water and dry matter contents gave good results, but estimates of other parameters contained high uncertainties.

After the EO-1 hyperspectral sensor was launched, researchers turned their attention to estimating vegetation biochemical compositions using hyperspectral data from satellites. Coops *et al*. (2003) used satellite-derived hyperspectral data to estimate eucalypt foliage nitrogen contents. Smith *et al*. (2003) estimated the nitrogen concentration of a temperate forest canopy and compared the results with the values estimated using the AVIRIS sensor. Townsend *et al*. (2003) also studied canopy nitrogen concentration. Ollinger and Smith (2005) simulated the nitrogen concentration of a temperate

forest canopy and estimated its net primary production by integrating field data with imaging spectroscopy. Zhang *et al*. (2008) developed a hyperspectral remote sensing algorithm to retrieve total leaf chlorophyll content, for both open spruce and closed forests, and tested for open forest canopies. At present, biochemical composition estimation from remote-sensing data usually includes three main methods: sensitive spectral bands (Huang *et al.* 2004), vegetation index (Wu *et al.* 2008) and model inversion (Kempeneers *et al.* 2008). A physical model is rarely used for biochemical contents estimation, for it is difficult to inverse and limited to several specific biochemical compositions.

Ustin *et al*. (2009) reviewed the recent advances in detecting plant pigments at the leaf level and discussed the successes of and reasons why challenges remain for robust remote observation and quantification. Kokaly *et al*. (2009) reviewed research into improving the application of imaging spectrometers to quantify non-pigment biochemical constituents of plants. Currently, hyperspectral data are widely used in estimating biochemical composition contents of plants, and more precise spectral resolution can help researchers to accurately retrieve vegetation biochemical compositions. Remote sensing offers a practical way to estimate foliar chemical concentrations, particularly when this must be done over large geographic areas.

Zarco-Tejada and Miller (1999) classified vegetated land cover based on red-edge spectral parameters, which were responsive to foliar chlorophyll pigment, and obtained a promising result. The basic idea of this article is that different plant species, co-existing in a certain area, have biochemical composition characteristics separable from each other, which can be retrieved by remote-sensing data and used to classify vegetation. In order to test this hypothesis, an experiment was conducted in north-western China, and an effective vegetation classification method was developed based on plant biochemical compositions estimated from Hyperion data.

2. Study area and data

2.1 *Study area*

The study area selected is a farm in north-western China's Gansu Province (see figure 1). It is a typical semi-arid farming area and has an average precipitation of 173.3 mm per year. Therefore, there is abundant sunlight and fine sunshine days, and it is relatively easy to acquire remote-sensing data. The farm is located in a flat valley at an average altitude of about 1900 m above sea level, and so uncertainty of classification accuracy caused by topographical facts will be reduced to the minimum. It is a producing base of legal opium poppy, which is used for medicine in China. The other dominant crops are wheat and sunflower.

2.2 *Remote sensing data*

An EO-1 Hyperion image and a Quickbird image were acquired in this study. The Hyperion image was acquired on 14 June 2005, at around 11:00 am local time (see figure 2). Hyperion is one of the three sensors on the NASA EO-1 platform, which was launched in November 2000. Hyperion is a push-broom imaging instrument that provides imagery with 242 spectral bands, with 10 nm spectral resolution and 30 m spatial resolution. Among its 242 spectral channels, channels 1–70 belong to the visible and NIR bands (400–1000 nm) and the others are shortwave infrared bands

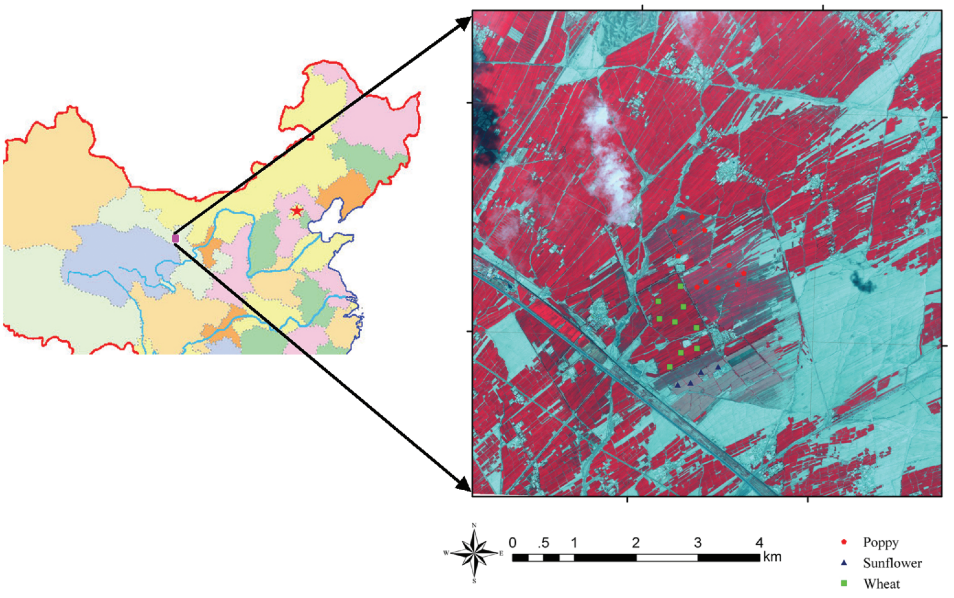

Figure 1. The pink square in the left image shows the geo-location of the study area in northwest China and the right image shows the distribution of the sampling points.

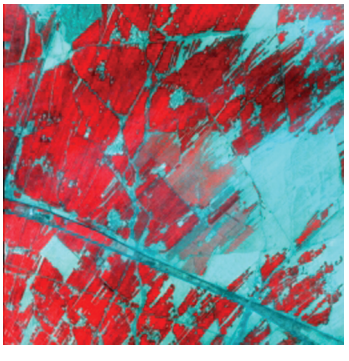

Figure 2. The Hyperion image used in this study.

(900–2500 nm). The satellite's data (digital numbers) were converted into radiances using the scaling approach proposed by Beck (2003).

In order to accurately position the sampling area and validate the classification accuracy of Hyperion, a high spatial resolution Quickbird image was obtained on 18 June 2005, at around 11:00 am local time. This multispectral image has four spectral bands: blue band, 450–520 nm; green band, 520–600 nm; red band, 630–690 nm; NIR band, 760–900 nm and a spatial resolution of 2.44 m.

2.3 *Field survey*

The field campaign was conducted on 14 June 2005 concurrently with acquired Hyperion data. The weather conditions were perfectly good, sunny and windless. They were suitable for acquiring spectral reflectance measurements and remote-sensing data. Twenty-five square sample sites (30 m × 30 m) within the study area were selected

based on the different crop distribution patterns and growth conditions (see figure 1). There were 12 poppy, 9 wheat and 4 sunflower sample sites. The centre of each sample site was determined using the Differential Global Positioning System (DGPS), with an accuracy of ±5 m. Within each sample site, there are five sample plots (each 3 m × 3 m), one at the centre of the site and the remaining four located at a distance of 10 m from each corner of the square site, along the diagonal of the square. Crop canopy spectral reflectance was measured at each sample plot, and then leaf samples were collected, as described below.

A FieldSpec Pro portable spectrometer (ASD Inc., Boulder, CO, USA) was used for the field spectral reflectance measurements. This spectrometer provided spectral coverage from 350 to 2500 nm at sampling intervals of 1.4 nm in the 350–1050 nm range and 2 nm in the 1050–2500 nm range. Pressed barium sulphate ($BaSO_4$) was used as the reference standard to calibrate the observed values. The spectrometer's probe, which has a 25° field of view, pointed straight downwards above the canopy. The measurement height was about 130 cm from the top of the canopy, and the field of view was about 60 cm wide. The spectral reflectance of each sample plot was measured 10 times. The overall spectral reflectance value for each sample plot was then calculated as the average of these 10 measurements, and the average of the five plots was the spectral value for each sample site. All spectra were converted into absolute reflectance values based on the measured value for the reference standard, which had known spectral properties.

At the same time as the spectral survey, leaf samples were collected at each sample plot for each crop. Three mature, fully expanded apical leaves were obtained, and the average of the five plots was the biochemical contents value for each sample site. The wet weight of the sampled leaves in each sample plot was about 200 g. Eight biochemical parameters (water content, protein, cellulose, lignin, chlorophyll, carotenoid, total nitrogen and total phosphorus) were measured in the laboratory, using the standard plant analysis methods in China (Shaanxi Normal University 1980, Zou 1995, Soil Science Society of China 2000).

3. Methods

3.1 *Data processing*

The Hyperion Level 1B data have 242 bands, of which 196 are valid bands; the remaining bands are the zero and overlapping bands, located in bands 1–7, 58–78 and 225–242. The 196 valid bands were analysed further. Several stripes (data columns of poor quality) in the Hyperion data contained no information or unusually low radiance values. These pixels were detected and replaced by the average radiance value of the immediately adjacent left and right pixels using the method proposed by Han *et al*. (2002). In addition, a minimum noise fraction (MNF) process was used to reduce the noise in the hyperspectral image (Green *et al*. 1988). Based on the eigenvalue profile, the effective bands that contained the most information were selected and an inverse MNF-transformed to obtain the Hyperion data for further analysis.

To obtain the hemispherical directional reflectance factor of the image, approximated by the surface reflectance (Schaepman *et al*. 2006), Atmospheric CORrection Now (ACORN) version 4.0 was used, which was based on the MODTRAN 4 radiative transfer model (AIG 2002). ACORN uses two water-absorption channels (940 and 1140 nm) to evaluate the amount of water vapour in combination with the visibility at the moment of data acquisition. Due to the low signal-to-noise ratio at the

beginning and end of the spectra (<436 and >2385 nm) and the significant water absorption in several spectral bands, 64 bands of the 196 valid bands were dropped, leaving 132 bands (located in bands 10–56, 87–96, 105–118, 135–162 and 189–221 of original Hyperion bands) for classification in this study.

Geometric corrections were performed using 40 ground control points from already geo-corrected Landsat Enhanced Thematic Mapper (ETM) data, which have a good consistency with the field GPS value, and the resulting geometric co-registration error was less than 1 pixel (30 m). A subset of the image that consisted of 256 columns × 256 lines × 132 spectral bands and that covered the area of interest was extracted from the Hyperion image (see figure 2) with World Geodetic System 84 (WGS-84) projection and at a 30 m spatial resolution. The comparison between the field survey and Hyperion spectral curve of a randomly selected poppy site is shown in figure 3. The Hyperion spectral curve has a similar trend to that of field survey data. It indicated a preferable image processing result.

To classify the hyperspectral data by means of MLC, endmembers that represent surface features were required. Endmembers were derived from known areas using the 'region of interest' (ROI) tools provided by ENVI version 4.5 (ITT Industries Inc., Boulder, CO, USA). In order to validate the classification accuracy, a number of sample pixels were randomly selected by the 'generate random sample' function of ENVI software based on the ground truth image, which was obtained by visual interpretation of the Quickbird image and ground survey. Table 1 summarizes the characteristics of the resulting endmember ROIs for training classifier and pixels used for validation.

In order to conduct the biochemical composition-based classification, biochemical parameters with obvious separability among the different plant species should be found out. One-way analysis of variance (ANOVA) was used to test whether the difference for the pairs of the three species (poppy vs. wheat, poppy vs. sunflower and wheat vs. sunflower) was significant for the eight biochemical parameter contents. The

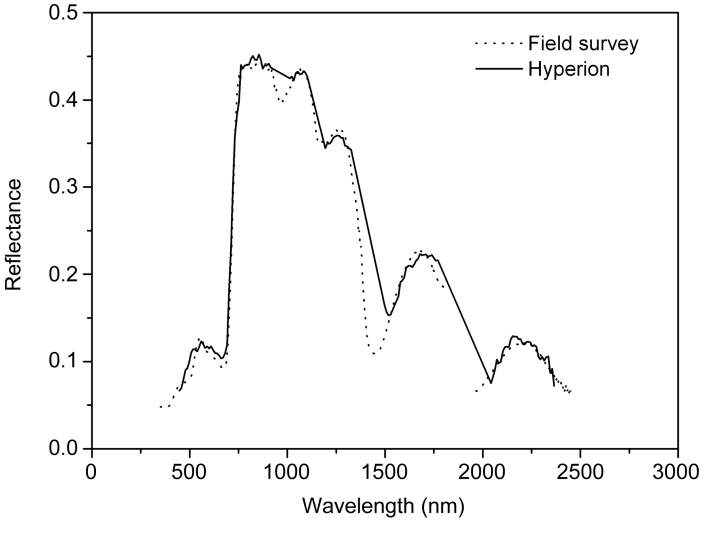

Figure 3. The comparison between the field survey and Hyperion spectral curve of a randomly selected poppy site.

Table 1. Number of regions of interest (ROIs) and pixels in each vegetation type used for training the classifier and number of pixels used for validation.

	Poppy	Wheat	Sunflower
NRT	8	6	6
NPT	300	287	245
NPV	105	98	112
TP	405	385	357

Notes: The sample pixels used for validation were generated by the 'generate random sample' function of ENVI software based on the ground truth image, which was obtained by visual interpretation of the Quickbird image and ground survey. NRT, number of ROIs used for training; NPT, number of pixels used for training; NPV, number of pixels used for validation; TP, total number of pixels used for training and validation.

ANOVA was tested with a 95% confidence level ($p < 0.05$). A significance test indicated that if the difference was significant then this biochemical parameter could be used for discriminating the two tested species. After ANOVA testing, the independent and appropriate biochemical parameters would be selected for further vegetation classification.

Some biochemical parameters of vegetation were strongly correlated, such as protein and total nitrogen. To identify the independent biochemical parameters, Pearson's correlations between the eight parameters were calculated. If the correlation between two parameters was higher than 0.8, only one of the two would be retained for further analysis. To do so, the parameter that had the lowest mean correlation with the other six parameters was chosen.

3.2 *Maximum-likelihood classifier method*

The MLC algorithm has been the most popular one used for classification of remote sensing imagery. As a parametric classifier, it assumes that a hyper-ellipsoid decision volume can be used to approximate the shape of the data clusters. For a given unknown pixel, described by a vector of features, the probability of membership in each class is calculated using the mean feature vectors of the classes, the covariance matrix and the prior probability (Duda and Hart 1973). The unknown pixel is considered to belong to the class with the maximum probability of membership. However, MLC classification performances are strongly related to the number of bands considered (Binagli *et al*. 2005). In this study, all bands of the Hyperion data and the bands selected for estimating biochemical parameter contents (biochemical used for biochemical content classifier, BCC) are provided for the MLC. The two classification results of the MLC will be compared with the result of the BCC to test whether the BCC is a credible classification method.

3.3 *Biochemical content classifier method*

The BCC algorithm is based on the principle that different plant species have separable biochemical compositions, and that these differences directly or indirectly influence the spectral reflectance of the vegetation. Therefore, the spectral reflectance can be used to estimate the biochemical composition of the vegetation at a canopy scale and can therefore be used to classify species based on their biochemical properties. The flowchart of BCC is shown in figure 4. The first step is to analyse the biochemical

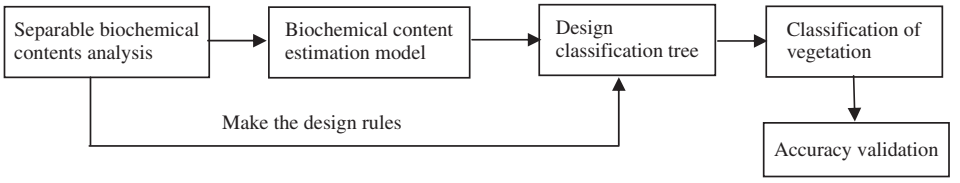

Figure 4. The flowchart of biochemical content classifier (BCC) method.

characteristics of different vegetation species/crop types and find out the biochemical compositions that can be used for classification. Next is to estimate the biochemical content from remote-sensing data using a model, which can be an empirical or a physical one. Then, based on the biochemical content and characteristics, it is possible to lay out a decision tree for classification. Finally, the classification accuracy will be validated.

3.3.1 Biochemical composition estimation. Biochemical contents estimation is the first step in the BCC classification method. In this study, the stepwise multivariable regression (SMR) method, which was most commonly used to predict crop variables in plants (Thenkabail *et al*. 2000, Curran *et al*. 2001, Haboudane *et al*. 2002), was applied to develop the biochemical contents estimation model using sample biochemical data and bands of Hyperion data. The SMR selected, in a stepwise manner, the appropriate Hyperion bands into the model of biochemical content estimation based on the experimental data. The regression equation would be used to estimate biochemical contents using Hyperion data.

3.3.2 Decision tree. The rationale of the decision-tree method is that starting from a set of examples described by a set of features, a binary decision rule can be defined that will split the data into two groups that are each more homogeneous than the original data. Each group is then iteratively subjected to a new split, generating increasingly homogeneous groups. In theory, the iteration continues until 'pure' subsets are obtained. Decision rules at each split are normally obtained by applying a threshold to the attribute that provides the best discrimination (a univariate tree) or by defining the best discriminant function based on linear combinations of attributes (a multivariate tree) (Brodley and Utgoff 1995). The choice of attributes to be used in each split is guided by a quality, which is applied to the generated subset. This step integrates the results of previous steps (i.e. to define the thresholds at each step in the tree) and uses the biochemical compositions to build the decision-tree classifier.

The BCC is only assumed to classify vegetation, so non-vegetated areas should be eliminated from further analysis, including roads, residential areas and water bodies. The elimination of non-vegetation areas was done by the normalized-difference vegetation index (NDVI), which revealed strong spectral differences between vegetated and non-vegetated areas. In this study, 890 nm (band 54 of the Hyperion data) and 670 nm (band 32 of the Hyperion data) were selected as the NIR band and the red band, respectively, to calculate the NDVI value. NDVI is defined as follows:

$$\mathrm{NDVI} = \frac{R_{\mathrm{NIR}} - R_{\mathrm{red}}}{R_{\mathrm{NIR}} + R_{\mathrm{red}}}, \tag{1}$$

where R_{NIR} is the reflectance in the NIR band and R_{red} is the reflectance in the red band.

Then, the biochemical parameters that could be used to distinguish between different plant species would be identified and used to define the decision rules applied to split the data into subsets. The decision rules were derived by expert knowledge analysing biochemical data values collected in this study.

3.4 *Validation*

To validate the BCC classification method, the classification results with those produced by visual interpretation of Quickbird image were compared. The objective was to determine whether the BCC was a feasible approach for classifying vegetation. Randomly selected sample pixels based on the ground truth image obtained from visual interpretation of Quickbird image and ground survey as described above were used to validate the accuracy of the classified maps. For each class, more than 90 pixels were obtained: 105 pixels for poppies, 98 pixels for wheat and 112 pixels for sunflowers. The classification accuracy and kappa statistic were then estimated (Congalton and Green 1999, Tso and Mather 2001). In order to access statistical differences between the accuracy measurements of the BCC and MLC, a Z-test was performed (Congalton 1991).

4. Results

4.1 *Biochemical compositions*

The ANOVA test results for each class pair are shown in table 2. The conclusions from the ANOVA test are that the mean biochemical contents between each class pair are significantly different in many measured biochemical compositions. All of the eight measured biochemical contents are significantly different for wheat versus sunflower. Seven are significant for poppy versus wheat; only total phosphorus is not. Only cellulose, carotenoid and total phosphorus contents are significantly different for poppy versus sunflower. It indicates many biochemical parameter contents are similar for poppy and sunflower. If only using one biochemical composition to classify the

Table 2. Result of statistical difference using ANOVA with 95% confidence level ($p < 0.05$) between the mean biochemical contents of all class pairs of poppy, wheat and sunflower.

Comparison	Water	Protein	Total nitrogen	Cellulose	Carotenoid	Lignin	Total phosphorus	Total chlorophyll
Poppy versus wheat	S	S	S	S	S	S	NS	S
Poppy versus sunflower	NS	NS	NS	S	S	NS	S	NS
Wheat versus sunflower	S	S	S	S	S	S	S	S

Note: S, significant; NS, not significant.

three species, cellulose and carotenoid will be the choice. There will be more choices if using two or more biochemical contents.

The Pearson's correlations between the different biochemical parameters were calculated (see table 3). Some biochemical parameters were strongly correlated, such as protein and nitrogen, and water and cellulose. And some biochemical parameters were independent of others; for example, total phosphorus was not obviously correlated with other biochemical parameters. Using all strongly correlated parameters to classify vegetation was not significant, for the classification results from using any one of these parameters was similar with any other one, even all of them. Thus, in the subsequent analyses, only one of the strongly correlated components would be considered. To do so, the parameter that had the lowest mean correlation with the other parameters was chosen. Five parameters, including cellulose, lignin, carotenoid, total phosphorus and chlorophyll, were selected for the subsequent analysis.

The statistical characteristics of the five biochemical parameters for the three plant species were calculated, including mean and standard deviation (SD) values (see table 4). The results showed that the mean biochemical compositions of the plants differed. For cellulose composition, the cellulose content of wheat was more than three times the values in the other species, and the content differed slightly between poppy and sunflower. So cellulose content was a significant biochemical composition factor for discriminating wheat from the other two species. Considering ANOVA test and the statistical characteristics of the five biochemical composition, lignin and chlorophyll content of sunflower and poppy were not significantly different, the carotenoid content of sunflower was higher than poppy and the difference of carotenoid contents between sunflower and poppy was larger than that between cellulose and phosphorus. So, carotenoid content was selected for distinguishing poppy and sunflower fields. Finally, the cellulose and carotenoid contents were selected as the biochemical composition factors to separate the three plants in this study. The wheat fields were discriminated from the other two species using cellulose contents, and then sunflower and poppy were distinguished by carotenoid contents.

After choosing the biochemical composition factors for classification, the next step was using Hyperion data to retrieve cellulose and carotenoid contents. The spectral reflectance data were extracted from the Hyperion image using the field survey GPS data. Then, the SMR was used for building the biochemical contents estimation model (see table 5). Of the nine wavebands selected for estimating cellulose and carotenoid contents, most were directly or indirectly related (located within $\pm$ 12 nm of absorption wavelength) to an absorption feature of the biochemical of interest or used by other researchers. For bands selected to estimate cellulose contents, R11 (457.34 nm), R96 (1104.18 nm) and R135 (1497.63 nm) were, respectively, related to absorption feature (Curran 1989) of chlorophyll (460 nm), lignin (1120 nm) and cellulose (1490 nm). Martin *et al*. (2008) found nitrogen content had a relationship to 720–730 nm, which was related to R37 (721.9 nm). Cellulose content had a high relationship with chlorophyll, lignin and nitrogen, so these bands were selected by the SMR to estimate that cellulose content was reasonable. For carotenoid content estimation, 510, 700, 710 (Gitelson *et al*. 2002, 2006), 708 and 860 nm (Datt 1998) bands were used. These bands were related to the bands selected (R16, R36 and R51) in this study to estimate carotenoid content. The regression model of cellulose and carotenoid estimation is shown in table 5.

Table 3. Correlations (Pearson's) among the eight biochemical parameters.

	Water	Protein	Total nitrogen	Cellulose	Carotenoid	Lignin	Total phosphorus	Total chlorophyll
Water	1							
Protein	0.905**	1						
Total nitrogen	0.905**	1.000**	1					
Cellulose	−0.956**	−0.932**	−0.933**	1				
Carotenoid	0.725**	0.854**	0.854**	−0.754**	1			
Lignin	−0.747**	−0.684**	−0.686**	0.747**	−0.464*	1		
Total phosphorus	0.254	0.346	0.350	−0.293	0.208	−0.233	1	
Total chlorophyll	0.687**	0.730**	0.732**	−0.672**	0.764**	−0.324	0.078	1

Notes: The number of sample data is 25.
**$p < 0.01$; *$p < 0.05$ (two-tailed test).

Table 4. The mean and standard deviation (SD) values of the five independent biochemical parameters from table 2 for the three species.

		Cellulose (g kg^{-1})	Lignin (g kg^{-1})	Carotenoid (mg (100 g)$^{-1}$)	Total phosphorus (g kg^{-1})	Total chlorophyll (g kg^{-1})
Poppy	Mean	73.76	221.43	43.03	4.31	1.49
	SD	3.68	29.73	4.71	0.66	0.17
Wheat	Mean	243.94	294.48	26.26	4.08	1.18
	SD	34.10	35.85	3.03	0.21	0.21
Sunflower	Mean	79.20	239.05	51.00	5.34	1.65
	SD	4.90	19.21	3.25	1.28	0.30

Note: The total sample number is 25.

Table 5. The results of stepwise multivariable regression (SMR) analysis between the wavebands of Hyperion and cellulose and carotenoid contents.

Biochemical	Number of bands selected	R^2	Selected wavebands of Hyperion, wavelength in parentheses (nm)	Model variable and coefficient in parentheses
Cellulose	5	0.955	R11 (457.34), R37 (721.9), R96 (1104.18), R135 (1497.63), R206 (2213.93)	R11 (3791.25), R37 (−3629.53), R96 (2265.18), R135 (−1482.42), R206 (2493.36), constant (–288.04)
Carotenoid	4	0.854	R16 (508.22), R36 (711.72), R51 (864.35), R118 (1326.05)	R16 (–1612.76), R36 (960.04), R51 (−554.59), R118 (−270.65), constant (–385.57)

Note: R^2 is the coefficient of determination.

4.2 *Decision tree*

NDVI was used to eliminate non-vegetated areas in the subsequent analysis, with NDVI >0.3 used to identify vegetated areas in the first step of the decision tree. Cellulose content was first used to separate wheat from poppy and sunflower, using a threshold value of 150 g kg^{-1}. Wheat had greater cellulose content than poppy and sunflower. Pixels with cellulose contents larger than 150 g kg^{-1} were classified as wheat. Then carotenoid content was used to separate poppy and sunflower with a threshold value of 47.7 mg (100 g)$^{-1}$. Pixels that had an estimated carotenoid content higher than these values were classified as sunflower, and others were classified as poppy. The resulting decision-tree classifier is shown in figure 5.

4.3 *Classification*

Vegetation classification based on the Hyperion data by applying the BCC and MLC methods was conducted using the ENVI software (see figure 6). A mask, generated using the NDVI values from the decision tree used in the BCC approach, was also

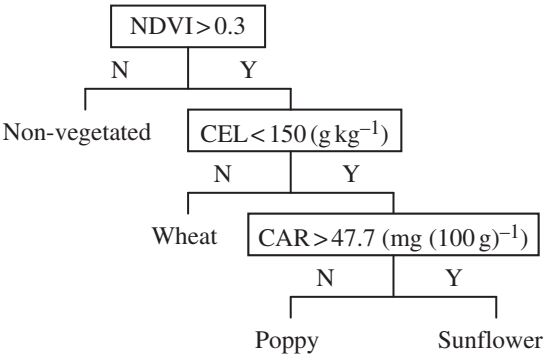

Figure 5. The decision-tree classifier developed to distinguish between the three crops based on Hyperion spectral reflectance data.
Note: CAR, carotenoid contents; CEL, cellulose content; Y, yes; N, no; NDVI, normalized-difference vegetation index.

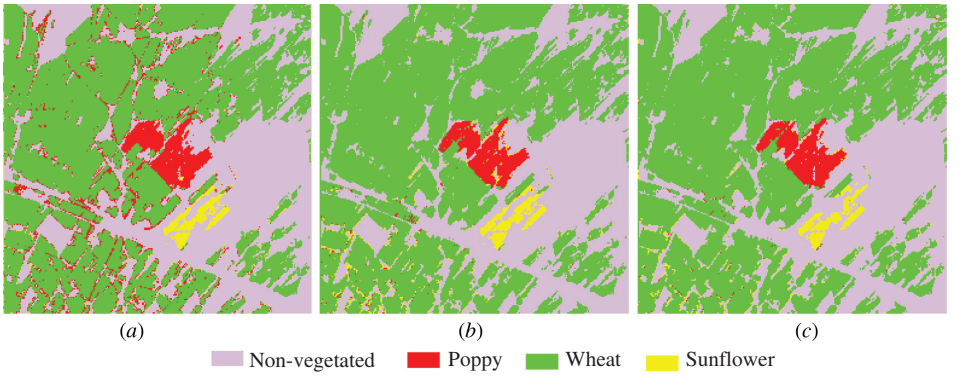

Figure 6. Classification results: (*a*) maximum-likelihood classifier (MLC) method using all Hyperion bands; (*b*) MLC using the bands which are selected for estimating cellulose and carotenoid contents; and (*c*) biochemical content classifier (BCC) method.

used in the MLC approach to eliminate non-vegetated areas, so that the classification results from the two methods could be compared.

All of the poppy and sunflower fields had been discriminated in the Hyperion image using both BCC and MLC classification methods. However, the classification results of the MLC using all the Hyperion bands (MLC_A) contained many small speckles at the edges of vegetated areas, such as pixels at many edges of wheat fields that were mis-classified as poppy pixels. This phenomenon might be caused by spectral mixture in the edges of the planted area, for the spatial resolution of Hyperion is not very fine. This phenomenon was also seen in the classification results of the BCC and MLC methods using the bands which are selected for estimating cellulose and carotenoid contents (MLC_S), but the number of speckles was less than that of the MLC method using all the Hyperion bands, and the edges of the vegetated areas were smoother. The mis-classification of spectral mixed pixels at the edges of vegetation areas greatly influence (i.e. decrease) the classification accuracy.

The BCC method has a similar classification result with MLC_S and the two classification results are comparative. It indicated that the BCC and MLC_S performed better than MLC_A in classification accuracy and in dealing with spectral mixed pixels, which could be leading to mis-classification. The exact accuracy evaluation is seen next.

4.4 *Accuracy evaluation*

The choice of the most appropriate classifier in terms of accuracy depends greatly on the objectives of the mapping project (Stehman 1997). Because the aim of this study was to evaluate whether a classification method based on plant biochemical compositions would be feasible, the overall classification accuracy and the kappa coefficient are both important indicators. Table 6 shows the confusion matrix achieved using the MLC and BCC methods for the same test data.

The overall performance of the BCC method (accuracy: 95.2%; kappa coefficient: 92.849%) was as good as MLC_S (accuracy: 95.2%; kappa coefficient: 92.845%) and better than MLC_A (accuracy: 91.1%; kappa coefficient: 86.637%). The Z-test was used to compare the error matrices (two at a time) to determine whether they are significantly different (see table 7). $Z > 1.96$ or $Z < -1.96$ would indicate the difference of the two error matrices being significant at the 5% significance level (Foody 2009). If the two error matrices are not significantly different, when given the choice of only these two approaches, one should use the easier, quicker or more efficient approach because the accuracy will not be the deciding factor (Congalton 1991). We can see from the Z-test results that the performance of MLC_A is significantly different to that of MLC_S and BCC, and BCC versus MLC_S is not significantly different. It indicated that better classification results can be achieved using bands which are selected for estimating biochemical factors. It is therefore concluded that vegetation classification based on plant biochemical compositions is a feasible vegetation classification method using remote-sensing data.

Table 6. Confusion matrix for the vegetation classification results of maximum-likelihood classifier (MLC) using all Hyperion bands (MLC_A) and using selected bands (MLC_S) and classification results based on the biochemical content classifier (BCC).

		Ground truth result (pixels)			
Mapped class		Poppy	Wheat	Sunflower	Total
Poppy	MLC_A	102	14	4	120
	MLC_S	104	5	3	112
	BCC	104	5	2	111
Wheat	MLC_A	2	81	4	87
	MLC_S	0	90	6	96
	BCC	0	91	6	97
Sunflower	MLC_A	1	3	104	108
	MLC_S	1	3	103	107
	BCC	1	2	104	107
Total	MLC_A	105	98	112	315
	MLC_S	105	98	112	315
	BCC	105	98	112	315

Table 7. *Z*-test results for comparison between error matrices for the different classification methods.

Comparison	*Z*-statistic	Result*
MLC_A versus MLC_S	2.0680	S
MLC_A versus BCC	2.0696	S
MLC_S versus BCC	0.0015	NS

Notes: MLC_A, maximum-likelihood classifier using all Hyperion bands; MLC_S, maximum-likelihood classifier using selected bands; BCC, biochemical content classifier. S, significant; NS, not significant.
*At the 95% confidence level.

5. Discussion and conclusions

Just like human vision discriminates vegetation types based on colour, remote sensing classifies vegetation types based on spectral reflectance. The pigment content of a plant determines the vegetation colour, and biochemical content influences the spectral reflectance of vegetation. If plant biochemical content can be used for vegetation classification of remote sensing, it will be more significant and will tally with human cognizance systems.

This study is to develop a vegetation classification method based on plant biochemical compositions estimated from remote-sensing images. The results indicated that if there was a separable biochemical characteristic for different vegetation species/crop types, and if using hyperspectral image plant biochemical content can be estimated over the vegetation canopy, then the estimated biochemical content can be used to classify vegetation species based on the separable biochemical characteristics. The classification results using the plant biochemical compositions-based method are more accurate than a traditional method based on an MLC using all the hyperspectral data, and thus represent a feasible tool for vegetation classification at a canopy scale.

Only a few spectral bands which were used to retrieve biochemical contents were needed for the biochemical-based classification method instead of all of the spectral bands of remote-sensing data. This reduced the data dimension and redundancy and gave high vegetation classification accuracy. The BCC method using decision trees for classification shares the same advantages as a decision-tree classifier compared with traditional probabilistic algorithms. Decision trees are strictly non-parametric, free from assumptions about the distribution of a parameter, can deal with non-linear relationships, are insensitive to missing values and can handle both numerical and categorical inputs and rapid calculation.

The biochemical contents estimation method in this study uses the SMR method, which is an empirical model and must be spatially and temporally limited. Although most of the selected bands are related to the absorption feature of the biochemical and make the estimation steady, the empirical model still needs to be built based on field survey data in specific regions. The biochemical content estimation accuracy can also influence the vegetation classification accuracy. An inaccurate biochemical content estimation result may lead to a false classification result. It is necessary to master the separable biochemical characters of different vegetation types in specific study areas, for it is important information to build the decision-tree classifier and the basic hypothesis of the BCC classification method.

The BCC method provides a significant vegetation classification method for remote-sensing data and will be used for vegetation classification on a larger scale, such as classification of crop types in agricultural regions and tree types in forest regions. This study provides an important illustration of the usefulness of the BCC method using hyperspectral remote sensing in an agricultural region. Most previous studies of plant biochemical content estimation using remote sensing were aimed at a single species. However, the BCC method must be able to estimate biochemical compositions for heterogeneous vegetation communities at a canopy scale, which is a large challenge for remote sensing. Despite this challenge, in this study, an empirical method was used to estimate plant biochemical compositions with accuracy at least as good as the MLC approach.

Although good results were achieved, accuracy should still be improved, and the empirical equations are limited both temporally and spatially. The selection of a physically based wavelength for biochemical content estimation would make the present work more robust, and this is a future objective of this research. The BCC method will be a useful tool in vegetation classification as it needs fewer spectral bands and has higher classification accuracy.

Acknowledgements

This work was supported by projects in the National Science & Technology Pillar Programme of China (No. 2006BAK09B01) and the Knowledge Innovation Programmes of the Chinese Academy of Sciences (No. KSCX1-YW-09-01). We are grateful for the comments and suggestions of two anonymous referees and for the help of the editor, Dr S. Paloscia, all of which have led to improvements in the presentation of this article.

References

AIG, 2002, *ACORN 4.0 User's Guide* (Boulder, CO: Analytical Imaging and Geophysics LLC).

Ankerst, M., Elsen, C., Ester, M. and Kriegel, H., 1999, Visual classification: an interactive approach to decision tree construction. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 15–18 August 1999, San Diego, CA (New York: Association for Computing Machinery), pp. 392–396.

Beck, R., 2003, *EO-1 User Guide—Version 2.3* (Cincinnati, OH: University of Cincinnati).

Binagli, E., Gallo, I., Boschetti, M. and Brivio, P.A., 2005, A neural adaptive algorithm for feature selection and classification of high dimensionality data. In *ICIAP 2005, LNCS 3617*, F. Roli and S. Vitulano (Eds.), pp. 753–760 (Berlin: Springer-Verlag).

Boschetti, M., Boschetti, L., Oliveri, S. and Casati, L., 2007, Tree species mapping with airborne hyper-spectral MIVIS data: the Ticino Park study case. *International Journal of Remote Sensing*, **28**, pp. 1251–1261.

Brodley, C.E. and Utgoff, P.E., 1995, Multivariate decision trees. *Machine Learning*, **19**, pp. 45–77.

Congalton, R.G., 1991, A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, pp. 35–46.

Congalton, R.G. and Green, K., 1999, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (Boca Raton, FL: Lewis Publishers).

Coops, N.C., Smith, M.L., Martin, M.E. and Ollinger, S.V., 2003, Prediction of eucalypt foliage nitrogen content from satellite-derived hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, pp. 1338–1346.

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

CURRAN, P.J., DUNGAN, J.L. and PETERSON, D.L., 2001, Estimating the foliar biochemical concentration of leaves with reflectance spectrometry testing the Kokaly and Clark methodologies. *Remote Sensing of Environment*, **76**, pp. 349–359.

DATT, B., 1998, Remote sensing of chlorophyll a, chlorophyll b, chlorophyll a + b, and total carotenoid content in eucalyptus leaves. *Remote Sensing of Environment*, **66**, pp. 111–121.

DE JONG, S.M. and RIEZEBOS, H.T., 1991, Use of a GIS database as "a priori" knowledge in multispectral land cover classification. In *Proceedings of the Second European Conference on Geographical Information Systems*, 2–5 April 1991, Brussels, Belgium (Utrecht: EGIS Foundation), pp. 503–508.

DUDA, R.O. and HART, P.E., 1973, *Pattern Classification and Scene Analysis* (New York: John Wiley & Sons).

FOODY, G.M., 2009, Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, **113**, pp. 1658–1663.

FOODY, G.M. and ARORA, M.K., 1997, An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, **18**, pp. 799–810.

FOURTY, T. and BARET, F., 1998, On spectral estimates of fresh leaf biochemistry. *International Journal of Remote Sensing*, **19**, pp. 1283–1297.

FRIEDL, M.A. and BRODLEY, C.E., 1997, Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, **61**, pp. 399–409.

GITELSON, A.A., KEYDAN, G.P. and MERZLYAK, M.N., 2006, Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. *Geophysical Research Letters*, **33**, L11402, doi: 10.1029/2006GL026457.

GITELSON, A.A., ZUR, Y., CHIVKUNOVA, O.B. and MERZLYAK, M.N., 2002, Assessing carotenoid content in plant leaves with reflectance spectroscopy. *Photochemistry and Photobiology*, **75**, pp. 272–281.

GREEN, A.A., BERMAN, M., SWITZER, P. and CRAIG, M.D., 1988, A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, **26**, pp. 65–74.

GREEN, R.O., EASTWOOD, M.L., SARTURE, C.M., CHRIEN, T.G., ARONSSON, M., CHIPPENDALE, B.J., FAUST, J.A., PAVRI, B.E., CHOVIT, C.J., SOLIS, M., OLAH, M.R. and WILLIAMS, O., 1998, Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, **65**, pp. 227–248.

HABOUDANE, D., MILLER, J.R., TREMBLAY, N., ZARCO-TEJADA, P.J. and DEXTRAZE, L., 2002, Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment*, **81**, pp. 416–426.

HAN, T., GOODENOUGH, D.G., DYK, A. and LOVE, J., 2002, Detection and correction of abnormal pixels in Hyperion images. In *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, 24–28 June 2002, Toronto, ON, Canada (New York: IEEE Press), pp. 1327–1330.

HEERMAN, P.D. and KHAZENIE, N., 1992, Classification of multispectral remote sensing data using a back-propagation neural network. *IEEE Transactions on Geoscience and Remote Sensing*, **20**, pp. 81–88.

HOFFBECK, J.P., 1995, Classification of high dimensional multispectral data. PhD thesis, Purdue University, West Lafayette, IN, USA.

HUANG, Z., TURNER, B.J., DURY, S.J., WALLIS, I.R. and FOLEY, W.J., 2004, Estimating foliage nitrogen concentration from HYMAP data using continuum removal analysis. *Remote Sensing of Environment*, **93**, pp. 18–29.

KANDRIKA, S. and ROY, P.S., 2008, Land use land cover classification of Orissa using multitemporal IRS-P6 AWIFS data: a decision tree approach. *International Journal of Applied Earth Observation and Geoinformation*, **10**, pp. 186–193.

Kavzoglu, T. and Mather, P.M., 2002, The role of feature selection in artificial neural network applications. *International Journal of Remote Sensing*, **23**, pp. 2919–2937.

Kempeneers, P., Zarco-Tejada, P.J., North, P.R.J., De, B.S., Delalieux, S., Sepulcre-Canto, G., Morales, F., Van Aardt, J.A.N., Sagardoy, R., Coppin, P. and Scheunders, P., 2008, Model inversion for chlorophyll estimation in open canopies from hyperspectral imagery. *International Journal of Remote Sensing*, **29**, pp. 5093–5111.

Kokaly, R.F., Asner, G.P., Ollinger, S.V., Martin, M.E. and Wessman, C.A., 2009, Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sensing of Environment*, **113**, pp. S78–S91.

Martin, M.E., Plourde, L.C., Ollinger, S.V., Smith, M.-L. and Mcneil, B.E., 2008, A generalizable method for remote sensing of canopy nitrogen across a wide range of forest ecosystems. *Remote Sensing of Environment*, **112**, pp. 3511–3519.

Myneni, R.B., Maggion, S. and Laquinta, J., 1995, Optical remote sensing of vegetation: modeling, caveats, and algorithms. *Remote Sensing of Environment*, **51**, pp. 169–188.

Nilsson, J.N., 1965, *Learning Machines: Foundations of Trainable Pattern-Classifying Systems* (New York: McGraw-Hill).

Ollinger, S.V. and Smith, M.L., 2005, Net primary production and canopy nitrogen in a temperate forest landscape: an analysis using imaging spectroscopy, modeling and field data. *Ecosystems*, **8**, pp. 760–778.

Pacifici, F., Chini, M. and Emery, W.J., 2009, A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment*, **113**, pp. 1276–1292.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C. and Roberts, D., 2008, Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. *Remote Sensing of Environment*, **112**, pp. 2272–2283.

Schaepman, S.G., Schaepman, M.E., Painter, T.H., Dangel, S. and Martonchik, J.V., 2006, Reflectance quantities in optical remote sensing—definitions and case studies. *Remote Sensing of Environment*, **103**, pp. 27–42.

Shaanxi Normal University, 1980, *Analyses Method in Common Use of Agricultural Chemistry* (Xi'an: Technology Press of Shaanxi) [in Chinese].

Skidmore, A.K., 1989, An expert system classifies eucalypt forest types using Thematic Mapper data and a digital terrain model. *Photogrammetric Engineering and Remote Sensing*, **55**, pp. 1449–1464.

Smith, M.L., Martin, M.E., Plourde, L. and Ollinger, S.V., 2003, Analysis of hyperspectral data for estimation of temperate forest canopy nitrogen concentration: comparison between an airborne (AVIRIS) and a spaceborne (Hyperion) sensor. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 1332–1333.

Soil Science Society of China, 2000, *Analyze Method of Soil and Agricultural Chemistry* (Beijing: Chinese Agricultural Press) [in Chinese].

Stehman, S.V., 1997, Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62**, pp. 77–89.

Thenkabail, P.S., Smith, R.B. and De Pauw, E., 2000, Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment*, **71**, pp. 158–182.

Tooke, T.R., Coopes, N.C., Goodwin, N.R. and Voogt, J.A., 2009, Extracting urban vegetation characteristics using spectral mixture analysis and decision tree classifications. *Remote Sensing of Environment*, **113**, pp. 398–407.

Townsend, P.A., Foster, J.R., Chastain, R.A. and Currie, W.S., 2003, Application of imaging spectroscopy to mapping canopy nitrogen in the forests of the Central Appalachian Mountains using Hyperion and AVIRIS. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, pp. 1347–1354.

Tso, B. and Mather, M.P., 2001, *Classification Methods for Remotely Sensed Data* (London: Taylor & Francis).

Ulfarsson, M.O., Benediktsson, J.A. and Sveinsson, J.R., 2003, Data fusion and feature extraction in the wavelet domain. *International Journal of Remote Sensing*, **24**, pp. 3933–3945.

Ustin, S.L., Gitelson, A.A., Jacquemoud, S., Schaepman, M., Asner, G.P., Gamon, J.A. and Zarco-Tejada, P., 2009, Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment*, **113**, pp. S67–S77.

Vaiphasa, C., Skidmore, A.K., Boer, W.F. and Vaiphasa, T., 2007, A hyperspectral band selector for plant species discrimination. *ISPRS Journal of Photogrammetry & Remote Sensing*, **62**, pp. 225–235.

Verstraete, M.M., Pinty, B. and Myneni, R.B., 1996, Potential and limitations of information extraction on the terrestrial biosphere from satellite remote sensing. *Remote Sensing of Environment*, **58**, pp. 201–214.

Wu, C.Y., Niu, Z., Tang, Q. and Huang, W.J., 2008, Estimating chlorophyll content from hyperspectral vegetation indices: modeling and validation. *Agricultural and Forest Meteorology*, **148**, pp. 1230–1241.

Zarco-Tejada, P.J. and Miller, J.R., 1999, Land cover mapping at BOREAS using red edge spectral parameters from CASI imagery. *Journal of Geophysical Research*, **104**, pp. 27 921–27 933.

Zhang, Y.Q., Chen, J.M., Miller, J.R. and Noland, T.L., 2008, Leaf chlorophyll content retrieval from airborne hyperspectral remote sensing imagery. *Remote Sensing of Environment*, **112**, pp. 3234–3247.

Zou, Y., 1995, *Experimental Guidance to Plant Physiology and Biochemistry* (Beijing: Chinese Agricultural Press) [in Chinese].