Spectral resampling based on user-defined inter-band correlation filter: C_3 and C_4 grass species classification

Clement Adjorlolo[a]*, Onisimo Mutanga[a], Moses A. Cho[a,b], Riyad Ismail[a]

[a]University of KwaZulu-Natal, School of Agriculture, Earth and Environmental Sciences, P. Bag X01, Scottsville 3209, Pietermaritzburg, South Africa
[b]Council for Industrial and Scientific Research, Meiring Naude Road, P.O. Box 395, Pretoria 0001, South Africa
*Corresponding author. Email: clement.adjorlolo@kzndae.gov.za

Abstract

In this paper, a user-defined inter-band correlation filter function was used to resample hyperspectral data and thereby mitigate the problem of multicollinearity in classification analysis. The proposed resampling technique convolves the spectral dependence information between a chosen band-centre and its shorter and longer wavelength neighbours. Weighting threshold of inter-band correlation (WTC, Pearson's r) was calculated, whereby $r = 1$ at the band-centre. Various WTC ($r = 0.99$, $r = 0.95$ and $r = 0.90$) were assessed, and bands with coefficients beyond a chosen threshold were assigned $r = 0$. The resultant data were used in the random forest analysis to classify C_3 and C_4 grass species. The respective WTC datasets yielded improved classification accuracies (kappa = 0.82, 0.79 and 0.76) with less correlated wavebands when compared to resampled Hyperion bands (kappa = 0.76). Overall, the results obtained from this study suggested that resampling of hyperspectral data should account for the spectral dependence information to improve overall classification accuracy as well as reducing the problem of multicollinearity.

Keywords: Spectral resampling; Inter-band correlation; Grass species classification; Random forests

1. Introduction

Discriminating grass species, which correspond to the 3-carbon (C_3) or 4-carbon (C_4) photosynthetic pathways, is consistent with the plant functional type (PFT) approach used in land surface modelling schemes (Tieszen et al., 1997; Ustin and Gamon, 2010). In general, C_3 and C_4 grasses differ significantly in a number of physiological and anatomical characteristic features. The C_4 type of grass species has more compact leaf mesophyll, higher proportion of vascular tissue and a lower interveinal distance, than those of C_3 grasses. In addition, several biochemicals such as intercellular air-moisture and nitrogen concentration are relatively lower in C_4 grass, compared to C_3 grass species (Oyarzabal et al., 2008). Such differences can manifest in the composition of C_3 and C_4 grasslands, with the dominant species strongly constituting the canopy reflectance. The fundamental principle is that C_3 and C_4 grass canopy reflectance is directly dependent on their spectral properties, which are in turn, controlled by the biophysical and biochemical characteristics of vegetation (Mutanga et al., 2004). Several empirical evidences have shown that the spectral variability between C_3 and C_4 grass species or groups of grasses is greater than the within group spectral information (Irisarri et al., 2009; Liu and Cheng, 2011; Smith and Blackshaw, 2003). For example, reflectance at 531 and 570 nm have been proposed as a set of spectral bands sensitive to differences in C_3 and C_4 species (Gamon et al., 1997). Slaton et al. (2001) modelled the near infrared (NIR) region and found that the reflectance around 800 nm is significantly different among species, which differ in intercellular structure. In a common garden experiment, Irisarri et al. (2009) demonstrated that it is possible, using reflectance centred around 820 nm to differentiate between C_3 and C_4 grass compositions.

The major challenge, however, is that spectral reflectance data obtained over many narrow contiguous channels (i.e. hyperspectral data) can represent multiple classes that are often mixed for a limited training-sample size (i.e. n < P: Chi and Bruzzone 2007). The problem of n < P is associated with the well-described "curse of dimensionality" or the Hughes phenomenon (Hughes, 1968). This phenomenon causes a decrease in a classifier ability to generalize accurately (Ham et al., 2005; Pal and Foody, 2010). Hence, very large training-samples are required to achieve a good description of data distribution (Dalponte et al., 2009). Besides, the Hughes phenomenon often introduces high degree of multicollinearity, caused by the use of highly-correlated predictor variables (Clevers et al., 2007). Multicollinearity is a prominent problem in processing hyperspectral data for vegetation applications, due to similarities in the reflectance properties of biophysical and biochemical characteristics (Ferwerda et al., 2005; Knox et al., 2010; Zhang et al., 2011). The problem of multicollinearity in the matrix of input spectral bands often leads to highly unstable parameter estimates and thus generalization error for a classifier (Bruzzone and Serpico, 2000; Clevers et al., 2007).

Attempts to solve the problems associated with spectral dimensionality and the related co-linearity include the use of feature reduction and feature selection techniques. The feature selection approach includes those based on a search strategy and on a separability measure. The Sequential Forward Floating Selection (Pudil et al., 1994) and the Steepest Ascent (Serpico and Bruzzone, 2001) are commonly used search strategy techniques, whereas the Bhattacharyya distance (Djouadi et al., 1990), Jeffries–Matusita distance (Bruzzone et al., 1995) and the transformed divergence distance (Su et al., 1990) are examples of the separability measures used in processing hyperspectral data. However, these feature selection techniques require estimation of some statistical properties at full dimensionality, in order to select optimum subset of the input spectral bands for a given classification task. If the training samples are insufficient, the parameterization may not be reliably adequate for the feature selection process (Chi and Bruzzone, 2007).

The studies by Schmidt & Skidmore (2003) and Becker et al. (2007) used the approach of analysing the most sensitive wavebands, considering the physical or spectroscopic meaning of each band across the spectrum. This approach often involves the resampling of high-dimensional spectra to wider bandwidths around a few chosen band-centres or to the spectral configuration of existing sensors. In this respect, the sensors respective spectral response functions or spectral resolutions (i.e. Full Width at Half Maximum, FWHM) are simulated. The major limitation of existing resampling methods is that an inherent property of vegetation spectral response is not considered. That is, the asymmetrical nature of correlation between a given waveband (λ) and its shorter and longer wavelength neighbours are not fully accounted for. In this regard, a more innovative approach can be followed, whereby the researchers consider the inter-band correlations around each band centre of interest. The approach has the advantage of linking the physical properties of the target vegetation and its characteristic spectral response function across the spectrum (Schmidt and Skidmore, 2003). When hyperspectral data are processed in this way, classifications are based on the spectroscopic interpretability of each set band (Becker et al., 2007; Faurtyot and Baret, 1997).

From this background, the present study sought to classify C_3 and C_4 grass canopies using resampled hyperspectral data obtained through an approach that convolves the spectral information around a chosen given band-centre. The resultant datasets were analysed using the random forest (Breiman, 2001) algorithm. Random forests are advanced non-parametric classifiers, which are increasingly becoming recognized in remote sensing applications involving classification of vegetation (Chan and Paelinckx, 2008; Ghimire et al., 2010; Ismail and Mutanga, 2010; Lawrence et al., 2006). Included in the random forest computation are embedded methods of assessing the generalization error and variable importance measures and the computation does not require tuning of many parameters (Breiman and Cutler, 2004).

2. Data acquisition and methods

2.1. Field spectral data measurements

Field data collection was conducted in the Cathedral Peak region of the Drakensberg Mountain Range, South Africa. The region consists of vegetation divided into altitudinal zones, which correspond closely with the physiographic features of the Drakensberg Mountains (Hill, 1996; Killick, 1963). Three zones namely, the Montane belt (1280 – 1829 m), the Sub-alpine belt (1930 – 2865 m) and the Alpine belt (2866 – 3353 m) are defined. These zones also coincide with three terraces in the Drakensberg. These include the river valley system, the foothills (also known as the Little Berg), and the summit areas, respectively. The sub-alpine belt, which is composed of C_3 and C_4 grass species, is also known as the *Themeda-Festuca* sub-alpine grassland (Hill, 1996). This zone is further divided into three grass communities denoted as the *Themeda triandra, Festuca costata* and the 'Mixed' grasslands. The so-called mixed community consists mainly of variable proportions of C_4 grasses, although the occurrence of *Rendlia altera* seems prevalent. Nonetheless, within the sub-alpine zone there are consociations of the *T. triandra* and the *F. costata* species occurrence on warm, northerly and cool, southerly slopes, respectively.

Reflectance measurements were collected during the December 2010 summer growing season, using a 2150 band (350–2500 nm resolution) Analytical Spectral Device (ASD), field spectroradiometer (FieldSpec®3 ASD, Inc., Boulder, CO, USA). This device uses a fibre optic cable set at 25^o field of view (FOV) to record reflected canopy radiation, which was individually calibrated against a barium sulfate ($BaSO_4$) white reference panel. Canopy reflectance measurements were collected to characterize the spectral separability among 1×1 m sample plots, represented by *F. costata* (C_3), *R. altera* (C_4) and *T. triandra* (C_4) dominant grass species. Spectral reflectance for these dominant grasses (i.e. in the 1×1 m plots) was measured at full canopy cover. Although the dominant grasses co-exist with other species, their respective canopy cover was consistently estimated at ≥ 80% in each target 1×1 m plot.

The field spectral measurements were consistently recorded, considering the recommendations in Thenkabail et al. (2000). The ASD optic sensor was held at about 1.5 m directly above the sampling plots, generating an instantaneous field of view of about 0.35 m^2. A minimum of three positions were randomly chosen within each 1×1 m plot and five spectral measurements were consistently acquired for each one of these positions. This process resulted in a minimum of 15 reflectance measurements per plot. There were no major issues with background effects, since average spectral ($\overline{r}_i$) measurements for each plot ($i \geq 15$) were taken at full canopy cover. A total of 110 plots were measured for each of the three categories of grass species. This process resulted in 330 sample plots, which were considered representative of the spectral variability within and among the grass species under investigation.

2.2. Resampling the spectral data

Spectral resampling of the ASD reflectance was conducted using ENVI's spectral resampling routine (ENVI Version 4.7, 2009 Edition, Copyright © ITT Visual Information Solutions). Initially, an ASCII file containing 10-nm-wide band spacing was created and used to aggregate the 1-nm-wide $\overline{r}_i$ spectral data, across the 400-2500 nm spectrum. The ENVI's resampling routine fits a Gaussian model with an FWHM equal to the specified band spacing to resample the data. This initial sampling of the data was carried to aid calculation of the inter-band correlation coefficient matrix of the input spectral bands. The degree of linear relationship between a band and its shorter and longer wavelength neighbours was calculated, using the well-known Pearson's *r* coefficient of correlation. The linear spectral dependence between two sample wavebands (Xi, Yi) was assessed, resulting in values between 0 and 1:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \qquad \text{(Eq. 1)}$$

where $\frac{X_i - \bar{X}}{s_X}, \bar{X},$ and s_X are the standard score, sample mean, and sample standard deviation, respectively (Eq. 1). The Pearson's *r* correlation coefficient was used, since in this experiment the data assume multimodal normal distribution of the response variables (Chi et al., 2008). The inter-band *r* coefficient matrix was computed using the R statistical software (R Development Core Team, 2010). The R routines output a spreadsheet file format and an x : y axis contour plot of the inter-band *r* values.

Spectral response curves were simulated for the predefined thirteen (13) band-centres. The band-centres were chosen on the basis of their known sensitivity to the biophysical and biochemical characteristics of vegetation. Table 1 shows the causal reflectance or absorption features associated with the chosen bands-centres. In addition, the wavelengths (i.e. band-centres) were selected, considering the observed pattern in the matrix of the data points, as depicted in Fig. 1a. Moreover, the chosen band-centres have been reported in the literature to be useful for C_3 and C_4 grass species discrimination. For example, Smith and Blackshaw, (2003) reported high frequencies (>15 out of 20 times) for the band-centres chosen within the visible and near infrared regions. Irisarri et al. (2009) found that the spectral channels around 820 nm were important for differentiating C_3 and C_4 grass compositions. Further, Noble et al. (2002) noted that the chosen bands-centres in the shortwave region are useful for C_3 and C_4 crop/weed species discrimination.

The inter-band Pearson's *r* correlation coefficient between each of the chosen band-centres and their shorter and longer waveband neighbours was calculated. The band-centres were located at the meeting point of the x: y axis (Fig. 1a), where $r = 1$, and bandwidths were estimated by considering the vertical or the horizontal distance across a given band-centre, as a function of wavelength. It is important to note that the inter-band correlation, *r* values are asymmetrical across the horizontal or the vertical lines. They are only symmetrical across the diagonal. Therefore, it was possible to capture the spectral response information around each band-centre, using a chosen weighting threshold of inter-band correlation. Various weighting thresholds of user-defined inter-band correlation (WTC r = 0.99, WTC r = 0.95 and WTC r = 0.90) were assessed. Fig. 1b illustrates the sizes of the respective user-defined inter-band correlations, using the 660 nm band-centre as an example. The shorter and longer wavelength sides of the sample band-centre (i.e. 660 nm) were calculated on the basis of $r \geq 0.99$, $r \geq 0.95$ and $r \geq 0.90$, whereby bands with coefficients beyond a specified WTC were assigned $r = 0$. The procedure was simulated for all the chosen 13 band-centres.

The resultant user-defined inter-band correlation filter functions were used in the ENVI's spectral resampling routine. The ENVI's routine assumes a critical spectral resampling when FWHM values are not provided by fitting a Gaussian model with a FWHM equivalent to specified band spacing. However, if a user-defined filter function is incorporated, the routine uses it to simulate each line of wavelengths as a multiplicative factor (i.e. weighting between 0 and 1) to resample the data (RSI, 2009). The size of the inter-band correlation for each band-centre varies depending of the chosen WTC and this accounted for the spectral dependence information in the original reflectance data.

In addition, conventional approach of resampling reflectance data to match the response of an existing instrument was assessed for comparative purposes, with the proposed user-defined inter-band correlation filter technique. This involves resampling the ASD data to match the Hyperion sensor's (on-board the Earth Observing-1 Satellite) spectral resolution or FWHM function. The procedure is analogous to that of the user-define spectral resampling described. However, the ENVI routine uses the pre-defined spectral library developed for the Hyperion sensor to resample the data. Since the canopy reflectance measurements were conducted under field conditions, the strong noisy incident radiation in 1350-1460 nm, 1790 -

1960 nm and inconsistent spectra below 400 nm were removed from all analysis (Thenkabail et al., 2004).

2.3. The Random forest's variable importance, classification and accuracy assessment

All the datasets were randomly split into 70% training and 30% holdout test sets (n = 77 and n =33 subsets, respectively), using the R statistical routine. The random forest algorithm was constructed to grow a large ensemble of classification trees. The resultant trees in the ensemble were used to assign each input spectral bands to a class membership of the response variables: *F. costata, R. altera or T. triandra*. Each tree is grown from a randomly and independently selected bootstrap sample of the training data, and about one-third, excluded samples, called the out of bag (OOB) samples were used to calculate an unbiased assessment of the classification accuracy (i.e. the OOB error). Since the OOB error is an unbiased assessment of the classification accuracy (Breiman, 2001; Prasad et al., 2006), it provides theoretical guarantee for the groups of C_3 and C_4 grass species detection and classification. Further to using the OOB error samples to assess the overall classification accuracy, the kappa coefficient analysis was performed. This was necessary, since the study involves a multiclass application and the goal is to account for actual agreement specified by each class versus the chance agreement. That is, it was important to determine if one OOB error matrix is significantly different from another (Stehman, 1997).

The random forest algorithm is easy to implement, because the user tunes only two parameters: (i) the number of trees (*ntree*) to grow and (ii) the number of variables to split at each node (*mtry*). The default value of the *mtry* parameter in the context of classification applications is denoted by the square root of the total number of input variables (Liaw and Wiener, 2002). In the current analysis, the OOB error samples for each class membership of the input spectral bands were used to optimize the *ntree* and *mtry* hyper-parameters (Ismail and Mutanga, 2011).

The random forest-based variable permutation mean decrease in accuracy (Strobl and Zeileis, 2008) was used to calculate the importance of each predictor variable. The variable rankings were calculated using all variables (i.e. 70% training set) and optimize *mtry* value based on the specified *ntree* (i.e.10 000 for all datasets) value. To decrease computing time the routine starts with the default *mtry* value for each dataset and then calculates to the right of the value; and then to the left of the value. For example, the default *mtry* for the resampled Hyperion dataset (n = 197) is 14, so the routine uses *mtry* values of 2, 7… to the left (deflate) of the default value; and *mtry* values of 28, 42… to the right (inflate) of the default *mtry* value. It then runs the random forest based on the optimized *mtry* and *ntree* values and determines variable rankings and the test dataset error.

2.3.1. Random forest-based fast forward variable selection

To calculate the greedy fast forward variable selection (FvS) using the OOB error rates (Adam et al., 2009; Dye et al., 2011), the routine uses the optimized random forest variable rankings calculated above to create different subsets of variables. It involves iteratively fitting the random forest model on the 70% training datasets, and at each iteration building a new model by adding the band with highest importance. Consequently, the routine optimizes the *mtry* and *ntree* values for each step of the variable selection process. To decrease computing time the routine was set to terminate at the iteration with subset OOB error less than the overall OOB error calculated when using all the variables. This can be calculated to include a percentage improvement of the overall OOB error. However, because the input Hyperion bands (n = 197) are high in dimension (more than 100 predictors variables), the percent improvement approach was not exploited for the current application.

Several empirical evidences have shown that the random forest algorithm shows significant preference towards highly correlated predictor variables (Nicodemus et al., 2010; Strobl et al., 2008). The authors reported that traditional random forest's preference for highly-correlated predictor variables can be carried forward to any significance test or variable selection processes constructed from the importance measures. In this respect, researchers have suggested the use of conditional variable importance approach to mitigate the problems associated with traditional random forest variable selection process. Despite the recommendation, in the current experiment assesses the random forest-based FvS process on the highly dimensional resampled Hyperion bands, for comparative purposes with the WTC datasets.

3. Results

3.1. The user-defined inter-band correlation filter technique of spectral resampling

The results obtained indicate that large portions of the C_3 and C_4 grass canopy reflectance exist in highly correlated wavelengths. In general, a decrease in spectral resolutions was observed for each of the 13 band-centres in relation to the user-defined weighting thresholds of inter-band correlation filter: $r = 0.99$, r = 0.95 and r = 0.90, respectively. For each derived waveband, the inter-band correlation coefficient $r = 1$ at the band-centres and generally decreases across the shorter or longer wavelengths neighbours, as quantified by a chosen WTC filter. Overall, the WTC $r = 0.99$ filter yielded higher spectral resolutions, among the three filters assessed. In addition, it appeared that the band-centres of specific regions (i.e. the visible, red-edge, near infrared and shortwave infrared spectra) showed varying degrees of inter-band correlations, resulting in variable spectral resolutions, for each of the spectral regions. Table 2 shows the results of the various spectral resolutions obtained for each band-centre of the WTC datasets.

3.2. C_3 and C_4 grass species classification using the WTC and resample Hyperion datasets

The random forest hyper-parameters were optimized using the OOB error rates. The WTC, r = 0.99 filter yielded the highest classification accuracy among the three user-defined thresholds of inter-band correlation filters assessed. Table 3 shows the overall accuracies (OOB error rates) and the kappa coefficients obtained for all datasets, including the resampled Hyperion. The OOB error and kappa coefficients increased substantially when the WTC of r = 0.90 filter dataset was analyzed. However, the results obtained showed that the proposed user-defined thresholds of inter-band correlation filter approach to resampling Hyperspectral data produced higher classification accuracies when compared with the conventional technique of resampling data to match the Hyperion sensors spectral resolution.

Although the WTC datasets yielded significantly variable classification accuracies, a similar variable importance ranking was obtained for these datasets. Consequently, only the variable importance ranks of the WTC r = 0.99 dataset is presented in Fig. 2. Since the random forest algorithm was initially run using all the resampled Hyperion bands (n =197), the variable importance measure (Fig. 3) was then exploited to evaluate whether the FvS process could improve the classification accuracy. This procedure yielded an optimal subset of 22 bands (Table 4), which were subsequently used to classify the C_3 and C_4 response variables. The results obtained showed that the resampled Hyperion band B7 yielded the highest mean decrease in accuracy (11.13%), and that was subsequently carried over to the variable (1/197 bands) selection process. It should be noted that only the optimum subset of the best ranked bands are reported in Table 4. Inter-band correlation coefficient matrix (Table 5) of the optimal subset of bands selection through the random forest-based FvS process was exploited. As expected, the results showed clearly that the FvS procedure yielded highly correlated resampled Hyperion bands and that the majority of the selected bands were concentrated in specific regions of the spectrum.

4. Discussion

Hyperspectral data are suitable for C_3 and C_4 grass species classification, since spectral fine features characteristic of vegetation is more discernible from narrowband sensors. However, it can be challenging to classify C_3 and C_4 grass species using their spectral reflectance data, due to the problem of hyper-dimensionality and associated multicollinearity phenomena (Pal and Foody, 2010). Among other factors (e.g. Phenological effects and solar illumination conditions), spectral similarity between C_3 and C_4 grasses and their co-existing species can have significant impacts on the classification capability of canopy reflectance data (Schmidt and Skidmore, 2001). Despite these challenges, recent studies have shown that subtle differences in structural and physiological properties, such as those described for C_3 and C_4 grasses may be detected by leaf or canopy reflectance (Irisarri et al., 2009; Liu and Cheng, 2011).

Although previous studies have used narrow-band spectral data to classify grasslands of C_3 and C_4 species composition, the present investigation explores the potential use of a user-defined inter-band correlation filter to resample hyperspectral data, for subsequent classification analysis. This study demonstrated the trade-offs between retaining narrow bands spectral data vs. the optimal reduction in spectral dimensionality for improved classification. The results obtained explained the spectroscopic interpretability of the chosen band-centres, in reference to their sensitivity to leaf or canopy surface properties, internal structure and biochemical concentrations. These characteristic features are known to significantly vary between C_3 and C_4 grass species. Hence, variations in pigments content, nitrogen, carbon compounds (lignin and fibre) and water components (inter-cellular air-moisture or leaf liquid water content) can be attributed to the good spectral separability obtained for the target grasses assessed in this study. In general, the results have shown that the proposed user-defined inter-band correlation filter technique yielded improved classification of *F. costata*, *R. altera* and *T. triandra* grass canopies. More detailed analyses of the results are presented next: the weighting thresholds of inter-band correlation filter approach to resampling hyperspectral data; the random forest classification and band subset selection of resampled Hyperion dataset, using a traditional method vs. prior dimensionality-reduction, using the WTC filter technique and; implications of the present investigation for applications involving C_3 and C_4 grass species.

4.1. The weighting thresholds of inter-band correlation filter approach

On the basis of the proposed resampling approach, this study has shown that highly correlated hyperspectral wavebands in specific regions can be optimally aggregated to reduce spectral dimension of the input spectral bands. The proposed spectral resampling technique takes advantage of the inherent property of vegetation reflectance, the asymmetrical nature of the inter-band correlation matrix of the collected wavebands. The resulted presented in Table 2 show the extent of spectral convolution using the highly correlated wavelengths around each of the selected band-centres across the 400 – 2500 nm spectrum. The vegetation spectral response property used to calculate the various WTC *r* values (i.e. 0.99, 0.95 and 0.90) can be attributed to reflectance or absorption features characteristic of the target C_3 and C_4 grasses (Ferwerda et al., 2005; Knox et al., 2010). In a previous study, Slanton et al. (2001) found 800 nm wavelength contained very strong discriminating power for plant species at the level of leaf internal structure. Further, Irisarri et al. (2009) reported that vegetation reflectance at the 820 nm spectral range is sensitive to even subtle differences among grass species or between groups of C_3 and C_4 grasses. Hence, in the present experiment, the use-defined inter-band WTC filters were used to assess the optimal spectral resolutions around chosen band-centres, including the 820 nm wavelength. In this regard, the proposed resampling procedure offers the potential for data dimensionality reduction and optimizes elimination of redundant spectral information by means of weighting thresholds of inter-band correlation criterion.

It is worth noting that the newly introduced resampling approach not only reduces dimensionality in hyperspectral data, it also preserves relevant spectral information for posterior classification of C_3 and C_4 grass canopy reflectance. In general, there is very close relationship between classifier sensitivity to data dimensionality and classification accuracy, under conditions of multiple correlations among the input spectral bands (Gomez-Chova et al., 2003). This suggests the concept of using spectral resampling techniques capable of reducing co-linearity problems in the input spectral space, for applications involving C_3 and C_4 grass species.

4.2. Random forest-based band subset selection vs. prior dimensionality-reduction

Random forests have been found attractive for the analysis of remotely sensed data for ecological applications (Chan and Paelinckx, 2008; Ham et al., 2005; Lawrence et al., 2006; Prasad et al., 2006). A number of studies have asserted that the method is insensitive to high-dimensionality and, therefore, does not require a dimensionality-reduction analysis in pre-processing (Breiman and Cutler, 2004; Ham et al., 2005). However, the assessment of random forest's variable importance measure in high-dimensional spectral space, has revealed that the algorithm thus show a preference to highly correlated predictor variable. Such a preference was also found to be manifest in the subsequent subset band selection process (Table 5). The results from the present experiment thus reaffirm the findings of the recent studies, which investigated random forests variable importance under predictor correlation and the generalization of parameter estimates (Nicodemus and Shugart, 2007; Strobl et al., 2008). In their studies, the authors recommended conditional variable importance approach for random forest-based variable selection and posterior classifications.

Critically, the results obtained from the present study showed that WTC r = 0.99 yielded the highest classification accuracy (kappa = 0.82) among the three inter-ban correlation thresholds assessed. This superior accuracy demonstrates clearly, the role of spectral resolutions on C_3 and C_4 grass classification and the classifier accuracy. However, the larger decrease in classification accuracy obtained for WTC $r = 0.90$ could be attributed to the very larger increase in wavelengths for each individual waveband, as represented in Table 2. The trend obtained among the classification of the three WTC *r* datasets compares well with Dalponte et al. (2009), who investigated the effect of changing spectral resolution upon different classifiers for forest applications. In their study, the authors found that as spectral resolutions were degraded from 4.6 nm to 36.8 nm, overall kappa accuracies dropped from ~ 89 % to ~ 84 %, respectively, using Support vector machines (SVM) algorithm (Vapnick, 1998). Furthermore, when compared with classification involving a simple parametric classifier such as LDA, Dalponte et al. (2009) recorded inferior kappa accuracies which also dropped from ~ 77 % to ~54 %, respectively. The authors concluded that advanced non-parametric classifiers, such as the random forest are more applicable for classifications involving complex vegetation feature spaces.

As depicted on Table 5, the random forest band selection process showed a significant bias toward the highly correlated Hyperion bands (e.g. B6 - B13 and B198 - B219). However, the random forest analysis on the prior dimensionality-reduction datasets offered distinct advantage, using the inter-band correlation WTC filters to aggregate the majority of the highly correlated wavebands. The novelty of the proposed method is that the bands contributing to the out reflectance data were weighted according to their linear relationship with a chosen band-centre. The resultant classification accuracies showed that the prior dimensionality-reduction approach considerably negates problems associated with spectral redundancy and thereby mitigated against the multicollinearity phenomenon (Gomez-Chova et al., 2003). Furthermore, the present experiment has demonstrated that even when a large training sample (i.e. 330 canopy spectra), compared to the number of spectral bands (n = 197) are used, spectral filtering may still be useful. This affirmation is supported by the accuracy derived from the use of the optimized 13 bands, which yielded superior classification accuracies (OOB = 0.14; kappa 0.82), compared to that derived from the use of a larger but high-correlated resampled Hyperion bands (OOB = 0.19; kappa = 0.76).

4.3. Implications of the present investigation and conclusion

The primary purpose of this study was to assess the spectral separability among C_3 and C_4 grasses, sampled from the Drakensberg Mountains of South Africa. The secondary goal was to address the issue of multicollinearity effect on the performance of the random forest variable importance and the subsequent band subset selection process under predictor correlation. The performance of the method, when applied to data derived by resampling spectra to the Hyperion sensor's spectral resolution, was compared to that of spectra resampled by weighting the inter-band correlations, as a function of wavelength. The overall implications for this investigation are related to various hyperspectral data application constraints: i) the trade-off between the number of spectral bands and the resolution of remotely sensed imagery; ii) the trade-off between higher spectral resolution and reduced signal-to-noise ratio, and iii) challenges associated with the optimal configuration of wavebands capable of providing sensitive information about a target vegetation (Price, 1994; Thenkabail et al., 2004). Therefore, in the present experiment, a technique has been proposed to reduce dimensionality, while preserving relevant spectral information for posterior classification task. It has been observed that the proposed resampling technique represents a potential method of reprogramming hyperspectral resolutions and band configurations. This potential also holds prospects in the development and configuration of future remote sensors to collect optimal spectral resolutions and configuration for specific vegetation applications. The results obtained in this study suggested that further studies addressing multicollinearity problem should consider techniques that account for the spectral dependence information contained in vegetation reflectance data. In summary, the current technique described in this paper yields the following distinct benefits:

- Reduces data dimensionality by accounting for the inter-band correlations around specific band-centres of interest and thereby mitigating against the multicollinearity phenomenon caused by highly correlated spectral bands.
- Optimizes the spectral resolutions useful for the separability among the dominant C_3 and C_4 grass species investigated.
- Assists the random forest, to achieve improved classification accuracy, thereby providing the potential to link each individual input band to the physical meaning of interaction effects in the structure of the acquired data.

5. Acknowledgements

Support was provided by the National Research Foundation (NRF), the KwaZulu-Natal Department of Agriculture and Environmental Affairs (KZNDAE) and the Ezemvelo KZN Wildlife.

6. References:

Adam, E., Mutanga, O., Rugege, D. and Ismail, R., 2009. Field spectrometry of papyrus vegetation (*Cyperus papyrus L.*) in swamp wetlands of St Lucia, South Africa Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009 Cape Town, pp. IV- 260 – IV- 263.

Becker, B.L., Lusch, D.P. and Qi, J., 2007. A classification-based assessment of the optimal spectral and spatial resolutions for Great Lakes coastal wetland imagery. Remote Sensing of Environment, 108(1): 111-120.

Breiman, L., 2001. Random Forests. Machine Learning, 45(1): 5-32.

Breiman, L. and Cutler, A., 2004. Random forests tools for predicting and understanding data, Interface workshop-April 2004.

Bruzzone, L., Roli, F. and Serpico, S.B., 1995. An extension to multiclass cases of the Jeffreys-Matusita distance. IEEE Transactions on Geoscience and Remote Sensing, 33(6): 1318-1321.

Bruzzone, L. and Serpico, S.B., 2000. A technique for feature selection in multiclass problems. International Journal of Remote Sensing, 21(3): 549-563.

Chan, J.C.-W. and Paelinckx, D., 2008. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment, 112(6): 2999-3011.

Chi, M. and Bruzzone, L., 2007. Semisupervised classification of hyperspectral images by SVMs optimized in the primal. Geoscience and Remote Sensing, IEEE Transactions on, 45(6): 1870-1880.

Chi, M., Feng, R. and Bruzzone, L., 2008. Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. Advances in Space Research, 41(11): 1793-1799.

Clevers, J.G.P.W., van der Heijden, G.W.A.M., Verzakov, S. and Schaepman, M.E., 2007. Estimating grassland biomass using SVM band shaving of hyperspectral data. Photogrammetric Engineering & Remote Sensing, 73(10): 1141-1148.

Dalponte, M., Bruzzone, L., Vescovo, L. and Gianelle, D., 2009. The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. Remote Sensing of Environment, 113(11): 2345-2355.

Djouadi, A., Snorrason, O. and Garber, F., 1990. “The quality of training-sample estimates of the Bhattacharyya coefficient”. IEEE Transactions on Pattern analysis and machine intelligence, 12(1): 92-97.

Dye, M., Mutanga, O. and Ismail, R., 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict *Pinus patula* age in KwaZulu-Natal, South Africa. Geocarto International, 26(4): 275-289.

Faurtyot, T. and Baret, F., 1997. Vegetation water and dry matter contents estimated from top-of-the-atmosphere reflectance data: A simulation study. Remote Sensing of Environment, 61(1): 34-45.

Ferwerda, J.G., Skidmore, A.K. and Mutanga, O., 2005. Nitrogen detection with hyperspectral normalized ratio indices across multiple plant species. International Journal of Remote Sensing, 26(18): 4083-4095.

Gamon, J.A., Serrano, L. and Surfus, J.S., 1997. The photochemical reflectance index: an optical indicator of photosynthetic radiation use efficiency across species, functional types, and nutrient levels. Oecologia, 112(4): 492-501.

Ghimire, B., Rogan, J. and Miller, J., 2010. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sensing Letters, 1(1): 45-54.

Gomez-Chova, L., Calpe, J., Camps-Valls, G., Martin, J.D., Soria, E., Vila, J., Alonso-Chorda, L. and Moreno, J., 2003. Feature selection of hyperspectral data through local correlation and SFFS for crop classification, Geoscience and Remote Sensing Symposium, 2003. IGARSS '03. Proceedings. 2003 IEEE International, pp. 555-557.

Ham, J., Yangchi, C., Crawford, M.M. and Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 43(3): 492-501.

Hill, T.R., 1996. Description, classification and ordination of the dominant vegetation communinities, Cathedral Peak, KwaZulu-Natal Drakensberg. South African Journal of Botany, 62(5): 263-269.

Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. Information Theory, IEEE Transactions on, 14(1): 55-63.

Irisarri, J.G.N., Oesterheld, M., Verón, S.R. and Paruelo, J.M., 2009. Grass species differentiation through canopy hyperspectral reflectance. International Journal of Remote Sensing, 30(22): 5959 - 5975.

Ismail, R. and Mutanga, O., 2010. A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa. International Journal of Applied Earth Observation and Geoinformation, 12: S45-S51.

Ismail, R. and Mutanga, O., 2011. Discriminating the early stages of *Sirex noctilio* infestation using classification tree ensembles and shortwave infrared bands. International Journal of Remote Sensing, 32(15): 4249-4266.

Killick, D.J.B., 1963. An account of the plant ecology of the Cathedral Peak area of the Natal Drakensberg. Memoirs of the Botanical Survey of South Africa, 34.

Knox, N.M., Skidmore, A.K., Schlerf, M., de Boer, W.F., van Wieren, S.E., van der Waal, C., Prins, H.H.T. and Slotow, R., 2010. Nitrogen prediction in grasses: effect of bandwidth and plant material state on absorption feature selection. International Journal of Remote Sensing, 31(3): 691-704.

Lawrence, R., Wood, S. and Sheley, R., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). Remote Sensing of Environment, 100(3): 356-362.

Liaw, A. and Wiener, M., 2002. Classification and regression by random forest. R News, 2/3: 18-22.

Liu, L. and Cheng, Z., 2011. Mapping C_3 and C_4 plant functional types using separated solar-induced chlorophyll fluorescence from hyperspectral data. International Journal of Remote Sensing, 32(24): 9171-9183.

Mutanga, O., Skidmore, A.K. and Prins, H.H.T., 2004. Predicting in situ pasture quality in the Kruger National Park, South Africa, using continuum-removed absorption features. Remote Sensing of Environment, 89(3): 393-408.

Nicodemus, K., Malley, J., Strobl, C. and Ziegler, A., 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics, 11(1): 110.

Nicodemus, K. and Shugart, Y., 2007. Impact of linkage disequilibrium and effect size on the ability of machine learning methods to detect epistasis in case-control studies. Abstract volume of the Sixteenth Annual Meeting of the International Genetic Epidemiology Society, North Yorkshire, UK, 31(6): 611.

Oyarzabal, M., Paruelo, J.M., Pino, F.d., Oesterheld, M. and Lauenroth, W.K., 2008. Trait differences between grass species along a climatic gradient in South and North America. Journal of Vegetation Science, 19(2): 183-192.

Pal, M. and Foody, G.M., 2010. Feature selection for classification of hyperspectral data by SVM. IEEE Transactions on Geoscience and Remote Sensing, 48(5): 2297-2307.

Prasad, A.M., Iverson, L.R. and Liaw, A., 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems, 9(2): 181-199.

Price, J.C., 1994. How unique are spectral signatures? Remote Sensing of Environment, 49(3): 181-186.

Pudil, P., Novovičová, J. and Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters, 15(11): 1119-1125.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RSI, 2009. ENVI (Environment for Visualizing Images) Online Help. ITT Visual Information Solutions. Boulder, USA.

Schmidt, K.S. and Skidmore, A.K., 2001. Exploring spectral discrimination of grass species in African rangelands. International Journal of Remote Sensing, 22: 3421-3434.

Schmidt, K.S. and Skidmore, A.K., 2003. Spectral discrimination of vegetation types in a coastal wetland. Remote Sensing of Environment, 85: 92-108.

Serpico, S.B. and Bruzzone, L., 2001. A new search algorithm for feature selection in hyperspectral remote sensing images. Geoscience and Remote Sensing, IEEE Transactions on, 39(7): 1360-1367.

Smith, A.M. and Blackshaw, R.E., 2003. Weed–crop discrimination using remote sensing: a detached leaf experiment1. Weed Technology, 17(4): 811-820.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment, 62(1): 77-89.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics, 9(1): 307.

Strobl, C. and Zeileis, A., 2008. Danger: high Power! – exploring the statistical properties of a test for random forest variable importance. Proceedings in Computational Statistics, 2: 59-66.

Su, H., Kanemasu, E.T., Ransom, M.D. and Yang, S.-S., 1990. Separability of soils in a tallgrasss prairie using SPOT and DEM data. Remote Sensing of Environment, 33(3): 157-163.

Thenkabail, P.S., Enclona, E.A., Ashton, M.S. and Van Der Meer, B., 2004. Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. Remote Sensing of Environment, 91(3-4): 354-376.

Tieszen, L.L., Reed, B.C., Bliss, N.B., Wylie, B.K. and DeJong, D.D., 1997. NDVI, C_3 and C_4 Production, and Distributions in Great Plains Grassland Land Cover Classes. Ecological Applications, 7(1): 59-78.

Ustin, S.L. and Gamon, J.A., 2010. Remote sensing of plant functional types. New Phytologist, 186(4): 795-816.

Vapnick, V.N., 1998. Statistical learning theory. John Wiley and Sons Inc.

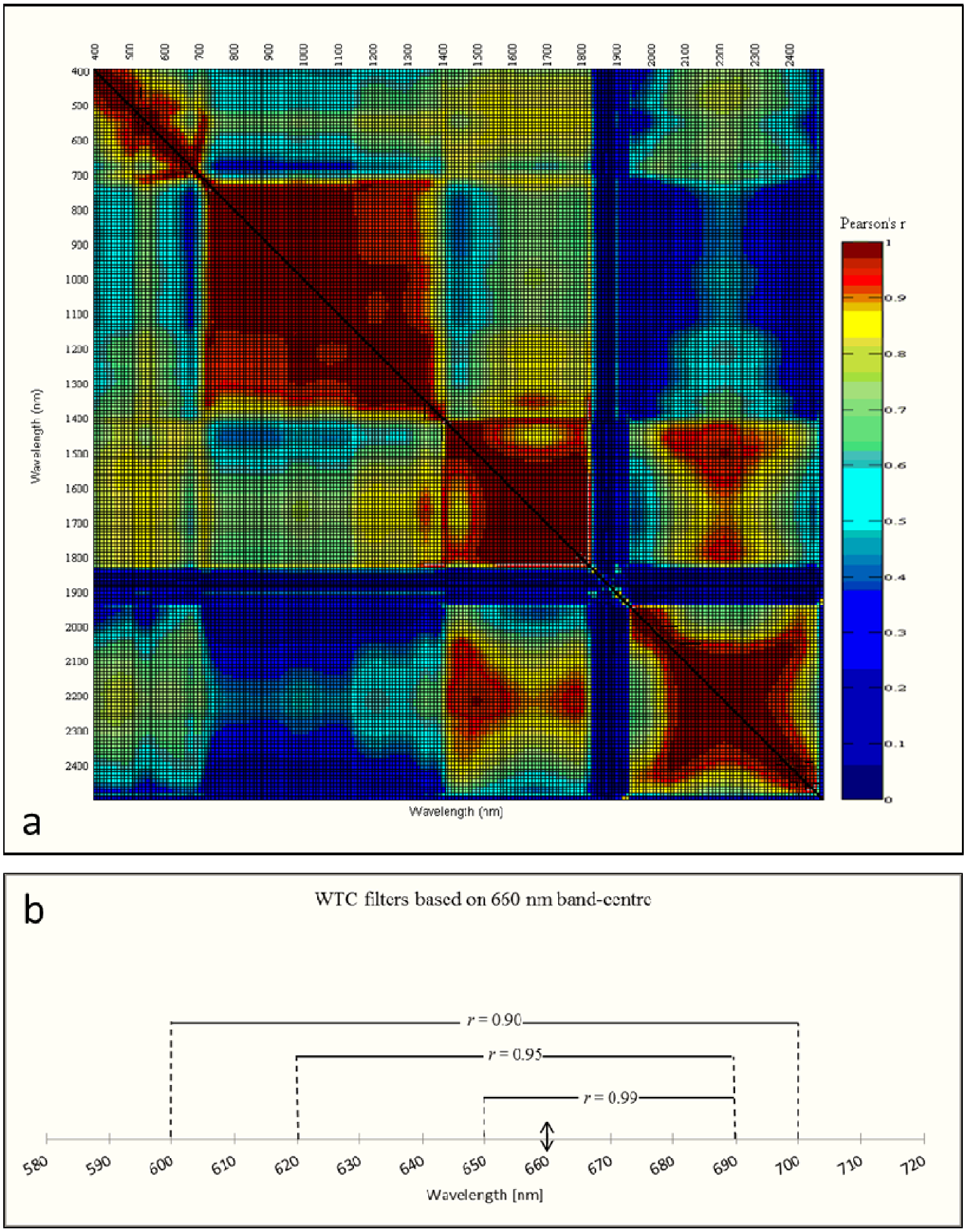
Zhang, J., Wu, J. and Zhou, L., 2011. Deriving vegetation leaf water content from spectrophotometric data with orthogonal signal correction-partial least square regression. International Journal of Remote Sensing: 1-18.

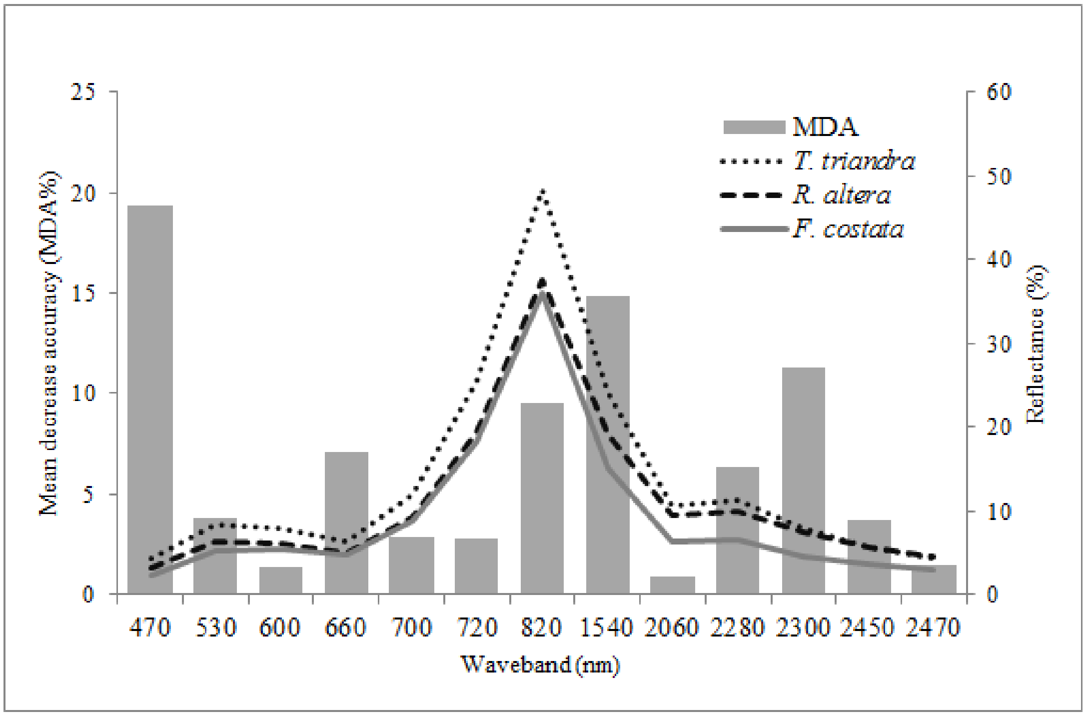
Figure captions

Fig. 1 Pearson's r correlation coefficients matrix (plot) of the input spectral bands, calculated using reflectance data aggregated into 10-nm-wide band intervals (a) and an illustration of the user-defined inter-band correlation filter for 660 nm band-centre (b).

Fig. 2 Random forests variable importance ranks for the WTC r = 0.99 dataset (n = 13 bands) based on the Mean Decrease in Accuracy values. The reflectance spectrum of the target grass species is shown.

Fig. 3 Random forests variable importance ranks for resampled Hyperion bands (n = 197) based on the Mean Decrease in Accuracy values. The reflectance spectrum of the target grass species is shown.


a
Pearson's r
Wavelength (nm)
Wavelength (nm)
b
WTC filters based on 660 nm band-centre
r = 0.90
r = 0.95
r = 0.99
Wavelength [nm]


MDA
T. triandra
R. altera
F. costata
Mean decrease accuracy (MDA%)
Reflectance (%)
470 530 600 660 700 720 820 1540 2060 2280 2300 2450 2470
Waveband (nm)

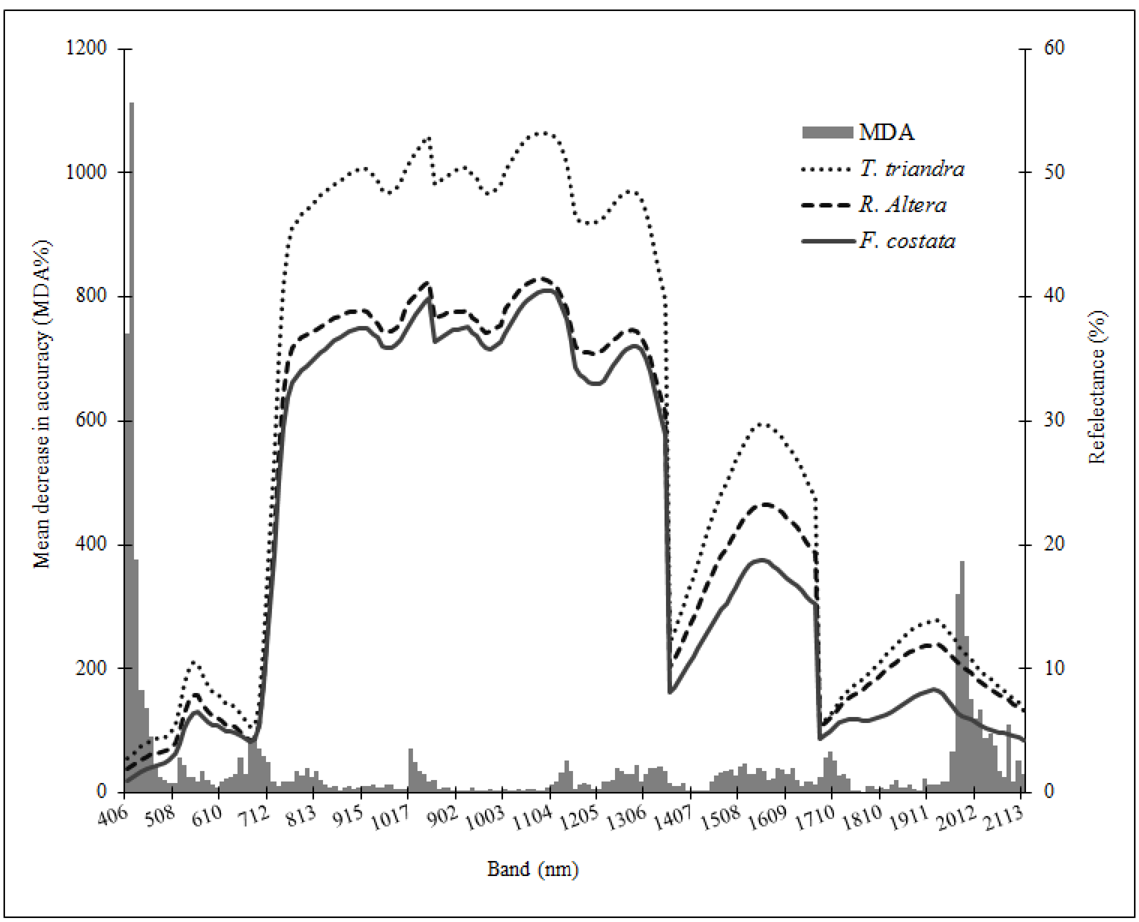

MDA
T. triandra
R. Altera
F. costata
Mean decrease in accuracy (MDA%)
Refelectance (%)
406 508 610 712 813 915 1017 902 1003 1104 1205 1306 1407 1508 1609 1710 1810 1911 2012 2113
Band (nm)

Table 1
Wavelengths corresponding to known absorption features, as described in previous studies to be highly sensitive to the properties of reflection or absorption of vegetation structural and biochemical characteristics.

No.	Band (nm)	Known Causal compound/ feature	Source
1	470	Total plant pigment concentration	Blackburn, 1998
2	530	Chlorophyll a absorption	Gamon et al., 1997
3	600	Nitrogen	Faurtyot and Baret, 1997
4	660	Nitrogen	Curran, 1989
5	700	Total Chlorophyll, Nitrogen	Carter, 1994
6	720	Total Chlorophyll, Leaf mass	Horler et al., 1983
7	820	Leaf mass, Leaf area index	Curran, 1989
8	1540	Cellulose, vegetation water content	Carter, 1994
9	2060	Protein	Carter, 1994
10	2280	Cellulose, Sugar, Starch, Leaf mass	Curran, 1989
11	2300	Leaf mass, vegetation water content	Carter, 1994
12	2450	Cellulose, Protein, Nitrogen	Carter, 1994
13	2470	Cellulose, Protein	Kumar et al., 2001

Table 2
Spectral resolutions obtained from the analysis of the user-defined inter-band correlation.

	Wavelength (nm) filter size/ Resampled datasets		
Band centre (nm)	WTC $r = 0.99$	WTC $r = 0.95$	WTC $r = 0.90$
470	430 - 500	410 - 520	400 – 530
530	520 - 570	500 - 600	470 – 600
600	580 - 630	550 - 650	530 – 660
660	650 - 690	620 - 690	600 – 700
700	700*	690 - 710	690 – 710
720	720*	710 - 730	710 – 740
820	740 - 1110	730 - 1190	730 – 1330
1540	1500 - 1630	1470 - 1780	1470 – 1780
2060	2040 - 2090	2020 - 2180	1990 – 2280
2280	2100 - 2300	2060 - 2300	2060 – 2300
2300	2300 - 2350	2280 - 2390	2289 – 2420
2450	2450*	2410 - 2460	2360 – 2470
2470	2470*	2460 - 2470	2450 – 2470

* Indicates wavelength = 10 nm of the input spectral band.

Table 3

Random forest model optimization and accuracy measures using OOB samples on the training dataset (70% sample). The Kappa-test set statistics were calculated using 30 % holdout samples.

Resampled datasets	Number of Bands	Optimized *mtry*	Optimized *ntree*	OOB error rate	Kappa Training set	Kappa-Test set
Resampled-Hyperion	197	56	500	0.19	0.71	-
Resampled-Hyperion	22	20	1500	0.19	0.72	0.76
WTC r = 0.99	13	6	4000	0.14	0.79	0.82
WTC r = 0.95	13	3	2500	0.21	0.68	0.79
WTC r = 0.90	13	9	4000	0.23	0.64	0.76

Table 4

Random forest-based forward best ranked band selection on resampled-Hyperion spectral resolution. The Kappa statistics were calculated on 70 % training sample.

Rank	Hyperion band	Average wavelength (nm)	FWHM (nm)	Optimized *mtry*	Optimized *ntree*	Accuracy: cumulative OOB
1	B7	416.64	11.39	1	500	0.518
2	B6	406.46	11.39	2	10000	0.595
3	B212	2274.42	10.43	3	1000	0.693
4	B211	2264.32	10.44	4	1500	0.68
5	B8	426.82	11.39	4	500	0.693
6	B213	2284.52	10.42	4	7500	0.699
7	B9	436.99	11.39	7	1500	0.706
8	B216	2314.81	10.41	1	7000	0.699
9	B214	2294.61	10.41	6	500	0.706
10	B215	2304.71	10.41	6	2500	0.693
11	B217	2324.91	10.41	9	1000	0.693
12	B10	447.17	11.39	12	500	0.706
13	B210	2254.22	10.46	1	500	0.699
14	B12	467.52	11.39	8	500	0.712
15	B218	2335.01	10.41	2	1000	0.725
16	B11	457.34	11.39	4	500	0.725
17	B219	2345.11	10.41	4	6500	0.732
18	B33	681.2	10.33	18	500	0.771
19	B198	2133.24	10.73	8	500	0.771
20	B200	2153.34	10.68	4	1000	0.764
21	B158	1729.7	11.56	20	1000	0.803
22	B13	477.69	11.39	20	1500	0.81

Table 5
High correlated variables from random forest's importance rank and selection process.

	Resampled Hyperion Wavelength (nm)																					
	406.46	416.64	426.82	436.99	447.17	457.34	467.52	477.69	681.2	1729.7	2133.24	2153.34	2254.22	2264.32	2274.42	2284.52	2294.61	2304.71	2314.81	2324.91	2335.01	2345.1
406.46	1																					
416.64	0.99	1																				
426.82	0.99	0.99	1																			
436.99	0.99	0.99	0.99	1																		
447.17	0.99	0.99	0.99	0.99	1																	
457.34	0.98	0.99	0.99	0.99	0.99	1																
467.52	0.98	0.98	0.99	0.99	0.99	0.99	1															
477.69	0.97	0.98	0.99	0.99	0.99	0.99	0.99	1														
681.2	0.74	0.77	0.79	0.81	0.82	0.83	0.85	0.87	1													
1729.7	0.87	0.88	0.89	0.9	0.9	0.9	0.91	0.91	0.78	1												
2133.24	0.85	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.78	0.93	1											
2153.34	0.85	0.86	0.86	0.87	0.87	0.88	0.88	0.88	0.79	0.94	0.99	1										
2254.22	0.86	0.87	0.87	0.87	0.88	0.88	0.88	0.88	0.77	0.95	0.99	0.99	1									
2264.32	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.76	0.94	0.99	0.99	0.99	1								
2274.42	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.76	0.93	0.99	0.99	0.99	0.99	1							
2284.52	0.85	0.86	0.86	0.86	0.86	0.87	0.87	0.87	0.76	0.93	0.99	0.99	0.99	0.99	0.99	1						
2294.61	0.85	0.85	0.86	0.86	0.86	0.86	0.86	0.86	0.77	0.93	0.99	0.99	0.99	0.99	0.99	0.99	1					
2304.71	0.84	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.76	0.92	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1				
2314.81	0.83	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.76	0.91	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1			
2324.91	0.83	0.83	0.84	0.84	0.84	0.84	0.84	0.84	0.76	0.9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1		
2335.01	0.82	0.82	0.83	0.83	0.83	0.83	0.84	0.84	0.76	0.9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	
2345.11	0.81	0.82	0.82	0.82	0.83	0.83	0.83	0.83	0.76	0.89	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1