 doi:10.20944/preprints201701.0001.v1

Vegetation Water Content Prediction: Towards More Relevant Explicatory Waveband Variables

Eric Ariel L. Salas*[1]
[1] Department of Fish, Wildlife and Conservation Ecology, New Mexico State University, Las Cruces, NM 88003, USA

* Author to whom correspondence should be addressed:
E-mail: easalas@nmsu.edu
Tel.: +1-575-646-2691

Abstract

Assessing vegetation water content (VWC) from hyperspectral reflectance dataset poses two foremost questions: what specific wavebands of the SWIR offer a good retrieval and what modeling methods have the best predictive ability. In this paper, we explored the application of multivariate statistical techniques such as stepwise multiple linear regression (SMLR) and partial least square regression (PLSR) for vegetation water content prediction using the absorption features. We also examined the recursive partitioning model of the dataset to illustrate relationships of splits among waveband predictors. Previously known wavelength features around 970 nm and 900 nm were leading predictors of water content. In the absence of actual field VWC data, the absorption area feature at 970 nm was computed and used to model other explicatory waveband variables that may boost the prediction. The results of this exploratory waveband-predictor study highlighted other essential locations around 956 nm, 922 nm, 976 nm, 935 nm, and 915 nm. The SMLR disqualified highly correlated bands that augment relatively little in the VWC prediction capability of the model. PLSR presented the 900 nm, 922 nm, and the 970 nm peaks affirming the results of the SMLR. The PLSR is the favored technique with RMSEcv = 0.002 (r^2=0.88) for the cross-validation, lower than the SMLR (RMSEcv = 0.023). Recursive partitioning method showed the 956 nm, surprisingly with the highest logworth among predictors. The overall r^2 after partitioning, when actual and predicted VWC were plotted against each other was a fair 0.59. This value is comparable with other empirical indices we previously analyzed. Recursive partitioning is a highly adaptive technique and care must be taken in the interpretation of results.

Keywords: vegetation water content, stepwise multiple linear regression, water content prediction, recursive partitioning, hyperspectral analysis

 doi:10.20944/preprints201701.0001.v1

Introduction

Reflectance spectra could present many possible water indices because of the existence of several water absorption features in the NIR and far-IR region of the spectrum (Serrano et al. 2000). Wavelength locations for water absorption features can be found at approximately 970, 1200, 1450, and 1950 nm in a vegetation spectrum (Carter, 1991; Kokaly and Clark, 1999; Tian et al., 2001; and Claudio et al., 2006).

Specific spectral positions have been used by different authors to foresee interesting possibilities of deriving information on canopy water contents: the ~1550 – 1750 nm (Tucker, 1980); 980 nm (Goetz et al., 1990); 960 nm (Roberts et al., 1997); 760 nm, 970 nm, 1190 nm, 1450 nm, and 1940 nm (Kokaly and Clark, 1999); 1430 nm and 1950 nm (Penuelas and Inoue, 1999); 970 nm and 1240 nm (Serrano et al, 2000); 970 nm, 1200 nm, and 1530 nm (Sims and Gamon, 2003); 820 nm and 1600 nm (Hunt and Rock, 1989 and Riggs and Running, 1991). The wavelength at 970 nm has been widely used, usually in combination of other wavelengths, in the estimation of vegetation water content.

Vegetation water indices typically are derived using ratio and difference of wavebands. A list of indices for canopy VWC estimation is shown in appendix 1. Still, researchers like Sims and Gamon (2003) have searched for the optimal mathematical statement and wavelengths that would tender a good measure of the water vegetation capacity.

As early as the 90s, Goetz et al. (1990) used the Spectral Curve Fitting technique to derive subtle information from vegetation spectra for biochemical constituents such as lignin. Linear least squares spectrum matching technique was employed by Gao and Goetz (1995) to retrieve equivalent water thickness (EWT) using AVIRIS imagery. A number of studies used empirical approaches that integrate spectral information of spectral wavelengths in assessing vegetation biophysical and biochemical properties (Kokaly and Clark, 1999; Lefsky et al., 2001; De Jong et al., 2003; Cho et al., 2007). Examples of empirical approaches are the univariate and multivariate regression models. In view of multivariate regression, partial least squares regression (PLSR) and stepwise multiple linear regression (SMLR) can be utilized to find the relationship between a target parameter (in this case the VWC absorption) and the spectral reflectance (in this case the hyperspectral reflectance data). SMLR serves as an exploratory tool to single out the potentially important predictors. If two independent variables are highly correlated, only one will end up in the model in a stepwise analysis, even though either one can be considered as a predictor. As one of our objectives, we examined the application of SMLR to quantify VWC at canopy level hyperspectral reflectance. PLSR, like the SMLR, can be used as an exploratory analysis tool to select appropriate predictors and to spot outliers before classical linear regression. However, PLSR is perhaps the least limiting of the several multivariate extensions of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

multiple linear regression models. PLSR accommodates all available spectral wavelengths simultaneously and studies such as Cho et al. (2007) and Nguyen and Lee (2006) used the potentials of the technique for estimating biophysical and biochemical properties of vegetation.

Infrared spectroscopy placed side by side with partial least squares and stepwise multiple linear regressions in predicting VWC using high spectral resolution data is not widely entertained by the scientific world so far. To defeat the curse of dataset dimensionality, one has to capture general trends in the dataset while eliminating extraneous information. This is done by employing predictive modeling techniques – a simple decision rule, for instance. The technique partitions the hyperspectral dataset into several cases based on the target VWC values. Groups unrelated to the target are considered less worth. Recursive partitioning, also known as CART, is used in this study to perform the data splitting, and then we interpreted results.

The number of times the recursive partitioning (splitting) process repeats can be thought of as a tuning parameter for the model. Each iteration subdivides the input dataset further and fosters training data accuracy. The model could give high degree of accuracy but also has the tendency to overfit (Vayssieres et al., 2000). To solve the over fitting tendency, a cross-validation procedure can be applied.

Hyperspectral dataset obtained by a field spectrometer include hundreds of narrow wavelengths that may not all be necessary for the characterization of VWC. Hence, the second objective of this study is to use recursive partitioning to know whether and to what extent hyperspectral wavebands may function as predictors of vegetation water content. We will note that the term VWC from hereon, in this paper, will mean the quantity of water content representing the area at the absorption feature around the 970 nm band.

Materials and Methods

Study Area Overview

The Sandhills of Nebraska is a unique ecosystem that covers 50,176 square kilometers of grass-covered sand dunes and 5,260 square kilometers of wetlands (Turner and Rundquist, 1980). The relationship between the land, water, wildlife, and people is what makes the Sandhills a truly unique place.

Spectral Measurements

The field dataset consisted of non-destructive spectral measurements that were measured under clear skies from vegetation and soil plots with 20 stations each using the Ocean Optics USB2000 spectrometer that covers the 350 nm to the 1025 nm wavelength range. The spectrometer provides resolution to 0.35 nm

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

full width at half maximum (FWHM). Full details of the equipment can be found at the Ocean Optics website (oceanoptics.com). For this study, plots 2, 3, 6 and 8 were utilized. Perennial grasses and numerous weeds dominated the plots. In each sampling station, four readings were taken representing the cardinal directions. To prevent shadow directly over the samples, the operator was made to stand with the sun in front.

Spectra Pre-processing

All field spectra went through filtering to attenuate sensor noise. A moving average with a frame size of 7 points was applied. This corresponds to about 2 nm difference (VIS) and less than 2 nm (NIR) between beginning and ending points. The smoothing procedure on the high spectral resolution data did not only attenuate noise, but also define the structure (shape) of the absorption features.

The 970 nm Absorption Feature

The absorption feature at the water absorption wavelength, 970 nm, was derived using the SAMS software (Spectral Analysis and Management System developed by the Center for Spatial Technologies and Remote Sensing at the University of California, Davis). SAMS calculates areas based on the continuum removed principle. The absorption feature equation (equation 1) is the ratio between the area under the function (in a specified spectrum interval) and the area under the straight line connecting the maxima.

$$a = 1 - \frac{Au}{Ac} \qquad [1]$$

where:

a = the absorption feature
Au = area under the curve
Ac = area under the continuum line

The feature around 970 nm is the only absorption feature for liquid water available for the hyperspectral field dataset.

Regression Models

Three approaches were seen to help model the relationship between the hyperspectral waveband predictors and the VWC: statistical multivariate techniques such as Stepwise Multiple Linear Regression and Partial Least Squares Regression, and the Recursive Partitioning model. No averaging per fraction of nanometer difference in bandwidth was made to the hyperspectral dataset to ascertain that each spectral band becomes part of the predictor set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

Stepwise Multiple Linear Regression

The SMLR was used to evaluate the relevance of each wavelength on the estimation of the vegetation water content at 970 nm. Stepwise selection employs the predictor variable selection of sequentially removing variables that do not meet entry and removal criteria.

The exclusion of most of the predictor variables from the model is due to high inter-correlation among them. High inter-correlation or multicollinearity among variables could introduce redundancy into the regression equation according to Dunagan et al. 2007. In this study, the SMLR looks into the tolerance (equation 2) of each variable; other researchers use the inverse of tolerance, which is the VIF or the variance inflation factor. A tolerance that is close to 0 means there is high multicollinearity of that variable with other independents and the beta coefficients will be unstable.

$$tolerance = (1 - R^2{}_i) \qquad [2]$$

where $R^2{}_i$ is the coefficient of determination of the regression that is produced when the i^{th} narrow spectral predictor band is regressed against the other predictor bands. The SMLR was applied to the hyperspectral field data between wavelengths 700 nm to 990 nm, 928 wavebands in total.

Partial Least Squares Regression

The Partial Least Squares Regression is probably the least restrictive of the known multiple linear regression models. This flexibility allows PLSR to be used in situations where the use of traditional multivariate methods is limited, such as when there are fewer observations than predictor variables. Furthermore, PLSR can be used as an exploratory analysis tool to interpret patterns of scores and loadings of the variables; to select suitable predictor variables (how much of the wavelength influence the VWC). Partial least squares regression has been used in various ways especially when a large number of predictors are involved.

For prediction, PLSR uses equation 3 below:

$$VWC = b_0 + b_1 x_1 + \ldots + b_k x_k + e \qquad [3]$$

where the y is the observed variable (VWC), x is spectral intensity, and b is beta coefficient. The b-coefficients are estimated from the observed y and PLSR scores for the optimal number of PLSR factors. The coefficients contain the spectral information necessary in steering the PLSR model. Each individual value of the coefficient is important in elucidating the important variables, which spectral intensity is contributing to the modeling of the configuration of the VWC.

 doi:10.20944/preprints201701.0001.v1

CART or Recursive partitioning

Recursive partitioning is a nonparametric technique that produces a tree of decision rules in which subjects are assigned to mutually exclusive subgroups according to a set of predictor variables creating a tree of partitions or splits. This technique is widely connected with the acronym CART (Classification and Regression Trees) as popularized by Breinam et al. (1984). In-depth examination of CART results may provide an alternative method to logistic regression for the selection of matching variables. Variable importance ranking (Van der Laan, 2006) is another attractive feature offered by recursive partitioning. Other advantages of the recursive partitioning include: good for exploring relationships without having a good prior model, handles large problems easily, and results are very interpretable.

The CART model partitions the space of the waveband variables into regions such that the variation in VWC within the same region is relatively small. The analysis produces two independent partitions of the space of waveband variables. The recursive partitioning algorithm works by allowing for all possible splits on all potential explanatory variables.

Recursive partitioning picks as the first split the one that does the best job of isolating distinct groups of values: low response values from high response values. Whichever is to be split first depends upon the worth of a partition (logworth) – the highest worth is selected first and the data is appropriately partitioned.

The logworth index in each parent node is a significance measure of the difference in mean values for the observations in each child node with regards to the VWC variable. Higher logworth means higher significant result, when it measures how well the input variable predicts the target values. Logworth is written in the form of $-\log 10\,(p)$, where p is the adjusted probability of the observed data under the hypothesis of the means being equal. The adjusted p-value takes into account the number of different ways splits can occur. It is fair compared to the unadjusted p-value and to the Bonferroni p-value (Sall, 2002).

The model reapplies the same procedure recursively, further splitting the sub-groups. Proceeding in this fashion, it generates a regression tree. Eventually, the realization of certain stopping rules terminates the process and the tree reaches its terminal node and the maximum number of split leaves.

Cross-validation is accomplished using the k-fold. To create a K-fold partition of the dataset, for each of K experiments, K-1 folds is used for training and the remaining one for testing. The advantage of this type of cross-validation over the hold-out method is that all the samples in the dataset are eventually used for both training and testing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

Results

Stepwise Multiple Linear Regression

Table 1 displays the statistical results of the eleven models from the SMLR. Note that we considered every possible decimal placement of the predictors to give us a factual impression of the predictive potential of the model. The predictor variable, 801.24 nm, having the largest correlation with the criterion variable was entered into the equation first (model 1). Depending on the contribution of each predictor to the VWC, the remaining 927 variables entered into the equation one at a time. Only nine met the entry and removal criteria following the input of all waveband predictors. Model 11 selected the best set of predictor variables into the regression equation. The final model consisted of seven specific waveband predictors: 970.53 nm, 956.26 nm, 900.1 nm, 922.77 nm, 976.3 nm, 935.02 nm, and 915.56 nm. These predictor bands can be generalized to its closest nanometer bands: 970 nm, 956 nm, 900 nm, 922 nm, 976 nm, 935 nm, and 915 nm when used further in the analysis.

The model with the highest r^2 value was used for the cross-validation process. The calibration procedure utilized another set of data that was purposely left out to cross-validate findings. Cross-validated predictions are shown in Table 2 along with the statistics previously computed for the combined vegetation water index (CVWI). The CVWI returned the lesser RMSEcv than the SMLR for the tested vegetation samples. To test the sensitivity of the SMLR to presence of soil spectra, a separate cross-validation was conducted using all reflectance samples. The SMLR prediction showed insensitivity to presence of soil spectra, with the almost equivalent RMSEcv values in Table 2. On the contrary, CVWI appeared to be susceptible to soil presence.

Table 1: Stepwise Multiple Linear Regression statistical summary of the eleven models. Model 11 gives the final set of predictors: 970.53, 956.26, 900.1, 922.77, 976.3, 935.02, and 915.56. The unit of measurement for each value is nanometer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

Model	r	r^2	Std. Error of the Estimate
1	0.396	0.157	0.00779
2	0.694	0.482	0.00614
3	0.860	0.739	0.00439
4	0.907	0.823	0.00364
5	0.915	0.838	0.00350
6	0.912	0.832	0.00354
7	0.936	0.876	0.00307
8	0.942	0.888	0.00293
9	0.949	0.900	0.00279
10	0.947	0.896	0.00282
11	0.951	0.904	0.00273

Table 2: Cross-validation statistics showing the selected accuracy indicators. Significance at 0.01 alpha levels are shown in parenthesis.

Index	Pearson Correlation (first set)	r^2cv (first set)	RMSEcv (first set)	RMSEcv (all samples)
SMLR	0.47 (<0.01)	0.22	0.024	0.023
CVWI	0.68 (<0.01)	0.46	0.013	0.171

Absolute values of the partial correlations for variables not in the equation were also examined. Figure 1 shows the values of the partial correlations for each predictor band. Spikes were associated with the best set of predictor variables model 11 resulted into. Significant positive correlations were observed for 970.53 nm, 900.1 nm, and 976.3 nm.

Partial Least Squares Regression

The relationship between the VWC absorption values and the set of hyperspectral reflectance data from 700 nm to 990 nm were modeled using the PLSR. Cross-validation outputs using the same dataset as the SMLR resulted in r^2cv = 0.89. Figure 2 shows a plot of the predicted against the actual VWC.

The optimum number of latent factors in the PLSR model preventing over-fitting was based on the cross-validated RMSE. The analysis revealed seven factors in the final model – the model with the best prediction RMSE (RMSE = 0.40). While seven factors were found to be optimal for the prediction of VWC, the first five were sufficient enough to account for more than 85% of the explained variance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

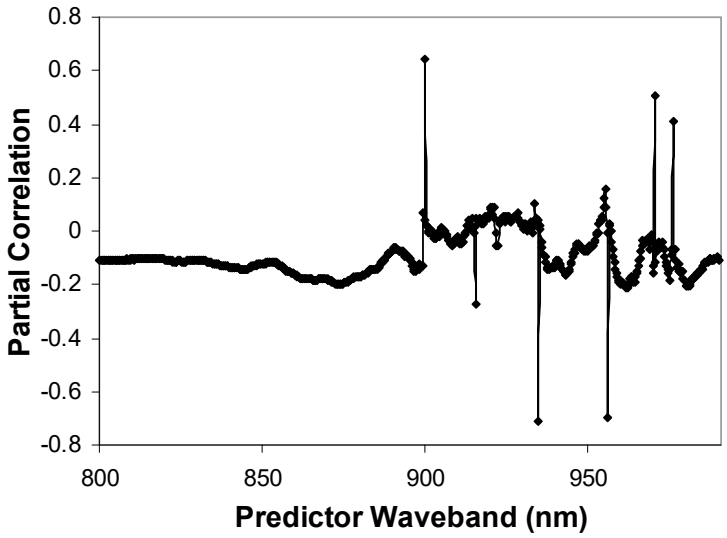

Figure 1: Partial correlation values plotted against the predictors for model 11.

To test the correlation between vegetation water content and spectral dependence of the PLSR model, mean values of the beta coefficients of the first five PLSR loadings for VWC were calculated. The average values of beta coefficients of the optimal seven latent factors were also computed and presented as graphs in Figure 3.

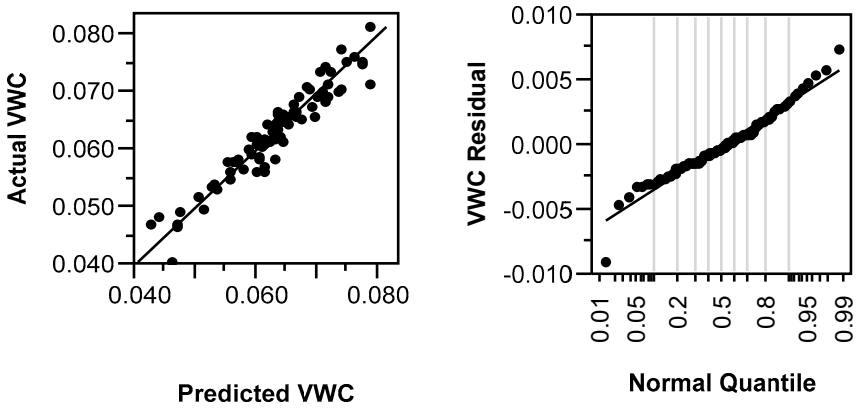

Figure 2: Cross-validated prediction of VWC versus hyperspectral reflectance data in Partial Least Squares Regression model.

Dominant peaks in Figure 3 were consistent with the characteristics of the water absorption features of the vegetation spectra. Using five or seven factors did not make any difference at all to the emergence of peaks or the formation of the valleys.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

The presence of the 900 nm (although not so defined), 922 nm, and the 970 nm peaks affirmed the results of the SMLR. The three wavebands were part of the 11 predictor variables used to model the VWC. A notable observation was the appearance of the dips at wavelengths that were strongly deemed significant predictors for the SMLR model. These wavelength dips were marked having indirect (inverse) effects to VWC than the wavelength peaks.

Another peak at around 940 nm can also be seen, however, this wavelength was not indicated by the final model of SMLR that represented the best set of predictor variables. Other models, on the contrary, had it. The visible peaks and lows were considered the major inputs of the PLSR model.

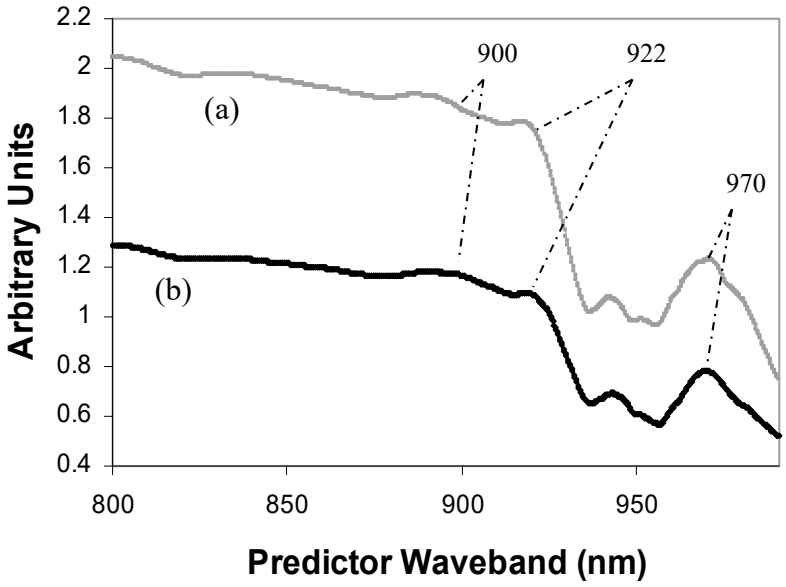

Figure 3: Beta coefficient spectra of the average of the 5 loadings (a) and 7 loadings (b) for the VWC using PLSR.

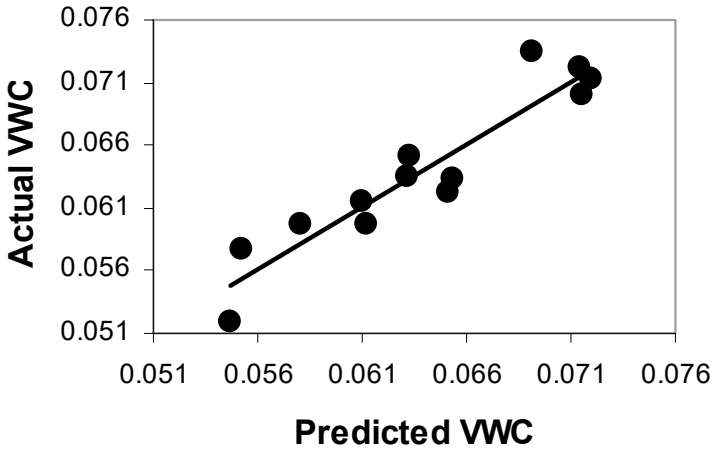

Figure 4: Cross-validated prediction of VWC versus hyperspectral reflectance data in Partial Least Squares Regression model with lesser sample size.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

To determine whether the prediction of the VWC absorption values can be improved using a smaller calibration set, another round of test was conducted. Prediction of these samples with 13 samples in the calibration set produced an r^2= 0.88 (Figure 4), slightly less than the r^2= 0.89 obtained using 928 samples. This demonstrated that even when considering a small number of samples, the calibration model could give reasonable results and could be robust in predicting new VWC unknowns. The RMSEcv of this new prediction was also lower at 0.002. The optimal number of latent factors for this PLSR model was 2. Figure 5 shows the dominant peaks that are also visible in Figure 3, although it tended to be flat around the 900 nm.

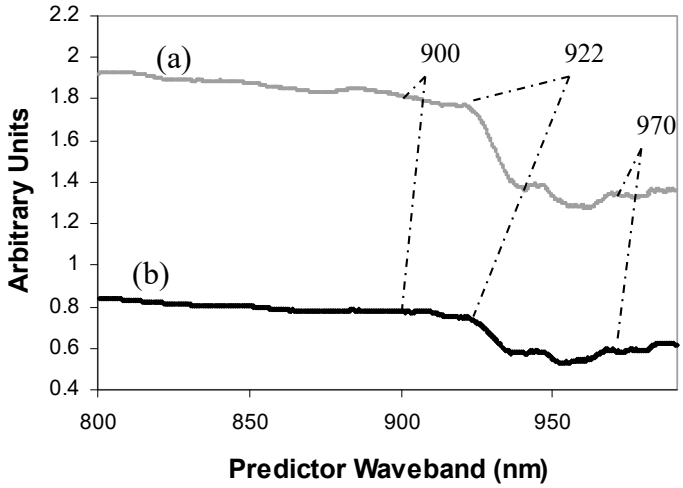

Figure 5: Beta coefficient spectra of the average of the 2 loadings (a) and 6 loadings (b) for the VWC using PLSR.

Recursive partitioning regression

The factor columns (X*s*) we used for this research is continuous. The recursive partitioning resulted into a maximal tree with thirteen leaves. Among the eight input predictors that maximized the information content (970.53 nm, 956.26 nm, 900.1 nm, 922.77 nm, 976.3 nm, 935.02 nm, 915.56 nm, and 940.06 nm), the 956.26 nm surprisingly came out the first input for partitioning the available data, having the highest logworth among predictors -- the best split for an input is the split that yields the highest logworth.

The second split comprised the 900.1 nm, an indication of its importance as a partitioning factor. Interestingly enough, a combination of 956.26 < 15.70 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

900.1 >= 20.99 resulted already into nine samples that could not be further split (Table 5) and with a very high mean VWC (0.068). The conditional values of the two wavelengths were understood to produce a larger difference of spectral reflectance between the two.

Lower mean values of VWC absorptions were spotted at the opposite side (left) of the decision tree. Splitting and increasing the nodes widened the tree further right with larger VWC values detected. As an illustration, a sample pruned tree with only seven leaves is shown in Figure 6. The figure shows the logworth as the tree is pruned further down. Manifestation of the variable importance through parent-child nodes was evident. Nonetheless, whether the number of leaves is seven or maximum thirteen, the 900.1 nm waveband was indicated as the next essential predictor of water content, regardless of the quantity of the VWC absorption.

The other water band, 970.53 nm, appeared one level down from the 900.1 nm as partitioning continued, ending at split 11. Something noteworthy to mention about the 970.53 nm was its emergence on the tree that was restricted only to the right end where larger VWC were identified. This illustrated a charming merging of the 900.1 nm and 970.53 nm: the former being the second partitioning factor, then further partitioned by 970.53 nm, in order to uncover larger values of VWC. A neighboring 976.3 nm also appeared on samples with high VWC.

The least contribution came from the 922.77 nm. This predictor input did not produce meaningful trees and may be eliminated. The same wavelength was weighted less (Table 4) in the Stepwise Multiple Linear Regression procedure (coefficient = -0.005). Table 4 shows how each predictor ranked between two statistical models. Apart from the 922.77 nm, the 915.56 also poorly ranked.

The 940.06 nm band appeared in the recursive partitioning but not in the SMLR.

Table 4: Ranking or importance of the predictors based on the SMLR and recursive partitioning results.

Waveband Predictor (nm)	SMLR Rank Based on Coefficients	Recursive Partitioning Tree Level Location
956.26	1	1
900.10	1	2
935.02	1	4
970.53	2	3
976.30	3	4
915.56	4	3
922.77	5	5
940.06	NA	3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

The 13-leaf tree demonstrated the opportunity that predictor-wavebands 956.26 nm, 900.1 nm, and 970.53 nm could be the three relatively strong predictors of VWC. In fact, the logworth on the right side of the tree at first split showed a very high measure of the worth of the split (logworth = 3.06) compared to the other end. In other words, the bands belong to the best inputs with the best split.

K-fold cross-validation r^2 showed a low value of 0.40. The overall r^2 after partitioning, when actual and predicted VWC were plotted against each other (Figure 7) was a fair 0.59. The cross-validation result value is comparable with other empirical indices previously analyzed.

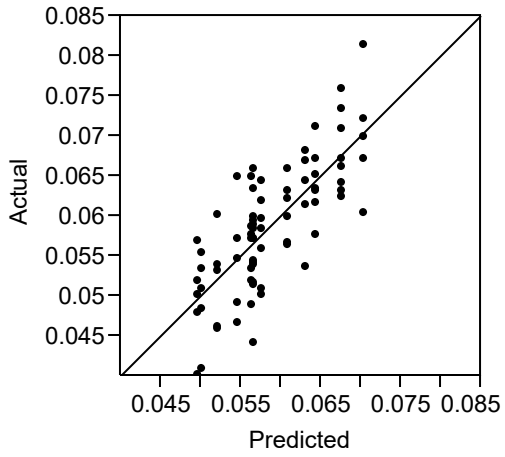

Figure 7: Plot of actual versus predicted VWC values showing how well is the goodness of fit. The abscissa values are the means predicted for each leaf.

Vegetation water content prediction using spectral feature as basis for the quantity of canopy water, such as the absorption area around the 970 nm, can be considered as an optional method, particularly on exploratory analysis when there is data deficiency. The SMLR, PLSR and the recursive partitioning (CART) were three of the methods we tested to further explore hyperspectral wavebands in a vegetation spectra that are central to picking wavelengths for future VWC studies.

The SMLR was implemented to a specified range of the hyperspectral field dataset covering the 700 nm to 990 nm and producing a total of 928 wavebands. When we discarded wavelengths in the visible range of the spectrum, there was efficiency and ease of computations. Also, the absence of water absorption feature present within the visible range, eliminating wavebands was helpful rather than impractical.

The SMLR was able to disqualify highly correlated bands that were present in the dataset. Eliminated bands had tolerance values very close to zero after inspection. Adding highly correlated bands augments relatively little in the

 doi:10.20944/preprints201701.0001.v1

prediction capability of the model. This reinforced the initial perception that no averaging is necessary within bands of almost the same wavelengths, e.g. 900.1 nm, 900.41 nm, and 900.71 nm. The SMLR showed the wavebands that have positive or direct relationship to the VWC through the positive sign of the beta coefficients.

PLSR substantiated the results of the SMLR. Important wavebands were present and were better explained by the scores and loadings plot. The method identified another wavelength for vegetation water prediction, at the 940 nm.

What recursive partitioning regression did to the hyperspectral data was split the spectral intensities using a cutting value and used it to partition the waveband. The recursive partitioning was capable in maximizing the difference in the responses between the two branches of the partition that resulted in wavebands defining particular levels of VWC.

However, there is a word of caution in the interpretation of the results of the recursive partitioning. The exploratory nature of subgroup analysis could possible create high optimism in the highly adaptive procedure in recursive partitioning. Being exploratory, crucial deductions should be reserved and extra steps be taken in interpreting the values resulting from the splitting method. The interactive characteristics of this nonparametric technique could confer optimistic findings most of the time, which could generally please researchers. Thus, a comparative study would be worthwhile to test the robustness of the partitioning results.

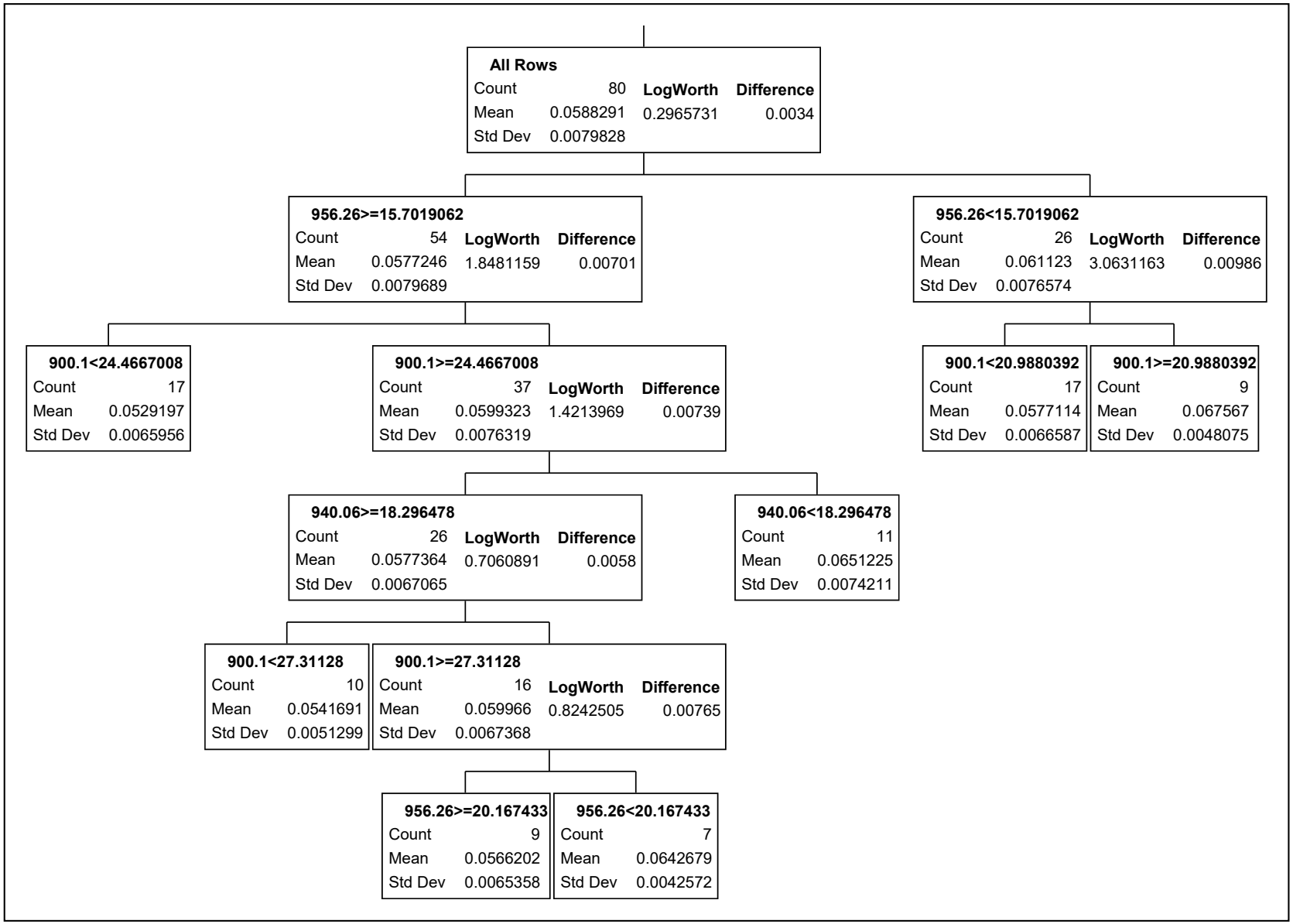

Figure 6: Pruned decision tree to predict vegetation water content from hyperspectral reflectance data using eight input predictors.

doi:10.20944/preprints201701.0001.v1

 doi:10.20944/preprints201701.0001.v1

Conclusions

The determination of the number of significant waveband predictors is essential for the application of statistical VWC models. This can be done even in the absence of actual water content dataset and by using known areas of water absorptions. In this research, we have focused on a major piece of the pie by presenting the matching waveband variables that were recognized to describe significantly the variations of VWC derived from the hyperspectral data. We used two multivariate methods, Partial Least Squares Regression and Stepwise Multiple Linear Regression, to uncover hidden wavebands that may help predict or estimate the canopy vegetation water content. The goal was fulfilled with the utilization of the spectral feature at the 970 nm absorption. Both statistical analyses showed the possibility of modeling the interaction between multiple wavelengths of vegetation spectra and VWC by capitalizing on the importance of specific bands and eliminating band redundancy.

The SMLR pinpointed the best band combination from the NIR region that could administer a strong contribution to the prediction and estimation of VWC. The PLSR affirmed the results of the SMLR by underscoring dominant peaks that were consistent with the characteristics of the water absorption features of vegetation spectra. The laurels of known absorption features such as 900 nm and 970 nm may have been given much attention by academicians due to their visibility on spectral curves that other wavebands, unfortunately, have been banked. From this study involving 80 samples and 928 wavebands, we detected unforeseen band predictors that could better boost VWC predictions only commonly predicted or estimated by the 900 nm and 970 nm.

The SMLR disclosed further the relationship of the wavelength to the VWC by the sign of the beta coefficients. While the models resulting from the SMLR indicated about seven best predictors, the PLSR needed only five factors to sufficient account for more than 85% of the explained variance.

PLSR is favored as a predictive technique over SMLR since the calibration model could give satisfactory results even at lesser number of samples. Compared to the previously investigated VWC indices (WBI, NDWI, NDVI), the multivariate statistical methods appeared to be more powerful, except for the CVWI, based on the RMSEcv accounts. The use of more than three bands or latent factors in the prognosis process has showed relatively better results than just adopting two, which existing VWC indices employed.

This study also applied a modeling scheme to acquire more understanding of the narrow band predictors and their correlations to the VWC. Recursive partitioning method was seen ideally suited to the role of initial predictive modeling methodology. The tree produced using the multidimensional dataset set boundaries that partitioned the VWC into several wavelength band classes

 doi:10.20944/preprints201701.0001.v1

or leaves. However, results of this study would not suggest the optimum tree partitioning.

The manifestation of the importance of few particular wavebands for VWC prediction was seen on our results; nevertheless, any tree branch produced by recursive partitioning may give the best grouping of samples or the best predictors. Also regression trees may have undesirable outcomes when applied to continuous variables. According to Clark and Pregibon (1992), regression trees are best when using categorical variables as predictors.

We want to highlight the point that the results of the recursive partitioning did not insinuate the level of relationships between predictors and water content values. The question of how strong the relationship cannot be deduced. The decision tree only displayed the interpretable actuality that wavelength locations were affected by water contents.

It is hoped that this paper will be instrumental in propagating more ideas in the direction of multivariate calibration model analyses for VWC. By utilizing the right wavebands and the right multivariate and modeling technique in a hyperspectral data with tremendous amount of narrow spectral bands, the biophysical characteristics of vegetation such as water content, can be salvaged with sufficient accuracy.

Acknowledgements

Credits go to Prof. Geoffrey M. Henebry for sharing the Nebraska Sandhills dataset.

 doi:10.20944/preprints201701.0001.v1

References

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and Regression Trees, The Wadsworth Statistics/Probability Series

Carter, G. A. 1991. Primary and secondary effects of water content of the spectral reflectance of leaves. *American Journal of Botany*, 78:916– 924.

Cho, M.A., Skidmore, A., Corsi, F., van Wieren, S.E., Sobhan, I. 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression, *International Journal of Applied Earth Observation and Geoinformation,* 9(4): 375–391.

Claudio, H.C., Y. Cheng, D. Fuentes, J.A. Gamon, A.F. Rahman, H. Qiu, D.A. Sims, H. Luo, and W.C. Oechel. 2006. Monitoring drought effects on vegetation water content and fluxes in chaparral with the 970nm water band index. *Remote Sensing of Environment*, 103: 304-311.

Clark, L.A., and D. Pregibon. 1992. Tree based models. In: Hastie, T.J. Statistical models in S. Pacific Grove, CA: Wadsworth and Brooks: 377-420.

de Jong, S.M., E.J. Pebesma, and B. Lacaze. 2003. Above-ground biomass assessment of Mediterranean forests using airborne imaging spectrometry: the DAIS Peyne experiment. *International Journal of Remote Sensing*, 24(7):1505-1520.

Dunagan, S.C., M.S. Gilmore, and J.C. Varekamp. 2007. Effects of mercury on visible/near-infrared reflectance spectra of mustard spinach plants (*Brassica rapa P.*). *Environmental Pollution,* 148(1): 301–311.

Gao, B.C. and A.F.H. Goetz. 1995. Retrieval of equivalent water thickness and information related to biochemical components of vegetation canopies from AVIRIS data. *Remote Sensing of Environment*, 52:155-162.

Goetz A. F. H., B.C. Gao, C.A. Wessman, W.D. Bowman. 1990. Estimation of biochemical constituents from fresh green leaves by spectrum matching techniques. In: Proceedings 10th International Geoscience and Remote Sensing Symposium (IGARSS'90), 2: 971–974.

Hunt, R. E., and B. N. Rock. 1989. Detection of changes in leaf water content using near-and middle-infrared reflectances. *Remote Sensing of Environment*, 30: 43-54.

Kokaly, R. and R.N. Clark. 1999. Determination of leaf chemical concentration using band-depth analysis of absorption features and stepwise linear regression. *Remote Sensing of Environment*, 67:267-287.

Lefsky, M.A., W.B. Cohen, and T.A. Spies. 2001. An evaluation of alternate remote sensing products for forest inventory, monitoring, and mapping of Douglas-fir forests in western Oregon. *Canadian Journal of Forest Research*. 31: 78-87.

Nguyen, H.T. and B.W. Lee. 2006. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. *European Journal of Agronomy,* 24:349-356.

Peñuelas, J. and Y. Inoue. 1999. Reflectance indices indicative of changes in water and pigment contents of peanut and wheat leaves. *Photosynthetica,* 36:355–360.

Riggs, G. A., and S.W. Running. 1991. Detection of canopy water stress in conifers using the Airborne Imaging Spectrometer (AIS). *Remote Sensing of Environment*, 35: 51–68.

Roberts D.A., R.O. Green, and J.B. Adams. 1997. Temporal and spatial patterns in vegetation and atmospheric properties from AVIRIS. *Remote sensing of environment*, 62:223-240.

Sall, J. 2002. "Monte Carlo calibration of distributions of partition statistics," SAS Institute, Technical Report [Online]. Available: jmp.com/software/whitepapers/pdfs/montecarlocal.pdf

Serrano, L., S.L. Ustin, D.A. Roberts, J.A. Gamon, and J. Peñuelas. 2000. Deriving water content of chaparral vegetation from AVIRIS data. *Remote Sensing of Environment*, 74:570– 581.

Sims, D. A. and J.A. Gamon. 2003. Estimation of vegetation water content and photosynthetic tissue area from spectral reflectance: a comparison of indices based on liquid water and chlorophyll absorption features. *Remote Sensing of Environment*, 84:526–537.

Tian, Q., Q. Tong, R. Pu, X. Guo, and C. Zhao. 2001. Spectroscopic determination of wheat water status using 1650–1850nm spectral absorption features. *International Journal of Remote Sensing*, 22:2329–2338.

Tucker, Compton J. 1980. Remote sensing of leaf water content in the near-infrared. *Remote Sensing of Environment,* 10: 23-32.

Van der Laan, M. J. 2006. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), Article 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

Appendix 1: Spectral indices that have been derived and used for estimation of vegetation water content (VWC) based on ratio, or simple mathematical formula of reflectance of two or more wavelengths.

WBI = Water Band Index; NDWI = Normalized Difference Water Index; EWT = Equivalent Water Thickness; WT = Water Thickness; MSI = Moisture Stress Index; NDII = Normalized Difference Infrared Index; PWI = Plant Water Index; SRWI = Simple Ratio Water Index; SR = Simple Ratio; NDVI = Normalized Difference Vegetation Index; CSI = Canopy Structure Index; GVMI = Global Vegetation Moisture Index; RDI = Relative Depth Index; LWCI = Leaf Water Content Index; CR= Continuum Removal

VW Index	Wavelength Used (nm)	Formulation /Equation	Application	Reference
WBI	895, 972	= R895/R972	AVIRIS imagery	Serrano et al. (2000)
	900, 970	= R900/R970	Unispec (350-1100nm) and GER 2600 (350-2500nm)	Sims and Gamon (2003)
	900, 970	= R900/R970	Field Spectrometer	Peñuelas et al. (1997)
NDWI	857, 1241	= (R857-R1241)/(R857+R1241)	AVIRIS imagery	Serrano et al. (2000)
	860, 1240	= (R860-R1240)/(R860+R1240)	MODIS	Zarco-Tejada et al. (2003)
EWT		Use R1400 through R2500	Earliest Definition of EWT	Knipling (1970)
		Use R867 through R1049	AVIRIS imagery	Serrano et al. (2000)
		Use R920 through R1070	Unispec (350-1100nm) and GER 2600 (350-2500nm)	Sims and Gamon (2003)
	1600	= -ln (1-a)/k where a = R^d1600-R1600, k = extension coefficient of leaf, d = dry state	VIRIS spectrometer	Hunt and Rock (1989)
WT		Use R867 through R1088	AVIRIS imagery	Serrano et al. (2000)
MSI	819, 1599	= MIR/NIR = R1599/R819	AVIRIS imagery	Serrano et al. (2000)
NDII	819, 1649	= (R819-R1649)/(R819+R1649)	AVIRIS imagery	Serrano et al. (2000)
PWI	970, 900	= R970/R900	MODIS	Zarco-Tejada et al. (2003)
SRWI	858, 1240	= R858/R1240	MODIS	Zarco-Tejada et al. (2003)
SR	680, 800	= R800/R680	Unispec (350-1100nm) and GER 2600 (350-2500nm)	Sims and Gamon (2003)
NDVI	680, 800	= (R800-R680)/(R800+R680)	Unispec (350-1100nm) and GER 2600 (350-2500nm)	Sims and Gamon (2003)
	675, 895	= (R895-R675)/(R895+R675)	AVIRIS imagery	Serrano et al. (2000)
	680, 800	= (R800-R680)/(R800+R680)	Field Spectrometer	Peñuelas et al. (1997)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2017 doi:10.20944/preprints201701.0001.v1

	677, 793	= (R793-R677)/(R793+R677)	AVIRIS imagery	Roberts et al. (1997)
CSI	680, 800, 900, 970	= 2sSR – sSR^2 + sWI^2 where: sWI = (WI1180 – 1) / (WI1180 – 1)max sSR = (SR680 – 1)/ (SR680 – 1)max WIxxx = R900/Rxxx	Unispec (350-1100nm) and GER 2600 (350-2500nm)	Sims and Gamon (2003)
GVMI	780-890 and 1580-1750	[(NIR+0.1)-(SWIR+0.02)]/ [(NIR+0.1)+(SWIR+0.02)]	SPOT-Vegetation	Ceccato et al. (2002)
RDI	1116, minimum between 1120 and 1250	[(Rmax-Rmin)/Rmax] where: Rmax = reflectance value at 1116nm; and Rmin = reflectance minimum between 1120 and 1250 nm	GER IRIS MK spectroradiometer	Rollin and Milton (1998)
LWCI	820, 1600	{-ln[1-(R820-R1600)]}/ {-ln[1-(R820-R^{FT}1600)]} where: R^{FT}1600 = reflectance factor at 1600 nm at full turgor	VIRIS spectrometer	Hunt and Rock (1989)
		Uses the RWC equation	AIS imagery	Riggs and Running (1991)
CR	Range 1650-1850		FieldSpec –FR (350-2500nm)	Tian et al. (2001)