Comparative analysis of different uni- and multi-variate methods for estimation of vegetation water content using hyper-spectral measurements

M. Mirzaie a, R. Darvishzadeh b,∗, A. Shakiba a, A.A. Matkan a, C. Atzberger c, A. Skidmore d

a RS & GIS Center, Shahid Beheshti University, Evin, Tehran, Iran
b University of Tehran, Faculty of Geography, RS & GIS Department, Vased Avenue, Tehran, Iran
c University of Natural Resources and Life Sciences (BOKU), Vienna, Institute of Surveying, Remote Sensing and Land Information, Peter Jordan-Strasse 82, 1190 Vienna, Austria
d Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, Hengelosestraat 99, P.O. Box 6, 7500 AA Enschede, The Netherlands

A R T I C L E   I N F O

Article history:
Received 22 December 2012
Accepted 22 April 2013

Keywords:
Hyper-spectral data
Vegetation water content (VWC)
PLSR
PCR
ANN
Narrow band indices

A B S T R A C T

Assessment of vegetation water content is critical for monitoring vegetation condition, detecting plant water stress, assessing the risk of forest fires and evaluating water status for irrigation. The main objective of this study was to investigate the performance of various mono- and multi-variate statistical methods for estimating vegetation water content (VWC) from hyper-spectral data. Hyper-spectral data is influenced by multi-collinearity because of a large number of (independent) spectral bands being modeled by a small number of (dependent) biophysical variables. Therefore, some full spectrum methods that are known to be suitable for analyzing multi-collinear data set were chosen. Canopy spectral reflectance was obtained with a GER 3700 spectro-radiometer (400–2400 nm) in a laboratory setting and VWC was measured by calculating wet/dry weight difference per unit of ground area (g/m²) of each plant canopy (n = 95). Three multivariate statistical methods were applied to estimate VWC: (1) partial least square regression, (2) artificial neural network and (3) principal component regression. They were selected to minimize the problem related to multi-collinearity. For comparison, uni-variate techniques including narrow band ratio water index (RWI), normalized difference water index (NDWI), second soil adjusted vegetation index (SAVIZ) and transferred soil adjusted vegetation index (TSAVI) were applied. For each type of vegetation index, all two-band combinations were evaluated to determine the best band combination. Validation of the methods was based on the cross validation procedure and using three statistical indicators: R², RMSE and relative RMSE. The cross-validated results identified PLSR as the regression model providing the most accurate estimates of VWC among the various methods. The result revealed that this model is highly recommended for use with multi-collinear datasets (R²CV = 0.94, RMSECV = 0.23). Principal component regression exhibited the lowest accuracy among the multivariate models (R²CV = 0.78, RMSECV = 0.41). © 2013 Elsevier B.V. All rights reserved.

1. Introduction

The quantification of biophysical and biochemical factors is a key element in vegetation monitoring and understanding terrestrial processes (Goetz et al., 1992). Remote sensing is regarded as a fast and non-destructive technique that has been widely used in quantifying biophysical (Cohen et al., 2003; Cho et al., 2007) and biochemical parameters (Curran et al., 2001; Hansen and Schjoerring, 2003; Darvishzadeh et al., 2008a,b) in different scales. Water content is one of the main properties of vegetation that can be investigated by remotely sensed data. Vegetation water status plays an important role in plant functioning, water and energy exchange with the atmosphere, as well as drought and fire risk (Penuelas et al., 1993, 1996). Vegetation water content (VWC) has been estimated in numerous studies (Clevers et al., 2010; Colombo et al., 2008; Sims and Gamon, 2003; Cecchato et al., 2001, 2002; Cheng et al., 2006; Danson and Bowyer, 2004). In the mentioned studies, different definitions were used for describing the VWC such as equivalent water thickness (EWT), fuel moisture content (FMC) gravity water content (GWC), and canopy water content (CWC). Remotely sensed VWC has been used to assess plant water status for irrigation purpose and was found helpful in crop estimation (Hunt and Rock, 1989; Penuelas et al., 1993; Jackson et al., 2004), retrieval of soil moisture (Jackson, 1993; Yilmaz et al., 2008),
assessing vegetation conditions related to drought (Tucker, 1980; Claudio et al., 2006) and monitoring the risk of forest fires (Pyne et al., 1996; Maki et al., 2004).

Numerous research efforts focus vegetation indices from broad spectral bands for estimating VWC and other vegetation parameters (Chen et al., 2005; Sims and Gamon, 2003; Fensholt and Sandholt, 2003). One of the most important drawbacks of these multispectral products is the use of average information over broadband widths that results in the loss of critical information provided by narrow bands (Thenkabail et al., 2000). Hyper-spectral remote sensing (or imaging spectroscopy) enables the characterization of vegetation optical properties in many small, contiguous spectral bands within the visible, NIR and shortwave infrared regions. It has demonstrated great potential for an accurate estimation of vegetation water content (Cleverson et al., 2010).

Previous studies have shown for example that narrow bands can be critical in providing essential information for quantifying the biophysical and biochemical characteristics of vegetation (Thenkabail et al., 2000; Broge and Leblanc, 2001; Darvishzadeh et al., 2008a; Cleverson et al., 2010; Zhang et al., 2003). However, it is complex and time consuming to analyze a large number of contiguous narrow bands provided by hyper-spectral data. Therefore, when hyper-spectral data is available researchers often try to utilize methods that can summarize a large number of these multi-collinear narrow spectral bands. Approaches involving uni-variate and multi-variate techniques have been widely used to estimate vegetation parameters and crop yield (Sims and Gamon, 2003; Atzberger et al., 2004; Nguyen and Lee, 2006; Cho et al., 2007; Ye et al., 2007; Li et al., 2008; Darvishzadeh et al., 2008b). Univariate techniques based on spectral indices are still widely used to predict vegetation biophysical properties such as vegetation water content (Sims and Gamon, 2003; Jackson et al., 2004; Thenkabail et al., 2000; Maki et al., 2004; Chen et al., 2005). At most, indices use two or three bands for detecting relationships between spectral measurements and may be limited in terms of exploiting the rich information in narrow bands of hyper-spectral data. Alternatively, several studies have focused on multi-variate models, such as multiple linear regression and stepwise regression, which use several spectral wavelengths for estimating biophysical and biochemical properties (Curran et al., 2001). However, these methods – when used with hyper-spectral data – are likely to suffer from multi-collinearity (Curran, 1989; Grossman et al., 1996).

Multi-collinearity is a common problem inherent to hyperspectral dataset. It arises when one or more of the independent variables (narrow bands) are highly correlated with one or more other independent variables (Curran et al., 2001; Nguyen and Lee, 2006; Van der Meer and Jia, 2012). Multivariate statistical models such as partial least square regression (PLSR), artificial neural network (ANN) and principal component regression (PCR) address this problem (Wold et al., 2001; Atzberger et al., 2010). However, by using PCR and PLSR models, the effects of the multi-collinearity problem can only be reduced but not completely removed (Da-Wen, 2010).

Partial least square regression and principal component regression are full-spectrum techniques that have been used for example to estimate vegetation parameters (Patel and Majumdar, 2010; Hansen and Schojorring, 2003; Nguyen and Lee, 2006; Ryu et al., 2011), crop yield (Yang, 2011; Ye et al., 2007) and soil properties (Ramanad et al., 2005; Farifteh et al., 2007; Nocita et al., 2011; Vohland et al., 2011). However, there are few studies that have investigated the potential of PLSR for estimating VWC (Li et al., 2008).

Artificial neural networks have been widely used in remote sensing (Atkinson and Tatnall, 1997; Walthall et al., 2004; Ramadan et al., 2005; Noh et al., 2006; Liu et al., 2010). Many applications use this non-parametric method for land cover/land use classifications (Bruzzone et al., 1997; Townshend et al., 1991; Serpico et al., 1996; Weng, 2012). Additionally, non-linear relationships and non-Gaussian distributions of data can be modeled, soil and vegetation parameters, as well as crop yield, can be predicted (Ahmad et al., 2010; Yang et al., 2009; Ye et al., 2006; Farifteh et al., 2007). Hence, the potential of neural networks for VWC estimation needs to be evaluated.

With the present study we examine the utility of hyper-spectral measurements to estimate vegetation water content by applying different uni-variate and multi-variate statistical models. Three full spectrum methods involving PLSR, ANN and PCR were used and compared with narrow band vegetation indices. The suitability of each method was analyzed and compared in terms of the relative cross validated root mean square error (RRMSECV) and cross-validated coefficient of determination (R2CV). To fully control the measurement conditions, the study was based on laboratory canopy spectral measurements using four types of plants with different leaf size and shape.

2. Methods and materials

2.1. Laboratory data collection

Four different plant species with different leaf shapes and sizes were selected for testing the utility of different methods for non-destructively estimating VWC. Fig. 1(a) illustrates the plant species and their variability in leaf size and shape. A total of 24 plants were used for the study, 6 plants per species. As different factors affect on canopy spectra, we generated variability within each species by inducing variation in canopy structure (i.e. LAI; see Section 2.2), canopy water content and soil background brightness. To obtain differences in leaf optical properties (in particular water content), plants were divided into two equal groups. One group was placed in nutrient rich soil with adequate irrigation, and the other group was placed in very poor soil, with inadequate irrigation. The detail of the data collection is explained in (Darvishzadeh et al., 2008b, 2009).

2.2. Canopy spectral measurements and VWC measurements

Spectra were measured in a remote sensing laboratory where all the walls and the ceiling were coated with black material to avoid ambient light or reflection. Eight replicates of canopy spectral measurements were collected from each plant using a GER 3700 spectro-radiometer (Geophysical and Environmental Research Corporation, Buffalo, New York.) The wavelength range of the instrument is 350–2500 nm, with a spectral sampling of 1.5 nm in the 350–1050 nm ranges, 6.2 nm in the 1050–1900 nm range and 9.5 nm in the 1900–2500 nm range. Two soil beds with two different types of soils (dark and light) were prepared. Three empty pots were fixed in each soil bed such that their centers were positioned on the border of the sensor’s field of view. In this way, lines drawn from center of each pot formed an equal-distance triangle. A plant canopy was made out of three plants of the same species and with identical treatment. Fig. 1(b and c) shows the arrangement of the pots in the experiment.

The average of the 8 replicates was used for each canopy measurement. For each sample, spectral measurements of bare and air-dried soils were acquired before initiating the canopy reflectance measurements.

To create variations in vegetation water content and spectral properties, the leaves on the inner side of the plants were harvested in six steps. At each step, approximately 1/6 of the total canopy (total leaves) was harvested. Immediately after the canopy reflectance measurements, the fresh weight (FW) of leaves was measured. Next, the leaves were dried at 70 C in an oven for 24 h,
Fig. 1. (a) The four plants type used in this study, (b) soil bed and location of pets and (c) schematic location of plants in our field of view.

Table 1
Summary statistics of vegetation water content parameters of 95 measured samples.

<table>
<thead>
<tr>
<th>Measured parameters</th>
<th>No. of obs.</th>
<th>Mean</th>
<th>St. dev.</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>VWC (g/m²)</td>
<td>95</td>
<td>57.30</td>
<td>52.6</td>
<td>252.7</td>
<td>2.75</td>
</tr>
</tbody>
</table>

until the dry weight (DW) was reached. Water content in canopy scale (CWC) was obtained by measuring the wet/dry weight difference of each plant canopy per unit of ground area (A) base on the following equation:

\[
CWC = \frac{FW - DW}{A} \text{ (g/m}^2\text{)}
\]  

A total of 95 canopy reflectance measurements and corresponding VWC were obtained (Darvishzadeh et al., 2008a, 2009). The summary statistics for canopy water content are given in Table 1. The variability of the spectral measurements in the four plant species is shown in Fig. 2. The effect of size and shape, water content and other differences in vegetation species and soil types are reflected in this figure.

2.3. Pre-processing of spectra

Due to a very high noise levels, bands below 400 nm and above 2400 nm were excluded. A moving Savitzky-Golay filter (Savitzky and Golay, 1964) with a frame-size of 17 data points and a 2nd polynomial was employed to smooth the spectra. The analysis and processing were performed using MATLAB 7.8 (Mathworks, Inc.).

2.4. Hyper-spectral vegetation indices

A number of vegetation indices have been developed to estimate vegetation parameters. Various algebraic combinations of remotely sensed spectral bands reveal useful information about vegetation
The most common indices are ratio-based and soil-based indices. Ratio-based indices used in the study included a ratio water index (RWI) (Pearson and Miller, 1972) and normalized vegetation water index (NDWI) (Gao, 1996) computed according to the following equations:

$$RWI = \frac{R_{\lambda 1}}{R_{\lambda 2}}$$

and

$$NDWI = \frac{R_{\lambda 1} - R_{\lambda 2}}{R_{\lambda 1} + R_{\lambda 2}}$$

The parameters $R_{\lambda 1}$ and $R_{\lambda 2}$ represent the reflectance at wavelengths $\lambda 1$ and $\lambda 2$, respectively.

Selected soil-based indices for this study involved a transformed soil-adjusted vegetation index (TSVI) (Baret et al., 1989) and second soil-adjusted vegetation index (SAVI2) (Major et al., 1990), which were computed using the following formulas:

$$TSVI = \frac{a(R_{\lambda 1} - aR_{\lambda 2} - b)}{aR_{\lambda 1} + bR_{\lambda 2}}$$

and

$$SAVI2 = \frac{R_{\lambda 1}}{R_{\lambda 2} + (a/b)}$$

where $a$ is the slope and $b$ is the intercept of the soil line for a given combination of two wavelengths.

Soil-based indices thus require the specification of the soil line for each wavelength combination. Originally, the soil line was defined by Richardson and Wiegand (1977) as the coefficients of the linear relationship between NIR and the red reflectance of bare soils. However, there are also soil lines for other wavelengths (Thenkabail et al., 2000; Baret et al., 1993). In our study, the soil line parameters $a$ and $b$ were determined experimentally for all wavelength combinations.

To determine the optimal two-spectral bands in each index, all possible two-band combinations using 584 spectral bands were calculated and compared. The 584 discrete narrow bands allowed for computation of $584 \times 584$ (341,056) band combinations for any of the indices (excluding $R_{\lambda 1} = R_{\lambda 2}$). The best wavelengths were detected based on the coefficient of determination ($R^2$) between the VWC and narrow band indices. Next, linear regression models were used to determine the relationships between VWC and optimum narrow-band indices.

### 2.5. PCR and PLSR

Principle component analysis is a standard technique for eliminating multi-collinearity in reflectance values (Martens and Naes, 1989). It involves an orthogonal transformation to convert a set of correlated observations into a set of values of linearly uncorrelated variables called principle components (PC) (Ye et al., 2007). The transformation is defined in such a way that the first principle component explains the largest part of the variance in the data as possible, and each succeeding PC in turn has the highest variance possible under the constraint that it be orthogonal with the preceding components. It is implicitly assumed that the later components contain more and more noise and can therefore be removed from the data set. Following this assumption, the (multiple) regression is thus run only with the first few PCs.

As a second full spectrum method, we used partial least square regression (PLSR). PLSR was introduced by Herman Wold in 1960s as an economic technique (Wold, 1966). It is commonly used in chemometrics as a modeling alternative to PCR and Ordinary Least Squares (OLS) regression when the predictor matrix is poorly conditioned (Ramadan et al., 2005). The PLSR approach can be used to model several response variables simultaneously while effectively addressing strong collinear and noisy independent variables (Wold et al., 2001). It is a bilinear calibration method that utilizes...
data compression by reducing a large number of measured collinear spectral variables to a few non-correlated factors (Hansen and Schjoerring, 2003; Cho et al., 2007; Darvishzadeh et al., 2008b). Similar to PCR, PLSR is used to perform decomposition on spectral data. However, whereas principle component analysis performs the decomposition on the spectral variables alone, PLSR uses the response variable information during the decomposition process (Ye et al., 2007). Thus, a PLS model will try to find the multidimensional direction in the space of predictor variables that explains at the same time a maximum of the variance in the response variable(s). For this reason, the first n PLS scores will always explain a higher amount of variance in the response variable compared to the same number n of PCs (Wold et al., 2001; Atzberger et al., 2010).

The main purpose of PLSR is to build the following linear model:

\[ Y = Xb + E \]

where \( Y \) is the mean-centered or auto-scaled matrix containing the response variable, \( X \) is the mean-centered or auto-scaled matrix of predictor variables, \( b \) is the matrix of the regression coefficient and \( E \) is the matrix of residuals.

As the data-scaling operation affect the result of PLSR, mean centering and auto-scaling were selected and applied to the dataset as alternative pre-processing methods. In mean centering operation the mean of each column is subtracted from the column while in auto-scaling one additionally reduces the standard deviation to unity. Both procedures are commonly used due to their effectiveness in eliminating the additive effects and/or multiplicative factors, such as soil background and illumination on reflectance (Li et al., 2008).

A limited number of latent factors in PCR and PLSR efficiently reduces the model architecture, and generally increase the accuracy of response variable prediction (Ye et al., 2007). It is therefore important to determine the optimal number of latent factors to prevent over-fitting. For PCR, the number of PCs was chosen applying two methods: (1) percent variance explained in the spectral data and (2) cross-validated RMSE. For selecting the number of latent variables for PLSR, visual inspection of cross-validated RMSE values versus the number of factor plots was applied. Additionally, for adding an extra factor to the model, the root mean square error of cross-validation (RMSECV) had to be reduced by >2% (Cho et al., 2007).

2.6. Artificial neural network

The multi-layer perceptron is one of the most widely used neural networks in the remote sensing community (Atkinson and Tatnall, 1997). A typical artificial neural network is composed of various layers (input, output and some hidden layers), and each layer contains a number of interconnected nodes and activation functions (Fig. 3).

Back-propagation is the most popular algorithm for modifying the weights and updating neuronal activities. This algorithm consists of forward and backward propagation processes. In the forward step, network biases and weights are updated from the input to output layer. In backward step, the inter-neuron weights are modified, starting from the output layer until the input layer is reached. The aim of updating the weight is to reduce any identification error.

Different variations of a 3-layer back-propagation ANN were applied in this study for the estimation of VWC. All ANN had a tan-sigmoidal transfer function in hidden layers and a linear transfer function in the output layer. The variants differed mainly regarding the input dataset (and corresponding number of input neurons). In addition, two different training algorithms were applied. As inputs we used (1) all available spectral bands (584 narrow bands), (2) the first ten principal components, and (3) the best narrow-band indices (NDWI, RWI, TSAVI and SAVI2) with pre-selected optimum bands. Principle component analysis and vegetation indices were applied to reduce the data dimension and to minimize the risk of over-fitting. The numbers of neurons in the hidden layer was optimized by trial-and-error. For network training we used (1) Levenberg-Marquardt (TrainLM), and (2) the Scaled Conjugate Gradient (TrainScg), two common training algorithms in back-propagation networks. Early stopping technique was used to avoid overfitting problems (Demuth et al., 2010).

As for the linear models, we used the cross-validation technique to model VWC. For each experiment, 94 samples were used for training; the out-of-bag data acted as a test sample and was simulated. Simple linear regression analyses between the network outputs and the actual VWC was performed to identify the best ANN model. To minimize undesired effects related to the random initialization of the optimization routine, the whole cross-validation procedure was repeated 1000 times. All results represent averages of these 1000 simulations.

2.7. Model performance

Three indices were employed to evaluate and compare the performance of the models used in this study: (1) coefficient of determination \( (R^2) \), (2) root mean square error \( (\text{RMSE}) \) and (3) relative root mean square error \( (\text{RRMSE}) \). \( R^2 \) and \( \text{RMSE} \) are two of the most useful and commonly used indices for evaluating performance of the models (Richter et al., 2012). The validation of the methods used in this study to estimate VWC was based on a cross-validation procedure also called the leave-one-out method. In this method, each sample is predicted based on a model calibrated over the remaining samples. In this study, we developed 95 individual models. Then, cross-validated values of \( \text{RMSE} \), \( 
\text{RRMSE} \) and \( R^2 \) were used for comparing the models. For the artificial neural networks, we additionally repeated the whole calibration process 1000 times to avoid artifacts related to the necessary (random) initialization at the start of the optimization. From the 1000 replicates, the average results were calculated.

3. Results

3.1. VWC estimation and hyper-spectral vegetation indices

Regions within the 2D-correlogram with coefficient of determination between VWC and narrow band indices higher than 75% are highlighted in Fig. 4. Each meeting point in the highlighted area of this plot corresponds to \( R^2 \) value (\( R^2 > 0.75 \)) from the vegetation index made of the reflectance values indicated on the corresponding X and Y axes. The 2D-correlogram shows that the SWIR region includes many band combinations with high \( R^2 \) values.
The best band position for all indices and $R^2$ values are depicted in Table 2. The narrow-band RWI involving bands at 720 and 1418 nm and SAVI2 with 825 and 1443 nm had somewhat higher correlation with VWC than NDWI and TSAVI ($R^2_{RWI} = 0.85$ and $R^2_{SAVI2} = 0.86$, $R^2_{NDWI} = 0.81$ and $R^2_{TSAVI} = 0.8$).

To predict VWC, different linear regression models using optimal indices and VWC were applied. The resulting cross-validated values of $R^2$ and RMSE are summarized in Table 3.

Among the four indices tested, the highest accuracies were found for SAVI2 ($R^2_{CV} = 0.84$ and RMSE$_{CV} = 0.34$) and RWI ($R^2_{CV} = 0.84$ and RMSE$_{CV} = 0.36$). All indices indicate good estimates of VWC.

Plotting the results of the four models reveals very few differences between the prediction accuracies (Fig. 5). Estimated and measured VWC group relatively well along the one-to-one line. The results indicate the potential of different narrow-band indices for estimating vegetation water content (VWC).

### 3.2. VWC estimation using PLSR and PCR

The key issue in performing the PLS and PC regressions was selecting the optimal number of factors. With respect to PCR, our results reveal that the first three principal components explain 99.75% of the spectral data (PC1 = 72.46, PC2 = 25.57, and PC3 = 1.72% of the total variance, respectively) (Fig. 6(a)). Fig. 6(b) shows that cross-validated RMSE in the estimated VWC was strongly reduced by utilizing the first three PCs; however, by adding more PCs, the RMSE increased. In fact, the fourth eigenvector in PCR led to an increase in the prediction error of the regression. Considering the selection criteria described in Section 2.5, the first three PCs were chosen as predictors.

Regarding PLS, seven and five PLS factors were found optimal for auto-scaled and mean-centered data, respectively. The selection was done based on visual inspection of cross-validated RMSE values versus the number of factor plots (Fig. 7(a and b)) and considering the condition of decreasing RMSE$_{CV}$ by >2% (Section 2.5). The different behavior of the two graphs shown in Fig. 7 indicates that the selected data-scaling approach affects the number of optimum PLS factors and model performance.

The comparative analysis of the estimation accuracies of PLSR models and PCR is summarized in Table 4. Both PLS models...
Fig. 5. Measured versus estimated VWC using linear regression models and best narrow-band indices.

Fig. 6. (a) Cumulated percentage variance in spectral data as a number of PCs and (b) changes in RMSEcv of the response variable (VWC) by increasing the number of PCs used in the model.

Fig. 7. Changes in RMSEcv by increasing the number of PLS factors used in the model (a) autoscaled dataset and (b) mean-centered data set.
yielded a better prediction of VWC compared to PC regression. Using PLSR, the auto-scaling method yielded better estimates than mean-centering techniques. The poorer accuracy of the PCR model may be due to the fact that the principal component mainly “explain/summarize” the predictive variable, not the response variable, while PLSR models – by definition – enhance the correlation between the predictive and response variables (Wold et al., 2001; Atzberger et al., 2010; Ye et al., 2007). The cross-validation results showed that the PLS model with seven and five latent variables explained 99.90 and 99.86% of the predictive variable variance (spectral data) and 94.29 and 91.52% of the response variable variance (VWC). For comparison, the three first PCs already explained 99.75% of the spectral variation.

The relationships between the measured and the estimated VWC are shown in Fig. 8(a and b) for PLSR and (c) for PCR. A curvilinear relationship between measured and estimated VWC appears in the PCR model, whereas in PLSR the estimates are much better distributed along 1-to-1 line.

3.3. VWC estimation using artificial neural networks

A comparative analysis of different artificial neural networks resulting from the two selected learning algorithms with different network inputs (and structures) is presented in Table 5.

Table 5 shows the prediction accuracy of the optimum ANN structure (= number of hidden neurons) for each dataset. The most accurate estimation of VWC was obtained with the first 10 principal components as input with TrainScg for network training ($R^2_{CV} = 0.86$ and RMSECV = 0.34). As the prediction performance of ANN depends on the number of neurons in the hidden layer, the best ANN size was identified by testing different number of neurons in the hidden layer (between 2 and 7 neurons). The best prediction accuracy of ANNs was obtained using 5, 6 and 3 neurons in hidden layer when all spectral bands (584 spectral bands), the 10 first PCs and 4 optimal indices, respectively, were applied as input datasets. When the number of neurons in the hidden layer increased, the accuracy of the cross validated dataset decreased due to over-fitting problems.

The reliability of ANN in predicting VWC was confirmed with the high values of the cross-validated $R^2$ and RMSE. The RMSECV and $R^2_{CV}$ values for the ANN models along with the relationship between predicted and measured VWC are shown in Fig. 8(d–f) for the Scg_Train training algorithm and using all 584 reflectance bands, the vegetation indices, and the 10 first PC as inputs, respectively. The figures demonstrate that the simulated data using PCs as inputs were most closely distributed along 1-to-1 line compared to the two other ANN models.

4. Discussion

Our research demonstrated that hyper-spectral measurements, can successfully predict vegetation water content across four structurally different plant types with different LAI and soil backgrounds when combined with multivariate linear and non-linear statistical methods. Within a laboratory setting, the vegetation water content estimation in this study involved two statistical techniques: uni- and multi-variate statistical methods. We have attempted to highlight the ability of multi-variate statistical techniques to deal with multi-collinearity. We found that these methods are superior to uni-variate methods for estimating VWC from hyper-spectral data. However vegetation indices utilizing narrow bands were also found suitable predictors of VWC.

In spite of our efforts, we were naturally limited in correctly simulating airborne or satellite remote sensing conditions or field-based studies. For example, the atmospheric absorption and scattering effects have been ignored in our data set leading to spectral reflectance data free of any kind of atmospheric perturbation. In addition, although in our dataset we tried to generate a wide range of canopy spectra measurements by inducing variation in canopy structure, LAI, leaf size and shape, soil background and water content, the dataset cannot represent all variability found in reality. For better simulating real word conditions, more species should be selected and the number of plants in each spectral measurement.
should be increased. For the above reasons, we confess that the results of our study are only of limited general validity. Nevertheless, the study gave some interesting insights regarding the usefulness of imaging spectroscopy for estimating VWC.

Simple uni-variate techniques involving narrow-band indices appeared to be relatively good predictors of VWC, in particular when including reflectance values from the SWIR. A common drawback toward VIs is that they only utilize information of two (or three) narrow bands available in hyper-spectral data sets, resulting in the loss of lots of important information. However, we made sure that all possible band combinations were tested. We can therefore claim that all available information was analyzed and used to find the best results. Our results demonstrate that despite the simplicity of this approach a reasonable estimation of VWC can be obtained.

Using optimized band settings, the best performance was obtained when SAVI2 and RWI were applied as the predictor variables. However, the performance of the other indices was almost as good. As SAVI2 could not improve the accuracy compared to RWI, it can be concluded that the soil background could be well handled through band ratioing in RWI.

The results showed that the common combination of red and NIR bands in most broad band vegetation indices (SAVI2, TSAVI and RWI) is not the best choice for estimating VWC (Table 2). The large amount of narrow bands available in hyper-spectral data allowed examining other band combinations to find optimal two bands. Selected bands shown in Table 2 and hot spots with high \( R^2_{CV} \) value in Fig. 4 demonstrate and confirm that the SWIR is a critical spectral region rich in information about VWC. Several previous studies have already indicated the strong relationship between SWIR bands and Leaf Area Index (LAI) (Brown et al., 2000; Cohen and Goward, 2004; Darvishzadeh et al., 2008b; Schlerf et al., 2003; Gong et al., 2003).

Another problem in Red/NIR based indices is that they suffer from saturation. In dense vegetation cover, there is a rise in NIR scattering while the red absorption gets quickly saturated; hence a slight change will happen in band ratios (Thenkabail et al., 2000). Examining all available 2-band combinations, we had the opportunity to find new indices reducing the saturation. Estimated values of samples with high VWC shown in Fig. 5 confirm that the SWIR-based indices could overcome problems commonly found in dense canopy.

No existing airborne or satellite sensor has exactly the same bands as used in our study. Therefore, in addition to showing the results of optimum bands for the various indices, we also highlighted the broad spectral areas with high correlation with VWC (Fig. 4). The highlighted regions showing a high correlation with VWC at canopy scale could be potentially applicable to existing airborne and satellite measurements.

Narrow bands indices were also used as inputs to estimate VWC using ANN. Compared to ANN fed with all 584 narrow spectral bands, not only the indices gave higher accuracies (see Table 5) but they also decreased the run time and complexity of ANN models significantly.

Principal components did not improve the prediction of VWC compared to vegetation indices or PLSR. On the contrary, using the first three principal components in the multiple linear regression led to the worst results amongst all methods. The poorer results obtained by PCR compared to PLSR is explained by the fact that decomposition in principal component regression is based solely on reflectance information, while PLSR performs decomposition based on both spectral and response variables (Darvishzadeh et al., 2008b; Ye et al., 2007).

Fig. 6(b) demonstrates that the fourth component in PCR increased the RMSECV of the PC-based regression. In a similar graph of RMSECV versus number of PLS factors (Fig. 7) the number of factors can be increased to 5 and 7 because decomposition in PLSR was performed based on variation in both independent and response variables.

By incorporating the PCs (first 10 PCs) into back-propagation neural networks (BP-ANNs) as input variables, we attempted to prevent the ANN of over-fitting while reducing the data dimension. The results indicated that performing this operation led to a better prediction of response variable (VWC) compared to applying the original dataset. By diminishing the model size using the orthogonal variables (principal components) as input vector, the running time and the model complexity decreased significantly.

After examining different ANN structures, the best neural network performance was found with a combination of PCs and ANN using the Scaled Conjugate Gradient training algorithm (TrainScg) \( R^2_{CV} = 0.86, \text{RMSECV} = 0.34 \). The same ANN input structure (combination of PCA with BP-ANN) was also successfully applied by Ramadan et al. (2005). They used 39 PCs for training a backpropagation (BP) ANN (Ramadan et al., 2005). In our study, PCs and optimal narrow band indices were almost equally effective ANN inputs. A high performance for all network structures and input dataset implies that the relationship between VWC and spectral measurements may be non-linear.

Overall, we found PLSR to be the most effective technique for estimating VWC. The PLSR model using auto-scaling was slightly superior to the mean-centered PLSR model (auto-scaling: \( R^2_{CV} = 0.94 \) and RMSECV = 0.23, mean centering: \( R^2_{CV} = 0.91 \) and RMSECV = 0.25). This difference confirmed that the data pre-processing can be important in improving PLSR performance. Statistical tests revealed, however, that these differences were not statistically significant. Based on the results obtained from this study it can be concluded that multivariate calibration models including PLSR and ANN provide the most accurate estimates of VWC.

5. Conclusion

Our results demonstrate the potential of combining hyper-spectral data and multi-variate linear/non-linear methods to predict vegetation water content. However, despite their simplicity, uni-variate statistical methods based on optimized narrow band indices also showed satisfactory performance. Combined with
ANN, optimized narrow band indices yielded even better results, because non-linearities are taken into account.

Comparing all methods used in this study, the highest correlation ($R^2 = 0.94$) between canopy spectral measurements and VWC was obtained when PLSR was applied. This confirms the potential of this chemometric method as a powerful tool for mapping various biophysical variables.

On the contrary, the use of principle components gave unsatisfactory results within linear models. This confirms, that data transformation has to be done with respect to the target variable (as in PLS), not the spectral data alone. The use of PCs was effective, however, when combined with the non-linear modeling capacity of ANNs.

References


