DE GRUYTER
OPEN

DOI: 10.1515/eces-2016-0008 ECOL CHEM ENG S. 2016;23(1):117-130

Wen-Zhi ZENG[1,2], Jie-Sheng HUANG[1*], Chi XU[1], Tao MA[1] and Jing-Wei WU[1]

HYPERSPECTRAL REFLECTANCE MODELS FOR SOIL SALT CONTENT BY FILTERING METHODS AND WAVEBAND SELECTION

WYKORZYSTANIE HIPERSPEKTRALNYCH MODELI WSPÓŁCZYNNIKA ODBICIA DO OCENY ZASOLENIA GLEBY METODAMI FILTROWANIA I SELEKCJI PASMA

Abstract: For improving the understanding of interactions between hyperspectral reflectance and soil salinity, in situ hyperspectral inversion of soil salt content at a depth of 0-10 cm was conducted in Hetao Irrigation District, Inner Mongolia, China. Six filtering methods were used to preprocess soil reflectance data, and waveband selection combined by VIP (variable importance in projection) and b-coefficients (regression coefficients of model) was also applied to simplify model. Then statistical methods of partial least square regression (PLS) and orthogonal projection to latent structures (OPLS) were processed to establish the inversion models. Our findings indicate that the selected sensitive wavebands for the 6 filtering methods are different, among which the multiplicative signal correction (MSC) and standard normal variate methods (SNV) have some similar sensitive wavebands with unfiltered data. Derivatives (DF1 and DF2) could characterize sensitive wavebands along the scale of VNIR (350-1100 nm), especially the second derivative (DF2). The sensitive wavebands for continuum-removed reflectance method (CR) have protruded many narrow absorption features. For orthogonal signal correction method (OSC), the selected wavebands are centralized in the range of 565-1013 nm. The calibration and evaluation processes have demonstrated the second order derivate filtering method (DF2) combined with waveband selection is superior to other processes, for it has high R^2 (larger than 0.7) both in PLS and OPLS models for calibration and evaluation, by choosing only 156 wavebands from the whole 700 wavebands. Meanwhile, OPLS method was considered to be more suitable for the analyzing than PLS in most of our situations.

Keywords: salinization, modeling, hyperspectral remote sensing

Introduction

Land salinization is a global environmental problem which has received considerable attention in recent years because it is increasing progressively worldwide, particularly in arid and semi-arid regions. Therefore, real-time monitoring of salinization on the basis of

[1] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China, phone +8602768774363
[2] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098, China, phone +8602583786606
[*] Corresponding author: huangjiesheng1962@gmail.com

remote sensing is important for detecting both the temporal and spatial variation of topsoil (0-10 cm) salt content [1, 2].

Superior to traditional multispectral data, hyperspectral data contains large amounts of high-resolution optical signatures to estimate salt content in saline soils and monitor regional salt distribution [3, 4]. All the electronic processes and photon vibrational processes of the overtones and combinations contributed by the fundamental modes of the minerals, water and carbonate are significant for the analyzing of saline-soil hyperspectral data. The spectral signature of saline soils can be a result of the salt itself, which is generated by some salt minerals during weathering process in nature, or added during agricultural management for soil reclamation [5]. Also, some other chromophores are indirectly related to the presence of the salt (*eg* organic matter, particle size distribution).

Salinity can be quantitatively identified in the scale from 350-1050 with hyperspectral data. For example, Csillag et al [6] find 550-770 nm and 900-1030 nm are efficient indicators for salt; Pang et al [7] find 400-900 nm can calibrate quantitative SSC (Soil Salt Content) and EC (Electrical Conductivity) models with R^2 up to 0.89 and 0.92, respectively. The mechanisms of spectral response with salt in the range of 350-1050 nm are mainly occupied by several reasons: the electronic processes, which are contributed by crystal-field effects (which are possibly deduced by Fe^{2+} and Fe^{3+}); the charge transfer by the migration of electrons; the color centers along with some color materials (halide); or the conduction band transition in some periodic lattices [8]. Meanwhile, some vibrational processes appear in the wavebands near 1000 nm, owing to the water molecules or hydroxyl groups essential to salty mineral structure, such as gypsum or montmorillonite [8]. In the meantime, some salt (halite) are spectrally featureless itself but can still be identified because the high affinity of salt to water molecules [9]. Besides, hyperspectral data also contribute a positive effect on salt signature identification, because subtle spectral changes appear in soil salt, which are contributed by the salty effect on the hydrogen bond in water molecules [10].

Hyperspectral data can be difficult to interpret owing to noise and collinearity among spectral bands. Preprocessing methods, such as multiplicative signal correction (MSC), standard normal variate correction or transformation (SNV), first and second order derivative filtering (DF1 and DF2), continuum-removed reflectance (CR), and orthogonal signal correction filtering (OSC), can efficiently minimize signal interference (*ie*, noise) and simplify the interpretation processes [11]. More specifically, MSC and the SNV can be used to remove solid particle scattering, DF1 and DF2 can reduce the baseline effect, CR is useful for comparing with a common baseline, and OSC is suitable for removing invalid information for all wavebands. Nevertheless, it remains unclear which of them is most efficient for hyperspectral reflectance analyses and salt data retrieval.

Novel hyperspectral analysis methods are being examined for extracting salinization information. The partial least square regression (PLS) has been progressively adopted in the fields of remote sensing as it can deal with strongly collinear, noisy data with numerous independent (X) variables [12]. Furthermore, combined with OSC, a new method, orthogonal projection to latent structures (OPLS), was developed [13, 14]. This method can preprocess unfiltered spectra by removing systematic orthogonal variation via OSC and integrate it with a regular PLS algorithm to produce spectra-chemical models [15]. Meanwhile, auxiliary products, variable importance in projection (VIP) and PLS regression coefficient (b-coefficients), are efficient for selecting spectral wavebands and calibrating a parsimonious model without sacrificing accuracy. The value of VIP can be regarded as an

indicator for the evaluation of the importance of variables, as a variable with higher VIP (normally higher than 1) will be more important or worthy of consideration than variables with lower VIP [16]. Also, wavebands with larger b-coefficients (normally larger than their standard deviation) are more important for incorporating in a model [17].

Our spectra-chemical research processes contain several steps: filtering out hyperspectral noise, spectra-chemical model establishment, waveband selection, simplifying the model, and developing a final model. To achieve this research we have the following objectives: (i) exploring the effects of different data pretreatment methods on the accuracy of hyperspectral inversion, (ii) analyzing the applicability and stability of PLS and OPLS methods in hyperspectral data modeling of salt affected soils, and (iii) developing a more accurate waveband selection method which can be used for soil salinity prediction by hyperspectral data.

Materials and methods

Soil spectral data acquisition and analysis

FieldSpec HandHeld (Analytical Spectral Devices, Inc. USA) was used for soil spectral acquisition. The spectral wavelength ranged from 325 nm to 1075 nm and the sampling interval was 1.5 m. The device could reduce the interval to 1 nm by cubic spline interpolation. The resolution ratio (FWHM) was 3.5 nm and the internal standard viewing angle was 25°. For this study, we chose the wavelength range from 351 to 1050 nm to alleviate the effect of noise on the primary wave band. Soil spectral acquisition was conducted in Hetao Irrigation District, Inner Mongolia, China. The sampling sites were randomly distributed within an area of about 26.7 km^2 (107°59'33" E and 41°1'21" N). We took 89 soil spectral samples in late April (sample set 1) and 101 soil spectral samples in late July (sample set 2).

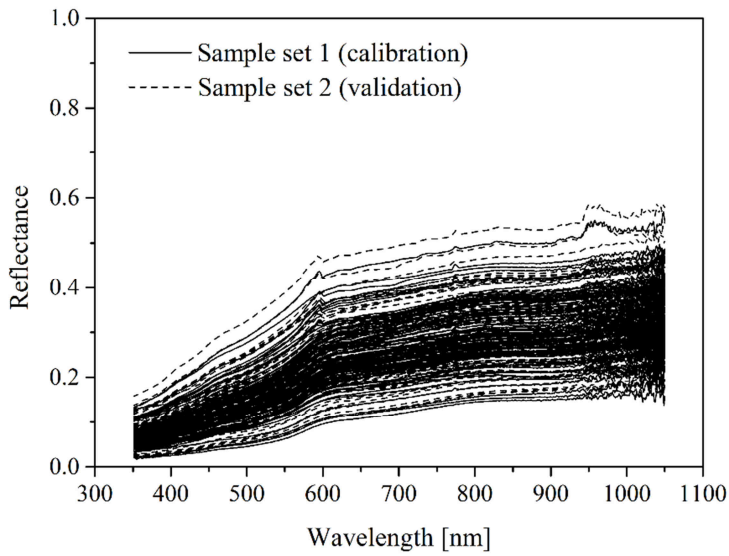

Fig. 1. Hyperspectral curves of unfiltered spectra (the solid lines are soil sample set 1 and the dash lines are soil sample set 2)

A white Spectralon panel by barium sulphate ($BaSO_4$) was used to transfer the relative reflectance (which is the quantity actually measured by instrument) to the absolute reflectance in post processing by multiplying the reflectance factor spectrum by the actual calibrated reflectance spectrum of the reference standard. Each spectral curve was achieved by taking the arithmetic mean of five measurements, and all samples were taken between 10 am to 2 pm (Fig. 1). Furthermore, the pistol grip was fixed in spider and keep the pistol grip facing towards the sun. The bubble on the pistol grip is used as an indicator to keep it level, and the optical fiber is kept 30 cm above the soil surface during the whole measurement.

Sample sets 1 and 2 (0-10 cm depth) underwent laboratory analysis to determine their water and salt contents (Table 1). Soil Water Content (SWC [g/g]) was obtained by oven drying at 105°C for 24 h. Soil Electrical Conductivity ($EC_{1:5}$ [dS/m]) was measured in a 1:5 soil: water suspension after 1 hour of end-over-end shaking at 25°C and was converted to Soil Salt Content (SSC [%]) according to an convincing empirical formula calibrated for local measured data (not presented here) by regression analysis ($SSC = 0.4EC_{1:5}-0.04$). According to Pang et al [7], the results of invert EC are almost identical to SSC. As a consequence, retrieve SSC is reasonable.

Table 1

Statistics for soil samples

Statistical index	Salt index/S [%]				Water index/θ [g/g]	
	S_1	lnS_1	S_2	lnS_2	θ_1	θ_2
Maximum	16.2	2.8	16.0	2.8	0.675	0.74
Minimum	0.2	-1.7	0.1	-2.5	0.016	0.01
Mean	2.2	0.2	3.1	0.4	0.114	0.11
Standard Deviation	3.3	1.0	4.0	1.3	0.084	0.12
Kurtosis	6.2	0.2	1.8	-0.6	21.286	15.61
Skewness	2.6	0.7	1.7	-0.0	3.375	3.43
Significance	0.0	0.5	0.0	0.7	0.126	0.00
Decision	reject	accept	reject	accept	reject	reject

The confidence interval was 95% and $p = 0.05$. S and θ indicate salt index and water index respectively, subscript 1 and 2 indicate soil sample 1 and soil sample 2, respectively. One sample in sample 2 lost soil moisture content

Pretreatment of data

We adopted the Kolmogorov-Smirnov test [18] to demonstrate that the results of natural logarithm processing of the soil salt content of sample sets 1 and 2 satisfied the normal distribution. Meanwhile, hyperspectral data was standardized by centering and scaling to unit variance to reduce the effects of dimensionality, in order to keep variables in the same scales. To minimize the impact of soil surface brightness, we applied 6 filtering methods: MSC, SNV, DF1, DF2, CR, and OSC to pretreat the data (Fig. 2).

MSC is a type of preprocessing method that has been used to separate absorption features from scattering features. Accordingly, it can calibrate the baseline effect, reduce the influence of scattering effects on the obtained spectra and decrease spectral discrepancies associated with sample inhomogeneity [19]. The SNV method is a row-oriented transformation that centers and scales individual spectra and employs an algorithm similar to MSC [20]. However, MSC calculates an ideal spectrum from the unfiltered data and uses this spectrum to modify the data, whereas the SNV method uses a normalizing

method to remove scattering effects. The methods of DF1 and DF2 can be used to find the inflection points of the spectral curve, detect subtle differences between changes and enhance relationships between spectral data and target parameters [21]. Furthermore, DF2 can enhance minor convexities and concavities in the reflectance curve to improve the elimination of baseline effects. The CR method aims to quantify the absorption of a given material at a specific wavelength, based on the assumption that other materials cannot affect the absorption features around a specific site. This method is designed to highlight departures and recognize absorption features by removing the non-absorbing continuum [22]. Finally, the OSC is a filtration method that can extract and remove orthogonal variation from independent variables of reflectance, allowing more accurate interpretation of the dependent data of SSC [23].

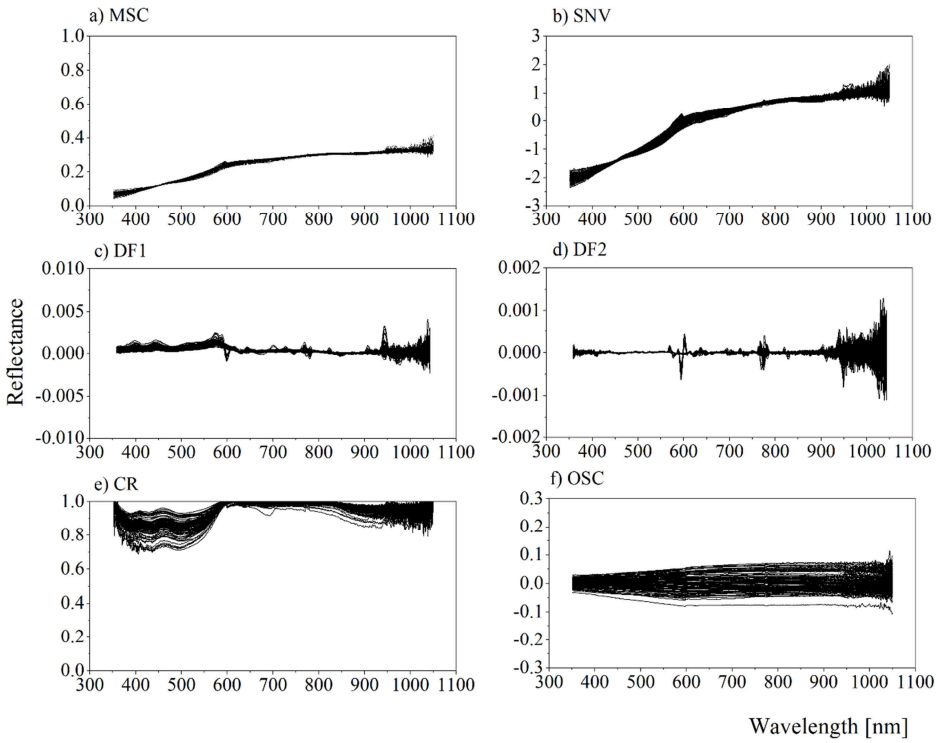

Fig. 2. Hyperspectral reflectance curves of different filter methods: a) multiplicative signal correction (MSC), b) standard normal variate (SNV), c) first-order derivate filter (DF1), d) second-order derivate filter (DF2), e) continuum - removed (CR), f) orthogonal signal correction (OSC)

Model calibration and evaluation

Both PLS and OPLS can be used to calibrate models. The OPLS is a combined algorithm incorporating both PLS and OSC, which first conduct preprocessing of the unfiltered spectra by removing systematic orthogonal variation just like OSC and then apply the PLS algorithm to the models.

In the present study, a model was calibrated using the SIMCA software (version 13.0). b-coefficients (coefficients of each waveband in PLS and OPLS models) were figured out simultaneously. Soil sample set 2 was used for model calibration while set 1 for evaluation. For both PLS or OPLS methods, it is important to select the appropriate number of spectral bands for the model according to Q^2, which is an effective index of determining the appropriate number of spectral bands [24]. The Q^2 represents the fraction of the total variation for dependent variables that can be predicted by a given component, while the Q^2 (cum) denotes the fraction that can be predicted by all components. Coefficients of determination, denoted R^2, measures how well the regression line approximates the real data points, should be as close to 1 as possible for the best model. *RMSEE* (Root Mean Square Error of Estimation) and *RMSEP* (Root Mean Square Error of Prediction) were used to indicate the fitness and predictive power of the models:

$$RMSEE = \sqrt{\frac{\sum (Y_{obs1} - Y_{pred1})^2}{N - 1 - A}} \tag{1}$$

$$RMSEP = \sqrt{\frac{\sum (Y_{obs2} - Y_{pred2})^2}{M}} \tag{2}$$

where Y_{obs1}–Y_{pred1} (Y_{obs2}–Y_{pred2}) and N (M) refer to the fitted residuals for the measured and simulated observations and the number of samples, respectively, in the calibration (evaluation) set. The root mean square error from cross-validation (*RMSECV*) applies to the calibration set (*ie* like *RMSEE*) but is an indicator of predictive power (*ie* like *RMSEP*). We calculated it by summarizing the cross-validation residuals of the observations in the calibration set.

To determine the significant wavebands of the established models, we used a parameter known as the *VIP* [25]. For an observed dependent variable Y, the *VIP* was calculated as follows:

$$VIP_k(a) = k \sum_a \omega_{ak}^2 \left(\frac{SSY_a}{SSY_t} \right) \tag{3}$$

where $VIP_k(a)$ is the importance of the k^{th} predictor variable based on a model with a factors, ω_{ak} is the corresponding loading weight of the k^{th} variable in the a^{th} PLS or OPLS factor, SSY_a is the sum of squares of dependent variable explained by a PLS or OPLS model with the α factors, SSY_t is the total sum of squares of dependent variables, and k is the total number of predictor variables.

Results

The soil samples collected in the study area had broad range of salinity with minor moisture. Specifically, the mean soil salt content are 2.2 and 3.1% for dataset sample 1 and sample 2 respectively, correspondent with the standard deviation to be 3.3 and 4.0 (Table 1). These soil samples was suitable for calibrating a representative model to retrieve soil salinity owing to the subtle disturbance by moisture, as well as broad range of measured salt content.

Filtering process

Hyperspectral data were treated with different filtering methods (Fig. 2) and the mean filtered reflectance curves are illustrated in Figure 3. It was clear that the trends of MSC, SNV, and the unfiltered spectra were similar. The DF1 curves exhibit considerable fluctuation over the range of 351-600 nm, compared to other wavebands. Conversely, the curves obtained using DF2 exhibit fluctuations throughout the entire spectrum, particularly pronounced around 600 and 780 nm and above 900 nm. The CR produced a spectral curve that could be divided approximately into 3 sections: the absorption characteristics were more pronounced in the first (351-600 nm) and third (820-1050 nm) sections than in the second section (600-820 nm). Meanwhile, there was a distinct peak at 600 nm and heavy fluctuations after 900 nm, and all 7 filtered curves had a fluctuation at 600 nm and large variations from 400 to 600 nm.

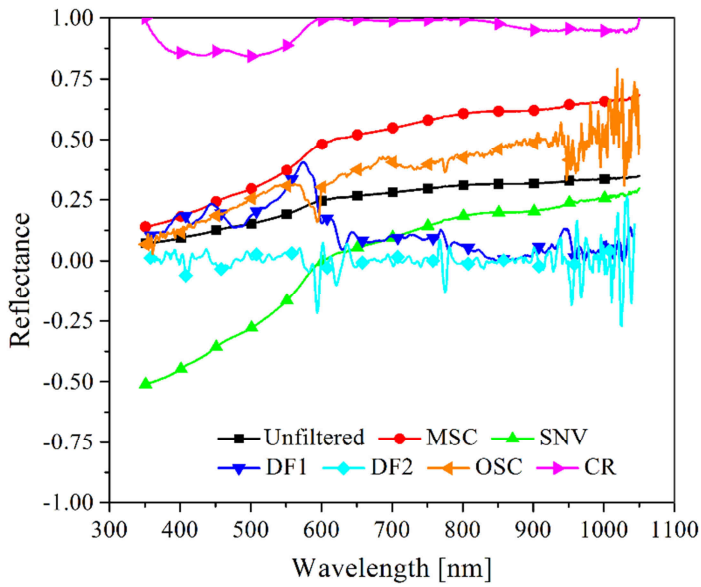

Fig. 3. Mean values of hyperspectral reflectance for unfiltered and filtered reflectance (manually multiply 2, 1/4, 300, 2000 and 300 for multiplicative signal correction (MSC), standard normal variate (SNV), first-order derivate filter (DF1), second-order derivate filter (DF2) and orthogonal signal correction (OSC), respectively, in order to compare patterns in the same graph)

Models established by PLS and OPLS

Both unfiltered and the 6 filtered spectra were used to calibrate the models according to the PLS method. Taking the Unfiltered-PLS model as an example: when the number of components reached 6, each component included was able to enhance the model's ability, as Q^2 (cum) increased from 0.0735 (not presented) to a maximum of 0.456. The numbers of components and other statistical indicators for the application of different filtering methods with the PLS model are presented in Table 2.

Table 2

Statistical results for partial least squares regression (PLS) models of unfiltered and filtered data, respectively (multiplicative signal correction (MSC), standard normal variate (SNV), first-order derivate filter (DF1), second-order derivate filter (DF2), continuum - removed (CR), and orthogonal signal correction (OSC) filter methods respectively)

Model	Calibration					Validation	
	NPC	Q^2(cum)	R^2	RMSEE	RMSECV	R^2	RMSEP
Unfiltered-PLS[1]	6	0.456	0.722	0.718	1.034	0.665	0.604
MSC-PLS[1]	6	0.552	0.746	0.686	0.883	0.655	0.615
SNV-PLS[1]	6	0.548	0.754	0.674	0.885	0.631	0.641
DF1-PLS[1]	5	0.536	0.759	0.664	0.903	0.416	0.836
DF2-PLS[1]	3	0.497	0.753	0.666	0.916	0.673	0.597
CR-PLS[1]	6	0.533	0.686	0.762	0.926	0.492	0.787
OSC-PLS[1]	1	0.787	0.793	0.604	0.605	0.719	0.564
Unfiltered-PLS[2]	6	0.547	0.724	0.595	0.781	0.675	0.565
MSC-PLS[2]	6	0.534	0.747	0.57	0.773	0.648	0.649
SNV-PLS[2]	6	0.538	0.747	0.569	0.761	0.656	0.635
DF1-PLS[2]	5	0.417	0.749	0.565	0.841	0.573	0.78
DF2-PLS[2]	3	0.55	0.799	0.5	0.784	0.741	0.529
CR-PLS[2]	6	0.524	0.755	0.56	0.764	0.684	0.593
OSC-PLS[2]	1	0.785	0.789	0.507	0.506	0.775	0.441

NPC indicates the components number; superscript 1 means original model and superscript 2 means new model after wavelength selection. Q^2(cum) means the fraction of the total variation of dependent variable (Y) that can be predicted by all components. R^2 means determination coefficient. RMSEE and RMSEP means root mean square error of estimation and prediction respectively. RMSECV means root mean square error from cross validation

Table 3

Statistical results for orthogonal projection to latent structures (OPLS) models of unfiltered and filtered data, respectively (multiplicative signal correction (MSC), standard normal variate (SNV), first-order derivate filter (DF1), second-order derivate filter (DF2) and continuum - removed (CR) respectively)

Model	Calibration					Validation	
	NPC	Q^2(cum)	R^2	RMSEE	RMSECV	R^2	RMSEP
Unfiltered-OPLS[1]	1+5	0.549	0.722	0.718	0.882	0.665	0.604
MSC-OPLS[1]	1+6	0.589	0.785	0.635	0.841	0.679	0.593
SNV-OPLS[1]	1+6	0.598	0.792	0.624	0.832	0.658	0.616
DF1-OPLS[1]	1+5	0.59	0.811	0.592	0.841	0.435	0.831
DF2-OPLS[1]	1+3	0.585	0.83	0.555	0.846	0.593	0.68
CR-OPLS[1]	1+5	0.53	0.686	0.762	0.9	0.492	0.787
Unfiltered-OPLS[2]	1+5	0.55	0.724	0.595	0.733	0.675	0.565
MSC-OPLS[2]	1+6	0.552	0.776	0.538	0.732	0.667	0.649
SNV-OPLS[2]	1+6	0.551	0.775	0.54	0.732	0.669	0.647
DF1-OPLS[2]	1+5	0.523	0.789	0.512	0.755	0.609	0.768
DF2-OPLS[2]	1+3	0.591	0.828	0.465	0.699	0.716	0.619
CR-OPLS[2]	1+5	0.558	0.755	0.56	0.726	0.684	0.593

NPC indicates the components number, in which the number before "+" (always 1) represents for predictive component and the number after "+" represents for orthogonal components; superscript 1 means original model and superscript 2 means new model after wavelength selection. Q^2(cum) means the fraction of the total variation of dependent variable (Y) that can be predicted by all components. R^2 means determination coefficient. RMSEE and RMSEP means root mean square error of estimation and prediction respectively. RMSECV means root mean square error from cross validation

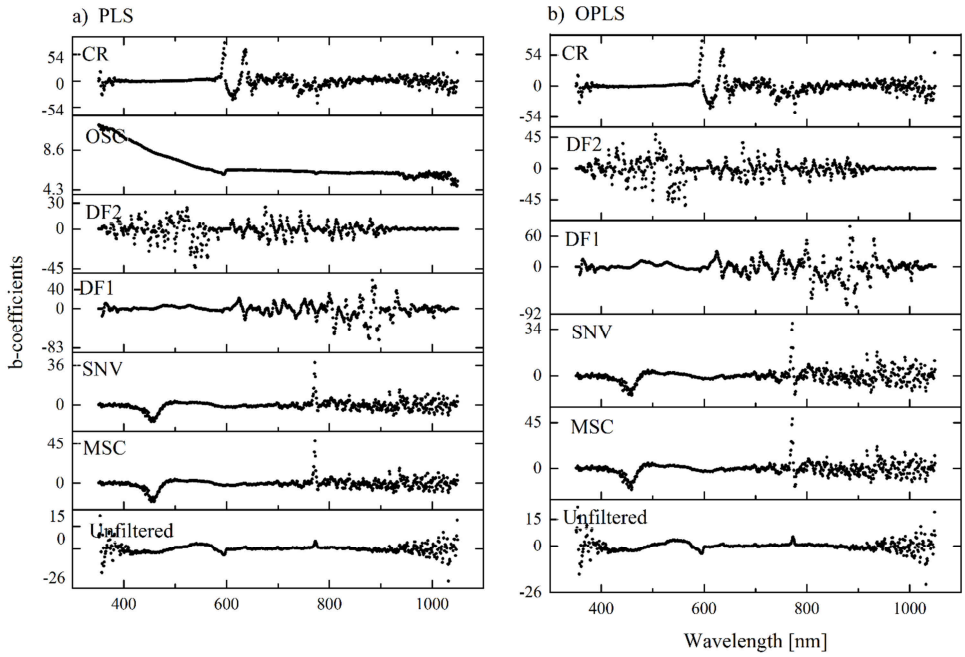

Fig. 4. Stack graphs of b-coefficient values of: a) partial least squares regression (PLS) and b) orthogonal projection to latent structures (OPLS) models for the hyperspectral data

The results of component analysis for the OPLS models were presented in Table 3. The OPLS method was able to conduct orthogonal projection filtering contained in orthogonal component, which is similar to OSC; therefore, we did not consider OSC in this part. Statistical analysis of our OPLS models demonstrated that the SNV-OPLS model produced the largest Q^2 (cum) and smallest *RMSECV* (Table 3). The coefficients of each waveband in the PLS and OPLS models (b-coefficients) are illustrated in Figure 4. It was clear that these coefficients are relatively similar for both the PLS and OPLS models.

Wavebands selection and new PLS and OPLS establishment

We calculated the VIP value (Fig. 5) using Eq. (3) and found similar VIP values for the OSC-PLS and Unfiltered-OPLS models. Furthermore, the VIP values obtained for MSC-PLS (or MSC-OPLS) was similar to SNV-PLS (or SNV-OPLS) model. Apart from that, no other obvious similarities are apparent between them.

VIP values (larger than 1) are commonly used with b-coefficients (larger than its standard deviation) together to identify the important variables. This selection theory was applied to all the data to select all the important wavebands (Fig. 6) as input variables in the new PLS or OPLS models. As a consequence, 94 wavebands for unfiltered data, 91 wavebands for MSC data, 88 wavebands for SNV data, 103 wavebands for DF1 data, 156 wavebands for DF2 wavebands, 438 wavebands for OSC data and 102 wavebands for CR data were retained, respectively (Fig. 6).

These were then utilized as variables to build new PLS (n-PLS) and new OPLS (n-OPLS) models (Tables 2 and 3). The calibrated new models showed better performance

than before (discussed later), which illustrated that the process of waveband selection was beneficial for calibrating more concise models without sacrificing accuracy.

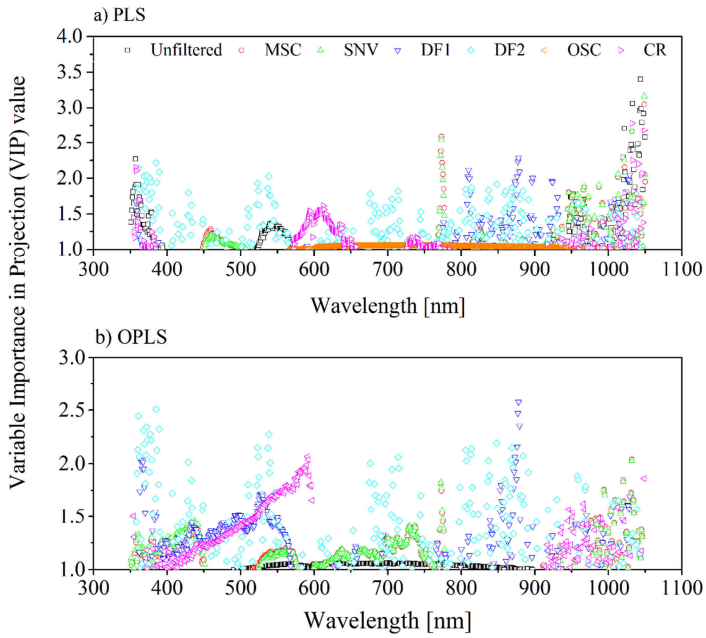

Fig. 5. Variable Importance in Projection (VIP) values of: a) partial least squares regression (PLS) and b) orthogonal projection to latent structures (OPLS) models for the hyperspectral data

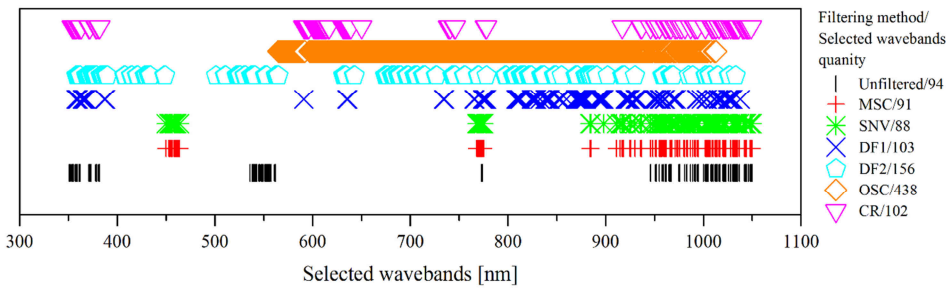

Fig. 6. Picture of selected wavebands for unfiltered and filtered reflectance, 94, 91, 88, 103, 156, 438 and 102 wavebands are selected out for Unfiltered, multiplicative signal correction (MSC), standard normal variate (SNV), first-order derivate filter (DF1), second-order derivate filter (DF2) and orthogonal signal correction (OSC) and continuum - removed (CR) reflectance, respectively

Discussion

Water in the soil pore space and the soil particle water film have a critical influence on reflectance spectra in the VNIR and SWIR wavelengths and are also influenced by the salinity-water content. According to research by Whiting, the fundamental stretching and

bending vibrations of water and hydroxyl bonds of soil in the region of 350 to 2500 nm mainly occurred in the SWIR such as 1400, 1900, and 2800 nm, except for a very weak absorption strength at VNIR (986 nm) [26]. The soil moisture is very low in samples, thus affect little on hyperspectral reflectance. It is interesting to notice that the sensitive wavebands selected in our process, especially after 950 nm, are meanwhile the relatively high-noise wavebands. This means, although the noise in the reflectance after 950 nm caused by the instrument is large, they are more sensitive and should be retained. This pattern is also detected by Csillag et al [6] who considered 900 to 1030 nm as indicators of soil salinity.

VIP combined with b-coefficients is a useful method of selecting important bands. For example, Elmasry et al [27] used VIP to select effective wavelengths from high spectral dimensionality data. In our study, we selected the wavebands with large VIP (greater than 1) and b-coefficients (larger than standard deviation) as valuable wavebands.

After waveband selection, different important wavebands were highlighted for different filtered data (Fig. 6). The selected bands for MSC (91 bands) were distributed in the range of 450-464 nm, 768-775 nm, and sparsely located after 884 nm. SNV had some similar selected bands (88 bands) with unfiltered data in the range of 773-774 nm and after 946 nm of, and some special bands in the range before 562 nm. Ninety-four bands were selected in unfiltered data, in which the wavebands of 351-382 nm and 536-562 nm were also important. The above wavebands mentioned for each method could be regarded as the selected sensitive wavebands (Fig. 6). When derivative data were selected, the former fashion of sensitive wavebands were changed to scattered distributed. Meanwhile, the DF2 data tended to be more decentralize than DF1 data (103 bands) and 156 bands are selected for DF2 data. It might be contributed by fact that DF2 derivative could characterize subtle, consistent variations caused by curvatures along the whole scale (Fig. 6), as a result, DF2 should be relative insensitive to variation caused by adverse effects such as sun angle and cloud cover, and thus contribute to a more precise model (Tables 2 and 3). Of all six selection processes, OSC selected the most number of the unfiltered wavebands of the 700 wavebands, reaching 438 wavebands. Moreover, they are centralized from 565 nm to 1013 nm [28]. CR reflectance spectra characterizes up to 102 bands, which indicated the absorption features well. In our study, the selected wavebands are located in the range of 354-363 nm, 371-382 nm, 591-617 nm, 630-639 nm, 739-746 nm, 777-778 nm, 936-1049 nm, which indicate the sensitive wavebands and absorption features located in the range of VNIR. This is similar to what Csillag et al [6] identified in a study in Hungary, where he characterized key spectral ranges in the visible (550-770 nm) and near infrared (900-1030 nm). All the selections for the seven hyperspectral data sets (Fig. 6) show the wavebands near 1000 nm are always retained for model calibration, possibly because the heavy vibrational, or sharp band transition process [8]. Meanwhile, the filtering methods significantly impacted the effect of sensitive bands, because the signature implied within a specific band can only be extracted out by some specific filtering algorithm. The complicated electronic processes, which is mainly constituted by crystal-field effects, charge-transfer, color centers and conduction band transitions, along with vibrational processes contributed by water, hydroxyl, carbonate or phosphate, are the reasons that produce differences among the waveband selection [8].

Our study shows that the predictive ability of PLS is similar with OPLS, evidenced by the similarities in b-coefficients (Fig. 4), extracted components and statistics (Tables 2 and 3). However, there were some disagreements about the advantages and disadvantages of

their applicability. Lin et al [29] demonstrated the OPLS methods are simpler, easier to interpret, and more accurate than the PLS method, while Tapp and Kemsley [30] opposed this. In our study, models established by OPLS were marginally more accurate than the PLS models when MSC, SNV, DF1 or DF2 pre-processing were applied. Of the indexes obtained, Q^2 (cum) were almost always higher for the OPLS method than for the PLS method (except for CR). This indicated that the OPLS method had a larger effect on detecting cross-validated components and a greater ability to interpret the models. Furthermore, OPLS extracted only one predictive component, which was an index reflecting salt content, that enhanced the model interpretation (Table 3). Model calibration showed most of the filtered reflectance data were better than unfiltered in the inversion process, showing the benefits of filtering.

After the wavebands selection, the new PLS or OPLS models had better results for most of the models. For PLS models, although only four of seven new calibration processes, which were Unfiltered, MSC, DF2 and CR, were better than old models, six new evaluation processes beyond MSC exhibited better performance than olds. The enhancement in evaluation for new models encourages us more. For example, the PLS models calibrated by selected DF2 data have higher R^2 values than unselected DF2 model (Table 2 - 0.799 versus 0.753), and the evaluation pattern is more obvious (0.741 versus 0.673). Besides, the validated PLS model by CR data improved the accuracy greatest when the selected wavebands were used (0.492 versus 0.684) [31]. In OPLS models, this phenomenon in evaluation processes is more clearly, which could be obviously observed in DF1, DF2 and CR data, with R^2 were 0.609, 0.716, and 0.684 for selected ones versus 0.435, 0.593, and 0.492 for previous ones respectively. As a consequence, the new models with less bands are potential for calibrating more representative and applicable models with better evaluation results.

Conclusions

Taking all into consideration, four models display great performance with R^2 values higher than 0.7 both in calibration and evaluation. The four models are the new PLS models of DF2 and OSC, the new OPLS model of DF2 and the old PLS model of OSC (without waveband selection). Among all the models, DF2 just selected one third of the wavebands compared to the OSC model. As a results, the DF2 filtering method and waveband selection are recommended in quantitative retrieval of salt content in arid lands. Overall, our study have successfully retrieved the salt content in saline soils by combining various filtering techniques and wavebands selection methods. This will be useful for the detection of soil salt content using the hyperspectral apparatus in the field and develop models for airborne and future satellite hyperspectral sensors such as AVIRIS and HYSPIRI.

Acknowledgements

This work was made possible by support from the State Natural Science Fund of China (grant No. 51379151), Open Foundation of State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (grant No. 2015490211), and China Postdoctoral Science Foundation (grant No. 2015M582274).

References

[1] Peragón JM, Delgado A, Díaz JaR, Pérez-Latorre FJ. A GIS-based decision tool for reducing salinization risks in olive orchards. Agr Water Manage. 2016;166:33-41. DOI:10.1016/j.agwat.2015.12.005.

[2] Benini L, Antonellini M, Laghi M, Mollema P. Assessment of water resources availability and groundwater salinization in future climate and land use change scenarios: A case study from a coastal drainage basin in Italy. Water Resour Manag. 2015;30(2):731-745. DOI: 10.1007/s11269-015-1187-4.

[3] Song D, Liu B, Li X, Chen S, Li L, Ma M, et al. Hyperspectral data spectrum and texture band selection based on the subspace-rough set method. Int J Remote Sens. 2015;36(8):2113-2128. DOI: 10.1080/01431161.2015.1034892.

[4] Kattenborn T, Maack J, Faßnacht F, Enßle F, Ermert J, Koch B. Mapping forest biomass from space-Fusion of hyperspectral EO1-hyperion data and Tandem-X and WorldView-2 canopy height models. International J Appl Earth Observation Geoinform. 2015;35:359-367. DOI: 10.1016/j.jag.2014.10.008.

[5] Dehaan R, Taylor G. Image-derived spectral endmembers as indicators of salinisation. Int J Remote Sens. 2003;24,(4):775-794. DOI: 10.1080/01431160110107635.

[6] Csillag F, Pásztor L, Biehl LL. Spectral band selection for the characterization of salinity status of soils. Remote Sens Environ. 1993,43(3):231-242. DOI: 10.1016/0034-4257(93)90068-9.

[7] Pang G, Wang T, Liao J, Li S. Quantitative model based on field-derived spectral characteristics to estimate soil salinity in Minqin County, China. Soil Sci Soc Am J. 2014;78,(2):546-555. DOI: 10.2136/sssaj2013.06.0241.

[8] Hunt GR. Spectral signatures of particulate minerals in the visible and near infrared. Geophysics. 1977;42,(3):501-513. DOI: 10.1190/1.1440721.

[9] Hick P, Russell W. Some spectral considerations for remote sensing of soil salinity. Soil Res. 1990;28,(3):417-431. DOI:10.1071/SR9900417.

[10] Hirschfeld T. Salinity determination using NIRA. Appl Spectrosc. 1985;39(4):740-741. DOI: 10.1366/0003702854250293.

[11] Tsai F, Philpot W. Derivative analysis of hyperspectral data. Remote Sens Environ. 1998;66(1):41-51. DOI: 10.1016/S0034-4257(98)00032-7.

[12] Andersson M. A comparison of nine PLS1 algorithms. J Chemometr. 2009;23(10):518-529. DOI: 10.1002/cem.1248.

[13] Dumarey M, Goodwin DJ, Davison C. Multivariate modelling to study the effect of the manufacturing process on the complete tablet dissolution profile. Int J Pharm. 2015;486(1):112-120. DOI: 10.1016/j.ijpharm.2015.03.040.

[14] Gabrielsson J, Jonsson H, Airiau C, Schmidt B, Escott R, Trygg J. OPLS methodology for analysis of pre-processing effects on spectroscopic data. Chemometr Intell Lab. 2006;84(1):153-158. DOI: 10.1016/j.chemolab.2006.03.013.

[15] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemometr. 2002;16(3):119-128. DOI: 10.1002/cem.695.

[16] Gosselin R, Rodrigue D, Duchesne C. A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. Chemometr Intell Lab. 2010;100(1):12-21. DOI: 10.1016/j.chemolab.2009.09.005.

[17] Haaland DM, Thomas EV. Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. Anal Chem. 1988;60(11):1202-1208. DOI: 10.1021/ac00162a021.

[18] Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc. 1967;62(318):399-402. DOI: 10.1080/01621459.1967.10482916.

[19] Martens H, Nielsen JP, Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Anal Chem. 2003;75(3):394-404. DOI: 10.1021/ac020194w.

[20] Dhanoa M, Lister S, Sanderson R, Barnes R. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. J Near Infrared Spectrosc. 1994;2(1):42-47. DOI: 10.1255/jnirs.30.

[21] Inoue Y, Sakaiya E, Zhu Y, Takahashi W. Diagnostic mapping of canopy nitrogen content in rice based on hyperspectral measurements. Remote Sens Environ. 2012;126:210-221. DOI: 10.1016/j.rse.2012.08.026.

[22] Vašát R, Kodešová R, Borůvka L, Klement A, Jakšík O, Gholizadeh A. Consideration of peak parameters derived from continuum-removed spectra to predict extractable nutrients in soils with visible and near-infrared diffuse reflectance spectroscopy (VNIR-DRS). Geoderma. 2014;232:208-218. DOI: 10.1016/j.geoderma.2014.05.012.

[23] Wang G, Yin S. Quality-related fault detection approach based on orthogonal signal correction and modified PLS. IEEE Trans Industr Informatics. 2015;11(2):398-405. DOI: 10.1109/TII.2015.2396853.

[24] Eastment H, Krzanowski W. Cross-validatory choice of the number of components from a principal component analysis. Technometrics. 1982;24(1):73-77. DOI: 10.1080/00401706.1982.10487712.
[25] Gomez C, Lagacherie P, Coulouma G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma. 2008;148(2):141-148. DOI: 10.1016/j.geoderma.2008.09.016.
[26] Whiting ML, Li L, Ustin SL. Predicting water content using Gaussian model on soil spectra. Remote Sens Environ. 2004;89(4):535-552. DOI: 10.1016/j.rse.2003.11.009.
[27] Elmasry G, Wang N, Vigneault C, Qiao J, Elsayed A. Early detection of apple bruises on different background colors using hyperspectral imaging. LWT-Food Science and Technology. 2008;41(2):337-345. DOI: 10.1016/j.lwt.2007.02.022.
[28] Griffin JL. Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. Curr Opin Chem Biol. 2003;7(5):648-654. DOI: 10.1016/j.cbpa.2003.08.008.
[29] Lin W-S, Yang C-M, Kuo B-J. Classifying cultivars of rice (Oryza sativa L.) based on corrected canopy reflectance spectra data using the orthogonal projections to latent structures (O-PLS) method. Chemometr Intell Lab. 2012;115:25-36. DOI: 10.1016/j.chemolab.2012.04.005.
[30] Tapp HS, Kemsley EK. Notes on the practical utility of OPLS. TrAC Trends Anal Chem. 2009;28(11):1322-1327. DOI: 10.1016/j.trac.2009.08.006.
[31] Cho MA, Skidmore A, Corsi F, Van Wieren SE, Sobhan I. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. Int J Appl Earth Observ Geoinformat. 2007;9(4):414-424. DOI: 10.1016/j.jag.2007.02.001.