Remote Sens. **2009**, *1*, 408-417; doi:10.3390/rs1030408

OPEN ACCESS

Remote Sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Communication

Evaluating Principal Components Analysis for Identifying *Optimal* Bands Using Wetland Hyperspectral Measurements From the Great Lakes, USA

Nathan Torbick [1],* and Brian Becker [2]

[1] Applied Geosolutions, 87 Packers Falls Rd, Durham, NH, 03857, USA

[2] Department of Geography, 292 Dow Science Complex, Central Michigan University, Mount Pleasant, MI, 48859, USA; E-Mail: becke1b@cmich.edu

* Author to whom correspondence should be addressed; E-Mail: torbick@agsemail.com; Tel.: +1-603-659-2392; Fax: +1-603-659-0419

Received: 13 May 2009; in revised form: 12 August 2009 / Accepted: 12 August 2009 / Published: 18 August 2009

Abstract: Mapping species composition is a focus of the wetland science community as this information will substantially enhance assessment and monitoring abilities. Hyperspectral remote sensing has been utilized as a cost-efficient approach. While hyperspectral instruments can record hundreds of contiguous narrow bands, much of the data are redundant and/or provide no increase in utility for distinguishing objects. Knowledge of the optimal bands allows users to efficiently focus on bands that provide the most information and several data reduction tools are available. The objective of this *Communication* was to evaluate Principal Components Analysis (PCA) for identifying optimal bands to discriminate wetland plant species. In-situ hyperspectral reflectance measurements were obtained for thirty-five species in two diverse Great Lakes wetlands. PCA was executed on a suite of categories based on botanical plant/substrate characteristics and spectral configuration schemes. Results showed that the data dependency of PCA makes it a poor, stand alone tool for selecting optimal wavelengths. PCA does not allow diagnostic comparison across sites and wavelengths identified by PCA do not necessarily represent wavelengths that indicate biophysical attributes of interest. Further, narrow bands captured by hyperspectral sensors need to be substantially re-sampled and/or smoothed in order for PCA to identify useful information.

Keywords: hyperspectral; principal components analysis; wetlands; invasives; optimal bands; Great Lakes

1. Introduction

The conservation and restoration of wetland ecosystems are goals of many levels of legislation and the management community. This requires assessment and the development of knowledge inventories concerning wetland extent, biological composition, and health for making management decisions and monitoring control efforts. Satellite remote sensing with multispectral sensors has been a useful tool in providing inventory information on wetlands types; however, both spatial and spectral resolutions have limited the level of detail ultimately required for comprehensive wetland assessments.

Advances in sensor technology and remote sensing science have developed an interest in hyperspectral data for mapping wetlands at the species level [1-4]. An underlying goal of hyperspectral imagery is to take advantage of its inherent high spectral resolution, and in the case of airborne sensors high spatial resolution, to better classify or discriminate objects. Analytical spectroscopic systems possess capabilities to capture data at narrow spectral bandwidths continuously covering 0.4 to 2.5 μm of the spectrum. This allows for small variations in plant/substrate absorptance and reflectance to be recorded [4]. Incorporating such relatively high spectral detail makes it possible to explore species separability [1,3].

As hyperspectral technologies become more commonplace, the wetlands end-user community will benefit from knowing the location of bands *optimal* for their specific area of application in order to achieve the potential improvements in classification accuracy these hyperspectral imagers present. Guidelines to assist end-users in the efficient selection of these optimal bands are crucial because the number of potential band combinations in a hyperspectral image is typically quite large; there are nearly 44,000 possible four-band combinations in a 16-band image.

In remote sensing, feature extraction involves the identification of those statistical characteristics of remotely sensed data that capture the most systematic variation [5,6]. Systematic variations, as opposed to non-systematic variations (i.e., noise), ultimately provide the foundation for target differentiation. Within multispectral data, often there is significant correlation between the information contained in closely adjacent bands. Inter-band correlation is especially problematic for hyperspectral data composed of a large number of narrow, contiguous bands. In principle, feature extraction can significantly reduce the magnitude of a dataset by isolating essential components (i.e., information) while discarding redundant or noisy data [5].

Principal components analysis (PCA) is a method in which original data is transformed into a new coordinate system, which acts to condense the information found in the original inter-correlated variables into a few uncorrelated variables, called principal components. With respect to hyperspectral data, PCA transforms large data sets into relatively few meaningful uncorrelated orthogonal variables/dimensions (i.e., the principal components) that represent most of the information present in the original image. In any principal components rotation, the first component or dimension accounts for the maximum proportion of the variance of the original image, and subsequent components account for maximum proportion of the remaining variance [7,8].

The overarching goal of this *Communication* is to summarize results evaluating PCA as a hybrid data reduction/band selection technique for identifying optimal bands in wetland hyperspectral measurements.

2. Experimental Section

2.1. Study Site

The investigation was carried out in fresh-water wetland areas within the central Lower Peninsula of Michigan, USA. The first focus area was an expansive marsh covering approximately 20 km^2 in the lower Muskegon River Watershed (MRW), located on the western side of central Michigan (W86°09'45", N43°16'10"). The National Wetland Inventory classifies the majority of the wetland complex as palustrine, with seasonally and semi-permanently flooded regions of scrub-shrub, forest, and emergent covers. The second focus area was a complex of semi-isolated, botanically diverse wetlands along a major highway (US127) in central Michigan (W84°31'02", N42°51'13"). The National Wetland Inventory classifies this wetland as a palustrine emergent wetland.

2.2. Data Collection

A field campaign was conducted during August in 2006, which represents the peak of the growing season. A portable spectroradiometer (FieldSpec Pro FR®, Analytical Spectral Devices (ASD), Inc., Boulder, CO, USA) was used to collect *in situ* radiance between 350 to 1,000 nanometers (nm). A subset of these radiance data were parsed from the original spectra, 1 nm wide bands ranging from 415 to 1,000 nm. This range was strategically selected as a number of previous investigations have highlighted the near infrared (NIR) region as possessing the greatest utility for distinguishing vegetative targets [1-4]. Viewing geometry utilized a 24 degree field-of-view (FOV) held approximately 1-meter above the target for measurements representing field-canopy conditions. The viewing geometry configuration approximately represents the spatial resolution current airborne hyperspectral sensors can achieve (~1m pixel). Therefore, this sampling scheme is representative of a typical configuration of airborne hyperspectral flights. Approximately 30-40 spectra were collected at nadir and averaged per sampling location to better capture inherent target variability.

Radiance measurements were taken of a Spectralon® reference panel (Labsphere, Inc., North Sutton, NH, USA) near-simultaneously with each plant spectra. A Lambertian reference panel was utilized for calibration. Sun-target-sensor geometry was repeated as best as possible under these difficult field conditions between 11:00-14:30 local time. Rapidly sequenced measurements were averaged over a homogeneous (one species) plot. The instrument was shifted within the patch during collection to capture inherent within species variability and ensure non-overlapping FOVs. This was repeated at four different patches for each species. During data acquisition, the sensor was first placed over the reference panel to record the panel-reflected radiance. Then the sensor was placed over the target to record the target-reflected radiance. Then, by ratioing the radiance measurements, surface reflectance factor was calculated. By definition, the term reflectance factor (Equation 1) is the ratio of radiant emittance of a target to that reflected into the same reflected-beam geometry and wavelength range by an ideal and diffuse standard surface irradiated under the same conditions [9]:

$$R\,(S_\mathrm{i}\,S_\mathrm{r}\,\lambda) \tag{1}$$

where S_i is the angular distribution of all incoming radiance and S_r is the reflected radiance measured by the sensor for a given λ wavelength [7].

In addition to the ASD spectral configuration, a baseline spectral library was created and re-sampled utilizing the band center/Full Width Half Maximum (FWHM) profiles emulating common airborne hyperspectral imagers (48 bands, avg. FWHM 5.8 nm). In addition, a second re-sampling was conducted that doubled the number of bands (95 bands, avg. FWHM 2.9 nm) to evaluate the impact of instrument configuration on selected bands. These two data sets emulate the FWHM of currently available airborne hyperspectral sensors (e.g., Compact Airborne Spectrographic Image (CASI)-1500).

Table 1. Sampled species from each study area. E:Emergent, S:Submergent, Fl:Floating, G:Graminoin, F:Forb, Sr:Shrub, R:Rushes/Segdes.

MUSKEGON		**U.S. 127**	
Eleocharis rostellata	E	*Asclepias incarnata*	F
Elodea canadensis	S	*Cephalanthus occidentalis*	Sr
Filamentous algae	Fl	*Cyperus esculentus*	R
Heteranthera dubia	S	*Eleocharis rostellata*	E
Iris versicolor	E	*Leersia orxyzoides*	G
Leersia orxyzoides	G	*Lemna minor*	Fl
Lemna minor	Fl	*Najas sp.*	S
Lythrum salicaria	E	*Phalaris arundinacea (Green)*	G
Myriophyllum spicatum	S	*Phragmites australis*	E
Nuphar lutea	Fl	*Sagittaria latifolia*	E
Nymphaea odorata	Fl	*Salix nigra*	Sr
Phragmites australis	E	*Schoenoplectus pungens*	R
Poa sp.	G	*Scirpus sp. (1)*	R
Polygonum hydropiperoides	F	*Scirpus sp. (2)*	R
Pontederia cordata	E	*Soldago gigantea*	F
Potamogeton crispus	S	*Sparganium androcladum*	E
Sagittaria latifolia	E	*Typha latifolia*	E
Salix nigra	S		
Schoenoplectus tabernaemontani	E		
Spaganium americanum	E		
Typha angustifolia	E		
Vallisnera americana	S		

2.3. Botanical Sub-Categorization

A set of sub-categorizations were developed to evaluate differences, if any, in optimal wavelengths identified via PCA based on plant types or biological communities. These subcategories were strategically chosen as species tend to grow in biological communities rather than random distributions. For this reason, we believe the categories are more logical than each species individually. The three data pools outlined above (i.e., ASD, 95-band, and 48-band) were further broken down into six categories based on their plant community and/or substrate type. A summary of these sub-categorizations and their associated abbreviations are contained in Table 1.

2.4. Analysis

A set of extensive scripts were developed utilizing the MATLAB™ environment in order to transform the three data pools and their six subcategories into their principal components (i.e., dimensions). A single category was selected within the dataset to perform a principal components transformation. Band number was established as the independent variable in order to characterize the explanative power of each band with respect to the sampled, botanical community (i.e., dependent variable).

Correlation-based PCA references standardized input variables (i.e., correlation matrices) that have a mean of zero and a variance of one. Standardization tends to inflate the contribution of variables whose variance is small, and reduce the influence of variables whose dimensions are large. Covariance-based PCA, on the other hand, is typically used when the relative magnitudes of the variables are important because its un-standardized format enhances magnitude differences and reduces the potential for an insignificant variable to exert a strong influence on the results [10]. Within the literature [i.e., 7-8], it remains unclear which methodology (covariance or correlation) was most applicable to band identification within the context of this research, so both were utilized. It was assumed that both provide a different, but meaningful, perspective.

Primary outputs of interest from PCA runs were eigenvalues and eigenvectors. Eigenvalues contain a synopsis of the percentage of the original data variance that is captured or explained by each principal component. Eigenvectors, which are by definition uncorrelated to each other and related to only one eigenvalue, provide information about data patterns within the new coordinate system. The eigenvalues, eigenvectors, and covariance/correlation matrices were further combined to yield component loadings (Equation 2). PCA component loadings represent a coefficient between each independent variable (i.e., band) and any one component:

$$\text{Component Loading}_{bp} = \frac{\text{Eigenvector}_{bp} \text{ x } (\text{Eiganvalue}_{p})^{0.5}}{(\text{COV}_{b} \text{ or } \text{COR}_{b})^{0.5}} \qquad (2)$$

where $_b$ is each original band and $_p$ is each principal component for covariance or correlation approach.

Stated differently, loadings measure the relative degree to which each original band explains the relationship between any one component and the body of dependant variables, in this case being botanical signatures. If any one component captured that portion of the overall data variance that was inherently related to the differentiation of these botanical signatures, then those bands loading highest on that component should also be well suited for botanical differentiation.

Although individual methods resulted in the identification of important loading values for an individual dimension, no method performed across the range of data. Thus, a more simplified approach to loading/band center selection was adopted. The top 10% positive and top 10% negative loading values were selected, when present. The 10% were calculated for each PC band, six for COR and 6 for COV based PCA. In some instances, especially the first dimensions, there are no negative loadings, so they are not present with respect to a single dimension.

The number of output factors (i.e., dimensions) generated by PCA is typically held equal to the number of substantively meaningful independent patterns (extracted features) among the variables tested [11]. In order to determine the dimensions deemed meaningful in this study, hyperspectral images of each study area was subjected to a PCA transformation. The resulting component images were systematically inspected in order to identify those dimensions that maintained landscape dependence. In both cases, the 7th dimension and beyond were found to be noisy, and the features within the wetland were not discernable. Thus, PCA dimensions 1-6 were deemed meaningful for this investigation.

The identification of the meaningful dimensions and the band specific (i.e., 20% of the 48 or 95 bands) loading values are the baseline data matrices upon which further analyses were conducted. Histograms were generated for each baseline data matrix, depicting the relative frequency of selected component loading values vs. wavelength (i.e., band centers). This methodology allowed the visual inspection of the loading response-curves associated with the 12 extracted dimensions. The extracted dimensions refer to both PCA approaches. Fundamentally, the key band centers identified through the visual examination of loading histograms should be most applicable to the differentiation of the botanical community from which the PCA-based signatures were generated.

The PCA results were then compared to another common band selection tool, namely 2nd derivative analysis (Equation 3). Derivative methodology has been used to distinguish wavelength locations where substantial inflection occurs [3, 4]. In Becker *et al*. [3] and Torbick *et al*. [4], second derivative approximations identified seven wavelengths (685, 731, 939, 514, 812, 835, 823, 560 nm) using contiguous data covering the visible and NIR regions coastal wetlands of Lake Erie and Lake Huron, USA. A set of scripts were developed using a piecewise cubic spline to smooth a non-continuous/unsmoothed spectra in order to create a polynomial from which true second derivative values could be calculated at each band location. The five highest magnitude positive and negative values were selected to identify wavelengths possessing distinct diagnostic spectral change. This percentage was chosen because inspection of the data shows that derivatives and their paired wavelengths resulting from inflection points caused by system noise and not botanical sources were more frequent in the "middle" 80% of the data which has been shown to be a useful approach [4]. The high magnitude values represent points of inflection that are located at the center of a reflectance (negative values capture convex features) and/or absorption feature (positive values capture concave features):

$$d^{1st} = (\rho_{n+1} - \rho_n) / (\lambda_{n+1} - \lambda_n) \quad (3)$$

where d^{1st} is the 1st derivative (line segment slope) and ρ is the percent reflectance factor at a given λ wavelength.

3. Results and Discussion

In an effort to circumvent repetitive in depth presentation of the results from numerous individual subcategory experiments that contributed little to the objectives of this research, only the noteworthy data manipulations and linked PCA output are presented in a sectional format.

3.1. Site/Location Specificity

Figure 1 compares the PCA and 2nd derivative approach for band selection across the two sites. Figure 1a illustrates the 2nd derivative results for each site with band ~825 and ~835 having the only noteworthy differences between the two sites. There were no substantial differences between the datasets for the 2nd derivative approach. The location of the most frequently occurring loading/band centers are clearly superimposed upon one another between the two sites. While small differences exist (575 nm), generally, the locations of high occurring loadings are the same. The NIR edge (665 and 715 nm) and the 525/560 regions are noted as locations of high occurrences.

Figure 1b is the 48-band re-sampled output (covariance and correlation combined) histograms for both MRW and US127 using PCA. Figure 1b also displays the two study site separately, now for covariance and correlation. The bands at 600, 850, and 950nm were found to have substantial differences between sites. Substantial defined as three or more orders of magnitude in terms of number of occurrences. This threshold was chosen to indicate the most noteworthy bands relatively. Based on similarity of these data, no further breakdown of the separate study sites are presented in the results, and subsequent results represent the output from MRW and US127 combined to increase the range of the various output histograms.

Figure 1. (a-top.) The similarity of the sites is near perfect match between the 2nd-derivative frequency-of-occurrence curves for both sites. (b-bottom.) The data dependency of PCA creates notably different frequency of occurrence curves for the two sites.

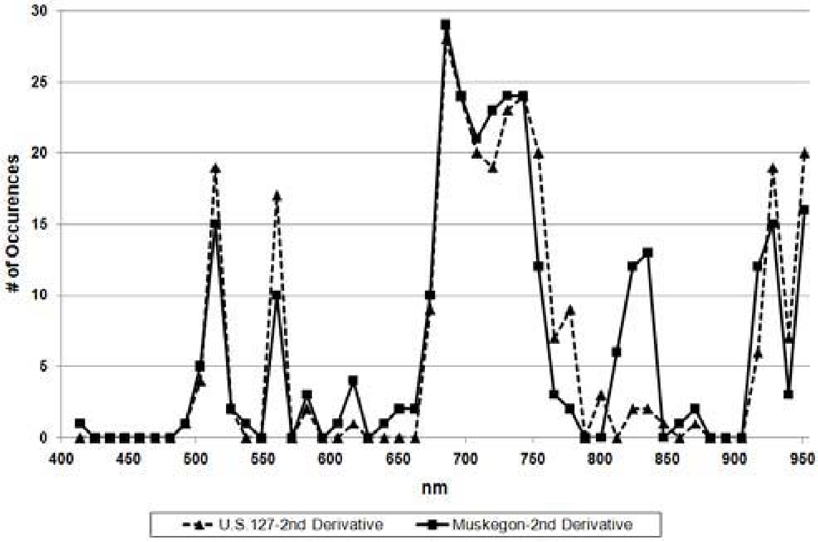
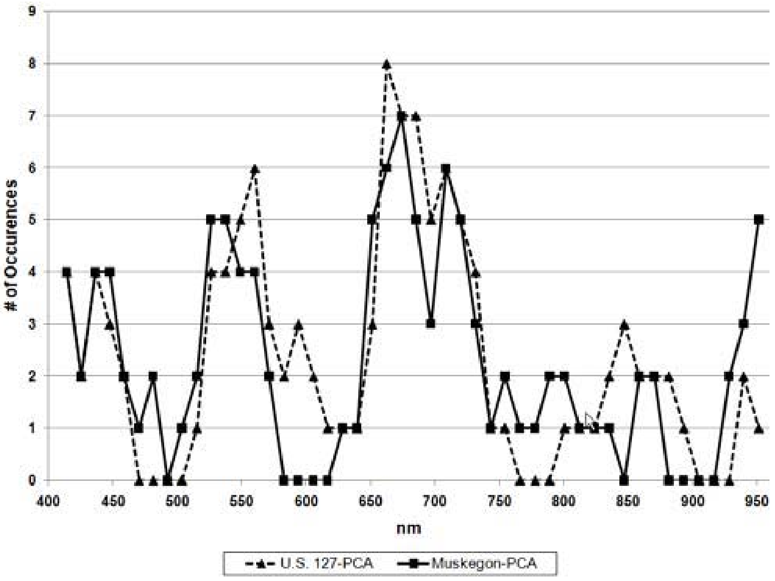


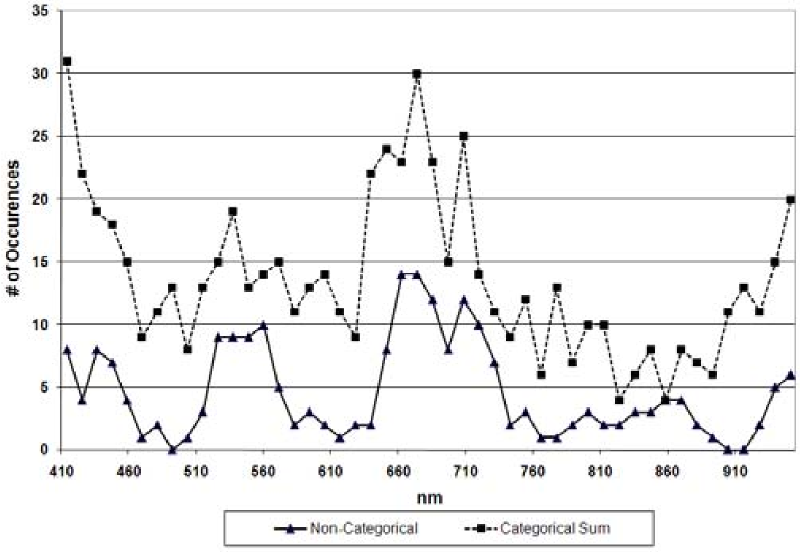
Figure 1. Cont.



3.2. Botanical Sub-Categories

Summarizing, botanical sub-categorization had little effect on the results, and individual categorical outputs are not presented. PCA is very sensitive to the nature and variability of the input data (i.e., high data dependency), so the summation of the histogram output related to each botanical subcategory would highlight any obvious differences when compared to the non-categorized results. Figure 2 contains the output histograms of the summed six botanical subcategories in relation to the non-categorized data, and the location of the most frequently occurring loading/band centers are clearly superimposed upon one another. No difference between sub-categories was found using PCA.

Figure 2. The number of times each band had a top five positive or negative loading value derived from the combined covariance and correlation PCA of all the spectra combined *vs*. the summation of the high-value occurrences from the PCA of each botanical category computed separately.



3.3. Covariance PCA versus Correlation PCA

Correlation-based PCA identifies slightly different bands of importance compared to covariance-based PCA using the same data set. For example, using the 48-band data set one-third (14 out of 48 bands) of the band locations produced occurrences of high-loading values that differed by two or more number of occurrences between correlation- and covariance-based PCA. The bands that displayed the greatest differences were 708, 939 and 951 nm, only one of which is found to be important via the derivative approach outlined by Becker *et al.* [3] and Torbick *et al.* [4]. Neither PCA method consistently identified the bands of importance selected by 2nd-derivative analysis. These results suggest that both types of PCA should be computed and used with other techniques to confirm utility. In addition, the sums of the occurrences of the high-value loadings from both methods may be more reliable to identify spectral zones of importance compared to using one approach alone. Therefore, computing both methods is recommended.

The large-magnitude sums of the occurrences of the high-value loadings from both covariance-based and correlation-based PCA generally identify only broad spectral zones of interest (i.e., NIR edge), many of which contain the optimum band locations determined Torbick *et al.* [4]. For both sites, the optimum band at 425 nm was bracketed by large-magnitude sums of occurrences. The optimum band pairs at 514/560 and 635/731 nm were either within or on the shoulder of a zone of large-magnitude sums of occurrences of the high-value loadings. The 812 and 939 nm optimum bands, on the other hand, are only weakly indicated as important by the combined PCA methods. In fact, PCA suggests that the bands centered at 847 and 858 are more important. Overall, the PCA results appear to contain more noise and are less diagnostic than the 2nd-derivative method. As a result, 2nd-derivative analysis appears to be better suited to the task of defining optimal band sets.

3.4. Re-Sampling Strategies

The PCA results from these data sets were compared in order to identify if re-sampling causes significant shift in the bands identified. Generally, the dominant peaks generated from all three data sets identify strikingly similar bands centers. The ASD data does display more variation (noise) due to the significantly higher number of bands, but all the dominant bands identified via all three data sets were largely identical. Re-sampling to better emulate the bandset of viable remote sensing platforms appears to have no major influence on the optimal bands identified via PCA.

4. Conclusions

We draw four overall conclusions from these results:

1. The data dependency of PCA makes it inappropriate for optimal band selection when used alone.
- It does not promote *diagnostic* comparison of multiple sites.
- It prohibits an analyst from exploring the biophysical reflectance differences between the categories without thorough interpretation of the bands identified.

2. Neither correlation-based nor covariance-based PCA consistently identified similar spectral bands other than the beginning of the NIR edge (~700), so both types of PCA should be computed. Combined PCA bands identified included 425, 514/560 and 635/731.

3. The large-magnitude sums of the occurrences of the high-value loadings from both covariance-based and correlation-based PCA identify general spectral zones of interest that are similar to the optimum band locations found in other studies [i.e., 1-4].
4. In-situ hyperspectral data with extremely narrow (i.e., 1-3 nm) bands need to be spectrally re-sampled in order to define useable, optimal bands via PCA.

Acknowledgements

This work has been funded in part by the Great Lakes Fisheries Trust, Society of Wetlands Scientists, Ducks Unlimited, and the Centre for Global Change and Earth Observations at Michigan State University.

References and Notes

1. Schmidt, K.; Skidmore, A. Spectral discrimination of vegetation types in a coastal wetland. *Remote Sens. Environ.* **2003**, *85*, 92-108.
2. Becker, B.; Lusch, D.; Qi, J. Identifying optimal spectral bands from in situ measurements of great Lakes coastal wetlands using second derivative analysis. *Remote Sens. Environ.* **2005**, *97*, 238-248.
3. Becker, B.L.; Lusch, D.P.; Qi, J. A classification-based assessment of the optimal spectral and spatial resolutions of coastal wetland imagery. *Remote Sens. Environ.* **2007**, *108*, 111-120.
4. Torbick, N.; Becker, B. Characterizing field-level hyperspectral measurements for identifying wetland invasive plant species. In *Invasive Species: Detection, Impact and Control Editors*; Wilcox, C.P., Turpin, R.B., Eds; Nova Science Publishers: New York, NY, USA, 2009; pp. 97-115.
5. Campbell, J. *Introduction to Remote Sensing*. Guilford Press: New York, NY, USA, 2007.
6. Lillesand, T.; Kiefer, R.; Chipman, J. *Remote Sensing and Image Interpretation*. John Wiley & Sons: New York, NY, USA, 2004.
7. Holden, H.; LeDrew, E. Spectral discrimination of healthy and non-healthy corals based on cluster analysis, principal components analysis and derivative spectroscopy. *Remote Sens. Environ.* **1998**, *65*, 217-224.
8. Zhao, G.; Maclean, L. A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 841-847.
9. Schaepman-Strub, G.; Schaepmen, M.; Painter, T.; Dangel, S.; Martonchik, J. Reflectance quantities in optical remote sensing – definitions and case studies. *Remote Sens. Environ.* **2006**, *103*, 27-42.
10. Davis, J.; *Statistics and Data Analysis in Geology*. John Wiley & Sons: New York, NY, USA, 1986.
11. Rummel, R.J. *Applied Factor Analysis*. Northwestern University Press: Evanston, IL, USA, 1970.