Novel Histogram Based Unsupervised Classification Technique to Determine Natural Classes From Biophysically Relevant Fit Parameters to Hyperspectral Data

Cooper McCann, Kevin S. Repasky, Mikindra Morin, Rick L. Lawrence, and Scott Powell

***Abstract*—Hyperspectral image analysis has benefited from an array of methods that take advantage of the increased spectral depth compared to multispectral sensors; however, the focus of these developments has been on supervised classification methods. Lack of *a priori* knowledge regarding land cover characteristics can make unsupervised classification methods preferable under certain circumstances. An unsupervised classification technique is presented that utilizes physically relevant basis functions to model the reflectance spectra. These fit parameters used to generate the basis functions allow clustering based on spectral characteristics rather than spectral channels and provide both noise and data reduction. Histogram splitting of the fit parameters is then used as a means of producing an unsupervised classification. Unlike current unsupervised classification techniques that rely primarily on Euclidian distance measures to determine similarity, the unsupervised classification technique uses the natural splitting of the fit parameters associated with the basis functions creating clusters that are similar in terms of physical parameters. The data set used in this work utilizes the publicly available data collected at Indian Pines, Indiana. This data set provides reference data allowing for comparisons of the efficacy of different unsupervised data analysis. The unsupervised histogram splitting technique presented in this paper is shown to be better than the standard unsupervised ISODATA clustering technique with an overall accuracy of 34.3/19.0% before merging and 40.9/39.2% after merging. This improvement is also seen as an improvement of kappa before/after merging of 24.8/30.5 for the histogram splitting technique compared to 15.8/28.5 for ISODATA.**

***Index Terms*—Agriculture, biophysics, clustering methods, remote sensing.**

Manuscript received November 21, 2016; revised March 28, 2017; accepted May 1, 2017. Date of publication May 22, 2017; date of current version September 20, 2017. This paper is based upon work supported the U.S. Department of Energy and the National Energy Technology Laboratory under Award Number DE-FC26-05NT42587. *(Corresponding author: Cooper McCann.)*

C. McCann is with Montana State University, Bozeman, MT 59717 USA (e-mail: cooper.mccann@montana.edu).

K. S. Repasky is with the Electrical and Computer Engineering Department, Montana State University, Bozeman, MT 59717 USA (e-mail: repasky@ece.montana.edu).

M. Morin, R. L. Lawrence, and S. Powell are with the Land Resources and Environmental Sciences Department, Montana State University, Bozeman, MT 59717 USA (e-mail: mikindra.morin@gmail.com; rickl@montana.edu; spowell@montana.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2017.2701360

I. Introduction

SUPERVISED classification techniques utilize training sets, which produce a spectral reflectance signature for the objects of interest. Once these spectral reflectance signatures are generated from the training sets, the reflectance spectra from each pixel in the entire image are compared to find the best matching class. Supervised classification techniques require some knowledge of the site, whether derived from other imagery or from site access, to define areas to be used for training sets. All potential classes must be known *a priori* in order to select a full set of training data. The success of the classification is largely dependent on the quality of the training sets [1] in terms of both the spectral accuracy and the statistical sampling used, meaning that the most accurate supervised classification technique depends on the specific data being investigated and the *a priori* knowledge available [2]–[4]. Unsupervised classification techniques cluster areas based on similar spectral features are typically used when training sets are not available and there is limited prior knowledge of the region under study [5].

Two common means of implementing unsupervised classifications involve ISODATA [6]–[8] and k-means [9], [10]. K-means is one of the simplest unsupervised classification techniques. It requires as an input the number of spectral clusters desired in a spectral image either through user input or through use of advanced preprocessing methods [9]–[13]. K-means can yield very different results depending on the number of clusters selected by the user, requiring user interaction in order to obtain physically meaningful clusters or classes. Additionally, depending on the specific user, different choices may be made as to which clusters to combine or which to split further, leading to different results in the final classes. Furthermore, k-means places equal values on all of the spectral channels, which can cause problems since channel-dependent noise could skew the classification. ISODATA has similar limitations, but being a more advanced algorithm has more user configurable parameters, and while the defaults work well for many clustering applications, the best values are dependent on the specific data being analyzed. Searching through this parameter space for the best clustering can be time consuming for a user, making this a powerful but difficult technique to use to its full potential.

In addition to ISODATA and k-means, many modern techniques exist that provide a high degree of accuracy on reference data [14]–[20].

This paper presents an unsupervised classification technique that addresses issues associated with hyperspectral imagery, primarily the issue of the large dimensionality is addressed by the introduction of a novel means of "band reduction" based on spectral fitting. The research goal was to develop an unsupervised classification technique that requires no user inputs in determining clusters, allowing for automated processing. The unsupervised classification scheme is a two-step process. The first step is to create a set of basis functions based on physically relevant fit parameters that model the major spectral features of interest. These basis functions provide noise and data reduction by fitting the spectral channels that make up the reflectance spectra with a modeled reflectance. Additionally, the physically relevant basis functions make it easier to interpret the results, since each cluster is based on overall spectral features not the reflectance value of particular spectral channels. The second step is to develop clusters by using the natural splitting associated with the histograms of the fit parameters. In contrast to the unsupervised classification techniques that rely on clustering based on *N*-dimensional Euclidian distances, the algorithm presented here is based on proximity of parameters via the usage of a histogram, but not on absolute measures such as Euclidian distance. The natural splits resulting from the histograms of the fit parameters can yield narrow or broad ranges of parameters depending on the parameter and the landscape. This type of clustering is fundamentally different than k-means or ISODATA, because it allows for multiple sizes, densities, and locations of clusters that are not subject to any type of Euclidian minimization.

This paper is organized as follows. Methodology is presented in Section II. A description of performance measures is presented in Section III. Results are presented and discussed in Section IV. Finally, some brief concluding comments are presented in Section V.

II. Methodology

A. Spectral Fitting Model

Data from the airborne visible infrared imaging spectrometer (AVIRIS) sensor consisted of 53 spectral channels covering the visible and near-IR spectral region (425–925 nm). This range was chosen as it represents the range of silicon detectors found in many commercial hyperspectral imagers. Extension of the basis functions to SWIR is possible with the development of a different set of basis functions making the technique very flexible. In this 425–925 nm spectral region, vegetation has a number of features that can be used for classification of land cover and vegetation health. Many of the differences between vegetation types can be subtle, so low noise instruments are important.

The fitting functions described below use the physical interpretation of the reflectance spectra to guide the model. The model has been designed to represent vegetation while at the same time having enough degrees of freedom to represent fallow regions. The technique presented here focuses on the classification of vegetation and fallow fields, but can be applied to a wider range of land cover types by using a different set of basis functions relevant to the physical content of the spectral image.

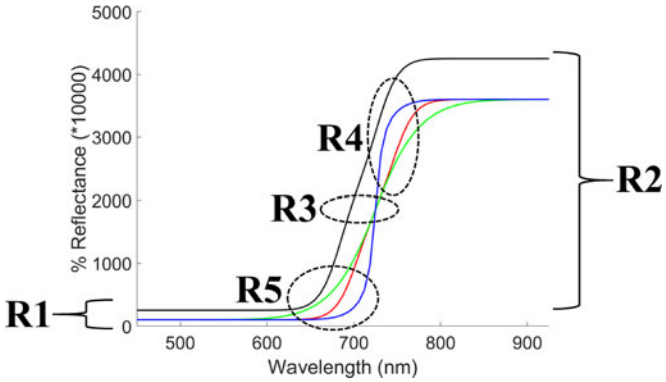

Fig. 1. Graphical representation of the different variables necessary to give the arctan function sufficient freedom to define the red edge well. R1 is the brightness of the pixel in the visible. R2 is similar to NDVI giving the difference between the brightness in the visible and the near-IR. R3 is the location of the inflection point giving the location of the red edge. R4 defines how steep the edge is in terms of reflectance versus wavelength. R5 defines the curvature of the function near the extremes.

The physically relevant basis functions used to model the spectral images presented in this work consist of two distinct functions that are summed to give the final modeled reflectance spectra. The two functions are referred to as the red edge function and the green peak (GP) function. The red edge function helps to define the baseline reflectance in the visible region, the location and behavior of the red edge, and the strength of the near-IR reflectance. The GP function defines the characteristics of the GP found in the visible region. The model uses a total of nine parameters to fit the reflectance spectra, and by reducing the 53 spectral channels of the AVIRIS data to these nine parameters reduced the data by a factor of 5.9.

With the success and widespread use of NDVI [21]–[23] and other vegetation indices [24]–[27] that rely on ratios between the near infrared and visible bands, the first function that was chosen was the inverse tangent (arctan) function (see Fig. 1) which serves as a type of step function. A step type function was chosen, because, in a broad sense, vegetation has a low reflectance in the visible and high reflectance in the near-IR. Given the spectral resolution of the hyperspectral camera, it is possible to determine significantly more information than simply a ratio between visible and near-IR as is done with an NDVI measurement. In order to have enough degrees of freedom to accurately represent the data, this part of the model requires five separate variables as shown in (1), which defines the red edge function. The first variable gives the baseline in the visible portion of the spectra, labeled $R1$. The second variable measures the difference between the visible baseline and the base level in the near-IR, labeled $R2$. The third variable, arguably the most important, is the location, in wavelength, of the inflection point of the arctan function, labeled $R3$. This value is a measure of the location of the red edge, and can be determined very precisely

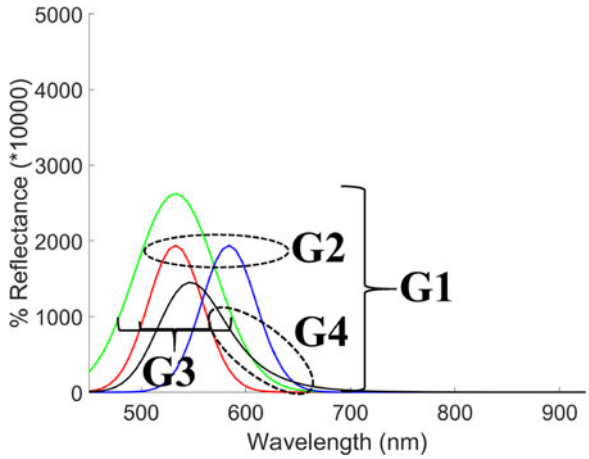

Fig. 2. Graphical representation of the different variables necessary to give the Gaussian sufficient freedom to define the green peak well. G1 defines the area, and height to a lesser degree of the peak. G2 defines the location of the center of the Gaussian peak. G3 defines the width of the Gaussian peak. G4 defines the exponential that modifies the Gaussian manifesting as an extended tail and an apparent shift in the peak location.

as compared to more simplistic means of determining the location of the red edge [28]–[34]. The fourth variable deals with how steep the rising edge of the arctan function is in terms of reflectance versus wavelength [35], labeled $R4$. The final variable is included to change the curvature of the function near its minimum and maximum to better match the curvature surrounding the red edge, labeled $R5$. Thus, the red edge function RE can be written in terms of the hyperspectral wavelengths λ_{HS} as

$$\mathrm{RE}(\lambda_{\mathrm{HS}}) = R2 * \left[\frac{\tan^{-1}\left((\lambda_{\mathrm{HS}} - R3) * R4 * e^{\frac{(\lambda_{\mathrm{HS}} - R3)^2}{R5}} \right)}{\pi} + \frac{1}{2} \right] + R1. \quad (1)$$

The red edge function with varying parameters is shown in Fig. 1 with the red line having average values and the green line having a decreased curvature near the edges determined by changing parameter *R*5. The blue line has steeper slope than the red line determined by parameter *R*4, and the black line is similar to the red line, but with larger parameters *R*1 and *R*2, and a shift in the red edge determined by parameter *R*3.

The GP function, the second basis function, is defined in (2) and shown graphically in Fig. 2. A normal cumulative distribution function given by $\mathrm{normcdf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$ is used in this GP basis function. As is common in natural phenomena, the reflectance peak was first assumed to have a Gaussian distribution as a function of wavelength; however, it was determined experimentally that it was insufficient to match the behavior seen in the vegetation spectra. In order to better model the "tail" of the reflectance peak that extends toward the near-IR, an exponentially modified Gaussian (exGaussian) was chosen. An exGaussian is a convolution of a Gaussian and an exponential. In the case of reflectance spectra from vegetation, this parameter more directly relates to the fact that Chlorophyll A and Chlorophyll B have different absorption profiles, especially in the red–yellow region. The parameters needed to define this function are similar to a Gaussian function with the first parameter related to the area of the peak, labeled $G1$. The second parameter is the location of the peak in wavelength space, labeled $G2$. The third is the width of the peak, labeled $G3$. The fourth and final parameter gives rise to the tail of the Gaussian, which causes a shift in the peak location, labeled $G4$. Thus, the GP function can be written in terms of the hyperspectral wavelengths λ_{HS} as

$$\mathrm{GP}(\lambda_{\mathrm{HS}}) = G1 * G4 * e^{\frac{(G3*G4)^2}{2} - (\lambda_{\mathrm{HS}} - G2) * G4} * \mathrm{normcdf}\left(\frac{\lambda_{\mathrm{HS}} - G2}{G3} - G3 * G4 \right). \quad (2)$$

A graphical representation of what changes in these parameters correspond to the model reflectance spectra is shown in Fig. 2 with the red line being normal values, the green line having a larger area determined by changing $G1$ and larger width determined by changing $G3$, the blue line is shifted by changing $G2$, and the black line has identical parameters to red with the exception of a tail because of a larger value of $G4$.

The strength of the model can be seen in Fig. 3 where the measured and modeled reflectance spectra are shown for pixels containing vegetation (upper left and upper right), sparse vegetation (lower left), and fallow (lower right). The black dots represent the measured reflectance spectra, the red-dotted line represents $\mathrm{RE}(\lambda_{\mathrm{HS}})$, the green-dashed line represents the $\mathrm{GP}(\lambda_{\mathrm{HS}})$, and the magenta line represents the total modeled reflectance spectra $\mathrm{RE}(\lambda_{\mathrm{HS}}) + \mathrm{GP}(\lambda_{\mathrm{HS}})$. As the model is designed for vegetation, the R^2 value is greater than 0.98 for most of the vegetative pixels. Reflectance spectra from pixels that are either partially or not vegetated show that the model has enough degrees of freedom to represent these reflectance spectra well, providing excellent fits with R^2 values greater than 0.95.

Fitting the reflectance spectra for each individual pixel with this model is a relatively time-consuming process on a standard desktop computer. For example, a quad core 2.7 GHz processor does fitting at a rate of approximately 30 pixels/s or 2.78 pixels/GHz/s. The fitting needs only be done once, is highly parallelizable, and could be done with a field programmable gated array if desired. After the fitting is complete, analysis of the entire image can be done in a number of ways with these physically relevant fit parameters. Specifically, the analysis examined here will center on an unsupervised classification scheme based on the natural splitting of histograms of the fit parameters.

B. Unsupervised Classification

An unsupervised classification technique was developed based on splitting the histograms associated with the fit parameters used to model the reflectance spectra. A histogram

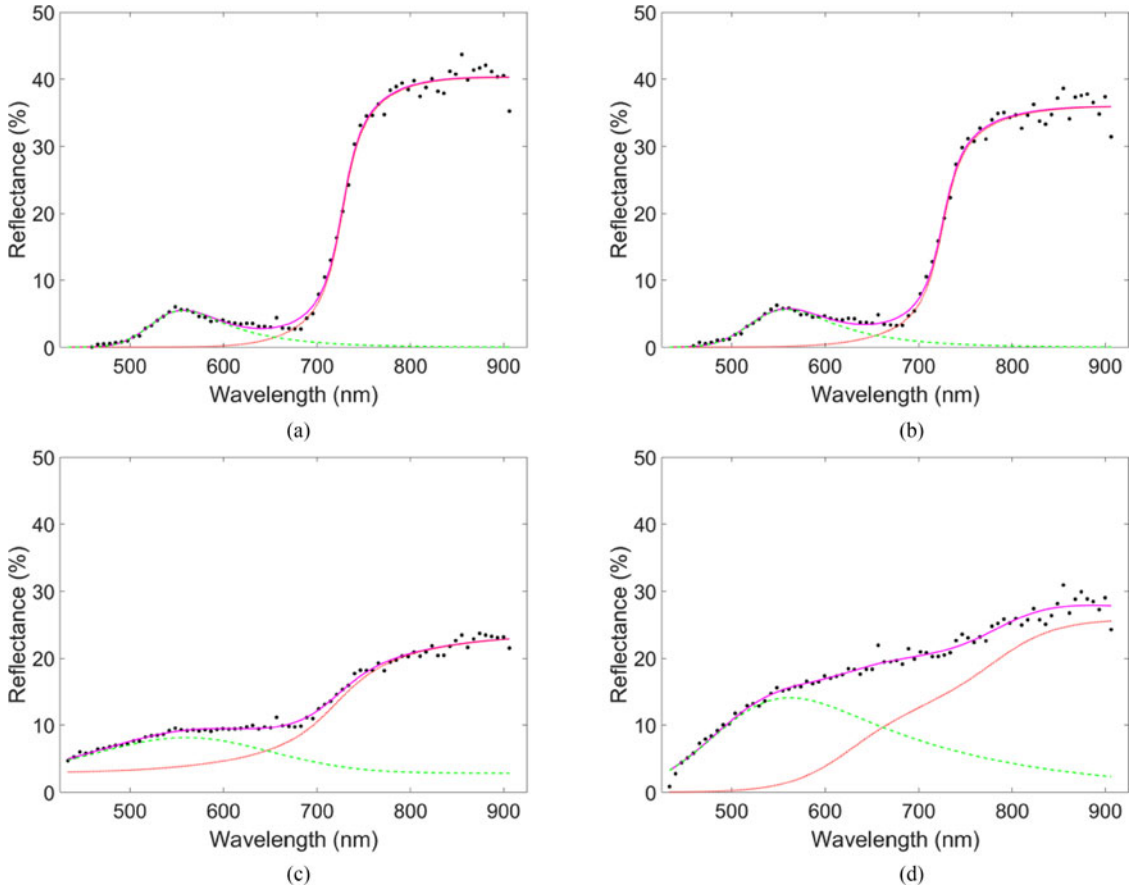

Fig. 3. Model fitting parameters for three different regions. Fits to a pixel containing vegetation (a, b), sparse vegetation (c), and fallow (d) are shown in solid magenta with the original spectra shown as individual black points. Individual fitting parameters are shown: Red-dotted line shows the red edge function. Green-dashed line shows the green function.

can be generated for each of the nine fit parameters by plotting the number of times a parameter falls within a certain range of values. Certain parameters will exhibit separable peaks that can be split into natural clusters. The unsupervised classification scheme begins by looking for a parameter with separable peaks to do an initial split into clusters. Within each cluster, another parameter with separable peaks is used to split the cluster into further subclusters. This process is repeated multiple times generating a clustered image that can be classified or analyzed as desired.

The first step in applying the unsupervised histogram splitting classification scheme is to provide an initial set of clusters for the spectral image. This can be accomplished by providing a set of broad seed clusters or defining the whole image as a single cluster. Next, one of the fit parameters is chosen, either randomly or simply by starting with the first parameter $R1$, and a histogram is generated for all of the pixels within this cluster. For this work, the histogram was generated by taking the range for the parameter of interest P_n, and dividing it by the number of parameter bins *m*, so that

$$\Delta P_n = (P_{n,\max} - P_{n,\min})/m \tag{3}$$

where $P_{n,\max}$ and $P_{n,\min}$ are the maximum and minimum values of P_n. The appropriate value for m is determined by the number of data points being examined, here taken to be five times the number of data points. The histogram is then a plot of the number of times a parameter value falls within the range $i\Delta P_n$ and $(i+1)\Delta P_n$ where i is an integer. In order to minimize the number of extraneous peaks or valleys that may be found in the data, the histogram is smoothed using a loess option of the smooth command in MATLAB. The loess option utilizes a local regression using weighted linear least squares and a second degree polynomial model. This choice, while it may change the specific values of the peaks and valleys, preserves their existence and location. The histograms generated for the various parameters plotted on a log scale are shown in Fig. 4. From these histograms, peaks and valleys are identified using the findpeaks command in MATLAB.

With the peaks and valleys in the smoothed histogram found, the next step is to determine where to split the histogram to generate subclusters. For each parameter value corresponding to a peak in the histogram, the higher and lower parameter values where the nearest valleys are located are determined. In a broad sense, these valleys are where the data could be split, but in practice there are three different cases that must be taken into account in order to intelligently split the data as shown in Fig. 5. The simplest and most ideal case is where there is a single valley between two adjacent peaks. This di-

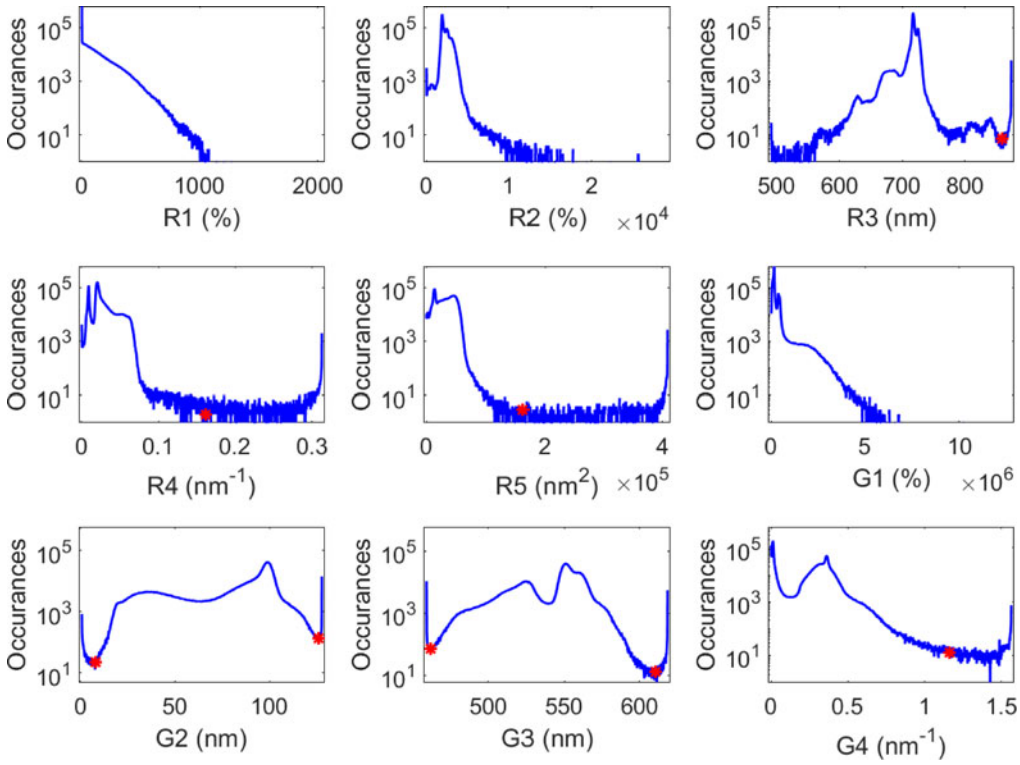

Fig. 4. Histograms of parameters found by fitting each individual pixel, note that the vertical axis is on a log scale. Labels refer to parameters described in Section II. Red asterisks show potential splitting locations. These histograms are for an entire image and become greatly simplified as splitting continues.

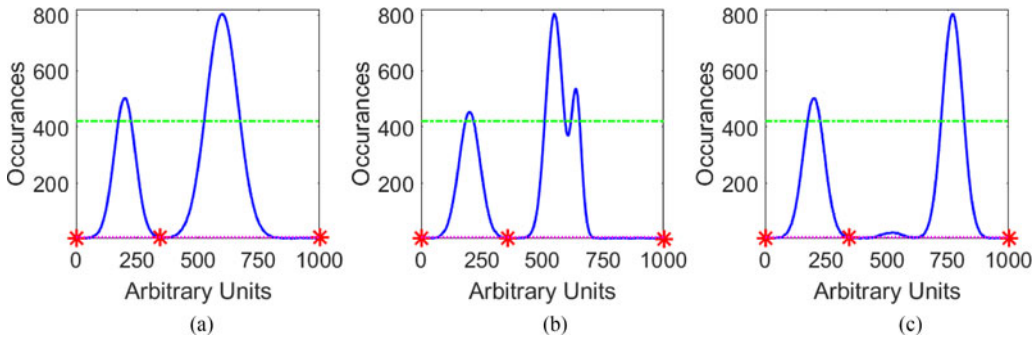

Fig. 5. Example data (blue) showing where the histogram splitting technique would split (red asterisks) different types of data. The left plot (a) shows the simplest case with a single peak between each set of valleys. The middle plot (b) shows how multiple peaks are contained between two valleys. Finally, the right plot (c) shows how the center of mass is used to determine that the small middle peaks should be associated with the rightmost peak since it is closer, but is too small to be considered an independent peak. In each plot the dotted magenta line shows the highest value that a valley can be as determined by 20% of the data being below this value, and the green-dashed line shows the lowest value a peak can be as determined by 20% of the data being above this value.

rectly leads to splitting the data at the location of the valley, Fig. 5 (left). The second case is where there are two or more peaks within the same set of valleys. This case only requires not taking into account the redundant peaks, Fig. 5 (middle). The final case is the most interesting and most common case, when there is a disagreement between where the split should occur, because there are multiple valleys between the peaks. The data between the valleys must be assigned to one of the adjacent peaks; this is accomplished by determining the center of mass (COM) of the data between the peaks and assigning it to the side it is closest to, Fig. 5 (right). The COM is defined to be

$$\text{COM} = \frac{\sum \Delta E * x}{\sum \Delta E} \tag{4}$$

where ΔE is the number of elements in the histogram bin centered at x as x takes on all of the values of the bin centers between the valleys of interest. While there could be other choices made in this regard, this was the simplest to implement, and is based on the idea that proximity of the parameters corresponds to being more physically similar.

With the best locations to split the data determined, the next step is to split the cluster into its subclusters. Care is taken to keep similar clusters near each other in cluster space. This is accomplished simply by maintaining the numbering scheme when a cluster is split. For example, an image initially has clusters 1–5 and then cluster 2 is split into two subclusters. These new clusters are inserted into the space where the original cluster was located, the new clustering now has clusters 1, (2, 3), 4, 5, 6, where (2, 3) was previously cluster 2.

This process is repeated for each cluster for the current parameter of interest. Then a new parameter is chosen, again randomly or numerically, and each cluster is tested against this new parameter. This process repeats until all of the parameters have been tested multiple times. One of the drawbacks of this approach is that the order in which the parameters are evaluated can make a difference as to whether a cluster can be split. To this end, the parameters are evaluated multiple times until no further natural splittings are possible, usually this takes 5–10 passes through the 9 parameters. Whether the parameters are evaluated sequentially or randomly does not matter as long as each parameter is examined enough times such that there are no further possible splits in the data.

C. Automatic Clustering

Combining clusters automatically can be done in a number of ways from spectral proximity to spatial location. Here, a simple method of spatial proximity is used to combine clusters. One of the features of the histogram splitting technique is that cluster numbers that are close directly correlate with clusters that are similar so there is an additional dimension that can be used to combine clusters. Determining spatial proximity was done by using a co-occurrence matrix with 20 nearest neighbors over the entire range of cluster numbers. The simplest choice in classification space is to only look at the spatial proximity of clusters with a single classification step (i.e., cluster 41 is only compared to cluster 40, and cluster 42). In order to give equal weight to each potential comparison, the co-occurrence matrix is scaled to the cluster size. This can be thought of as a percent co-occurrence so a value of 20 would correspond to every pixel in a class being completely surrounded by 20 pixels of a second class and a value of 0 would correspond to no pixel being within two pixels of the second class.

Looping through the percent co-occurrence matrix and recombining classes which are above a certain threshold reduces the number of classes, but does not handle small classes effectively. Looping through the remaining classes with a co-occurrence matrix and combining small clusters based on their spatial proximity only (ignoring the cluster number) and removing clusters below a certain size greatly reduces the number of classes. Due to the nature of the histogram splitting, a large number of classes have only 10 s of pixels each. This is due to the splitting nature of the technique, and why the histogram technique excels in datasets with large numbers of pixels.

After these two loops though the co-occurrence matrix the histogram splitting technique clusters have been reduced from approximately 200 down to approximately 20. Since ISODATA does not have a classification dimension we cannot apply the first loop directly, instead a measure of contrast is used. Contrast is determined by taking a ratio between 95% and 50% in a given row of the matrix, this determines how strongly two clusters are related spatially while diminishing the contribution of large randomly distributed classes. This value is again scaled to the size of a cluster to equally weight all clusters. The largest contrast value is determined and those two clusters are merged. After looping through the co-occurrence matrix and combining clusters based off of this scaled contrast the number of clusters is reduced to approximately 20 for both the band-based and parameter-based ISODATA clustering. Classes are then assigned based on maximizing the trace of the confusion/error matrix. A single error matrix is shown in Table I for the merged histogram splitting.

A comparison of clusters before and after merging is shown in Fig. 6 for the three different classification methods. Qualitatively the differences between using parameters versus bands for ISODATA give very different results, but after merging begin to converge on clusters. A number of fields are clustered similarly between the three methods potentially pointing to variability within the fields that is not expressed in the reference data. Qualitative measurements of the accuracy of the methods are discussed in Section IV.

III. Performance Measures

In order to evaluate the performance of the histogram method versus ISODATA, a classification matrix was first determined based off of location of ground reference data to clusters contained within the area. Since the number of clusters found is not constrained the matrix will be $n * c$ where n is the number of reference data classes and c is the number of clusters found. A typical entry q_{ij} shows how many samples belonging to class i have been assigned to class j. A perfect method would produce an $n * n$ matrix with nonzero values only along the diagonal.

Performance measures evaluated are: individual class accuracy (PA_i) or producer's accuracy, reliability (UA_i) or user accuracy, average accuracy $(\overline{PA})$, average reliability $(\overline{UA})$, overall accuracy (A_{tot}), and Cohen's kappa (κ) [36], [37]. These measures are defined to be

$$PA_i = \frac{q_{ii}}{\sum_{j=1}^{n_c} q_{ij}} \tag{5}$$

$$UA_i = \frac{q_{ii}}{\sum_{i=1}^{n_c} q_{ij}} \tag{6}$$

$$\overline{PA} = \frac{1}{n_c} \sum_{i=1}^{n_c} PA_i \tag{7}$$

$$\overline{UA} = \frac{1}{n_c} \sum_{i=1}^{n_c} UA_i \tag{8}$$

$$A_{\text{tot}} = \frac{1}{N} \sum_{i=1}^{n_c} q_{ii} \tag{9}$$

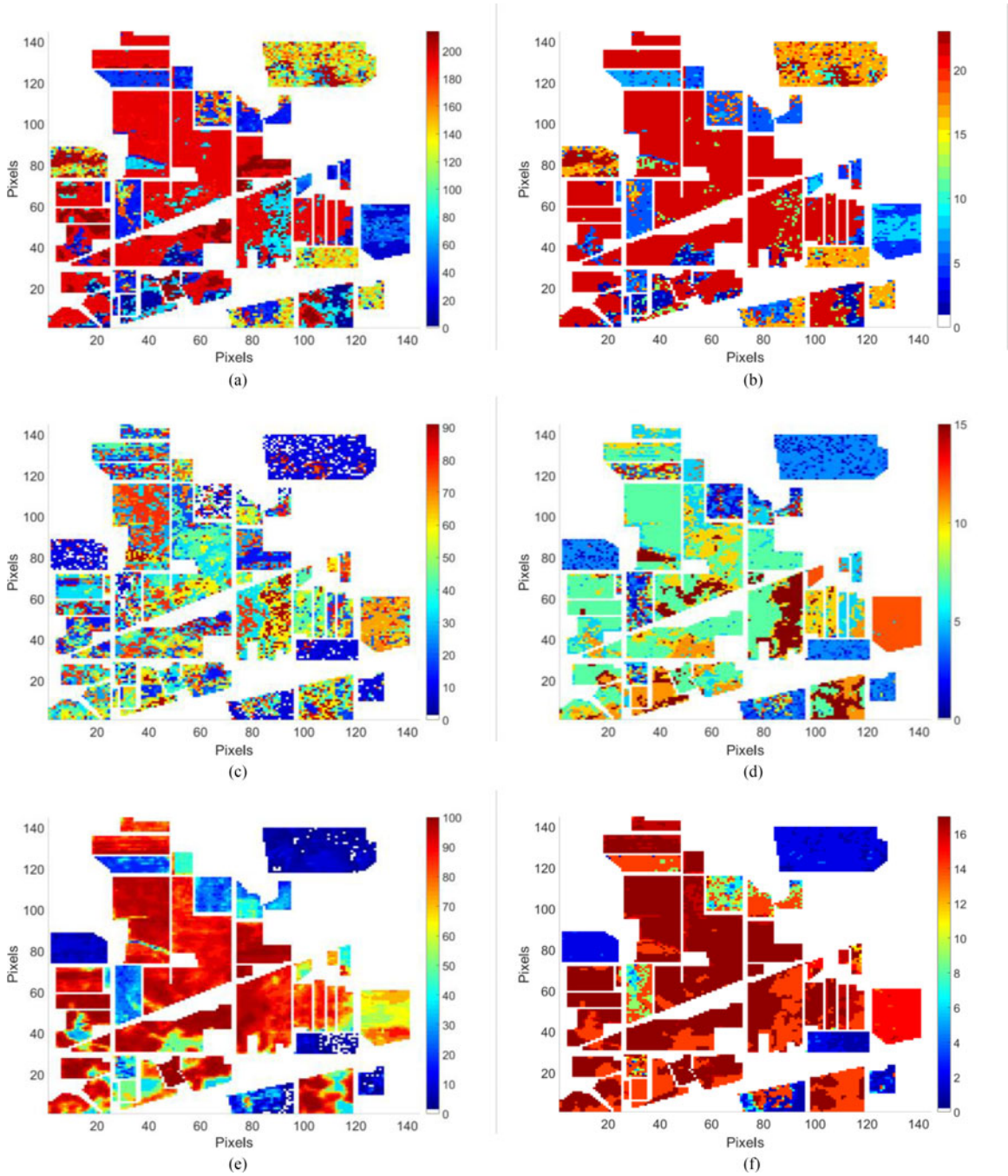

Fig. 6. Indian Pines classified using histogram splitting (a, b), ISODATA on the fit parameters (c, d), and ISODATA on the spectral bands (e, f); before (a, c, e) and after (b, d, f) automatically combining clusters are shown with a mask corresponding to available reference data. Vertical color scale corresponds to cluster numbers.

where q_{ii} are correctly classified samples, n_c is the number of samples in class c_i, and N is the number of samples in the entire dataset. The entire dataset of pixels with reference data was used for performance measures.

Cohen's kappa coefficient is a statistic which measures inter-rater agreement for categorical items and is generally thought to be a more robust measure than simple percent agreement calculation, since it takes into account the agreement occurring by chance.

IV. Results and Discussion

The publicly available Indian Pines data set [38] was acquired with the AVIRIS sensor over the Indian Pines test site in Northwestern Indiana. This data set was 145 $\times$ 145 pixels (pixels correspond to an area of approximately 20 m $\times$ 20 m) in extent and originally contained 224 spectral reflectance bands in the wavelength range 400–2500 nm, which was reduced to 200 bands after removal of bands covering water absorption. The Indian Pines dataset is designated into 16 classes

TABLE I
ERROR MATRIX FOR MERGED HISTOGRAM SPLITTING

Error matrix for Merged Histogram Splitting. Class labels (A1 – A16) are defined in Table II and A17 being unclassified pixels.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	PA
A1	0	0	0	0	0	11	0	27	0	0	0	0	0	0	0	0	8	0.00
A2	0	216	0	0	0	47	0	0	0	0	946	0	0	2	0	0	217	15.1
A3	0	61	9	0	0	21	0	42	0	16	654	9	2	1	0	0	15	1.08
A4	0	56	0	2	0	35	0	13	0	3	118	4	2	0	0	0	4	0.84
A5	0	0	1	0	7	124	0	7	0	0	23	1	5	294	0	0	21	1.45
A6	0	0	0	0	0	453	0	0	0	0	0	0	0	265	0	0	12	62.1
A7	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0.00
A8	0	0	0	0	0	249	0	199	0	0	0	0	0	0	0	0	30	41.6
A9	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0.00
A10	0	70	0	0	0	19	0	0	0	211	603	0	0	0	0	0	69	21.7
A11	0	0	0	0	0	0	0	0	0	0	1899	0	0	0	0	0	556	77.4
A12	0	99	0	0	0	56	0	85	0	85	205	28	0	0	0	0	35	4.72
A13	0	0	0	0	0	181	0	0	0	0	10	0	14	0	0	0	0	6.83
A14	0	0	0	0	0	0	0	0	0	0	1	0	0	1149	0	0	115	90.8
A15	0	0	0	0	1	161	0	8	0	0	15	0	2	173	3	0	23	0.78
A16	0	61	3	0	0	0	0	6	0	15	3	1	0	0	0	1	3	1.08
UA	NaN	38.4	69.2	100.	87.5	32.2	NaN	51.4	NaN	63.9	42.4	65.1	56.0	61.0	100.	100.		

Class labels (A1–A16) are defined in Table II and A17 being unclassified pixels.

approximately two-third agriculture land and one-third forest or other natural perennial vegetation and includes two major dual lane highways, a rail line, low density housing, other built structures, and smaller roads. The scene was acquired in June, so some of the crops present such as corn and soybeans are in early stages of growth with less than 5% coverage.

As a means of validating the unsupervised classification technique based on the histogram splitting method, data were clustered by the both the histogram method and ISODATA using the default settings in ERDAS Imagine (Number of classes = 1 to 200, minimum size = 0.01%, minimum distance = 4, maximum SD = 5, Max merges = 1, maximum iterations = 50, convergence threshold = 0.99). ISODATA was run on both the 53 spectral bands and the parameters from the fit model described in this paper. The resulting data are quantitatively shown in Table II showing individual class accuracy PA_i, reliability UA_i, $\overline{PA}$, $\overline{UA}$, $A_{\rm tot}$, and κ. The final clustering is displayed qualitatively in Fig. 6 to better show the spatial distribution of the resultant clusters.

The histogram splitting method was shown to be better than ISODATA on the reference Indian Pines data before any type of clustering in terms of Cohen's kappa and overall accuracy. After automated clustering/merging, both ISODATA and the histogram method were improved, with significant improvement to the ISODATA technique in terms of overall accuracy and Cohen's kappa. Average accuracy suffered as a number of small classes were not identified at all. Small classes such as "Alfalfa" and "Grass-pasture-mowed" are handled better before merging while medium to large classes, above approximately 400 pixels, are generally improved due to the nature of merging small clusters into larger clusters which is why average accuracy is reduced after merging.

Extensions to the histogram splitting technique would be to examine mixed dimensions as there may be splittings that do not occur strictly along the parameter axes, or possibly shaping *N*-dimensional surfaces around clusters in parameter space. Also further work on the automated clustering algorithm could be done to examine both parameter space and physical proximity at the same time could benefit any clustering technique as the simple implementation here showed.

Finally, due to the small number of pixels in this data set, the histograms generated and subsequently split are sparse, making splitting difficult to accurately determine. The histogram splitting technique is better suited for larger data sets where the histograms become denser. This denser histogram, Fig. 4 for example, allows for a more accurate determination of where the biophysically relevant parameters differ and can be split.

TABLE II
CLASSIFICATION EFFICACY FOR INDIAN PINES DATA SET COMPARING THE HISTOGRAM SPLITTING METHOD, ISODATA ON BIOPHYSICAL FIT PARAMETERS, ISODATA ON HYPERSPECTRAL BANDS BOTH BEFORE AND AFTER AUTOMATIC CLUSTER MERGING

Class type	Sample Size	Raw Histogram Splitting PA/UA%	Merged Histogram Splitting PA/UA%	Raw ISODATA on Parameters PA/UA%	Merged ISODATA on Parameters PA/UA%	Raw ISODATA on Bands PA/UA%	Merged ISODATA on Bands PA/UA%
A1: Alfalfa	46	15.2/77.8	0.00/NaN	50.0/100.	0.00/NaN	41.3/100.	0.00/NaN
A2: Corn-notill	1428	12.5/41.5	15.1/38.4	15.6/55.1	22.3/57.6	15.8/50.2	0.07/100.
A3: Corn-mintill	830	2.29/95.0	1.08/69.2	10.2/58.6	11.3/44.6	16.0/84.7	7.59/30.1
A4: Corn	237	1.69/66.7	0.84/100.	22.8/87.1	2.53/100.	18.6/100.	0.00/NaN
A5: Grass-pasture	483	9.32/52.3	1.45/87.5	5.59/100.	5.59/100.	31.9/99.4	3.52/73.9
A6: Grass-trees	730	55.1/42.9	62.1/32.2	20.6/60.2	20.1/60.2	11.5/97.7	30.8/91.5
A7: Grass-pasture-mowed	28	3.57/100.	0.00/NaN	64.3/100.	3.57/100.	60.7/100.	0.00/NaN
A8: Hay-windrowed	478	32.0/79.3	41.6/51.4	30.8/99.3	49.2/90.7	20.3/99.0	96.7/88.9
A9: Oats	20	20.0/100.	0.00/NaN	20.0/100.	20.0/100.	25.0/100.	0.00/NaN
A10: Soybean-notill	972	16.7/61.8	21.7/63.9	15.4/78.5	37.0/65.5	14.9/78.0	42.5/24.8
A11: Soybean-mintill	2455	68.4/42.1	77.4/42.4	16.1/46.0	47.7/36.9	20.3/55.1	74.2/37.8
A12: Soybean-clean	593	7.25/48.3	4.72/65.1	9.44/78.9	26.1/52.4	7.59/100.	0.00/NaN
A13: Wheat	205	15.6/54.2	6.83/56.0	31.7/84.4	31.7/84.4	41.5/89.4	0.00/NaN
A14: Woods	1265	60.7/58.6	90.8/61.0	52.2 /76.3	74.0/59.5	24.0/75.8	72.2/71.5
A15: Buildings-Grass-Trees-Drives	386	2.07/100.	0.78/100.	17.9/75.0	19.4/71.4	5.96/100.	26.9/96.3
A16: Stone-Steel-Towers	93	3.23/100.	1.08/100.	38.7/100.	0.00/NaN	68.8/65.3	0.00/NaN
	$\overline{PA}/\overline{UA}$	20.4/66.4	20.3/61.9	26.3/76.4	23.2/73.1	26.5/87.2	22.2/68.3
STATISTICS	A_{tot}	**34.3**	**40.9**	**21.1**	**35.1**	**19.0**	**39.2**
	κ	**24.8**	**30.5**	**17.6**	**26.9**	**15.8**	**28.5**

NaN values of the user accuracy are due to the respective algorithm not classifying any pixels in that class.

V. CONCLUSION

The ability to effectively utilize hyperspectral data for a variety of applications depends on the ability to develop processing algorithms that are efficient and lend themselves to automation. In this work, a two-step classification technique was presented. The first step in the unsupervised classification technique involves fitting the reflectance spectra to a set of physically relevant basis functions using a set of fit parameters. The second step in the unsupervised classification technique involves using the natural splitting of the histograms of the fit parameters. Additionally, a simple means of automated clustering involving spatial and cluster proximity was examined.

Fitting reflectance spectra as a technique is relatively new and the model presented, while simple, shows the power of such a technique. With improved signal-to-noise, finer structural detail can be resolved and with extended spectral range more structural features become accessible. It is feasible with a more advanced model and better sensors to glean information about plant structure, health, and composition without having to analyze hundreds of bands at a time.

The histogram splitting method was shown to be better than ISODATA on the reference Indian Pines data before any type of clustering, and with automated clustering both ISODATA and the histogram method were improved in relation to the reference data with the histogram method exceeding ISODATA in many of the classes.

ACKNOWLEDGMENT

Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

[1] C. Palaniswami, A. K. Upadhyay, and H. P. Maheswarappa, "Spectral mixture analysis for subpixel classification of coconut," *Curr. Sci.*, vol. 91, no. 12, pp. 1706–1711, 2006.

[2] G. Bogdan and L. Petrakieva, "Combining labelled and unlabelled data in the design of pattern classification systems," *Int. J. Approx. Reason.*, vol. 35, no. 3, pp. 251–273, 2004.

[3] K. Perumal and R. Bhaskaran, "Supervised classification performance of multispectral images," *J. Comp.*, vol. 2, no. 2, Feb. 2010.

[4] D. I. M. Enderle and R. C. Weih Jr., "Integrating supervised and unsupervised classification methods to develop a more accurate land cover classification," *J. Arkansas Acad. Sci.*, vol. 59, pp. 65–73, 2005.

[5] J. A. Richards. *Remote Sensing Digital Image Analysis*, vol. 3. Berlin, Germany: Springer, 1999.
[6] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recog. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[7] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, pp. 32–57, 1973.
[8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.
[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
[10] J. A. Hartigan, *Clustering Algorithms (Probability & Mathematical Statistics)*. Hoboken, NJ, USA: Wiley, 1975.
[11] D. Pelleg and A. W. Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Proc. 17th Int. Conf. Mach. Learn.*, vol. 1, 2000, pp. 727–734.
[12] G. Hamerly and E. Charles, "Learning the k in k-means," *Adv. Neural Inf. Process. Syst.*, vol. 16, 2004, Art. no. 281.
[13] Z. Selim Shokri and A. Mohamed Ismail, "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 1, pp. 81–87, Jan. 1984.
[14] X. Han, Y. Zhong, and L. Zhang, "Spatial-spectral unsupervised convolutional sparse auto-encoder classifier for hyperspectral imagery," *Photogram. Eng. Remote Sens.*, vol. 83, no. 3, pp. 195–206, 2017.
[15] A. Mughees, X. Chen, and L. Tao, "Unsupervised hyperspectral image segmentation: Merging spectral and spatial information in boundary adjustment," in *Proc. 2016 55th Annu. Conf. Soc. Instrum. Control Eng. Jpn.*, Sep. 2016, pp. 1466–1471.
[16] A. K. Gautam and M. R. Bhutiyani, "Performance evaluation of Hyperspectral image segmentation implemented by recombination of PCT and bilateral filter based fused images, "in *Proc. IEEE 2016 3rd Int. Conf. Signal Process. Integr. Netw.*, Feb. 2016, pp. 152–156.
[17] A. Mehta and O. Dikshit, "Comparative study on projected clustering methods for hyperspectral imagery classification, "*Geocarto Int.*, vol. 31, no. 3, pp. 296–307, 2016.
[18] J. P. Papa, L. P. Papa, D. R. Pereira, and R. J. Pisani, "A hyperheuristic approach for unsupervised land-cover classification, "*IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 9, no. 6, pp. 2333–2342, Jun. 2016.
[19] B. B. Damodaran, N. Courty, and S. Lefèvre, "Unsupervised classifier selection approach for hyperspectral image classification, "in *Proc. 2016 IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2016, pp. 5111–5114.
[20] A. C. Karaca and M. K. Güllü, "Comparison of traditional and recent unsupervised band selection approaches in hyperspectral images, "in *Proc. 2016 24th Signal Process. Commun. Appl. Conf.*, May 2016, pp. 785–788.
[21] T. N. Carlson and D. A. Ripley, "On the relation between NDVI, fractional vegetation cover, and leaf area index," *Remote Sens. Environ.*, vol. 62, no. 3, pp. 241–252, 1997.
[22] R. S. DeFries and J. R. G. Townshend, "NDVI-derived land cover classifications at a global scale," *Int. J. Remote Sens.*, vol. 15, no. 17, pp. 3567–3586, 1994.
[23] M. Wójtowicz *et al.*, "Application of remote sensing methods in agriculture," *Commun. Biometry Crop Sci.*, vol. 11, pp. 31–50, 2016.
[24] J. M. Peña-Barragán, F. López-Granados, M. Jurado-Expósito, and L. García-Torres, "Mapping Ridolfia segetum patches in sunflower crop using remote sensing," *Weed Res.*, vol. 47, no. 2, pp. 164–172, 2007.
[25] J. Chang, D. E. Clay, K. Dalsted, S. Clay, and M. O'Neill, "Corn (L.) yield prediction using multispectral and multidate reflectance," *Agronomy J.*, vol. 95, no. 6, pp. 1447–1453, 2003.
[26] F. Li *et al.*, "Estimating N status of winter wheat using a handheld spectrometer in the North China Plain," *Field Crops Res.*, vol. 106, no. 1, pp. 77–85, 2008.
[27] H. Genc, L. Genc, H. Turhan, S. E. Smith, and J. L. Nation, "Vegetation indices as indicators of damage by the sunn pest (Hemiptera: Scutelleridae) to field grown wheat," *Afr. J. Biotechnol.*, vol. 7, no. 2, pp. 173–180, 2008.
[28] D. N. H. Horler, M. Dockray, and J. Barber, "The red edge of plant leaf reflectance," *Int. J. Remote Sens.*, vol. 4, no. 2, pp. 273–288, 1983.
[29] G. Carter and A. Knapp, "Leaf optical properties in higher plants: Linking spectral characteristics to stress and chlorophyll concentration," *Amer. J. Botany*, vol. 88, no. 4, pp. 677–684, 2001.
[30] W. Smith *et al.*, "Leaf form and photosynthesis," *BioScience*, vol. 47, no. 11, pp. 785–793, 1997.
[31] P. Zarco-Tejada, "Chlorophyll fluorescence effects on vegetation apparent reflectance I. Leaf-level measurements and model simulation," *Remote Sens. Environ.*, vol. 74, no. 3, pp. 582–595, 2000.
[32] A. A. Gitelson, M. N. Merzlyak, and H. K. Lichtenthaler, "Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm," *J. Plant Physiol.*, vol. 148, no. 3, pp. 501–508, 1996.
[33] T. P. Dawson and P. J. Curran, "Technical note; A new technique for interpolating the reflectance red edge position," *Int. J. Remote Sens.*, vol. 19, pp. 2133–2139, 1998.
[34] C. M. Azong and A. K. Skidmore, "A new technique for extracting the red edge position from hyperspectral data: The linear extrapolation method," *Remote Sens. Environ.*, vol. 101, no. 2, pp. 181–193, 2006.
[35] J. R. Miller, E. W. Hare, and J. Wu, "Quantitative characterization of the vegetation red edge reflectance 1. An inverted-Gaussian reflectance model," *Remote Sens.*, vol. 11, no. 10, pp. 1755–1773, 1990.
[36] J. Senthilnath *et al.*, "Crop stage classification of hyperspectral data using unsupervised techniques," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 6, no. 2, pp. 861–866, Apr. 2013.
[37] J. Cohen, "Kappa: Coefficient of concordance," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960.
[38] M. Baumgardner, L. Biehl, and D. Landgrebe, "220 band aviris hyperspectral image data set: June 12, 1992 Indian pine test site 3," Purdue University Research Repository, 2015. Doi: 10.4231/R7RX991C.

Cooper McCann received the B.S., M.S., and Ph.D. degrees in physics from Montana State University, Bozeman, MT, USA, in 2004, 2015, and 2017, respectively.

He has worked with high power pulsed laser, nonlinear optics, and for several small startup optics companies. He is continuing his work with hyperspectral data in industry. His research interests include hyperspectral imagery, "big data" processing, general optics, and teaching.

Kevin S. Repasky received the B.E. degree in mechanical engineering from Youngstown State University, Youngstown, OH, USA, in 1987, and the M.S. and Ph.D. degrees in physics from Montana State University, Bozeman, MT, USA, in 1992 and 1996, respectively.

He is currently a Professor in the Electrical and Computer Engineering Department, Montana State University and has recently been appointed a Visiting Scientist in the National Center for Atmospheric Research. His research interests include laser and photonic development, nonlinear optics, optical remote sensing, and atmospheric science.

Mikindra Morin received the B.S. degree in health science from the Clemson University, Clemson, SC, USA, in 2010 and the J.D. degree with an emphasis on environmental law from the University of Arizona, Tucson, AZ, USA, in 2013.

She is a Master's Candidate in the Land Resources and Environmental Science Department at Montana State University working under the guidance of Dr. Rick Lawrence in the Spatial Sciences Center. Her research focuses on using remote sensing applications to detect vegetation stress over agricultural management areas.

Rick L. Lawrence received the B.A. degree in political science from the Claremont McKenna College, Claremont, CA, USA, the J.D. degree from the Columbia University School of Law, New York, NY, USA, and the M.S. and Ph.D. degrees in forest resources from Oregon State University, Corvallis, OR, USA, in 1976, 1979, 1995, and 1998, respectively.

He practiced law from 1979 to 1993 with the law firms or Hughes, Hubbard, & Reed, Adams, Duque, and Hazeltine, and LeBoeuf, Lamb, Leiby, & McCrea. In 1998, he joined the Land Resources and Environmental Sciences Faculty, Montana State University, Bozeman, MT, USA, where he is currently a Professor and the Director of the Spatial Sciences Center. His research interests include a wide range of applications related to natural resource management, from carbon sequestration, to crop and range issues, and to forest and wildlife. He also has a strong interest in algorithm development for classification of remotely sensed imagery and specializing in classification tree analysis and its variants.

Dr. Lawrence is a member of the American Society of Photogrammetry and Remote Sensing, the Society of American Foresters, the American Geophysical Union, and the California Bar Association (inactive).

Scott Powell received the B.A. degree in biology and environmental science from Macalester College, St. Paul, MN, USA, in 1993, the M.E.M. degree in resource ecology from the Nicholas School of the Environment, Duke University, Durham, NC, USA, in 1997, and the Ph.D. degree in ecology from Montana State University, Bozeman, MT, USA, in 2004.

He then completed a 3-year postdoctoral position with the U.S.D.A. Forest Service, Pacific Northwest Research Station, Corvallis, OR. In 2009, he joined the faculty as an Assistant Research Professor in the Department of Land Resources and Environmental Sciences, Montana State University, transitioning to an Assistant Professor in 2014. His research interests and teaching focus on the use of geospatial data to characterize ecosystem and landscape processes, including quantification and monitoring of carbon sequestration at broad spatial and temporal scales, from forest ecosystems to dryland agricultural systems, to human-engineered geologic sequestration sites.