METHODOLOGY FOR HYPERSPECTRAL BAND SELECTION

Peter Bajcsy and Peter Groves
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign
605 East Springfield Avenue, Champaign, IL 61820
pbajcsy@ncsa.uiuc.edu and pgroves@ncsa.uiuc.edu
Published in Photogrammetric Engineering and Remote Sensing journal, Vol. 70, Number 7, July 2004, pp. 793-802.

ABSTRACT

While hyperspectral data are very rich in information, processing the hyperspectral data poses several challenges regarding computational requirements, information redundancy removal, relevant information identification, and modeling accuracy. In this paper we present a new methodology for combining unsupervised and supervised methods under classification accuracy and computational requirement constraints that is designed to perform hyperspectral band (wavelength range) selection and statistical modeling method selection. The band and method selections are utilized for prediction of continuous ground variables using airborne hyperspectral measurements. The novelty of the proposed work is in combining strengths of unsupervised and supervised band selection methods to build a computationally efficient and accurate band selection system. The unsupervised methods are used to rank hyperspectral bands while the accuracy of the predictions of supervised methods are used to score those rankings. We conducted experiments with seven unsupervised and three supervised methods. The list of unsupervised methods includes information entropy, first and second spectral derivative, spatial contrast, spectral ratio, correlation and principal component analysis ranking combined with regression, regression tree and instance based supervised methods. These methods were applied to a data set that relates ground measurements of

soil electrical conductivity with airborne hyperspectral image values. The outcomes of our analysis led to a conclusion that the optimum number of bands in this domain is the top 4 to 8 bands obtained by the entropy unsupervised method followed by the regression tree supervised method evaluation. Although the proposed band selection approach is demonstrated with a data set from the precision agriculture domain, it applies in other hyperspectral application domains.

1 INTRODUCTION

Recent development of advanced hyperspectral sensors has enabled better class discrimination of objects due to a higher spectral resolution than one could achieve with standard electro-optical (EO) and infrared (IR) sensors. Hyperspectral sensors generate imagery that captures surface and sub-surface properties of objects, e.g., 1-2 mm depth in fine textured soils and 1-2 cm in coarse sands (Lee, 1978), at a fine spectral resolution, e.g., 10 nm using AVIRIS data (Campbell, 1996), and provide non-invasive and non-intrusive reflectance measurements. Hyperspectral image analysis has been applied in several GIS application areas (Campbell, 1996; Miller and Han, 1999) including environmental monitoring (Czillag *et al.,* 1993; Yamagata, 1996; Warner *et al.,* 1999; Merenyi *et al*., 2000), sensor design (Wiersma and Landgrebe, 1980; Price, 1994), geological exploration (Hughes, 1968; Benediktsson, *et al.,* 1995; Merenyi *et al.,* 1996), agriculture (Gopalapillai and Tian, 1999), forestry (Pu and Gong, 2000), security (Healey and Slater, 1999), cartography and military (Jia and Richards, 1994; Withagen, 2001). Common problems in the area of hyperspectral analysis involving data relevancy include optimal selections of wavelength, number of bands, and spatial and spectral resolution (Wiersma and Landgrebe, 1980; Price, 1994; Jasani and Stein, 2002). Additional issues

include the modeling issues of scene, sensor and processor contributions to the measured hyperspectral values (Warner *et al.,* 1999), finding appropriate classification methods (Benediktsson, *et al.,* 1995), and identifying underlying mathematical models (Hughes, 1968). Every problem formulation is usually also associated with multiple application constraints. For example, communication bandwidth, data storage, discrimination or classification accuracy, minimum signal-to-noise ratio, sensor and data acquisition cost must be addressed.

In almost all application areas, the basic goal of hyperspectral image analysis is to classify or discriminate objects. Driven by classification or discrimination accuracy, one would expect that, as the number of hyperspectral bands increases, the accuracy of classification should also increase. Nonetheless, this is not the case in a model-based analysis (Hughes, 1968; Benediktsson, *et al.,* 1995). Redundancy in data can cause convergence instability of models. Furthermore, variations due to noise in redundant data propagate through a classification or discrimination model. The same is true of spectral information that has no relation to the feature being classified in the underlying mathematical model. Such information is the same as noise to any statistical model, even if it is unique and accurate. Thus, processing a large number of hyperspectral bands can result in higher classification inaccuracy than processing a subset of relevant bands without redundancy. In addition, computational requirements for processing large hyperspectral data sets might be prohibitive and a method for selecting a data subset is therefore sought. Although a method for band selection leads to data compression, we would like to emphasize that the performance objective of data compression is based on

data size (communication bandwidth), which is different from classification or discrimination accuracy.

In this work, we will address the issue of hyperspectral band and method selection using unsupervised and supervised methods driven by classification accuracy and computational cost. The problem is formulated as follows: Given N unsupervised band selection methods and M supervised classification methods, how would one obtain the optimal number of bands and the best performing pair of methods that maximize classification accuracy and minimize computational requirements? This formulation is a variant of the problem definition in (Swain and Davis, 1978, Chapter 3-8) and has been researched for a single band selection method in the previous work (Jia and Richards, 1994; Merenyi *et al*., 1996; Fung *et al*., 1999; Warner *et al.,* 1999; Witthagen, 2001; Shettigara *et al*., 2002). While previous work has evaluated individual methods, for example, stepwise discriminant analysis in (Fung *et al.*, 1999), maximum likelihood classification in (Jia and Richards, 1994) spatial autocorrelation analysis in (Warner *et al.,* 1999), or principal component analysis (PCA) jointly with artificial neural network (ANN) analysis in (Pu and Gong, 2000), our formulation is more general by optimizing not only over all methods but also over all combinations of supervised and unsupervised methods. Our reasoning for doing so is based on the no-free-lunch (NFL) theorem (Duda *et al.*, 2001, pp. 454), which states that no single supervised method is superior over all problem domains; methods can only be superior for particular data sets. As we assume no prior knowledge of the underlying structure of the data, and there is no universally accepted 'best' supervised method by the NFL theorem, we experiment over a range of methods and implementations to find which is superior for hyperspectral data. To limit

the computational complexity of trying all $\sum_{i=1}^{nb}\binom{nb}{i}$ combinations of bands with each supervised method, where *nb* is the number of bands, we deploy a set of unsupervised methods to trim the search space to include only the bands that the various unsupervised methods respectively deem most informative and least redundant. We exclude from the problem, however, any formation of new features such as creating eigenvectors (principle component analysis), averaging adjacent bands, or searching for basis functions in a subspace of data intrinsic dimensionality (Bruske and Merenyi, 1999) because otherwise the type of feature formation would be another degree of freedom for the proposed analysis and the search space would become computationally prohibitive. The outcome is expected to answer basic questions about which wavelength ranges should be used given a hyperspectral image and a specific application. The research objective of this paper is to investigate a methodology for combining unsupervised and supervised methods under classification accuracy and computational requirement constraints that can provide the answers to the band selection questions described above.

Next in Section 2 is the methodology for evaluating band selection methods. In Section 3 we present an overview of both unsupervised and supervised band selection methods. Experimental results and discussion of the results are presented in Section 4. In Section 5 we summarize our work and address future directions.

2 EVALUATION METHODOLOGY

The tradeoff between accuracy and computational requirements is related to the choice of bands and classification methods. Thus, there is a need for a methodology for choosing hyperspectral bands that provide sufficient, but not redundant, information to classification or prediction algorithms using a practical amount of computational

resources. We can use unsupervised methods to compute rank ordered lists of bands in a computationally efficient way thereby pre-filtering bands based on their redundancy and information content. Because direct comparisons of scores obtained by unsupervised methods are not valid due to different score scales, we use the predictive accuracy of supervised methods that use the top ranked bands from the unsupervised methods to evaluate both the quality of the top ranked bands and, indirectly, the quality of the unsupervised methods. Furthermore, supervised methods can be applied to a variable number of top ranked bands obtained from unsupervised methods. The trend of model errors as a function of the processed number of ranked bands will demonstrate local (or global) minima that will identify the optimal number of bands S maximizing model accuracy. Lastly, the problem of selecting a supervised method is addressed by choosing the method that forms the most accurate model with respect to the training data.

To evaluate a supervised method given a set of bands and control parameters, cross-validation is used (Duda *et al.*, 2001). The process of n-fold cross-validation involves splitting the data set into n non-overlapping, exhaustive subsets, and then building n models, one for each subset being withheld from training. Averaging the error calculated for each withheld set then scores each method. The error is defined as the difference between the actual values and predicted values of the set's respective model. This assigned error is a function not only of the method and data set, but also on any control parameters of the method. To account for this, we performed an optimization of the control parameters for each combination of top ranked band sets and supervised methods. In our study, the error assigned to a set of bands was based on the mean absolute error of the test cases using 12-fold cross-validation. Appropriate models were

tried with 300 different randomly selected control parameter sets at each band set evaluation. The control parameter set with the best performance using 8-fold cross-validation was then used to compute the final 12-fold cross-validation error (the smaller value of *n=8* required less computation for this step). Once the model error is computed for band count varying from 1 to N, where N is the maximum number of bands, we evaluate a discrimination measure (*DM)* as defined below in order to establish the quality of the optimal number of bands *S* that maximizes the model accuracy.

$$DM = \left\| Error(S) - Error(S+\Delta) \right\| + \left\| Error(S) - Error(S-\Delta) \right\| \quad (1)$$

In a nutshell, the process can be described as running unsupervised methods to rank the best bands followed by testing those band choices with the supervised methods to see which combinations are best for a particular application. The same process should also reveal the optimal number of bands for the application in question.

3 OVERVIEW OF BAND SELECTION METHODS

Our methodology involves two types of band selection methods, unsupervised and supervised. Unsupervised methods order hyperspectral bands without any training and the methods are based on generic information evaluation approaches. Unsupervised methods are usually very fast and computationally efficient. These methods require very little or no hyperspectral image pre-processing. For instance, there is no need for image geo-referencing or registration using geographic referencing information, which might be labor-intensive operations.

In contrast to unsupervised methods, supervised methods require training data in order to build an internal predictive model. A training data set is obtained via registration of calibrated hyperspectral imagery with ground measurements. Supervised methods are usually more computationally intensive than unsupervised methods due to an arbitrarily

high model complexity and an iterative nature of model formation. Another requirement of supervised methods is that the number of examples in a training set should be sufficiently larger than the number of attributes (bands, in this case). This requirement might be hard to meet as the number of hyperspectral bands grows and the collection of each ground measurement has an associated real-world cost. If taken alone, the unsupervised methods can, at best, be used to create classes by clustering of spectral values followed by assigning an average ground measurement for each cluster as the cluster label. Supervised methods therefore provide more accurate results than unsupervised methods.

We developed seven unsupervised methods described in Section 3.1 including entropy, contrast, 1st and 2nd spectral derivative, ratio, correlation and principal component analysis ranking based algorithms. We chose three supervised methods described in Section 3.2 including regression, instance based (k-nearest neighbor) and regression tree algorithms because they represent methods for prediction of continuous input/output variables with global, local, and hybrid modeling approaches, as discussed in the following sections. A brief outline of all band selection methods used in this work follows.

3.1 Unsupervised Band Selection Methods

Information Entropy: This method is based on evaluating each band separately using the information entropy measure (Russ, 1999, Chapter 3) defined below.

$$H(\lambda) = -\sum_{i=1}^{m} p_i \ln p_i \qquad (2)$$

H is the entropy measure, p is the probability density function of reflectance values in a hyperspectral band and m is the number of distinct reflectance values. The probabilities

are estimated by computing a histogram of reflectance values. Generally, if the entropy value H is high then the amount of information in the data is large. Thus, the bands are ranked in the ascending order from the band with the highest entropy value (large amount of information) to the band with the smallest entropy value (small amount of information).

First Spectral Derivative: The bandwidth, or wavelength range, of each band is a variable in a hyperspectral sensor design (Price, 1994; Wiersma and Landgrebe, 1980). This method explores the bandwidth variable as a function of added information. It is apparent that if two adjacent bands do not differ greatly then the underlying geo-spatial property can be characterized with only one band. The mathematical description is shown below, where I represents the hyperspectral value, x is a spatial location and λ is the central wavelength. Thus, if D_1 is equal to zero then one of the bands is redundant. In general, the adjacent bands that differ significantly should be retained, while similar adjacent bands can be reduced.

$$D_1(\lambda_i) = \sum_x \left\| I(x, \lambda_i) - I(x, \lambda_{i+1}) \right\| \tag{3}$$

Second Spectral Derivative: Similar to the first spectral derivative, this method explores the bandwidth variable in hyperspectral imagery as a function of added information. If three bands are adjacent, and the two outside bands can be used to predict the middle band through linear interpolation, then the band is redundant. The larger the deviation from a linear model, the higher the information value of the band. The mathematical description of this method is shown below, where D_2 represents the measure of linear deviation, I is a hyperspectral value, x is a spatial location and λ is the central wavelength.

$$D_2(\lambda_i) = \sum_x \left\| I(x,\lambda_{i-1}) - 2I(x,\lambda_i) + I(x,\lambda_{i+1}) \right\| \quad (4)$$

Contrast Measure: This method is based on the assumption that each band could be used for classification purposes by itself. The usefulness of a band would be measured by a classification error achieved by using only one particular band and minimizing the error. In order to minimize a classification error, it is desirable to select bands that provide the highest amplitude discrimination (image contrast) among classes. If the class boundaries were known *a priori* then the measure would be computed as a sum of all contrast values along the boundaries. However, the class boundaries are unknown *a priori* in the unsupervised case. One can evaluate contrast at all spatial locations instead assuming that each class is defined as a homogeneous region (no texture variation within a class). The mathematical description of the contrast measure computation is shown below for a discrete case.

$$ContrastM(\lambda) = \sum_{i=1}^{m} \left\| f_i - E(f) \right\| * f_i \quad (5)$$

f is the histogram (estimated probability density function) of all contrast values computed across one band by using Sobel edge detector (Russ, 1999, Chapter 4), *E(f)* is the sample mean of the histogram *f* and λ is the central wavelength. *m* is the number of distinct contrast values in a discrete case. The equation includes the contrast magnitude term and the term with the likelihood of contrast occurrence. In general, bands characterized by a large value of *ContrastM* are ranked higher (good class discrimination) than the bands with a small value of *ContrastM*.

Spectral Ratio Measure: In many practical cases, band ratios are effective in revealing information about inverse relationship between spectral responses to the same phenomenon (e.g., living vegetation using the normalized difference vegetation index

(Campbell, 1996, Chapters 16.6 and 17.7). This method explores the band ratio quotients for ranking bands and identifies bands that differ just by a scaling factor. The larger the deviation from the average of ratios *E(ratio)* over the entire image, the higher the *RatioM* value of the band. The mathematical description of this method is shown below, where *RatioM* represents the measure, *I* is a hyperspectral value, *x* is a spatial location and λ is the central wavelength.

$$RatioM(\lambda_i) = \sum_x \left\| \frac{I(x,\lambda_i)}{I(x,\lambda_{i+1})} - E\left(\frac{I(x,\lambda_i)}{I(x,\lambda_{i+1})}\right) \right\| \tag{6}$$

Correlation Measure: One of the standard measures of band similarity is normalized correlation (Duda *et al.*, 2001). The normalized correlation metric is a statistical measure that performs well if a signal-to-noise ratio is large enough. This measure is also less sensitive to local mismatches since it is based on a global statistical match. The correlation based band ordering computes the normalized correlation measure for all adjacent pairs of bands similar to the spatial autocorrelation method (Warner *et al.,* 1999) applied to all ratios of pairs of image bands. The mathematical description of the normalized correlation measure is shown below, where *CorM* represents the measure, *I* is a hyperspectral value, *x* is a spatial location and λ is the central wavelength. *E* denotes an expected value and σ is a standard deviation.

$$CorM(\lambda_i) = \frac{E(I(\lambda_i) * I(\lambda_{i+1})) - E(I(\lambda_i)) * E(I(\lambda_{i+1}))}{\sigma(I(\lambda_i)) * \sigma(I(\lambda_{i+1}))} \tag{7}$$

After selecting the first least correlated band based on all adjacent bands, the subsequent bands are chosen as the least correlated bands with the previously selected bands. This type of ranking is based on mathematical analysis of Jia and Richards, 1994, where spectrally adjacent blocks of correlated bands are represented in a selected subset.

Principal Component Analysis Ranking (PCAr): Principal component analysis has been used very frequently for band selection in the past (Campbell, 1996, pp. 289). The method transforms a multidimensional space to one of an equivalent number of dimensions where the first dimension contains the most variability in the data, the second the second most, and so on. The process of creating this space gives two sets of outputs. The first is a set of values that indicate the amount of variability each of the new dimensions in the new space represents, which are also known as eigenvalues (ε). The second is a set of vectors of coefficients, one vector for each new dimension, that define the mapping function from the original coordinates to the coordinate value of a particular new dimension. The mapping function is the sum of the original coordinate values of a data point weighted by these coefficients. As a result, the eigenvalue indicates the amount of information in a new dimension and the coefficients indicate the influence of the original dimensions on the new dimension. Our PCA based ranking system (PCAr) makes use of these two facts by scoring the bands (the "original" dimensions in the above discussion) as follows.

$$PCAr(\lambda_i) = \sum_j \left| \varepsilon_j c_{ij} \right| \tag{8}$$

λ_i is the central wavelength, ε_j is the eigenvalue for the j^{th} principal component, and c_{ij} is the mapping coefficient of the i^{th} central wavelength in the j^{th} principal component. As the procedure for computing the eigenvalues and coefficients is both complex and available in most data analysis texts (Duda, *et al*, 2001), it is omitted.

3.2 Supervised Prediction Methods

Using the proposed approach requires choosing classification methods according to the type of input (here, hyperspectral) and output (predicted) variables. In general, any

of these variables can be either continuous or discrete (also commonly referred to as numeric and categorical, or scalar and nominal). In this application all supervised methods predict a continuous variable (soil electrical conductivity) using all continuous inputs (values representing hyperspectral measurements).

Regression: The regression method is based on a multivariate regression (Gill *et al*., 1991; Han and Kamber, 2001) that is used for predicting a single continuous variable *Y* given multiple continuous input variables $\{X_1 ... X_n\}$. The model building process can be described as follows. Given a set of training examples *T*, find the set of coefficients $\beta = \{\beta_0 ... \beta_n\}$ that gives the minimum value of *g(T),* where

$$g(T) = \min \sum_{e \in T} (Y_e - Y'_e)^2 \tag{9}$$

Y_e is the observed output variable of a training example *e* and

$$Y'_e = \beta_0 + \beta_1 X_1^e + \beta_2 X_2^e + \ldots + \beta_n X_n^e \tag{10}$$

Y'_e is therefore the predicted value for Y_e given values for $\{X_1^e ... X_n^e\}$ which, in this case, are reflectance values at varying wavelengths for the training example *e*. The problem as stated can be solved numerically using well-known matrix algebra techniques. Further details for finding $\beta = \{\beta_0 ... \beta_n\}$ are therefore omitted for the sake of brevity.

Instance Based Method: The instance based method uses inverse Euclidean distance weighting of the k-nearest neighbors to predict any number of continuous variables (Witten and Frank, 2000; Han and Kamber, 2001). To predict a value Y' of the example being evaluated *e*, the *k* points in the training data set with the minimum distance (see Eq. (11)) to the point *e* over the spectral dimensions $\{X_1 ... X_n\}$ are found.

$$d = \sqrt{\sum_{i=1}^{n} (X_i' - X_i)^2} \tag{11}$$

The weighted average of the observed *Y* values of these *k* closest training points is then computed where the weighting factor is based on the inverse of the distance from each of the *k* points to the point *e* according to Eq. (12). Furthermore, the weighting factor is raised to the power *w*. Altering the value of *w* therefore influences the relationship between the impact of a training point on the final prediction and that training point's distance to the point being evaluated. The user must set the values of the control parameters *k* and *w*. In our study, these parameters were selected using the optimization procedure described in Section 2..

$$Y' = \frac{\sum_{i=0}^{k} \left(\frac{1}{d_i^w} \right) \cdot Y_i}{\sum_{i=0}^{k} \left(\frac{1}{d_i^w} \right)} \tag{12}$$

Regression Tree: A regression tree is a decision tree that is modified to make continuous valued predictions (Breiman *et al.,* 1984). They are akin to binary search trees where the attribute used in the path-determining comparison changes from node to node. The leaves then contain a distinct regression model used to make the final prediction.

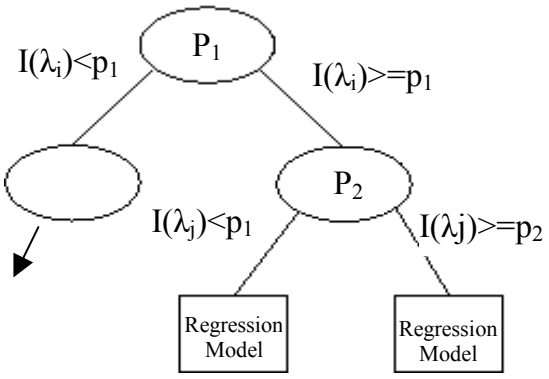

Figure 1: A simple regression tree where spectral values determine the path and the leaves contain regression models.

To evaluate (or test) an example using a regression tree, the tree is traversed, starting at the root, by first comparing the reflectance value at a single wavelength requested by the node and compared to the split-point (in Fig. 2, the *P* values). Particular wavelengths may be used by several nodes or none at all. If the reflectance value of the example at the appropriate wavelength is less than the split point, the left branch is taken, if greater than or equal to the split-point, the right. This splitting procedure based on reflectance values continues until a tree leaf is encountered, at which time the prediction can be made based on data in the leaf.

To build a model, one must select what bands and what reflectance values for those bands are necessary to split the examples into sets that have similar target variables. To do this, a greedy approach is employed based on minimization of the target variable's variance (defined in Eq. (13)). More precisely, at every node, find central wavelength λ and corresponding split point *p* such that the average variance of the targets of the two portions of the data set *s* after being split is minimized. This average variance is weighted based on how many training examples take the left or right branch, respectively (see Eq. (13)).

$$BestSplit(s) = \min_{\lambda,p} \left(\frac{\frac{\text{var}(t,n_l)}{|n_l|} + \frac{\text{var}(t,n_r)}{|n_r|}}{|n_l| + |n_r|} \right) \tag{13}$$

where $n_l = \{examples\ e : e_\lambda \le p\}$, $n_r = \{examples\ e : e_\lambda > p\}$ and the variance of the variable *Y*, in the set of examples *s* is given by

$$\text{var}(Y,s) = \sum_{i=0}^{|s|} (Y_i - \overline{Y})^2 \tag{14}$$

To find the values of λ and *p*, the algorithm only tries the mean of the reflectance values at each wavelength and selects the (λ, *p*) combination according to Eq. (13). Future experiments may deploy a more comprehensive search of optimal values of (λ , *p*).

The algorithm halts when one of two criteria is met. The first is that the number of examples that evaluate to a node falls below *m,* the minimum allowed examples per node. The other is that the improvement (reduction) in variance that would be obtained by doing the best possible split is below some improvement threshold, *t*. In either case, the node at which the halting criteria are met is marked as a leaf and a regression model is built on the training examples that evaluate to that node. Both *t* and *m* are control parameters which are optimized via the procedure from Section 2..

3.3 Expected Trends

Assuming that each unsupervised method sorts the bands based on band redundancy in ascending order, our expectation is to see the following trends in the resulting function (see Figure 2). First, the regression-based supervised method is using a global modeling approach where very few bands (insufficient information) or too many bands (redundant information) will have a negative impact on the model accuracy. Thus, we expected the trend of a parabola with one global minimum. Second, the instance-based method exploits local information and adding more bands will either decrease an error or preserve it constant. The expected trend is a down-sloped staircase curve with several plateau intervals. The beginning of each plateau interval can be considered as a local minimum for selecting the optimal number of bands (see crosses in Figure 2). Lastly, the regression tree based method uses a hybrid approach from a standpoint of local versus global information. It is expected to demonstrate a trend of the instance-

based method for a small number of processed bands (band count) and a trend of the regression-based method for a large number of processed bands.

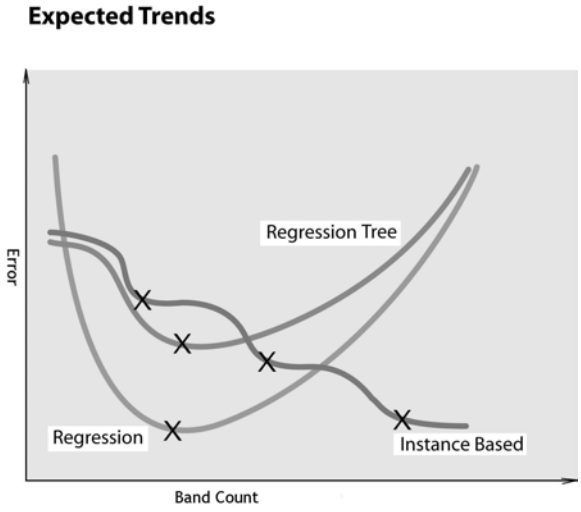

Figure 2: Expected trends based on the models of three supervised methods.

4 EXPERIMENTAL RESULTS

The proposed methods for band selection were applied to hyperspectral data collected for precision farming applications. Detailed information about the hyperspectral data, experimental results from unsupervised and supervised band selection methods, and interpretation of the obtained results are provided next.

4.1 Hyperspectral Data

The hyperspectral image data used in this work were collected from an aerial platform with a Regional Data Assembly Centers Sensor (RDACS), model hyperspectral (H-3), which is a 120-channel prism-grading, push-broom sensor developed by NASA. Each image has 2500 rows, 640 columns, and 120 bands per pixel. The 120 bands correspond to the visible and infrared range of 471 to 828 nm, recorded at a spectral resolution of 3 nm. The motivation for choosing the wavelength range came from the agricultural application domain where the 400-900 nm wavelength range responds to

plant characteristics very well (Swain and Davis, 1978, Chapters 2-2 and 5-2) and has been used for vegetation sensing in the past (Gopalapillai and Tian, 1999). By selecting this wavelength range, the data analysis avoids issues related to water absorption bands (1400 nm and 1900 nm). In the particular range we compensated only for low reflectance in the blue (450 nm) and red (650 nm) wavelength sub-ranges due to the two chlorophyll absorption bands (Campbell, 1996, Chapter 17.4) during reflectance calibration. While our experiments dealt with images of bare soil, we used a sensor that is optimal for vegetation observation as that is what is likely to be available in agricultural applications (for the reasons given above). Indeed, the experimental data set used in this study came from a series of images taken over the entire growing season that were collected to study the relationship between hyperspectral information and both bare soil properties before crops had emerged and crop properties when they were present. For application specific interpretations of data, each band index of the hyperspectral image was converted to the band central wavelength by applying the following formula: $\lambda_b = 471 + 3*(b-1) \quad [nm]$.

The images were collected from altitudes in the range of 1200 m to 4000 m on April 26, 2000. The spatial resolution of the images is approximately 1-m for the processed Gvillo field located near the city of Columbia in the central part of Missouri. The images were pre-processed to correct for geometrical distortions, calibrated for sensor noise and illumination, and geo-registered (Swain and Davis, 1978, Chapter 2-7). However, the images were not pre-processed for any atmospheric corrections (Campbell, 1994, (Chapter 10-4)]. An image of the Gvillo site is shown in Figure 3.

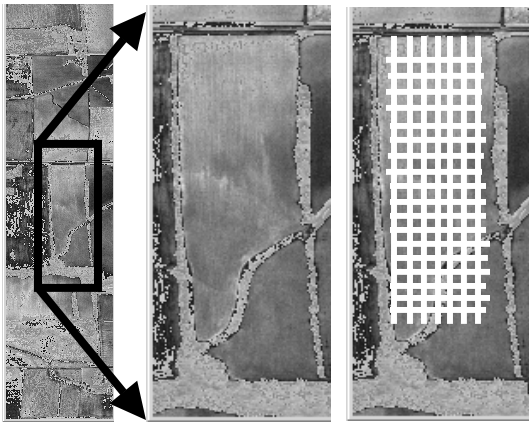

Figure 3: A hyperspectral image (left) obtained in April 26, 2000 at 4000 m altitude and the Gvillo site of interest (middle) with associated grid-based locations of ground measurements (right). The display shows combined bands with central wavelengths 471 nm, 660 nm and 828 nm.

Ground measurements of several variables (e.g., conductivity, elevation, organic matter, phosphorous content) were collected by the Illinois Laboratory Agricultural Remote Sensing (ILARS) using the Veris profiler 3000 made by Veris Technologies, Salina, KS, and the data were provided by Dr. Tian. The hyperspectral images provided by Spectral Visions, a non-profit research organization funded by the NASA Commercial Remote Sensing Program, were geo-registered with the ground measurements by Dr. Gopalapillai (Department of Biological and Agricultural Engineering, University of Arkansas) and both ground and aerial measurements formed a training data set covering about 19,000 m^2 of the Gvillo field. We used the training data with 190 examples from the hyperspectral imagery collected at 4000 m altitude for evaluating the band selection methods. The training data contained these hyperspectral values and associated ground values of soil electrical conductivity. The field coverage on the date of data collection was bare soil.

Among all ground variables, we anticipated to find relationships between hyperspectral values (reflected part of the electro-magnetic (EM) waves in the wavelength range [471nm, 828 nm]) and surface/field characteristics that change electric and magnetic properties according to the EM theory of wave propagation (Balanis, 1989, Chapter 5). Thus, electrical conductivity appeared as the number one candidate among other variables. We verified with a simple linear correlation method that there exists a significant enough correlation (around 0.5) between the conductivity variable and hyperspectral values (190 conductivity values were correlated with 190 hyperspectral values for each band to obtain 120 correlation values averaging near 0.5). The conductivity values ranged from [22.4262, 52.66] miliSiemens per meter with the sample mean equal to 36.10836 and the standard deviation equal to 5.212215. Based on the known classification of soil properties (Veris Technologies, 2003) as a function of conductivity with approximate class conductivity ranges of sand (0,2], silt [2, 20] and clay [10, 1000], we concluded that the ground soil consisted of silt and clay soil types. Soil electrical conductivity is an important characteristic considered for crop yield prediction in the agricultural application. Electrical conductivity indirectly characterizes several important soil characteristics including soil texture (the relative amount of sand-silt-clay) and salinity, which affects the crops ability to acquire water.

4.2 Results from Unsupervised Band Selection Methods

The results of seven unsupervised band selection methods are shown in Table 1. The processed hyperspectral data came from the training set without using any ground measurements (only 120 hyperspectral band values). The unsupervised methods were

implemented in Java and documented for interested users of hyperspectral analysis tools (Bajcsy, 2002).

Table 1: Top 15 bands selected by seven unsupervised band selection methods and reported by their central wavelength in nm.

Order	Entropy	Contrast	1st Der.	2nd Der.	Ratio	Correl.	PCAr
1	741	741	741	741	741	741	588
2	795	594	738	744	498	486	591
3	822	597	669	738	744	828	582
4	669	600	747	669	492	588	585
5	615	603	666	747	501	471	594
6	825	606	639	672	747	825	579
7	819	609	498	642	669	603	636
8	636	612	792	639	639	822	648
9	627	615	744	699	486	579	600
10	654	591	696	501	489	819	642
11	612	639	699	498	738	474	597
12	666	666	750	795	522	600	603
13	645	669	801	471	483	816	576
14	828	570	642	801	513	501	645
15	651	585	636	666	636	813	630

In this experimental evaluation, the contrast based unsupervised method utilized the fact that the hyperspectral examples extracted from a hyperspectral image were spatially ordered along a geo-spatial line (row). Based on the contrast measure definition in Section 3.1, computing an amplitude contrast using Sobel edge detector requires spatially adjacent amplitude locations. This requirement was satisfied by selecting hyperspectral values at the locations of grid-based ground measurements (see Figure 3

right). Spatial adjacency is not an issue in the case a hyperspectral image since it is well defined by an underlying image grid. We have not encountered any other problem during this part of the experiment.

The unsupervised methods ran on a Dell PC, Dimension 4100 with a single Intel Pentium III processor and Windows 2000 operating system. The order of unsupervised methods based on their algorithmic computationally efficiency was (1) 1st spectral derivative, (2) 2nd spectral derivative, (3) ratio, (4) contrast, (5) entropy, (6) correlation, and (7) PCAr based methods. The maximum time for processing 190 examples with 120 bands did not exceed 2 seconds.

4.3 Results from Supervised Band Selection Methods

We processed seven rank-ordered lists of bands obtained using unsupervised methods by three supervised methods. Figures 5, 6 and 7 were formed by computing an error value for two, four, six, …120 top bands from each rank-ordered list using Regression (Fig. 5), Instance based (Fig. 6) or Regression tree (Fig. 7) supervised algorithms with the ground measurement of soil electrical conductivity.

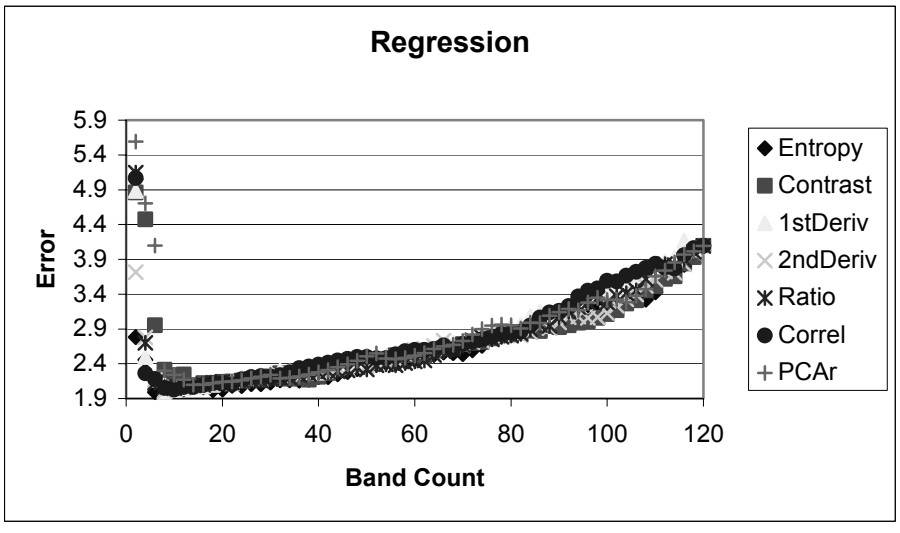

Figure 4: Results obtained from the regression based supervised method using rank ordered bands from unsupervised methods.

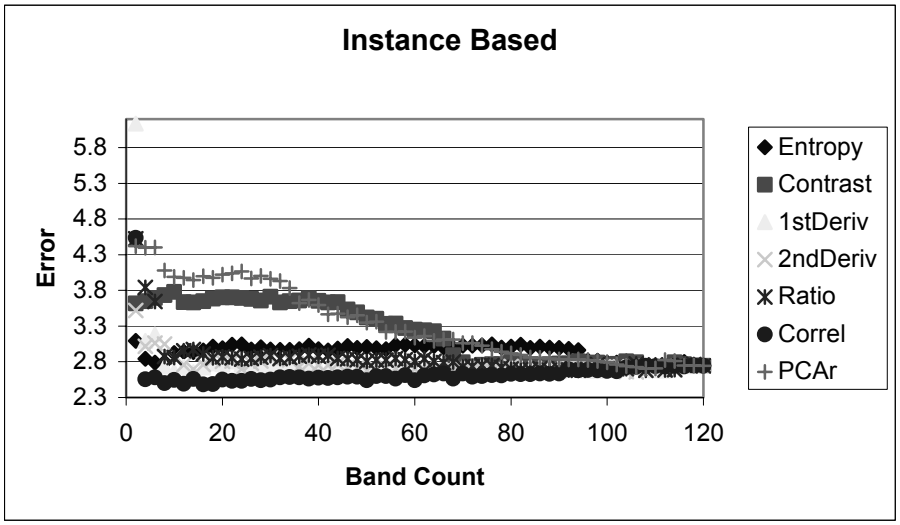

Figure 5: Results obtained from the instance based supervised method using rank ordered bands from unsupervised methods.

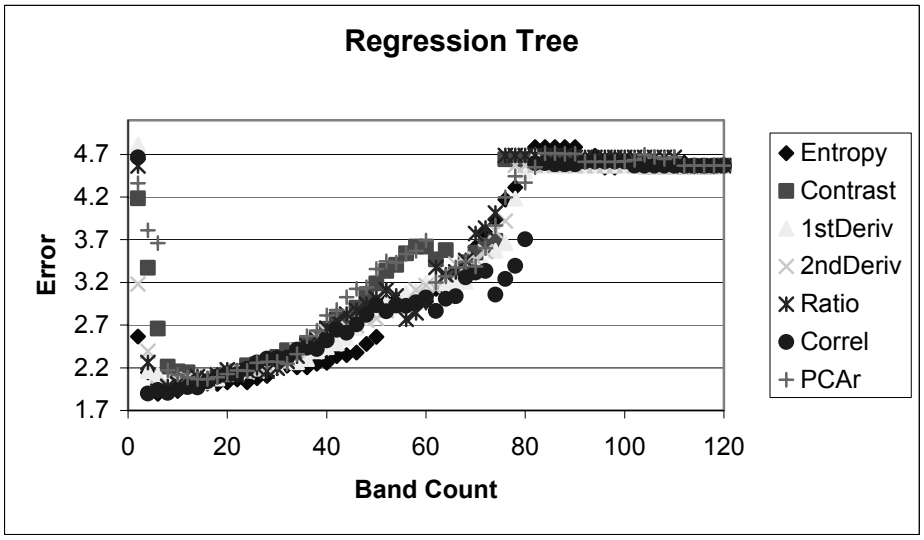

Figure 6: Results obtained from the regression tree based supervised method using rank ordered bands from unsupervised methods.

We also processed lists of bands obtained by random and incremental ranking along the spectral axis. The incremental ranking sub-divides spectral bands into the ordered set {b64, b1, b120, b32, b96, b16, b112, b48, …}. The result for the incrementally created set and the average result for seven randomly generated band sets are shown in Figure 7. These results were generated as a baseline for quantitative comparison with the results from Figures 5, 6, and 7. Based on the comparative summary provided in Table 2, we concluded that the results obtained from the best unsupervised ranking always lead to smaller prediction error, smaller or equal number of optimal bands *S* and higher discrimination DM (see Eq. (1)) of *S* for regression and instance based methods.

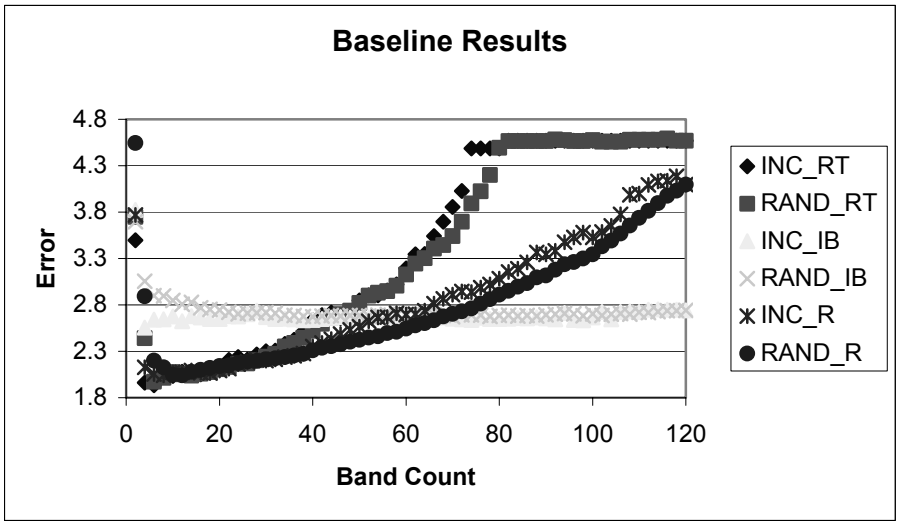

Figure 7: Baseline results obtained for randomly (RAND) and incrementally (INC) selected bands with regression tree (RT), instance based (IB) and regression (R) based supervised methods.

Table 2: Comparison of the results from baseline random (RAND) and incremental (INC) rankings with the best results from unsupervised (BEST UNSUP) rankings.

	RAND			INC			BEST UNSUP		
	Error	S	DM	Error	S	DM	Error	S	DM
Regression	2.0405	12	0.4489	2.0422	8	0.0418	1.9940	6	0.4912
Instance Based	2.7044	24	0.01523	2.5548	4	1.3544	2.5543	4	2.0101
Regression Tree	1.9713	6	0.5107	1.9332	6	0.1186	1.8953	6	0.2727

In all these experimental evaluations, we had to overcome several limitations of the supervised algorithms. For example, the multivariate regression must have at least as many examples as there are bands (attributes) in order for the linear algebra routines it makes use of to be valid. Furthermore, as the number of bands approaches the number of examples, the algorithm begins to perform poorly even if it does not fail. Our regression tree algorithm was modified so that it would return a mean model if the regression were to fail. The so-called mean model simply predicts the average output value of the training examples for any testing case, ignoring all input (spectral) information. The regression tree model also gave similar results of rapid accuracy decline when the leaves of a regression tree contained very few examples relative to the number of bands being evaluated. As a consequence of Eq. (10), the number of examples in each leaf has to be greater than or equal to the number of unknowns, which is equal to the number of bands. This accuracy decline can be observed in Figure 6. Because the mean model is more accurate than regression models built with inadequate data, we see the trend around 80 bands where the regression models fail and are replaced by mean models that give static accuracy. There were no limitations in the case of instance-based algorithm.

The supervised methods, also implemented in Java in a data flow programming environment called D2K (Welge *et al*., 2000), ran on a Sun Ultra- Enterprise machine with 16 processors and Solaris 5.7 operating system. Processing the results of all seven unsupervised methods with an increment of two bands {2, 4, 6,…, 120} took

approximately 22 hours. If we disregard the fact that the evaluations of large numbers of bands took longer than the evaluations with smaller numbers of bands then the average time of each evaluation was around 3.14 minutes. All supervised methods ran in parallel. The most computationally efficient method is regression-based method, followed by the regression tree and finally the instance based method. It is also important to mention that the majority of the time was spent finding the optimal control parameters for the regression tree and instance based algorithms.

4.4 Interpretation of Results

In this experiment, the goal was to select a combination of unsupervised and supervised methods, the optimal number of bands, and band indices subject to model accuracy and computational requirement considerations. Following the methodology in Section 2, the results of supervised methods and the trends in Figures 5, 6 and 7 were investigated.

We concluded that the trends for all seven unsupervised methods followed the predicted trends in Figure 2 for supervised regression (Figure 4) and regression tree (Figure 6) based evaluations quite well. Some trend deviation is observed in the regression tree evaluation for band count variable larger that 80 due to the sample size limitation in tree leaves as it was explained in the previous section. This deviation could be removed by increasing the number of training examples. The trends of the instance based algorithm observed in Figure 5 are present for contrast, ratio and PCA ranking based-methods but are less pronounced for other unsupervised methods. For other than these three methods, the error values reach a value near-global minimum after the band count variable becomes 4-6 and then decrease by only a very small amount (an

approximate error gradient for curves at band count larger than 6 is less than 0.002). Thus, this particular trend is a case shown in Figure 2 with only one plateau and can be explained by optimal band ranking of the other four methods.

By analyzing the results in Figures 5, 6 and 7, the minimum error per figure was achieved by (a) the entropy based unsupervised method evaluated with the regression-based supervised method (error = 1.9940 in Fig. 5), (b) the correlation based method with instance-based supervised method (error = 2.5543 in Fig. 6) and (c) the entropy based unsupervised method with regression tree based supervised method (error = 1.8953 in Fig. 7). The number of bands *S* (the smallest *S*) at the local minima of error was reported in Table 2 for each pair of unsupervised and supervised methods. The optimal numbers of bands *S* that were reported the most often were 4 and 6. The highest discrimination measure *DM* defined in Eq.(1), with $\Delta = 2$, for the optimal number of bands *S* was achieved by instance based supervised method and 1st derivative based unsupervised method for *S* equal 4. We used the discrimination measure *DM* to quantify our confidence in finding the true error minimum and the corresponding *S.* While higher DM means higher confidence, the absolute values of *DM* can be compared only for the same supervised method since the range of *DM* values depends on the error range and can theoretically reach twice the difference between maximum and minimum error values.

Table 3: The lowest optimal number of bands *S* and its discrimination score *DM* determined for each combination of unsupervised and supervised methods based on the local minima of error as a function of processed band count.

	Entropy		Contrast		1st Der.		2nd Der.		Ratio		Correl.		PCAr	
	S	DM	S	DM	S	DM	S	DM	S	DM	S	DM	S	DM

Regression	6	0.4912	8	0.5643	8	0.2060	8	0.2022	10	0.0427	10	0.0531	14	0.0184
Instance based	6	0.0579	14	0.0225	4	3.1283	4	0.5445	10	0.1121	4	2.0101	4	0.0255
Regression tree	6	0.2727	14	0.1425	8	0.1838	6	0.4696	6	0.3962	4	2.8082	16	0.0232

In summary, our recommendation is to select approximately the top 4 to 8 bands with the entropy based unsupervised method followed by a classification model using the regression tree based supervised method. The recommendation is based on computing a weighted average of optimal bands per each supervised method according to the equation $S_{\text{supervised\_method}} = \frac{1}{\sum_{i=1}^{\text{N}} DM(i)} \sum_{i=1}^{\text{N}} DM(i) * S(i)$, where N is the number of unsupervised methods, and leading to $S_{\text{Regression}} = 7.568$, $S_{\text{Instance Based}} = 4.172$ and $S_{\text{Regression Tree}} = 5.098$. The most frequently selected bands (by more than two methods) are 498, 501, 600, 603, 636, 639, 642, 666, 669, 738, 741, 744 and 747 nm, as it is shown in Figure 8.

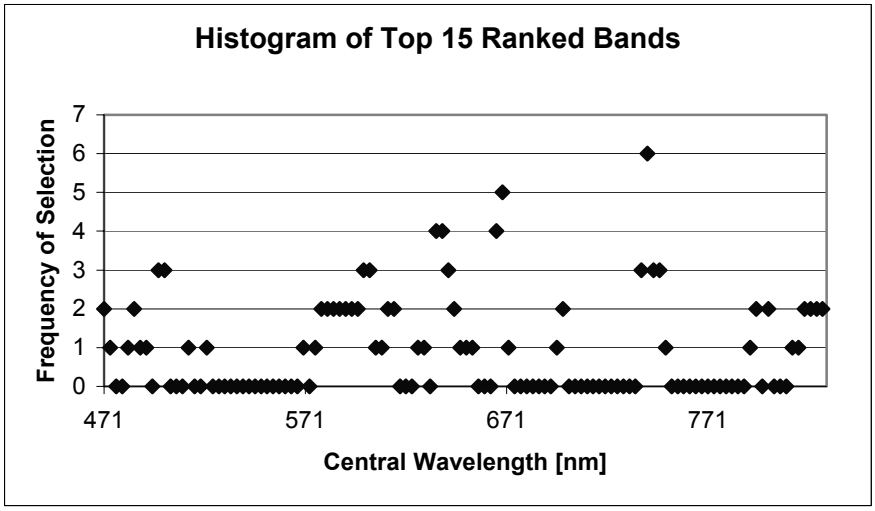

Figure 8: Histogram of top 15 ranked bands by all unsupervised methods according to Table 1.

5 SUMMARY

In this paper we have presented a new methodology for combining unsupervised and supervised methods under classification accuracy and computational requirement constraints that was used for selecting hyperspectral bands and classification methods. The novelty of the work is in combining strengths of unsupervised and supervised band selection methods to build a computationally efficient and accurate band selection system. We have developed and combined seven unsupervised and three supervised methods to test the proposed methodology. The methodology was applied to the prediction problem between airborne hyperspectral measurements and ground soil electrical conductivity measurements. While analyzing soil electrical conductivity is important for soil characterization and crop yield prediction, the airborne hyperspectral data collection represents more economical and efficient way of soil information gathering than ground data measurements. We conducted a study based on the experimental data that demonstrated the process of obtaining the optimal number of bands, band central wavelengths and the selection of classification methods under classification accuracy and computational requirement constraints. The study concluded that there are about 4-8 most informative bands for the electrical conductivity variable including 7 bands in the red spectrum (600, 603, 636, 639, 642, 666 and 669 nm), 4 bands in the near infrared spectrum (738, 741, 744 and 747 nm) and 2 bands close to the border of blue and green spectra (498 and 501 nm). We believe that this result is in accordance with our empirical observations (soils with ferrites would appear reddish), as well as electromagnetic theory (phenomenological and atomic models according to Balanis, 1989, Section 2.8.3) that derive dependency of electric conductivity as a function

of wavelength from Maxwell equations. The proposed band selection methodology is also applicable to other application domains requiring hyperspectral data analysis.

In the future, we would like to improve the unsupervised ranking procedure with ratio measures by iteratively selecting and removing the best bands and then selecting the subsequent best bands from the remaining set (Jia and Richards, 1994). We would also like to include in our analysis other linear, e.g., correlation, and non-linear, e.g., artificial neural network (Kavzoglu and Mather, 2000), supervised methods, and evaluate their performance within our band selection framework. Another direction to pursue is finding the true optimal band set through an exhaustive search over a data set verified by a domain expert. We also plan to investigate the problem of combining all constraints into a more rigorously formulated mathematical framework. The tradeoffs between classification accuracy versus computational requirements are currently loosely integrated and a rigorous quantitative analysis might be useful.

REFERENCES

Bajcsy, P., 2002. Image To Knowledge (I2K)," Software Overview and Documentation, Automated Learning Group at NCSA, University of Illinois at Urbana-Champaign, IL, URL: http://alg.ncsa.uiuc.edu/tools/docs/i2k/manual/index.html (last date accessed: 17 April, 2003).

Balanis C. A., 1989. *Advanced Engineering Electromagnetics,* John Wiley and Sons, USA.

Benediktsson, J. A., J. R. Sveinsson and K. Arnason, 1995, Classification and feature extraction of AVIRIS data. IEEE Transactions on Geoscience and Remote Sensing 33, 1194-1205.

Breiman L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Monterey, CA, Wadsworth International Group.

Bruske J. and E. Merényi, 1999. Estimating the Intrinsic Dimensionality of Hyperspectral Images, *Proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium, April 21-23, pp. 105-110.

Campbell, B.J., 1996. *Introduction to Remote Sensing*, second edition, The Guilford Press, New York, NY.

Csillag, F, L. Pasztor, and L. Biehl, 1993, Spectral band selection for the characterization of salinity status of soils. Remote Sensing of Environment 43, 231-242.

Duda, R., P. Hart and D. Stork, 2001. *Pattern Classification,* Second Edition, Wiley-Interscience.

Fung, T., F. Ma and W.L. Siu, 1999. Band Selection using Hyperspectral Data of subtropical Tree Species, *Proceedings of Asian Conference on Remote Sensing,* Poster Session 3, November 22-25, Hong-Kong, China URL: http://www.gisdevelopment.net/aars/acrs/1999/ps3/ps3055pf.htm (last date accessed: 27 September 2002).

Gill, P.E., W. Murray, and M.H. Wright, 1991. *Numerical Linear Algebra and Optimization,* Volume 1, Addison-Wesley Publishing Company, pp. 223.

Gopalapillai S., and L. Tian, 1999. In-field variability detection and yield prediction in corn using digital aerial imaging, *Transactions of the ASAE* 42(6), pp. 1911-1920.

Han, J., and Kamber, M., 2001. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA.

Healey, G., and D.A. Slater, 1999. Invariant Recognition in Hyperspectral Images, *IEEE Proceedings of CVPR99,* pp. 438-443.

Hughes, G. F., "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, Vol. IT-14, No. 1, January 1968.

Jasani, B. and G. Stein, 2002. *"Commercial Satellite Imagery: A tactic in nuclear weapon deterrence*," Springer Praxis Publishing Ltd., Chichester, UK.

Jia, X. and J. A. Richards, 1994, Efficient maximum likelihood classification for imaging spectrometer data sets. IEEE Transactions on Geoscience and Remote Sensing 32, 274-281.

Kavzoglu, T., and P.M. Mather, 2000. The Use of Feature Selection Techniques in the Context of Artificial Neural Networks, *26th Annual Conference of the Remote Sensing Society*, September 12-14.

Lee R. 1978. *Forest Microclimatology*, New York, Columbia University Press, pp. 276.

Merényi, E., R. B. Singer, and J. S. Miller, 1996. Mapping of Spectral Variations On the Surface of Mars From High Spectral Resolution Telescopic Images, *ICARUS 1996,* 124, pp. 280-295.

Merényi, E., W.H. Farrand, L.E. Stevens, T.S. Melis, and K. Chhibber, 2000. Mapping Colorado River Ecosystem Resources In Glen Canyon: Analysis of Hyperspectral Low-Altitude AVIRIS Imagery, *Proceedings of ERIM, 14th Int'l Conference and Workshops on Applied Geologic Remote Sensing*, November 4-6, Las Vegas, Nevada, pp. 44-51.

Miller, H.J., and J. Han, 1999. *Discovering geographic knowledge in data-rich environments,* Specialist Meeting Report, National Center for Geographic Information and Analysis Project Varenius, March 18-20, also at URL: http://www.spatial.maine.edu/~max/varenius/KDreport.pdf.

Price, J. C., 1994. Band selection procedure for multispectral scanners. Applied Optics 33, 3281-3288.

Pu, R. and P. Gong, 2000. Band Selection From Hyperspectral Data For Conifer Species Identification, *Proceedings of Geoinformatics'00 Conference,* Monterey Bay, June 21-23, pp.139-146.

Russ, J. C., 1999. "*The Image Processing Handbook,"* Third Edition, CRC Press LLC.

Shettigara, V.K., D. O'Mara, T. Bubner and S. G. Kempinger, 2002. Hyperspectral Band Selection Using Entropy and Target to Clutter Ratio Measures, URL: http://www.cs.uwa.edu.au/~davido/export/hyperspectral.pdf (last date accessed: 29 March 2002).

Swain, P. H. and S. M. Davis, 1978. *Remote Sensing: The Quantitative Approach,* McGraw-Hill, New York.

Veris Technologies, 2003, 601 N. Broadway, Salina, KS 67401, http://www.veristech.com

Warner, T., K. Steinmaus, and H. Foote, 1999. An evaluation of spatial autocorrelation-based feature selection. International Journal of Remote Sensing 20 (8): 1601-1616.

Welge, M., W.H., Hsu, L.S., Auvil, T.M., Redman, and D. Tcheng, 2000. High-Performance Knowledge Discovery and Data Mining Systems Using Workstation Clusters. *12th National Conference on High Performance Networking and Computing (SC99)*, Portland, OR, November 2000.

Wiersma, D. J. and D. A. Landgrebe, 1980, Analytical design of multispectral sensors. IEEE Transactions on Geoscience and Remote Sensing GE-18, 180-189.

Withagen, P.J., E. den Breejen, E.M. Franken, A.N. de Jong, and H. Winkel, 2001. Band selection from a hyperspectral data-cube for a real-time multispectral 3CCD camera, *Proceedings of SPIE AeroSense*, Algorithms for Multi-, Hyper, and Ultraspectral Imagery VII, April 16-20, Orlando, Florida, vol. 4381.

Witten L., and E. Frank, 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufmann Publishers.

Yamagata, Y., 1996. Unmixing with Subspace Method and Application to Hyper Spectral Image, *Journal of Japanese Society of Photogrammetry and Remote Sensing,* Vol.35, pp. 34-42.