INTEGRATION OF HYPERSPECTRAL DATA FOR RANGELAND CHARACTERIZATION

Gregg Petrie[*], Val Cullinan, Larry Cadwell, Janelle Downs, Eileen Perry, and Jim Turner[**]
[*]Battelle Northwest Division, Richland, Washington USA
gregg.petrie@pnl.gov
[**]U.S. Department of the Interior, Bureau of Land Management, National Applied Resource Sciences Center, Denver, Colorado USA

Working Group IV/2

KEY WORDS: Rangeland Health, Hyperspectral Imagery, Landscape Characterization.

ABSTRACT

Grazing lands in the United States consist of approximately 312 million hectares of federal, state, and private land. A recent review by the National Research Council indicates a need to collect comprehensive, consistent, and uniform data to evaluate rangeland health. Pacific Northwest National Laboratory is assisting the U.S. Bureau of Land Management, and working cooperatively with the U.S. Forest Service and the U.S. Natural Resource Conservation Service to address these data needs using new remote sensing methods designed to characterize large areas quickly and at minimal cost. Anticipated application for new remote sensing protocols range from conducting national assessments of the ecological condition of rangelands to monitoring change in the condition of individual grazing allotments. Our work focuses on the spatial characterization of soil stability and the biotic (plant) community using an array of remote sensing data (fine-resolution hyperspectral and multispectral data and coarse-resolution multispectral data), spatial modeling, and statistical analysis. The general strategy is to fuse low-cost Landsat-7 imagery with geographically explicit physical data to derive a first-order spatial characterization of soil and vegetation for large areas of rangelands. A key component of this strategy is the use of hyperspectral data to model/understand the spectral, spatial, and temporal information contained in both the Landsat and new high-resolution systems. Some preliminary results from summer and fall 1999 hyperspectral images (3- to 5-meter resolution) taken over test areas in southwestern Colorado are presented to illustrate the development of a new endmember selection algorithm to unmix hyperspectral data.

1 INTRODUCTION

Many large areas of public lands existing across the Intermountain West are multi-use lands that support various levels of livestock grazing. In 1989, the Board of Agriculture of the National Research Council convened a committee to examine the scientific basis for methods used by several federal agencies to inventory, classify, and monitor rangelands. That committee concluded that effective management of rangelands is hampered because inventory and monitoring data cannot be readily analyzed and integrated to provide information about current rangeland health (NRC, 1994). The health of these vast rangelands is defined by the degree to which the integrity of the soil, vegetation, water and air as well as the ecological processes of rangeland ecosystems are balanced and sustained. Direct measures of the integrity of ecological processes are difficult or expensive to measure due to the complexity of the processes and their interrelationships. Therefore, biological and physical attributes are often used as indicators of the functional status of ecological processes. The need for commonly used indicators was identified (NRC, 1994) and seventeen ecological attributes were selected that could be used to assess soil stability, hydrologic function, and integrity of the biotic community (NRCS, 1997). Current approaches to monitoring rangelands commonly use locally collected field data that are then aggregated to describe rangeland conditions at the landscape or the national scale.

The U.S. Department of Energy's (DOE) Pacific Northwest National Laboratory (PNNL) is assisting the Bureau of Land Management (BLM), and cooperating with the National Resource Conservation Service (NRCS) and the U.S. Forest Service (USFS) to adapt and demonstrate new remote sensing methods for monitoring rangelands in a cost-effective and timely manner. Of 17 rangeland health indicators (NRCS, 1997), we have identified seven that would be effective in monitoring and assessing rangeland health and have the potential of being obtained using advanced remote sensing technology. These parameters include 1) percentage of bare ground, 2) gullies, 3) plant community type, 4) plant functional/structural groups, 5) plant mortality/decadence, 6) annual production, and 7) plant species of interest (particularly noxious/invasive). Our initial efforts focus on examining bare soil and vegetation-related indicators at various spatial and temporal scales using imagery for a study site in southwestern Colorado. We emphasize a wide range of scales because remote sensing applied to grazing lands must serve a wide range of rangeland stewardship functions, from local to regional or national assessments. Furthermore, a number of practical factors must be

considered during the development of remote sensing monitoring methods. Limited budgets are available for monitoring rangelands, yet vast areas need to be characterized in a timely manner. Also, the results need to be made available to multiple endusers, and they must be useful within the context of existing management procedures.

2 REMOTE SENSING DATA FUSION STRATEGY

To meet these complex objectives, the PNNL team developed a strategy based on a stratified hierarchical spatial and temporal design (Figure 1). The hierarchical component of the strategy is designed to allow the most effective combination of sensors to be applied at the appropriate scale for monitoring rangeland health indicators. The stratification component of the design is achieved by identifying relatively homogenous blocks in the landscape according to either physical or biological characteristics. We are currently approaching this stratification of the landscape with two approaches. The first approach is directly analogous to an unsupervised classification. Specifically, we are using slope, aspect, soils, monthly climatic rainfall, and temperature to create an 'image' data cube that is then subject to classical unsupervised classification techniques. One enhancement to this unsupervised methodology is the development of data protocols to first calculate the optimum number of classes to be selected for the unsupervised classification, which is discussed in detail in this paper. The other approach, being lead by BLM at the National Applied Research Center, is a methodology that is analogous to the supervised classification approach. The same data cube is used but with expert information to partition the landscape into homogenous zones with explicit biological meaning. In both cases, the landscape is divided into homogenous zones that can then be further characterized with remotely sensed information. This strategy allows us to apply the appropriate sensors and/or methods to each zone. For example, homogenous zones with relatively dense vegetation cover might best be analyzed using the Normalized Difference Vegetation Index (NDVI) for biomass characterization, whereas arid areas with sparse vegetation might require an index such as the Soil Adjusted Vegetation Index (SAVI), as seen in Figure 2. In addition to different indices or methods, different sensors may be indicated. For instance, using a 1-meter panchromatic system and analyzing spatial texture may permit us to identify change in a sparse vegetation zone, whereas a zone with dense vegetation might be better characterized using a coarser resolution multispectral system.

The hierarchical component of the strategy is designed to use the optimum combination of sensors to provide the most information about rangeland health while taking into account limited operational budgets. Although a single high-resolution hyperspectral airborne sensor has the potential to characterize many of the features of interest by itself, both cost (Cost for the imagery per square kilometer can be 1000 times more expensive for hyperspectral imagery than Landsat) and operational considerations (e.g., massive disk space requirements) preclude this option. Thus, a low-cost, wide area coverage system (currently Landsat fills this role) is first used to roughly characterize the entire study area. This characterization will allow us to pinpoint specific areas of interest that can then be investigated in greater detail using more expensive, but spectrally and spatially rich systems. This hierarchical strategy also allows supporting information to be fused using techniques that have proven particularly useful in characterizing subtle features important to national security issues (Petrie et al., 1997). These features are analogous to particular rangeland health indicators of interest. We intend to apply an array of remotely sensed imagery (fine to coarse spatial resolution multispectral imagery, and fine spatial resolution hyperspectral imagery), spatial modeling, data fusion, and statistical analysis methodology.

The stratified hierarchical strategy is designed to meet both technical demands and practical considerations, but it does require a thorough understanding of the advantages and disadvantages of each potential sensor and auxiliary data set required for the analysis. For instance, optimizing the tradeoffs between using a 30-meter multispectal Landsat image and a 1-meter panchromatic IKONOS satellite image involve not only different spectral, spatial, temporal, and area coverage, but drastically different cost structures, important customer support, and historical acceptance. Furthermore as new satellite systems are launched, it is important to evaluate how these new systems will fit into an overall strategy to avoid costly retrofitting. In addition to understanding relationships between the data sets, these new methods need to be tested and refined to ensure they capture changes in subtle rangeland features. Hyperspectral data sets can be used to address these construction issues. For example, hyperspectral data can be used to model new satellite systems. The rich spectral content available in hyperspectal data sets can often be used to test (i.e., provide a wide variety of test cases), demonstrate, and refine the application of new platforms and analysis and display software.

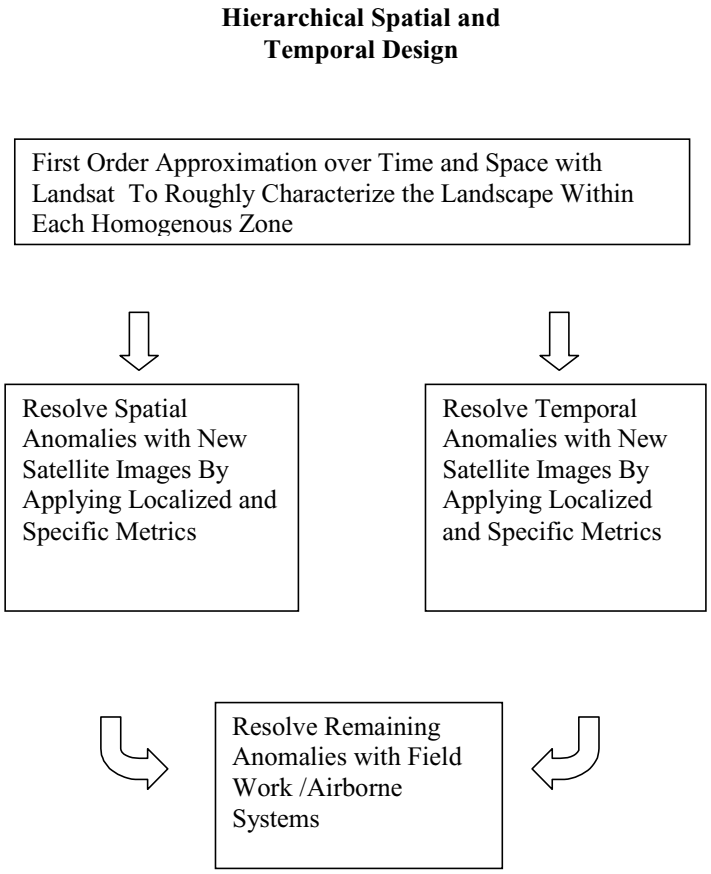

Figure 1. Diagrammatic representation of the Stratified Hierarchical Approach.

We present here two examples of how hyperspectral data are being used to support complex environmental characterization and monitoring. The first example illustrates a method developed to create optimal endmembers. Our goal is to refine the standard pixel unmixing methodology with improved endmember characterization to better support simple multispectral systems, particularly Landsat. The second case illustrates how hyperspectral data are applied to determine the optimum number of classes to use in a classification procedure. The number of cover classes selected for identification in a classification analysis can be less than optimal, particularly if subjective methods are used to select the number of classes used to stratify the landscape. An inappropriate selection of the number of classes can result in over- or under-defining the imagery.

3 OPTIMUM ENDMEMBER ANALYSIS

To detect and characterize subtle rangeland health signatures we are investigating the use of linear unmixing algorithms (e.g., Smith et al., 1990; Nash and Conel, 1974) to help reduce the confounding effect of mixed pixel signatures. Particularly with pixels of Landsat size, the observed spectra can be a combination of several individual spectral signatures associated with several different materials found within one 30-meter pixel. However, with appropriate characterization of the 'pure' spectra of each of contributing material (i.e., the endmember) it is possible to calculate the proportion of each endmember under the appropriate conditions in a given pixel. One critical condition, that is not always easily met, is the need to accurately define both the spectral signature and the number of endmembers to physical and mathematically describe the problem space. Failure to do so can significantly confound the analysis. This can be particularly true for the sometimes subtle changes in rangeland health that are of direct interest to rangeland managers. For these reasons, we have been exploring new methods to objectively select correct endmembers. We would also like this metric to be accessible to a wide variety of potential users in a way that would allow them to apply it to a variety of problem sets. One method that is currently very promising and may have general application, is to exploit the fact that we are normally only interested in characterizing only one factor at a time. Thus, the strategy is to fix this one ecological attribute of interest (e.g., % bare ground) and have the algorithm select both the signature and number of the other endmembers that are mathematically optimal. This strategy is designed to provide the required information (e.g., % bare ground) and the most realistic mathematical description (i.e., the calculated endmembers) at the same time.

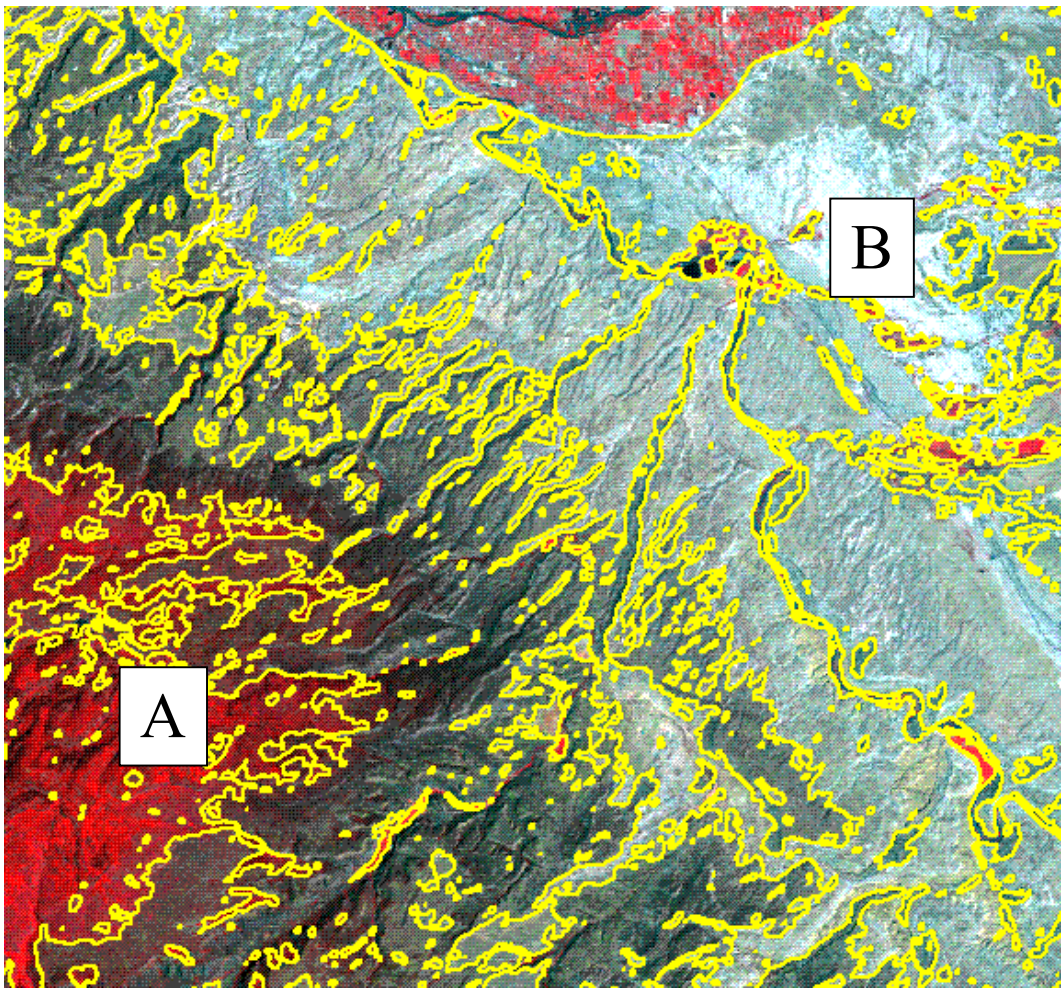

Figure 2. Landsat imagery near some of the rangeland test sites, showing landscape zones which might be analyzed with different indices, or even different sensors. Areas of greater (A) and lesser (B) vegetation are noted.

In general, the methodology is relatively straightforward and closely follows ideas first presented in Harsanyi (1994). In his paper, Harsanyi presented a method that can be used to partition the entire endmember set into two groups: an endmember of direct interest (vector d in the Harsanyi paper) and the rest into 'interfering' ones. One could then optimize the characterization of vector d by projecting each pixel vector onto a subspace that is orthogonal to the undesired signatures and then, once the interfering signatures have been nulled, the residual is projected onto the signature of interest, d. However, implicit in the development was the idea that the users would supply the perfect spectral characterization of all endmembers. We have experimented with extending this methodology to calculate the endmembers by first partitioning an image of interest into sets of pixels dominated by the feature of interest (e.g., a noxious weed) and other pixels (e.g., paved roads, water, shade, grass, trees.) that are potential candidates for endmembers. Typically one might reuse the training sets used in a supervised classification. Given the spectral characterization of the vector d (e.g., the spectral vector that idealizes a particular noxious weed) the procedure then proceeds iteratively:

Step 1: Calculate the single band depicting the fractional amount of the feature of interest following the procedure in Harsanyi (1994). If you have picked your endmember set well, then the pixels identified in your training set has no containing material d (e.g., no noxious weed present); it should be zero under ideal conditions.

Step 2: Within the set of candidate pixels (i.e., no noxious weed present), find the pixel that is farthest from the ideal, and add this vector to the set of interfering endmembers that should be accounted for.

Step 3: Continue this procedure until either: all candidate values are near zero, the maximum number of endmembers is reached, or the most of the candidate pixels are reduced to noise.

The initial application of this methodology (for a worked example see Petrie, 1997) with hyperspectral data sets has produced good results. More importantly, the previous work has laid the groundwork for exploiting lower spectral resolution data sets (e.g., Landsat, IKONOS) for rangeland health assessment. This hyperspectral experimentation identified several factors that should be considered when attempting the more difficult problem of using multispectral images (e.g., severe constraints imposed on the number of endmembers possible). First, the methodology can be conceptually viewed as a numerical optimization search for the best subspace; with no guarantee that the best endmembers have been found. There is always the danger of finding a local minimal. Second, using the most extreme signature in an image as a candidate endmember could present problems if it is not representative (e.g., using a signature from a small anomalous man-made structure in an otherwise natural scene).

4 OPTIMUM LANDSCAPE CLASSIFICATION

The number of cover classes used in a classification analysis can be less then optimal. Selecting too many or too few classes can cause either the consideration of more classes than the data can support or can be the cause for ignoring useful information inherent in the data set. Moreover, determining the error derived from specifying the number of classes in the model incorrectly has been largely ignored. Therefore, selecting the correct number of classes for the rangeland health protocols developed for this project could be particularly important for applications with similar objectives. The approach consists of two steps, based on statistical analysis with few user-selected thresholds.

Step 1: Determine the maximum number of classes based on hierarchichal clustering.
Data in the form of reflectance from each band for a given pixel are sampled using a random sample of n = 200 pixels (user selected) to reduce spatial correlation (i.e., correlations between neighboring pixels) and to reduce the dimension of the matrix to be inverted later during discrimination of classes. A joining tree clustering algorithm using Euclidean distance as a measure of the difference between observations (i.e., the square root of the sum of squared differences between two observations for all variables measured) and a complete linkage rule for combining clusters (i.e., clusters are combined based on the Euclidean distance between the points farthest apart) was used to determine the range of clusters evaluated. The number of clusters achieved at 10 or 15% (user selected) of the maximum linkage distance such that the percent of classes with less than five observations remains less than 40% defines the set of classes to be considered.

Step 2: Determine the optimum number of classes based on k-means clustering.
The optimum number of classes to identify is determined by the resulting analysis of variance, class assignments, and the Euclidean distance from class centroids produced from a k-means clustering analysis for k = 2 to the maximum number of classes considered (from Step 1). The analysis of variance results (i.e., sum of squares between = SSB and sum of squares within = SSE) from each band are totaled to produce a weighted grand F-statistic. Weights are the proportion of the total sum of squares associated with a given band corrected for the mean. The weighted F-statistic is then equal to the weighted SSB divided by (k-1) (i.e., MSB) divided by the weighted SSE divided by (n-k) (i.e., MSE). The F-statistic will be at a maximum when the number of classes produces a maximum distance between classes (MSB) and a minimum distance between observations within a class (MSE). In order to maximize the number of classes identified, we have allowed the optimum number of classes to be greater than the number of classes that achieved the maximum F-statistic. However, to constrain the optimum number of classes, we have limited the number of classes that have less than 5 observations. Further, we have constrained the maximum Euclidean distance from an observation to its class centroid to less than the 75th percentile achieved from all observed distances assuming only two possible classes. Thus, our protocol for choosing the optimum number of classes is conducted by stepping from k = 2 to k = maximum number of classes, and calculating the F-statistic, until achieving the highest number of classes (k) that meets the following three criteria:

1. The percent difference between the maximum F-statistic (i.e.,, between classes mean square divided by the within classes mean square, using the weighted total sum of squares, where the weights are the proportion of the total sum of squares associated with each band) and the achieved F-statistic is less than or equal to 20% (user selected);
2. The percent of classes out of the total considered with less than five observations is less than or equal to 20% (user selected); and
3. The maximum distance of any observation from its class centroid is less than the 75th percentile of all distances observed using only two classes.

A test of this protocol was conducted using hyperspectral data (5-meter pixel resolution) and a Landsat image (30-meter pixel resolution) from the same region. We anticipated that the optimum number of classes should decrease as pixel size increases, although the difference in extents of the imagery added a confounding effect. For direct comparison of

the results, the hyperspectral bands corresponding to the Landsat 7 band centers were selected, using 172 pixels randomly selected across the hyperspectral image.

Joining tree clustering of the Landsat subset of the Whitewater region resulted in 43 classes at 10% and 22 classes at 15% of the maximum linkage distance. Analysis of 22 classes yielded over 40% of the classes with less than five observations; thus, 22 was the maximum number of classes evaluated with k-means clustering. The maximum F-statistic was achieved at 3 classes; however 12 classes met all the criteria (Figure 3). From the diagnostic plot, it is apparent that analysis of a number of classes greater than 12 resulted in a weighted F-statistic that was over 20% smaller than that achieved with three classes. Also, we can see that an analysis of a number of classes greater than 19 yielded too many classes with less than five observations. Thus, 12 classes were chosen as the optimum number of classes for identification. The same analysis was applied to the hyperspectral imagery, resulting in 17 classes being chosen as the optimum number of class sizes for identification. The higher spatial resolution did yield a greater optimum number of classes. This result is likely due to the smaller pixel size of hyperspectral data, which includes greater variability in spectral response (less smoothing) than found in the larger pixel size imagery.

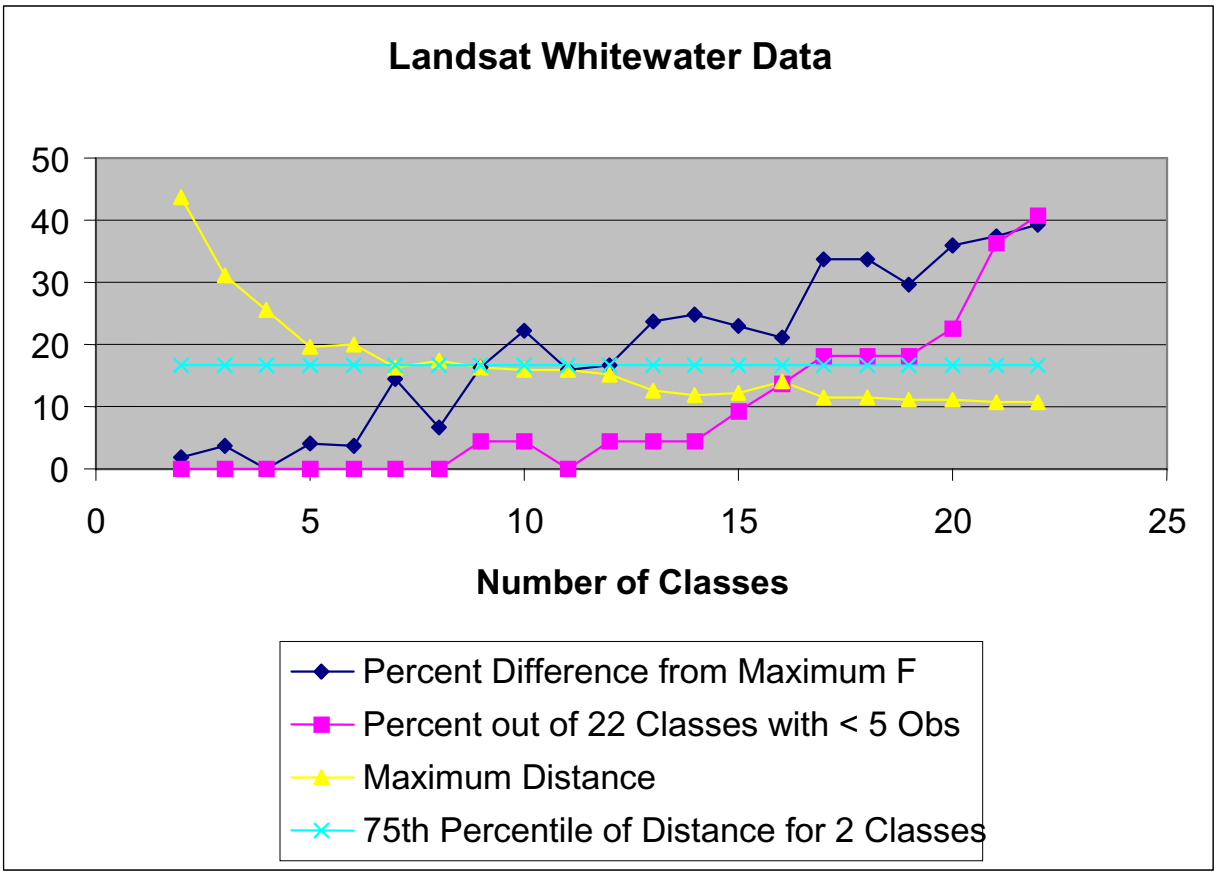

Figure 3. Diagnostic plot to determine the optimum number of classes using the seven bands from 200 pixels randomly sampled from the Landsat footprint of the Whitewater Region. For this example, a k value of 12 (12 classes) met all of the criteria.

The protocol developed to determine the number of classes for identification is not limited to image data, and thus can provide the basis for the optimal physical stratification of the landscape with Geographic Information System layers. This is an important aspect of our strategy for effectively exploiting Landsat and other imagery. Further, the protocol also aids the collection of field data by supporting the optimal classification of a landscape into homogenous blocks for stratified random sampling. This protocol may have applications for supervised classification as well, by guiding the user to select the optimum number of training sets.

5 CONCLUSIONS

Cost, data storage considerations, limited availability, and other practical considerations preclude the use of hyperspectral data to explicitly cover large land areas associated with rangeland characterization. However, both the examples given above, and ongoing work with the use of hyperspectral data sets, suggest that hyperspectral data sets can play an important role in solving complex environmental problem sets both directly and indirectly. Hyperspectral data sets can be used as a test platform to model new satellite systems (e.g., Orbview 4) and their use in rangeland characterization. In the best case, they can be used to help design new systems that are optimal for rangeland characterization. Hyperspectral data can be used to model tradeoffs between high and low-resolution systems to find the optimal mix of sensors in our hierarchical strategy. These analyses provide information needed to refine, test, and understand new methods for selecting cost-effective sensor arrays. These methods also help to direct the application of hyperspectral sensors to those critical areas identified with broad area coverage systems. Our experience to date indicates that there is an important role for hyperspectral data in characterizing large land areas such as those required for monitoring indicators of rangeland health.

ACKNOWLEDGMENTS

Funding for this work was provided by the Bureau of Land Management, U.S. Department of the Interior.

REFERENCES

Harsanyi, J.C., and C-I. Chang, 1994. Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. Trans. On Geosci and Rem. Sensing, (32), pp. 779-785.

Nash, E. B., and J.E. Conel, 1974. Spectral reflectance systematics for mixtures of powdered hypersthene, labradorite, and ilmenite, J. Geophys. Res., (79), pp.1615-1621.

NRC (National Research Council), 1994. Rangeland Health: New Methods to Classify, Inventory, and Monitor Rangelands. National Academy Press, Washington, D.C.

NRCS (National Resources Conservation Service), 1997. National range and pasture handbook. 190-VI, NRCS, U.S. Department of Agriculture, Washington, D.C.

Petrie, G.M., E.M. Perry, R.R. Kirkham, D.W. Slater, P.G. Heasler, and H.P. Foote, 1997. Advanced Image Collection, Information Extraction, and Change Detection in Support of NN-20 Broad Area Search and Analysis. PNNL-11729, Pacific Northwest National Laboratory, Richland, WA-USA

Smith, M.O., S.L. Ustin, J.B. Adams, and A.R. Gillespie, 1990. Vegetation in deserts I. A regional measure of abundances from multispectral images. Remote Sens. Environ., (29), pp.1-26.