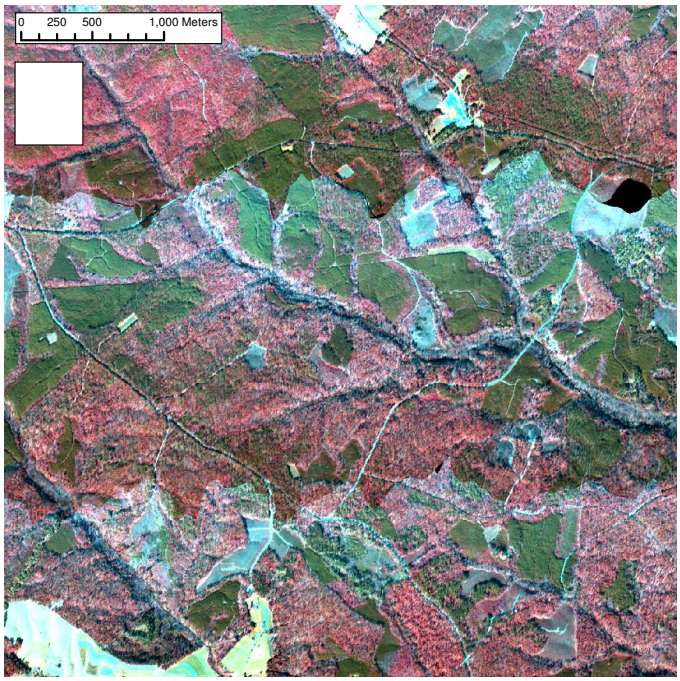

Fig. 9. AVIRIS image (three flight lines) taken over Appomattox Buckingham State Forest in Virginia, USA.

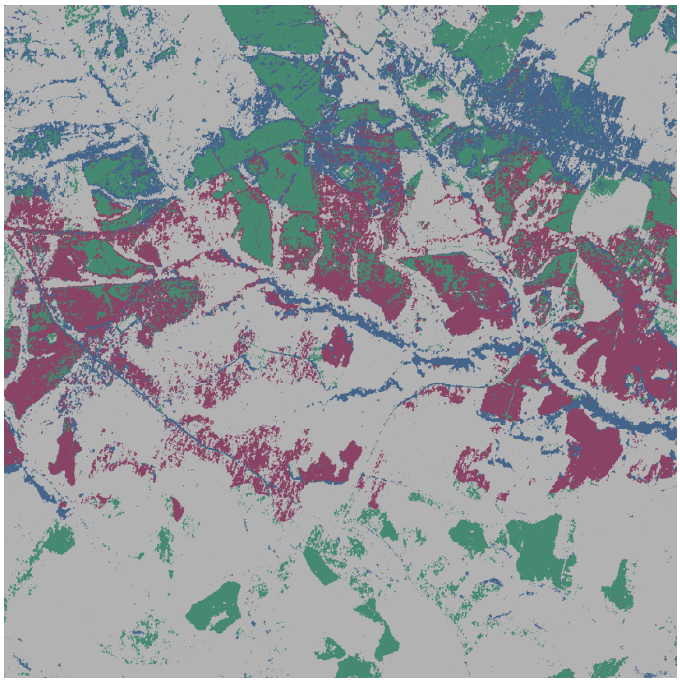

Fig. 10. IGSCR IS classification of ABSF.

the upper half on the zoomed image (Figs. 6, 7, and 8). Fig. 6 indicates a likelihood that there is insufficient training data for these regions. Ultimately these water and shadow regions are misclassified using the decision rule in IGSCR (not pictured), and these regions are classified incorrectly using CIGSCR with Euclidean distance squared. However, notice in Fig. 8 that the CIGSCR IS using Euclidean distance to the fourth power correctly classified the river and the shadow regions. With soft clustering, different clusters were formed, allowing these features to potentially be correctly placed in similar clusters, even though these clusters likely contained small percentages of the training data. In this case, it is potentially useful to know that these features are unclassified (in IGSCR) allowing for modification of the training data, and unfortunately CIGSCR does not

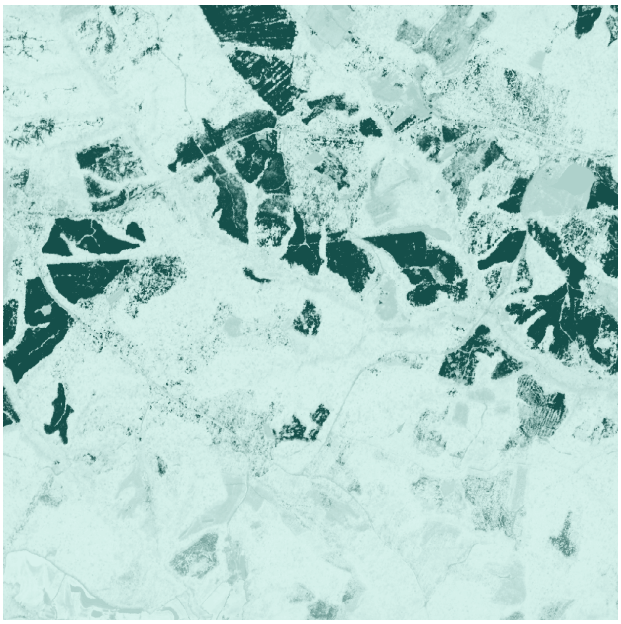

Fig. 11a. CIGSCR IS classification (loblolly pines).

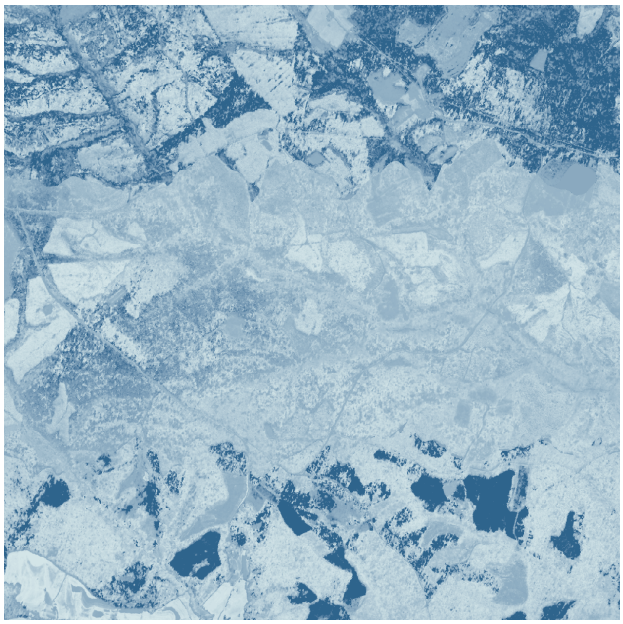

Fig. 11b. CIGSCR IS classification (shortleaf pines).

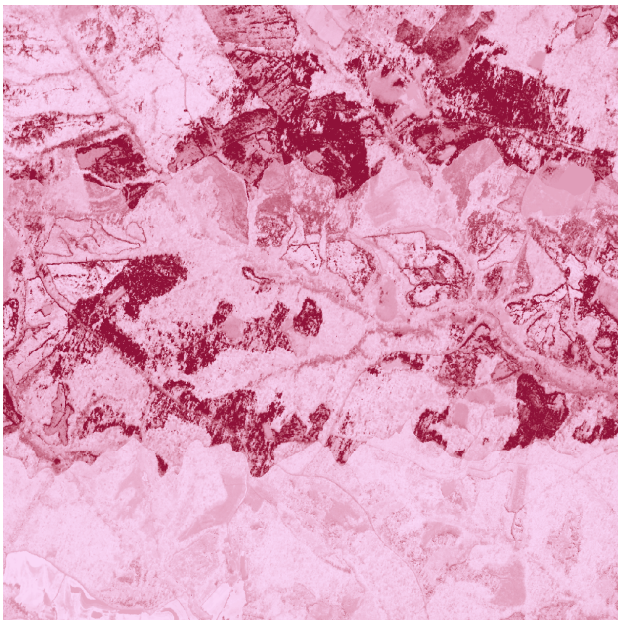

Fig. 11c. CIGSCR IS classification (Virginia pines).

Fig. 11d. CIGSCR IS classification (nonpine).

have this capability. However, when more training samples are not available, CIGSCR can potentially provide a better estimate of the correct class for these data that are not well represented in the training data (although this is obviously not guaranteed as CIGSCR using two different distance functions produced different classification results). Also of interest is that the uncertainty in the soft classifications (regions in beige) does not necessarily match the unclassified regions in Fig. 6. There does not appear to be a correlation between samples that are not part of pure clusters in IGSCR and samples that may belong to multiple classes in CIGSCR.

The accuracies reported for the classification of ABSF tend to be lower than the classification accuracies reported for VA1734, which is reasonable considering the classification of ABSF is attempting to discriminate between spectrally similar pine species, ABSF is noisy, and ABSF contains several heterogeneous areas, making training difficult. Also note that the VA1734 DR classifications were almost always more accurate than corresponding IS classifications, but ABSF DR classifications are often less accurate than corresponding IS classifications. All ABSF classifications (IGSCR DR and IS and CIGSCR DR and IS)

TABLE 3
IGSCR AND CIGSCR DECISION RULE (DR) CLASSIFICATION ACCURACIES FOR ABSF.

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$)			clustering
clusters	$p = .5$	$p = .9$	$\rho = \|\|x-U\|\|_2^2$	$\rho = \|\|x-U\|\|_2^4$	$\rho = e^{\|\|x-U\|\|_2}$	(no iteration)
10	83.50	*	47.50	79.50	72.50	*
15	*	*	62.50	83.50	79.75	*
20	*	*	66.75	73.50	74.25	*
25	51.00	51.00	63.00	75.00	78.75	*

TABLE 4
IGSCR ITERATIVE STACKED PLUS (IS+) AND CIGSCR ITERATIVE STACKED (IS) CLASSIFICATION ACCURACIES FOR ABSF.

no. init.	IGSCR ($\alpha = .01$)		CIGSCR ($\alpha = .0001$)			clustering
clusters	$p = .5$	$p = .9$	$\rho = \|\|x-U\|\|_2^2$	$\rho = \|\|x-U\|\|_2^4$	$\rho = e^{\|\|x-U\|\|_2}$	(no iteration)
10	83.75	*	51.75	84.50	72.75	*
15	*	*	51.00	84.50	83.25	*
20	*	*	51.00	84.00	81.50	*
25	91.00	75.25	51.00	76.75	83.00	*

TABLE 5
FOR VA17 IGSCR, NUMBER OF PURE CLUSTERS. FOR VA17 CIGSCR, THE PAIRS (A,B) = (NUMBER OF CLUSTERS PRODUCED, NUMBER OF ASSOCIATED CLUSTERS).

no. init.	IGSCR		CIGSCR		
clusters	$p = .5$	$p = .9$	$\rho = \|\|x-U\|\|_2^2$	$\rho = \|\|x-U\|\|_2^4$	$\rho = e^{\|\|x-U\|\|_2}$
10	19	6	15,13	11,11	12,12
15	15	6	20,16	20,19	20,20
20	20	18	25,21	21,21	24,24
25	52	17	30,25	30,28	30,29

TABLE 6
FOR ABSF IGSCR, NUMBER OF PURE CLUSTERS. FOR ABSF CIGSCR, THE PAIRS (A,B) = (NUMBER OF CLUSTERS PRODUCED, NUMBER OF ASSOCIATED CLUSTERS).

no. init.	IGSCR		CIGSCR		
clusters	$p = .5$	$p = .9$	$\rho = \|\|x-U\|\|_2^2$	$\rho = \|\|x-U\|\|_2^4$	$\rho = e^{\|\|x-U\|\|_2}$
10	16	8	15,15	10,10	11,11
15	14	11	20,19	15,15	15,15
20	19	9	25,24	20,20	20,20
25	23	15	30,29	25,25	26,26

reasonably separated pines from nonpines, but IGSCR and CIGSCR differed in the identification of individual pines species. Both classification methods identified individual pines in mixed hardwood/pine stands in the top left corner of the image (Figs. 10 and 11a–d). A visual inspection of the classification images reveals that IGSCR and CIGSCR classifications disagree on loblolly (IGSCR has underestimated those stands) and shortleaf (both overestimated). IGSCR incorrectly picked out patches of shortleaf along the "veins" of the image, and both classifications overestimated Virginia pines.

Another potential advantage of CIGSCR with an alternative radial function is the ability to locate clusters associated with classes, even when there is overlap between classes or there is a small amount of training data for a class. IGSCR failed to locate enough pure clusters to perform classification, indicated by an asterisk in Tables 3 and 4, in most ABSF classification attempts. CIGSCR using Euclidean distance squared produced classifications, although the accuracies are low. CIGSCR using alternative radial functions performed reasonable classifications no matter the number of initial clusters. In highly heterogeneous sites like this where limited training data is available for multiple classes, IGSCR has difficulty locating pure clusters. Since multiple classes are spectrally similar, soft clustering allows for small differences between classes in a cluster to be detected.

Hard clusters containing one species would be likely to contain a significant amount of the other species, and would therefore fail the hypothesis test (for reasonable p and α. With soft clustering, portions of both species would be attributed to a soft cluster, but if there is statistical significance of the difference in the memberships of the species, the cluster can be associated and used for training purposes. Furthermore, soft clustering allows for alternative functions to be used to determine cluster assignments. Recall that these radial functions magnify the difference between small and large probabilities, allowing clusters containing these less well represented classes to be formed and allowing samples to have high probabilities of belonging to those clusters.

Finally, perhaps the most important question about this semisupervised clustering scheme is whether using the combination of the association significance test and the iteration improves the clustering for the purposes of classification. Each cluster is labeled with the class that has the highest average membership in the cluster. Observe in experimental runs in Tables 1 and 2 that **all** classification accuracies using just clustering are lower than corresponding classification accuracies using CIGSCR with Euclidean distance. In Tables 3 and 4, iterative refinement was necessary to locate enough clusters (such that each class was represented by at least one cluster) for classification using Euclidean distance squared. Accuracies are much higher using alternative distance functions, but little or no iterative refinement was used. Based on the available results in Tables 1–4, the semisupervised clustering scheme in CIGSCR improves classification accuracies when training data are available to influence clustering.

V. Conclusions

This paper presented a continuous analog to IGSCR that rejects and refines clusters to automatically classify a remotely sensed image based on informational class training data. This new algorithm addressed specific challenges presented by remotely sensed data including large datasets (millions of samples), relatively small training datasets, and difficulty in identifying spectral classes. The resulting classifications are fundamentally different from IGSCR (the discrete predecessor to CIGSCR) classifications, even when converting the CIGSCR soft classifications to hard classifications. CIGSCR has many advantages over IGSCR, such as the ability to produce soft classification, less sensitivity to certain input parameters, ability to use alternative distance functions that often produce more accurate classifications, potential to correctly classify regions that are not amply represented in training data, and a better ability to locate clusters associated with all classes. The semisupervised clustering framework within CIGSCR has been shown here to improve classification accuracies over clustering alone. This semisupervised clustering framework could be incorporated into many classification algorithms that use clustering. The radial functions used in CIGSCR resulted in consistently accurate classifications.

The highly automated CIGSCR classification algorithm is a contribution to the remote sensing community that has few if any automated semisupervised soft classification algorithms analogous to the many automated semisupervised hard classification algorithms that exist. Future work includes using this soft classifier for many applications of classification in remote sensing.

References.

[1] J.P. Wayman, R.H. Wynne, J.A. Scrivani, and G.A. Reams, "Landsat TM-based forest area estimation using Iterative Guided Spectral Class Rejection," *Photogrammetric Engineering & Remote Sensing*, vol. 67, pp. 1155–1166, 2001

[2] R.F. Musy, R.H. Wynne, C.E. Blinn, J.A. Scrivani, and R.E. McRoberts, "Automated Forest Area Estimation via Iterative Guided Spectral Class Rejection," *Photogrammetric Engineering & Remote Sensing*, vol. 72, pp. 949–960, 2006

[3] R.D. Phillips, L.T. Watson, and R.H. Wynne, "Hybrid image classification and parameter selection using a shared memory parallel algorithm," *Computers & Geosciences*, vol. 33, no. 7, pp. 875–897, 2007

[4] J.C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 1, pp. 1–8, 1980

[5] R. D. Phillips, "A probabilistic classification algorithm with soft classification output," PhD Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2009

[6] J.A.N Van Aardt and R.H. Wynne, "Examining pine spectral separability using hyperspectral data from an airborne sensor: An extension of field-based results," *International Journal of Remote Sensing*, vol. 28, no. 2, pp. 431–436, 2007