# Enrichment Procedures for Soft Clusters: A Statistical Test and its Applications

Rhonda D. Phillips *Member, IEEE*, Layne T. Watson *Fellow, IEEE*,
Randolph H. Wynne *Member, IEEE*, and Naren Ramakrishnan *Member, IEEE*

**Abstract**—Clusters, typically mined by modeling locality of attribute spaces, are often evaluated for their ability to demonstrate 'enrichment' of categorical features. A cluster enrichment procedure evaluates the membership of a cluster for significant representation in pre-defined categories of interest. While classical enrichment procedures assume a hard clustering definition, in this paper we introduce a new statistical test that computes enrichments for soft clusters. We demonstrate an application of this test in refining and evaluating soft clusters for classification of remotely sensed images.

## 1 Introduction

Clustering is an unsupervised process that models locality of data samples in attribute space to identify groupings: samples within a group are closer to each other than to samples from other groups. To assess whether the discovered clusters are meaningful, a typical procedure is to see if the groupings capture other categorical information *not originally used during clustering*. For instance, in microarray bioinformatics, data samples correspond to genes and their expression vectors, clusters capture locality in expression space, and they are evaluated to see if genes within a cluster share common biological function/annotations. (Observe that the functional annotations are not used during clustering). In text mining, data samples correspond to documents and their text vectors, clusters capture locality in term space, and are evaluated

R.D. Phillips is with the MIT Lincoln Laboratory, Lexington, MA, 02420.

L.T. Watson and N. Ramakrishnan are with the Departments of Computer Science and Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

R.H. Wynne is with the Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

for their correspondence with *a priori* domain information such as topics. In remote sensing, data samples correspond to pixels in an image, clusters capture locality of pixel intensities, and are evaluated for their correspondence with land cover classifications.

All of the above applications are essentially determining whether locality in one space preserves correspondence with information in another space, also referred to as the *cluster assumption* [1]. While cluster evaluation is typically conducted as a distinct post-processing stage after mining, recently developed clustering formulations blur this boundary. For instance, in [2], locality information is used along with background knowledge to influence the clustering. Such background knowledge takes the form of constraints, some of which dictate that certain samples should appear in the same cluster, while others specify that two samples should be in different clusters. Similarly, in [3], clusters are designed using an objective function that balances compression of the primary random variable against preservation of mutual information with an auxiliary variable. With the advent of semi-supervised clustering [1], more ways to integrate labeled and unlabeled information are rapidly being proposed.

The design of both classical and the newer clustering algorithms is predicated on the ability to evaluate clusters for enrichment and using this information to drive the refinement and subsequent discovery of clusters. However, classical statistical enrichment procedures (e.g., using the hyper-geometric distribution [4]) assume a hard clustering formulation. Our focus here is on soft clusters where the groupings are defined by portions of individual samples. We present a new statistical test to enrich soft clusters and demonstrate its application to a remote sensing context.

## 2 Clustering

### *2.1 Hard Clustering*

Hard clustering produces clusters that are a collection of individual samples. Let the $i$th sample be denoted by $x^{(i)} \in \Re^b$ where $i = 1, \ldots, n$. A cluster is typically represented by a prototype, such as the mean of the samples contained in the cluster, and let the $j$th cluster prototype be $U^{(j)} \in \Re^b$ where $j = 1, \ldots, K$. All clusters taken together form a partition of the data, defined by a partition matrix, $w$ with $w_{ij} = 1$ indicating that the $i$th sample belongs to the $j$th cluster, $w_{ij} = 0$ otherwise, and $\sum_{j=1}^{K} w_{ij} = 1$ for all $i$. Each sample is a member of exactly one cluster.

A classic example of a simple hard clustering method is the $K$-means clustering algorithm that locates a local minimum point of the objective function

$$J(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \rho_{ij} \tag{1}$$

subject to

$$\sum_{j=1}^{K} w_{ij} = 1, \quad \text{for } i = 1, \ldots, n,$$

where $\rho_{ij} = \|x^{(i)} - U^{(j)}\|_2^2$ [5]. In this case, $\rho_{ij}$ is a measure of dissimilarity or distance between the $i$th sample and the $j$th cluster. The $K$-means clustering algorithm attempts to find the ideal partition that minimizes the sum of squared distances between each sample and the prototype of the cluster to which the sample belongs. The algorithm for $K$-means requires $K$ initial cluster prototypes and iteratively assigns each sample to the closest cluster using

$$w_{ij} = \begin{cases} 1, & \text{if } j = \operatorname*{argmin}_{1 \le j \le K} \rho_{ij}, \\ 0, & \text{otherwise,} \end{cases}$$

for each $i$, followed by the cluster prototype (mean) recalculation

$$U^{(j)} = \sum_{i=1}^{n} (w_{ij} x^{(i)}) \Big/ \sum_{i=1}^{n} w_{ij}$$

once $w$ has been calculated. This process, guaranteed to terminate in a finite number of iterations, continues until no further improvement is possible, terminating at a local minimum point of (1).

In hard clusters, such as those produced by $K$-means, the collection of samples that belong to a particular cluster can be evaluated to determine a cluster's eligibility to perform classification. The class memberships of the labeled samples in a particular cluster can be modeled using discrete random variables generated from binomial, multinomial, or hypergeometric distributions, for example. These random variables form the basis of statistical tests used to evaluate clusters for classification. For example, let $V_{ic}$ be a Bernoulli random variable where success ($V_{ic} = 1$) indicates the $i$th labeled sample is labeled with the $c$th class. The number of labeled samples labeled with the $c$th class in a particular cluster would be a binomial random variable $V_{c,j} = \sum_{i \in I_j} V_{ic}$ where $I_j$ is the index set of labeled samples belonging to the $j$th cluster. This binomial random variable can be used as the basis for a statistical hypothesis test to determine if the number of samples labeled with the $c$th class (as opposed to all other classes) in the $j$th cluster is significant. In practice, the $c$th class that would be tested would be the class that is most represented in the $j$th cluster, or mathematically, $c = \text{argmax}_{1 \leq c \leq C} V_{c,j}$ for a particular $j$ where $C$ is the number of classes.

### *2.2 Soft Clustering*

Soft clusters are clusters that instead of containing a collection of individual samples, contain portions of individual samples. Another view of soft clustering is that each sample has a probability of belonging to a particular cluster. Soft clustering has advantages over hard clustering in that a sample is not simply assigned to the closest cluster, but information is preserved about relationships to other clusters as well. Furthermore, these continuous assignments are less constrained that discrete assignments, resulting in a less constrained objective function. Like in hard clustering, $w_{ij}$ indicates cluster membership, but instead of being either zero or one, $w_{ij} \in (0, 1)$, and like in hard clustering, $\sum_{j=1}^{K} w_{ij} = 1$ for all $i$. Some versions of fuzzy clustering do not impose this requirement, but those non-probabilistic methods will not be considered here.

An example of a soft clustering method analogous to $K$-means is fuzzy $K$-means that locates a local minimum point of the objective function

$$J(\rho) = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij}^{p} \rho_{ij} \tag{2}$$

subject to

$$\sum_{j=1}^{K} w_{ij} = 1$$

where $\rho_{ij}$ is still the squared Euclidean distance between $x^{(i)}$ and $U^{(j)}$ and $p > 1$ [6]. The algorithm that minimizes this objective function is similar to that of $K$-means in that it first calculates

$$w_{ij} = \frac{(1/\rho_{ij})^{1/(p-1)}}{\displaystyle\sum_{k=1}^{K} (1/\rho_{ik})^{1/(p-1)}}$$

for all $i$ and $j$ followed by calculating updated cluster prototypes

$$U^{(j)} = \sum_{i=1}^{n} w_{ij}^{p} x^{(i)} \Bigg/ \sum_{i=1}^{n} w_{ij}^{p}.$$

The cluster prototype is a weighted average. This iteration (recalculation of the weights followed by recalculation of cluster prototypes, following by recalculation of the weights, etc.) is guaranteed to converge (with these definitions of $\rho_{ij}$, $U^{(j)}$, and $w_{ij}$) for $p > 1$ [7].

## 3 Soft Cluster Evaluation

Evaluation of soft clusters requires taking cluster weights into account when examining class memberships of the labeled samples. Each labeled sample will have some positive membership in each cluster, and a new type of evaluation will be necessary to directly evaluate soft clusters. Soft cluster memberships could be converted to hard cluster memberships for the purpose of cluster evaluation, but if soft clustering is warranted, those soft clusters should be evaluated directly.

Hard cluster evaluation (for classification) is based on the composition of the cluster, or what type of samples are making up the cluster. The question of whether a cluster should be used for classification can be answered when some of the samples within the cluster have labels and there are a sufficient number of samples to draw statistical conclusions. Because soft clusters no longer "contain" samples, the more important question is whether the relative magnitudes of memberships between samples of a particular class and the cluster are significantly different. In other words, if the magnitude of cluster memberships for samples of a particular class appear to be significantly higher than memberships for other classes, then the cluster is demonstrating characteristics of that class. With hard clusters, a cluster is pure if only one class is contained in the cluster; no samples labeled with another class are present in the cluster. This is impossible in soft clustering as all types of samples will have positive memberships in all clusters, and in practice, these memberships, although possibly small, will be nonnegligible.

Just as hard clusters that are ideal for classification contain only one class, soft clusters that are ideal for classification will be representative of just one class. The goal in using soft clustering for classification is to assign a class label to an entire cluster (the same goal for hard clusters), but just as each sample has a soft membership in a particular cluster, each sample will have soft membership in a class. The samples demonstrate characteristics of multiple classes, justifying soft classification, but the clusters (logical grouping of similar data) should not contain or represent multiple classes. The goal of this work is to associate a soft cluster to one particular class if that class is clearly dominant within the cluster. Probability will determine how clearly a particular cluster is composed of one class, and if this probability passes a predetermined threshold test, the cluster will be associated with a class.

### *3.1 Hypothesis Test*

The statistical tests used to evaluate clusters in this paper are statistical hypothesis tests, where a null hypothesis is proposed. If observed evidence strongly indicates the null hypothesis should be rejected, the alternate hypothesis will be accepted. In the absence of compelling evidence to the contrary, the null hypothesis cannot be rejected.

The first hypothesis test is based on the average cluster weights in the cluster of interest, the $j$th cluster. In order to associate the $j$th cluster to the $c$th class, the average cluster weight for the $c$th class

$$\overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij},$$

where $n_c$ is the number of samples labeled with the $c$th class and $J_c$ is the index set of samples labeled with the $c$th class, should be statistically significantly higher than other cluster weights for the $j$th cluster. If the weights for samples labeled with the $c$th class are higher in general than samples from arbitrary classes, the cluster is demonstrating a tendency to the $c$th class, and can be used to discriminate the $c$th class from other classes.

The null hypothesis is that the average cluster weights for samples from the $c$th class in the $j$th cluster is not significantly different from the average cluster weight for samples from all classes in the $j$th cluster. The alternate hypothesis is that the average weight for samples from the $c$th class in the $j$th cluster is significantly different (higher) than the average cluster weight for all samples. Note that in practice, only the class with the highest average cluster weight for the $j$th cluster would be considered. Suppose that a test statistic derived for this test is normally distributed, and is in fact a standard normal random variable $Z$. Then if the observed value is $\hat{z}$, if $P(Z \geq \hat{z}) \leq \alpha$ for $0 < \alpha < 1$, the null hypothesis is rejected. The following sections derive appropriate test statistics to use in this hypothesis test.

### *3.2 Test Statistic 1*

Suppose a dataset $x$ contains $n$ samples $x^{(i)} \in \Re^B$, $i = 1, \ldots, n$. For $K$ fixed cluster centers $U^{(k)} \in \Re^B$, $k = 1, \ldots, K$, the assigned weight of the $i$th pixel to the $j$th cluster is

$$w_{ij} = \frac{1/\|x^{(i)} - U^{(j)}\|_2^2}{\sum_{k=1}^{K} 1/\|x^{(i)} - U^{(k)}\|_2^2},$$

which is the inverse of the distance squared over the sum of the inverse squared distances. (Such inverse distance weights are widely used, e.g., by Shepard's algorithm for sparse data interpolation.) Note this is the specific case in the soft clustering algorithm described above when $p = 2$. In many practical applications where a dataset is to be clustered (such as the clustering of a remotely sensed image), it is reasonable to assume that $x^{(i)}$, $i = 1, \ldots, n$ are generated from a finite number of multivariate normal distributions. The

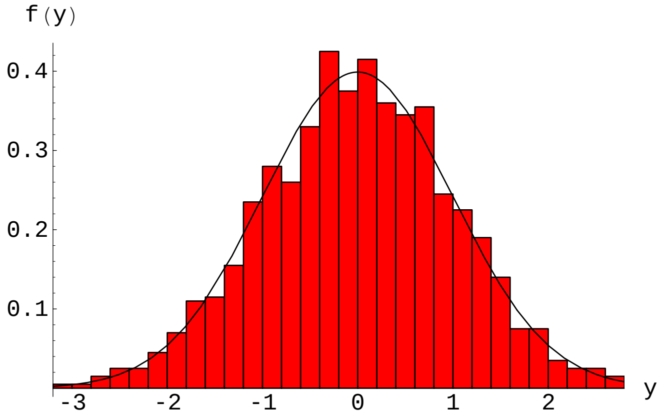

Fig. 1. Distribution of sums of weights in one soft cluster out of two.

act of clustering assumes that the data are generated from a finite number of distributions. The following theorem from [8] demonstrates that under these assumptions (samples are generated from a finite number of normal distributions), the Lindeberg condition is satisfied and therefore the central limit theorem applies to the sum of a sequence of cluster weight random variables $\sum_{i=1}^{n} W_{ij}$. Let $q = \psi(i)$ denote the distribution from which the random vector $X^{(i)}$ was sampled.

*Theorem:* Let $X^{(i)}$ , $i = 1, 2, \ldots$, be $B$-dimensional random vectors having one of $Q$ distinct multivariate normal distributions. For $i = 1, 2, \ldots$ and $j = 1, \ldots, K$ define the random variables

$$W_{ij} = W_j(X^{(i)}) = \frac{1/\|X^{(i)} - U^{(j)}\|_2^2}{\sum_{k=1}^{K} 1/\|X^{(i)} - U^{(k)}\|_2^2},$$

where $K$ is the number of clusters and $U^{(k)} \in \Re^B$ is the $k$th cluster center (and is considered fixed for weight calculation). Then for any $j = 1, \ldots, K$,

$$P\left\{\frac{1}{B_{nj}} \sum_{i=1}^{n} (W_{ij} - a_{ij}) < x\right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$$

as $n \rightarrow \infty$, where $a_{ij} = \mathrm{E}[W_{ij}]$, $b_{ij}^2 = \mathrm{Var}[W_{ij}]$, and $B_{nj}^2 = \sum_{i=1}^{n} b_{ij}^2$.

*Remark:* The assumption that the $X^{(i)}$, $i = 1, 2, \ldots$, are generated from a finite number of normal distributions is stronger than necessary. The proof in [8] holds if $X^{(i)}$, $i = 1, 2, \ldots$, are generated from a finite number of arbitrary distributions.

Experimental clustering results using a dataset described in Section 4.2 of this paper match this theoretical result, as illustrated by one experiment in Fig. 1. This illustration shows the distribution of sums of cluster weights for one particular cluster (when $K = 2$).

Starting with the normal approximation for the sum of the cluster weights, the standard normal test statistic would be

$$\hat{z} = \frac{\sum_{i \in J_c} \left( w_{ij} - \mathrm{E}[W_{ij}] \right)}{\sqrt{\sum_{i \in J_c} \mathrm{Var}[W_{ij}]}},$$

where $\mathrm{E}[W_{ij}]$ is the expected value of $W_{ij}$ and $\mathrm{Var}[W_{ij}]$ is the variance of $W_{ij}$ for the $j$th cluster. $\mathrm{E}[W_{ij}]$ and $\mathrm{Var}[W_{ij}]$ are unknown, but can be reasonably approximated using the sample mean

$$\overline{w}_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}$$

and sample standard deviation

$$S_{\overline{w}_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (w_{ij} - \overline{w}_j)^2}.$$

The Wald statistic is then

$$\hat{z} = \frac{\sqrt{n_c}(\overline{w}_{c,j} - \overline{w}_j)}{S_{\overline{w}_j}}, \tag{3}$$

where

$$\overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}.$$

Since $\hat{z}$ is generated (approximately) by the standard normal distribution, this test statistic can be used in the proposed hypothesis test.

### *3.3 Test Statistic 2*

One potential issue with the above statistic is that the sample mean and standard deviation calculations assume the sample is identically distributed, which is specifically *not* the assumption in this case (clustering assumes that the data are generated from a number of distributions, where the true number of clusters is equal to the number of distributions, which is unknown apriori) . A better statistic acknowledges that the data are not identically distributed, but are generated from a finite number of distributions. Since the number of distributions and the distributions are unknown, the number of classes and the individual class labels, which are assumed to correspond to inherent structure of the data, are used to approximate the true mean and variance of multiple clusters. Precisely, assume that all labeled sample indices $i$ with distribution index $\psi(i) = q$ correspond to the same class label $\phi(i) = c$. If $i \in \psi^{-1}(q)$, then $i \in \phi^{-1}(c)$, but $i \in \phi^{-1}(c)$ does not imply $i \in \psi^{-1}(q)$ (more than one distribution can correspond to one class), and $J_c = \phi^{-1}(c) = \{i \mid \phi(i) = c, 1 \leq i \leq n\}$. The above statistic requires modification to use class information. In the previous statistic,

$$\sum_{i \in J_c} w_{ij} = \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c},$$

$$\hat{z} = \frac{\displaystyle\sum_{i=1}^{n}\big(w_{ij}\delta_{\phi(i),c} - \mathrm{E}[W_{ij}\delta_{\phi(i),c}]\big)}{\sqrt{\displaystyle\sum_{i=1}^{n} \mathrm{Var}[W_{ij}\delta_{\phi(i),c}]}},$$

$$\begin{aligned}
&\sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - \mathrm{E}[W_{ij}\delta_{\phi(i),c}]) \\
&= \sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - a_{ij}\delta_{\phi(i),c}) \\
&= \sum_{i=1}^{n}(w_{ij}\delta_{\phi(i),c} - \alpha_{qj}\delta_{\phi(i),c}),
\end{aligned}$$

recalling that $\mathrm{E}[W_{ij}] = a_{ij} = \alpha_{qj}$ for $i \in I_q$. Assume when $\phi(i) = c$, and distribution index $q = \psi(i)$ corresponds to $c = \phi(i)$, then $\alpha_{qj}$ can be approximated by $\gamma_{cj}$, the mean of class $c = \phi(i)$. Ideally $\alpha_{qj}$ should be approximated directly, but there is no way to know $\psi^{-1}(q)$, so essentially $\psi^{-1}(q) \subset \phi^{-1}(c)$ is being approximated by $\phi^{-1}(c)$. Unfortunately, using the sample mean of the $c$th class and the $j$th cluster to approximate $\gamma_{cj}$ and therefore $\alpha_{qj}$ breaks down because the sample mean of the $c$th class and the $j$th cluster is both the random variable on the left side and the approximation of the expected value on the right side of the minus sign. This is illustrated below. Approximating $\gamma_{cj}$ (and $\alpha_{qj}$) with the sample mean for the $c$th class,

$$\gamma_{cj} \approx \overline{w}_{c,j} = \frac{\displaystyle\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\displaystyle\sum_{k=1}^{n} \delta_{\phi(k),c}},$$

the numerator of the test statistic $\hat{z}$ becomes

$$\sum_{i=1}^{n}\big(w_{ij}\delta_{\phi(i),c} - \overline{w}_{c,j}\delta_{\phi(i),c}\big)$$

$$= \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \frac{\displaystyle\sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c}}{\displaystyle\sum_{k=1}^{n} \delta_{\phi(k),c}} \sum_{i=1}^{n} \delta_{\phi(i),c}$$

$$= \sum_{i=1}^{n} w_{ij}\delta_{\phi(i),c} - \sum_{k=1}^{n} w_{kj}\delta_{\phi(k),c} = 0.$$

Thus this test statistic does not work because the value being tested is the same as the estimated mean for the $c$th class.

In order to make use of class information to estimate distribution statistics (mean and variance), it is necessary to modify the random variable to model class labels as well as cluster memberships. Consider each labeled sample's membership in a particular class, say the $c$th class, to be a Bernoulli trial $V_{ic}$, where $V_{ic} = 1$ indicates the $i$th sample is labeled with the $c$th class, and $W_{ij}$ is defined above. Define

$$Y_{c,j} = V_{1c}W_{1j} + V_{2c}W_{2j} + \cdots + V_{nc}W_{nj},$$

where $n$ is the total number of labeled samples as the random variable for the sum of weights for samples in the $c$th class to the $j$th cluster. The Central Limit Theorem applies to this sum of bounded random variables with finite mean and variance (see Theorem 1), and $Y_{c,j}$ is approximately normal.

Consider now the test statistic

$$\hat{z} = \frac{y_{c,j} - \mathrm{E}[Y_{c,j}]}{\sqrt{\mathrm{Var}[Y_{c,j}]}}.$$

Fixing $j$ and $c$, assuming $W_{ij}$ and $V_{ic}$ are independent, and defining $m_q = |I_q|$, the number of indices $i$ for which $X^{(i)}$ has the $q$th distribution,

$$\mathrm{E}[Y_{c,j}] = \mathrm{E}\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] = \sum_{i=1}^{n} \mathrm{E}[W_{ij}V_{ic}]$$

$$= \sum_{i=1}^{n} \mathrm{E}[W_{ij}]\mathrm{E}[V_{ic}] = \sum_{q=1}^{Q} m_q \alpha_{qj} p_c = p_c \sum_{q=1}^{Q} m_q \alpha_{qj},$$

where $p_c$ is the probability that $V_{ic} = 1$. Assuming all the samples are independent and recalling that

$\text{Var}[W_{ij}] = b_{ij}^2 = \beta_{qj}^2$ where $i \in I_q$,

$$\begin{aligned}
\text{Var}[Y_{c,j}] &= \text{Var}\left[\sum_{i=1}^{n} W_{ij}V_{ic}\right] = \sum_{i=1}^{n} \text{Var}[W_{ij}V_{ic}] \\
&= \sum_{i=1}^{n} \big(\text{E}[W_{ij}^2 V_{ic}^2] - \text{E}[W_{ij}V_{ic}]^2\big) \\
&= \sum_{i=1}^{n} (p_c \text{E}[W_{ij}^2] - p_c^2 a_{ij}^2) \\
&= \sum_{i=1}^{n} (p_c(b_{ij}^2 + a_{ij}^2) - p_c^2 a_{ij}^2) \\
&= \sum_{q=1}^{Q} m_q \big(p_c(\beta_{qj}^2 + \alpha_{qj}^2) - p_c^2 \alpha_{qj}^2\big) \\
&= p_c \sum_{q=1}^{Q} m_q(\beta_{qj}^2 + (1 - p_c)\alpha_{qj}^2).
\end{aligned}$$

In the above formula, $p_c$ would be approximated by its maximum likelihood estimate $n_c/n = |J_c|/n$. In order to estimate $\alpha_{qj}$, assume that the $q$th distribution corresponds to the $c$th class, $\psi^{-1}(q) \subset \phi^{-1}(c)$, and

$$\alpha_{qj} \approx \overline{w}_{c,j} = \frac{1}{n_c} \sum_{i \in J_c} w_{ij}, \quad c = 1, \ldots, C,$$

where $C$ is the number of classes. Then

$$\begin{aligned}
\text{E}[Y_{c,j}] = p_c \sum_{q=1}^{Q} m_q \alpha_{qj} &\approx p_c \sum_{d=1}^{C} n_d \cdot \frac{1}{n_d} \sum_{i \in J_d} w_{ij} \\
&= \frac{n_c}{n} \sum_{i=1}^{n} w_{ij} = n_c \overline{w}_j,
\end{aligned}$$

and

$$\begin{aligned}
\text{Var}[Y_{c,j}] &= p_c \sum_{q=1}^{Q} m_q(\beta_{qj}^2 + (1 - p_c)\alpha_{qj}^2) \\
&\approx p_c \sum_{d=1}^{C} n_d(S_{\overline{w}_{d,j}}^2 + (1 - p_c)\overline{w}_{d,j}^2),
\end{aligned}$$

where

$$S_{\overline{w}_{d,j}}^2 = \frac{1}{n_d - 1} \sum_{i \in J_d} (w_{ij} - \overline{w}_{d,j})^2.$$

Using these expressions for the mean and variance of $Y_{c,j}$, the Wald statistic for the $c$th class and $j$th cluster is

$$\hat{z} = \frac{y_{c,j} - n_c \overline{w}_j}{\sqrt{p_c \sum_{d=1}^{C} n_d \big( S^2_{\overline{w}_{d,j}} + (1 - p_c) \overline{w}^2_{d,j} \big)}}, \tag{4}$$

and the null hypothesis is rejected if $P(Z \geq \hat{z}) \leq \alpha$.

## 4 Application to Remote Sensing

One way to evaluate the above proposed cluster enrichment approach is to use it in conjunction with a classification algorithm that uses soft clusters as a basis for class predictions. Clusters that pass the test (the null hypothesis is rejected) are used as the basis for classification, and clusters that fail the test (the null hypothesis is not rejected) are not used in classification. Some methods refine these latter clusters, iteratively, to arrive at better clusters. Although our enrichment strategy applies to any algorithm that outputs (and/or refines) soft clusters, we demonstrate its use with CIGSR, the continuous iterative guided spectral class rejection (CIGSCR) classification method that is popular in remote sensing [8], [9]. CIGSCR provides an example of how soft cluster evaluation can be used in the classification of remotely sensed images. We hasten to add that our choice of CIGSR is driven by our focus on remote sensing applications and that other clustering algorithms (that output soft clusters) and classification algorithms (that use soft clusters) can be readily plugged into our framework.

### *4.1 CIGSCR*

CIGSCR is a classification method used in remote sensing to assign a label to each pixel or object in a remotely sensed image using a small set of labeled pixels/objects within the image. CIGSCR uses clustering to generate a classification model $p(c_i|x)$ where $x$ is a multivariate sample to be classified and $c_i$, $i = 1, \ldots, C$, is the $i$th class where there are $C$ classes in the classification scheme. CIGSCR uses clustering to estimate $p(k_j|x)$ in the expression

$$p(c_i|x) = \sum_{j=1}^{K} p(c_i, k_j|x) = \sum_{j=1}^{K} p(c_i|k_j, x) p(k_j|x), \tag{5}$$

where $k_j$, $j = 1, \ldots, K$, is the $j$th cluster out of $K$ total clusters. CIGSCR also uses the clusters to train a decision rule using Bayes' theorem [10]

$$p(k_j|x) = \frac{p(x|k_j)p(k_j)}{\sum_{i=1}^{K} p(x|k_i)p(k_i)}. \tag{6}$$

The prior probabilities of the clusters $p(k_j)$ are assumed to be equal. While different soft clustering algorithms could be used in this algorithm, for simplicity, fuzzy $k$-means is used for this work.

CIGSCR uses labeled data to locate clusters that correspond to classes in a given classification scheme. CIGSCR requires a labeled set of training data comprised of individual samples and corresponding class labels. Rather than using the labeled data to train a decision rule directly, the entire image is clustered, thereby capturing the inherent structure of all the data and not just the labeled samples. The clusters represent spectral classes, and in remote sensing, each spectral class ideally corresponds to exactly one class in the final classification scheme. Once clusters are generated, each cluster must be associated with one class or rejected as impure. Impure clusters are rejected and can be further refined in the iterative part of the algorithm. The test for cluster purity is the cluster evaluation test presented earlier and is performed using the labeled training set.

In CIGSCR, the iterative cluster refinement takes place by selecting a target cluster in the set of existing clusters and then creating a new cluster using information contained in the target cluster. The target cluster is selected in one of two ways. First, if a class is not represented by an associated cluster, a cluster that contains the best information for that particular class is selected. If the $c$th class is not represented in the associated clusters, the cluster that is closest to being associated with the $c$th class is used to generate a new cluster. The "closest" cluster is determined to be the cluster with the highest ratio of the average membership the $c$th class to the average membership of the majority class. If all classes are represented by associated clusters, if at least one cluster failed the association significance test, the cluster with the lowest value of $\hat{z}$ is selected for further refinement. This is accomplished by adding one new cluster using information contained in a target cluster, which effectively splits the cluster into two clusters. When using a clustering algorithm based on objective function (2), adding a new cluster guarantees a smaller function value when $p = 2$. The new cluster mean is determined using

$$U^{(K+1)} = \frac{\sum_{i \in \phi^{-1}(c_k)} w_{ik} x^{(i)}}{\sum_{i \in \phi^{-1}(c_k)} w_{ik}}, \tag{7}$$

where the target cluster is the $k$th cluster, the class of interest is the $c_k$th class (in the $k$th cluster), and recall that $\phi^{-1}(c)$ is the index set of labeled samples whose label is $c$.

Once the iterative clustering is complete, one or more classifications is performed. The first classification is called the iterative stacked (IS) classification and is the result of using cluster assignments to directly produce a classification. The IS assignment for a pixel $x = x^{(m)}$ using (5) is a vector IS(x) with $i$th component

$$p(c_i|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x), \tag{8}$$

where $p(k_j|x)$ is estimated using $w_{mj}$ and

$$p(c_i|k_j, x) = \begin{cases} 1, & \text{if } k_j \text{ is labeled } c_i, \\ 0, & \text{otherwise.} \end{cases}$$

The second possible classification, the decision rule (DR) classification, uses the associated clusters to form a decision rule. Recall in (6) that

$$p(k_j|x) = \frac{p(x|k_j)}{\sum_{i=1}^{K} p(x|k_i)}$$

when all the $p(k_j)$ are equal. Traditionally, the maximum likelihood decision rule, assuming a multivariate normal distribution

$$p(x|k_j) = 2\pi^{-B/2}|\Sigma_j|^{-1/2}e^{-\frac{1}{2}(x-U^{(j)})^T\Sigma_j^{-1}(x-U^{(j)})},$$

is used where $\Sigma_j$ is the covariance matrix of the $j$th cluster [11]. The DR classification function is a vector DR(x) with the $i$th component

$$p(c_i|x) = \sum_{j=1}^{K} p(c_i|k_j, x)p(k_j|x)$$
$$= \frac{\displaystyle\sum_{j=1}^{K} p(c_i|k_j, x)\left[\frac{2e^{-\frac{1}{2}(x-U^{(j)})^T\Sigma_j^{-1}(x-U^{(j)})}}{\pi^{B/2}|\Sigma_j|^{1/2}}\right]}{\displaystyle\sum_{l=1}^{K}\left[\frac{2e^{-\frac{1}{2}(x-U^{(l)})^T\Sigma_l^{-1}(x-U^{(l)})}}{\pi^{B/2}|\Sigma_l|^{1/2}}\right]}. \tag{9}$$

See [8], [9] for more details on the CIGSCR algorithm.

### *4.2 Test Data*

The first dataset used to obtain experimental results for IGSCR and CIGSCR is a mosaicked Landsat Enhanced Thematic Mapper Plus (ETM+) satellite image taken from Landsat Worldwide Reference System

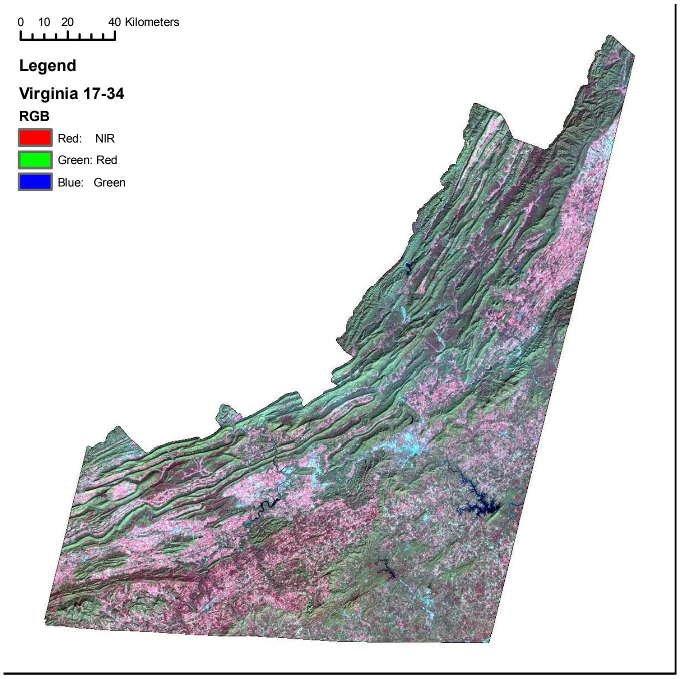

Fig. 2. Landsat ETM+ path 17/row 34 over Virginia, USA.

(WRS) path 17, row 34, located in Virginia, USA, shown in Fig. 2. This image, hereafter referred to as VA1734, was acquired on November 2, 2003 and consists largely of forested, mountainous regions, and a few developed regions that are predominantly light blue and light pink in Fig. 2. Fig. 2 contains a three color representation of VA1734 where the red color band in Fig. 2 corresponds to the near infrared wavelength in VA1734, the green color band in Fig. 2 corresponds to the red wavelength in VA1734, and the blue color band in Fig. 2 corresponds to the green wavelength in VA1734.

The training data for this image was created by the interpretation of point locations from a systematic, hexagonal grid over Virginia Base Mapping Program (VBMP) true color digital orthophotographs. A two class classification was performed (forest/nonforest), and classification parameters and results are given in Table 1 (DR classification) and Table 2 (IS classification). Fig. 3 is a DR classification image of this study area using 10 initial clusters.

Validation data in the form of point locations at the center of USDA Forest Service Forest Inventory and Analysis (FIA) ground plots were used to assess the accuracy of this classification. Since these validation data are typically used to evaluate crisp classifications, only homogeneous FIA plots were used (either 100 percent forest or nonforest), and these plots were obtained between 1997 and 2001. Accuracy was assessed based on an error matrix where classification results for specific points (not included in the training data set) are compared against known class values.

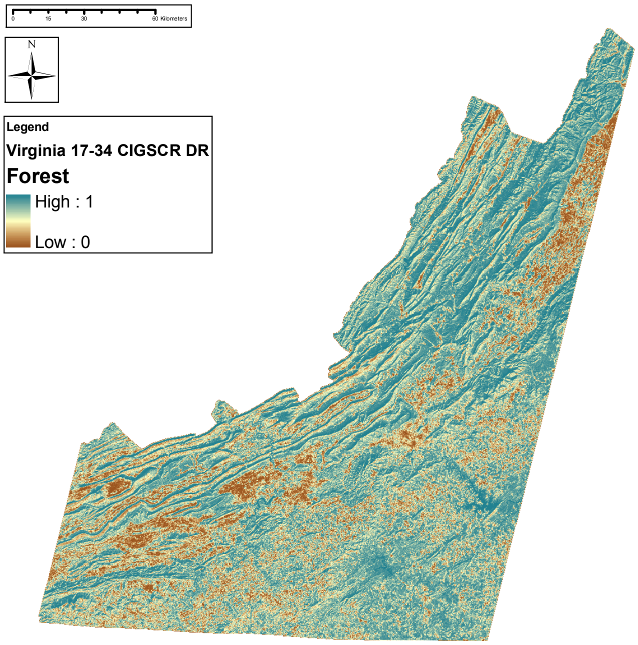

Fig. 3. CIGSCR DR classification using 10 initial clusters.

The second dataset used to obtain experimental results for IGSCR and CIGSCR is a hyperspectral image of the Appomattox Buckingham State Forest in Virginia, USA. The AVIRIS 224-band, low-altitude flight lines were acquired in the winter of 1999 and ranged from approximately 400-2500nm (10nm spectral resolution) with 3.4m spatial resolution [12]. The AVIRIS data were geometrically and radiometrically corrected (to level 1B at-sensor radiance, units of microwatts per square centimeter per nanometer per steradian) by the Jet Propulsion Laboratory (JPL; Pasadena, California, USA). The three flight lines used for this study were registered (8–12 control points per flight line) to an existing 0.5m orthophoto of the area. Resampling resulted in root mean square errors (RMSE) ranging between 0.23 and 0.24 pixels [12].

Training data were acquired by collecting 142 field locations [12] surrounded by homogeneous areas of single pine species (64 loblolly (*Pinus taeda*), 30 shortleaf (*Pinus echinata*), and 48 Virginia pine (*Pinus virginiana*)) with differentially corrected global positioning system (GPS) coordinates. These locations were used in a region growing algorithm to obtain a sufficient number of points for training and validation, and nonpine training data were acquired using knowledge of the area and maps of known stands in the region. The image (shown in Fig. 4 and hereafter referred to as ABSF) contains various tree stands that include the three species of pines listed above, hardwoods, and mixed (evergreens and hardwoods). 400 points were randomly selected to serve as validation data for these four classes (loblolly, shortleaf, and Virginia pines, and nonpine). The IS classification of this image is shown in Fig. 5.

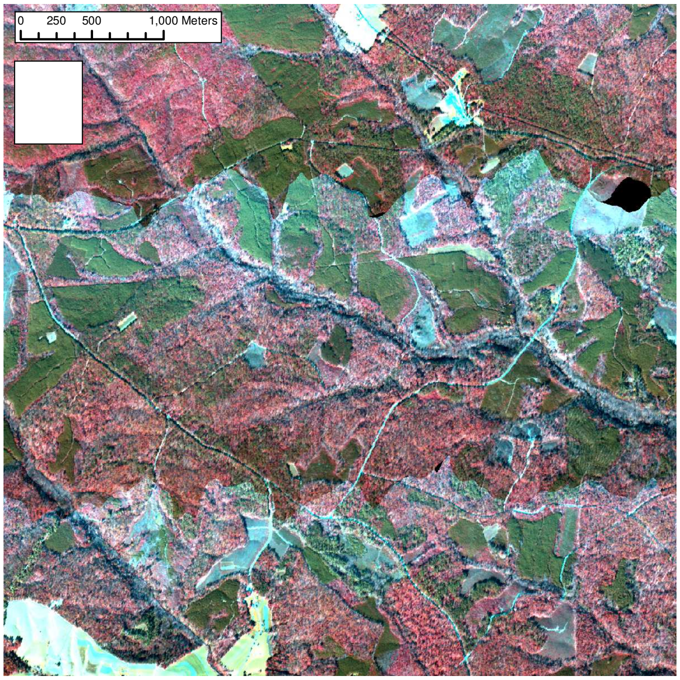

Fig. 4. AVIRIS image (three flight lines) taken over Appomattox Buckingham State Forest in Virginia, USA.

### *4.3 Results*

The accuracies reported in Tables 1–2 were obtained by first converting all soft classifications to hard classifications for the purpose of comparing hard classification values to hard ground truth values. The classification results reported in Tables 1–2 used 10, 15, 20, and 25 initial clusters for CIGSCR. Experimental runs of CIGSCR used $\alpha = .0001$ (values of $\hat{z}$ tend to be high for the association significance test). All reported CIGSCR classifications used test statistic (4). In practice, few classifications are different when using test statistic (3) instead of (4). Values of $\hat{z}$ are slightly smaller using (4) than (3), resulting in more potential for cluster refinement. Classification was performed using just clustering without the cluster refinement framework in CIGSCR to evaluate the effect of the combination of the association significance test and iteration in CIGSCR on classification accuracies. An asterisk (*) indicates that the classification failed because at least one class had no associated clusters.

In all but one experimental run (IS classification using 20 initial clusters), the combination of soft cluster evaluation and refinement in CIGSCR significantly improved final classification accuracies. Note that for the DR classification in Table 1, classification accuracies improved by at least 3.3%, a significant improvement. In practice, the DR classification often has been found to be the more accurate CIGSCR and

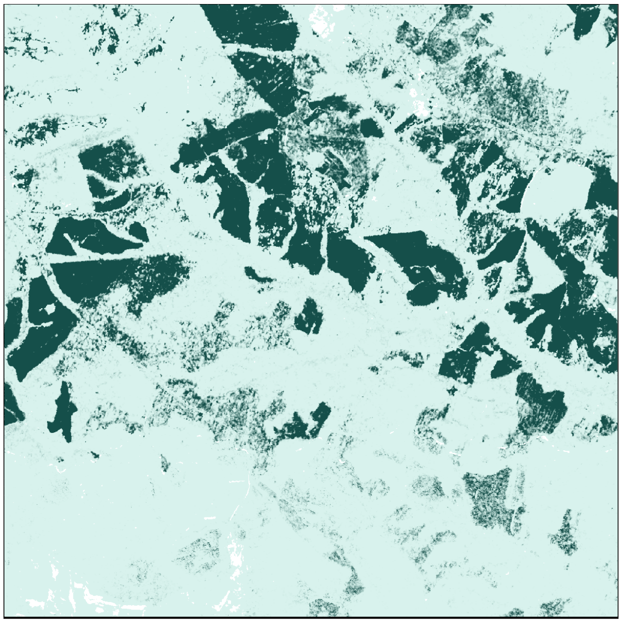

Fig. 5a. CIGSCR DR classification (loblolly pines).

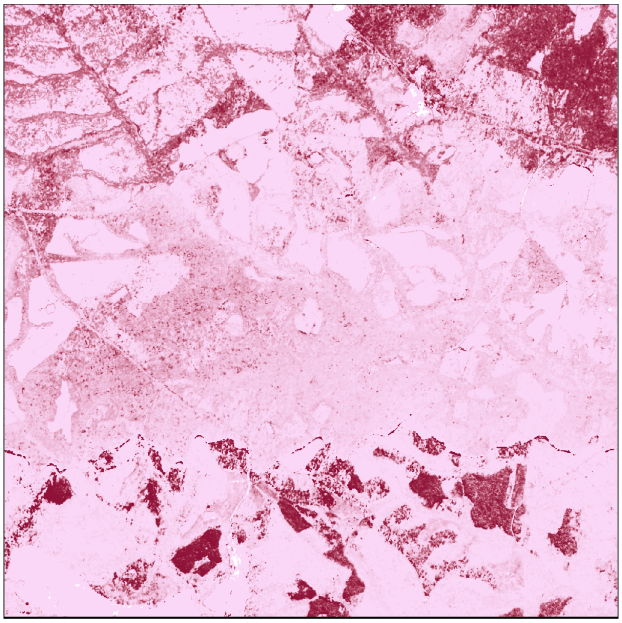

Fig. 5b. CIGSCR DR classification (shortleaf pines).

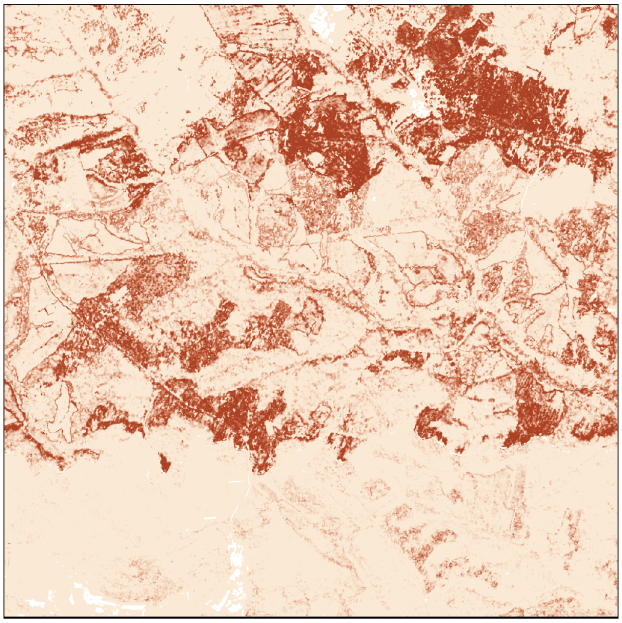

Fig. 5c. CIGSCR DR classification (Virginia pines).

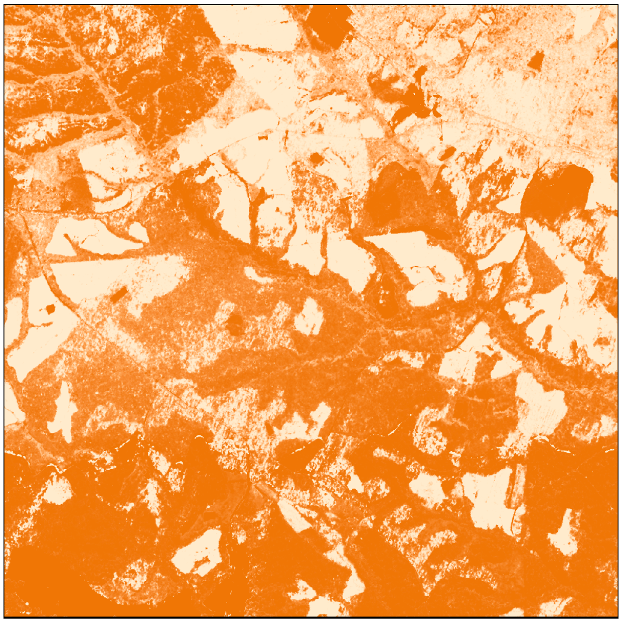

Fig. 5d. CIGSCR DR classification (nonpine).

IGSCR (the predecessor to CIGSCR that uses hard clustering) classification ([13], [14], [8], [9]). In some cases, such as when 10 initial clusters are used, the classification accuracy improvement is dramatic **(over 16%)**. This evaluation may be especially critical in cases such as this where the initial clustering does not

Table 1

VA1734 classification accuracies.

| no. init. | no. clusters | DR |  | IS |  |
|---|---|---|---|---|---|
| clusters | produced, associated | CIGSCR | clustering | CIGSCR | clustering |
| 10 | 15,13 | 88.74 | 72.26 | 83.63 | 72.26 |
| 15 | 20,16 | 80.50 | 73.72 | 76.96 | 72.99 |
| 20 | 25,21 | 79.87 | 76.54 | 75.60 | 76.85 |
| 25 | 30,25 | 81.44 | 77.58 | 78.52 | 76.75 |

Table 2

ABSF classification accuracies.

| no. init. | no. clusters | DR |  | IS |  |
|---|---|---|---|---|---|
| clusters | produced, associated | CIGSCR | clustering | CIGSCR | clustering |
| 10 | 15,15 | 47.50 | * | 51.75 | * |
| 15 | 20,19 | 62.50 | * | 51.00 | * |
| 20 | 25,24 | 66.75 | * | 51.00 | * |
| 25 | 30,29 | 63.00 | * | 51.00 | * |

produce an accurate classification, likely because the clusters do not conform to a “cluster assumption.”

The asterisks in Table 2 indicate that simply using clustering for classification failed because one or more classes was not represented by a cluster. This illustrates the fundamental problem with using clusters for classification without evaluation and/or refinement—there is no guarantee that a set of clusters will match up with a classification scheme. In this case, there are classes in the scheme that do not have a particularly strong presence in any clusters. This classification is a result of both adding new clusters that will be representative of the missing classes and performing the cluster evaluation and refinement described above. The classification results are less accurate than those for the VA1734 dataset, but classifying ABSF (a noisier image than VA1734) into four classes, three of which have some overlap between spectral classes, is a much harder problem that classifying VA1734 into two distinct classes. Regardless, the comparatively lower classification accuracies of the ABSF dataset are a better result than failing to produce a classification using clustering alone.

## 5 Conclusions

We have demonstrated a new cluster enrichment strategy suitable for soft clusters and applied it in the context of a remote sensing classification algorithm (CIGSCR). Using the cluster evaluation presented here, CIGSCR produced significantly better classifications (measured by classification accuracy) than clustering without enrichment evaluation and refinement. Since few soft cluster enrichment methods exist, we argue that our framework contributes a key methodology for clustering and cluster evaluation research.

## References

[1] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[2] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, “Constrained $K$-means clustering with background knowledge,” *Proc. of the 18th International Conference on Machine Learning (ICML ’01)*, pp. 577-584, 2001.

[3] N. Tishby, F.C. Pereira, and W. Bialek, “The information bottleneck method,” *Proc. 37th annual Allerton Conference on Communication, Control, and Computing*, pp. 368-377, 1999.

[4] W.J. Ewens and G.R. Grant, *Statistical Methods in Bioinformatics*, Springer, 2001.

[5] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L.M. Le Cam and J. Neyman, vol. 1, University of California Press, Berkeley, CA, pp. 281-297, 1967.

[6] J. Bezdek, “Fuzzy mathematics in pattern classification,” PhD Thesis, Cornell University, Ithaca, NY, 1974.

[7] 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.

[8] R.D. Phillips, L.T. Watson, R.H. Wynne, and N. Ramakrishnan, “Continuous Iterative Guided Spectral Class Rejection Classification Algorithm: Part 1,” Technical Report TR-09-09, Computer Science, Virginia Tech, Blacksburg, VA, USA, 2009.

[9] R.D. Phillips, L.T. Watson, R.H. Wynne, and N. Ramakrishnan, “Continuous Iterative Guided Spectral Class Rejection Classification Algorithm: Part 2,” Technical Report TR-09-10, Computer Science, Virginia Tech, Blacksburg, VA, USA, 2009.

[10] B.V. Gnedenko, *Theory of Probability (sixth ed.)*, Gordan and Breach Science Publishers, The Netherlands, 1997.

[11] J.A. Richards and X. Jia, *Remote Sensing Digital Image Analysis (third ed.)*, Springer-Verlag, Berlin, 1999.

[12] 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.

[13] 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.

[14] 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.