1. INTRODUCTION

A common task in analyzing remote sensing imagery is supervised classification, where the objective is to construct a classifier based on few labeled training samples and then to assign a label (e.g., forest, water, urban) to each pixel (vector, whose elements are spectral measurements) in the entire image. The commonly used maximum likelihood classifier (MLC) has two well known limitations. First, it works well if the land cover classes are spectrally separable. In reality, the classes under investigation are often spectrally overlapping as the reflectance recorded by remote sensing satellites for many of these thematic classes is dependent on several extraneous factors like terrain, soil type, moisture content, acquisition time, atmospheric conditions, etc. The usefulness of ancillary data for improving classification accuracy is well known, but there is no convenient multivariate statistical tool for modeling this multi-source data (i.e., images and ancillary geo-spatial data together). Previous studies [1], [2] have focused on incorporating ancillary information into the MLC (typically via a priori term).

Second, MLC uses maximum likelihood estimation (MLE) technique for estimating class probability distribution parameters which requires large amounts of accurate training data. Collecting ground truth data for large number of samples is very difficult. Apart from time and cost considerations, in many emergency situations like forest fires, land slides, floods, it is impossible to collect accurate training samples. As a result, supervised learning is often carried out with small number of training samples, which leads to large variance in parameter estimates and thus higher classification error rates. Several approaches can be also be found in the literature that specifically deal with small sample size problems in supervised learning [3, 4]. These methods are aimed at designing appropriate classifiers, feature selection, and parameter estimation so that classification error rates can be minimized while working with small sample sizes. However, only recently attempts have been made to incorporate unlabeled samples in supervised learning, which gave raise to new breed of techniques, collectively known as semi-supervised learning methods. Well-known studies in this area include, but not limited to [5, 6]. The common thread between many of these methods is the Expectation Maximization (EM) [7] algorithm. Many of the semi-supervised learning methods pose class labels as the missing data and use the EM algorithm to improve initial (either guessed or estimated from small labeled samples) parameter estimates. Although several previous studies showed that adding unlabeled training samples improves overall classification accuracy, semi-supervised learning techniques also suffer from the same limitations when dealing with multsource geospatial data for the lack of convenient GMM.

2. OUR APPROACH

In this paper, we provide a new hybrid semi-supervised learning method based on a mixture of discrete and continuous distributions. In typical semi-supervised approach, the population is assumed to be generated by a mixture of multivariate normal distributions for continuous attributes (e.g., remote sensing images), or mixture of multinomial distributions for categorical attributes (e.g., text documents, ancillary geospatial data such as soil types, upland and lowlands). However, for multsource data we defined a mixture model that admits both continuous and categorical attributes. To reduce the complexity we used a knowledge based approach to stratify the geographic region into three broad categories, viz., uplands, lowlands and developed area. The main objective for this stratification is to split the geographic region into different spatial units where each spatial unit contains classes that are easily discriminable. Finally we used expectation maximization algorithm to estimate the model parameters.
parameters, where our model consists of six continuous attributes (corresponding to six channels in the ETM image) and a categorical attribute (generated using a knowledge based classification algorithm). We have conducted several experiments to evaluate the usefulness of our method in thematic classification of multisource geospatial datasets. We will present detailed description of the algorithm in full paper. We now briefly present the results in the following section.

![Bivariate Density Plot: Band 4 and 6](image)

(a) ML  
(b) SSL(EM)

**Fig. 1.** Parameters Estimated using (a) ML, (b) SSL(EM).

### 3. EXPERIMENTAL RESULTS AND CONCLUSIONS

We used a spring Landsat 7 scene, taken on May 31, 2000 over Cloquet town located in Carlton County, Minnesota. We designed two different experiments to validate our hypothesis that adding ancillary geospatial datasets and unlabeled training samples improve the classification performance. The labeled training data consists of 14 plots (2 plots per class), and unlabeled training data consists of 50 plots. For both of these experiments the test dataset was fixed and consisted of 205 plots. We trained three classifiers: MLC, Semi-supervised Classifier (SSL), and Multisource Semi-supervised classifier (SSL-MS). The estimates obtained by maximum likelihood and semi-supervised approaches (using expectation maximization) are summarized (in the form of bivariate density plots) in Figure 1. The individual class accuracy and overall classification accuracies were summarized in Figure 2. This figure (table) shows the great potential of our proposed classification scheme in small sample and multisource data classification problems. Further research is needed to automatically discover the stratified units from ancillary data.

### 4. REFERENCES


