Available online at www.sciencedirect.com
SCIENCE DIRECT
Remote Sensing of Environment 96 (2005) 62–77
Remote Sensing of Environment
www.elsevier.com/locate/rse

# On the choice of spatial and categorical scale in remote sensing land cover classification

Junchang Ju[a,*], Sucharita Gopal[a], Eric D. Kolaczyk[b]

[a]Department of Geography, Boston University, MA 02215, United States
[b]Department of Mathematics and Statistics, Boston University, MA 02215, United States

Received 22 July 2004; received in revised form 12 January 2005; accepted 16 January 2005

**Abstract**

Our interest in this paper is on the choice of spatial and categorical scale, and their interaction, in creating classifications of land cover from remotely sensed measurements. We note that in discussing categorical scale, the concept of spatial scale naturally arises, and in discussing spatial scale, the issue of aggregation of measurements must be considered. Therefore, and working towards an ultimate goal of producing multiscale, multigranular characterizations of land cover, we address here successively and in a cumulative fashion the topics of (1) aggregation of measurements across multiple scales, (2) adaptive choice of spatial scale, and (3) adaptive choice of categorical scale jointly with spatial scale. We show that the use of statistical finite mixture models with groups of original pixel-scale measurements, at successive spatial scales, offers improved pixel-wise classification accuracy as compared to the commonly used technique of label aggregation. We then show how a statistical model selection strategy may be used with the finite mixture models to provide a data-adaptive choice of spatial scale, varying by location (i.e., multiscale), from which classifications at least as accurate as those of any single spatial scale may be achieved. Finally, we extend this paradigm to allow for jointly adaptive selection of spatial and categorical scale. Our emphasis throughout is on the empirical quantification of the role of the various elements above, and a comparison of their performance with standard methods, using various artificial landscapes. The methods proposed in this paper should be useful for a variety of scale-related land cover classification tasks.


*Keywords:* Spatial scale; Categorical hierarchy; Multiscale; Multigranular; Finite mixture model; Complexity-penalized maximum likelihood; Translation invariance

## 1. Introduction

### 1.1. '*Scale*' *in land cover classification*

Remote sensing land cover classification is essential for a variety of tasks, such as land resource management or the retrieval of biophysical parameters of land cover, like leaf area index, albedo and surface roughness. Choice of spatial scale and categorical scale are two of the central issues in remote sensing land cover classification. The use of the term 'spatial scale' in this context is mostly concerned with the manner in which the image information content is determined by its spatial resolution and with the way that spatial resolution is handled in the image processing stages, with respect to land cover patches of varying size and shape. The term 'categorical scale' (also called 'categorical resolution' by other authors) refers to the level of detail in the categories used in classification (e.g., Franklin & Woodcock, 1997), and is often discussed with reference to a categorical hierarchy. For example, a land cover category such as *forest* in a regional classification system has a coarser categorical scale than a *conifer* or *hardwood* category. Numerous studies have examined the effect of spatial scale on classification accuracy (e.g., Irons et al., 1985; Marceau et al., 1994b; Markham & Townshend, 1981; Raptis et al.,

* Corresponding author. Tel.: +1 617 3580210; fax: +1 617 3538399.
*E-mail address:* jju@bu.edu (J. Ju).

0034-4257/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2005.01.016

2003). Categorical scale has been, in comparison, examined to a lesser extent.

The dominant paradigm in land cover classification, for most of its history, has been that of classification at a single spatial scale and a single categorical scale (i.e., with category labels from only one level of a potential categorical hierarchy). However, increasingly the nature of current applications demands that land cover classification be conducted at multiple spatial scales and, potentially, at multiple categorical scales. That is, there is growing need for methods of land cover classification that are both 'multiscale'—in the sense of being able to incorporate information across multiple spatial scales in constructing a map–and also 'multigranular'—in the sense that class labels of different granularity, or categorical scale, similarly co-exist in the same map.

Earlier studies on the effects of scale on remote sensing land cover classification sought a single 'optimal' spatial scale for classification (e.g., Irons et al., 1985; Markham & Townshend, 1981). However, it was shown that in fact a single 'optimal' scale cannot accurately represent all classes in a complex scene, due to the contrasting sizes, shapes, and internal variation of the patches for different land cover classes (e.g., Marceau et al., 1994a, 1994b; Raptis et al., 2003). As a result, recent research has emphasized the development of multiscale classification tools. The wide variety of methods proposed for 'multiscale' classification in remote sensing, and in the imaging and computer vision communities more widely, are of course too numerous to be cited here in any significant detail. Instead, it is more useful for our purposes to remark that, broadly speaking, at the heart of these methods, their multiscale nature tends to rest on the adoption of either of two paradigms: (i) aggregation, or (ii) wavelet transformations.

The traditional practice of aggregating measurement information, from an initial fine scale to successively coarser scales, is arguably at the center of a large fraction of proposed methods of multiscale analysis in remote sensing. This practice includes two principle variations. The first involves the explicit aggregation of spectral measurements, through simple or weighted averaging on moving windows. The second involves the implicit aggregation of the information in these measurements by instead aggregating to coarser spatial scales the class labels assigned by a classifier of choice to pixels at the finest spatial scale. We refer to these two variations as spectral and label aggregation, respectively. The approach of label aggregation is more widely used (e.g., Moody & Woodcock, 1994; Wu & David, 2002), likely because spectral aggregation requires that a classifier be trained at each scale, a non-trivial task.

The units of aggregation in these approaches typically are the cells in successively coarser uniform grids over the image region, which can induce a number of problems. For example, the resulting land cover maps themselves can end up looking blocky. In addition, since the block boundaries are artificially configured and most likely do not match patch boundaries in the scene, artifacts can result (Barnsley et al., 1997). Moreover, in the conventional method of label aggregation, error in the original per-pixel classification, whose accuracy is typically around 70%, is inherited and integrated into the classifications at coarser scales (Foody, 2002; Gurney, 1983). More sophisticated implementations of the aggregation principle make use of texture and context information at multiple spatial resolutions (e.g., Gong & Howarth, 1992), to exploit spatial variation in the image, but still suffer from limitations imposed by the underlying system of grids. A particularly simple, but widely used, ad hoc way of contextual classification is post-classification relabeling with an odd-sized moving window filter (Gurney, 1983).

Complementing the methods based on aggregation are image classification methods based on the mathematical tool of wavelet analysis (e.g., Bouman & Shapiro, 1994; Choi & Baraniuk, 2001; Sveinsson et al., 1998). In such approaches a transformation(s) of the original measurements is computed, by convolving the measurements with a scale-indexed sequence of wavelet-based filters. These filters have the property of enhancing the information in the data local to a range of combinations of spatial scale and position. They also are designed in such a way that, upon the application of statistical analysis procedures and classifiers, the resulting maps in the original image space tend to suffer notably less from the types of artifacts experienced by aggregation methods. However, because the use of wavelet transformations in more than two dimensions quickly becomes prohibitive from a computational perspective, it is not typically the case that multi-spectral remote sensing measurements are analyzed as a unit. Rather, two-dimensional wavelet transforms are applied either to some transformation of the spectra into scalar quantities or to the data in each spectral band separately.

In addition to the explicit use of aggregation or wavelets, multiple scales can arise in land cover classification methods implicitly as well, such as when incorporating certain image segmentation techniques. Segmentation techniques based on region-growing principles, for example, 'grow' candidate landscape patches in a scale-adaptive manner. Such methods have been found in many cases to be superior to simple per-pixel approaches to classification (e.g., Barlow et al., 2003; Tilton, 1998). However, when the within-class spectral variation is high, the use of certain segmentation techniques has been found to be less appropriate, as it does not increase classification accuracy (e.g., Stuckens et al., 2000). Image segmentation techniques based on quad-trees, a popular data structure, also share the sensitivity of the aggregation techniques due to the use of block-like spatial regions.

On a side note, we mention that of course there is also recent work on the 'fusion' of data for the same spatial region from instruments of different spatial resolutions (e.g., Cohen et al., 2003). However, since these methods often incorporate the same sort of multiscale structures described

for single-instrument measurements above, combined in some sort of coherent fashion, we shall focus our attention here on the single-instrument case.

Regardless of the manner in which multiple spatial scales are incorporated into a land cover classification framework, there remains the important issue of choosing an appropriate categorical scale to be used in assigning labels to pixels. This issue arises through the simple fact that, as spatial scale increases beyond certain points (relative to the patch sizes, positions, and shapes in the underlying landscape), labels of increasingly coarser categorical scale become viable, due to correspondingly increased levels of mixing. In fact, most multiscale approaches use labels from only one level of categorical scale, i.e., they are categorically monoscale (or 'monogranular'). A few studies that do incorporate multiple scales of class labels with multiple spatial scales do so simply by lumping classes into more general classes as the spatial scale gets coarser (e.g., Marceau et al., 1994b). However, such lumping of categories on a global basis can lead to considerable loss of categorical information for a large area where only one specific class is present. This practice is typical in the use of a hierarchical land cover/use classification system like that described in Anderson et al. (1976), in which the categorical scale and spatial scale are not explicitly connected.

Recently, more sophisticated approaches have sought to link the use of coarser class labels adaptively with certain relevant aspects of the coarser regions to which they might be applied. For example, the method of Tilton (1998) incorporates the labeling of regions directly within the overall process of growing regions, in extending classical region-growing segmentation methods. Alternatively, Wu and David (2002) have developed a systems-level paradigm for modeling complex ecological systems with hierarchical structures, based on the ideas of 'hierarchical patch dynamics' and 'scaling ladders', which has been successfully adapted to the study of heterogeneous landscapes.

### *1.2. Overview of the proposed framework*

The ultimate goal of this paper is to present and evaluate the elements of a new statistical framework for multiscale, multigranular land cover classification. Spatial scale will be represented in an efficient manner through the use of a quad-tree data structure, in analogy to conventional aggregation methods and many methods of image segmentation. A translation invariant (TI) implementation, analogous to those used commonly in the wavelet literature (e.g., Coifman & Donoho, 1995), helps overcome the block effects that would otherwise accompany the quad-trees. To this basic data structure we add the use of statistical finite mixture models, through which we implement a form of implicit aggregation using maximum likelihood procedures. Finally, categorical scale is incorporated through the presence of a user-defined categorical hierarchy, after the hierarchical scene model proposed by Strahler et al. (1986). The use of such hierarchies in this paper emphasizes spatial nesting, as in Woodcock and Harward (1992), in that a patch is labeled as a general class only when at a certain spatial scale it is judged to contain sufficient information likely to correspond to patches of more than one specific class. Overall, classification maps are produced in our framework through the implementation of formal statistical model selection procedures, and a simple post-processing step that allows for the enhancement of user-specified preferences.

The various modeling components and statistical tools used in this paper are borrowed from the multiscale, multigranular image segmentation framework introduced by the authors in Kolaczyk et al. (submitted for publication). There a rigorous mathematical derivation and analysis of the framework can be found, including a result showing that the underlying model selection strategy is capable of producing characterizations of an underlying landscape that are nearly optimal in the sense of a certain well-defined decision-theoretic criterion. The primary contribution of the current paper is to provide a detailed empirical assessment of the degree to which the resulting framework and its individual components can contribute to improved remote sensing land cover classification on a practical level. We proceed through the use of three separate numerical simulation experiments, which allows us to compare results closely against an actual 'truth'. Our model is benchmarked against a conventional label aggregation approach and modifications thereof.

More specifically, we present our work cumulatively in three stages, addressing first the topic of aggregation, then the topic of adaptive choice of spatial scale, and lastly, the topic of adaptive choice of categorical scale jointly with spatial scale. In Section 2, we show that statistical finite mixture models can be used with groups of un-aggregated spectral measurements, at each of a range of individual scales, to produce gains in classification accuracy over an analogous aggregation-based approach. We then demonstrate in Section 3 how the principle of statistical model selection may be used with the finite mixture models to provide a method that makes locally adaptive choices of spatial scale, and we quantify the effect of this capability on classification accuracy. Lastly, we extend this paradigm in Section 4 to allow for jointly adaptive selection of spatial and categorical scale, yielding our overall multiscale, multigranular framework, and we similarly quantify the resulting effects. In Section 5, we finish with additional discussion of these results.

## 2. First experiment: spectral measurements vs. class labels in aggregation

As the first of our three experiments, this section presents an aggregation approach that uses statistical finite mixture models in conjunction with original pixel-scale spectral

measurements at successively coarser scales and compares this approach with conventional aggregation of pixel-wise labels.

### 2.1. Methodology

Throughout this paper we represent the notion of consecutive spatial scales through the use of windows of different sizes, starting with the original pixels and then moving up to windows of size $2\times2$, $4\times4$, $8\times8$, ..., doubling with each change of scale and nesting windows at one scale within those at coarser scales. The nesting of scales is convenient algorithmically for the material that follows in the next two sections. In this section alone, it also will be of interest to refine this set of scales to use a more continuous progression of scales, such as through the incorporation of $3\times3$, $5\times5$, $6\times6$, and $7\times7$, and so on.

Now at each scale, we consider two methods by which to estimate the fraction of each land cover class within a given window: the conventional method of counting labels resulting from pixel-scale classification (i.e., label aggregation), and a method based on statistical finite mixture models. For the former, the original spectral measurements are used directly only for the classification at the finest spatial scale and the resulting labels are then used as input at each successively coarser scale. In contrast, the mixture model approach uses the measurements directly at every scale, which intuitively should result in a gain in information.

Statistical finite mixture models dictate the measurements of some population as having in fact come from a finite number of some sub-populations. Accordingly, the population probability density can be modeled as the weighted sum of the probability densities of the sub-populations (components), i.e.,

$$f(x) = \sum_{i=1}^{g} \pi_i f_i(x), \tag{1}$$

where $f(x)$ is the mixture density for the measurement $x$, $g$ is the number of components, $f_i(x)$ is the density for the $i$-th component, and the weights $\pi_i$ are non-negative quantities that sum to one. See McLachlan and Peel (2000), for example.

Finite mixture models can be useful in the context of land cover classification because at coarser scales a window often contains a number of land cover categories that can be identified at finer spatial scales. Hence, as a group, the spectral measurements $x$ corresponding to the pixels in the window can be modeled as deriving from a mixture of densities like that in Eq. (1), where $g$ is number of classes, the component densities $f_i$ correspond to the spectral distributions for each possible individual land cover category and the weights $\{\pi_i\}$ can be interpreted as the fraction of each category in the window. For areas with only one class we would expect the weight for that class to be close to one and the weights for all other classes to be close to zero. It is possible therefore to model the land cover across an image, at a given choice of scale, if we let the set of all possible land cover classes inform our choice of mixture components $f_i$ and the weights $\pi_i$ to potentially vary from window to window. This principle is the same as that underlying the use of 'mixture analysis' in the literature on the remote sensing sub-pixel proportion problem (e.g., Smith et al., 1990), except that here the focus is on 'supra-pixel' (i.e., above pixel scale) proportions across a succession of coarser scales. Of course, in order to use mixture models as we describe, the component probability densities $f_i$ must be known or learned beforehand. In this paper we specify them to be themselves Gaussian mixtures and fit the corresponding parameters for each component to pure-class training data using a standard EM algorithm, as described in Ju et al. (2003). Our choice to model the pure-class densities as Gaussian mixtures is driven by the fact that such mixtures have the capability of approximating complex densities arbitrarily well (e.g., pages 11–14, McLachlan & Peel, 2000). The fitting of these mixtures entails estimation of the number of Gaussian components in a class and each component's proportion, mean and covariance. Taking these inferred $f_i$ as fixed, we then analyze an image at a given scale by estimating a set of weights $\{\hat{\pi}_i\}$ for each window from the measurements for all pixels in that window, again using an EM algorithm (see McLachlan & Krishnan, 1997 for detail on EM).

#### 2.1.1. Translation invariance implementation

In performing aggregation-based analysis of remote sensing data, it is common that at a given scale the window is moved across the image in a non-overlapping fashion. The resulting representation will look 'blocky', particularly as the window size increases, due to the monoscale nature of the method. Additionally, results produced in this way lack translation invariance (TI). This limitation is illustrated in Fig. 1, where aggregation of an $8\times8$ image at scale $2\times2$ splits the pure *conifer* patch in the middle and very likely makes it disappear at this scale. The patch would be unsplit and likely preserved if shifted off the central axis. These problems stem from using a collection of windows that induce a fixed partitioning of the image, but they can be overcome with a translation invariant implementation. The basic idea is to allow for the use of an enlarged collection of windows corresponding to a fully redundant system of all possible partitionings and to then average the representations from these partitionings.

In this section a TI implementation is applied to both the label aggregation approach and the finite mixture model approach, at each scale except the coarsest scale, in the following manner. At a scale with window size $m\times m$, the image is shifted $m^2$ times, each time by $i\in\{0, 1,\ldots, m-1\}$ rows and $j\in\{0, 1,\ldots, m-1\}$ columns, and the rows and columns that are shifted out are appended to the bottom and to the right of the image in the original order (wrap

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| h | h | h | h | h | h | h | h |
| h | h | h | h | h | h | h | h |
| h | h | h | h | h | h | h | h |
| h | h | h | c | c | h | h | h |
| h | h | h | c | c | h | h | h |
| h | h | h | h | h | h | h | h |
| h | h | h | h | h | h | h | h |
| h | h | h | h | h | h | h | h |

Fig. 1. A simple illustration of lack of translation invariance in aggregation using a fixed partition. Label 'c' represents conifer class and 'h' represents hardwood class. Note the pure conifer patch is split by this fixed partition.

around). For each shifted image at a scale, the land cover fractions within all windows are estimated with both approaches, and then the columns and rows of the resulting image are shifted back to the original positions. Finally, the land cover fractions for each pixel over all $m^2$ shifts are averaged. It is obvious that TI would have no effect at the coarsest scale in a monoscale aggregation.

The redundant nature of the TI algorithm makes the averaged vector of fractions at a given scale actually available for each pixel at the original $1 \times 1$ scale. The vector contains information for a spatial neighborhood of each such pixel proportional in size to the scale of the window used. In the experiment of this section, we apply a majority rule to the vector of each pixel to obtain a final label. The 'wrapping' used in our algorithm can sometimes create artifacts in a narrow strip of pixels on the borders of the image, like many other approaches handling the border problem (e.g. Schowengerdt, 1997), but it helps enormously to improve the characterization for the interior of the image. TI implementations used throughout this paper, although the algorithms for the methods of the second and third experiments will be slightly more involved. See Coifman and Donoho (1995) for more details on TI in statistics and Sveinsson et al. (1998) for its application to de-noising in remote sensing.

As a final note on this topic, we mention that, on a scale-by-scale basis, the method resulting from combining the conventional aggregation scheme with a TI implementation has an interesting connection to the commonly used overlapping moving window filter of odd sizes (e.g., Gurney, 1983). Specifically, this combination provides land cover fractions for each pixel by averaging the fractions within all possible windows containing that pixel. This approach is equivalent to using a moving, overlapping window with a particular set of weights decaying from the central pixel. Take scales $2 \times 2$ or $3 \times 3$, for example. The method is equivalent to using a moving window of $3 \times 3$ or $5 \times 5$, with respective weights

$$1/16 \cdot \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \text{ and } 1/81 \cdot \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 4 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 4 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix}.$$

The use of weighted moving windows is common in the filtering of the original spectral image, but is rarely seen in post-classification contextual relabeling. One exception is Thomas (1980), who used a weighted $3 \times 3$ window accounting for the rectangular shape of the early Landsat MSS pixels.

### 2.1.2. Measures of accuracy

We use two measures of accuracy throughout this paper, corresponding to two different ways the vector of estimated land cover fractions for each pixel can be treated, i.e., either as input for a majority vote rule (or similar) or as output themselves. We define our first accuracy measure in relation to the former case, as the percentage of the land cover categories in the 'truth' that are correctly recovered.

Our second accuracy measure quantifies the performance of our two approaches more precisely in taking into account the actual estimated fractions of all classes. That is, we will use the mean absolute error (MAE),

$$\text{MAE} = 1/N \sum_{i=1}^{N} \sum_{j=1}^{l_c} |f_{e,i,j} - f_{t,i,j}|/l_c \ , \qquad (2)$$

where $N$ is the number of pixels in the image, $l_c$ is the number of classes in the landscape, $f_{e,i,j}$ is the estimated fraction for class $j$ at the $i$-th pixel, $f_{t,i,j}$ is the 'true' fraction for class $j$ at the same pixel (which is either 1 or 0, by the assumption of 'pure' classes at the pixel scale). This error measure indicates how close the estimated fraction vector is to the 'true' fraction on average. It potentially preserves more information than the first accuracy measure, since as long as the majority class is correctly identified for a pixel the first accuracy measure is the same regardless of the exact fractions of all land cover classes.

## 2.2. Experiments and results

A simple artificial landscape of $128 \times 128$ pixels and 9 patches with pre-determined patch boundaries was constructed, in which each patch is composed of a pure land cover class (Fig. 2(a)). The patches range from 93 to 5894 pixels in size, with a mean of 1820 pixels. Four land cover classes are present in this landscape, i.e., *conifer*, *hardwood*, *brush*, and *grass*. The spectral data used in this simulation are derived from field data of Plymouth County (which is south of Boston, Massachusetts); the section of the county included in this study is mostly rural with a State Forest and cultural features such as golf courses. Specifically, the

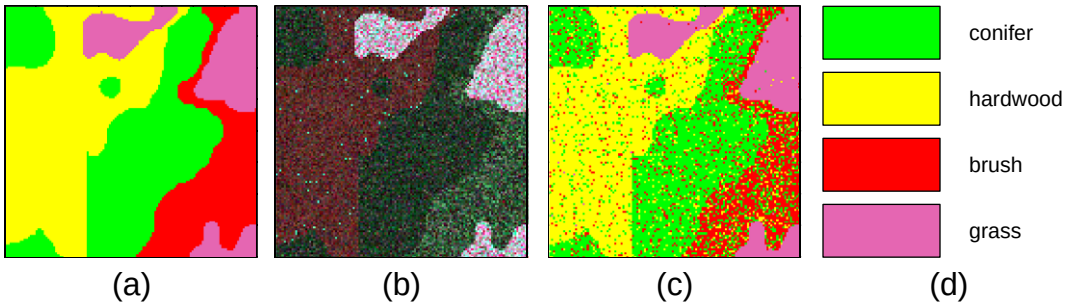


Fig. 2. Simulated landscape consisting of pure patches exclusively. (a) The 'true' class labels; (b) color composite with TM bands 432 in RGB; (c) per-pixel classification; (d) map legend. Note (b) and (c) are based on one trial.

*conifer* class consists mostly of white pine plantations, the *hardwood* class includes stands of a few oak species, the *brush* class is the growth after severe forest fires, and the *grass* class is mainly cultured grass on golf courses. Training sites for the four classes were identified in a Landsat TM image acquired on September 27, 2000, and visited in the field in the Summer of 2003. Land cover of the four classes can be considered constant during this period because although there might be inter-annual variation there was no human-induced land cover change in the State Forest. Class-specific probability densities were then fitted to the 6 reflective TM spectral bands of training data using Gaussian mixtures, in the manner described in Ju et al. (2003). Finally, simulated spectral measurements in the six bands were generated for each pixel in the artificial image by sampling from the appropriate fitted density, according to the 'true' class label of the pixel. The reason that we sample from the fitted densities, rather than using real, validated pixel data, is simply that in the latter case a simulation involving an image of the size of the one used in this experiment would require an unrealistically large number of validation sites.

The simulation was run 30 times with the underlying 'truth' (Fig. 2(a)) unchanged. For each trial, these simulated data were analyzed using the label aggregation method and the finite mixture model method described previously. Both methods require class-specific densities for the data, which were taken to be the densities previously fit to the field data. In other words, for these simulations there is no uncertainty as to the underlying densities, since the same densities are used both to generate the data and analyze it, thus placing both methods on an equal and ideal footing.

Fig. 2(b) shows a sample of simulated TM image in color composite with bands 432 in RGB. Fig. 2(c) shows a maximum likelihood classification of this image at the pixel scale using the fitted densities; the 'salt and pepper' appearance is typical. The finite mixture method, using a majority rule on the estimated weights at the $1 \times 1$ scale, would reproduce the results of this classifier, simply by definition. The label aggregation method takes this classification as input. Fig. 3 shows the land cover maps created by label aggregation and finite mixture model methods with TI for this particular image at nested spatial scales $2 \times 2$, $4 \times 4$, …, and $128 \times 128$. Visually, the two methods produced similar maps at each scale. At the coarsest scale, both methods label the entire image as the most dominant class, *hardwood*. One difference is that the 'salt and pepper' appearance at the $2 \times 2$ and $4 \times 4$ scales is more severe for label aggregation method than for the finite mixture model method.

The classification accuracy at all the spatial scales (nested scales $2 \times 2$, $4 \times 4$, …, $128 \times 128$, and additional scales $3 \times 3$, $5 \times 5$, $6 \times 6$, $7 \times 7$) for both approaches, based on 30 trials, are shown in Fig. 4(a). The mean classification accuracy at the $1 \times 1$ scale is 84%. Both the finite mixture and label aggregation methods dramatically improved accuracy in moving from the pixel scale to coarser scales, and they achieve the best accuracies, 99% and 98% respectively, both at the $3 \times 3$ scale. The finite mixture model method has slightly higher accuracy than the conventional aggregation at all spatial scales. It is not surprising that the accuracy for both approaches is poor beyond the $16 \times 16$ scale, which is too coarse to represent fine spatial details in this landscape e.g., patch boundaries. The two approaches have exactly the same, low classification accuracy at the coarsest scale where both of them identify the most dominant class (hardwood in this case) correctly, although the information at the coarsest scale may not be useful in many contexts. The models have been evaluated at scales three through seven, in addition to the spatially nested scales, to check whether the conventional approach out-performed the mixture model approach in the neighborhood of the optimal scale, but it is apparent that it did not, as displayed in Fig. 4(a). The accuracy performance of the two approaches should be viewed only in a relative sense; the specific accuracies will depend more generally on the variability in training data, size and shape of the patches. In reality it is rare to achieve an accuracy of 84% at the pixel scale and virtually impossible to obtain an accuracy as high as that around the $3 \times 3$ scale in this experiment.

At most of the scales, the two approaches showed similar classification accuracy for this particular simulation. However, this similarity does not necessarily mean that the two approaches are equivalent in their performance. Specifically, the label aggregation approach consistently has a higher error rate when measured by the mean absolute error (MAE)

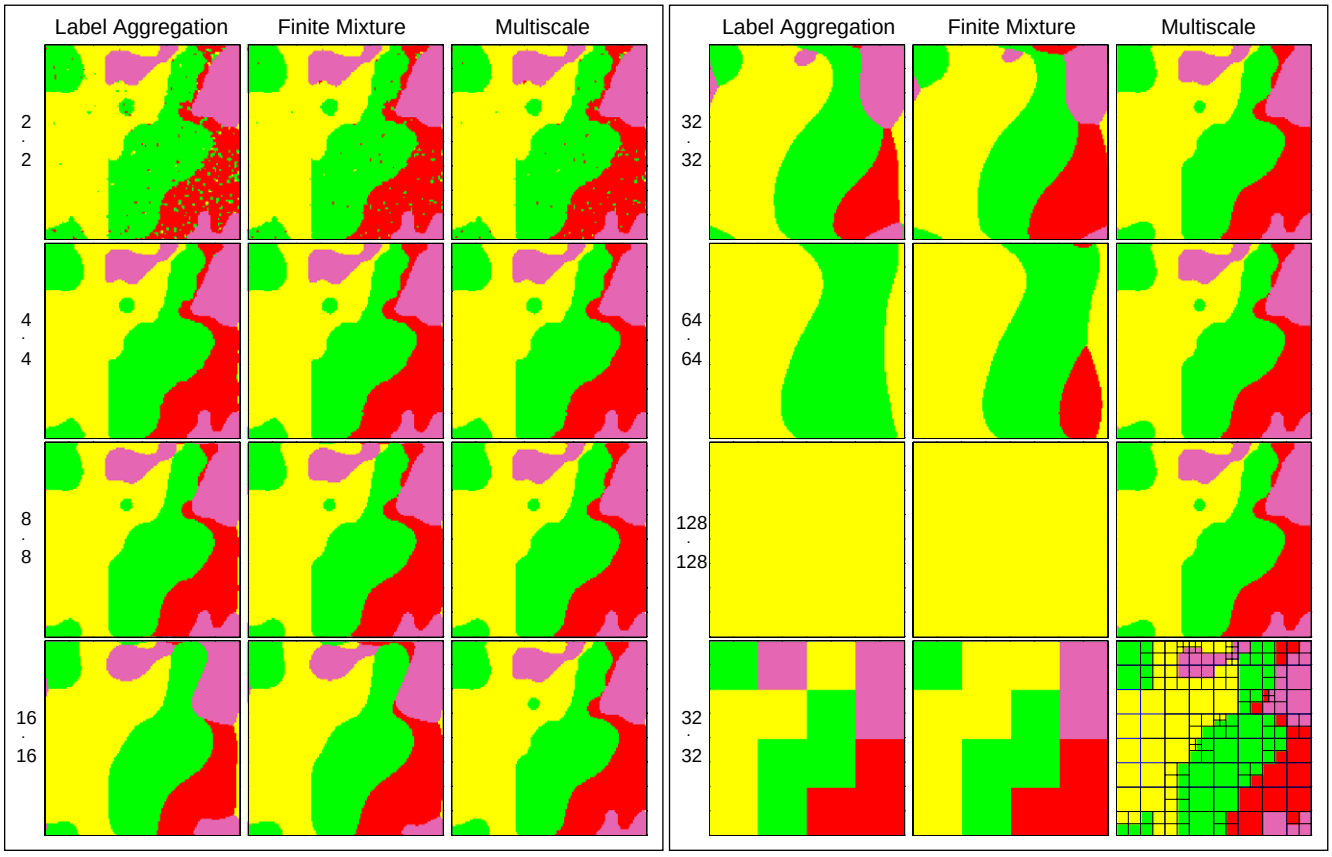


Fig. 3. The land cover maps for the landscape with pure patches exclusively, produced by label aggregation, finite mixture model, and multiscale model approaches, at the nested spatial scales $2 \times 2$, $4 \times 4$, . . . , $128 \times 128$. Note the last row in the right panel shows maps without translation invariance (TI) implementation at the $32 \times 32$ scale. See Fig. 2(d) for map legend.

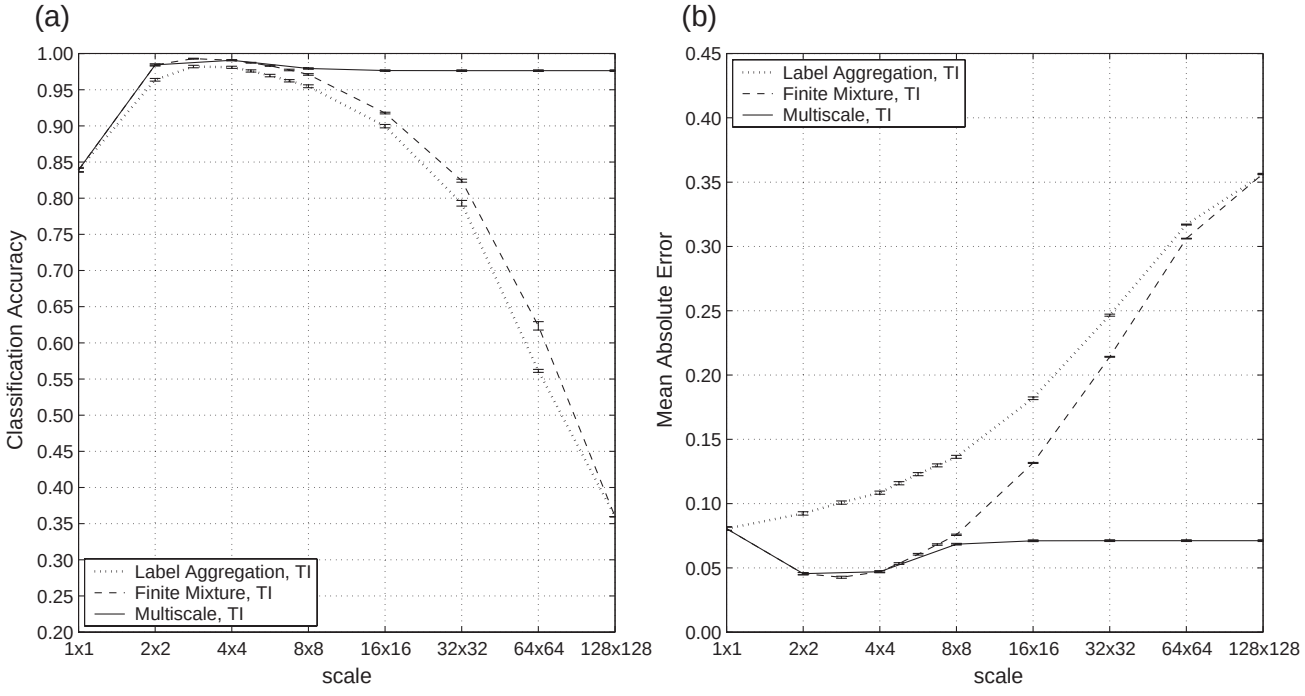


Fig. 4. Accuracy of label aggregation, finite mixture model, and multiscale model approaches for the landscape with pure patches exclusively: mean and $\pm 1$ standard deviation based on 30 trials. (a) Classification accuracy at each scale for the three methods; (b) the mean absolute error in fraction estimation by the three methods. For the multiscale method, the results are shown at nested spatial scales only, i.e., $2 \times 2$, $4 \times 4$, . . ., $128 \times 128$; for the other two methods, results at additional scales $3 \times 3$, $5 \times 5$, $6 \times 6$ and $7 \times 7$ are also shown.

in Eq. (2) and shown in Fig. 4(b). Examining this figure, we can see that at the $3 \times 3$ scale, where both approaches achieved their best accuracy, the MAE for conventional aggregation is 6% higher. The majority rule and the corresponding classification accuracy measure mask this difference when the largest values in the fraction vectors estimated by the two methods for a pixel correspond to the 'truth' label. But the fraction vector estimated by the finite mixture model approach is closer to the 'truth' by the MAE measure.

To summarize, the use of statistical finite mixture models with the original, pixel-scale spectral measurements improves classification accuracy at each of the successive spatial scales as compared to aggregation of pixel-scale class labels, which uses spectral measurements only at the pixel scale.

## 3. Second experiment: adaptive selection of spatial scales

As shown in the first experiment, coarsening spatial scale initially improves classification accuracy over per-pixel classification, but that accuracy drops off after a certain scale. This tradeoff between scale and accuracy surely varies with the content of landscape. Moving up to coarser scales initially helps to smooth out the internal variation within pure patches, but the boundaries between patches can be poorly represented with further coarsening. Therefore, a location-sensitive selection of optimal scale is desirable, so as to preserve the boundaries between patches at finer scales while smoothing out the variation within patches at coarser scales. This suggests an adaptive, multiscale method.

### 3.1. Methodology for adaptive selection of spatial scale

A commonly used multiscale data structure is the quad-tree (e.g., Samet, 1990). At the coarsest scale, the entire image is treated as a single pixel and at the next finer scale, the image can be split into four square quads. Each of the quads can be further split into four sub-quads and so on, until the finest scale is reached. This recursive dyadic partitioning results in a hierarchical representation of the image. For an image of $2^J \times 2^J$ pixels, a maximal $J+1$ layers of representation is produced. If one of these layers is used in its entirety, the representation is monoscale (i.e., of a uniform quad size), as in the previous section. However, the nesting of these layers may be exploited to allow the quad size to vary across the image, in which case the representation is multiscale. This feature is appealing for producing land cover characterizations, since it is desirable to represent patch boundaries with smaller quads and patch interiors with larger quads.

Note that the layers of quads in a quad tree correspond to successive tilings of an image region with nested, non-overlapping windows of increasingly larger dyadic size (i.e., $2 \times 2$, $4 \times 4$, etc.). It has been shown in the previous section that the land cover content of a given window can be captured accurately through the use of statistical mixture models. Therefore, it is natural to turn to formal tools of statistical model selection for the task of choosing an appropriate set of quads for the content of an image region. Our class of candidate models is indexed by the set of all recursive dyadic partitionings, which in turn corresponds to the collection of all possible prunings of the complete quad-tree.

Principles of model selection dictate that the overall quality of a 'good' model should balance (a) its goodness of fit, against (b) its complexity. Note that both measures are needed to avoid choosing a model that fits a given set of measurements excessively well but generalizes poorly. When using finite mixture models, the standard criterion for goodness of fit is the statistical likelihood. Assuming statistical independence between spectral measurements among pixels, given the structure of the underlying landscape (i.e., conditional on the set of labels), the likelihood for our context will be simply the product over all pixels of densities like that in Eq. (1), where the weights $\{\pi_i\}$ are restricted to be the same for all pixels within each quad (window) of a candidate model.

If we let $\ell(\mathbf{x}|\mathcal{M})$ denote the logarithm of the likelihood of a model $\mathcal{M}$, based on measurements $\mathbf{x}$ and we let pen $(\mathcal{M})$ be a penalty associated with $\mathcal{M}$, our model selection procedure can then be expressed as one of seeking

$$\hat{\mathcal{M}} \equiv \arg\max_{\mathcal{M}} \{\ell(\mathbf{x}|\mathcal{M}) - 2\ \mathrm{pen}(\mathcal{M})\}, \tag{3}$$

where the optimization is over all models $\mathcal{M}$ defined through some pruning of the quad-tree and the equipping of the resulting quads with arbitrary weights $\{\pi_i\}$. While the penalty function conceivably can take on all manners of forms, we choose to use a simple, concise representation for spatial complexity, given by

$$\mathrm{pen}(\mathcal{M}) = m\{4/3\ \log\ 2 + (l_c - 1)\beta\ \log\ N\}, \tag{4}$$

where $m$ is the number of quads used in the quad-tree underlying $\mathcal{M}$, $l_c$ is the total umber of possible land cover classes across the image (e.g., $l_c = 4$ in the experiments of this and the previous sections), $\beta$ is a tuning parameter (e.g, we have used $\beta = 1/8$ in our experiments), and $N$ is the number of pixels in the image.

The genesis of this particular choice of penalty is rooted in a formal mathematical argument combining elements of statistical and information theory, in the spirit of those underlying the penalties associated with traditional methods of model selection (also called 'subset selection') in linear regression, such as the $C_p$ statistic and the AIC criterion (e.g., Weisberg, 1985, Ch. 8.8). Its derivation parallels that found in the appendix of Kolaczyk et al. (submitted for publication) and is omitted. However, the effect of this penalty is straightforward to see, in that for a given number of classes $l_c$ and an image of a given size $N$, the gains in likelihood obtainable through the use of more quads is directly offset

by the total number $m$ used. Therefore, models $\mathcal{M}$ that are more complex than others but offer only a modest gain in likelihood will be less appealing under this criterion.

The expression in Eq. (3) specifies that an optimization is to be carried out. In fact, this optimization may be solved exactly using an efficient dynamic programming algorithm, analogous to the type of bottom-up optimal tree-pruning algorithms used in fitting decision trees (e.g., Breiman et al., 1984). On the simulated image data used in this section, this algorithm typically ran in 30 seconds on a Dell Precision 450 workstation running Linux. The essence of the algorithm is that, at each of the dyadic scales, recursively, a decision is made on whether to use a single label for a quad (i.e., merging) or to keep the models chosen for the four sub-quads at the previous scale (i.e., splitting), based on complexity-penalized maximum likelihood. Specifically, the following steps are carried out:

- The contribution to the penalized log-likelihood from each block at the $2 \times 2$ scale, for those blocks that entail merging, is the logarithm of the likelihood for the mixture model from Eq. (1), with all four pixels using the same class label, minus a term equal to the penalty from Eq. (4) with $m = 1$ for using a single block. Similarly, when splitting, this contribution is the sum of the logarithms of the likelihoods of the four individual pixels at the $1 \times 1$ scale, minus a penalty from the penalty function with $m = 4$ for using four blocks. The results of this comparison are then saved, i.e., whether the quad is merged or split, the corresponding penalized log-likelihood, and the set of fitted weights.
- Recursively, at the $4 \times 4$ scale and coarser, the penalized log-likelihood for merging is calculated the same way as at the $2 \times 2$ scale, and the penalized log-likelihood for splitting is simply the sum of that for the four sub-quads obtained from the previous scale.

We note that, given the scale-recursive nature of the algorithm, moving from fines scales (bottom of the quad-tree) to coarse (root of the quad-tree), one in fact obtains at each of the successive scales a model that is optimal among that subset of models $\mathcal{M}$ allowing quads up to but no coarser than the current scale of iteration. Since in practice the underlying structure of the landscape may be such that only a certain limited range of scales play a role, a user may find it useful to only run the algorithm up through this range, thus saving some computational time. In addition, even if the complete algorithm is run, the succession of optimal models generated in this manner can communicate useful information as a group, as we shall see below.

Given $\hat{\mathcal{M}}$ at any of the successive scales, one could then in principle label each quad using a majority rule or some other rule on the weights. However, the resulting land cover characterization will likely look blocky and lack translation invariance. Fortunately, a TI implementation of this framework is possible, with only a modest increase in computational complexity. The details, which are slightly more involved than those described in our first experiment, may be found in Kolaczyk et al. (submitted for publication). To produce a land cover classification in this extended framework, the vectors of weights corresponding to each pixel are averaged over all possible shifts of the image. The pixel is then assigned the label corresponding to the largest fraction in this average weight vector. Implementation of this procedure on the simulated image data used in this paper typically took 48 minutes if the algorithm was run to the coarsest scale, and 19 min if to the $32 \times 32$ scale where the algorithm automatically stopped aggregation, on the same computer mentioned above.

### 3.2. Experiments and results

We used the same landscape and simulated data of the first experiment and applied the multiscale approach just described. The results, in the form of the sequence of optimal models at the successive scales just described, may be compared to those of the monoscale approaches studied in our first experiment. Recall that these approaches were 'monoscale' in the sense that at each scale the resulting map involved aggregation on quads only of that scale. In contrast, in our multiscale method, the maps at each scale may contain information adaptively chosen from any of the finer scales as well. Note that the results below are presented only at the dyadic scales, given the nested nature of the quad-tree.

Fig. 3 shows the land cover maps corresponding to the sequence of optimal multiscale models just described, along with maps produced by the two monoscale models described in Experiment one, for the image shown in Fig. 2(b). At the $2 \times 2$ and $4 \times 4$ scales, the multiscale maps are exactly the same as those by the finite mixture model. However, note that at coarser scales, i.e., $8 \times 8$ and onwards, there is very little degradation in quality for the multiscale method when compared to the $4 \times 4$ scale, in contrast to the two monoscale methods. These results provide a clear illustration of the ability of the multiscale approach to maintain finer detail at coarser resolutions; for example, the 63-pixel conifer patch in the middle of hardwood is well represented even at the $32 \times 32$ scale. The maps of the $64 \times 64$ and $128 \times 128$ scales stay the same as that of the $32 \times 32$ scale, due to an automatic termination of spatial coarsening induced by the penalty in Eq. (4). The incongruous segments of labels with a width of one or two pixels lining the edges of the image are artifacts created by the 'wrapping' in the TI implementation, but they appear trivial, compared to the improvement of TI in the overall representation. An examination of the maps at scale $32 \times 32$ based on a fixed quad-tree (the last row of the right panel) may help to better visualize the multiscale nature of the method and to appreciate the effect of TI.

The classification accuracy for our multiscale method based on all 30 trials is plotted with the accuracy curves

from the first experiment, in Fig. 4(a), and the error in estimation of land cover fractions, MAE, with the others in Fig. 4(b). At the $2 \times 2$ and $4 \times 4$ scales, the multiscale model and the finite mixture model have identical accuracies. It is because for this particular set of data, the multiscale method did not keep details finer than $2 \times 2$ and $4 \times 4$ at these two scales, i.e., it was effectively a monoscale finite mixture method. The difference, however, is that the multiscale method is not restricted to use only $2 \times 2$ or $4 \times 4$ quads—it chose to do so. At the $8 \times 8$ scale and beyond, the multiscale method has greater accuracy than either of the monoscale approaches. Moreover, the accuracy is essentially maintained from its peak at the $8 \times 8$ scale. This is due to the ability of the method to inherit finer spatial scales and in fact, there is no further coarsening beyond the $32 \times 32$ scale. These results indicate that this choice is made in an effective manner.

We note that while there are other multiscale methods for image processing in general (e.g., Bouman & Shapiro, 1994), and even in land cover characterization (e.g. Franklin & Wilson, 1992), our approach links in an intuitive manner both to the basic building blocks of pixel-wise methods and the notion of scale linked to quad-trees.

To summarize, localized adaptive selection of spatial resolution using methods of statistical model selection is able to not only match but improve upon the optimal scale-specific accuracy of the monoscale approaches described in the last section.

## 4. Third experiment: adaptive selection of spatial and categorical scales

While our use of mixture models was found to yield certain gains in information over aggregation methods in Section 2, and while they were a correspondingly fundamental component of the multiscale framework presented in Section 3, it is in allowing for the notion of categorical scale to be incorporated in a simple but effective manner in a statistical modeling paradigm that they become particularly valuable to us. By expanding the class of models from Section 3 in a natural manner, and generalizing the notion of complexity accordingly, the multiscale framework just described extends to a fully multiscale, multigranular (MSMG) framework for remote sensing image classification.

### *4.1. Methodology for jointly adaptive choice of spatial and categorical scales*

A mixture of all possible pure classes is used, for all quads at all scales, throughout the first two experiments. In reality, however, not all classes will be present in all regions of an image at all scales. From the perspective of our statistical modeling, this means that different mixtures of different numbers of pure classes can be expected to best fit different quads, depending on the scale of a quad and the local composition of the landscape. Just which combinations of classes best fit which quads therefore can be expected to provide useful information on the categorical complexity, or granularity, of the corresponding land cover.

Not all possible mixtures of pure classes may be both realistic and of interest to the user. In our framework, candidate mixtures are to be identified by the user based on prior knowledge of the landscape, and structured in a hierarchical fashion, with more specific classes spatially nested within more general classes, along the lines of Woodcock and Harward (1992). We use this type of hierarchy to explicitly constrain the class of possible finite mixture models fit to any given quad in the procedure described below, with each candidate mixture model possessing a name of its own, corresponding to the appropriate general class label in the hierarchy. The hierarchy should reflect the characteristics of the scene, and can also be modified based on intended use of the map.

Consider, for example, the simple categorical hierarchy shown in Fig. 5(b), which was used for the experiments presented later in this section. It has the four classes used in the first two experiments as specific classes, at the leaf level of the hierarchy, and three general classes, i.e., *forest* and *non-forest* and then *vegetation*. All mixtures are ultimately characterized by their respective specific classes and the proportions thereof. Thus, we allow for regions to be potentially called *forest* (or *non-forest*) where the two classes *conifer* and *hardwood* (or *grass* and *brush*) are spatially mixed in sufficient proportions. Similarly, where all four specific classes are present, the label *vegetation* can be one of the candidate classes.

As in the multiscale framework described in Section 3, we use statistical model selection to choose an appropriate multiscale, multigranular representation for a set of remote sensing measurements. The set of candidate models $\mathcal{M}$ is now indexed by the set of all labeled recursive dyadic partitionings, which in turn corresponds to the collection of all possible labeled prunings of the complete quad-tree. The possible labels themselves are simply those to be found on the specified categorical hierarchy. Our task of model selection again can be expressed in the form of the optimization in Eq. (3), but we modify the penalty function to account for not only complexity of spatial scale, as in Eq. (4), but also complexity of categorical scale. Using arguments analogous to those producing the previous penalty, this new penalty has the form

$$\mathrm{pen}(\mathcal{M}) = \left(\frac{4m}{3}\right)\log 2 + m \log L_c + \sum_{i=1}^{m} [l(R_i) - 1]\beta \log N. \tag{5}$$

Here $m$, $\beta$, and $N$ are the same as in Eq. (4), the new variable $L_C$ refers to the number of labels in the hierarchy, and $l(R_i)$ is the number of specific (pure) classes in a mixture used for region (i.e., quad) $R_i$. Note that a specific

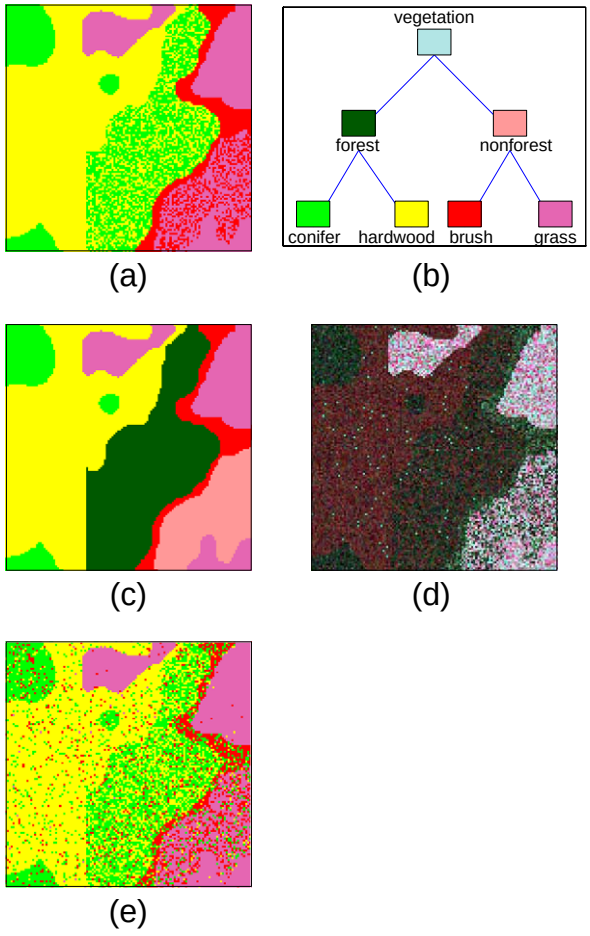


Fig. 5. Simulated landscape with two of the patches mixed. (a) The 'monogranular truth'; (b) categorical hierarchy; (c) the 'multigranular truth' with respect to the categorical hierarchy; (d) color composite with TM bands 432 in RGB; (e) per-pixel classification. Note (d) and (e) are based on one trial.

class is considered to be a mixture with just one component. The parameter $L_C$ has a constant component, i.e., the number of specific classes, and a variable component, i.e., the number of general classes, which can differ with applications for the same set of specific classes and image data. So both the second term and the third term in the penalty function jointly penalize spatial and categorical complexity.

The optimization in Eq. (3) can again be solved exactly and efficiently, using a variation on the dynamic programming algorithm described in Section 3 for the multiscale framework. The algorithm maintains the same scale-recursive structure, considering the splitting or merging of each quad at each scale. The only difference is that, for each quad, an additional step is now included in which all possible mixtures of classes are compared, through their likelihood and the appropriate penalty of their categorical complexity (all models share the same spatial complexity on a given quad). In more detail, the following procedures are carried out:

- At the $2\times2$ scale, the calculation of penalized log-likelihood for merging considers all classes in the hierarchy. The log-likelihood for merging using a mixture class is the logarithm of the likelihood from Eq. (1), minus penalty from Eq. (5) with $m=1$ for using a single block and $l(R_i)$ the number of the specific classes in this mixture. This is also true for merging using a specific class. The log-likelihood for splitting is the logarithm of the sum of the likelihood of the four finest pixels, minus a penalty with $m=4$ and $l(R_i)=1$. The results of this model selection step are saved, i.e., the class label(s) for the quads, the penalized log-likelihood, and the set of fitted weights.
- Then, at the $4\times4$ scale and coarser, the penalized log-likelihood for merging is calculated the same way as at the previous scale, and for splitting, the penalized log-likelihood is simply the sum of those for the models chosen the four sub-quads of the previous scale.

Note that, similar to the algorithm for our multiscale framework, the recursive pruning at each of the successive scales produces an optimal multiscale, multigranular model which include quads up to but no coarser than the current scale. Additionally, a TI version of this procedure can be implemented in direct analogy to that for our multiscale framework; we again refer the reader to Kolaczyk et al. (submitted for publication) for details. Actual computational times scale like those reported in Section 3, in proportion to

the size of the categorical hierarchy. For our simulated image, the complete MSMG model with TI took 61 min, while execution up to the $32 \times 32$ scale took 25 min. For comparison, these numbers and those cited earlier in this paper, have been collected in Table 1.

It should also be noted that, in contrast to the output of our multiscale method, which consists of optimally chosen quads and weights, the output of our MSMG method additionally includes the labels selected. This opens the possibility of using either the labels, the weights, or both in producing a final land cover map. We have found the weights to contain more useful information than the class labels themselves, which is not surprising in light of the results of Section 2. Therefore, in post-processing the results of our statistical model selection procedure to produce a land cover map, we apply to the weights a modification of the majority-vote rule employed with the multiscale framework of Section 3, adjusted to account for multiple granularities in a manner guided by the categorical hierarchy.

Specifically, in our TI implementation we average the weights over all possible shifts, as in Section 3. Then, each pixel is assigned a category as specific as possible, under the condition that the category's areal fraction in terms of summed fractions of its specific classes is greater than 50% and it contains at least 75% of the areal fraction of its parent. The reason for this two-staged criteria is that, as can be expected, the presence of even a small amount of a given specific class typically will cause the model selection procedure to choose a mixture model using a more general class that includes that specific class. However, an expert might judge that this particular specific class is not present in sufficient quantity to merit the label of the general class. The simple labeling strategy we adopt allows such judgment to be incorporated in a straightforward fashion, within the context of the same categorical hierarchy used in the original model selection. An outline of our overall MSMG framework can be found in Fig. 6.

To benchmark this multiscale, multigranular approach, the two monoscale, monogranular aggregation approaches described in the first experiment (i.e., label aggregation and finite mixture model) were extended to be multigranular by interpreting the estimated fraction vectors produced by each

Table 1
The computational time of the multiscale model (MS) and multiscale, multigranular model (MSMG) on the simulated data. For the non-TI case, the results for running the model to the coarsest scale, $128 \times 128$, are shown. For the TI case, additional results up to the intermediate $32 \times 32$ scale at which the respective models automatically stopped aggregation are also reported

| | Non-TI | TI | |
|---|---|---|---|
| | | $32 \times 32$ | $128 \times 128$ |
| MS | 30 sec | 19 min | 48 min |
| MSMG | 55 sec | 25 min | 61 min |

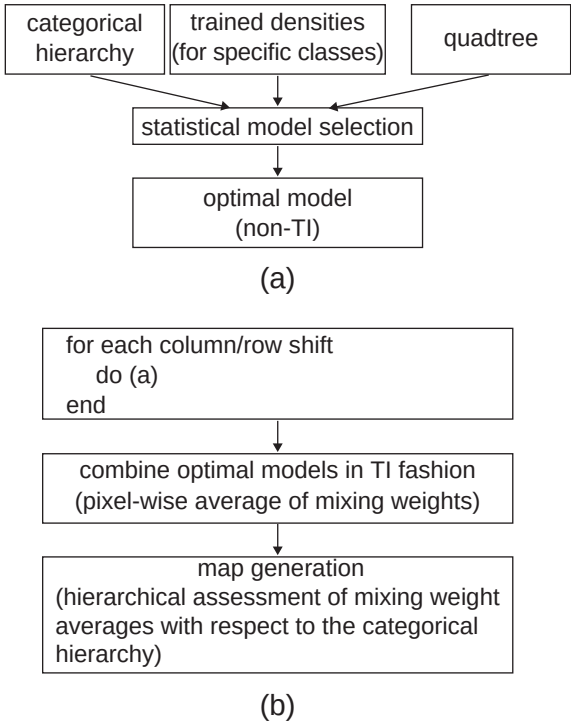


Fig. 6. Schematic of the MSMG framework: (a) the core of MSMG, without TI implementation; (b) MSMG with TI. Note that the for-loop in the TI implementation is only implicit, and is instead done in a computationally efficient manner.

in the same manner as just described, with respect to the same categorical hierarchy and the same specified thresholds of areal fractions.

### 4.2. Experiments and results

The landscape used in the first two experiments was modified by making two of the patches mixed and keeping the others pure. One of the mixed patches is a *conifer–hardwood* mixture (*forest*), and the another is a *brush–grass* mixture (*non-forest*). Mixed patches were constructed by first randomly assigning a pixel in the patch a label of one of the two specific classes with certain probability and then generating spectral measurements from the fitted density for that class. The probabilities (which are simply the weights in the corresponding mixture model) used for the *conifer–hardwood* mixture are 0.6 and 0.4, respectively, and are 0.55 and 0.45 for the *brush–grass* mixture. Note that the mixing pattern of *brush* and *grass* is hypothetical since here *grass* density is fitted with cultured grass from golf courses, which is never mixed with *brush* in reality. But this mixing setup in this experiment has practical implications. The resulting landscape represented with specific class labels ('monogranular truth') is shown in Fig. 5(a). Since the mixed patches are highly fragmented with specific classes, it may be nonsensical to label the mixed patches using the specific class labels at the pixel scale, even if a per-pixel classification could achieve an accuracy up to 100%. The

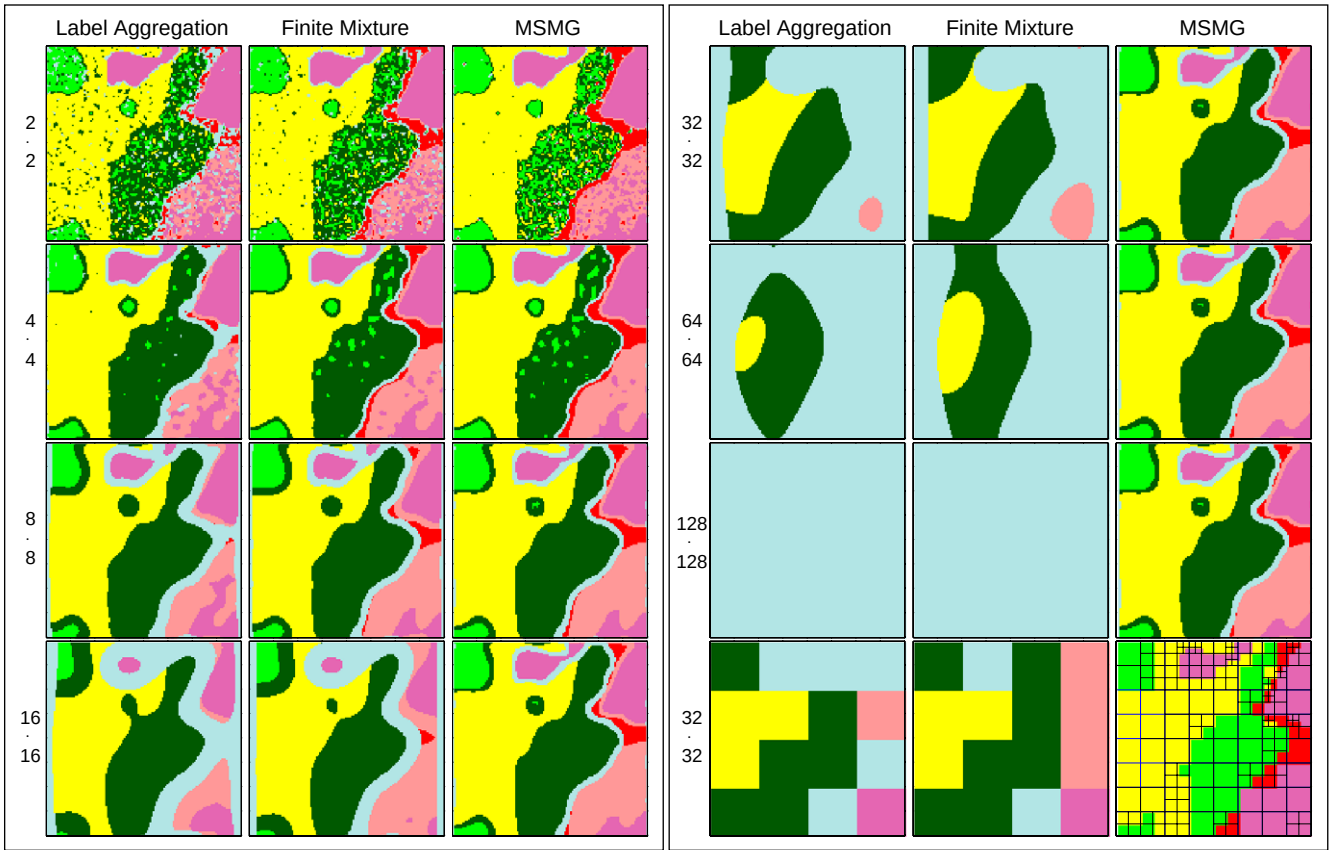


Fig. 7. The land cover maps for the landscape with two of the patches mixed, produced by label aggregation, finite mixture model, and MSMG approaches, at the nested spatial scales $2 \times 2$, $4 \times 4$, ..., $128 \times 128$. Note the last row on the right panel shows maps without translation invariance (TI) implementation at the $32 \times 32$ scale. See Fig. 5(b) for map legend.

use of a more general class can greatly simplify the pattern and be desirable for many purposes. The corresponding multigranular representation ('multigranular truth'), conforming to the hierarchy presented in Fig. 5(b) is given in Fig. 5(c).

The simulation was run 30 times and the simulated images were analyzed with the methods just described. Fig. 5(d) shows a simulated TM image in color composite from one of the trials. The per-pixel classification of this image is shown in Fig. 5(e) with an accuracy of 84% when evaluated against the 'monogranular truth'. However, our real interest is in capturing the 'multigranular truth'. The classification of this image by MSMG and the two monoscale methods with multigranular extension, at successive coarser scales, can be found side by side in Fig. 7. At the $2 \times 2$ scale, the underlying 'multigranular truth' for the mixed patches is already beginning to appear in the maps produced by all three methods. By the $4 \times 4$ scale, the pure patches are labeled in a more or less uniform fashion, while the mixed patches are beginning to solidify. From the $8 \times 8$ scale onwards, both pure patches and mixed patches appear in solid color. As in the second experiment, the MSMG maps at the $64 \times 64$ and $128 \times 128$ scales stay the same as that at the $32 \times 32$ scale, as a result of the adaptive choice of optimal spatial and, in this case, categorical scale. The artifacts at the edge of the image created by 'wrapping' in the TI implementation, albeit trivial, are visible. The classification based on a fixed partition at the $32 \times 32$ scale is shown in the last row of the right panel of Fig. 7, which can be compared with the TI version at this scale to see the effect of the translation invariant implementation. In contrast, the finite mixture model and the label aggregation methods keep coarsening when moving up to coarser spatial scales and produce poor results at the $16 \times 16$ scale and beyond.

Fig. 8(a) shows a comparison of classification accuracy, based on all 30 trials, for the MSM approach of this section against the multigranular extensions of the label aggregation and finite mixture model methods of Section 2. The MSMG approach is clearly superior to the other two, and to the label aggregation in particular, in analogy to what was observed for the multiscale, monogranular method in the experiment of Section 3. All three approaches performed best at the $4 \times 4$ scale, with mean accuracies of 89%, 87%, and 81%, respectively. The accuracy for the multiscale, multigranular approach does not drop much from the peak accuracy at coarser scales, while that for the other two approaches drops considerably. Compared to their monogranular counterparts in Section 3, the multiscale, multigranular approach and the label aggregation approach with multigranular extension

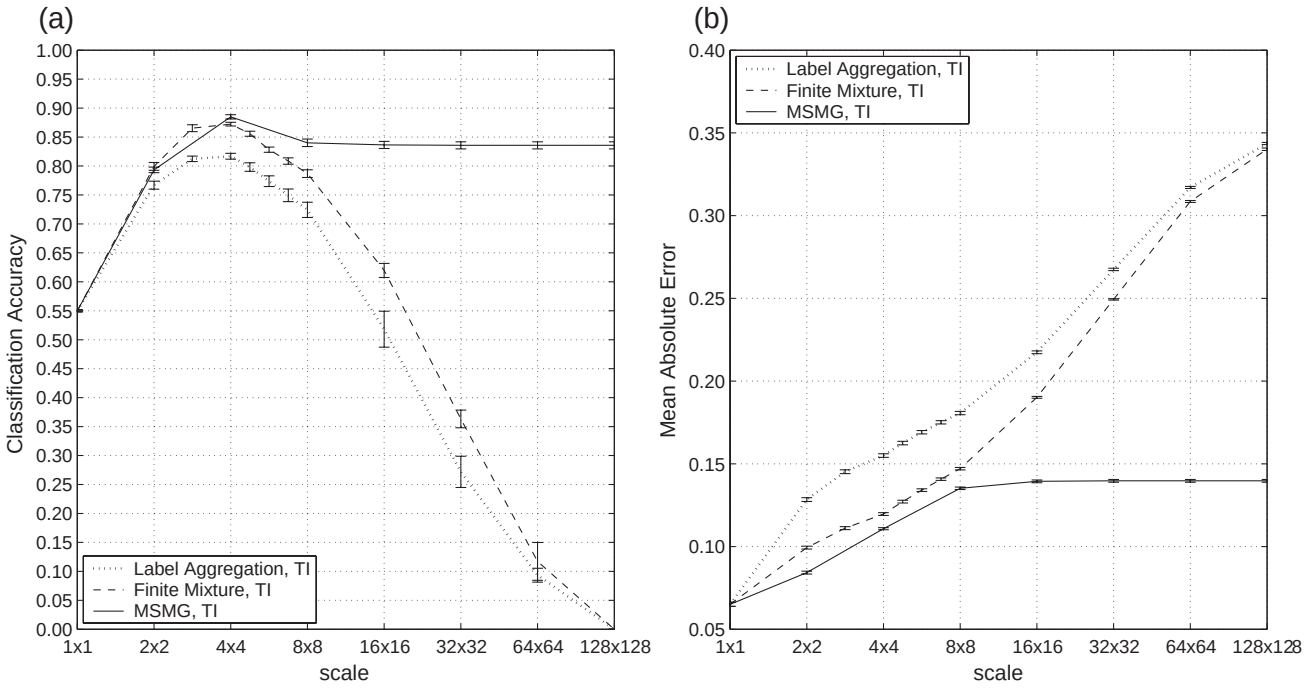


Fig. 8. Accuracy of the multiscale multigranular method (MSMG) and the multigranular extensions of finite mixture model and label aggregation methods for the landscape with two of the patches mixed: mean and $\pm 1$ standard deviation based on 30 trials. (a) Classification accuracy at each scale for the three methods; (b) the mean absolute error in fraction estimation by the three methods. For MSMG, the results are shown at nested spatial scales only i.e., $2 \times 2$, $4 \times 4$, . . ., $128 \times 128$; for the other two methods, results at additional scales $3 \times 3$, $5 \times 5$, $6 \times 6$ and $7 \times 7$ are also shown.

have a smaller difference in MAE at most of the scales (Fig. 8(b)), but they differ more in their classification accuracies. We note that the multiscale multigranular approach has a lower accuracy than the multiscale approach in the second experiment, because of the greater difficulty of the new problem.

One noticeable phenomenon that arises here in the maps produced in this experiment, and not in the previous experiments, is the presence of 'bands' straddling adjacent patches along the boundary. In reality our landscape is designed with an abrupt shift between classes across boundaries. But our method labels a narrow strip along the boundaries as a mixture of the classes on either side. This phenomenon is due in part to the fact that the method generally seemed to operate in this particular landscape at scales no finer than a slight coarsening of the original pixel scale (e.g., the $4 \times 4$ scale in the experiment of Section 3), in an effort to balance accuracy against noise in the data. The mathematical analysis in Kolaczyk et al. (submitted for publication) suggests that the width of these boundaries will in general be roughly on the order of $\log N$ for an image of $N$ pixels using the methodology proposed here, although the precise width will vary with specifics of the landscape, the 'pure' classes and hierarchy used. This effect is similar to the so-called boundary effect described in the GIS literature (e.g., Ehlschlaeger & Goodchild, 1994).

The degree to which these bands are apparent in the final land cover map depends on the nature of the post-processing that is applied to the results of the model selection stage. For example, while the multiscale (monogranular) method of Section 3 also uses blocks slightly coarser than the pixel resolution near boundaries, we force the method, in our post-processing step, to choose a 'pure' class label for each quad. Here, for our multiscale, multi-granular method, our post-processing algorithm allows the use of labels of mixed classes anywhere in the image. A more sophisticated post-processing strategy, one that attempts to distinguish between interiors and boundaries of subregions, should be able to alleviate this effect, if desired. Although we do not pursue this extension here, we remark that, since it is in 'bands' labeled as general classes where most of the classification error was observed to occur, accuracies even better than those described above should be achievable.

To summarize, the multiscale, multigranular approach produces sensible, visually appealing characterizations of the general pattern in the more complex landscape introduced in this section, with higher accuracies at all scales compared with the multigranular extension of the conventional label aggregation approach and the finite mixture model approach.

## 5. Discussion and conclusions

The overall methodology presented in this paper addresses the issue of choosing spatial and categorical scale in remote sensing land cover classification. As a technical

framework, it has the flexibility to incorporate the statistical components of generic conventional monoscale, monogranular classification methods, if the probability density models of those methods for the specific classes are available at the pixel scale. The progressive nature of the presentation of our results was designed to facilitate the exposition of each of the key components of our overall methodology: mixture models, multiscale modeling, and multiscale, multigranular modeling. Accordingly, we have shown that both the components themselves and the complete framework have important implications for addressing some fundamental issues in land cover remote sensing.

First, the use of statistical finite mixture models, as illustrated in Section 2 (Experiment 1), can improve the accuracy of pixel-wise classification over the conventional label aggregation approach. This result can impact, for example, the validation of coarse-scale land cover classification or sub-pixel unmixing, where typically the 'truth' at the coarser scale is obtained from conventional aggregation of class labels derived at a finer spatial scale, i.e., what we have called label aggregation.

Second, our multiscale method in Section 3 (Experiment 2) addresses the limitation of monoscale approaches to land cover classification that has been highlighted in many studies, and responds to calls for methods that make adaptive choice of scale in correspondence with local complexities in the image (Barnsley et al., 1997; Milne & Cohen, 1999; Moody & Woodcock, 1995). While some rudimentary multiscale approaches based on the quad-tree data structure respond similarly to some degree (e.g., Csillag, 1997; Franklin & Wilson, 1992), they lack the statistical rigor of our approach and the translation invariant implementation. Our multiscale method adaptively chooses spatial scale and automatically stops aggregation, in a location-sensitive manner, at the resolution beyond which no further accuracy can be gained, inheriting the details from finer scales where necessary.

Interestingly, our multiscale approach with TI implementation actually takes non-trivial steps towards addressing the modifiable areal unit problem (MAUP) in remote sensing. MAUP, as a more general problem in spatial data analysis, originates from the fact that a large number of ways exist in which a study area can be divided into non-overlapping areal units for data collection and analysis (Marceau et al., 1994a, 1994b; Openshaw & Taylor, 1981), both across different choices of scale (the scaling problem) and within a given scale (the zoning problem). Remote sensing represents a particular case of MAUP in that the pixel sizes of remote sensing images are generally arbitrary, when considered across various types of instruments and applications, and the use of quads at a given scale in the analysis of fine-scale images is just one choice of areal units. Our multiscale approach allows, through the use of quad-trees, for the partitioning of the study area across a range of scales and, through the use of model selection, chooses scales in a locally sensitive manner across the image scene, thus addressing the scaling problem. In addition, the TI implementation considers all possible ways of partitioning allowed by the multiscale approach, thus addressing the zoning problem. These aspects of our method would see to merit additional study.

Thirdly, the multiscale, multigranular framework presented in Section 4 (Experiment 3), in allowing for adaptive choice of both spatial scale and categorical scale, can be viewed as approaching the task of land cover classification as a version of the 'cartographic generalization' problem in cartography, wherein one seeks to best represent selected classes of features at different map scales. Our framework provides a statistically grounded procedure for generating maps that are potentially more intuitive and visually appealing than those created using monogranular methods. The results are much more like a 'hand-drawn' map a cartographer might produce, showing both generalizations and details, as warranted.

Ongoing work includes the evaluation of our framework with real TM image data of Plymouth County, Massachusetts, with a greater number of specific classes and a more complicated categorical hierarchy than used here. In addition, a related focus of this work is a deeper study of the necessary principles and practice needed to make effective choice of categorical hierarchies for our framework. Relevant issues include the choice of basic scene elements, the effect of varying the structure of the categorical hierarchy, and the effect of class mixing at the sub-pixel scale.

## Acknowledgments

This research was supported in part by National Science Foundation grants BCS 0079077 and 0318209 and by Office of Naval Research award N00014-99-1-0219. We thank Curtis Woodcock for guidance in field collection and comments. We thank the four anonymous reviewers for their valuable suggestions and comments that greatly improved the quality of this paper.

## References

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). A land use and land cover classification system for use with remote sensor data. USGS Professional Paper 964.

Barlow, J., Martin, Y., & Franklin, S. E. (2003). Detecting translational landslide scars using segmentation of Landsat ETM+ and DEM data in the northern Cascade Mountains, British Columbia. *Canadian Journal of Remote Sensing*, *29*(4), 510–517.

Barnsley, M. J., Barr, S. L., & Tsang, T. (1997). Scaling and generalization in land cover mapping from satellite sensors. In P. R. Van ardingen, G. M. Foody, & P. J. Curran (Eds.), *Scaling up, from cell to landscape* (pp. 173–199). Cambridge University Press.

Bouman, C. A., & Shapiro, L. (1994). A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on Image Processing*, *3*, 162–177.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

Choi, H., & Baraniuk, R. (2001). Multiscale image segmentation using wavelet-domain hidden Markov models. *IEEE Transactions on Image Processing*, *10*(9), 1309–1321.

Cohen, W. B., Maiersperger, T. K., Yang, Z. Q., Gower, S. T., Turner, D. P., Ritts, W. D., et al. (2003). Comparisons of land cover and LAI estimates derived from ETM plus and MODIS for four sites in North America: A quality assessment of 2000/2001 provisional MODIS products. *Remote Sensing of Environment*, *88*, 233–255.

Coifman, R. R., & Donoho, D. L. (1995). Translation-invariant de-noising. In A. Antoniadis, & G. Oppenheim (Eds.), *Wavelets and statistics* (pp. 125–150). Springer-Verlag.

Csillag, F. (1997). Quadtrees: Hierarchical multiresolution data structures for analysis of digital images. In D. A. Quattrochi, & M. F. Goodchild (Eds.), *Scale in remote sensing and GIS* (pp. 247–271). Boca Raton, FL: CRC Lewis.

Ehlschlaeger, C. R., & Goodchild, M. F. (1994). Dealing with uncertainty in categorical coverage maps: Defining, visualizing, and managing errors. *Proceedings of workshop on geographic information systems at the conference on information and knowledge management* (pp. 86–91). Gaithersburg MD.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*, 185–201.

Franklin, J., & Woodcock, C. E. (1997). Multiscale vegetation data for the mountains of Southern California: Spatial and categorical resolution. In D. A. Quattrochi, & M. F. Goodchild (Eds.), *Scale in remote sensing and GIS* (pp. 141–168). Boca Raton, FL: CRC Lewis.

Franklin, S. E., & Wilson, B. A. (1992). A three-stage classifier for remote sensing of mountain environments. *Photogrammetric Engineering and Remote Sensing*, *58*, 449–454.

Gong, P., & Howarth, P. J. (1992). Frequency-based contextual classification and gray-level vector reduction for land-use identification. *Photogrammetric Engineering and Remote Sensing*, *58*(4), 423–437.

Gurney, C. M. (1983). The use of contextual information in the classification of remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *49*, 55–64.

Irons, J. R., Markham, B. L., Nelson, R. F., Toll, D. L., Williams, D. L., Latty, R. S., et al. (1985). The effects of spatial resolution on the classification of Thematic Mapper data. *International Journal of Remote Sensing*, *6*, 1385–1403.

Ju, J., Kolaczyk, E. D., & Gopal, S. (2003). Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. *Remote Sensing of Environment*, *84*, 550–560.

Kolaczyk, E. D., Ju, J. C., & Gopal, S., (submitted for publication). Multiscale, multigranular statistical image segmentation. *Journal of the American Statistical Association.*

Marceau, D. J., Gratton, D. J., Fourier, R. A., & Fortin, J. (1994a). Remote sensing and the measurement of the geographical entities in a forested environment: 2. The optimal spatial resolution. *Remote Sensing of Environment*, *49*, 105–117.

Marceau, D. J., Howarth, P., & Gratton, D. J. (1994b). Remote sensing and the measurement of the geographical entities in a forested environment: 1. The scale and spatial aggregation problem. *Remote Sensing of Environment*, *49*, 93–104.

Markham, B. L., & Townshend, J. R. G. (1981). Land cover classification accuracy as a function of sensor spatial resolution. *Proceedings of the Fifteenth International Symposium on Remote Sensing of Environment*, Ann Arbor, Michigan (pp. 1075–1090).

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: John Wiley.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley, John & Sons.

Milne, B. T., & Cohen, W. B. (1999). Multiscale assessment of binary and continuous landcover variables for MODIS validation, mapping, and modeling applications. *Remote Sensing of Environment*, *70*, 82–98.

Moody, A., & Woodcock, C. E. (1994). Scale-dependent errors in the estimation of land-cover proportions: Implications for global land-cover datasets. *Photogrammetric Engineering and Remote Sensing*, *60*, 585–594.

Moody, A., & Woodcock, C. E. (1995). Calibration-based methods for correcting coarse resolution estimates of land-cover proportions. *Proceedings ASPRS conference*, *vol. 3*. (pp. 684–693).

Openshaw, S., & Taylor, P. J. (1981). The modifiable areal unit problem. *Quantitative geography, the British view* (pp. 60–70). London: Routledge.

Raptis, V. S., Vaughan, R. A., & Wright, G. G. (2003). The effect of scaling on land cover classification from satellite data. *Computers & Geosciences*, *29*, 705–714.

Samet, H. (1990). *The design and analysis of spatial data structures*. Reading, MA: Addison-Wesley.

Schowengerdt, R. A. (1997). *Remote sensing: Models and methods for image processing* (pp. 237–238). Reading, MA: Academic Press.

Smith, M. O., Ustin, S. L., Adams, J. B., & Gillespie, A. R. (1990). Vegetation in deserts: I. A regional measure of abundance from multispectral images. *Remote Sensing of Environment*, *31*, 1–26.

Strahler, A. H., Woodcock, C. E., & Smith, J. A. (1986). On the nature of models in remote sensing. *Remote Sensing of Environment*, *20*, 121–139.

Stuckens, J., Coppin, P. R., & Bauer, M. E. (2000). Integrating contextual information with per-pixel classification for improved land cover classification. *Remote Sensing of Environment*, *71*, 282–296.

Sveinsson, J. R., Hilmarsson, O., & Benediktsson, J. A. (1998). Translation invariant wavelets for speckle reduction of SAR images. *Proceedings of the international geoscience and remote sensing symposium (IGARSS'98)*, *vol. 2*. (pp. 1121–1123).

Thomas, I. (1980). Spatial post processing of spectrally classified Landsat data. *Photogrammetric Engineering and Remote Sensing*, *46*, 1201–1206.

Tilton, J. C. (1998). Image segmentation by region growing and spectral clustering with a natural convergence criterion. *Geoscience and remote sensing symposium proceedings*, *vol. 4*. (pp. 1766–1768).

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: John Wiley & Sons.

Woodcock, C. E., & Harward, V. J. (1992). Nested-hierarchical scene models and image segmentation. *International Journal of Remote Sensing*, *13*, 3167–3187.

Wu, J. G., & David, J. L. (2002). A spatially explicit hierarchical approach to modeling complex ecological systems: Theory and application. *Ecological Modelling*, *153*, 7–26.