Multiscale, Multigranular Statistical Image Segmentation

Eric D. Kolaczyk, Junchang Ju, and Sucharita Gopal*

Abstract

We consider a problem in image segmentation, in which the goal is to determine and label a relatively small number of homogeneous subregions in an image scene, based on multivariate, pixel-wise measurements. Motivated by current challenges in the field of remote sensing land cover characterization, we introduce a framework that allows for adaptive choice of both the spatial resolution of subregions and the categorical granularity of labels. Our framework is based upon a class of models we call *mixlets,* a blending of recursive dyadic partitions and finite mixture models. The first component of these models allows for the sparse representation of spatial structure at multiple resolutions. The second component provides a natural mechanism for capturing the varying degrees of mixing of pure categories that accompany the use of different resolutions, and enables them to be related to a user-specified hierarchy of labels at multiple granularities in a straightforward manner. A segmentation is produced in our framework by selecting an optimal mixlet model, through complexity-penalized maximum likelihood, and summarizing the information in that model with respect to the categorical hierarchy. Both theoretical and empirical evaluations of the proposed framework are presented.

Index Terms - Land cover characterization; mixture models; recursive dyadic partition; remote sensing.

*Eric D. Kolaczyk (E-mail: *kolaczyk@math.bu.edu*) is Associate Professor, Department of Mathematics and Statistics, Boston University, Boston, MA 02215. Junchang Ju is a Post-doctoral Research Associate and Sucharita Gopal is Professor, Department of Geography and Center for Remote Sensing, Boston University, Boston, MA 02215. This research was supported in part by NSF awards BCS 0079077 and BCS 0318209, and by ONR award N00014-99-1-0219.

1 Introduction

Image segmentation is a task wherein one typically seeks to partition the region underlying an image into subregions that are homogeneous as individual elements and yet meaningful as parts of the overall scene. In addition, this task may take on an aspect of classification if one also wishes to equip each of the subregions with one of a set of pre-determined labels. A wide variety of techniques have been proposed over the years for this problem, in fields ranging from statistics to pattern recognition to computer vision. See Haralick and Shapiro (1985) or Pal and Pal (1993) for reviews. The application of these techniques has been quite varied as well, with their usage arising in diagnostic imaging in medicine (e.g., Bose and O'Sullivan 1997), natural scene analysis (e.g., Zhu 2003), and the analysis of remotely sensed image data (e.g., Tilton 1998), to name just a few.

Yet despite this substantial body of work, there remains much to be done in attacking the image segmentation problem in its broadest sense, due in no small part to the difficulties inherent in equipping with a precise mathematical formulation a problem that at its core is influenced in a nontrivial fashion by cognitive elements like vision and semantics (Marr 1982). The work in this paper seeks to address one such difficulty – the problem of effectively modeling the interaction of what we will call 'spatial' and 'categorical' scale. More precisely, we introduce a framework for producing a labeled image segmentation from pixel-wise measurements that allows for simultaneous, adaptive choice of both the spatial resolution of subregions and the categorical granularity of their class labels.

Although our framework is broadly applicable, our motivation for this work derives primarily from the task of remote sensing land cover characterization, in which the goal is to provide a meaningful description of the vegetation and/or land use in a given spatial region, using measurements obtained from remote sensing instruments. In this field, and indeed in the field of geo-spatial sciences more generally, a well-known issue of fundamental importance is that of *scale* and its effects on a wide range of inter-related problems involving spatial measurement and inference. In fact, the critical importance of this issue was highlighted recently through its being listed as one of a set of top ten long-term research challenges by the University Consortium for Geographical Information Science, an umbrella group with a broad representation of universities and professional associations in this area (http://www.ucgis.org/). Our work here may be seen as addressing one particular aspect of this challenge, in focusing on the interaction of spatial and categorical scale. In turn, it has the potential to impact user communities in a variety of areas that depend on accurate, informative characterizations of land cover for their own work, such as agriculture, biodiversity, climate modeling, ecology, forestry, and urban planning.

Organization of this paper is as follows. In Section 2, we present background on the motivating problem of land cover characterization, relevant issues of spatial and categorical scale, and a brief review of related literature. In Section 3, we introduce a new class of models, which we call *mixlets*, that allow for the efficient representation of joint spatial and categorical information in an image simultaneously at multiple scales, through the blending of recursive dyadic partitions and finite mixtures. Additionally, we present a strategy for selecting an optimal mixlet model for a given set of measurements, using a complexity-penalized maximum likelihood criterion, and we show that this strategy has desirable risk properties. In Section 4, we describe how multiscale, multigranular segmentations of images may be produced using mixlets, by directly extracting and combining information on subregions and class labels from a redundant system of fitted mixlet models. We confirm the practical potential of the proposed framework in Section 5, in the context of remote sensing land cover characterization, using a dataset for the Sierra region of central California. Finally, some discussion may be found in Section 6. All proofs and derivations of related formulas may be found in the appendix.

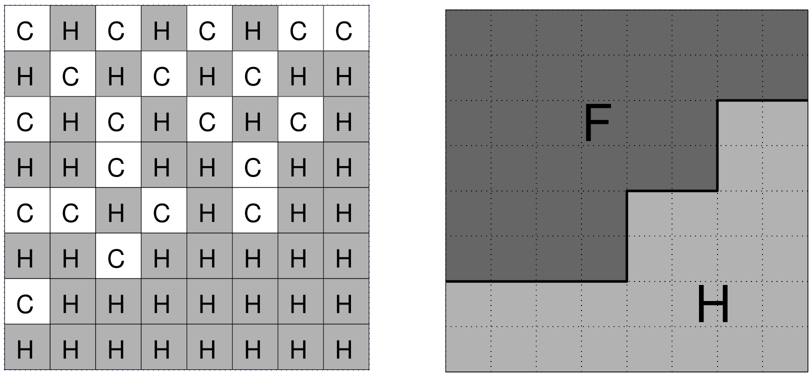
2 Background

2.1 Spatial and Categorical Scale

Consider a finite spatial region of interest, which without loss of generality will be taken to be the unit square $[0,1]^2 \subset \mathbb{R}^2$. In the land cover characterization problem, which we will use to motivate our discussion in this section and the work that follows, this region corresponds to a portion of the Earth and the goal is to produce a meaningful representation of the corresponding land cover from available data. Typically this region, viewed as an image, is accompanied by a digitization into pixels, say $\{I_{i_1,i_2}\}_{i_1,i_2=1}^{n}$, and associated with each pixel I_{i_1,i_2} is a measurement vector x_{i_1,i_2}. The measurements generally are based on reflected or emitted radiation in various bands of the electro-magnetic spectrum (e.g., visible, near-infrared, infrared), gathered on board aircraft or satellites orbiting the earth. The digitization derives from a combination of pre-processing of the radiation measurements and considerations of the instrumental resolution. As an approximation to reality we will assume, as is standard, that the land cover content within any given pixel may be described adequately by one of a set of pre-specified labels, say $c_{i_1,i_2} \in \{1,\ldots,C\}$. Characterization of land cover may then be taken to mean the process of creating an inferred and typically visual description of the collection $\{c_{i_1,i_2}\}$ from the data $\boldsymbol{x} = \{x_{i_1,i_2}\}$. Hence, this task is fundamentally one of image segmentation.

Now consider the two diagrams in Figure 1, designed to show two possible manners in which one might characterize the same land cover in a given spatial region. In Figure 1(a), each of the individual pixels is labeled as consisting of either of two classes of trees i.e., *hardwood* (H) or *conifer* (C). This diagram exemplifies the type of output obtained from standard methods, in which a label $\hat{c}_{i_1,i_2}$ is inferred for each pixel I_{i_1,i_2} as the solution to a classification problem. There are a variety of methods of this nature that have been proposed, including those based on maximum likelihood, decision trees, nearest neighbors, and neural networks. See Landgrebe (2003), for example. However, in looking at this characterization, note that one may argue that the underlying spatial region actually is composed of just two subregions, separated by a fairly smooth boundary, with the lower right-hand subregion filled with *hardwood* and the upper left-hand subregion filled with both *hardwood* and *conifer* interwoven in a somewhat regular fashion. Figure 1(b) shows an alternative characterization along these lines in which, while the label (H) has been used for the *hardwood* in the lower right-hand subregion, a new label (F) for *forest* has been used for the upper left-hand subregion.

From the perspective of image segmentation, in comparing these two characterizations, the second can be said to combine subregions of coarser spatial resolution with the use of a new class label of coarser categorical granularity. That is, the first characterization is a segmentation in which all subregions are restricted to the spatial resolution of the original

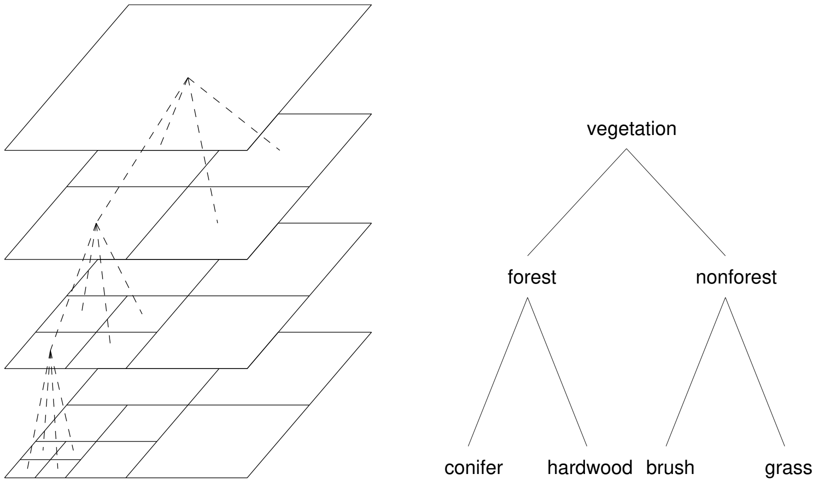


(a) (b)

Figure 1: Schematic illustration of two possible manners in which to characterize land cover in a given spatial region. (Symbols: H = *hardwood*, C = *conifer*, and F = *forest*.)

pixels I_{i_1,i_2}. The labels are therefore similarly restricted to those that are assumed to describe the possible content of a single pixel, which are generally called 'pure class' labels. On the other hand, the second characterization aggregates the original pixels into larger subregions, and accommodates for the mixing of pure classes in one of those subregions by adopting a composite label.

At the heart of this illustration is the notion of a trade-off between spatial complexity and categorical complexity that seems to be managed quite naturally by the human cognitive and visual systems. We propose in this paper a framework for statistical image segmentation that is designed for modeling interactions in the complexity of spatial resolution and categorical granularity and which allows, with appropriate user input, for the adaptive choice of each in producing a final segmentation of an image. Our framework is built upon two corresponding hierarchical structures, which are then linked in a particular fashion. These are shown in Figure 2. The first structure is a so-called quad-tree, which captures the notion of multiple spatial resolutions within a given spatial region by defining pixels at each of a discrete, nested sequence of resolutions as the union of four pixels at the next finer resolution. The second structure is a categorical tree, which serves the analogous purpose of capturing a user's notion of multiple categorical granularities. Just like the quad-tree has at its base nodes representing units of the finest spatial resolution, so too the categorical tree has at its base nodes corresponding to class labels of the greatest specificity or finest granularity. Increasingly higher levels of this tree indicate the manner in which labels of finer granularity



(a) (b)

Figure 2: Illustration of two key hierarchical structures in the mixlet modeling framework: (a) quad-tree induced by recursive dyadic partitions, and (b) categorical tree displaying relationship among labels used in the example of Section 5.

may be subsumed into labels of coarser granularity, such as through the relation $\{H, C\} \rightarrow \{F\}$ in the example of Figure 1. Of course, both of these structures are substantial, but useful, simplifications of the reality inherent in their respective notions of scale. We postpone additional discussion of this issue until Section 6.

2.2 Related Literature

The inter-relationship between spatial resolution and categorical granularity has been recognized and studied for some time. For example, again considering the problem of remote sensing land cover characterization, studies in this area have shown that the accuracy with which regions of a given class may be labeled varies with spatial resolution of the underlying pixels (e.g., Marceau, Howarth, and Gratton 1994). Similarly, landscape ecologists recognize that the identification of spatial patterns and their relationship with underlying ecological processes depends critically on spatial resolution (e.g., Pickett and Cadenasso 1995). However, obtaining precise quantitative characterizations of the relationships between spatial resolution and categorical granularity, across the numerous possible combinations of each, is a subject of ongoing and active research in fields like remote sensing, landscape ecology, and geographical information science.

From a broader perspective, spatial resolution, categorical granularity, and the interaction of the two, may be viewed as facets of the more general problem of *scale.* The term 'scale' has tended to be used in a catch-all manner in imaging and the spatial sciences to capture a variety of inter-related issues. In the context of the land cover characterization problem, such issues include the choice of cartographic units, domain extent, and measurement resolution, the intrinsic dimensionality of structures in the underlying landscape, and the granularity of the categorical labels used. These issues in turn are all intimately related to the so-called modifiable areal unit problem (MAUP), which refers to the impact that choice of modifiable units may have on the collection of data and the inferences that may be made based on such data. See the edited volume by Quattrochi and Goodchild (1998) for an overview of such issues, or Gotway and Young (2002), for a recent review within the statistics literature.

Nevertheless, despite the long-standing presence of the two notions of scale considered in this paper and the literature surrounding them, the framework we propose for modeling them is unique. Specifically, although each of the two hierarchical structures illustrated in Figure 2 separately can be said to be quite standard in various areas of relevant literatures, to the best of our knowledge this paper represents a first in combining them in a conceptually and mathematically precise fashion. For example, there has been much work on multiscale image segmentation algorithms in the past decade, say in the spirit of Bouman and Shapiro (1994). However, while the data generally are modeled at multiple spatial resolutions in such frameworks, using quad-trees or more generally wavelets of various types, the resulting segmentations typically are based only on monogranular sets of categorical labels. Conversely, categorical hierarchies generally seem to be developed from a conceptual, ontological perspective, based on principles and rules of the corresponding sciences, and largely in isolation from spatial hierarchies like quad-trees. Such is the case, for example, in the so-called Anderson system in remote sensing land cover classification, developed by the US Geological Survey (Anderson, Hardy, and Roach 1976). At present it would appear that the only existing methods attempting to integrate quantitative assessment of both spatial resolution and categorical granularity are some of those in the literature on hierarchical segmentation, such as Tilton (1998), which combine canonical region-growing methods of segmentation with clustering of the measurements x_{i_1,i_2} in subregions according to various rules, in a spatially nested fashion. However, such methods are arguably still fairly ad hoc in nature, as compared to the framework developed in Sections 3 and 4, and therefore also less amenable to theoretical evaluation, such as is presented in Section 3.3.

3 Modeling Spatial and Categorical Scale Using Mixlets

Our image segmentation framework, to be described in Section 4, is based upon a new class of models we call *mixlets.* These models are described in this section, and can be seen to be essentially local mixture models obtained by grafting standard finite mixture models onto recursive dyadic partitions (quad-trees). Choice and fitting of an appropriate mixlet model is accomplished through the use of a complexity-penalized maximum likelihood criterion, which attempts to balance fidelity to the data with complexity of spatial and categorical scale. The optimization in this setting may be done using a computationally efficient, tree-based algorithm. A theoretical evaluation of our mixlet models and model selection strategy shows near-optimal asymptotic risk rates within the context of an extension of the classical horizon model (defined below).

3.1 Mixlet Models

Consider an $n \times n$ discretization of $[0,1]^2$, with a corresponding set of $N = n^2$ (multivariate) observations $\{x_{i_1,i_2}\}$, as described in Section 2. Without loss of generality, we will take n to be a power of two, as is common in many areas of application. When the goal of an analysis is to produce a segmentation of an image region with labeled subregions, a common approach is to take the x_{i_1,i_2} to be conditionally independent draws from a finite set of component densities i.e., $x_{i_1,i_2} \sim g(x|c_{i_1,i_2})$, given a collection $\{c_{i_1,i_2}\}$ of true class labels for the pixels $\{I_{i_1,i_2}\}$. Traditionally in such contexts, however, it is assumed that the subregions sought each are to be homogeneous in one of the 'pure' classes in some set $\{c_1, \ldots, c_L\}$. But here we entertain the notion that the subregions may be homogeneous in certain specific inter-woven combinations of pure classes as well. Towards this end, we assume that the set of all possible class labels is specified in association with a categorical tree, say $\mathcal{T}_{Cat}$, like that shown in Figure 2(b). And we express this collection of labels in the form $\mathcal{C}(\mathcal{T}_{Cat}) \equiv \{\{c^{(d,l)}\}_{l=1}^{L_d}\}_{d=0}^{D}$, where d denotes depth, from root to leaves, and l denotes position within depth, from left to right. The class label at the root of $\mathcal{T}_{Cat}$ is thus $c^{(0,1)}$. Additionally, we will use $\alpha_{d'}(c^{(d,l)})$ to refer to the ancestor of $c^{(d,l)}$ at depth $d' \leq d$, while $\delta_{d'}(c^{(d,l)})$ similarly will refer to the set of all descendants of $c^{(d,l)}$ at depth $d' \geq d$ i.e., $\delta_{d'}(c^{(d,l)}) = \{c^{(d',l')} : \alpha_d(c^{(d',l')}) = c^{(d,l)}\}$. In the case that $d' = d$, note that $\alpha_{d'}(c^{(d,l)}) = \delta_{d'}(c^{(d,l)}) = c^{(d,l)}$.

The labels in $\mathcal{C}(\mathcal{T}_{Cat})$ will be used to assign classes to subregions R. In order to allow for subregions whose sizes can vary across a range of scales in an orderly fashion, which is important for both analytical and computational reasons, we restrict the set of possible subregions to those that may be produced through recursive dyadic partitioning of the original region $[0,1]^2$. A recursive dyadic partition (RDP) $\mathcal{P}$ is any partition of $[0,1]^2$ that may be produced through application of the following two rules: (i) $\mathcal{P}_0 \equiv [0,1]^2$, the trivial RDP, and (ii) if $\mathcal{P}$ is an RDP, composed of m squares $R_1, \ldots, R_m$, then any partition obtained by splitting one of the squares R into four equal sub-squares is also an RDP. Note that with n a power of two and our use of dyadic splits on the horizontal and vertical axes (due to the constraint that 'parent' squares be split into four identical 'children' squares), the original $n \times n$ pixelization of the unit square is itself an RDP, which typically is called a complete recursive dyadic partition (C-RDP) and denoted by $\mathcal{P}_n^*$. By $\mathcal{P} \preceq \mathcal{P}_n^*$ we shall mean any RDP that may, upon (possibly) repeated application of rule (ii), be refined to produce $\mathcal{P}_n^*$. To indicate that a subregion R results from a particular RDP, we will write $R \in \mathcal{P}$.

The set of all RDP's $\mathcal{P} \preceq \mathcal{P}_n^*$ can be associated in one-to-one correspondence with the set of all quad-trees, as in Figure 2(a), that are sub-trees of the full quad-tree associated with $\mathcal{P}_n^*$ and share the same root. The set of all RDP's also is closely related to the two-dimensional Haar wavelet basis, in that any RDP representation has an equivalent representation using a Haar basis with hereditary constraints on the non-zero wavelet coefficients (Donoho 1997). As a result, RDPs are a common structure for multiscale modeling in imaging applications, balancing richness of representation with parsimony. The use of rigid dyadic boundaries, however, is known to have the potential for inducing artifacts in results (analogous to the so-called Gibbs phenomenon in Fourier analysis), but these can be overcome using an extension to a redundant system of square subregions, as we describe in Section 4.1, which also has the benefit of producing translation-invariant results. Alternatively, one might use other multiscale systems such as anisotropic Haar bases (e.g., Donoho 1997) or wedgelets (e.g., Donoho 1999), for which extensions of our framework are straightforward.

Our mixlet model class may now be defined as follows.

Definition 1 (Mixlets) *Let $\mathcal{P}_n^*$ be a C-RDP of $[0,1]^2$ and let $\mathcal{T}_{Cat}$ be a categorical tree of labels $\mathcal{C}(\mathcal{T}_{Cat}) \equiv \{\{\, c^{(d,l)}\,\}_{l=1}^{L_d}\}_{d=0}^{D}$. A mixlet model $\mathcal{M}$ is a triple $\{\mathcal{P}, \boldsymbol{c}(\mathcal{P}), \boldsymbol{\pi}(\mathcal{P})\}$, where $\mathcal{P} \preceq \mathcal{P}_n^*$, $\boldsymbol{c}(\mathcal{P}) \equiv \{\, c(R)\,\}_{R\in\mathcal{P}}$, and $\boldsymbol{\pi}(\mathcal{P}) \equiv \{\, \boldsymbol{\pi}(R)\,\}_{R\in\mathcal{P}}$, along with densities $g(x|c)$, for $c \in \left\{c^{(D,1)}, \ldots, c^{(D,L_D)}\right\}$, such that the measurements $x_{i_1,i_2} \in I\!R^k$ are sampled independently according to the mixture densities*

$$f_{\mathcal{M}}\,(x\,|\,c_{i_1,i_2}) = \sum_{c\in\delta_D(c_{i_1,i_2})} \pi_c(R(i_1,i_2)) \;\times\; g\,(x\,|\,c) \quad . \tag{1}$$

Here $R(i_1,i_2)$ is that subregion $R \in \mathcal{P}$ containing the pixel I_{i_1,i_2}, $c_{i_1,i_2} = c(R(i_1,i_2))$ is that label assigned to I_{i_1,i_2} through its association with $R(i_1,i_2)$, $\delta_D(c_{i_1,i_2})$ is a set of size $l_{i_1,i_2} \equiv l(R(i_1,i_2))$ consisting of those labels at depth D of $\mathcal{T}_{Cat}$ that are descendants of the label c_{i_1,i_2}, and the values $\boldsymbol{\pi}(R(i_1,i_2)) \equiv \left(\pi_1(R(i_1,i_2)), \ldots, \pi_{l_{i_1,i_2}}(R(i_1,i_2))\right)$ are mixing weights corresponding to those l_{i_1,i_2} labels, which are non-negative and sum to one.

In other words, for each subregion R of an RDP $\mathcal{P}$, our mixlet model specifies that the measurements x_{i_1,i_2} for all pixels $I_{i_1,i_2} \subseteq R$ be sampled independently from a density $f(x|c(R))$ that is a finite mixture of component densities $g(x|c)$ corresponding to those 'pure' classes in the leaves of the sub-tree $\mathcal{T}(R)$ of $\mathcal{T}_{Cat}$ rooted at $c(R)$. Letting $n(R) = |\{(i_1,i_2) \,:\, I_{i_1,i_2} \subseteq R\}|$ and $l(R) = |\delta_D(c(R))|$, we make the natural restriction that $l(R) \leq n(R)$. From this restriction, combined with (i) mild assumptions on the densities $g(\cdot|c)$ – for example, most continuous densities qualify (Titterington, Smith, and Makov 1985) – and (ii) the ordering inherent among these densities through $\mathcal{T}_{Cat}$, it follows that the mixlet models are identifiable. That is, if $\mathcal{M} = \{\mathcal{P}, c(\mathcal{P}), \pi(\mathcal{P})\}$ and $\mathcal{M}' = \{\mathcal{P}, c'(\mathcal{P}), \pi'(\mathcal{P})\}$, then $f_{\mathcal{M}} = f_{\mathcal{M}'}$ implies that $c(\mathcal{P}) = c'(\mathcal{P})$ and $\pi(\mathcal{P}) = \pi'(\mathcal{P})$.

To illustrate the manner in which our mixlet model class couples spatial and categorical scale, consider again the two land cover characterizations in Figure 1. Recall that the two relevant 'pure' classes are *hardwood* (H) and *conifer* (C). The schematic in Figure 1(a) corresponds to a model that makes use of only the finest spatial and categorical scales, in which the measurement x_{i_1,i_2} of each pixel I_{i_1,i_2} is sampled independently according to the density $g(x|c_{i_1,i_2})$, where c_{i_1,i_2} is either H or C. In contrast, the schematic in Figure 1(b) may be represented efficiently by a mixlet model. To begin with, the 28 pixels that constitute the region labeled with a large 'H' can be described spatially by just four quads – one 4×4 quad in the lower righthand corner, one 2×2 quad immediately above and to the right of that, and two adjacent 2×2 quads to the left. Similarly, the 36 pixels that constitute the region labeled with a large 'F' can be described spatially by just six quads. These 10 quad regions R can be represented by an RDP $\mathcal{P}$ deriving from an appropriate pruning of a C-RDP like that in Figure 2(a). Note that the reduction in spatial complexity of the upper region follows from having first subsumed the categories H and C into a category *forest* (F) of coarser granularity. The corresponding categorical tree $\mathcal{T}_{Cat}$ is simply the bottom leftmost branch of the binary tree in Figure 2(b). The sampling of pixel-wise measurements x_{i_1,i_2} is now determined under the mixlet model by the label of the quad region R in which each pixel I_{i_1,i_2} falls. For those pixels I_{i_2,i_2} falling in one of the four quads making up the 'H' region, x_{i_1,i_2} is sampled according to $g(x|\mathrm{H})$. But for those pixels falling in the 'F' region, the sampling is done with respect to the mixture density $\pi_{\mathrm{C}}(R)\, g(x|\mathrm{C}) + \pi_{\mathrm{H}}(R)\, g(x|\mathrm{H})$, where $R = R(i_1, i_2)$ is that one of the six 'F' quads that contains I_{i_1,i_2}, and $\pi_{\mathrm{C}}(R)$ and $\pi_{\mathrm{H}}(R) = 1 - \pi_{\mathrm{C}}(R)$ are quad-specific mixing weights.

Implicit in the definition of our model is the assumption that the component densities $g(x|c)$ are known. In practice, of course, these densities need to be learned from measurements. Often such learning is done previously off-line, in a supervised fashion, from samples of pixels corresponding to spatial regions that are judged by expert opinion to be sufficiently pure and representative of those to be analyzed. This step actually, while difficult in its own right, is a standard part of many current image segmentation frameworks. For example, significant amounts of time and energy are invested in this and related tasks by researchers working on remote sensing land cover classification, for various combinations of satellite instruments, imaging resolutions, and land cover types. In fact, we view it as a strength of our proposed multiscale, multigranular framework that one can easily incorporate into it the component densities g learned for any given monoscale, monogranular framework. In the applications presented later in this paper, for example, we use mixtures of multivariate Gaussian densities to approximate each g, in the spirit of the mixture discriminant analysis (MDA) framework of Hastie and Tibshirani (1996), where the mixtures were learned as described in Ju, Kolaczyk, and Gopal (2003).

3.2 Selecting an Optimal Mixlet Model

Our approach for producing a segmentation of an image is based on extracting the information in a particular mixlet model, say $\hat{\mathcal{M}} \equiv \hat{\mathcal{M}}(\boldsymbol{x})$, that is optimal in some well-defined sense with respect to the measurements $\boldsymbol{x}$. In order to evaluate the relative merits of each candidate model $\mathcal{M}$, we use a complexity-penalized maximum likelihood criterion, defining

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

where $\ell(\boldsymbol{x} \,|\, \mathcal{M})$ is the log-likelihood of the data under model $\mathcal{M}$ and $\mathrm{pen}(\mathcal{M})$ is a penalty function. Through this criterion we seek to balance fidelity to the data with parsimony of the model. Towards this end, for a given model $\mathcal{M} = \{\mathcal{P}, \boldsymbol{c}(\mathcal{P}), \boldsymbol{\pi}(\mathcal{P})\}$ we let the penalty take the form

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

where m is the number of subregions R defined by $\mathcal{P}$, $L_{\mathcal{C}} = |\mathcal{C}(\mathcal{T}_{Cat})|$ is the total number of class labels associated with $\mathcal{T}_{Cat}$, and $l(R_i) = |\delta_D(c(R_i))|$ is the number of component densities $g(\cdot|c)$ fit to the pixels in the i-th subregion R_i. Recall that $N = n^2$ is the number of pixels in the image. The parameter $\beta > 0$ is an additional tuning parameter, relating to a discretization of the interval $[0, 1]$ over which the mixing weights π_c vary, some settings for which are given in Sections 3.3 and 5.

Our choice of penalty is informed by the statistical risk analysis described in the theorem of Section 3.3, the derivation of which may be found in the appendix. As such, it is analogous to the penalties in Kolaczyk and Nowak (2004b), for a class of multiscale nonparametric generalized linear models based on piecewise polynomials, or more generally it is in a spirit similar to those of Birgé and Massart (1998), for minimum contrast estimation on sieves. Note that the penalty in (3) summarizes the complexity of spatial and categorical scale in the model $\mathcal{M}$ in a simple, straightforward manner. Spatial complexity is assessed through counting of the number of subregions m defined by $\mathcal{P}$, and has an effect both alone in the first term and through its interaction with the effects of categorical complexity in the second two terms. The effects of categorical complexity are determined both through the number of component densities $l(R)$ fit to each subregion R and the overall size $L_{\mathcal{C}}$ of the categorical tree $\mathcal{T}_{Cat}$.

From a Bayesian perspective, it is evident that in (2) the penalty plays the role of the logarithm of a prior distribution on the collection of mixlet models $\mathcal{M}$, and hence $\hat{\mathcal{M}}$ is analogous to the corresponding maximum a posteriori (MAP) estimator. Thus at a minimum, the function pen$(\cdot)$ should incorporate expert knowledge of the manner in which complexity of spatial resolution and categorical granularity interact. More generally, information on such aspects as the geometry or morphology of subregions might be included. However, for many areas of application these expectations remain overly ambitious. For example, within the field of remote sensing land cover characterization, the development of a sufficiently sophisticated formal language (ontology) for describing such information is an ongoing effort. Therefore, as a result, penalties like that in (3), guided primarily by a combination of statistical and information theoretic considerations, are appealing. Additional discussion of this issue may be found in Section 6.

Calculation of the optimization in (2) may be done using standard algorithms for mixture models within the framework of a traditional bottom-up tree pruning algorithm, in which optimal sub-models from successively coarser spatial resolutions are compared in a recursive fashion. Beginning at the finest spatial resolution (i.e., at the base of the quad-tree corresponding to $\mathcal{P}_n^*$), fitting of an optimal mixlet sub-model simply corresponds to selecting that class c for each pixel I with the largest log-likelihood among those of the finest granularity (i.e., at the base of the categorical tree $\mathcal{T}_{Cat}$). Next, for each quad of four pixels, we compare the complexity-penalized likelihood of two particular sub-models: (i) the union of the four most likely single-pixel models, with penalty $(16/3)\log 2$ for a single spatial split (i.e., of the quad into four pixels) plus $4 \log L_{\mathcal{C}}$ for using four separate labels, and (ii) that single model for the set of four pixels that is most likely among all allowable $c \in \mathcal{C}(\mathcal{T}_{Cat})$, with penalty $(4/3)\log 2$ for a single spatial region, plus $\log L_{\mathcal{C}}$ for using one single label, plus possibly a third term if a mixture is used (i.e., with $l(R) > 1$). This process then continues in a recursive fashion, at each scale comparing the union of the optimal sub-models in the four sub-quads of a quad with that best model for the pixels in the quad as a whole.

More concretely, consider the example of the class of mixlet models corresponding to the land cover characterization in Figure 1(b), as described immediately following Definition 1. There $\mathcal{C}(\mathcal{T}_{Cat}) = \{\text{C,H,F}\}$, $L_{\mathcal{C}} = 3$, and $l(R) = 1$ for regions R in which the label $c(R) =$C or H, or $l(R) = 2$ otherwise. Write $\boldsymbol{x}_R$ for the set of measurements x_{i_1,i_2} of pixels $I_{i_1,i_2} \subseteq R$, and let $\ell(\boldsymbol{x}_R|c)$ be the log-likelihood under the model that the elements of $\boldsymbol{x}_R$ are drawn from $g(x|c)$. Let CPL_R^{opt} denote the complexity penalized likelihood of the optimal submodel on region R. Finally, let $ch(R)$ be a set consisting of the subregions $R_1, \ldots, R_4$ (i.e., 'children') resulting from the splitting of R into four quads. The recursive nature of the algorithm can be captured by writing $CPL_R^{opt} = \max\{CPL_R^{split}, CPL_R^{merge}\}$, where $CPL_R^{split} = \sum_{R_i \in ch(R)} CPL_{R_i}^{opt}$ and CPL_R^{merge} is the maximum over $c \in \mathcal{C}(\mathcal{T}_{Cat})$ of $\ell(\boldsymbol{x}_R|c) - 2[(4/3)\log 2 + \log 3 + 1_{\mathrm{F}}\, \beta \log N]$, with $1_{\mathrm{F}} = 1$ if $c =$ F and 0 otherwise. The recursion is initialized by starting at the finest spatial scale, where R ranges over the set of pixels $\{I_{i_1,i_2}\}$, and setting $CPL_{I_{i_2,i_2}}^{opt} = CPL_{I_{i_1,i_2}}^{merge}$.

Upon completion of the algorithm, one is left with a single optimal model for the entire spatial region $[0,1]^2$, which is simply $\hat{\mathcal{M}}$. As there are $\mathcal{O}(N)$ nodes on the quad-tree corresponding to an $n \times n$ image, where $N = n^2$, there are only $\mathcal{O}(N)$ such comparison stages needed in finding $\hat{\mathcal{M}}$. Within each comparison stage, comparisons are made among no more than $L_{\mathcal{C}}$ various mixture models, the parameters for each of which may be fit using a standard EM algorithm (e.g., McLachlan and Peel 2000). This general type of bottom-up, tree pruning algorithm is by now fairly common in the literature on multiscale methods and is possible here primarily due to the additivity of both the log-likelihood and our penalty function in the coordinates of the underlying quad-tree. The reader is referred to Kolaczyk and Nowak (2004b), for example, or the seminal paper of Donoho (1997), for more detailed discussions on such algorithms.

3.3 Theoretical Evaluation of $\hat{\mathcal{M}}$: Horizon Model

The 'model of boundary fragments' of Korostelev and Tsybakov (1993), more recently termed the 'horizon model' by Donoho (1999), is a standard theoretical framework by which to evaluate the properties of proposed image processing methodologies. We describe now a natural extension of this framework to the context of this paper, and show that the adaptively chosen mixlet model $\hat{\mathcal{M}}$ possesses properties analogous to those of the type of near-optimal, multiscale methods designed for the standard setting.

A horizon model consists of a partitioning of the domain $[0,1]^2$ into two compact regions, say G and $\bar{G}$, separated by a boundary ∂G of some requisite smoothness, and a function or density with two distinct pieces, on G and $\bar{G}$, respectively, that defines the underlying structure in the image. As such, it is in some sense the simplest nontrivial image structure one might face. Similarly, in more complex images, it can be viewed as arising possibly after sufficient magnification of some local area around a boundary (hence the alternative name 'model of boundary fragments'). We will define a sampling model for the measurements $\boldsymbol{x}$ in this setting through the specification of a particular density $f(x)$. We then provide a quantification of the degree to which our mixlet model and model selection methodology can successfully learn this sampling model, by comparing the density $f_{\hat{\mathcal{M}}}$ of an adaptively selected optimal mixlet model $\hat{\mathcal{M}}$ to f in a precise manner (to be defined below). Since our overall image segmentation framework, described in Section 4, derives from appropriate post-processing of the components of $f_{\hat{\mathcal{M}}}$, our results provide assurance as to the degree to which we are post-processing with an object near to the unknown truth (i.e., f).

We begin our definition of f by assuming that $G \in \mathcal{G}$, where

$$\mathcal{G} \equiv \{ (t_1, t_2) \in [0,1]^2 : t_1 \in [0,1],\, 0 \le t_2 \le h(t_1),\, h \in \mathrm{Lip}_A[0,1] \} \tag{4}$$

and $\mathrm{Lip}_A[0,1]$ denotes the set of Lipschitz functions on the unit interval i.e., $|h(u) - h(v)| \le A|u - v|, \forall u, v \in [0,1]$ and some $A > 0$. Next we define $\pi : [0,1]^2 \to [0,1]^{L_D}$ to be a function of L_D components on the unit square such that (i) each component $\pi_c(\cdot)$ is constant on each of the regions G and $\bar{G}$, (ii) π differs in at least one of its components between the two regions, and (iii) $\pi_1(t) + \cdots + \pi_{L_D}(t) = 1, \forall t \in [0,1]^2$. Lastly, we will suppose that the measurements x_{i_1,i_2} are drawn independently according to the density

$$f(x) = \sum_{c=1}^{L_D} \pi_c(I_{i_1,i_2}) \times g(x|c) \ \ , \tag{5}$$

where $\pi_c(I_{i_1,i_2})$ is obtained by averaging the c-th component of the function π over the pixel I_{i_1,i_2}. Let $\mathcal{F}$ be the collection of all horizon model densities (5).

Note that under our model class $\mathcal{F}$, the sampling on the two subregions G and $\bar{G}$ may be done with respect to densities f with any combination of the L_D component densities $g(x|c)$. Therefore, as could be expected in reality, not only may this sampling have been done with respect to any of the combinations dictated by the user-specified tree $\mathcal{T}_{Cat}$, but it may also have been done with respect to combinations not anticipated by the user in specifying $\mathcal{T}_{Cat}$. However, the set of L_D component densities specified by the user are assumed to contain that set underlying the true state of nature. In the case where the combinations of components used on G and/or $\bar{G}$ are not among those specified in $\mathcal{T}_{Cat}$, such combinations can be represented in our modeling framework by models $\mathcal{M}$ using the finest scale category $c \in \mathcal{C}(\mathcal{T}_{Cat})$ that still contains all of the components. For example, consider the tree $\mathcal{T}_{Cat}$ in Figure 2(b), and suppose that the subregion G consists of the specific classes *conifer, hardwood,* and *grass* in proportions of 1/3 each. Then although there are no elements of *brush* in the image, the coarse class *vegetation* must be used to include all three of the other classes, and any such models will penalize the usage of the *vegetation* class with $l(R) = 4$ in (3), even those assigning a weight of zero to the class *brush.*

We now provide an evaluation of the quality of the adaptively selected mixlet model $\hat{\mathcal{M}}$ through an analysis of its statistical risk, where loss is defined through squared Hellinger distance i.e., $H^2(p,q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$ for any two densities p and q, and the risk in estimating f by an estimator $\hat{f}$ is quantified as $Risk(f,\hat{f}) = (1/N)E[H^2(f,\hat{f})]$.

Theorem 1 *Suppose $\boldsymbol{x} \sim f$, for $f \in \mathcal{F}$. Define $L_{G,\bar{G}} = L_{G,\bar{G}}(f)$ to be the total number of true component densities $g(x|c)$ corresponding to G and $\bar{G}$ under f, and let $L'_{G,\bar{G}} = L'_{G,\bar{G}}(f,\mathcal{T}_{Cat}) \geq L_{G,\bar{G}}$ denote the total number of components among the finest scale categories $c(G), c(\bar{G}) \in \mathcal{T}_{Cat}$ that contain these true components. Let $f_{\mathcal{M}}$ be the density corresponding to a mixlet model $\mathcal{M}$, as in (1). Consider the model $\hat{\mathcal{M}} \equiv \hat{\mathcal{M}}(\boldsymbol{x})$ chosen with respect to the* *complexity-penalized maximum likelihood criterion in (2), where the optimization is over a collection $\mathcal{M}_N^*$ produced through discretization of mixing weights to an accuracy of $N^{-\beta}$, for $\beta = 1/2$ (as described in the appendix). Finally, suppose that the breadth of the categorical tree $\mathcal{T}_{Cat}$ at its base is constrained so that $L_D \leq \mathcal{O}(\log N)$, and that the component densities g satisfy the condition $sup_{c,c',x} g(x|c)/g(x|c') \leq B$ for some $B < \infty$. Then*

$$Risk(f, f_{\hat{\mathcal{M}}}) \ \leq \ L'_{G,\bar{G}} \times \mathcal{O}\left(\left(N^{-1}\log N\right)^{1/2} \right) \tag{6}$$

for each $f \in \mathcal{F}$, and

$$\sup_{f\in\mathcal{F}} Risk(f, f_{\hat{\mathcal{M}}}) \ \leq \ \mathcal{O}\left(\left(N^{-1}\log^3 N\right)^{1/2} \right) \quad . \tag{7}$$

Proof of this theorem may be found in the appendix, while here we confine ourselves to a few relevant comments. First, we note that the rate $\mathcal{O}\left((N^{-1}\log N)^{1/2} \right)$ is the sort of rate that may be obtained by adaptive, multiscale estimators in the standard horizon model (e.g., Donoho 1999), for which it is known that the minimax optimal rate is $\mathcal{O}\left(N^{-1/2}\right)$. The bound in (6) shows that for any given $f \in \mathcal{F}$ we obtain essentially the same rate. Second, whereas in the standard horizon model the rate $\mathcal{O}\left((N^{-1}\log N)^{1/2} \right)$ continues to hold over the entire class $\mathcal{F}$, here this rate is increased by a factor of $\log N$, as indicated in (7). This increase may be seen as the price paid for not knowing the number of component densities g used by f on G and $\bar{G}$. Finally, we note that it would seem that the constraint $L_D \leq \mathcal{O}(\log N)$ is necessary not only for (7), of course, but also for (6). This constraint can be interpreted as enforcing identifiability of any mixture model possible under $\mathcal{T}_{Cat}$ at the spatial resolution optimal for isolating the boundary ∂G in a parsimonious manner, which in doing so uses subregions R containing $n(R) = \mathcal{O}(\log N)$ pixels.

4 Image Segmentation Using Mixlets

The modeling and model selection framework described in Section 3 yields an adaptively selected model $\hat{\mathcal{M}} = \{\hat{\mathcal{P}}, \hat{\mathbf{c}}(\hat{\mathcal{P}}), \hat{\boldsymbol{\pi}}(\hat{\mathcal{P}})\}$ that can be expected to capture well both the boundaries and the content of subregions in an image, according to a Hellinger-based measure of loss. A corresponding segmentation can be produced simply through the reporting of the RDP $\hat{\mathcal{P}}$ and its labels $\hat{\mathbf{c}}(\hat{\mathcal{P}})$. By design, such a segmentation will in principle accomplish our goal of simultaneously incorporating subregions at multiple scales and labels at multiple granularities, chosen in such a way as to balance their relative complexities. However, the visual quality of this segmentation will be comparatively crude, as is to be expected with the use of a single RDP. Therefore, we re-cast our basic framework within a redundant formulation, using a system of multiple, interwoven mixlet models, and describe the corresponding algorithmic implications. A multiscale, multigranular segmentation of an image may then be produced from such a system by extracting and combining the information on subregions and class labels from each of the component mixlet models, some methods for which we describe.

4.1 Redundant Systems of Mixlets

In creating a modeling framework that is inherently multiscale in nature, we have made the not-uncommon choice to use recursive dyadic partitions. Despite the obvious advantages and convenience of this choice, however, there are also some well-known artifacts accompanying it that are less desirable. For example, segmentation subregions defined through $\hat{\mathcal{M}}$ will tend to be 'blocky' in shape, not surprisingly, and the overall modeling framework is not invariant to simple translations of the underlying scene. In the literature on multiscale methods, a standard solution for eliminating such artifacts is equivalent to replacing the use of a single C-RDP $\mathcal{P}_n^*$ with a redundant system of multiple C-RDPs. This paradigm is often used in the wavelet literature, where it is sometimes called 'cycle-spinning' and can be shown to correspond to the use of a redundant system of wavelets (e.g., Coifman and Donoho 1995) or, alternatively, can be interpreted as a type of model averaging (e.g., Kolaczyk 1999). In our context, given that we are dealing with partitions and not wavelets *per se,* the details of the implementation differ slightly, although the spirit is the same. Therefore, we simply sketch the most relevant aspects below.

Begin by considering the process of shifting the rows and columns of an image. Specifically, by an s-step periodic row (column) shift of an image we will mean the process by which all rows (columns) are shifted down (right) s steps, and the s rows (columns) that 'fall off' the bottom (right-hand side) of the image are placed at the top (left-hand side) in their original order. There are a total of n^2 unique combinations (s_r, s_c) of such row-column shifts for an $n \times n$ image, each of which may be associated with a different C-RDP. In principle, we can envision performing the following three-step procedure for every one of these row-column shifts: (i) shift the original image by s_r rows and s_c columns; (ii) calculate an optimal mixlet model, say $\hat{\mathcal{M}}^{(s_r,s_c)}$ for the shifted image; (iii) un-shift the information in $\hat{\mathcal{M}}^{(s_r,s_c)}$ to the coordinates of the original image. The information in the $\hat{\mathcal{M}}^{(s_r,s_c)}$ can then be extracted and combined to produce an image segmentation that is translation-invariant and defined by relatively smooth boundaries, as we describe in a moment in Section 4.2.

Explicit implementation of this procedure, as defined, would be prohibitive. Fortunately, it may be implemented in a manner more efficient than that suggested by its description, by exploiting redundancies inherent among the C-RDPs associated with each of the various shifts (s_r, s_c). In particular, if we consider square subregions R of size $n(R) = 2^j \times 2^j$, for fixed $j = 0, \ldots, J-1$, where $J = \log_2(n)$, then it is easy to see that there are only n^2 unique subregions of this size across the entire collection of row-column shifted images for each j, and therefore only $n^2 \log_2(n)$ unique subregions in total. Hence, among the $N = n^2$ models $\hat{\mathcal{M}}^{(s_r,s_c)}$ there are only $\mathcal{O}(N \log N)$ unique sets of labels $\hat{\mathbf{c}}^{(s_r,s_c)}$ and mixing weights $\hat{\boldsymbol{\pi}}^{(s_r,s_c)}$ to be calculated. And following reasoning analogous to that at the end of Section 3.2 in the case of a single C-RDP, it is clear that such calculations may be done in a scale-recursive fashion, proceeding from fine spatial scales to coarse. This recursion will require $\mathcal{O}(N \log N)$ comparison stages, within each of which a set of four optimal sub-sub-models will be compared to the optimal sub-model on a given quad. The storage requirements scale similarly. Overall this figure represents a vast improvement on the apparent $\mathcal{O}(N^2)$ figure accompanying the original description of this procedure above, and compares favorably with the $\mathcal{O}(N)$ figure for a single C-RDP. The relevant implementation details are straightforward, although tedious in terms of 'book-keeping', and therefore omitted here. The corresponding code is available at *http://math.bu.edu/people/kolaczk/software.html.*

4.2 Extracting an Image Segmentation

To produce an image segmentation in our redundant framework we must first extract and combine the relevant information in each of the n^2 shift-optimal mixlet models $\hat{\mathcal{M}}^{(s_r,s_c)} = \left(\hat{\mathcal{P}}^{(s_r,s_c)}, \hat{\mathbf{c}}^{(s_r,s_c)}, \hat{\boldsymbol{\pi}}^{(s_r,s_c)}\right)$. Note that while the (un)shifting in our framework renders the partitions $\hat{\mathcal{P}}^{(s_r,s_c)}$ less meaningful, their role is still implicit in producing the class labels $\hat{\mathbf{c}}^{(s_r,s_c)}$, which remain informative on a pixel-wise basis. For each pixel I_{i_1,i_2} we thus end up with n^2 labels, $\{\hat{c}_{i_1,i_2}^{(s_r,s_c)}\}_{s_r,s_c=1}^{n}$, from which a single set of class labels $\{\hat{c}_{i_1,i_2}\}_{i_1,i_2=1}^{n}$ for the image may be produced through application of a pixel-wise majority-vote or similar rule. Upon assigning each label an (artificial) numerical value, a segmentation of the image can be derived from the corresponding level sets.

However, we note that there is additional information in the $\hat{\mathcal{M}}^{(s_r,s_c)}$ that is not being used in this approach, in the form of the vectors of weights $\{\hat{\boldsymbol{\pi}}_{i_1,i_2}^{(s_r,s_c)}\}$. In analogy to the class labels, these vectors may be combined on a pixel-by-pixel basis through simple averaging, for example, to produce a single set of pixel-wise weight vectors $\{\hat{\boldsymbol{\pi}}_{i_1,i_2}\}_{i_1,i_2=1}^{n}$. The information in these pixel-wise averages can then be used to further refine the segmentation produced from the original class labels, if desired. For example, consider the case of a land cover image composed almost entirely of pixels of a single pure class, such as *conifer*, with only some small fraction of some other class present, such as *hardwood.* Then if the component densities g for *conifer* and *hardwood* are sufficiently well-separated in their supports, the presence of contaminating pixels of *hardwood* can be expected eventually to force the algorithm to use subregions R containing those pixels that are labeled $\hat{c}(R) =$ *forest,* regardless of the degree of contamination, its spatial extent and pattern, or the character of the larger region containing it – all of which would be important criteria to an expert in deciding whether a label of *forest* were truly appropriate. This effect has been observed to various degrees in our empirical experience, but can be modified by incorporating information in the weight vectors $\hat{\boldsymbol{\pi}}_{i_1,i_2}$.

For the numerical results in Section 5 we have used a simple, greedy, coarse-to-fine algorithm over the categories in the tree $\mathcal{T}_{Cat}$, applied on a pixel-by-pixel basis. More precisely, for each $c \in \mathcal{C}(\mathcal{T}_{Cat})$ let $\pi_c \equiv \sum_{c' \in \delta_D(c)} \pi_{c'}$, let $\alpha(c)$ and $\delta(c)$ denote the immediate ancestor and descendants of c, respectively, and let $\epsilon_1, \epsilon_2 \in [0,1]$. Then to each pixel I_{i_1,i_2}

we assign the first label c encountered, in moving from the inferred label $\hat{c}_{i_1,i_2}$ toward the leaves of $\mathcal{T}_{Cat}$, for which the following criteria are satisfied: (i) $\hat{\pi}_c > \epsilon_1$, (ii) $\hat{\pi}_c/\hat{\pi}_{\alpha(c)} > \epsilon_2$, and (iii) the same cannot be said of any of the immediate descendants $c' \in \delta(c)$. In words, we allow for the refinement of the label $\hat{c}_{i_1,i_2}$ assigned to I_{i_1,i_2} originally, by seeking the label of that descendant class c of as fine a granularity as possible that both contains a significant fraction of the overall mass in $\hat{\boldsymbol{\pi}}_{i_1,i_2}$ and for which that mass is shared sufficiently among its immediate descendants. We have found the values $\epsilon_1 = 0.5$ and $\epsilon_2 = 0.75$ to work well in practice. Our final segmentation is then defined as above, through the level sets of the collection, say $\{\hat{\hat{c}}_{i_1,i_2}\}$ that result from this post-processing procedure.

On a final note, we point out that the core of our framework for image segmentation is the class of mixlet models, which may be used to produce fitted data objects with information directly relevant to the task of segmentation using multiple spatial and categorical scales. The methods that we have presented for extracting and using such information are the most obvious, and chosen partially for their illustrative value. It is clearly possible to envision additional methods of increasing sophistication and complexity.

5 Empirical Illustration: Cartographic Generalization

In the field of Geography, *cartographic generalization* refers to the principle of reducing and filtering the geographical information in a set of data to a particular scale and theme, typically for the purpose of producing a map or similar such visual representation (Buttenfield and McMaster 1991). The characterization of land cover from remotely sensed measurements is therefore a particular instance of cartographic generalization. The process behind cartographic generalization is generally complex and time consuming, even with the now-established role of geographical information systems (GIS), and it is often also rather subjective. The mixlet model framework proposed in this paper can be used to capture, in an objective and computationally efficient manner, information relevant to the characterization of land cover, at multiple spatial scales and categorical granularities. Visual representations of this information may then be produced by summarizing this information, such as through the segmentation procedures just described, in accordance with user-defined preferences and priorities.

Figure 3(a) shows a red-green-blue (RGB) composite image representation of remote sensing measurements taken by the Landsat TM satellite instruments in the early 1990's for a certain portion of the Sierra region of central California. The Landsat instrument resolution is 30 meters by 30 meters. The 128×128 image shown in this figure corresponds to an area of roughly 14.7 square kilometers that has been studied by a number of authors in the remote sensing literature (e.g., Franklin, Woodcock, and Warbington 2000). Measurements $\boldsymbol{x}$ are based on aggregate spectral reflectances in six spectral bands (i.e., 0.45-0.53 μm, 0.52-0.60 μm, 0.63-0.69 μm, 0.76-0.90 μm, 1.55-1.75 μm, and 2.08-2.35 μm), registered and resampled with respect to our pixelization according to standard pre-processing procedures. The image itself is created by mapping the second, third, and fourth bands to the colors blue, green, and red, respectively, which has the effect of highlighting dense vegetation.

A regular consumer of this type of data is the United States Forest Service (USFS), whose mandate includes balancing such goals as the responsible harvesting of timber and the effective management of wildlife habitats in national parks and preserves. Different land cover is relevant to these goals to different degrees. For example, the harvesting of timber revolves primarily around regions of conifer trees, such as ponderosa pine, Jeffrey pine, and red/white fir (Woodcock *et al.* 1994). Conversely, the management of habitats for a species such as the spotted owl, upon which much attention and resources have been focused in the past decades, centers mainly on stands of hardwood trees or mixed stands of hardwood and conifer trees (Franklin and Stephenson 1996). Hence, land cover characterizations that clearly delineate regions of each type, based on the relatively cheap and accessible measurements gained through remote sensing, are useful in balancing these two tasks. However, complicating matters is the fact that in this region of the United States there are inter-woven with stands of conifer and hardwood large quantities of relatively open land covered with little more than brush (e.g., chaparral), which is relatively useless with respect to either the harvesting of conifer timber or the management of hardwood stands. And, unfortunately, the spectral signature of pixels dominated by such brush tends to be quite similar to that of pixels dominated by hardwood; therefore, 'brush' is a nuisance both scientifically and statistically.

Figures 3(b) and (c) show two possible characterizations of the land cover in the spatial

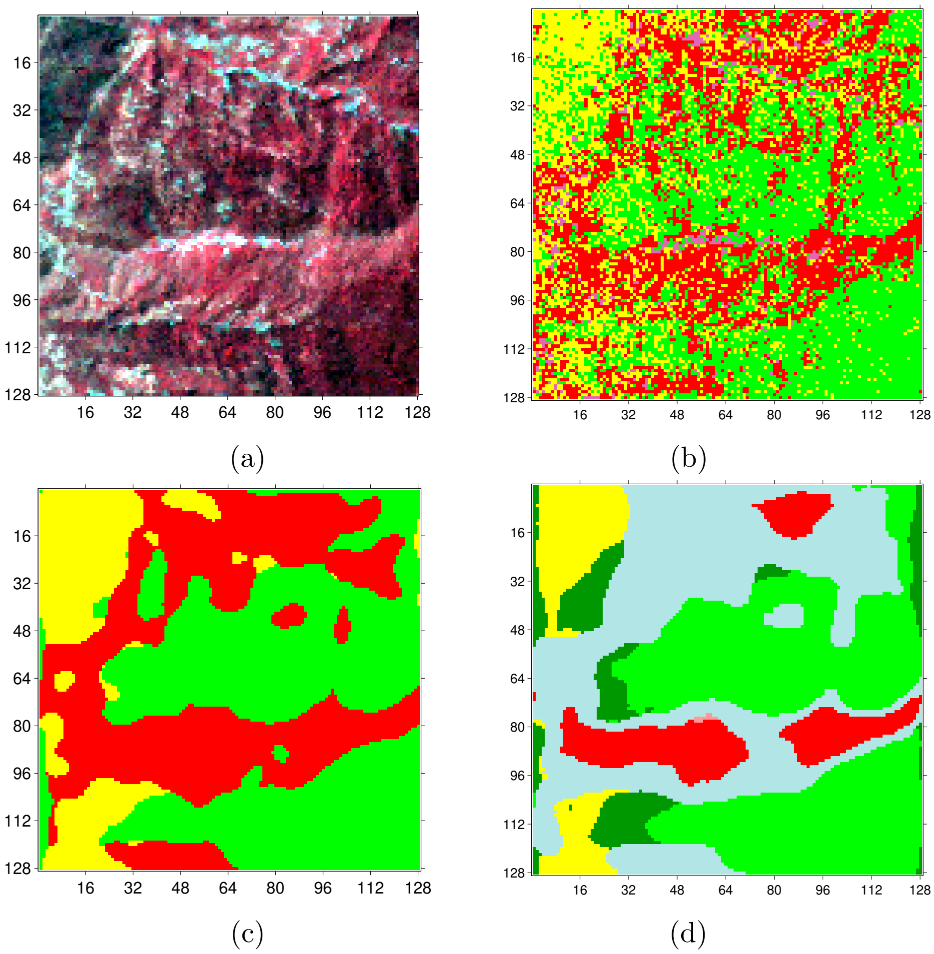


Figure 3: (a) composite (RGB) image of Landsat TM measurements for Sierra region in California; (b) pixel-wise classification based on maximum likelihood; (c) smoothed version of pixel-wise classification; (d) land cover characterization produced by multiscale, multi-granular image segmentation. In (b), (c), and (d), the colors correspond to categories as follows: light green=conifer, yellow=hardwood, red=brush, purple=grass; dark green=forest, pink=non-forest; blue=vegetation.

region underlying the measurements in Figure 3(a). Class-specific densities $g(x|c)$ were trained, for the classes *conifer, hardwood, brush,* and *grass*, as a mixture of Gaussian densities for each, following the procedures described in Ju, Kolaczyk, and Gopal (2003). A standard maximum-likelihood classifier was then applied, based on these densities, on a pixel-by-pixel basis, to produce the characterization in Figure 3(b). This image is representative of raw, baseline results in remote sensing land cover characterization. These are then often further refined by 'smoothing' the map of inferred pixel labels through the application of various standard post-processing tools. In Figure 3(c) is shown the result of one such technique, in which we (i) calculate the fraction of pixels labeled as each land cover class in each 8×8 square; (ii) compute for each pixel the average fraction for each land cover class over all 8×8 squares containing that pixel; and (iii) replace the original pixel label by that label corresponding to the largest average fraction.

Note that both of these land cover characterizations are monogranular, as is standard, and monoscale as well, with two different choices of scale. In considering Figure 3(b), the results suggest the presence of broad swaths of conifer in two horizontal bands across the middle and lower portions of the image region. Similarly, there seems to be evidence of a moderately uniform region of hardwood in the upper left-hand corner. As for the rest of the landscape, it would appear to be rather fragmented, with various degrees of conifer, hardwood, and brush combining throughout, with a hint of grass here and there. In contrast, with respect to the smoothing used in producing Figure 3(c), while it maintains the labeling of conifer and hardwood in the uniform regions just mentioned, it imposes a similarly uniform choice of labels in regions that previously appeared fragmented. For example, the *brush* class is now used extensively throughout much of the image, and a new region of hardwood is determined to be present in the lower left-hand corner.

However, to the degree that the results in Figure 3(b) may be taken as indicative of truth at the finest scale, there is something artificial in relabeling fragmented regions with whatever pure class happens to be most prevalent. On the other hand, the graininess of the representation at the finest scale is less than appealing as well, particularly from the perspective of cartographic generalization, which seeks a concise reduction of the information to the most relevant elements. Figure 3(d) shows the results of applying to the data our framework for multiscale, multigranular image segmentation. Specifically, we applied the estimator (2), using the penalty in (3) with $\beta = 1/8$, and post-processed the results using the method of Section 4.2, within the redundant framework described in Section 4.1. The corresponding categorical hierarchy is as shown in Figure 2(b), allowing for the aggregation of the *conifer* and *hardwood* classes into a single *forest* class, the aggregation of the *brush* and *grass* classes into a *non-forest* class, and the aggregation of all four pure classes into a single *vegetation* class. A few points are interesting to note. First, the relatively pure regions of conifer and hardwood suggested by the results in Figure 3(b) are still clearly evident.

Second, there now is confirmation of the previous impression of only a moderate degree of uniformity in the hardwood region of the upper left-hand corner, as indicated by the appearance of a subregion of *forest* class in that area. Third, the majority of the remainder of the image region, which came across as being fragmented in Figure 3(b) and was declared to be filled uniformly with brush in Figure 3(c), is represented now quite simply using broad stretches of the *vegetation* class. For the purposes of timber harvesting and habitat management described above, the characterization in Figure 3(d) would seem to successfully present the relevant information on land cover in a manner that is simultaneously spatially and categorically concise.

Additional numerical evaluations of the behavior of our multiscale, multigranular image segmentation procedure, using various artificial landscapes, may be found in a companion paper (Ju, Gopal, and Kolaczyk 2005). These evaluations further confirm the ability of our method to adaptively select both spatial resolution and categorical granularity.

6 Discussion

In this paper, we have formulated a particular variation of the image segmentation problem and have offered what is intended to be an initial, prototype solution. Our framework is mathematically intuitive and computationally feasible, and is demonstrated both to have desirable theoretical properties and to produce empirical results of practical relevance. However, the general challenge of incorporating and evaluating information at multiple spatial and categorical scales in an image in a quantitatively formal and precise fashion seems one largely unaddressed in the literature, despite its long-standing visibility at a qualitative level in certain disciplines. As a result, there would seem to be ample opportunity for further contributions from a number of different technical communities, including those in statistics and mathematics, image processing, and computer vision.

With respect to our own work in this paper, the development of a framework for incorporating substantive prior information on the spatial and categorical scale of image content in a formal manner, as well as using a proper class of probability distributions defined over the set of mixlet models $\mathcal{M}$ or similar, as compared to the penalty in (3), should allow for improved segmentations. Extensions of the composition systems of Stuart Geman and colleagues (e.g., Geman, Potter, and Chi 2002), or related work of Bouman and colleagues (e.g., Pollak, Siskind, Harper, and Bouman 2003), which merge multiscale modeling structures (e.g., recursive partitions) with probabilistic grammars on hierarchically defined object sets, would seem a promising direction. Work of Patil and colleagues also is likely to be relevant in this context (e.g., Patil and Taille 1999). A corresponding segmentation could then be produced based on the posterior $\Pr(\mathcal{M} \,|\, \boldsymbol{x})$.

Similarly, the issue of accuracy of segmentations produced in multiscale, multigranular contexts bears examining. That is, the introduction of a hierarchy of possible labels, in conjunction with multiple spatial resolutions, as captured by the two hierarchical data structures in Figure 2, allows for the possibility of multiple representations of ostensibly equal 'accuracy', as illustrated in Figure 1. Our approach in this paper to distinguishing between competing representations of an image scene in this context is to adopt the perspective of model selection and to penalize unnecessary excess in either spatial or categorical complexity. Yet this approach does not provide any quantification of the level of accuracy of the final product. The risk theory accompanying the estimator $\hat{\mathcal{M}}$, described in Section 3.3, is arguably only an indirect indication of the accuracy one may expect from the segmentation framework as a whole. Extensions of some of the current measures of accuracy assessment in the literature on remote sensing land cover characterization would be of interest in this setting. These measures typically involve simple aggregate summaries, such as conditional frequencies capturing so-called "user's" and "producer's" accuracy (see Congalton 1991, for example), of pixel-wise classification accuracy with respect to pure class labels, and therefore by definition are only monoscale, monogranular assessment tools at the finest spatial and categorical scales.

7 Appendix

Here we sketch the proof of Theorem 1. We begin with the observation that, by Theorem 7 of Kolaczyk and Nowak (2004a), for estimators of the form given in (2) we have

$$Risk(f, f_{\hat{\mathcal{M}}}) \leq \min_{\mathcal{M} \in \mathcal{M}_N^*} \{ (1/N)\, K(f, f_{\mathcal{M}}) + (2/N)\ \mathrm{pen}(\mathcal{M}) \} \quad , \tag{8}$$

where $K(p,q) = \int p(x) \log\left(p(x)/q(x)\right) dx$ is the Kullback-Leibler divergence between densities p and q, $\mathcal{M}_N^*$ is a finite set of models $\mathcal{M}$, and the penalty $\mathrm{pen}(\cdot)$ satisfies the constraint

$$\sum_{\mathcal{M} \in \mathcal{M}_N^*} e^{-\mathrm{pen}(\mathcal{M})} \leq 1 \quad . \tag{9}$$

The collection $\mathcal{M}_N^*$ is produced from the set of all possible mixlet models, for a given categorical tree $\mathcal{T}_{Cat}$ and its associated component densities g, simply by discretizing the unit interval $[0,1]$ (on which the mixing parameters π_c are defined) to a grid of N^β equi-spaced values. Such a step is a natural approximation of actual practice, and is necessary for the argument producing (8). The expression in (9) is essentially the Kraft inequality from information theory (if re-expressed in base 2), and has a fundamental connection with results on coding that may be exploited here to argue that the penalty in (3) satisfies this criterion.

More precisely, (9) is satisfied because $\mathrm{pen}(\mathcal{M})/\log_e 2$ is the code-length for a uniquely decodable binary code over $\mathcal{M} \in \mathcal{M}_N^*$. (See Cover and Thomas 1991, Theorem 5.5.1, for example, for a succinct overview of this topic area.) That this assertion holds true may be argued by construction. Let $\mathcal{M} = \{\mathcal{P}, \boldsymbol{c}(\mathcal{P}), \boldsymbol{\pi}(\mathcal{P})\}$. Any RDP $\mathcal{P}$ of size $m = |\mathcal{P}|$ may be described through its corresponding quad-tree, pruned from that corresponding to $\mathcal{P}_N^*$, by using m ones for the leaf nodes and $(m-1)/3$ zeros for the internal nodes, for a total of $(4m-1)/3$ bits. Furthermore, as it requires $\log_2 L_{\mathcal{C}}$ bits to code the $L_{\mathcal{C}}$ elements of $\mathcal{C}(\mathcal{T}_{Cat})$, a total of $m\, \log_2 L_{\mathcal{C}}$ bits are needed to describe all m of the class labels in $\boldsymbol{c}(\mathcal{P})$. Finally, since there are a total of $l(R) - 1$ free parameters π_c associated with the mixture model in each region $R \in \mathcal{P}$, and it requires $\log_2 N^\beta$ bits to code the N^β elements in the discretization of $[0,1]$, the total number of bits needed to describe the mixing parameters in $\boldsymbol{\pi}(\mathcal{P})$ is obtained by summing the quantities $[l(R) - 1] \log_2 N^\beta$ over the m regions $R \in \mathcal{P}$. Upon adding up these various terms, and switching from base 2 to base e, the form of $\mathrm{pen}(\mathcal{M})$ then follows accordingly.

It therefore remains to bound the behavior of the right-hand side of (8). We will do so by bounding $(1/N)K(f, f_{\mathcal{M}}) + (2/N)\mathrm{pen}(\mathcal{M})$ for a particular choice of $\mathcal{M}$, which will then imply a bound for the minimum of this quantity over $\mathcal{M}_N^*$. Let the subregion $G \in \mathcal{G}$ and the L_D-variate weight function π underlying the true density f in (5) be defined as in Section 3.3. Now consider just a single component function $\pi_c : [0,1]^2 \to [0,1]$ of π that is piecewise constant on the subregions G and $\bar{G}$, and picture covering the boundary ∂G with square subregions R of side-length $\mathcal{O}(1/m)$. Under the condition that the horizon function h is Lipschitz, it is known that such a covering may be obtained through an RDP, say $\tilde{\mathcal{P}}$, of size $\mathcal{O}(m)$ and that the accuracy of the optimal corresponding approximation of π_c by functions piecewise constant on the subregions $R \in \tilde{\mathcal{P}}$, call this $\tilde{\pi}_c$, is $||\pi_c - \tilde{\pi}_c||_{L_1} = \mathcal{O}(m^{-1})$. See Donoho (1999), for example. As this holds for each of the $L_{G,\bar{G}}$ nontrivial π_c, we can approximate π itself to a total accuracy of $L_{G,\bar{G}}\,\mathcal{O}(m^{-1})$. Next let the values $\tilde{\pi}_c(I_{i_1,i_2})$ be the result of average-sampling the functions $\tilde{\pi}_c$ on the pixels I_{i_1,i_2}, and let the quantities $[\tilde{\pi}_c(I_{i_1,i_2})]$ be the result of quantizing these values to the N^{β}-point discretization of $[0,1]$ underlying our definition of $\mathcal{M}_N^*$. This quantization can be done so as to produce non-negative weights summing to one on each pixel in a straightforward manner, for the details of which we refer the reader to Section 5.3 of Kolaczyk and Nowak (2004a). Finally, let $\tilde{f}$ be the density function of the mixlet model obtained from this process.

By independence of the x_{i_1,i_2}, the Kullback-Leibler divergence $K(f, \tilde{f})$ is simply the sum of Kullback-Leibler divergences $K(f^{(i_1,i_2)}, \tilde{f}^{(i_1,i_2)})$ at each pixel, and so we consider these terms individually for the moment. Given the condition of boundedness of the ratios of the component densities g, the same holds true for any mixture densities formed from these as convex combinations, and so we may bound each $K(f^{(i_1,i_2)}, \tilde{f}^{(i_1,i_2)})$ by $(2 + \log(B))\, H^2(f^{(i_1,i_2)}, \tilde{f}^{(i_1,i_2)})$, using a result of Yang and Barron (1998). But we also have that

$$H^2(f^{(i_1,i_2)}, \tilde{f}^{(i_1,i_2)}) \;\le\; \int \left| f^{(i_1,i_2)}(x) - \tilde{f}^{(i_1,i_2)}(x) \right| dx \;\le\; \sum_{c=1}^{L_D} \left| \, \pi_c(I_{i_1,i_2}) - [\tilde{\pi}_c(I_{i,i_2})] \, \right| \quad , \qquad (10)$$

where the first inequality is well-known and the second follows from straightforward calculations. Here the $\pi_c(I_{i_1,i_2})$ are as in model (5), produced through average sampling on the pixels I_{i_1,i_2} of the true underlying weight function π. As a result of these observations, we find that we can bound $(1/N)K(f,\tilde{f})$ by the constant $2+\log(B)$ times

$$\begin{aligned}
& (1/N)\sum_{i_1,i_2}\sum_{c=1}^{L_D} |\,\pi_c(I_{i_1,i_2}) - [\tilde{\pi}_c(I_{i,i_2})]\,| \\
\leq\ & (1/N)\sum_{i_1,i_2}\sum_{c=1}^{L_D} |\,\pi_c(I_{i_1,i_2}) - \tilde{\pi}_c(I_{i,i_2})\,| \ +\ (1/N)\sum_{i_1,i_2}\sum_{c=1}^{L_D} |\,\tilde{\pi}_c(I_{i_1,i_2}) - [\tilde{\pi}_c(I_{i,i_2})]\,| \\
\leq\ & \sum_{c'\in\delta_D(c(G,\bar{G}))} ||\pi_{c'} - \tilde{\pi}_{c'}||_{L_1} \ +\ \sum_{c'\in\delta_D(c(G,\bar{G}))} \mathcal{O}(N^{-\beta}) \\
\leq\ & L_{G,\bar{G}}\left[\mathcal{O}(m^{-1}) \ +\ \mathcal{O}(N^{-\beta})\right] \ ,
\end{aligned}$$

where we write $\delta_D(c(G,\bar{G}))$ for $\delta_D(c(G)) \cup \delta_D(c(\bar{G}))$.

Note that for the density $\tilde{f}$ to be a proper mixlet model, as defined in Section 3.1, it must obey the constraint $l(R) \leq n(R), \forall R \in \tilde{\mathcal{P}}$. Since $\tilde{\mathcal{P}}$ is constructed to have regions R of side-length at least $\mathcal{O}(1/m)$, this means that for each such region $n(R) \geq \mathcal{O}(n^2/m^2)$. And since $l(R) \leq L_D \leq \mathcal{O}(\log N)$, we are led to restrict our attention to $m \leq \mathcal{O}(\,(N\log^{-1}N)^{1/2})$. To complete our proof, we seek to optimize our construction of the density $\tilde{f}$ over m in this range. Taking into account now the form of the penalty $\mathrm{pen}(\cdot)$ in (3) and the bounds derived above on $(1/N)K(f,\tilde{f})$, it is found that the right-hand side of (8) may be bounded by a quantity whose relevant terms are of the form

$$L_{G,\bar{G}}\,\mathcal{O}(m^{-1}) + L'_{G,\bar{G}}\,\mathcal{O}(mN^{-1}\log N) \leq L'_{G,\bar{G}}\left[\mathcal{O}(m^{-1}) + \mathcal{O}(mN^{-1}\log N)\right] \ , \qquad (11)$$

the right-hand side of which is optimized with $m \sim (N\log^{-1}N)^{1/2}$. Substitution of this choice of m and setting $\beta = 1/2$ then yields the result in (6); that in (7) follows trivially from the uniformity of the $O\left((N^{-1}\log N)^{1/2}\right)$ term in (6) and the constraint $L_D \leq \mathcal{O}(\log N)$.

References.

Anderson, J.R., Hardy, E.E., and Roach, J.T. (1976), " Land Use and Land Cover Classification System for Use with Remote Sensing Data. Geological Survey Professional Paper 964. A revision of the land use classification system as presented in the U.S. Geological 671. U. S. Government Printing Office, Washington, D.C.

Birgé, L. and Massart, P. (1998), "Minimum contrast estimators on sieves: exponential bounds and rates of convergence," *Bernoulli,* 4:3, 329-375.

Bose, S. and O'Sullivan, F. (1997), "A region-based segmentation method for multichannel image data," *Journal of the American Statistical Association,* 92, 92-106.

Bouman, C.A. and Shapiro, L. (1994), "A multiscale random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing,* 3:2, 162-177.

Buttenfield, B.P. and McMaster, R.B., Eds. (1991), *Map Generalization: Making Rules for Knowledge Representation.* London: Longman's, Inc.

Coifman, R.R. and Donoho, D.L. (1995), "Translation-invariant de-noising." In *Wavelets and Statistics,* Antoniadis, A. and Oppenheim, G. (eds.): Springer-Verlag.

Congalton, R.G. (1991), "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment,* 37, 35-46.

Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory.* New York: John Wiley & Sons.

Donoho, D.L. (1997), "CART and Best-Ortho-Basis: a connection," *Annals of Statistics,* 25, 1870-1911.

Donoho, D.L. (1999), "Wedgelets: nearly minimax estimation of edges," *Annals of Statistics,* 27:3, 859–897.

Franklin, J. and Stephenson, J. (1996), "Integrating GIS and remote sensing to produce regional vegetation databases: attributes related to environmental modeling," *Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling,* Santa Fe, New Mexico.

Franklin, J., Woodcock, C.E., and Warbington, R. (2000). "Multi-attribute vegetation maps of forest service lands in California supporting resource management decisions," *Pho-*

togrammetric Engineering and Remote Sensing, 66, 1209-1217.

Geman, S., Potter, D., and Chi, Z. (2002), "Composition systems," *Quarterly of Applied Mathematics,* LX, 707-736.

Gotway, C.A. and Young, L.J. (2002), "Combining incompatible spatial data," *Journal of the American Statistical Association,* 97, 632 - 648.

Haralick, R. and Shapiro, L. (1985), "Image segmentation techniques," *Computer Vision, Graphics and Image Processing,* 29, 100-132.

Hastie, T. and Tibshirani, R. (1996), "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society, Series B,* 58, 155–176.

Ju, J., Kolaczyk, E.D., and Gopal, S. (2003), "Gaussian mixture discriminant analysis and sub-pixel land cover classification in remote sensing," *Remote Sensing of Environment,* 84(4),550-560.

Ju, J., Gopal, S., and Kolaczyk, E.D. (2005), "On the choice of spatial and categorical scale in remote sensing land cover characterization," *Remote Sensing of Environment,* (to appear).

Kolaczyk, E.D. (1999), "Bayesian multi-scale models for Poisson processes," *Journal of the American Statistical Association,* 94, 920-933.

Kolaczyk, E.D. and Nowak, R.D. (2004a), "Multiscale likelihood analysis and complexity-penalized estimation," *Annals of Statistics,* 32, 500-527.

Kolaczyk, E.D. and Nowak, R.D. (2004b), "Multiscale generalized linear models for non-parametric function estimation," *Biometrika,* (accepted for publication).

Korostelev, A.P. and Tsybakov, A. (1993), *Minimax Theory of Image Reconstruction. Lecture Notes in Statistics,* 82. New York: Springer.

Landgrebe,D.A. (2003), *Signal Theory Methods in Multispectral Remote Sensing,* Hoboken, New Jersey: John Wiley & Sons, Inc.

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

Marr, D. (1982), *Vision,* San Francisco: Freeman.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models,* New York, New York: John

Wiley & Sons, Inc.

Pal, X and Pal, Y (1993), "A review of image segmentation techniques," *Pattern Recognition,* 26:9, 1277-1294.

Patil, G.P. and Taille, C. (1999), "A Markov model for hierarchically scaled landscape patterns," *Bulletin of the International Statistical Institute,* 58:1, 82-92.

Pickett, S.T.A. and Cadenasso, M.L. (1995), "Landscape ecology: spatial heterogeneity in ecological systems," *Science,* 269: 331-334.

Pollak, I., Siskind, J.M., Harper, M.P., and Bouman, C.A. (2003), "Spatial random trees and the center-surround algorithm," Technical Report TR-ECE-03-03, School of Electrical and Computer Engineering, Purdue University.

Quattrochi, D.A. and Goodchild, M.F., Eds. (1997), *Scale in Remote Sensing and GIS.* Boca Raton, FL: CRC Lewis.

Tilton, J.C. (1998), "Image segmentation by region growing and spectral clustering with a natural convergence criterion," *Geoscience and Remote Sensing Symposium Proceedings,* 4, 1766 - 1768.

Titterington, D.M., Smith, A.F.M., and Markov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester: John Wiley & Sons.

Woodcock, C.E., Collins, J.B., Gopal, S., Jakabhazy, V.D., Li, X., Macomber, S., Ryherd, S., Wu, Y., Harward, V.J., Levitan, J., and Warbinton, R. (1994), "Mapping forest vegetation using Landsat TM imagery and a canopy reflectance model," *Remote Sensing of Environment,* 50, 240-254.

Yang, Y. and Barron, A. (1998), "An asymptotic property of model selection criteria," *IEEE Transactions on Information Theory,* 44:1, 95-116.

Zhu, S-C. (2003), "Statistical modeling and conceptualization of visual patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 25, 1-22.