BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

HIGH DIMENSIONAL LAND COVER INFERENCE USING REMOTELY SENSED MODIS DATA

by

HUNTER S. GLANZ

B.S., California Polytechnic State University, 2009
M.A., Boston University, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2014

Approved by

First Reader

Luis Carvalho, PhD
Assistant Professor of Statistics

Second Reader

Mark Friedl, PhD
Professor of Earth and Environment

Third Reader

Eric Kolaczyk, PhD
Professor of Statistics

Acknowledgments

Thank you to Maddie Schroth, my love, for everything that you are and your constant support. Thank you to Luis Carvalho, an unparalleled advisor, mentor, and friend that I will always be learning from. Thanks to Mark Friedl and Damien Sulla-Menashe for providing one of the best collaborative experiences of my life thus far. Thank you to Ian Johnston and Lijun Peng, the best officemates and colleagues I could have asked for. To my committee, I would like to thank you for all of the help and patience you have shown me. Thanks to the rest of my Boston University family: those friends, however close, that made my time in Boston and at BU so wonderful.

HIGH DIMENSIONAL LAND COVER INFERENCE USING REMOTELY SENSED MODIS DATA

(Order No.)

HUNTER S. GLANZ

Boston University, Graduate School of Arts and Sciences, 2014

Major Professor: Luis Carvalho, Assistant Professor

ABSTRACT

Image segmentation persists as a major statistical problem, with the volume and complexity of data expanding alongside new technologies. Land cover classification, one of the most studied problems in Remote Sensing, provides an important example of image segmentation whose needs transcend the choice of a particular classification method. That is, the challenges associated with land cover classification pervade the analysis process from data pre-processing to estimation of a final land cover map. Many of the same challenges also plague the task of land cover change detection. Multispectral, multitemporal data with inherent spatial relationships have hardly received adequate treatment due to the large size of the data and the presence of missing values.

In this work we propose a novel, concerted application of methods which provide a unified way to estimate model parameters, impute missing data, reduce dimensionality, classify land cover, and detect land cover changes. This comprehensive analysis adopts a Bayesian approach which incorporates prior knowledge to improve the interpretability, efficiency, and versatility of land cover classification and change detection. We explore a parsimonious, parametric model that allows for a natural application of principal components analysis to isolate important spectral characteristics while preserving temporal information. Moreover, it allows us to impute missing data and estimate parameters via expectation-maximization (EM). A significant byproduct of our framework includes a suite of training data assessment tools. To classify land cover, we employ a spanning tree ap-

proximation to a lattice Potts prior to incorporate spatial relationships in a judicious way and more efficiently access the posterior distribution of pixel labels. We then achieve exact inference of the labels via the centroid estimator. To detect land cover changes, we develop a new EM algorithm based on the same parametric model. We perform simulation studies to validate our models and methods, and conduct an extensive continental scale case study using MODIS data. The results show that we successfully classify land cover and recover the spatial patterns present in large scale data. Application of our change point method to an area in the Amazon successfully identifies the progression of deforestation through portions of the region.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Expansion
MODIS	Moderate Resolution Imaging Spectroradiometer
SPOT	Satellite for observation of Earth
MERIS	Medium Resolution Imaging Spectrometer
AVHRR	Advanced Very High Resolution Radiometer
EM	Expectation-Maximization
STEP	System for Terrestrial Ecosystem Parameterization
IGBP	International Geosphere-Biosphere Programme
Matrix-N	Matrix Normal Distribution
MLE	Maximum Likelihood Estimate
PCA	Principal Components Analysis
PC	Principal Component
CH	Calinski-Harabasz
Dir	Dirichlet Distribution
MN	Multinomial Distribution
N	Normal Distribution
MAP	Maximum *a posteriori*
MI	Mutual Information
KL	Kullback-Leibler
NASA	National Aeronautics and Space Administration
EVI	Enhanced Vegetation Index
EBF	Evergreen Broadleaf Forests
BRDF	Bidirectional Reflectance Distribution Function
NBAR	Nadir BRDF-adjusted Reflectance
PRODES	Basin Restoration Program
INPE	Brazil's National Institute for Space Research
LCCS	Land Cover Classification Scheme
ICM	Iterated Conditional Modes
SA	Simulated Annealing
MRF	Markov Random Field

List of Symbols

Symbol	Meaning
$N(v)$	Neighborhood of node v in a graph
$X \sim f$	X has probability density function f
$x \propto y$	x is proportional to y
X_i	Random variable, X, for observation i
$\text{vec}(X)$	Vectorized version of (matrix) X by appending columns
$\otimes$	Kronecker product
$\text{tr}(X)$	The trace of matrix X
$\partial Q/\partial \mu$	Partial derivative of Q with respect to μ
$\circ$	Hadamard (element-wise) product

Chapter 1

Introduction

1.1 Motivation

Interaction with and manipulation of the surface of the Earth necessitates an understanding of the processes and environments potentially affected. Understanding of global and local systems in both the biosphere and the atmosphere, as well as how they interact, continues to grow in importance as a foundation for resource management and land-use policies. The role of humans as members of many of Earth's ecosystems and contributors to many others has become less difficult to study in recent decades. Multitemporal, remotely sensed data of the entire Earth has become ubiquitous and with certain technological advances during the 1990s, continental and global scale land cover was mapped for the first time using remotely sensed data [21, 22, 47, 66, 105]. As newer remote sensing data sources have emerged (e.g. MODIS, SPOT-VEGETATION, MERIS), beyond older sensors such as AVHRR and Landsat, resulting data sets have grown in spectral, temporal, and spatial resolution. To continue informing Earth system models [8, 27, 94, 104] and providing Earth scientists with an accurate picture of global land cover, methods for land cover classification and change detection need to adapt to and mirror the sophistication of remote sensing data and technology.

The processes of land cover classification and change detection both involve a series of procedures built to deal with missing data, perform some sort of data compression, and finally conduct land cover inference. Therefore, attempts at these particular Remote Sensing tasks necessarily pursue a full analysis pipeline to overcome all of these obstacles. The

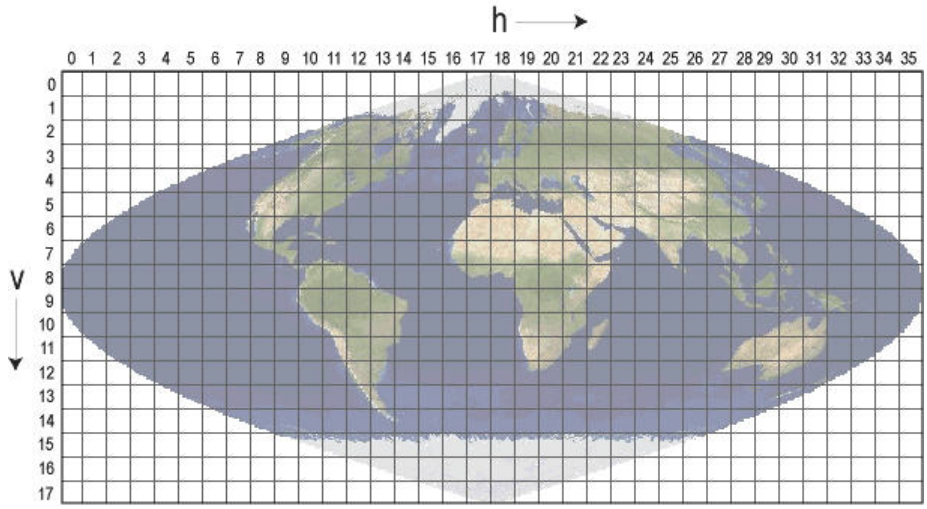


Figure 1.1: MODIS tile map.

contributions detailed throughout this document comprise a series of models and methods, which addresses each one of these sub-issues in a novel way. We aim at computationally feasible and yet representative approach which results in a pipeline that can be applied to these two problems in a number of ways. Before we review existing approaches to these problems, the following section introduces the nature of our data.

1.2 The Data

We focus our attention on a fairly new sensor launched aboard the Terra (1999) and Aqua (2002) spacecraft, called MODIS (Moderate Resolution Imaging Spectroradiometer). This instrument images the entire surface of the Earth every 1-2 days in 36 spectral bands at three different spatial resolutions: 250 meter, 500 meter, and 1,000 meter [1]. Thus, we possess land surface data at roughly 1.8 billion pixels world-wide for around 11 years. Figure 1.1 shows the globe broken down into *MODIS tiles*. MODIS tiles can be used to partition analysis of the surface of the Earth into more manageable sub-problems, with the caveat that tile boundaries can persist into analysis results.

A host of issues plague these raw data such as noise, atmospheric contamination, and variable view geometry. Our collaborators made a series of decisions in pre-processing

Table 1.1: MODIS bands 1-7 (land bands) [1].

Band	Range (nm)	Key Use
1	620-670	Absolute Land Cover Transformation, Vegetation Chlorophyll
2	841-876	Cloud Amount, Vegetation Land Cover Transformation
3	459-479	Soil/Vegetation Differences
4	545-565	Green Vegetation
5	1230-1250	Leaf/Canopy Differences
6	1628-1652	Snow/Cloud Differences
7	2105-2155	Cloud Properties, Land Properties

these data to circumvent these issues, consistent with traditional methods and approaches as we will see in the next section. We use the MODIS 500 meter Nadir BRDF-adjusted Reflectance (NBAR) product, which is designed to minimize noise due to bidirectional reflectance effects arising from varying solar and view geometry [95]. Additionally, we composite the daily data to 8-day weekly observations. For the purposes of land cover inference, we restrict our attention to the 7 MODIS *land bands* given in Table 1.1.

We begin work with a multispectral, multitemporal observation (7 spectral values recorded at 46 times points) at each pixel, for each year in our dataset. Land cover classification and detection of changes in land cover require a scheme of land cover classes which encompasses all major land cover types. We employ a carefully established set of land cover classes constructed under the International Geosphere-Biosphere Programme (IGBP) [18]. According to the IGBP this collection of land cover classes "was chosen to be exhaustive," mutually exclusive, and "structured so that [they] were equally interpretable with 1 km data." Detailed information about these land cover classes can be found in Table 1.2 with a corresponding color legend in Figure 1.2. In the interest of brevity, the land cover class numbers in Table 1.2 will be used in *lieu* of the actual names in plots throughout the remainder of this document.

The following two sections describe the most recently used approaches to land cover classification and change detection respectively, and involve explicit handling of each of the aforementioned sub-problems.

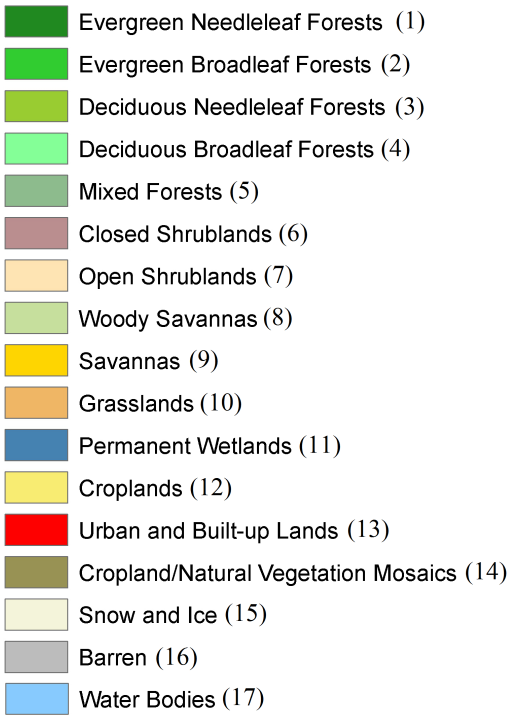


Figure 1.2: International Geosphere-Biosphere Programme (IGBP) land cover color scheme.

1.3 Previous Work

Were we sitting alongside MODIS, circling the Earth, the northeast portion of North America might look like the image in Figure 1.3 (without large water bodies masked out).

The tasks of land cover classification and change detection involve analysis of the data at the pixels in such images. In the case of classification, we aim to produce a map with each pixel colored according to a discrete set of land cover labels (see Figure 1.4).

The remainder of this document outlines the ways in which we move from Figure 1.3 to Figure 1.4. The next two sections detail the "state-of-the-art," currently employed methods for classifying land cover and detecting change.

Figure 1.3: True color composite of northeast portion of North America.

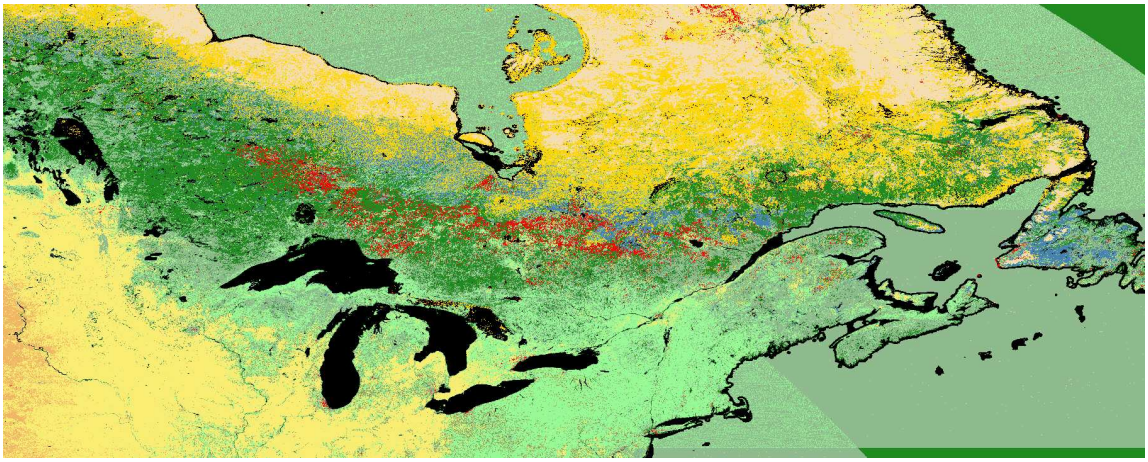


Figure 1.4: Example of land cover map of northeast portion of North America.

MODIS Collection 5 Land Cover Product [32]

The two previous MODIS land cover products, dubbed Collection 4 and Collection 5, benefit from a rich history of land cover classification efforts throughout the past 20-30 years. In an effort to take advantage of some of the improved Remote Sensing capabilities afforded by MODIS, our collaborators set out to establish a pipeline that could be reliably used to classify land cover from data with high temporal resolution and low spatial resolution.

The process begins with the establishment of high quality training data to be used in a supervised classification algorithm. Training sites are defined using Landsat or higher

resolution imagery via manual interpretation. These sites can range in size from 1 to 376 pixels, and are constructed to homogeneously represent one IGBP class. Because the frequency distribution land cover classes across Earth's land areas is not uniform and because certain classes exhibit more variability than others, the training data do not contain an equal amount of sites from each land cover class. These sites, along with previously defined sites, contribute to a MODIS land cover training site database whose maintenance and growth plays a crucial role in land cover analysis.

For the remainder of this section we will focus on the production of MODIS Collection 5. The original data consists of 8-day weekly, NBAR reflectance values [95] of MODIS bands 1-7, land surface temperature [116], and the enhanced vegetation index (EVI)[53]. These data are aggregated to 32-day averages to reduce data volume and the prevalence of missing values. The final input to the classification algorithm consists of these averages as well as annual metrics (minimum, maximum and mean values) for the NBAR bands, land surface temperature, and EVI for a grand total of 135 features for each pixel in a given year.

Decision tree algorithms detailed in [87] (C4.5) provide a robust way of performing classification while coping with missing data. However, if the number of missing features at a pixel exceeds 84, the pixel is not classified. Instead the pixel is assigned the most recent Collection 5 label (since this product consists of labels per pixel, per year) or Collection 4 label (if the problem persists). Two assumptions of the ensemble decision tree algorithm contribute to the need for post-classification adjustments: 1) that the distribution of the training data is representative of the population, and 2) that the features are able to distinguish the classes in the training data. Violations of the first assumption result from the inability of the training site database to completely capture the complete range of variability in the land cover classes. Ambiguous class definitions, as seen in Table 1.2, and low spectral-temporal separability of many classes breaks the second assumption.

To overcome these two issues, the class-conditional probabilities estimated by the boosted decision trees are adjusted in multiple ways. First, to correct the sample bias

associated with the limitations of the training site database the class conditional probabilities are multiplied by prior probabilities defined to be inversely proportional to the number of training samples in each class. Second, inadequate class separability is addressed by applying a set of spatially explicit prior probabilities. Friedl et al. use previous land cover products and cropping intensity [89], with a moving window algorithm to compute the regional proportion of each class, and estimate these second prior probabilities for each pixel. To minimize the influence of these spatial priors, a tuning parameter, $c \in [0, 1]$, controls how heavily the prior probabilities, $P(i)$ for class i, get weighted with respect to an original prior $P_0(i)$:

$$P(i) = P_0(i) + (1 - P_0(i)) \times (1 - c). \tag{1.1}$$

These get normalized to sum to 1 after using c. The Urban, Wetlands, and Deciduous Needleleaf Forest classes continue to be problematic even after the post-classification application of priors. If the posterior probability of Deciduous Needleleaf Forest does not exceed 0.7 then the next most likely class gets assigned. Similarly, if the posterior probability of Wetlands does not exceed 0.75 then the next most likely class gets assigned. Later, in Section 3.3.2, we discuss our ability to formally adjust for these problematic classes using confusion (gain) matrices in the posterior inference. The urban class was mapped using an ecoregion-based stratification with eighteen strata, where training data and supervised classifications were developed and tuned to each stratum. Finally, the classification results get stabilized across years. That is, if the pixel label in a given year differs from the previous year then the new label is only assigned if its posterior probability exceeds the probability associated with the previous year's label. They carry out this final adjustment using three-year windows.

The execution of this pipeline produces reasonable and accurate maps of land cover. However, a host of items warrant attention and further effort. The aggregation of spectral features to monthly values represents a major sacrifice of potentially important temporal

information. Missing data does not receive formal attention outside of the aggregation of features. The classification method, decision trees via the C4.5 algorithm, not only exhibits excessive sensitivity to training data warranting post-classification adjustments, but assigns labels to pixels independent of the information at or labels of neighboring pixels. Spatial information contributes to the final classification, only in the form of previous land cover products and cropping intensity. In Section 1.4 we enumerate our novel approaches to these key issues.

Distance Metric-Based Change Detection [51]

Change point detection describes a broad class of problems involving some process we wish identify a change in. In Remote Sensing, land cover change detection can still take many forms. With more than 12 years of global observations from MODIS, Huang and Friedl aim to take advantage of this fairly new data source in an attempt to identify areas of Evergreen Broadleaf Forest which have changed in some way such as burned, logged, or converted to cropland. They build an analysis pipeline to detect change at the scale of years that begins with data pre-processing.

Training data summarizing the assumed pre-change class of Evergreen Broadleaf Forest comes, again, from the MODIS training site database. In this work, 8-day weekly observations of MODIS band 7 and EVI2, a second slightly different enhanced vegetation index, from May through September constituted the data at each pixel per year. Even after subsetting the full MODIS data at each pixel, in a given year, down to 38 (2 spectral features at 19 times points) features there still exists missing data.

For each missing value, they first searched for 'candidate' fill values on the same date from nearby pixels of the same class. If suitable pixels were not found, the search window was expanded until suitable fill values were found. The median of high quality candidate values was then used to fill the data gap. If gap-filled values differed substantially from current and previous year time series, they were replaced using interpolation. A 'despiking' procedure was then applied to eliminate sudden changes in all values at each pixel. Finally,

a 3-point median filter was applied to further reduce noise.

With missing data addressed they proceed to identify change in a year for a given pixel using two distance metrics: 1) distance from its assumed land cover class in each year and 2) distance between preceeding and following year time series. The first, within-class distance, constitutes a Mahalanobis distance [72]. They then proceed to test and establish distance thresholds. With these set, change detection was performed on a per pixel basis and when the thresholds were exceeded a pixel was flagged as 'potential change.' Changed pixels were required to exceed the selected thresholds in two consecutive years, and the timing of the change was defined as the first year when the pixel was flagged.

This method successfully mapped regional forest change in the Xingu River Basin in Mato Grosso, Brazil and in the Pacific Northwest region of the United States. However, this series of procedures contains some user-defined components that do not lend themselves to easy and interpretable adjustment. For example, the distance thresholds used and the requirement of 2 consecutive 'potential change' flags could depend on the type of changes being analyzed, the assumed pre-change class, or the region of interest. In Section 1.4 we propose an approach to change detection that builds on the machinery established for land cover classification and is well suited for conversion from one land cover class to another.

The two aforementioned pipelines for land cover classification and change detection have evolved from numerous attempts in the field of Remote Sensing, but a number of more general, statistical techniques have been brought to bare in other ways. While we also propose a pipeline, we put forth new methodologies and methods that each carry new contributions to Statistics. The next section reviews the relevant literature for each sub-problem of these analysis pipelines, and then introduces our novel contribution.

1.4 Novel Contributions

Statistically, land cover classification and change detection attracts a great deal of attention and a number of tools, both novel and well-studied. Despite being a classic problem,

classification and change detection here involve three significant challenges: large data, spatio-temporal structure, and missing data as we saw in Section 1.1. To overcome these challenges and successfully classify land cover and detect changes, we employ a series of models and methods which provide physically satisfying and interpretable results with an eye toward computational efficiency.

Our first reduction in the size of the data comes from the decision to trim the ends of the year off. While we ultimately plan to produce global land cover maps, our regions of interest in this document will live in the northern hemisphere. Because the ends of the year, or winter months, contain extra amounts of noise and missing data we focus our attention on the middle 28 time points (or 60%) of the year.

Parametric Model

Traditional maximum likelihood approaches to land cover classification [108, 65, 56, 113, 60] often assume that the data at a given pixel, X_v, are multivariate normal (1.2) conditional on the land cover class at pixel v, θ_v.

$$X_v \,|\, \theta_v = c \overset{\text{ind}}{\sim} N(\mu_c, \Sigma_c) \tag{1.2}$$

More specifically, given the land cover class of a particular pixel v, θ_v, the data at pixel v, X_v, has a multivariate normal distribution with a class-specific mean and covariance. We propose a novel refinement of this model (Chapter 2) in the form of a Kronecker structure covariance (1.3), unprecedented in the Remote Sensing literature. Because these data consist of two distinct dimensions, spectral and temporal, models that can exploit this structure will yield increased interpretability and potentially simpler estimation procedures. To this end we model the single-year, spectral-temporal data at pixel v as a matrix, X_v, with the rows representing bands and the columns representing time points:

$$\text{vec}(X_v) \,|\, \theta_v \overset{\text{ind}}{\sim} N(\text{vec}(\mu_c), \Sigma_s \otimes \Sigma_{t,c}) \tag{1.3}$$

where $\otimes$ denotes the Kronecker product and $\text{vec}(X_v)$ denotes the vectorization of the matrix X_v. This Kronecker structure naturally fits the spectral-temporal structure of the data and represents a reduction in the number of model parameters from (1.2). Because seasonal variation is a first-order property that helps to distinguish many land cover classes, multitemporal information provides critical information that our model isolates.

More generally, the Kronecker covariance structure has been explored in many other situations. Multivariate repeated measures data, for example, provides this type of framework [7, 80, 92, 93]. That is, the response variables and time are two separate *dimensions* of the data that can be characterized independently. Thus, a straightforward way to model the full covariance is via the Kronecker product of a covariance of the responses and a covariance of time. Similarly, multivariate time series of other types such as longitudinal data [13, 38] and spatio-temporal data [97, 36] lend themselves to this kind of covariance decomposition. To this end, classic work has been done on parameter estimates for the matrix normal distribution [19, 25, 101, 102, 3].

Because of the reduction in parameters it represents, the Kronecker structure and linear combinations of Kronecker structure matrices have been used to approximate high-dimensional covariances [110]. Dubbed the "ML flip-flop algorithm," the approach established in [25] has become a standard way to estimate a covariance with a Kronecker structure. Convergence studies in [111] confirm the usefulness of available algorithms, but people persist in developing faster, non-iterative methods of estimation [117]. While maximum likelihood estimates have been derived along with tests of whether a covariance matrix has this structure, the issue of missing data in this context has not been fully addressed. Work by Allen and Tibshirani [3] treats a specific variation of this problem as an application of variance selection. Key differences include constraints they impose on the mean, μ, and the number of observations used to estimate the parameters. Additionally, work in [64] attempts to estimate a Kronecker structure covariance in the presence of missing values, but focuses on estimation of the overall covariance instead of the two components.

Because we plan to target $\widehat{\Sigma}_s$ with principal components analysis, we need to estimate

the individual components of the Kronecker structure (i.e. the parameters of the matrix normal distribution). For this we derive a statistically novel expectation-maximization (EM) algorithm [23] to estimate the parameters of this model, in the presence of missing data, using training data from the System for Terrestrial Ecosystem Parameterization (STEP) database [79] (Chapter 2.2).

Missing Data Imputation

Clouds, snow and other disruptive phenomena prevent clean, high quality images of the Earth's surface. As a result, missing values exist throughout all MODIS data. In the previous section we mentioned the implementation of an EM procedure to estimate the matrix normal parameters in the presence of missing data. Ideally, we could use an EM on all data to infer the land cover labels, again treating the missing data as latent. However, the iterative nature of this classification method and size of the data make it computationally prohibitive.

Though expectation-maximization iterates between optimizing two things, in the case of the previous section parameter estimates and missing data values, we restrict our use of this procedure to estimation of the model parameters. The reason being, imputation of missing data via this procedure could only work for training data. We need a way to handle missing data in every pixel we wish to analyze. In Remote Sensing, some typical procedures include compositing of multiple dates of data, interpolation, or other ad-hoc gap-filling methods as described above in [32, 51]. To formally address the presence of missing data in almost all pixels we therefore propose a pre-processing step to impute all missing data before attempting classification.

With the matrix normal parameter estimates assumed known, we propose a statistically **novel** way to impute missing data which incorporates information from the observed data and all of the land cover classes. That is, we derive an EM algorithm which now treats the pixel label as latent and estimates the missing data in an iterative fashion, for each pixel.

Data Compression and Training Data Assessment

At this point in our data at a given pixel in a given year consists of a 7×28 matrix with the 7 MODIS land band values recorded at each of 28 times points during the middle 60% of the year. The size of the data remains cumbersome, and so we seek to further reduce its dimensionality. Principal components analysis (PCA) [58] provides an intuitive way to identify linear combinations of features that contain large portions of variability in the data.

Historically, people construct additional features such as EVI or annual metrics to assist in land cover inference procedures. With the advent of hyperspectral data and other high dimensional, Remote Sensing datasets, feature selection or reduction has become an important part of any analysis pipeline. To this end, principal components analysis has been used extensively to simplify the feature space both for classification and visualization purposes [14, 78, 50, 90, 57, 26]. As a by-product of such an analysis, the first few principal components can often be interpreted in terms of physical characteristics of a pixel [63].

Experts in the field attest to the correlation between spectral bands and the consequent redundancy of information among them [17]. To isolate the most useful information present in the data, in a lower dimensional space, we utilize a principal components analysis. Because spectral variation transcends the differences between land-cover classes, we target $\widehat{\Sigma}_s$ with PCA. The main novelty in our application of PCA stems from our targeting $\widehat{\Sigma}_s$ as opposed to the full covariance. We still recover physically interpretable components and decide to use the first three principal components instead of all 7 MODIS land bands (Chapter 2.4)

Additionally, exploiting the increased interpretability of the matrix normal likelihood and this application of PCA we develop a novel set of tools to assess the quality of training data sites. In particular, we create a measure of training site *homogeneity*, an estimate of training site *composition*, and a metric to describe the evidence that a pixel belongs to a given training site (i.e. the training site *membership* of a pixel).

Land Cover Classification

The problem of land cover classification has in some way eluded any effort made to conquer it. The variety of classification schemes and data sources make it a very challenging task to say the least. Whether a particular remote sensing instrument was built to map land cover or not, the task inevitably gets attempted. Different spectral, temporal and spatial resolutions unfailingly affect the classification results. The most popular methods applied in recent years consist of supervised classification algorithms. People cite the complexity of this high-dimensional, spectral-temporal data as reason for finding and using a host of tools to enhance the classification procedure.

Traditional methods train a classifier on high quality data from diverse, but hopefully representative training sites. As mentioned above, previous maximum likelihood approaches assume the data follow a multivariate normal distribution and merely classify pixels independent of one another according to the land cover class with the highest likelihood. While these models performed well, they were eventually expanded to incorporate prior information about the distribution of land cover classes [20, 76, 106]. This took the form of simple, global prior probabilities being multiplied by the likelihood of the spectral data at each pixel to produce posterior probabilities and classification via the maximum *a posteriori* (MAP) estimator. These prior probabilities could represent the frequency distribution of the classes in the training data, but were other times estimated using independent data.

To take advantage of spatial information when performing classification, many people utilized indicator kriging to estimate pixel-specific prior probabilities that incorporated data from nearby pixels [42, 83, 43]. The traditional MAP estimators described above could then take advantage of these spatially explicit prior probabilities.

While some of these previous likelihood based methods did not always assume a multivariate normal distribution, much of the recent literature tends toward the use of machine learning techniques such as neural networks [2, 5, 44, 48] and decision trees [31, 33, 46, 81,

35, 34] in order to escape from parametric assumptions that they believe may limit the abilities to capture complex relationships within possibly multimodal data.

Despite being flexible and quick, techniques like decision trees can suffer from a sensitivity to training data (e.g. overfitting); do not formally handle missing data; still require a number of post-classification adjustments; and, perhaps most importantly, do not include spatial information when classifying pixels. Grander attempts exploit both spectral and spatial information when performing land cover classification involve Markov Random Fields [55] (MRF). Specifying a graphical model on the lattice of pixels in an image allows for statistical inference of the pixel labels to take advantage of previously developed machinery in Statistical Mechanics. These methods specify some likelihood for the spectral data and some cost associated with neighboring pixels being different land cover classes. An energy function combines these two components of information for the entire lattice. The set of pixel labels that minimizes this energy function corresponds to a global MAP estimator. For a thorough application of this approach to remote sensing images, we refer the reader to [77].

Because the lattice is highly connected and can be quite large, finding the set of pixel labels that globally maximize the posterior is computationally intractable even for low dimensions. Two classic, iterative approaches to minimizing the energy function of the Markov Random Field are Iterated Conditional Modes (ICM) [6] and Simulated Annealing (SA) [39, 107]. Simulated Annealing can converge, under certain assumptions, to the global energy minimum but usually involves long computation times. Iterated Conditional Modes is a deterministic method which does not usually take long to run, but only converges to a local energy minimum and, as such, is very sensitive to starting values.

In this document we demonstrate two classification techniques. The first involves no spatial information and classifies pixels independently via a MAP estimator which takes advantage of our novel matrix normal likelihood and prior probabilities elicited from our collaborators. The second involves building a hierarchical graphical model, not unlike the MRF methods described above. Because of the previously mentioned computational

intractability of the lattice in this Bayesian setting, we propose approximating the prior distribution of labels on the lattice with a distribution on a minimally connected graph, a spanning tree of the lattice. We derive a statistically novel EM algorithm for finding a representative spanning tree to use during inference. With this tree in hand we conduct exact posterior inference of the pixel labels via the *centroid* estimator [12], a novel estimator in the field of Remote Sensing.

Change Point Detection

Though land cover classification comprises one of the largest problems in Remote Sensing, another similarly significant problem consists of detecting *changes* in land cover. A host of methods have been developed to address this challenge [67], from comparing the raw spectral data to comparing land cover classification results of images at two different points in time.

Change point detection methods have been applied extensively in various fields of environmental and climate monitoring, to problems involving rates of Tropical cyclone activity, precipitation and temperature trends, and fishery population regime change [28, 16, 91, 99]. In the field of Remote Sensing, various change detection techniques were developed using bi-temporal or multi-temporal imagery for mapping changes including deforestation, forest mortality, and urban expansion (see [98, 68]). As MODIS time series grow, more studies have focused on better exploitation of the temporal information in MODIS data for change detection [e.g., 114, 88, 52]. However, due to the nature of optical remote sensing (susceptible to cloud and atmospheric contamination), it remains challenging to pre-process and fully utilize the time series data. Methods that better address missing data and are robust to noise are needed.

Statistically, the general change point problem can be broken down into on-line (real time) [30] and off-line (retrospective) frameworks. Additionally, approaches to change point detection typically involve specifying which types of change to look for. Previous methods for detecting change vary by the following change types: mean-type shifts [96, 69], variance

change [37], or distribution change [4, 61, 109, 100, 40]. Popular approaches include time series models, purely likelihood-based methods, special forms of regression or Bayesian techniques [9, 29, 75, 84, 103].

Most existing methods for change detection in the presence of missing data attempt to impute or estimate missing data first and then proceed to identify changes [10]. Estimation can proceed in a number of ways, including nearest neighbor interpolation, linear interpolation, polynomial interpolation, or spline interpolation. Missing values can be imputed using multiple imputation or, as we have already been exposed to, EM. Honaker and King propose a way to impute missing data in time-series [49]. For a thorough review of handling missing data in statistical analyses, we refer the reader to Rubin's book [62].

We maintain our likelihood from (1.3) with the same set of IGBP land cover classes. Our aim here consists of identifying which year, if any, constitutes a change point for each pixel independently of other pixels. The high dimensionality of the parameters of (1.3) and short time series prohibit the estimation of pre- and post-change model parameters for all of the change years we want to consider. Consequently, we propose a method which models conversion-type changes that does not rely so heavily on inference from the pixels of interest. After estimating class-specific parameters for (1.3) we devise a model which describes the pre- and post-change data in terms of our known land cover classes as opposed to relying solely on inference about the pre- and post-change distributions from the data of interest. We incorporate subject-matter knowledge about the number and locations of the change points via prior distributions on these values and then estimate the change point locations in the presence of missing data with another EM procedure. Chapter 4 contains the details of our new approach and subsequent results.

Our aforementioned spectral-temporal model established, we develop a progression of novel change detection algorithms which identify change points in the presence of missing data. The culmination of these approaches utilizes class parameter estimates based on STEP training data and another EM procedure to detect conversion-type changes in individual pixel time series.

1.5 Summary of Contributions

Our land cover classification and change detection efforts will begin with the specification of a parametric model that explicitly characterizes the spectral-temporal structure of our data in a novel way. Because the presence of missing data presents a non-trivial issue to overcome, we derive a novel EM algorithm for estimating model parameters (Sections 2.1 and 2.2) that accounts for missing data.

We derive a missing data imputation pre-processing step in Section 2.3 due to the computational infeasibility of inferring land cover labels with an iterative EM procedure. Section 2.4 describes our novel application of PCA to compress the data spectrally, resulting in the use of physically interpretable components as opposed to the original, correlated MODIS land bands. Taking advantage of the increased interpretability of our model, we build novel measures of training data quality in Section 2.7. We present simulation and case study results of our parameter estimation procedure, data compression method, and these training data assessments in Section 2.5.

With high quality training data, we apply both independent pixel and hierarchical graphical classification methods to imputed, transformed data in North America. We conduct exact posterior inference of land cover labels while incorporating spatial information with a novel construction of a Potts prior on the lattice of pixels and use of the centroid estimator in Chapter 3. We give classification results and comparisons in Sections 3.1.1, 3.4.2, and 3.4.4.

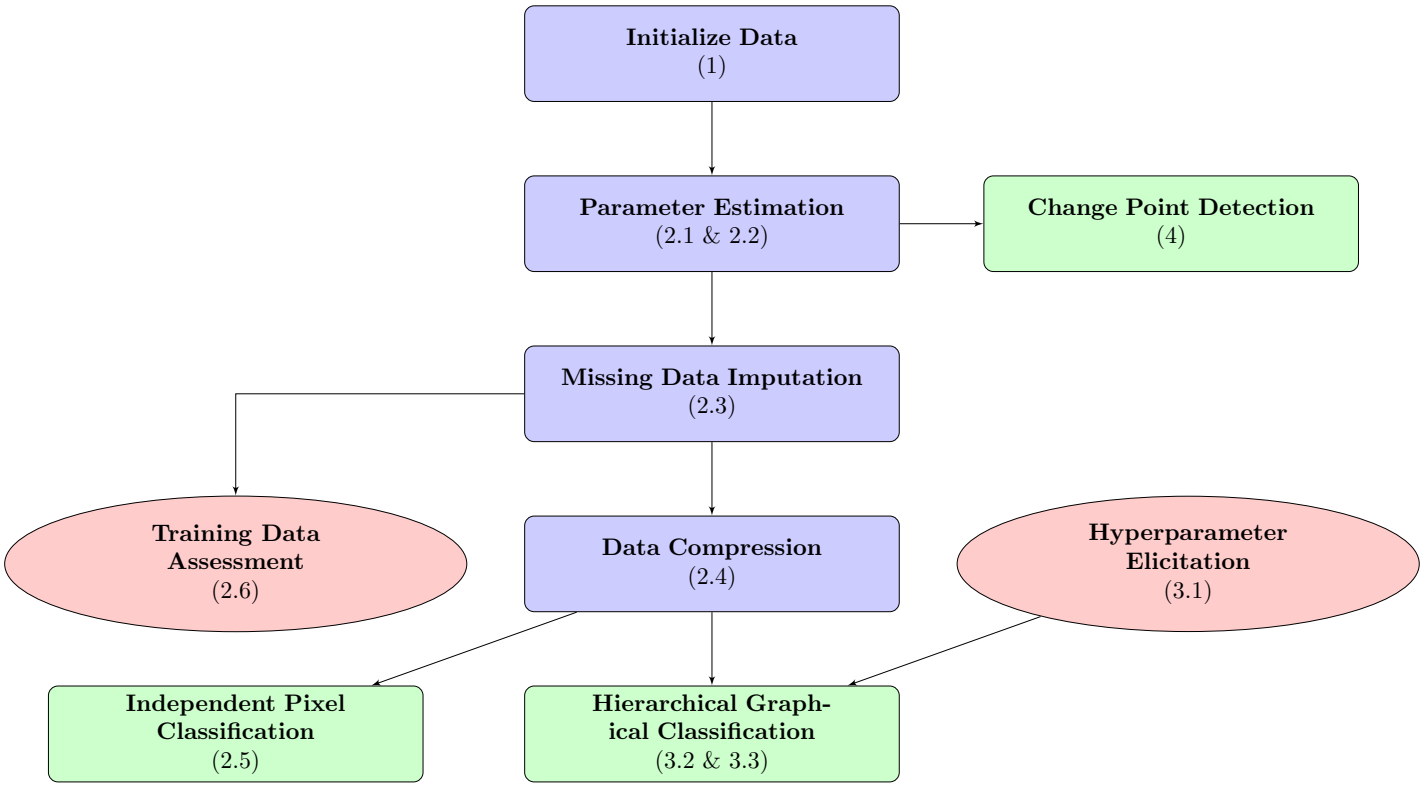
Finally, derive a novel EM algorithm to identify conversion-type changes in Remote Sensing time series in Section 4.3.1. We present simulation and case study results for this change point detection technique in Sections 4.4 and 4.5.

Figure 1.5 gives a schematic of the unified analysis pipeline we develop here for the tasks of land cover classification and land cover change detection.

Table 1.2: Land cover class definitions within the International Geosphere-Biosphere Programme.

CLASS	CLASS NAME	DESCRIPTION
1	Evergreen Needleleaf Forests	Lands dominated by trees with a percent canopy cover > 60% and height exceeding 2 meters. Almost all tree remain green all year. Canopy is never without green foliage.
2	Evergreen Broadleaf Forests	Lands dominated by trees with a percent canopy cover > 60% and height exceeding 2 meters. Almost all tree remain green all year. Canopy is never without green foliage.
3	Deciduous Needleleaf Forests	Lands dominated by trees with a percent canopy cover > 60% and height exceeding 2 meters. Consists of seasonal needleleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
4	Deciduous Broadleaf Forests	Lands dominated by trees with a percent canopy cover > 60% and height exceeding 2 meters. Consists of seasonal broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
5	Mixed Forests	Lands dominated by trees with a percent canopy cover > 60% and height exceeding 2 meters. Consists of tree communities with interspersed mixtures or mosaics of the other four forest cover types. None of the forest types exceeds 60% of landscape.
6	Closed Shrublands	Lands with woody vegetation less than 2 meters tall and with shrub canopy cover is > 60%. The shrub foliage can be either evergreen or deciduous.
7	Open Shrublands	Lands with woody vegetation less than 2 meters tall and with shrub canopy cover is $10-60\%$. The shrub foliage can be either evergreen or deciduous.
8	Woody Savannas	Lands with herbaceous and other understorey systems, and with forest canopy cover between $30-60\%$. The forest cover height exceeds 2 meters.
9	Savannas	Lands with herbaceous and other understorey systems, and with forest canopy cover between $10-30\%$. The forest cover height exceeds 2 meters.
10	Grasslands	Lands with herbaceous types of cover. Tree and shrub cover is less than 10%.
11	Permanent Wetlands	Lands with a permanent mixture of water and herbaceous or woody vegetation that cover extensive areas. The vegetation can be present in either salt, brackish, or fresh water.
12	Cropland	Lands covered with temporary crops followed by harvest and a bare soil period (e.g. single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type.
13	Urban and Built-Up	Lands covered by building and other man-made structures.
14	Cropland/Natural Vegetation Mosaics	Lands with a mosaic of croplands, forest, shrublands, and grasslands in which no one component comprises more than 60% of the landscape.
15	Snow and Ice	Lands under snow and/or ice cover throughout the year.
16	Barren	Lands exposed soil, sand, rocks, or snow and never has more than 10% vegetated cover during any time of the year.
17	Water Bodies	Oceans, seas, lakes, reservoirs, and rivers. Can be either fresh or salt water bodies.

Figure 1.5: Schematic of our proposed analysis pipeline.



Chapter 2

Likelihood Model and Parameter Elicitation

The first practical problem of interest here involves labeling an image domain pixel-wise with a given set of discrete labels, $\mathcal{S} = \{1, \ldots, C\}$, which come from our IGBP scheme. The data consists of multivariate (or matrix) observations at each pixel in the image. We confine our view to a single year of data (7 spectral bands observed at 28 time points) when performing land cover classification. In general, let P be a set of pixels in the image L (denoted L because the array of pixels naturally form a square lattice) of size $n = n_r \times n_c$, $\mathcal{S} = \{1, \ldots, C\}$ a set of C labels, and for each node $v \in \mathcal{P}$ is defined as a set $N(v) \subset P$ called the *neighborhood* of v. The classification problem consists of assigning a label from the set $\mathcal{S}$ to each node in the set of nodes, P. In pursuit of a Bayesian approach which incorporates *a priori* knowledge and conducts land cover inference via a posterior distribution, we begin by specifying a new parametric likelihood for the data.

For the data, X_v, given the land cover class, θ_v, at pixel v we specify

$$X_v \,|\, \theta_v = c \overset{\text{ind}}{\sim} \text{Matrix} - N(\mu_c, \Sigma_{t,c}, \Sigma_s) \tag{2.1}$$

where Matrix-N denotes the matrix normal distribution. In (2.1), X_v and μ_c are matrices with the rows and columns corresponding to the spectral bands and time points, respectively. Here, then, $\Sigma_{t,c}$ represents a class-specific temporal covariance or covariance of the columns of X_v, and Σ_s represents a spectral covariance or covariance of the rows of X_v.

The model in (2.1) is equivalent to specifying

$$\text{vec}(X_v) \,|\, \theta_v \overset{\text{ind}}{\sim} N(\text{vec}(\mu_c), \Sigma_s \otimes \Sigma_{t,c}). \tag{2.2}$$

In the sections that follow we derive an EM algorithm for estimating the parameters of (2.2) in the presence of missing data. We then impute missing values and compress the data in a physically interpretable way. The chapter concludes with simulation and case study results for these methods.

2.1 Likelihood Model

Technological advances in recent decades have ushered in a new era in data management and analysis. The dimension of data sets continues to grow alongside the number of observations. Consequently, the estimation of parameters or characteristics of these data remains a significant challenge. Specifically, the covariance matrix of such high dimensional data can be extremely difficult to estimate and handle. An increasingly common simplification is the assumption that this covariance has a Kronecker product structure.

For our remotely sensed data we begin with N independent observations from a mixture of matrix normal distributions,

$$X_i \,|\, \theta_i = c \overset{\text{ind}}{\sim} \text{Matrix} - N(\mu_c, \Sigma_{t,c}, \Sigma_s), \quad i = 1, \ldots, N, \tag{2.3}$$

where X_i and μ_c are $B \times T$ matrices, Σ_s is $B \times B$ and $\Sigma_{t,c}$ is $T \times T$. The use of this structure reduces the number of parameters by explicitly describing the covariance between the rows and the covariance between the columns as opposed to an individual covariance in each cell of the upper triangle of the full, $BT \times BT$, covariance matrix. Besides this simplification, the partitioning of the covariance follows naturally from a setup involving two *physical*, or *separable*, dimensions such as spectral (spectral bands) and temporal (time points) with our remotely sensed data.

In this model, two matrices will characterize class-specific temporal properties: a B-by-T matrix that provides the mean spectral-temporal profile for each land cover class c (μ_c), and an associated temporal covariance matrix ($\Sigma_{t,c}$, which is of order T.) The covariance structure between spectral bands is captured by Σ_s (of order B.) Since spectral and temporal effects are orthogonal, the model requires fewer parameters; moreover, this separation between spectral and temporal sources of variance allows dimensionality reduction to be focused in the spectral bands, as detailed in Section 2.4. Because we want to isolate the spectral and temporal sources of variation in our remotely sensed data, this matrix normal distribution provides a natural choice for our data likelihood. Clouds and other physical phenomena induce missing values in MODIS data, necessitating procedures that can properly accommodate missing data. To this end we derive an expectation-maximization (EM) algorithm for estimating the parameters of the matrix normal distribution in the presence of missing data.

2.2 Parameter Elicitation

Until Section 2.5 we will focus on estimating the parameters of a single matrix normal distribution: μ, Σ_s, and Σ_t. For the sake of generality we will denote the order of Σ_s by p and the order of Σ_t by q.

2.2.1 Parameter Estimation with Complete Data

It is straightforward to obtain maximum likelihood estimates of the parameters of a matrix normal distribution when there is no missing data. If $X_i \overset{\text{iid}}{\sim} \text{Matrix} - N(\mu, \Sigma_c, \Sigma_s)$ then, equivalently,

$$\text{vec}(X_i) \overset{\text{iid}}{\sim} N(\text{vec}(\mu), \Sigma_s \otimes \Sigma_c)$$

and so

$$\mathbb{P}(X_i;\mu,\Sigma_c,\Sigma_s) = (2\pi)^{-\frac{pq}{2}}|(\Sigma_s\otimes\Sigma_c)^{-1}|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\mathrm{vec}(X_i-\mu)^\top(\Sigma_s\otimes\Sigma_c)^{-1}\mathrm{vec}(X_i-\mu)\right\}.$$

As usual in multivariate normal densities, in the exponential term we have the Mahalanobis distance $D_\Sigma(X_i,\mu)$, with $\Sigma := \Sigma_s\otimes\Sigma_c$, between X_i and μ,

$$D_\Sigma(X_i,\mu) \doteq \mathrm{vec}(X_i-\mu)^\top(\Sigma_s\otimes\Sigma_c)^{-1}\mathrm{vec}(X_i-\mu).$$

This distance can be worked through known identities of the vec operator and Kronecker product to yield a simpler expression for the matrix normal density:

$$\begin{aligned} D_\Sigma(X_i,\mu) &= \mathrm{vec}(X_i-\mu)^\top(\Sigma_s\otimes\Sigma_c)^{-1}\mathrm{vec}(X_i-\mu) \\ &\overset{\text{(i)}}{=} \mathrm{vec}(X_i-\mu)^\top(\Sigma_s^{-1}\otimes\Sigma_c^{-1})\mathrm{vec}(X_i-\mu) \\ &\overset{\text{(ii)}}{=} \mathrm{vec}(X_i-\mu)^\top\mathrm{vec}[\Sigma_c^{-1}(X_i-\mu)\Sigma_s^{-1}] \\ &\overset{\text{(iii)}}{=} \mathrm{tr}\left[(X_i-\mu)\Sigma_c^{-1}(X_i-\mu)^\top\Sigma_s^{-1}\right] \\ &= \mathrm{tr}\left[\Sigma_s^{-1}(X_i-\mu)\Sigma_c^{-1}(X_i-\mu)^\top\right] \end{aligned}$$

where (i), (ii), and (iii) are applications of identities (488), (496), and (497) in [85], respectively. Furthermore, since

$$|(\Sigma_s\otimes\Sigma_c)^{-1}| = |\Sigma_s^{-1}|^p|\Sigma_c^{-1}|^q$$

by identity (492) in [85], we recover the characterization of [101]: $X_i \sim Matrix{-}N(\mu,\Sigma_c,\Sigma_s)$ if and only if the density of X_i is given by

$$\mathbb{P}(X_i;\mu,\Sigma_c,\Sigma_s) = (2\pi)^{-\frac{pq}{2}}|\Sigma_s|^{-\frac{q}{2}}|\Sigma_c|^{-\frac{p}{2}}$$

$$\exp\left\{ -\frac{1}{2}\mathrm{tr}\left\{ \Sigma_s^{-1}(X_i-\mu)\Sigma_c^{-1}(X_i-\mu)^\top \right\} \right\}. \quad (2.4)$$

The log likelihood of the parameters $\Theta = (\mu, \Sigma_s, \Sigma_c)$ is then

$$\begin{aligned} \log \mathbb{P}(X_1, \ldots, X_n; \Theta) &= \sum_i \log \mathbb{P}(X_i; \Theta) \\ &= \frac{pN}{2}\log|\Sigma_c^{-1}| + \frac{qN}{2}\log|\Sigma_s^{-1}| - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i D_\Sigma(X_i, \mu) \quad (2.5) \end{aligned}$$

up to a normalizing constant. The form in (2.4) simplifies the matrix derivatives of (2.5) considerably leaving us with the following maximum likelihood estimates (MLEs):

$$\begin{aligned} \widehat{\mu} &= \frac{1}{N}\sum_{i=1}^N X_i \\ \widehat{\Sigma}_c &= \frac{1}{pN}\sum_{i=1}^N (X_i-\widehat{\mu})^\top \widehat{\Sigma}_s^{-1}(X_i-\widehat{\mu}) \quad (2.6) \\ \widehat{\Sigma}_s &= \frac{1}{qN}\sum_{i=1}^N (X_i-\widehat{\mu})\widehat{\Sigma}_c^{-1}(X_i-\widehat{\mu})^\top \end{aligned}$$

The two covariance estimates depend on each other and thus their estimates must be computed in an iterative fashion until convergence.

Handling non-identifiability

If for any $\kappa \neq 0$ we define $\tilde{\Sigma}_c := \kappa\Sigma_c$ and $\tilde{\Sigma}_s := \Sigma_s/\kappa$ then $\tilde{\Sigma}_s \otimes \tilde{\Sigma}_c = \Sigma_s \otimes \Sigma_c$, and so both estimates yield the same overall covariance matrix. To resolve this non-identifiability issue we propose the following amendment to the model:

$$\mathrm{vec}(X_i) \sim N(\mathrm{vec}(\mu), \sigma^2\Sigma_s \otimes \Sigma_c) \quad (2.7)$$

and require that $(\Sigma_c)_{11} = 1$ and $(\Sigma_s)_{11} = 1$ (the choice of the top-left entry is arbitrary.) In this way we fix the scale of Σ_c and Σ_s, and estimate the scale of the overall covariance in σ^2.

The MLE of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{pqN} \sum_{i=1}^{N} (X_i - \widehat{\mu})^\top (\widehat{\Sigma}_s \otimes \widehat{\Sigma}_c)^{-1} (X_i - \widehat{\mu})$$

and depends on the other estimates. The MLE for μ is clearly the same as in (2.6) since we only changed the variance of the model. However, since the variance scale is now captured by σ^2 we need to scale the MLEs for Σ_c and Σ_s by their top-left entry at each iteration: if $\widehat{\Sigma}_c^*$ and $\widehat{\Sigma}_s^*$ are the estimates from (2.6) for Σ_c and Σ_s, then the respective MLEs for (2.7) are $\widehat{\Sigma}_c = \widehat{\Sigma}_c^*/(\widehat{\Sigma}_c^*)_{11}$ and $\widehat{\Sigma}_s = \widehat{\Sigma}_s^*/(\widehat{\Sigma}_s^*)_{11}$.

Finally, we remark that, according to [102, Theorem 3.1], if $N > \max(p, q)$ then the maximum likelihood estimates are unique.

2.2.2 Parameter Estimation with Missing Data

Estimation of the parameters in (2.7) is challenging because surface reflectance data from MODIS include significant proportions of missing values, which are largely caused by clouds (especially in the tropics) and large solar zenith angles at high latitudes. Missing data presents a difficult, albeit well-studied challenge in parameter estimation. Traditional methods, such as the EM algorithm, can usually handle missing data in a straightforward way. As the dimensionality increases, as in our case, the method can become quite computationally expensive. Naturally, we aim to assess different ways of achieving accurate parameter estimates with an eye towards reducing computation time. In the following section we derive and EM algorithm for obtaining parameter estimates in the presence of missing data. We assume that data is missing at random.

EM Algorithm for Matrix Normal Distribution

In the situation where missing data exists the EM algorithm is a convenient way to estimate the parameters $\Theta = (\mu, \Sigma_c, \Sigma_s, \sigma^2)$ in (2.7). Let us denote the data at a single pixel by $X = (Y, Z)$ where Z is the missing portion of X and Y is the observed portion of X. For the E-step, we need:

$$\begin{aligned} Q(\Theta; \Theta^{(t)}) &\doteq \mathbb{E}_{Z \mid Y; \Theta^{(t)}}[\log \mathbb{P}(X_1, \ldots, X_n; \Theta)] \\ &= \frac{pN}{2} \log |\Sigma_c^{-1}| + \frac{qN}{2} \log |\Sigma_s^{-1}| - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i \mathbb{E}_{Z \mid Y; \Theta^{(t)}}[D_\Sigma(X_i, \mu)], \end{aligned}$$

while the M-step updates Θ by maximizing Q,

$$\Theta^{(t+1)} \doteq \arg\max_{\Theta} Q(\Theta; \Theta^{(t)}),$$

via matrix differentiation in our case.

The Mahalanobis distance obeys a Pythagorean relationship: if $\tilde{\mu}_i^{(t)} := \mathbb{E}_{Z \mid Y; \Theta^{(t)}}[X_i]$, then

$$\mathbb{E}_{Z \mid Y; \Theta^{(t)}}[D_\Sigma(X_i, \mu)] = \mathbb{E}_{Z \mid Y; \Theta^{(t)}}[D_\Sigma(X_i, \mu_i^{(t)})] + D_\Sigma(\mu_i^{(t)}, \mu).$$

From here the update for μ follows from $\partial Q / \partial \mu = 0$:

$$\widehat{\mu}^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\mu}_i^{(t)},$$

similarly to the plain MLE case in (2.6).

Updating Σ_c, Σ_s and σ^2 requires a bit more work. To this end we focus, first, on the

following term:

$$\begin{aligned} R_i(\Theta;\Theta^{(t)}) &\doteq \mathbb{E}_{Z\,|\,Y;\Theta^{(t)}}[D_\Sigma(X_i,\mu_i^{(t)})] \\ &= \mathbb{E}\left[\mathrm{tr}[(\Sigma_s\otimes\Sigma_c)^{-1}\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})^\top]\right] \\ &= \mathrm{tr}\left[(\Sigma_s^{-1}\otimes\Sigma_c^{-1})\underbrace{\mathbb{E}[\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})^\top]}_{V_i^{(t)}}\right]. \end{aligned}$$

where we define the expected outer product

$$V_i^{(t)} \doteq \mathbb{E}_{Z\,|\,Y;\Theta^{(t)}}[\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})^\top].$$

To get the partial derivatives of Q with respect to Σ_c we need

$$\begin{aligned} \frac{\partial R_i}{\partial(\Sigma_c^{-1})_{kl}} &= \mathrm{tr}\left\{\frac{\partial}{\partial(\Sigma_c^{-1})_{kl}}[(\Sigma_s^{-1}\otimes\Sigma_c^{-1})V_i]\right\} \\ &= \mathrm{tr}\left\{\left(\Sigma_s^{-1}\otimes\underbrace{\frac{\partial\Sigma_c^{-1}}{\partial(\Sigma_c^{-1})_{kl}}}_{S_{kl}}\right)V_i\right\} \end{aligned}$$

where S_{kl} is the structure matrix [85] of a symmetric matrix, that is, $S_{kl}=[\delta_{ik}\cdot\delta_{jl}+\delta_{il}\cdot\delta_{jk}]_{ij}$. Thus, $(\Sigma_s^{-1}\otimes S_{kl})$ is a block matrix.

We can now look at V_i as a $q\times q$ block matrix where each element is a $p\times p$ matrix in the following way, for example:

$$V_{i,kl} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overset{\overset{b}{\downarrow}}{\begin{bmatrix} \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \circ & \circ & \circ & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \circ & \circ & \circ & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \circ & \circ & \circ & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \end{bmatrix}} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ b' \end{array} \tag{2.8}$$

So the matrix V_i being a block matrix leads to $V_{i,kl}$ being a symmetric $p \times p$ matrix with zeros at the entries corresponding to non-missing data in (2.8) (empty circles) and $\mathrm{Cov}_{Z_i \mid Y_i}(Z_{kb}, Z_{lb'})$ at the entries corresponding to missing entries (filled circles). Thus,

$$\begin{aligned} \frac{\partial R_i}{\partial (\Sigma_c^{-1})_{kl}} &= \mathrm{tr}[\Sigma_s^{-1}(V_{i,kl} + V_{i,kl}^\top)] \\ &= \textstyle\sum_{b,b' \in \mathsf{miss}(k,l)} (\Sigma_s^{-1})_{b,b'} \mathrm{Cov}_{Z_i \mid Y_i, \Theta^{(t)}}[Z_{kb}, Z_{lb'}]. \end{aligned}$$

Here $\mathsf{miss}(k,l)$ are the row-column pairs for which there are missing entries in V_i, and $V_{i,kl}$ is the $p \times p$ block submatrix of V_i from rows $(k-1)p+1$ to kp and columns $(l-1)p+1$ to lp. Note that $\partial R_i / \partial (\Sigma_c^{-1})_{k,l}$ does not depend on Z_i. Moreover, the conditional covariances in $\mathrm{Var}_{Z_i \mid Y_i, \Theta^{(t)}}[Z_i]$ above can be obtained by applying the SWEEP operator [41] to the rows of $\Sigma_s^{-1} \otimes \Sigma_c^{-1}$ that correspond to missing values.

Thus, from solving $\partial Q / \partial (\Sigma_c^{-1}) = 0$, we have

$$\widehat{\Sigma}_c^{(t+1)} = \frac{1}{\widehat{\sigma}^{2(t)} pN} \sum_i \left(\frac{\partial R_i}{\partial \Sigma_c^{-1}} + (\tilde{\mu}_i^{(t)} - \mu^{(t)})^\top \Sigma_s^{-1^{(t)}} (\tilde{\mu}_i^{(t)} - \mu^{(t)}) \right). \tag{2.9}$$

Similarly, for $\widehat{\Sigma}_s$,

$$\widehat{\Sigma}_s^{(t+1)} = \frac{1}{\widehat{\sigma}^{2(t)} qN} \sum_i \left(\frac{\partial R_i}{\partial \Sigma_s^{-1}} + (\tilde{\mu}_i^{(t)} - \mu^{(t)}) \Sigma_c^{-1^{(t)}} (\tilde{\mu}_i^{(t)} - \mu^{(t)})^\top \right). \tag{2.10}$$

Finally, for σ^2,

$$\widehat{\sigma}^{2^{(t+1)}} = \frac{1}{pqN} \sum_i R_i + \mathrm{vec}(\tilde{\mu}_i^{(t)} - \widehat{\mu}^{(t)})^\top (\widehat{\Sigma}_s^{(t)} \otimes \widehat{\Sigma}_c^{(t)})^{-1} \mathrm{vec}(\tilde{\mu}_i^{(t)} - \widehat{\mu}^{(t)}). \tag{2.11}$$

Just as in the last section, we normalize these covariance estimates by their upper-left entry; i.e. fixing $(\widehat{\Sigma}_c)_{11} = 1$ and $(\widehat{\Sigma}_s)_{11} = 1$. The algorithm for this EM procedure can be seen in Algorithm 1, with `R` code included in the supplementary material. In Algorithm 1 we denote by $\circ$ the Hadamard (element-wise) product. We further denote by $A[m_s(i)]$ the

Algorithm 1: Matrix Normal Expectation-Maximization

Initialize: $\mu^{(1)} \leftarrow 0_{p\times q}$; $\Sigma_c^{(1)} \leftarrow I_q$; $\Sigma_s^{(1)} \leftarrow I_p$;
for $t \leftarrow 1, \ldots$ *(until convergence)* **do**
  $\mu^{(t+1)} \leftarrow 0_{p\times q}$; $\Sigma_c^{(t+1)} \leftarrow 0_q$; $\Sigma_s^{(t+1)} \leftarrow 0_p$;
  for $i \leftarrow 1, \ldots, n$ **do**
    `// Initialize missing entry indices and conditional covariance:`
    Set $\mathsf{miss}(i)$ as the indices of missing entries in the i-th observation;
    Set $m_s(i)$ and $m_c(i)$ as the row and column indices of the entries in $\mathsf{miss}(i)$;
    SWEEP the $\mathsf{miss}(i)$ rows of $R = \Sigma_c^{-1^{(t)}} \otimes \Sigma_s^{-1^{(t)}}$; `// Compute` $\mathrm{Var}_{Z_i \mid Y_i}[X_i]$

    `// Set auxiliary variables:`
    $X_m \leftarrow \begin{bmatrix} Y_i \\ \mu^{(t)}[\mathsf{miss}(i)] + R[-\mathsf{miss}(i), \mathsf{miss}(i)]^\top (Y_i - \mu^{(t)}[-\mathsf{miss}(i)]) \end{bmatrix}$;
    `//` $\mathbb{E}_{Z_i \mid Y_i}[X_i]$
    $R_m \leftarrow R[\mathsf{miss}(i), \mathsf{miss}(i)]$;
    $S_{s,m} \leftarrow \Sigma_s^{-1^{(t)}}[m_s(i), m_s(i)]$;
    $S_{c,m} \leftarrow \Sigma_c^{-1^{(t)}}[m_c(i), m_c(i)]$;

    `// Update:`
    $\mu^{(t+1)} \leftarrow \mu^{(t+1)} + X_m$;
    $\Sigma_c^{-1^{(t+1)}} \leftarrow$
    $\Sigma_c^{-1^{(t+1)}} + \mathbf{e}(m_c(i))^\top \cdot (R_m \circ S_{c,m}) \cdot \mathbf{e}(m_c(i)) + (X_m - \mu^{(t)})^\top \Sigma_s^{-1^{(t)}} (X_m - \mu^{(t)})$;
    $\Sigma_s^{-1^{(t+1)}} \leftarrow$
    $\Sigma_s^{-1^{(t+1)}} + \mathbf{e}(m_s(i))^\top \cdot (R_m \circ S_{s,m}) \cdot \mathbf{e}(m_s(i)) + (X_m - \mu^{(t)}) \Sigma_c^{-1^{(t)}} (X_m - \mu^{(t)})^\top$;
    $\sigma^{2^{(t+1)}} \leftarrow \sigma^{2^{(t+1)}} + \sum_j \mathrm{vec}(R_m \circ S_{c,m} \circ S_{s,m}) +$
    $\mathrm{tr}(\mathrm{vec}(X_m - \mu^{(t)})^\top (\Sigma_s^{-1^{(t)}} \otimes \Sigma_c^{-1^{(t)}}) \mathrm{vec}(X_m - \mu^{(t)}))$;
  end

  `// Scale:`
  $\mu^{(t+1)} \leftarrow \mu^{(t+1)}/n$;
  $\Sigma_c^{(t+1)} \leftarrow \Sigma_c^{(t+1)}/\Sigma_{c,11}^{(t+1)}$;
  $\Sigma_s^{(t+1)} \leftarrow \Sigma_s^{(t+1)}/\Sigma_{s,11}^{(t+1)}$;
  $\sigma^{2^{(t+1)}} \leftarrow \sigma^{2^{(t+1)}}/(nBT)$;
end

rows of A indexed by $m_s(i)$, and by $A[-m_s(i)]$ the rows of A there are *not* in $m_s(i)$. A similar notation is used to subset columns based on index sets.

In the context of land cover classification the likelihood for the data at pixel v actually takes the following form:

$$\text{vec}(X_v) \,|\, \theta_v = c \sim N(\text{vec}(\mu_c), \sigma_c^2 \Sigma_s \otimes \Sigma_c). \tag{2.12}$$

As alluded to before, the mean profile μ_c and the temporal covariance Σ_c characterize class specific information. Because we revised the model to handle non-identifiability, the scale of the covariance (σ^2) will also depend on the land cover class. We use training data, independent of the image of interest, to estimate the parameters in (2.12) prior to the label assignment procedure.

2.3 Missing Data Imputation

Ideally, we could obtain posterior probabilities of the pixel labels using

$$\mathbb{P}(\theta_v \,|\, X_v) = \int \mathbb{P}(\theta_v, z_v) dz_v. \tag{2.13}$$

Because integrating out the missing values this way is infeasible, we could obtain a pixel label by approximating (2.13) with an EM algorithm with the following M-step where we condition on the observed data, Y_v:

$$\theta_v^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{Z_v \,|\, Y_v, \theta^{(t)}}[\log(X_v; \theta^{(t)})] \tag{2.14}$$

Performing classification in this way remains computationally prohibitive due to the tedious, iterative nature of this approach. To alleviate a portion of the computational burden associated with classifying pixels while accounting for missing data, we propose a pre-processing step to impute missing values. A one-time imputation pre-processing step enriches our data and allows for an independent choice of classification method should we

desire it. Using the estimates for the parameters in (2.7), we derive a second expectation-maximization procedure for imputing the missing values in each pixel. This procedure will impute missing data for each pixel independent of the other pixels in the image and in the pre-established Bayesian vein, utilizes a prior distribution on the land cover classes, $\mathbb{P}(\theta_v)$. Treating the pixel label, θ_v, as latent we compute an update for the missing data, Z_v, for pixel v and iterate until convergence. Let us again denote the data at pixel v by $X_v = [\, Y_v \quad Z_v \,]$.

Using the likelihood in (2.7) along with the global prior $\mathbf{h}$ ($\mathbb{P}(\theta_v = l) = h_l$) we have the following E-step:

$$\begin{aligned} Q(Z, Z^{(t)}) &= \mathbb{E}_{\boldsymbol{\theta}\,|\,Y,Z^{(t)}}[\log(\textstyle\prod_v \prod_l (\mathbb{P}(X_v \,|\, \theta_v = l)\mathbb{P}(\theta_v = l))^{\mathbf{I}(\theta_v = l)})] \\ &= \textstyle\sum_v \sum_l \mathbb{P}(\theta_v = l \,|\, Y, Z_v^{(t)}) \mathbb{E}_{\theta\,|\,Y,Z^{(t)}}[\log \mathbb{P}(X_v \,|\, \theta_v = l)] + \mathbb{P}(\theta_v = l \,|\, Y, Z_v^{(t)}) \log \mathbf{h}_l \\ &= \textstyle\sum_v \sum_l \mathbb{P}(\theta_v = l \,|\, Y, Z_v^{(t)}) \cdot \left[\log \mathbb{P}(X_v \,|\, \theta_v = l) + \log \mathbf{h}_l \right] \end{aligned}$$

Let $\Sigma_l = \sigma_l^2 \cdot (\Sigma_s \otimes \Sigma_{t,l})$. Then the derivative of $Q(Z, Z^{(t)})$ with respect to Z is

$$\frac{\partial Q(Z, Z^{(t)})}{\partial Z_v} = -\sum_l \mathbb{P}(\theta_v = l \,|\, Y, Z^{(t)}) \cdot \left[(\Sigma_l^{-1})_{YZ}^\top (Y - \mu_{l,Y}) + (\Sigma_l^{-1})_{ZZ} (Z - \mu_{l,Z}) \right].$$

With some algebra we arrive at our corresponding M-step:

$$\begin{aligned} Z_v^{(t+1)} = \left[\sum_l \mathbb{P}(\theta_v = l \,|\, Y_v, Z_v^{(t)}) (\Sigma_l^{-1})_{zz} \right]^{-1} \Bigg[\sum_l \mathbb{P}(\theta_v = l \,|\, Y_v, Z_v^{(t)}) (\Sigma_l^{-1})_{zz} \cdot \\ \Big(\mu_{l,z_v} - (\Sigma_l^{-1})_{zz}^{-1} (\Sigma_l^{-1})_{yz}^\top (Y_v - \mu_{l,y_v}) \Big) \Bigg] \end{aligned} \tag{2.15}$$

While the imputation of missing data involved computing posterior probabilities, $\mathbb{P}(\theta_v = l \,|\, Y_v, Z_v^{(t)})$, we did not utilize these to make inference of the pixel labels. We will, however, make use of them later in Section 2.7 to assess the quality of training data. With the

observation at each pixel completed with this imputation, we proceed to further address the size of the data.

2.4 Data Compression

Land cover classification has been approached in many different ways for quite some time. Traditionally, principal components analysis has been used to reduce the dimensionality of the data. However, usually all spectral-temporal features are considered distinct features. Consequently, spectral and temporal variation cannot be isolated from an eigen-decomposition of the full $(\Sigma_s \otimes \Sigma_c)$ covariance matrix, as required from PCA.

We ultimately seek labels for pixels across the globe, roughly 1.8 billion pixels. Due to significant correlation between the 7 MODIS land bands and a desire to further reduce the dimensionality of the data, we envision utilizing a small combination of the spectral features that possess most of the variation in the data. To isolate the most useful information present in the data, in a lower dimensional space, we pursue a special application of principal components analysis. Because spectral variation transcends the differences between land-cover classes (our C labels), we target $\widehat{\Sigma}_s$ with PCA, as opposed to the entire model covariance.

Because seasonal variation is a first-order property that helps to distinguish many land cover classes (e.g., deciduous broadleaf forests), multitemporal information is important for land cover classification. Thus, here we target multicollinearity in spectral bands for reducing feature dimensionality. Specifically, because of the variance structure in (2.7), it is possible to eigen-decompose only Σ_s (the spectral band covariance). In doing so, our goal is to isolate linear combinations of bands that maximize spectral information and minimize correlation between spectral bands, while at the same time retaining the variance associated with temporal dynamics in the original data. Let P be the $B \times K$ matrix of eigenvectors of $\widehat{\Sigma}_s$. Then, using $K < B$ principal components (PC)

$$X_{v,pc} = P^T X_v$$

and the covariance matrix becomes

$$(P^T \Sigma_s P) \otimes \Sigma_c,$$

where the transformed observation $X_{v,pc}$ is a $K \times T$ matrix. Therefore, with $\mu_{c,pc} \doteq P^T \mu_c$ (the PC transformed mean) and $\Sigma_{s,pc} \doteq P^T \Sigma_s P$ (the PC transformed band variance matrix—a diagonal matrix with the eigenvalues of Σ_s on the diagonal) the model in (2.12) becomes:

$$\text{vec}(X_{v,pc}) \,|\, \theta_v = c \sim N(\text{vec}(\mu_{c,pc}), \sigma_c^2 \Sigma_{s,pc} \otimes \Sigma_c). \tag{2.16}$$

This model reduces the dimensionality related to correlation among spectral bands while retaining the variance associated with multitemporal patterns in the data. For the remainder of this document we will drop the *pc* subscript in (2.16) and refer only to the likelihood in (2.12) under the assumption that the data has been imputed and transformed into the space of the first K principal components. That is, X and μ_c will now be $K \times 28$ matrices. The remote sensing results in Section 2.6 indicate that the first three ($K = 3$) principal components capture most of the variation in the data.

2.5 Parameter Estimation Simulation Study

In this section we compare our derived EM procedure ("*EM*"), in Algorithm 1, to two other popular choices for parameter estimation in the presence of missing data. The first approach (which we label "*MM*") applies a maximization in two ways: (1) "imputation" of missing values and (2) maximum likelihood parameter estimation. In particular, the missing values get replaced by the most recent estimate of the mean in step (1). The next iteration of mean and covariance estimates come from the same maximum likelihood expressions in (2.6), with the addition of $\widehat{\sigma}^2$, based on the fully imputed data. The ease and simplicity of this method make it a natural first step in handling missing data, but also hinder its robustness and ability to capture all of the uncertainty associated with missing

data.

The second approach ("GEM") applies the EM algorithm to the most general version of the model. As opposed to estimating the parameters of (2.7), the EM algorithm provides parameter estimates for the following model:

$$\text{vec}(X_i) \sim N(\text{vec}(\mu), \Sigma). \tag{2.17}$$

These multivariate normal EM estimates have the same form of those found in [74]. This approach does not simplify the original problem since it requires more parameters to be estimated by not assuming the Kronecker structure. The simple form of (2.17) attracts much attention, but its complexity far exceeds that of (2.7). Where sources of variation in the data can be naturally partitioned, such as in space or time, the Kronecker structure surpasses (2.17) in both interpretability and computational efficiency.

To empirically assess the model and algorithm, we simulated data from a matrix normal distribution with randomly chosen parameters of dimensions: $p = 3$ and $q = 5$, and $p = 3$ and $q = 7$. Three sample sizes were used: 250, 500, and 1000. Four different proportions of missing data were used: 5%, 10%, 15% and 20%. Data was simulated 100 times at each combination of sample size and proportion of missing data to evaluate how the accuracy of the estimates vary. The three different algorithms described above (MM, GEM, and EM) were run in each of these combinations to provide a richer comparison. In order to compare these methods, the covariance errors were always measured with respect to the full (Kronecker product) covariance matrix.

The relative errors of the mean estimates across the three methods and the four different proportions of missing data were consistently low, as seen in Figure 2.5. The methods differ very little when it comes to the estimate of the mean, and so we focus on the variance estimates. Indeed, the models and estimation procedure vary most when dealing with the covariance matrix.

Figure 2.2 tells a rich story about how these three methods differ most. Since the MM

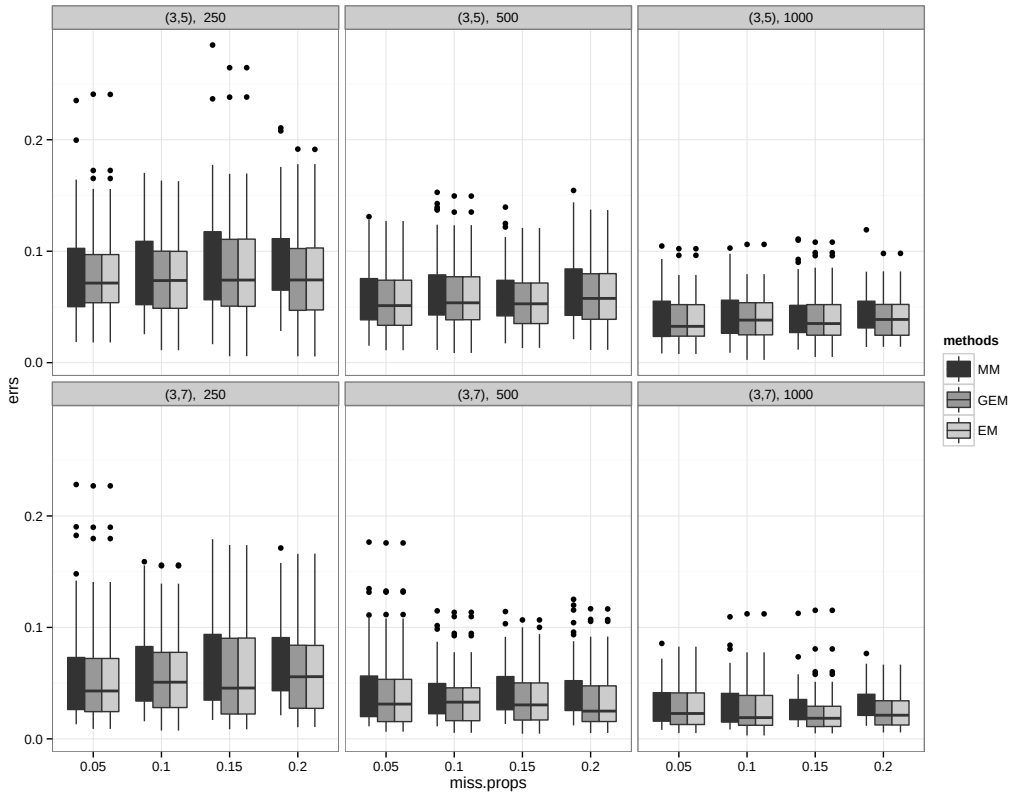


Figure 2.1: Boxplots of the relative errors in the estimates of μ, in order of MM, GEM, and EM.

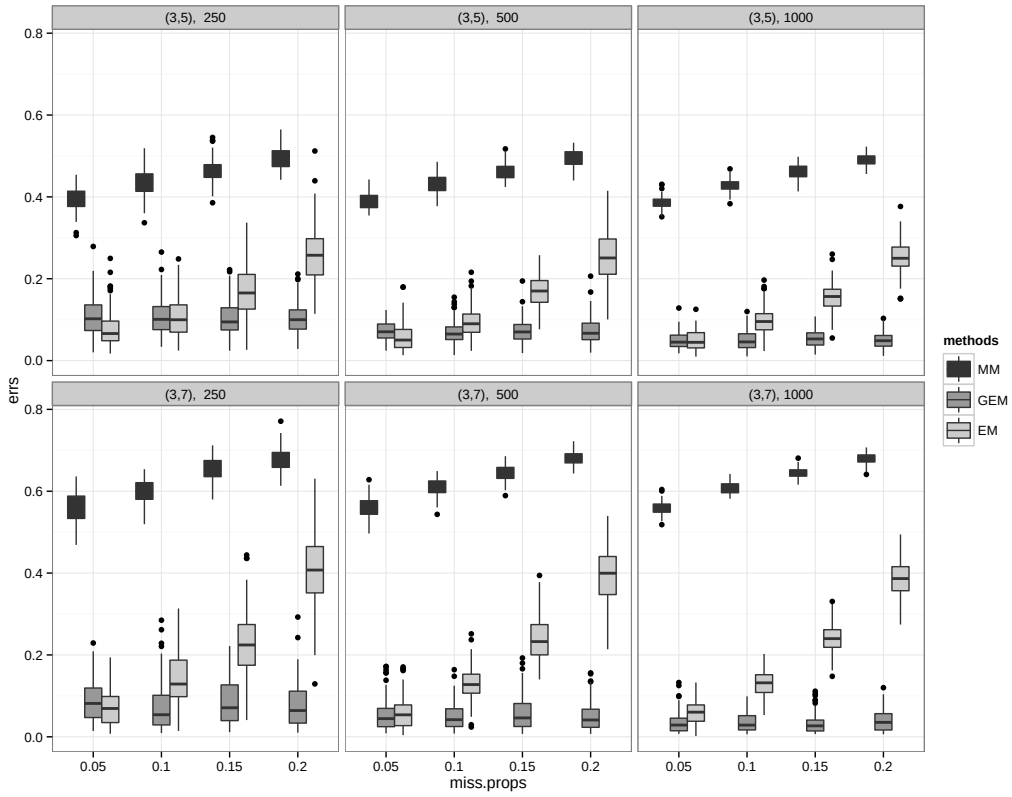


Figure 2.2: Boxplots of the relative errors in the estimates of Σ, in order of MM, GEM, and EM.

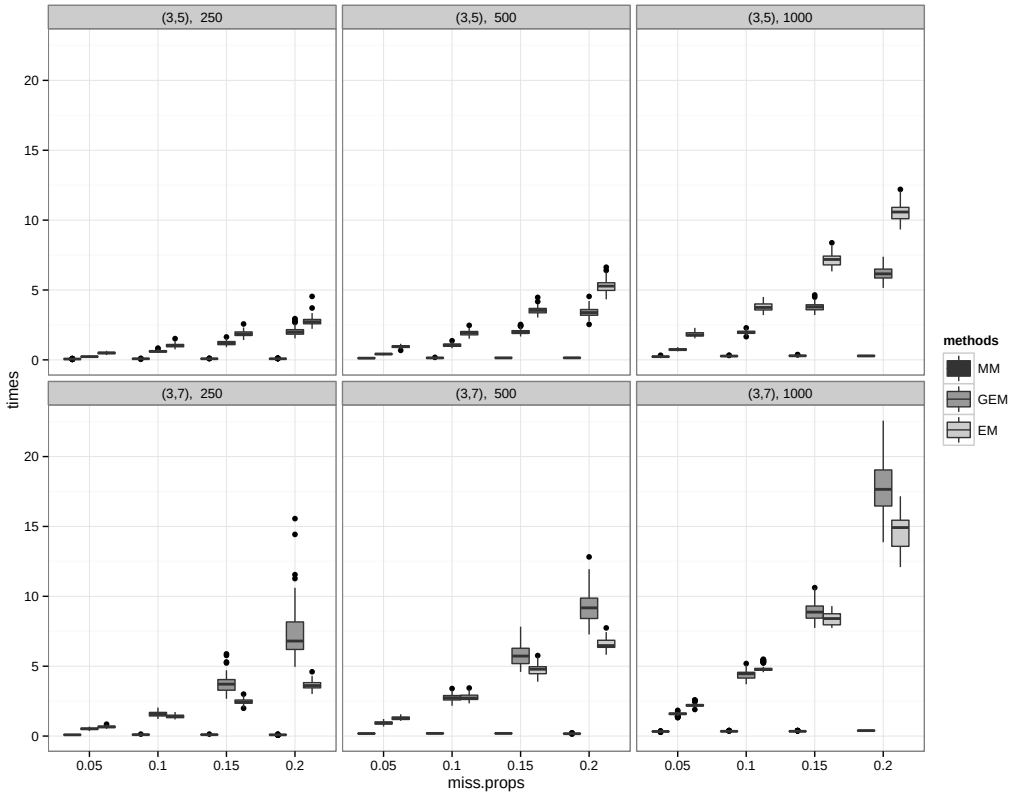


Figure 2.3: Boxplots of the run times (in seconds), in order of MM, GEM, and EM.

method essentially treats the imputed values as actual data and fails to account for all of the uncertainty present, the errors for this method top all of those from the GEM and EM. As the sample size increases the estimates appear to improve slightly, but the most interesting feature lies in the difference between the GEM and EM methods. Since (2.17) contains more parameters to be estimated, this model can achieve better resolution and accuracy than (2.7), albeit needing more samples to identify parameters. The significant cost lies in the computation time, making the Kronecker structure a worthwhile consideration since it noticeably reduces the complexity of the model and still achieves accurate parameter estimates.

Figure 2.3 introduces the computational differences between the methods. The MM method, while still requiring some iterating, takes very little time in all scenarios. Interestingly, the EM method requires the most time for lower dimensions such as the $p = 3$, $q = 5$ situation simulated here because of the iterative algorithm needed to estimate the two components of the Kronecker product. When the dimensions increase, we begin to see the gains in the Kronecker model. Naturally, as the discrepancy in the number of parameters being estimated by GEM and EM increases, the computational advantages become more significant. Additionally, the difference in sample sizes necessary to estimate the covariance grows, with GEM requiring more.

Of course, this presumes the choice between the two models. In a situation where the physical dimensions of the data imply a Kronecker structure, we can take comfort in the above results. One such example is our application in Remote Sensing.

2.6 Parameter Estimation and Data Compression Case Study

Our rich MODIS datasets are both multispectral and multitemporal. For the analysis that follows we used a subset of the MODIS Land Cover Training site database that includes 204 sites located over the northeastern portion of the conterminous United States. These sites include 2,733 MODIS pixels and encompass most major biomes and land cover types in the

Table 2.1: IGBP classes and their representativeness in the northeast training dataset. The last column lists the prior probability of a pixel belonging to each class.

IGBP	Description	# Pixels	Global Prior (α)
1	Evergreen Needleleaf Forests	211	0.17
4	Deciduous Broadleaf Forests	131	0.02
5	Mixed Forests	184	0.14
7	Open Shrublands	95	0.08
8	Woody Savannas	66	0.05
9	Savannas	93	0.03
10	Grasslands	327	0.17
11	Permanent Wetlands	177	0.04
12	Croplands	918	0.12
13	Urban and Built-Up Lands	207	0.01
14	Cropland/Natural Veg. Mosaics	124	0.07
15	Permanent Snow and Ice	7	-
16	Barren/Sparsely Veg. Lands	34	-
17	Water Bodies	159	0.10

lower 48 United States [95]. Table 2.1 summarizes the frequency of land cover classes in our training dataset along with the prior to be used in missing data imputation. Classes 15 and 16 were excluded because they are rare in the study area. The final data set included 12 land-cover classes across 2,692 pixels. Table 2.2 and Figure 2.4 give the results of our PCA.

About 5% of our data is missing, making the Kronecker structure and our proposed EM algorithm even more ideal, from a purely computational standpoint, as evidenced by the simulation study above. Using the proposed EM algorithm we estimate the parameters of (2.7) and perform PCA on our estimate of Σ_s. To, again, provide a comparison which confirms the quality of our EM we compare the PCA results just described to those computed using the MM method as well as to a principal components analysis on the full (196×196) covariance matrix.

Figures 2.5 and 2.6 show the data projected into the space of the first two principal components using the MM and EM methods. The class separability seems reasonable and

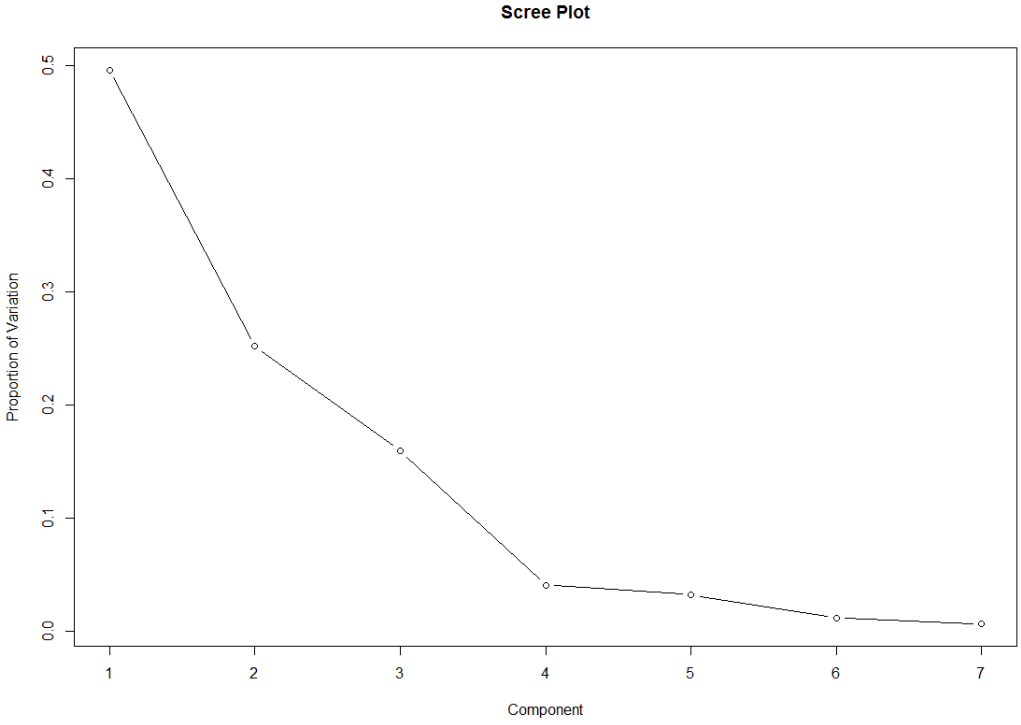


Figure 2.4: Scree plot for PCA of northeast training data. Plot shows proportion of variation versus the principal components.

Table 2.2: First three PCs for training dataset and the proportion of variance explained by each.

Band	PC_1	PC_2	PC_3
1	-0.252	0.209	-0.519
2	-0.563	-0.645	-0.022
3	-0.179	0.085	-0.512
4	-0.233	0.089	-0.485
5	-0.555	-0.102	0.349
6	-0.391	0.484	0.309
7	-0.269	0.530	0.126
Prop. Var.	0.497	0.252	0.160

the different amounts of variation in each class are certainly identified. A closer inspection reveals that the *EM* method achieves better separability of the land cover classes. This can be seen in the PC1 versus PC2 plots (Figures 2.5, 2.6, and 2.7) as well as their dendrograms

Figure 2.5: Northeast, North America training data projected into the space of the first two principal components, colored by their respective land color classes using *MM*.

in Figure 2.8.

The dendrograms in Figure 2.8 resulted from a hierarchical clustering of the classes using the following distance metric, d_{ij}:

$$d_{ij} = (\mu_i - \bar{\mu}_{ij})^\top \Sigma_i^{-1} (\mu_i - \bar{\mu}_{ij}) + (\mu_j - \bar{\mu}_{ij})^\top \Sigma_j^{-1} (\mu_j - \bar{\mu}_{ij}).$$

where the $\bar{\mu}_{ij}$ are "in-between" centers for clusters i and j,

$$\bar{\mu}_{ij} = \Sigma_i^{-1}(\Sigma_i^{-1} + \Sigma_j^{-1})^{-1}\mu_i + \Sigma_j^{-1}(\Sigma_i^{-1} + \Sigma_j^{-1})^{-1}\mu_j.$$

The log distances are plotted above along with an overall log distance, $D = \sum_{i<j} d_{ij}$. As a measure of the overall separability of the classes in the space of the first two principal components, the greater value of $\log(D)$ achieved by the EM confirms that this method

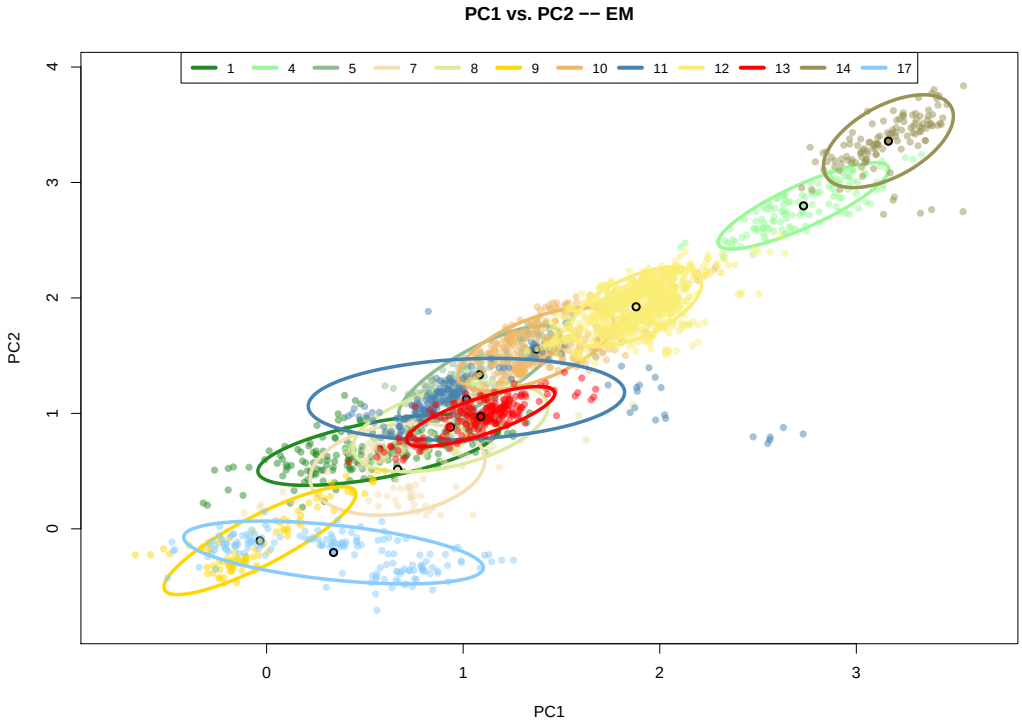


Figure 2.6: Northeast, North America training data projected into the space of the first two principal components, colored by their respective land color classes using *EM*.

helped to better distinguish land cover classes.

Although the decomposition of the covariance matrix in (2.12) is new, the results of the PCA were consistent with established correlation structure among the seven MODIS land bands. Specifically, $K = 3$ principal components captured 90.9% of the spectral variance (Table 2.2) and correspond, roughly, to the 'Brightness', 'Wetness', and 'Greenness' of the Tasseled-Cap Transformation for MODIS data [63]. As evidenced by the scree plot in Figure 2.4, subsequent components capture significantly less spectral variation. Therefore, we happily proceed with transforming the data into the space of the first three ($K = 3$) principal components.

We use this framework below to perform land cover classification, to assess the homogeneity and composition of training sites used in this analysis, and to assess site membership of pixels in the training data.

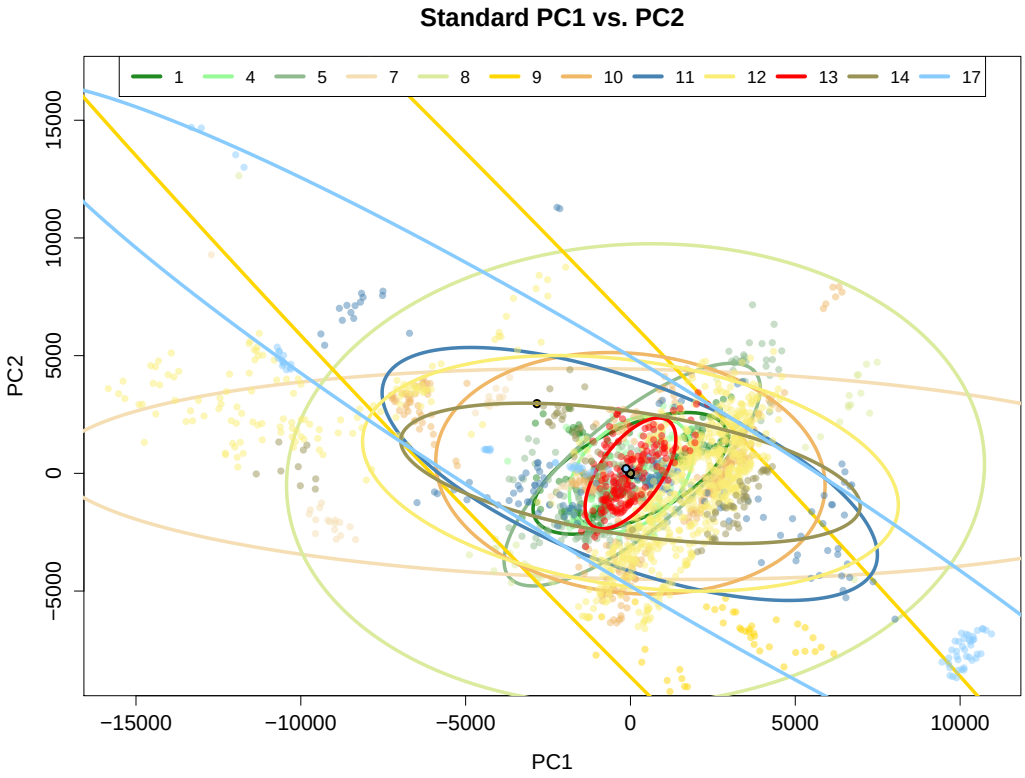


Figure 2.7: First and second principal component scores for northeast, North America training pixels performed using a standard PCA based on the full covariance matrix with all 196 features. Note the scale of the PC scores.

2.7 Training Data Assessment

Supervised classification methods, which are widely used to map land cover from remotely sensed data, require high quality exemplars to train classification algorithms. Conventionally, individual pixels or training sites composed of multiple pixels are selected for this purpose. In many cases, however, selection of training data is not straightforward because class definitions are ambiguous or because landscape heterogeneity introduces sub-pixel variability in land cover. Both of these issues are relevant to the case we consider here: classification of 500-m spatial resolution MODIS pixels based on the IGBP classes used by the MODIS land cover type product [32]. Specifically, the degree to which training sites provide good exemplars of the classes to be classified exerts strong influence on classification

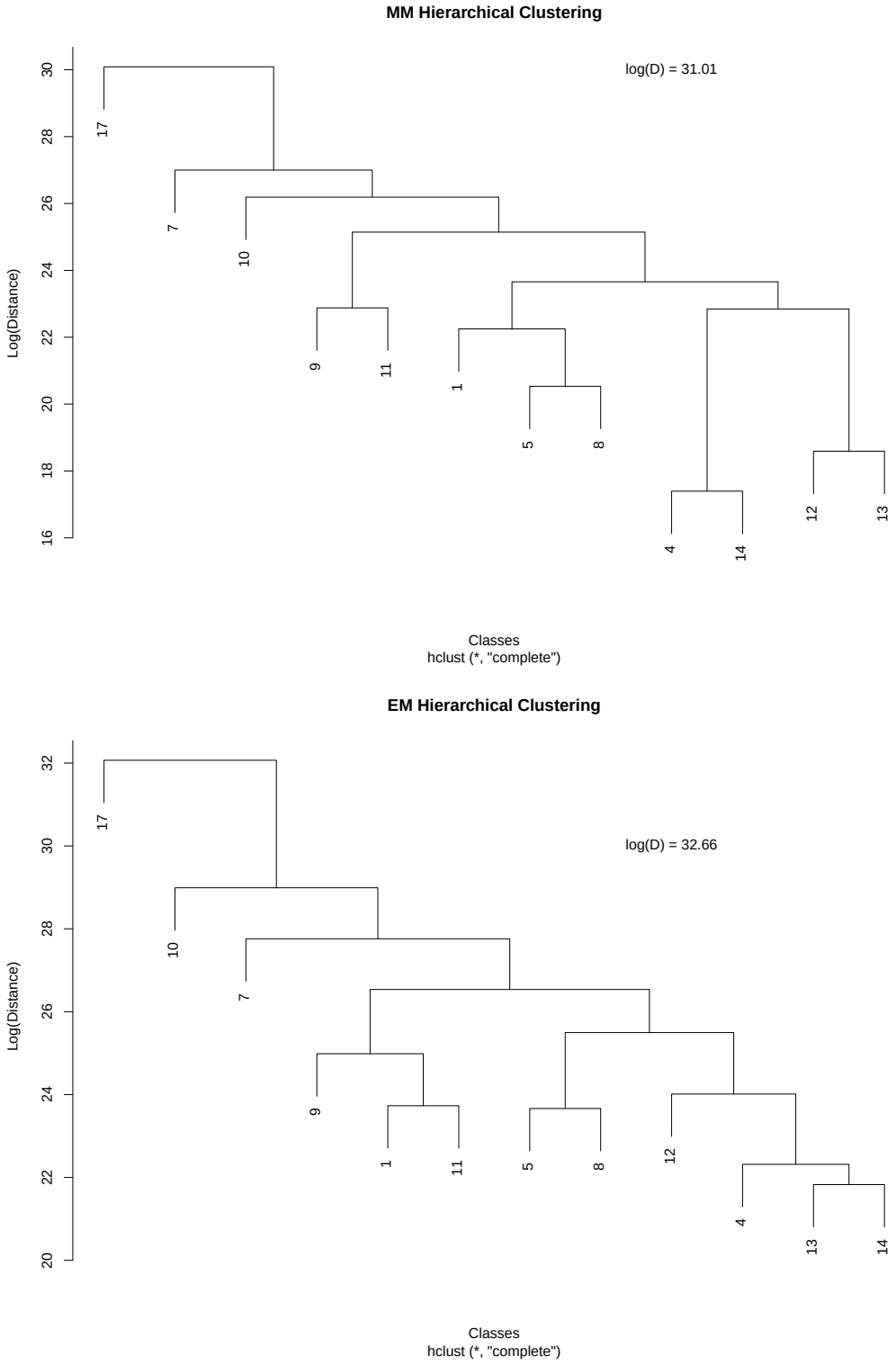


Figure 2.8: Hierarchical clustering of classes using *MM* and *EM*.

results.

We designate a *single* testing dataset comprised of 36 sites (3 from each class) and train the parameters in (2.12) from the remainder of the data in our northeast, North America training data set (as described in Section 2.6. Sampling three sites from each class yielded a testing data set of 447 pixels and a training data set of 2,245 pixels. We assess site composition, site membership, and site homogeneity for this testing dataset in Section 2.7.4.

2.7.1 Site Composition

To address the issues described above, we propose a strategy that characterizes the class composition and internal variability of a particular site using a vector (π), where π_c (c^{th} component of π) represents the proportion of a site that corresponds to land-cover class c. To do this, we use EM to estimate π for each site using the previously estimated parameters in (2.12). The estimated land cover class for the site is then based on its composition π; a common choice, for instance, is the most likely class (i.e., $\arg\max_c \pi_c$).

Similar to how we estimated the posterior probabilities for land cover classes, we treat the site composition π as random. We would expect π to be, on average and *a priori*, similar to the prior probabilities $\mathbf{h}$ in the previous section; that is, $\mathbb{E}[\pi] = \mathbf{h}$. To specify a (prior) distribution for the site composition we still need to characterize its variability. To this end, we define another parameter n_0, which measures the "strength" of prior belief. We then adopt a Dirichlet distribution for the site composition π based on these two parameters:

$$\pi \sim \mathsf{Dir}(n_0, \mathbf{h}).$$

Given observations $X_1, \ldots, X_n$ for n pixels in a site, our goal is to characterize the posterior distribution of π given $X_1, \ldots, X_n$. To do this, we first obtain the posterior mode

π^* and assume that

$$\pi \,|\, X_1, \ldots, X_n \sim \mathsf{Dir}\left(n_0 \mathbf{h} + \sum_v P(\theta_v \,|\, X_v; \pi^*) \right). \tag{2.18}$$

Our method then iterates by updating the site class composition at step t, $\pi^{(t)}$, for each class, until convergence,

$$\pi_c^{(t+1)} = \frac{\sum_{v=1}^{n} \mathbb{P}(\theta_v = c \,|\, X_v; \pi^{(t)}) + n_0 h_c - 1}{n + n_0 - C}, \tag{2.19}$$

where the conditional posteriors $P(\theta_v = c \,|\, X_v; \pi^{(t)})$ evaluated at t are specified as in (3.3), and C is the number of land-cover classes. We can arrive at (2.19) via the following derivation.

Our model is the following,

$$\begin{aligned} X_v \,|\, \theta_v = c &\sim N(\mu_c, \Sigma_c) \\ \theta_v \,|\, \pi &\sim \mathsf{MN}(1; \pi) \\ \pi &\sim \mathsf{Dir}(\mathbf{h}^*) \end{aligned}$$

where $\mathbf{h}^* = n_0 \mathbf{h}$ and $\Sigma_c = \Sigma_s \otimes \Sigma_{t,c}$. The log-likelihood is then

$$\begin{aligned} \log \mathbb{P}(X, \theta; \pi) = \log \left\{ \prod_v \mathbb{P}(X_v \,|\, \theta_v) \mathbb{P}(\theta_v \,|\, \pi) \mathbb{P}(\pi) \right\} \\ = \sum_v \sum_c I(\theta_v = c) \Big[\log \mathbb{P}(X_v \,|\, \theta_v = c) \\ + \log \mathbb{P}(\theta_v = c \,|\, \pi) \Big] + \sum_c (h_c^* - 1) \log \pi_c \end{aligned}$$

up to a constant. We now define

$$Q(\pi; \pi^{(t)}) := \mathbb{E}_{\theta \,|\, X; \pi^{(t)}} \Big[\log \mathbb{P}(X, \theta; \pi) \Big]$$

$$= \sum_v \sum_c \mathbb{P}(\theta_v = c \,|\, X; \pi^{(t)}) \Bigg[\log \mathbb{P}(X_v \,|\, \theta_v = c) + \log \pi_c \Bigg] + \sum_c (h_c^* - 1) \log \pi_c$$

to obtain the M-step as

$$\pi^{(t+1)} := \underset{\pi : \sum_c \pi_c = 1}{\arg\max}\, Q(\pi; \pi^{(t)}).$$

Adopting a Lagrange multiplier formulation, we next define

$$\tilde{Q}(\pi; \pi^{(t)}) := Q(\pi; \pi^{(t)}) + \lambda \left(1 - \sum_c \pi_c \right).$$

Then,

$$\frac{\partial \tilde{Q}}{\partial \pi_c} = \frac{\sum_v \mathbb{P}(\theta_v = c \,|\, X; \pi^{(t)}) + h_c^* - 1}{\pi_c} - \lambda,$$

and so, by setting $\partial \tilde{Q} / \partial \pi_c = 0$, we get

$$\pi_c^{(t+1)} = \frac{\sum_v \mathbb{P}(\theta_v = c \,|\, X; \pi^{(t)}) + h_c^* - 1}{\lambda}.$$

Finally, solving for λ when $\sum_c \pi_c^{(t+1)} = 1$ we have

$$\lambda^{-1} = \sum_c \sum_v \mathbb{P}(\theta_v = c \,|\, X; \pi^{(t)}) + h_c^* - 1 = n + n_0 - C$$

and thus

$$\pi_c^{(t+1)} = \frac{\sum_v \mathbb{P}(\theta_v = c \,|\, X; \pi^{(t)}) + h_c^* - 1}{n + n_0 - C}$$

as in (2.19).

2.7.2 Site Membership

The presence of outliers in training data presents a number of challenges. Most importantly, depending on the algorithm being used to perform the classification, outliers in training

data can introduce error to classification results. In addition, training data are widely used to assess classification accuracies via cross-validation. Hence, the presence of outliers can bias both classifier and cross-validation results.

Fortunately, in addition to providing a powerful way to characterize the class composition of training sites, the model we describe above provides a straightforward way to assess the site membership of pixels and suggest potential outliers in training data. Specifically, we first note that if a pixel v belongs to class c, then, since X_v follows a normal distribution according to (2.12), the Mahalanobis distance

$$D_c(X_v) = \text{vec}(X_v - \mu_c)^T(\Sigma_s \otimes \Sigma_c)^{-1}\text{vec}(X_v - \mu_c)$$

follows a chi-square distribution with KT degrees of freedom. In fact, the statistic above is just the sum of squares of observations that have been decorrelated across both multispectral and multitemporal space, as detailed here.

If the observations $\text{vec}(\tilde{X}) \sim N(\text{vec}(\mu), \Sigma_s \otimes \Sigma)$ were classified as realizations from class c we can test $H_o : \mu = \mu_c, \Sigma = \Sigma_c$ against $H_a : \mu \neq \mu_c$ or $\Sigma \neq \Sigma_c$. Since under the null hypothesis $\text{vec}(X) \sim N(\text{vec}(\mu_c), \Sigma_s \otimes \Sigma_c)$, then

$$D_c(\tilde{X}) = \text{vec}(\tilde{X} - \mu_c)^T(\Sigma_s \otimes \Sigma_c)^{-1}\text{vec}(\tilde{X} - \mu_c) \sim \chi^2_{KT},$$

and so we have a chi-square test. The test statistic above is just the sum of squares of observations that have been decorrelated across time and principal component scores; if

$u = X - \mu_c$ and L_c and L_s are the (lower triangular) Cholesky factors of Σ_c and Σ_s, then

$$\begin{aligned}
\mathrm{vec}(u)^T(\Sigma_s \otimes \Sigma_c)^{-1}\mathrm{vec}(u) = & \\
&= \mathrm{vec}(u)^T(\Sigma_s^{-1} \otimes \Sigma_c^{-1})\mathrm{vec}(u) \\
&= \mathrm{vec}(u)^T\mathrm{vec}(\Sigma_s^{-1}u\Sigma_c^{-1}) \\
&= \mathrm{vec}(u)^T\mathrm{vec}(L_s^{-T}\underbrace{L_s^{-1}uL_c^{-T}}_{w}L_c^{-1}) \\
&= \mathrm{vec}(u)^T(L_s^{-T} \otimes L_c^{-T})\mathrm{vec}(w) \\
&= [(L_s^{-1} \otimes L_c^{-1})\mathrm{vec}(u)]^T\mathrm{vec}(w) \\
&= \mathrm{vec}(\underbrace{L_s^{-1}uL_c^{-T}}_{w})^T\mathrm{vec}(w) \\
&= \mathrm{vec}(w)^T\mathrm{vec}(w) = \sum_{t=1}^{T}\sum_{b=1}^{K} w_{bt}^2.
\end{aligned}$$

The matrix $w = L_c^{-1}(L_s^{-1}u^T)^T$ is the decorrelated version of u.

For any given pixel with observed data $\tilde{X}$, we can calculate the "evidence" of $\tilde{X}$ coming from class c, which we denote by $e(\tilde{X} \in c)$, as follows:

$$e(\tilde{X} \in c) = \mathbb{P}(D_c(X_v) > D_c(\tilde{X})) = \mathbb{P}(Q > D_c(\tilde{X})),$$

where the probability is over all possible X_v for any pixel v in class c, and $Q \sim \chi^2_{KT}$ from above. Formally, this probability evaluates how extreme $\tilde{X}$ is compared to a typical X_v from class c, and is thus equivalent to a p-value for testing if $\tilde{X}$ belongs to class c. It is also similar in spirit to quadratic discriminant analysis (see e.g., [73], and also above for a formal test.)

For this work we wish to assess if a pixel $\tilde{X}$ is potentially an outlier relative to a specific training site composed of multiple pixels, or rather if pixel $\tilde{X}$ should indeed be a member of a specific training site. Thus, we must also consider the site composition. To this end, a pixel does not belong to a particular training site if its transformed spectral-temporal

profile differs from the site composition (i.e., as opposed to the majority land cover class or even a user-defined site label). Using the evidence of a pixel belonging to class c along with the posterior distribution for π in (2.18), we propose the following formulation for the evidence of $\tilde{X}$ belonging to site s, conditional on all the data $X_1, \ldots, X_n$ in site s, $e(\tilde{X} \in s \,|\, X_1, \ldots, X_n)$:

$$\begin{aligned} e(\tilde{X} \in s | X_1, \ldots, X_n) &= \sum_c \int e(\tilde{X} \in c) P(\tilde{\theta} = c \,|\, \pi, X_1, \ldots, X_n) P(\pi \,|\, X_1, \ldots, X_n) d\pi \\ &= \sum_c e(\tilde{X} \in c) \mathbb{E}[\pi_c \,|\, X_1, \ldots, X_n]. \end{aligned}$$

That is, to estimate the evidence that $\tilde{X}$ belongs to each class we compute a weighted average that accounts for all possible "candidate" classes $\tilde{\theta}$ for $\tilde{X}$ and all possible site compositions π according to the Dirichlet distribution in (2.18). In this way, we estimate a pixel's class composition and then judge whether it belongs to a site if that composition differs from the composition of that site. Note that according to this distribution

$$\begin{aligned} \mathbb{E}[\pi_c \,|\, X_1, \ldots, X_n] &= \frac{\sum_{v=1}^{n} P(\theta_v = c \,|\, X_v; \pi^*) + n_0 h_c}{n + n_0} \\ &= \pi_c^* + \frac{1 - C\pi_c^*}{n + n_0} \end{aligned}$$

can increasingly differ from the mode π^* as the number of pixels in the site becomes smaller. Based on this, we might consider $\tilde{X}$ an outlier if $e(\tilde{X} \in s \,|\, X_1, \ldots, X_n)$ is below a prescribed threshold, which is similar to defining a significance level. This approach accounts for the fact that our ability to detect outliers in a site depends, in part, on both the size and composition of the site.

2.7.3 Site Homogeneity

Finally, we can use the framework described above to investigate the *homogeneity* of any given site. When training sites are created it is generally assumed that every pixel (in sites that contain more than one pixel) belongs to the same class. Here we propose to address

this question more rigorously than by simply examining the predicted class at each pixel within sites or by evaluating whether each pixel belongs to a specific site. Specifically, if we define $\mathfrak{H} > 0.5$ to be a user-prescribed probability threshold, we can define the evidence

$$\begin{aligned} e(\text{Site is homogeneous}) &= P\left(\max_c \pi_c > \mathfrak{H} \,|\, X_1, \ldots, X_n \right) \\ &= \textstyle\sum_c P(\pi_c > \mathfrak{H} \,|\, X_1, \ldots, X_n) = \sum_c P(B_c > \mathfrak{H}), \end{aligned}$$

where, if $\alpha_c^* = n_0 \alpha_c + \sum_{v=1}^{n} P(\theta_v = c \,|\, X_v; \pi^*)$, $B_c \sim \mathsf{Beta}(\alpha_c^*, n_0 + n - \alpha_c^*)$ is the marginal distribution for each π_c. Using this, we can assess whether a site is homogeneous or not without specifying which land cover class the site belongs to.

2.7.4 Training Data Assessment Results

Using the procedures described above, we estimate the composition of each test site, the evidence that each pixel is a member of its corresponding site, and the probability that each test site is homogeneous. Table 2.3 summarizes these results. In these calculations we assumed that $n_0 = \max\{0.05n, C\}$: the prior "counts" are only 5% of the number of pixels in the site, but are no smaller than the number of classes C to avoid discontinuities in (2.19). The homogeneity probabilities vary widely, along with the site membership evidences. Recall that these *evidences* represent the evidence that a pixel belongs to the site it is a member of. Hence, lower values of evidence indicate potential outliers. Additionally, our measure of homogeneity considers all classes instead of just the class we may see dominating the site composition estimate. Results shown in Table 2.3 reflect patterns observed later in the classification accuracy assessment (Section 3.1.1) in Tables 3.1 and 3.2. The estimated composition for most sites reflect the class chosen by the expert, with classes 7, 8, and 9 clearly problematic. Identification of representative and homogeneous land-cover training sites can be quite challenging. Hence, the methods we describe here have significant potential for assessing the quality of manually interpreted training data.

Figure 2.9 show results from an analysis designed to evaluate how well our proposed

Table 2.3: Each row in the table is a site in the testing data set. The estimated site compositions are in the 12 columns between *Size* and P(Hom), corresponding to the land-cover classes. Grayed cells indicate the class labeled by the experts. P(Hom) gives the probability that the site is homogeneous for a threshold of $h = 0.5$. The last two columns are the minimum and maximum site membership evidences, across pixels in the site.

Site	Size	1	4	5	7	8	9	10	11	12	13	14	17	P(Hom)	Min Out	Max Out
1	13	1.00												0.85	0.91	0.98
2	13	0.77		0.23										0.43	0.99	1.00
3	6	1.00												0.32	0.72	0.81
4	14		1.00											0.69	0.69	0.78
5	13		1.00											0.62	0.69	0.78
6	4		0.50	0.50										0.01	0.68	0.72
7	12	0.17		0.83										0.45	0.91	1.00
8	10			1.00										0.62	0.98	1.00
9	10		0.30	0.70										0.16	0.79	0.90
10	12		0.09	0.91	0.00									0.57	0.21	0.28
11	16				0.00			1.00						0.94	0.97	0.97
12	6				0.00				1.00					0.11	0.41	0.44
13	12			1.00		0.00								0.76	0.95	1.00
14	4			1.00		0.00								0.12	0.85	0.91
15	14			1.00		0.00								0.86	0.72	0.86
16	12						0.00	0.42		0.33	0.25			0.02	1.00	1.00
17	2						0.00				1.00			0.01	0.28	1.00
18	10						0.80		0.20					0.13	0.41	0.42
19	12							1.00						0.81	0.98	1.00
20	12							1.00						0.81	0.94	1.00
21	14							1.00						0.89	0.97	1.00
22	14								1.00					0.72	0.68	0.73
23	2	1.00							0.00					0.05	0.63	0.63
24	6						1.00		0.00					0.10	0.06	0.10
25	7									1.00				0.31	0.48	0.62
26	12									1.00				0.73	0.58	0.59
27	51									0.98		0.02		1.00	0.81	0.96
28	12										1.00			0.51	1.00	1.00
29	20										1.00			0.93	1.00	1.00
30	11			0.52							0.48			0.04	1.00	1.00
31	10		0.38									0.62		0.04	0.16	0.42
32	5		1.00									0.00		0.05	0.85	0.88
33	6									0.17		0.83		0.06	0.93	1.00
34	16												1.00	0.89	0.53	0.74
35	39	0.15											0.85	0.99	0.00	0.96
36	15							0.07					0.93	0.75	0.00	0.74

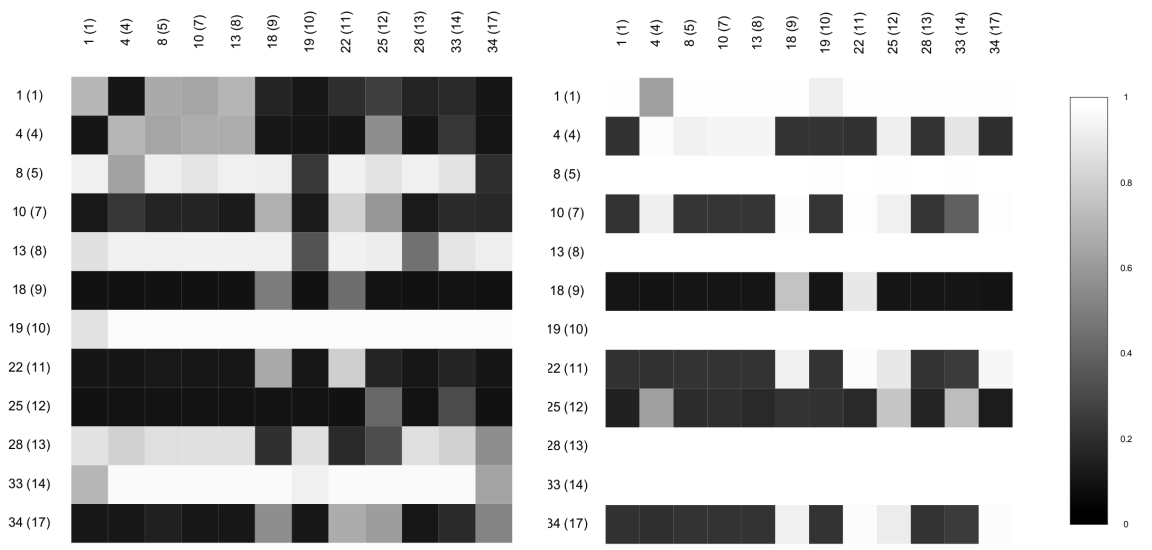


Figure 2.9: Heat maps of site membership evidence. On the left is the minimum membership evidence among pixels within the given site. On the right is the maximum membership evidence among pixels within the given site. The labels are formatted as "Site (Class)" where "Site" is the number from Table 2.3 and "Class" is the class of that particular site.

site membership evidence metric is able to suggest outlier pixels. To do this, we first selected the test site from each class that was most unimodal in its estimated composition. We then compute the membership evidence for the pixels in each of these sites as if they belonged in each of the other sites. The rows in Figure 2.9 represent the source (original) site and the columns represent the target sites. Ideally, the off-diagonal cells would be close to black indicating low *evidence* that the row site pixels belong to the column site. Sites 8, 13, 19, 28 and 33 display noticeably higher *minimum membership evidence*. Despite the unimodal compositions of these sites, the *evidence* captures more of the distribution of the composition and better reflects the ambiguity in the class definitions. Specifically, recall that the estimated site composition represents the posterior mode (π^*), whereas the membership evidence involves the posterior mean, which differs from the mode. This variability in the posterior distribution of π_c, in addition to ambiguous definitions for classes such as Mixed Forests, Woody Savannas, Grasslands, Urban, and Croplands, leads to difficulties distinguishing the correct label for pixels belonging to these classes. Not

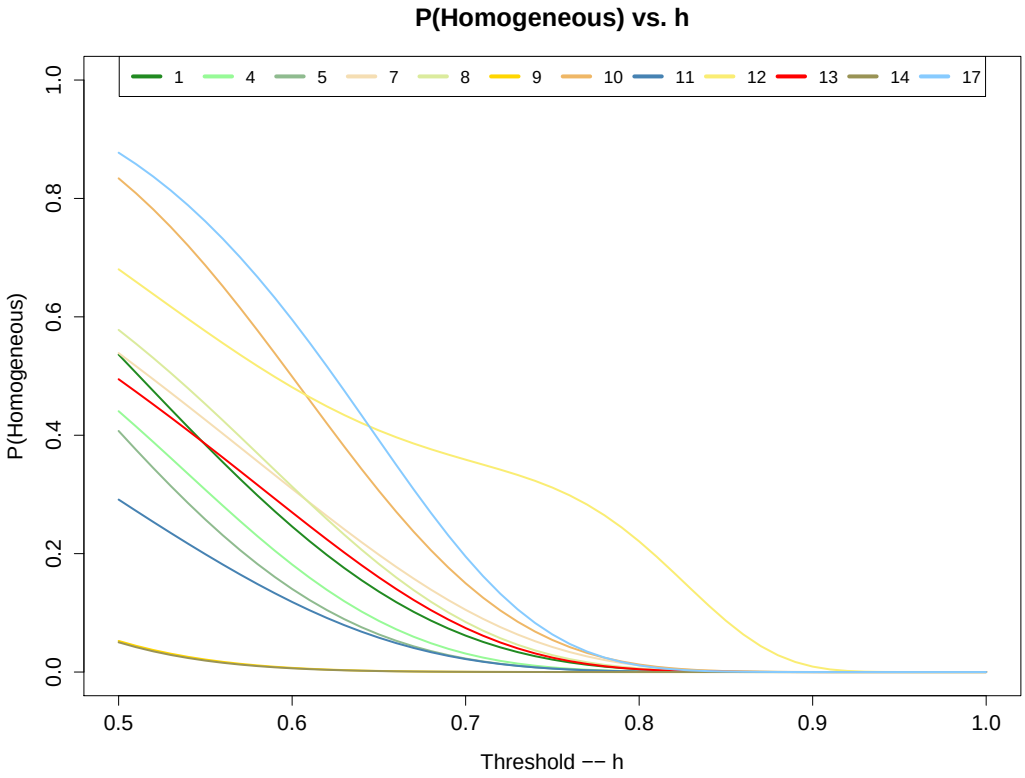


Figure 2.10: Probability of homogeneity for each class as a function of the homogeneity threshold $\mathfrak{H}$. Each curve is the average over three testing sites for each class.

surprisingly in Figure 2.9, the membership evidence is low for water pixels placed in other sites. In either case, both heat maps align with our expectations and further support the patterns reported in the results of Section 3.1.1: the weaknesses of the model are mostly restricted to particular classes, but otherwise the model describes the data well.

Finally, Figure 2.10 investigates how the proportion of a site belonging to the same class affects our estimated probability of a site being homogeneous. Recall that $\mathfrak{H}$ describes the proportion of the site that needs to be the same in order for us to call it homogeneous. Figure 2.10 summarizes, for each class, the average probability that the test sites are homogeneous. Naturally, the probability of a site being homogeneous decreases as we require more and more of the site to be same, since this demands more from the data. While this behavior is consistent across all of the classes, certain classes are clearly more

homogeneous than others. As expected, water sites exhibit some of the highest probabilities of being homogeneous and classes such as Wetlands and Natural Vegetation Mosaics (which are mixture classes by definition) display some of the lowest.

Chapter 3

Land Cover Classification

We arrive at the original problem of land cover classification after estimating model parameters, imputing missing data, and using principal components analysis to reduce dimensionality. Because we have *a priori* knowledge about the distribution of land cover classes, we pursue a Bayesian classification framework in two ways:

$$\mathbb{P}(\theta_v \,|\, X_v) \propto \mathbb{P}(X_v \,|\, \theta_v)\mathbb{P}(\theta_v) \tag{3.1}$$

and

$$\mathbb{P}(\boldsymbol{\theta} \,|\, \mathbf{X}) \propto \mathbb{P}(\mathbf{X} \,|\, \boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}). \tag{3.2}$$

where $\mathbf{X}$ and $\boldsymbol{\theta}$ denote the data and labels for all of the pixels in the image, respectively. In both approaches the pixel data are conditionally independent given the land cover labels. In (3.1), classification occurs pixel-by-pixel without using information from neighboring pixels. In (3.2), we conduct posterior inference for all of the pixels in the lattice with the help of data in adjacent pixels and prior knowledge about the spatial distribution of land cover classes.

In Section 3.1 we apply an independent pixel classification, via (3.1), in a spirit similar to most traditional classification techniques. However, we originally sought the incorporation of extensive prior information and spatial relationships into our classification procedure. Popular methods for the modeling of this type of spatial data include kriging and other uses of Gaussian processes. Our core objection to these methods consists of the use of a spatially varying distribution for the data. We believe that the land cover classes themselves

completely capture the variability in the data. That is, by modeling the spectral-temporal variance for each land cover class and describing the spatial relationships between the classes through a prior distribution we successfully include all of the meaningful spatial information. The remainder of the chapter, after Section 3.1, introduces a hierarchical graphical model to achieve this goal in both an interpretable and computationally efficient way.

3.1 Independent Pixel Classification

The model in (2.12) specifies the distribution, in principal components space, of the annual profile for any pixel, v. Based on this framework, and using the fitted parameters estimated by EM, it is straightforward to naïvely classify the land cover type for each pixel in the data set, independent of the other pixels. Specifically, if pixel v belongs to each class c with *a priori* probabilities h_c, then the posterior probability that pixel v belongs to class c is given by Bayes rule:

$$P(\theta_v = c \,|\, X_v; \mathbf{h}) = \frac{h_c P(X_v \,|\, \theta_v = c)}{\sum_{\tilde{c}} h_{\tilde{c}} P(X_v \,|\, \theta_v = \tilde{c})} \tag{3.3}$$

where the likelihood $P(X_v \,|\, \theta_v = c)$ is given by (2.12) and the sum in the denominator is over all land cover classes.

3.1.1 Case Study

To test our proposed model, we use data from the MODIS Land Cover Training site database, described in Section 2.6. Table 2.1 lists the number of pixels in each IGBP class in the data set, along with the prior probability for each class. Classes 15 and 16 were excluded because they are rare in the study area. The final data set included 12 land-cover classes across 2,692 pixels.

We employ two different types of assessment procedures to investigate the results from the methods we propose. In the first case we use a *n-fold cross-validation* procedure

in which two sites are randomly sampled from each class and withheld as testing data. The parameters of (2.12) were then estimated from the remainder of the data. Repeated 100 times, this process yields a more robust assessment of the classification accuracy and precision broken down by land cover class.

The second procedure uses a *single* testing dataset comprised of 36 sites (3 from each class) and trains the parameters in (2.12) from the remainder of the data. Sampling three sites from each class yielded a testing data set of 447 pixels and a training data set of 2,245 pixels (the same as in Section 2.7).

To assess the accuracy of our classification method, we used the estimated principal components to assign the class with the highest posterior probability based on (3.3) in Section 3.1 to each pixel. Below we present results from both the *cross-validation procedure* and the *single testing dataset procedure*. We compare the single testing dataset results to classification of the imputed PC data via C4.5 decision trees as well as Random Forests.

Cross-Validation Procedure

The cross-validation procedure we used withholds two sites from each class, estimates the parameters of the model from the remaining data, and then classifies the test pixels in the withheld sites (100 times). At each iteration we compute class accuracies and precisions, which yields a set of 100 accuracies and 100 precisions for each class, as well as 100 overall accuracies. Figure 3.1 summarizes class-specific and overall classification results from this analysis. Overall classification performance is reasonably good (67%), but significant variability across classes exists. The second type of assessment sheds more light on this variability across classes.

Single Testing Dataset Procedure

In this second procedure, classification proceeds in exactly the same way but parameter estimates were computed using a single training data set composed of 2,245 pixels. Table 3.1 presents a confusion matrix of the results for the training data. Overall, the

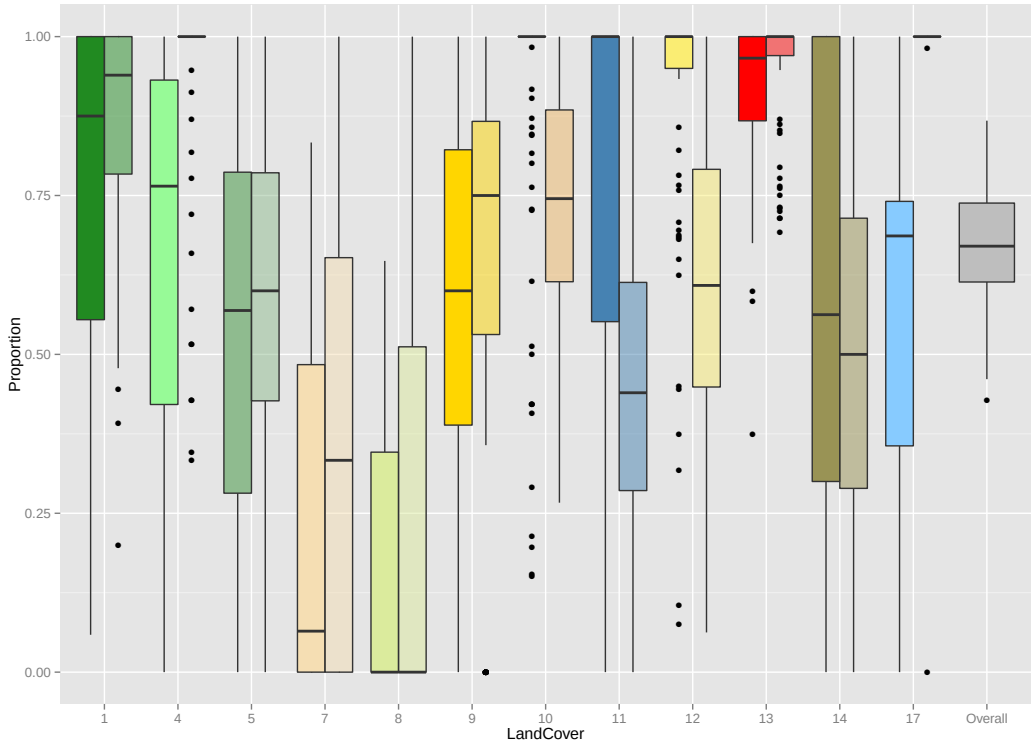


Figure 3.1: Classification accuracies and precisions broken down by class. The first (left) boxplot for each class describes accuracy and the second (right) describes precision. The mean overall accuracy was 67.2%. The dots above represent potential outlier accuracy values.

classifier performs well (78% correctly classified). However, class-specific accuracies were quite variable, and some classes (e.g., class 8: woody savanna) were clearly problematic. Figure 3.2 displays a comparison of the accuracy of the classifier when using the principal components versus the untransformed data and clearly illustrates that with the exception of classes 7, 8, 9, and 11 accuracies improved using our approach. Note, however, that these three classes had many fewer data values overall relative to the other classes. Thus, under sampling of these classes probably explains part of this result.

The test data set for this stage of the assessment consisted of 447 MODIS pixels from 36 sites (3 from each class), where all of the pixels in each site were assumed to have the same class. Table 3.2 presents a confusion matrix for model predictions for each pixel in the test set, and shows that results on the test data are consistent with results from the training

Table 3.1: Confusion matrix for classification of training pixels using the first three principal components. The rows represent our classification; the columns represent the true classification.

Class	1	4	5	7	8	9	10	11	12	13	14	17	Total	Precision
1	158	0	18	0	19	1	0	2	0	0	0	0	198	0.80
4	0	95	29	0	10	0	0	0	15	0	40	0	189	0.50
5	15	5	99	0	2	9	0	16	0	0	0	0	146	0.68
7	0	0	0	25	0	0	0	9	0	0	0	0	34	0.74
8	5	0	6	0	0	5	0	7	0	0	0	0	23	0.00
9	0	0	0	7	0	29	0	16	0	0	0	0	52	0.56
10	0	0	0	10	0	0	194	0	42	0	0	0	246	0.79
11	0	0	0	2	0	10	0	103	0	0	0	0	115	0.90
12	0	0	0	17	0	0	69	0	740	0	16	0	842	0.88
13	1	0	0	0	4	15	26	2	7	164	0	0	219	0.75
14	0	0	0	0	1	0	0	0	44	0	47	0	92	0.51
17	0	0	0	0	0	0	0	0	0	0	0	89	89	1.00
Total	179	100	152	61	36	69	289	155	848	164	103	89	2245	
Accuracy	0.88	0.95	0.65	0.41	0.00	0.42	0.67	0.66	0.87	1.00	0.46	1.00		**0.78**

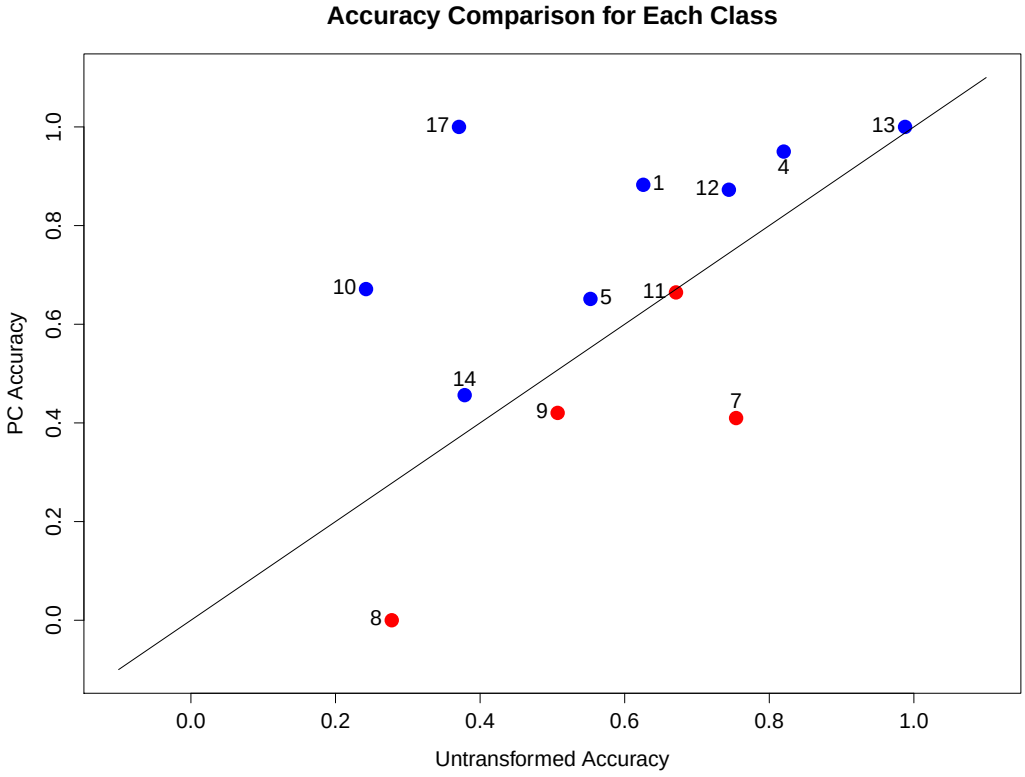


Figure 3.2: Scatterplot of class accuracies. Each point corresponds to a class accuracy using untransformed data (horizontal axis) and a class accuracy using the PCs (vertical axis). Most points are above the 1:1 line (blue), indicating increased classification accuracies when using the PCs.

data. Classification accuracies, both overall and for specific classes, were similar (although slightly lower) to those for the training data, suggesting that the overall approach is robust.

Table 3.2: Confusion matrix for classification of testing pixels using the first three principal components. The rows represent our classification; the columns represent the true classification.

Class	1	4	5	7	8	9	10	11	12	13	14	17	Total	Precision
1	29	0	2	0	0	0	0	2	0	0	0	6	39	0.74
4	0	29	3	1	0	0	0	0	0	0	9	0	42	0.69
5	3	2	27	11	30	0	0	0	0	6	0	0	79	0.34
7	0	0	0	0	0	0	0	0	0	0	0	0	0	-
8	0	0	0	0	0	0	0	0	0	0	0	0	0	-
9	0	0	0	0	0	8	0	6	0	0	0	0	14	0.57
10	0	0	0	16	0	5	38	0	0	0	0	0	59	0.64
11	0	0	0	6	0	2	0	14	0	0	0	1	23	0.61
12	0	0	0	0	0	4	0	0	69	0	1	0	74	0.93
13	0	0	0	0	0	5	0	0	0	37	0	0	42	0.88
14	0	0	0	0	0	0	0	0	1	0	11	0	12	0.92
17	0	0	0	0	0	0	0	0	0	0	0	63	63	1.00
Total	32	31	32	34	30	24	38	22	70	43	21	70	447	
Accuracy	0.91	0.94	0.84	0.00	0.00	0.33	1.00	0.64	0.99	0.86	0.52	0.90	-	**0.73**

Tables 3.3 and 3.4 give the classification results for the test set from C4.5 Decision Trees and Random Forests, respectively. These methods appear a bit more accurate than our independent pixel classification. The mistakes appear common across all three methods though. That is, it appears that some of the same classes and pixels were problematic for all three algorithms. Despite maintaining higher accuracy here, recall that two primary concerns with methods like C4.5 and Random Forests were the ways in which they handle missing data as well as their failure to incorporate spatial information.

The remainder of this section gives some mapped results, comments on the issue of training data selection, and re-emphasizes our ultimate goal of incorporating spatial information via a hierarchical graphical model. It should be noted, however, that C4.5 or Random Forests could just as easily be used for classification after our proposed missing data imputation and data compression. The series of methods we develop here does not need to be applied in its entirety or not at all.

Because the region surrounding Montreal, Canada contains a good mixture of land cover classes and is well known to our collaborators, it serves as a useful testing area for our methods. A small, 500 pixel $\times$ 500 pixel area centered on Montreal provides an informative glimpse at the utility and efficiency of our proposed analysis pipeline. So, after estimating the parameters of (2.12) with the training data from the northeast portion of North America and imputing the data for these 250,000 pixels around Montreal we generate

Table 3.3: Confusion matrix for classification of testing pixels using C4.5 decision trees trained on imputed data. The rows represent our classification; the columns represent the true classification.

Class	1	4	5	7	8	9	10	11	12	13	14	17	Total	Precision
1	32	0	0	0	0	0	0	0	0	0	0	9	41	0.78
4	0	27	0	0	0	0	0	0	0	0	0	0	27	1
5	0	4	32	0	15	0	0	0	0	0	0	0	51	0.63
7	0	0	0	1	0	0	0	0	0	0	0	0	1	1
8	0	0	0	0	0	0	0	0	0	0	0	0	0	-
9	0	0	0	0	1	10	0	0	0	0	0	0	11	0.91
10	0	0	0	13	0	4	38	0	0	0	0	0	55	0.69
11	0	0	0	17	13	0	0	22	0	0	0	0	52	0.42
12	0	0	0	3	1	8	0	0	70	0	1	0	83	0.84
13	0	0	0	0	0	2	0	0	0	43	0	0	45	0.96
14	0	0	0	0	0	0	0	0	0	0	20	0	20	1
17	0	0	0	0	0	0	0	0	0	0	0	61	61	1
Total	32	31	32	34	30	24	38	22	70	43	21	70	447	-
Accuracy	1	0.87	1	0.03	0	0.42	1	1	1	1	0.95	0.87	-	**0.80**

Table 3.4: Confusion matrix for classification of testing pixels usingRandom Forests trained on imputed data. The rows represent our classification; the columns represent the true classification.

Class	1	4	5	7	8	9	10	11	12	13	14	17	Total	Precision
1	32	0	1	0	0	0	0	0	0	0	0	0	33	0.97
4	0	27	0	0	0	0	0	0	0	0	0	0	27	1
5	0	4	31	0	19	0	0	0	0	0	0	0	54	0.57
7	0	0	0	0	0	0	0	0	0	0	0	0	0	-
8	0	0	0	0	0	0	0	0	0	0	0	0	0	-
9	0	0	0	0	0	10	0	0	0	0	0	0	10	1
10	0	0	0	16	0	10	38	0	0	0	0	0	64	0.59
11	0	0	0	18	10	0	0	22	0	0	0	0	50	0.44
12	0	0	0	0	0	2	0	0	70	0	3	0	75	0.93
13	0	0	0	0	1	2	0	0	0	43	0	0	46	0.93
14	0	0	0	0	0	0	0	0	0	0	18	0	18	1
17	0	0	0	0	0	0	0	0	0	0	0	70	70	1
Total	32	31	32	34	30	24	38	22	70	43	21	70	447	-
Accuracy	1	0.87	0.97	0	0	0.42	1	1	1	1	0.86	1	-	**0.81**

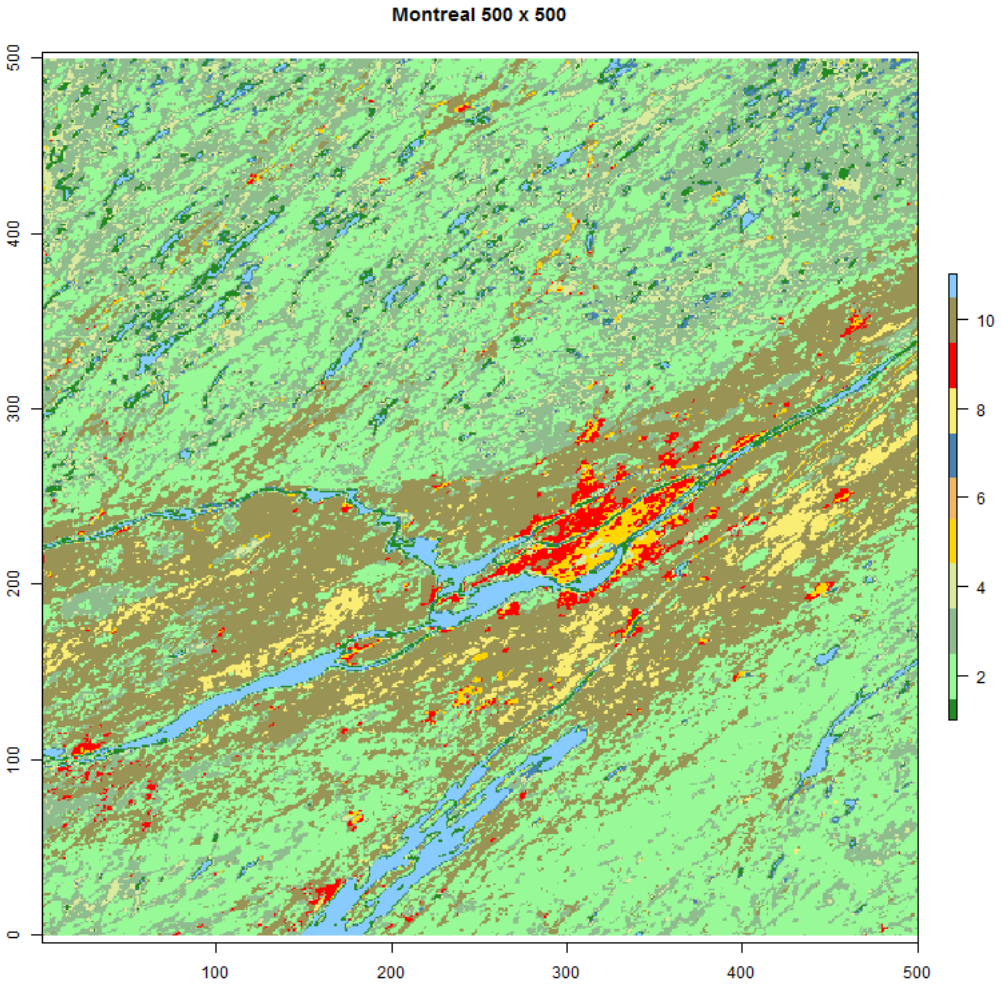


Figure 3.3: Independent pixel classification of region surrounding Montreal, Canada.

a land cover map of the region in Figure 3.3 (use the legend in Figure 1.2 to interpret the color scheme).

To explore the usefulness of our approach in a starkly different biome, we also applied our pipeline to a 500 pixel $\times$ 500 pixel region centered on Bakersfield, California. Figure 3.4 gives our result. Oceans are masked out of the original data set so the results for these pixels (left side of the image) should be ignored. We estimated the parameters to use for this region from the training sites in the seven MODIS tiles (tile map in Figure 1.1) surrounding

Bakersfield. This seems unfair and reasonable at the same time. The training site database contains data from across the globe. However, land cover classes look different in different climates or on different continents. For these reasons, we began our foray into land cover classification using training data we think most appropriate for the regions we classify. In Section 3.1.2 we further explore the issue of training data selection and, near the end of this section, give another example of it. Figure 3.5 gives a result for Bakersfield excluding the agricultural and urban classes. Even in the MODIS tiles surrounding Bakersfield, the agricultural and urban classes contain a significant amount of diversity and consequently take over the other classes to some degree, and so we exclude them here. A full independent pixel classification of North America can be seen in Figure B.1 in Appendix B.

Beyond these assessment tools, the importance of relevant and good quality training data cannot be overstated. The original training data used above (for the *single testing dataset* procedure) from the northeast portion of North America was delivered to us by our collaborators. We independently selected a subset of the training site database from within the northeast portion of North America which contained 1,677 pixels (around 1,000 less than the previous extract of data). We ran through the analysis pipeline (parameter estimation, missing data imputation, data transformation, and independent pixel classification) to reproduce the map of Montreal and gauge the influence of the training data.

The map seen in Figure 3.6 does not differ much from our original map in Figure 3.3. However, if we use the training site data from **all** of North America we see something completely different in Figure 3.7. The result in Figure 3.7 not only differs greatly from our previous result, but from our collaborators' previous results and what they expect to see. We need to be careful about where our training data come from relative to where we perform our classification, or consider involving information about the climate or biome a given pixel lives in.

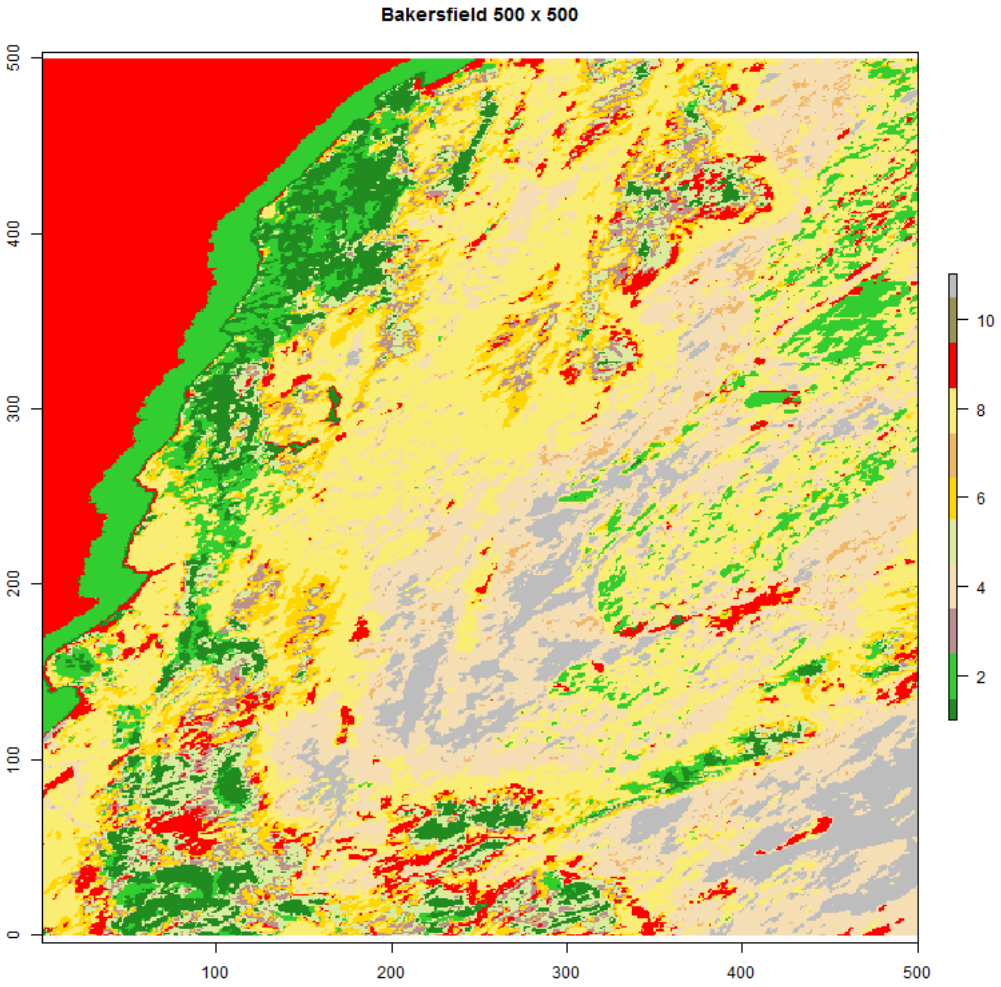


Figure 3.4: Independent pixel classification of region surrounding Bakersfield, California.

3.1.2 Clustering Training Data Using Climate Metrics

The land cover classification results in Figures 3.6 and 3.7 display the significant within-class variation present in our training site database. To address this issue we propose using climate features, which we have access to at the same spatial resolution as our original data, to identify the multiple modes or sub-classes of our established IGBP land cover classes. The climate data consist of eight pixel-wise annual metrics: mean temperature, standard deviation of monthly temperatures, minimum temperature, maximum temper-

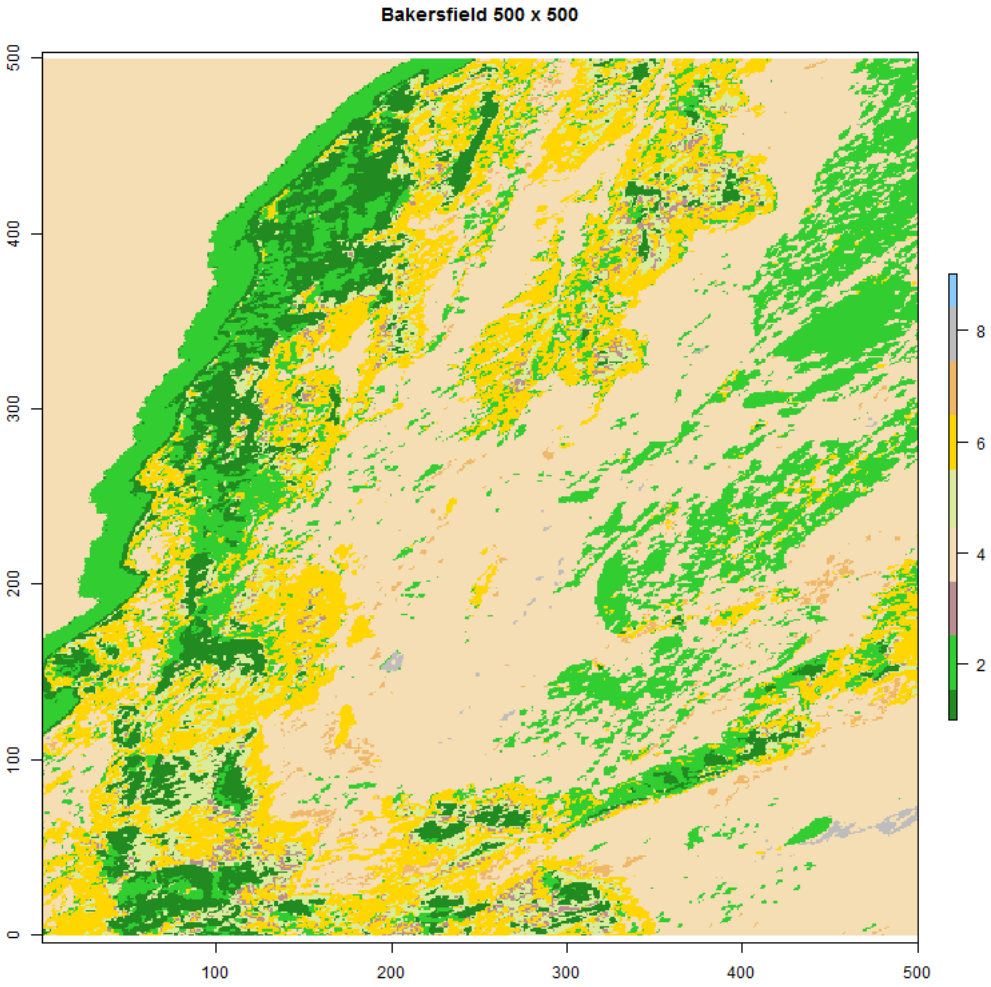


Figure 3.5: Independent pixel classification of region surrounding Bakersfield, California.

ature, annual precipitation, minimum precipitation, maximimum precipitation, and the coefficient of variation for precipitation. Using these climate metrics we cluster our training data into sub-classes that correspond to the same land cover classes broken across different biomes. We use hierarchical clustering with our own distance metric to determine the cluster assignments.

Let us denote by x_o our *original* spectral-temporal data that have been decorrelated temporally and averaged across time, by x_{clim} the data on the eight climate features spec-

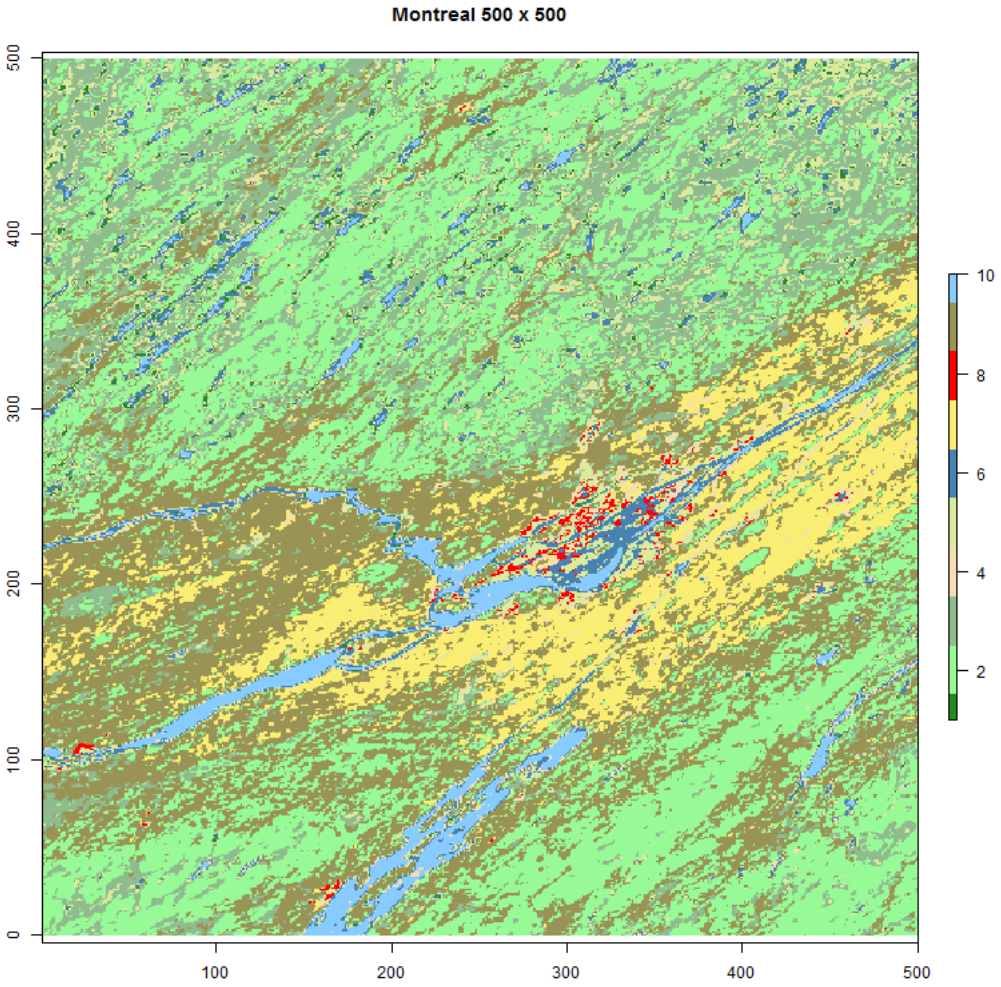


Figure 3.6: Land cover map of Montreal, Canada using different training data in the northeast portion of North America.

ified above, and by $\Sigma_{c,clim}$ the class specific covariance of the climate features. We will use the following distance metric, with tuning parameter λ, to perform the clustering.

$$d(x,y) \doteq (x_o - y_o)^\top \Sigma_s^{-1}(x_o - y_o) + \lambda(x_{clim} - y_{clim})^\top \Sigma_{c,clim}^{-1}(x_{clim} - y_{clim}) \tag{3.4}$$

Since we are clustering training data, the land cover class of pixel x and pixel y is known to be c. We can vary the importance of the climate features in computing our distances

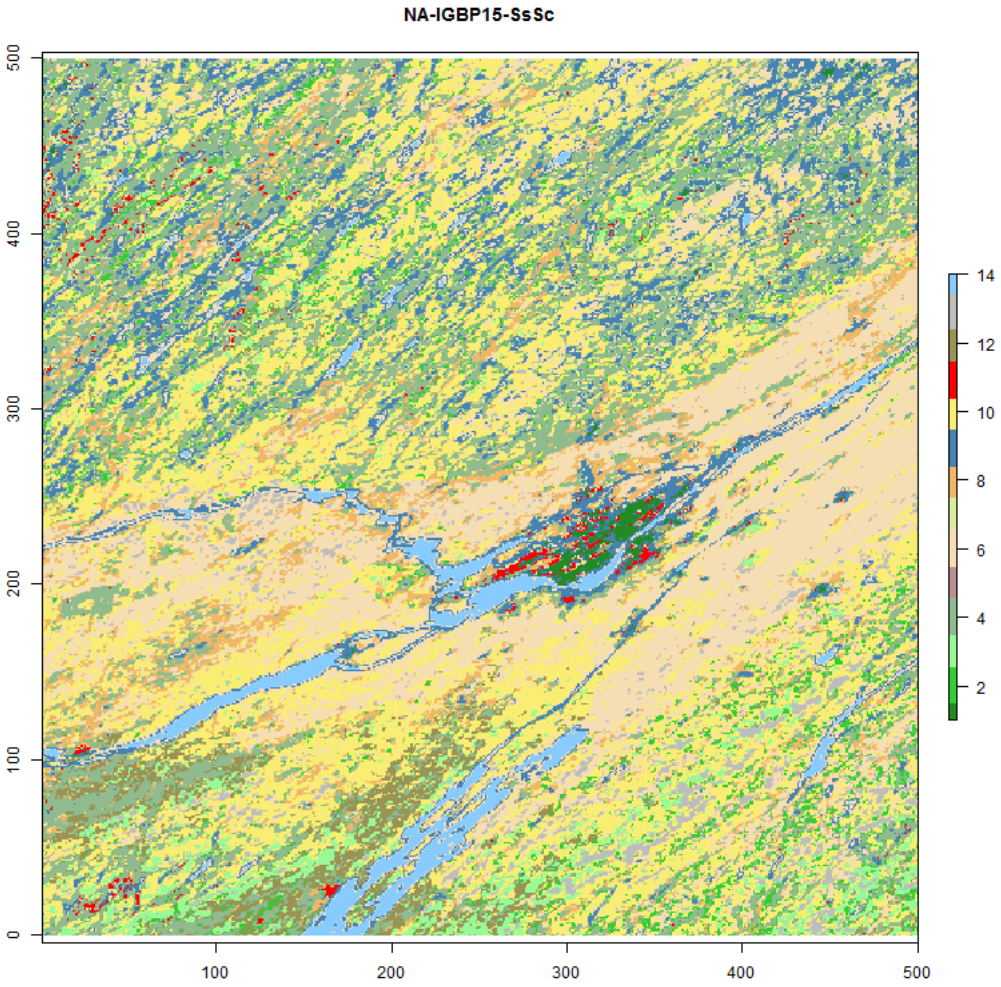


Figure 3.7: Land cover map of Montreal, Canada using training data from all of North America.

by specifying different values of λ. We cluster data within each land cover class, separate from the other classes. We use the Calinski-Harabasz (CH) index in (3.5) [11] to formally decide on an ideal number of clusters for each land cover class. If $B = \sum_{k=1}^{K} n_k(\mu_k - \mu)^2$ and $W = \sum_{k=1}^{K} \sum_{i=1}^{n_k}(x_{i,k} - \mu_k)^2$ are the between-cluster dispersion and within-cluster dispersion, respectively, then the CH index is given by

$$\mathcal{C} = \frac{B/(K-1)}{W/(N-K)}. \tag{3.5}$$

The clusters decided on would become sub-classes and this set of sub-classes would become our new classification scheme. The CH index cannot be used alone to determine the number of clusters because the parameters of (2.12) must be estimated for each cluster. Therefore, each cluster must contain a minimum number of training pixels. The number of clusters to consider ($K = 2, \ldots, 7$) resulted from this consideration as well. Because a high CH index represents better clustering, we considered each number of clusters in turn from the highest CH index to the lowest. The value of K that yielded the highest CH index *and* a partitioning that would still allow for the estimation of the model parameters dictated the number of clusters for each land cover class.

Motivated in part by the use of all of the North America training data to classify Montreal, the same 7,139 pixels from 15 IGBP classes in North America were clustered. The conditions above could not be satisified for some classes and so those classes were not broken down into sub-classes. Ultimately, this procedure resulted in a new set of 24 land cover sub-classes. Appendix A.1 contains a principal components plot for these data in the original 15 colors (i.e. sub-classes are colored by their parent class). Figures 3.8 and 3.9 give independent classification results using these 24 new sub-classes and colored by the original 15.

The map of Montreal in 3.8 vastly surpasses the map in 3.7. However, weaknesses not present in 3.3 still reside in 3.8. For example, wetlands pervade too much of the region and three forest classes exist where there previously existed only two. In Figure 3.9, the urban class seems to be dominating again. Full independent pixel classification results for North America using these clustered sub-classes can be seen in Figures B.3 and B.4 in Appendix B, and do represent an improvement in our attempts to classify pixels independently. However, mapping global land cover will likely require further exploration and clustering as we move to using global training data.

The remainder of this chapter details the construction of a hierarchical graphical model that incorporates spatial information into a Potts prior and allows for exact posterior inference of the pixel labels.

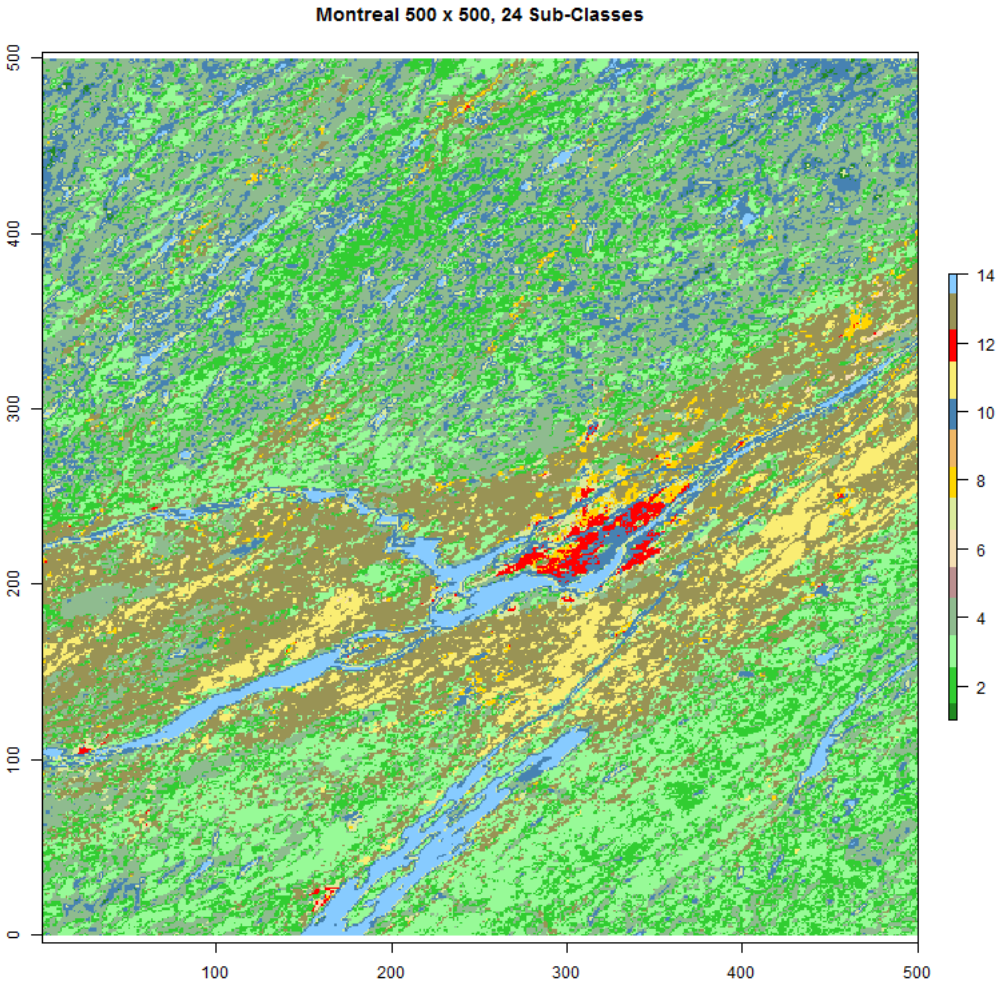


Figure 3.8: Land cover map of Montreal, Canada using training data from all of North America after clustering data into 24 sub-classes using climate metrics.

3.2 Potts Prior Model and Hyperparameter Elicitation

Ultimately we wish to incorporate spatial information in our classification method as opposed to assigning labels to pixels independent of surrounding data. To this end we specify a hierarchical graphical model that utilizes the aforementioned likelihood, in (2.12), but now defines a prior on the lattice of pixels in the image of interest.

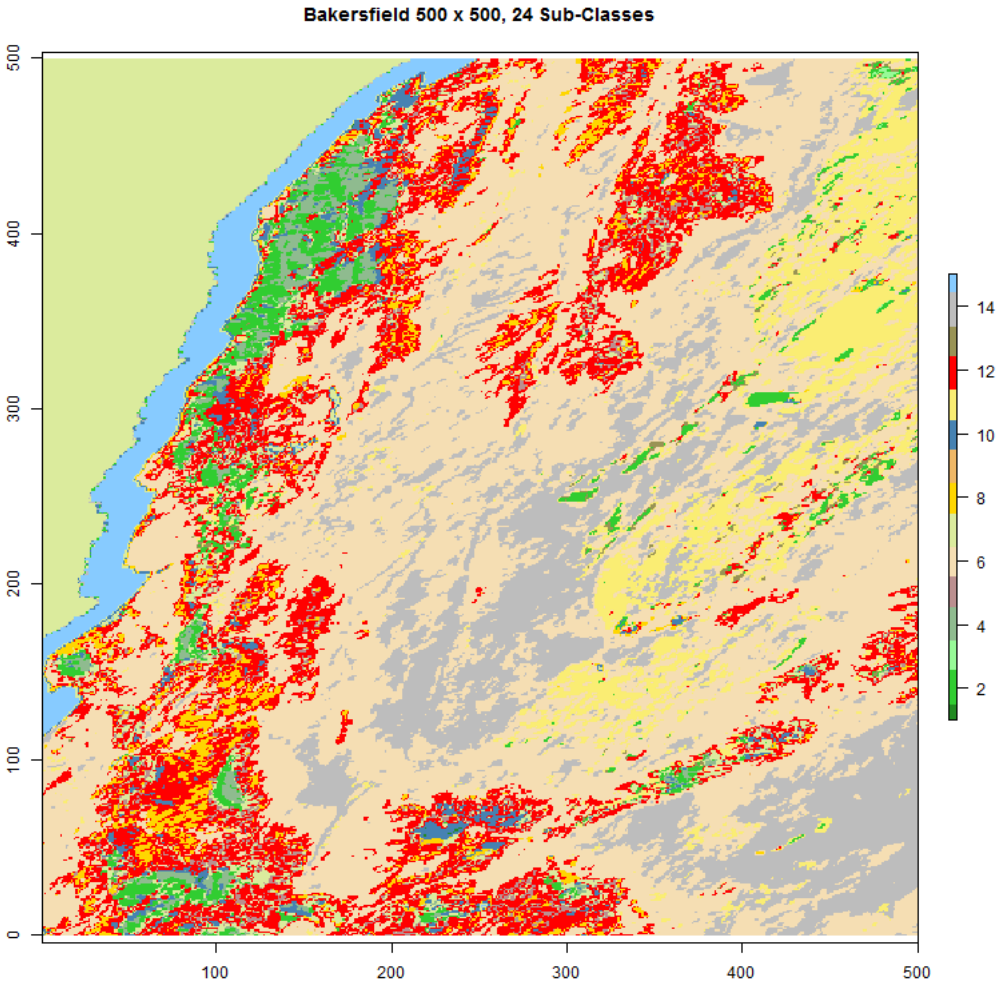


Figure 3.9: Land cover map of Bakersfield, California using training data from all of North America after clustering data into 24 sub-classes using climate metrics.

3.2.1 Prior Model

We denote a complete set of labels for the pixels in the image by $\boldsymbol{\theta}$. Whereas in Section 3.1 our prior on the land cover classes for the entire image amounted to $\mathbb{P}(\theta) = \prod_v \mathbb{P}(\theta_v) = \exp\{\sum_v \mathbf{h}^\top \mathbf{e}(\theta_v)\}$, we wish to extend this in a way that incorporates spatial information. A Markov Random Field accommodates this interest and so we employ a Potts model [86]

prior on $\boldsymbol{\theta}$,

$$\mathbb{P}(\boldsymbol{\theta}) \propto \exp\left\{ \beta \sum_{v \in \mathcal{P}} \mathbf{h}^\top \mathbf{e}(\theta_v) + \eta \sum_{(u,v) \in L} \mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v) \right\} \tag{3.6}$$

where the image, L, consists of a two-dimensional lattice on the grid of pixels. In (3.6), $\mathbf{h}$ characterizes the same global distribution of labels we saw in Section 3.1 and $\mathbf{J}$ describes the relationship between neighboring pixels. More specifically, $\mathbf{J}_{r,s}$ corresponds to an empirical estimate of the joint probability of observing labels r and s in adjacent pixels. In an Empirical Bayes approach we obtain the values of $\mathbf{h}$ and $\mathbf{J}$ from training data or previous land cover products similar to, but independent of, the image of interest.

The hyperparameters β and η, specified by the user, control the strength of the corresponding pieces of this prior distribution.

3.3 Posterior Inference

Traditional posterior inference via the maximum *a posteriori* (MAP) estimator necessitates the computation of the label configuration which maximizes

$$\mathbb{P}(\boldsymbol{\theta} \,|\, X) = \exp\left\{ \sum_{v} l(X_v \,|\, \theta_v) + \beta \sum_{v \in P} \mathbf{h}^\top \mathbf{e}(\theta_v) + \eta \sum_{(u,v) \in L} \mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v) \right\} / Z(L). \tag{3.7}$$

As we will see in Section 3.3.1, our estimator of choice necessitates the computation of the posterior marginal probabilities which means we need to compute

$$Z(L) \doteq \sum_{\theta \in S^n} \exp\left\{ \sum_{v \in P} l(X_v \,|\, \theta_v) + \beta \sum_{v \in P} \mathbf{h}^\top \mathbf{e}(\theta_v) + \eta \sum_{(u,v) \in L} \mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v) \right\}. \tag{3.8}$$

Because of the *high connectivity* of the two-dimensional lattice, the computation of (3.8) is intractable. A popular method in this framework, ICM maximizes a joint probability and thus does not encounter this difficulty the lattice presents. However, the value reached

by ICM does not necessarily correspond to even the optimal, MAP estimate using (3.7). The *centroid* estimator represents the posterior space in a superior way and utilizes the information built into our hierarchical model.

3.3.1 The Centroid Estimator

Perhaps the most popular estimation technique in this Bayesian framework, maximum *a posteriori* estimation assigns to the set of pixels the following:

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\tilde{\boldsymbol{\theta}} \in \Theta_\theta} \mathbb{P}(\tilde{\boldsymbol{\theta}} \,|\, X). \tag{3.9}$$

In a high-dimensional situation such as this, however, the MAP estimator can myopically find a solution that does not represent the posterior space well. That is, the zero-one loss function used by the MAP estimator does not differentiate between solutions that are different by one pixel label or different by all but one of the pixel labels. A more suitable estimator would measure these finer differences in an element-wise way. Therefore we assign a label to each pixel in the image with the *centroid* estimator [12] using the Hamming loss [45], $H(\cdot, \cdot)$,

$$\widehat{\boldsymbol{\theta}}_C = \arg\min_{\tilde{\theta} \in \Theta} \mathbb{E}_{\theta|X}\Big[H(\tilde{\theta}, \theta)\Big] = \arg\max_{\tilde{\theta} \in \Theta} \sum_v P(\theta_v = \tilde{\theta}_v | X). \tag{3.10}$$

Use of this loss function means assigning labels not with the full posterior joint, but with the posterior marginal distribution at each pixel. Our centroid estimator assigns the pixel label which maximizes the posterior marginal distribution. While this better represents the posterior space, the computation of the centroid proves problematic due to the connectivity of the lattice. For this reason we pursue an approximation to the lattice in Section 3.4 which greatly simplifies the computation. In particular, message-passing algorithms [70] will allow for quick calculation of the posterior marginal distribution at each pixel. We can expand our inference in Section 3.3.2, with the use of a gain matrix, to take advantage

of knowledge we possess about the frequency with which certain land cover classes get confused.

3.3.2 Centroid Estimator with a Gain Matrix

The risk in (3.10) can be written, in general, as

$$\mathbb{E}_{\theta \mid X}[L(\theta, \tilde{\theta})] = \mathbb{E}_{\theta \mid X}[\sum_{v} \mathbf{e}(\tilde{\theta}_v)^\top \mathbf{L}\mathbf{e}(\theta_v)] \tag{3.11}$$

where the matrix $\mathbf{L}$ encodes our choice of loss function. In (3.10), the Hamming loss corresponds to $\mathbf{L} = \mathbf{1}_C\mathbf{1}_C^\top - \mathbf{I}_C$ where $\mathbf{I}_C$ indicates the $C \times C$ identity matrix. So, minimizing our risk to obtain our centroid estimator involves the following:

$$\begin{aligned} \arg\min_{\tilde{\theta}\in\Theta} \mathbb{E}_{\theta \mid X}[L(\theta, \tilde{\theta})] &= \arg\min_{\tilde{\theta}\in\Theta} \textstyle\sum_v \mathbb{E}_{\theta \mid X}[\sum_v \mathbf{e}(\tilde{\theta}_v)^\top \mathbf{L}\mathbf{e}(\theta_v)] \\ &= \arg\max_{\tilde{\theta}\in\Theta} \textstyle\sum_v \mathbb{E}_{\theta \mid X}[\sum_v \mathbf{e}(\tilde{\theta}_v)^\top (-\mathbf{L})\mathbf{e}(\theta_v)] \end{aligned} \tag{3.12}$$

Letting $\mathbf{G} = -\mathbf{L}$ we formulate the centroid estimator in terms of a *gain matrix* which allows us finer control over the final inference. We then write the centroid estimator in this new way:

$$\widehat{\boldsymbol{\theta}}_C = \arg\max_{\tilde{\theta}\in\Theta} \sum_{v} \mathbf{e}(\tilde{\theta}_v)^\top \mathbf{G}\mathbf{p}_v. \tag{3.13}$$

In (3.13), the entries of the vector $\mathbf{p}_v$ represent the marginal posterior probabilities at pixel v. We can think of $\mathbf{G}$ as a re-weighting of the posterior probabilities since we can normalize $\mathbf{G}\mathbf{p}_v$ and still preserve the arg max. With this re-weighting we aim to address ambiguities in the land cover class definitions (Table 1.2) and spectral similarities between some of the classes.

Construction of the gain matrix can happen in a multitude of ways, but we will only detail two here. The first involves conceptualizing $\mathbf{G}$ as a type of similarity matrix and specifying the $ij-$entry to reflect, say, the probability that a pixel of land cover class i gets classified as land cover class j. Subject matter experts could build $\mathbf{G}$ manually or base it

on the confusion matrix of a previously approved classification result.

The second method still involves $\mathbf{G}$ being a type of similarity matrix, but now based almost entirely on the spectral data in our training set. Recall the PC plot in Figure 2.6 and the fact that we retained the first *three* principal components. We could construct $\mathbf{G}$ using the volumes found in this 3-dimensional principal components space (see Figure 3.11). Imagine taking the distances of every observation to its class' mean. To prevent extreme observations from influencing these volumes too much we use only the points whose distance to their mean is under the $90^{\text{th}}-$percentile of distances in that class. Evidence for this choice of percentile appears as a reasonable *kink* in the curves in Figure 3.10. Figure 3.12 confirms the ability of this approach to exclude extreme observations. With 90% convex hulls established in 3-dimensional PC space, we determine the $ij-$entry of $\mathbf{G}$ by taking the volume of the intersection of the convex hull for land cover class i and land cover class j. In this way we not only consider the spectral similarity of the classes, but also the representativeness of each class in our training dataset.

Again, we encode with $\mathbf{G}$ the similarities of classes or in another sense the specific severity of different classification errors. For example, while the diagonal of $\mathbf{G}$ should contain the highest entries of each row, spectrally similar classes may have entries very close in magnitude. We want to allow mistaking one type of forest for another more often than we want to allow mistaking a forest for water, and so forth. We present two examples of gain matrices in Tables 3.5 and 3.6. Table 3.5 gives a gain matrix constructed using all of the northeast, North America training data. Table 3.6 gives a gain matrix constructed as the average of gain matrices computed for each of 500 bootstrap samples from the original northeast, North America training data.

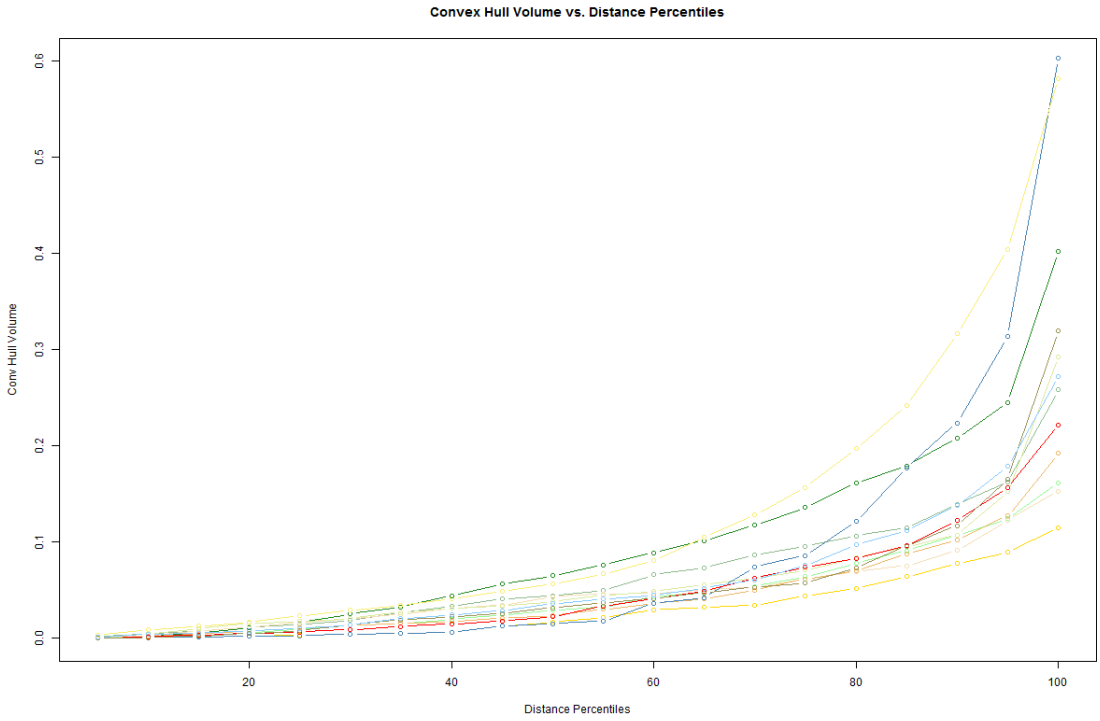


Figure 3.10: Class convex hull volumes using only the points whose distance to their mean fall below the given distance percentile.

Table 3.5: Gain matrix, **G**, determined using all of the northeast training data.

Class	1	4	5	7	8	9	10	11	12	13	14	17
1	0.21	0	5×10^{-4}	0	.05	0	0	4×10^{-3}	0	0.04	0	0
4	0	0.11	0	0	0	0	0	0	0	0	0.02	0
5	5×10^{-4}	0	0.14	0	0.02	0	0	0.05	0	0.01	0	0
7	0	0	0	0.09	0	0	0	0	0	0	0	0
8	0.05	0	0.02	0	0.11	0	0	0.05	0	0.06	0	0
9	0	0	0	0	0	0.08	0	0	0	0	0	0.03
10	0	0	0	0	0	0	0.10	0	2×10^{-3}	0	0	0
11	4×10^{-3}	0	0.05	0	0.05	0	0	0.22	0	0.05	0	0
12	0	0	0	0	0	0	2×10^{-3}	0	0.32	0	0	0
13	0.04	0	0.01	0	0.06	0	0	0.05	0	0.12	0	0
14	0	0.02	0	0	0	0	0	0	0	0	0.12	0
17	0	0	0	0	0	0.03	0	0	0	0	0	0.14

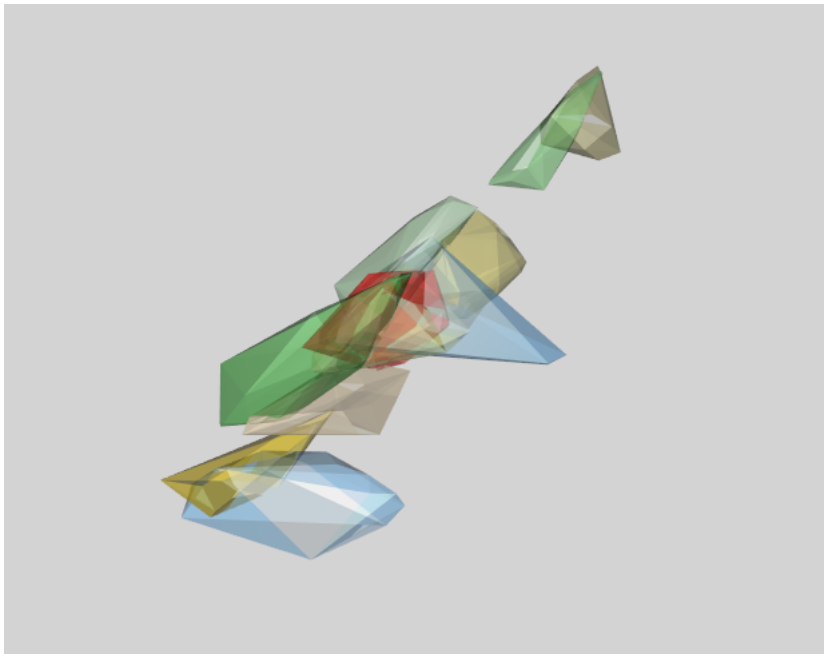


Figure 3.11: Convex hulls of the 12 land cover classes in our *single tesing dataset* framework, in the space of the first three PCs.

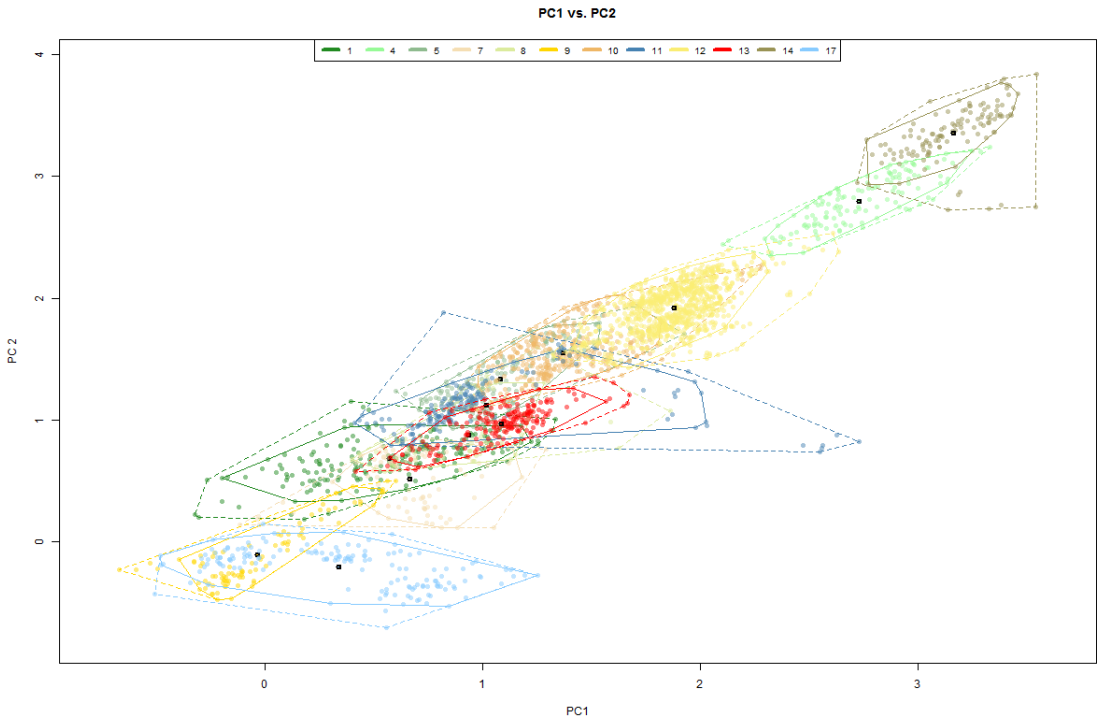


Figure 3.12: Plot of the northeast training data in the space of the first two PCs. The solid lines indicate convex hulls around all training points from that class. Dashed lines indicate convex hulls around points whose distance to their class mean falls below the $90^{\text{th}}-$percentile.

Table 3.6: Gain matrix, **G**, determined using bootstrap samples of the northeast training data and the second method described.

Class	1	4	5	7	8	9	10	11	12	13	14	17
1	0.14	0	3×10^{-6}	9×10^{-5}	0.05	0	0	5×10^{-3}	0	0.02	0	0
4	0	0.08	0	0	0	0	0	0	0	0	4×10^{-3}	0
5	3×10^{-6}	0	0.10	0	0.01	0	0	0.03	0	3×10^{-3}	0	0
7	9×10^{-5}	0	0	0.08	0	2×10^{-6}	0	0	0	0	0	0
8	0.05	0	0.01	0	0.11	0	0	0.03	0	0.05	0	0
9	0	0	0	2×10^{-6}	0	0.07	0	0	0	0	0	0.02
10	0	0	0	0	0	0	0.06	0	2×10^{-5}	0	0	0
11	5×10^{-3}	0	0.03	0	0.03	0	0	0.15	2×10^{-6}	0.01	0	0
12	0	0	0	0	0	0	2×10^{-5}	2×10^{-6}	0.13	0	0	0
13	0.02	0	3×10^{-3}	0	0.05	0	0	0.01	0	0.07	0	0
14	0	4×10^{-3}	0	0	0	0	0	0	0	0	0.08	0
17	0	0	0	0	0	0.02	0	0	0	0	0	0.10

3.4 Tree Approximation

As described above, computing the centroid on the lattice is intractable and so we propose graphically approximating the lattice. Variational approaches might attempt to approximate the distribution on the lattice, (3.6), with a distribution that eliminates the computational intractability of computing (3.8). In a vein similar to this, we pursue a graph approximation to the lattice which retains the most important features of the lattice structure while being simpler to compute on.

So that we can compute (3.8) and continue with posterior inference, we approximate the two-dimensional lattice with a spanning tree. Minimally connected, this spanning tree approximation allows for more efficient computations. This tree needs to retain the most important spatial information in order to ensure the quality of the approximation as well as preserve the purpose of the hierarchical graphical model. As alluded to above, variational approaches might seek the spanning tree in a way that minimizes the difference between the distribution on the tree and the distribution on the original lattice [115]. We pursue this spanning tree approximation in two alternate ways, using *mutual information* and EM.

3.4.1 Mutual Information Spanning Tree

The original classification approach presented in Section 3.1 assigns pixel labels independently. A natural first departure from this model, via a tree approximation to the lattice, seeks to retain only the edges between pixels that represent the biggest difference between modeling adjacent pixels jointly or independently. One of the most popular measures of difference between two probability distributions is the Kullback-Leibler divergence [59]:

$$D_{KL}(\mathbb{P}\|\mathbb{Q}) = \sum_i \mathbb{P}(i) \ln\left(\frac{\mathbb{P}(i)}{\mathbb{Q}(i)}\right), \tag{3.14}$$

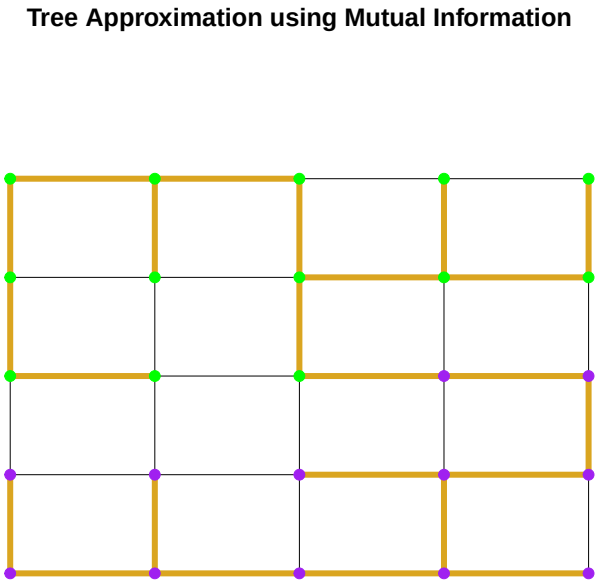


Figure 3.13: Toy example of mutual information spanning tree approximation.

where $\mathbb{P}$ and $\mathbb{Q}$ are discrete probability distributions. We will actually use what is commonly known as *mutual information* (MI) [71], a specific example of the KL-divergence:

$$I(\theta_u;\theta_v) = \sum_{\theta_u\in\mathcal{S}}\sum_{\theta_v\in\mathcal{S}} \mathbb{P}_1(\theta_u,\theta_v \mid X)\log\left(\frac{\mathbb{P}_1(\theta_u,\theta_v \mid X)}{\mathbb{P}_2(\theta_u \mid X)\mathbb{P}_2(\theta_v \mid X)}\right), \tag{3.15}$$

where $(u,v)\in L$ and $\mathbb{P}_1$ and $\mathbb{P}_2$ are naïve marginal and joint distributions assuming full independence. Using the mutual information between adjacent pixels as edge weights we approximate L with T as determined by a Maximum Weighted Spanning Tree algorithm [82]. Figure 3.13 displays a toy example, consisting of a 5×5 lattice with a simple two-class configuration denoted by the green and purple nodes. Because the overwhelming measure of a spatial relationship comes from the entries of $\mathbf{J}$ in (3.6), we can affect the edges that get picked up using mutual information by choosing to maximize either $I(\theta_u;\theta_v)$, or $-I(\theta_u;\theta_v)$.

Maximizing $I(\theta_u; \theta_v)$ will emphasize higher entries in $\mathbf{J}$, or class adjacency probabilities, and lead to the type of tree seen in Figure 3.13 where there exists only one edge on the boundary of the label regions. This makes sense since land cover classes are most likely seen adjacent to themselves.

Using mutual information to help approximate the distribution on the lattice, L, has been previously explored by Chow and Liu [15]. However, our approach differs in two significant ways. First, they do not use a specific graphical model (the lattice in our case) as the original distribution to be approximated. Second, their algorithm needs full marginal and joint probabilities. That is, our expression in (3.15) uses conditional probabilities and only requires joint probabilities for adjacent nodes in the lattice whereas their algorithm needs unconditional probabilities and joint probabilities for *all pairs* of nodes.

3.4.2 MI Tree Classification Results

Mutual information can be used to encode the edges of our lattice, L, with the spatial relationships we want to incorporate into our classification. Figure 3.13 gave a small example of such a tree, T. Figure 3.14 gives a land cover classification result using the centroid from (3.10) based on a spanning tree identified using mutual information. The region consists of a 100 pixel $\times$ 100 pixel region just above Florida. While noticeable differences exist between our result and the provided "reference" image, many of the differences are acceptable and actually demonstrate that we have achieved our aim. Our map is much more spatially homogeneous while still retaining the overall spatial patterns present in the reference. Additionally, the land cover classes less prevalent in our map (than the reference) are spectrally very similar to the ones most common.

3.4.3 Spanning Tree via EM

While we demonstrated that mutual information could be used to capture spatial relationships, we will in fact focus our attention on a different path forward. That is, approximating

the lattice with a spanning tree can introduce an additional layer to the model, yielding

$$\mathbb{P}(\boldsymbol{\theta} \mid X) \propto \sum_{T \in \tau(L)} \mathbb{P}(X \mid \boldsymbol{\theta}, T)\mathbb{P}(\boldsymbol{\theta} \mid T)\mathbb{P}(T) \tag{3.16}$$

where we assume $\mathbb{P}(T) \propto 1$ and $\tau(L)$ corresponds to the space of all spanning trees on L. Despite the approximation of L with T in (3.6), the sum over T in (3.16) presents a new computationally intractable piece of the model. To circumvent this, we propose the following approximation:

$$\begin{array}{rl} \mathbb{P}(\boldsymbol{\theta} \mid X) & \propto \sum_{T \in \tau(L)} \mathbb{P}(X \mid \boldsymbol{\theta}, T)\mathbb{P}(\boldsymbol{\theta} \mid T)\mathbb{P}(T) \\ & \approx \mathbb{P}(X \mid \boldsymbol{\theta}, T^*)\mathbb{P}(\boldsymbol{\theta} \mid T^*). \end{array} \tag{3.17}$$

Here, T^* represents a spanning tree optimized to capture the most important spatial relationships in the lattice. Formally, we desire

$$T^* = \arg\max_T \mathbb{P}(T \mid X) \quad \text{with} \quad \mathbb{P}(T \mid X) = \sum_{\boldsymbol{\theta}} \mathbb{P}(\boldsymbol{\theta}, T \mid X). \tag{3.18}$$

However, the expression in (3.18) still requires considering all possible spanning trees of and label combinations on L. We can, instead, obtain T^* by utilizing our model directly via an expectation-maximization procedure. Conveniently we have

$$\begin{array}{rl} T^{(t+1)} & = \arg\max_T \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}}[\log \mathbb{P}(\boldsymbol{\theta}, T \mid X)] \\ & = \arg\max_T \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}}[\log \mathbb{P}(\boldsymbol{\theta}, X, T)] \end{array} \tag{3.19}$$

With the parameters of (2.12) estimated and considered known, we can treat T as the parameter of our model to estimate in the presence of unknown (or latent) pixel labels, $\boldsymbol{\theta}$. To this end we have the following E-step:

$$\begin{array}{rl} Q(T, T^{(t)}) & = \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}}[\log \mathbb{P}(\boldsymbol{\theta}, X, T)] \\ & = \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}}[\log \mathbb{P}(X \mid \boldsymbol{\theta}, T) + \log \mathbb{P}(\boldsymbol{\theta} \mid T)] \\ & = K + \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}}[\sum_{(u,v) \in T} \mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v)] \\ & = K + \sum_{(u,v) \in T} \sum_{\theta_u, \theta_v} \mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v) \cdot \mathbb{P}(\theta_u, \theta_v \mid X, T^{(t)}). \end{array} \tag{3.20}$$

In (3.20), K contains all of the terms that do not depend on T. The M-step involves assigning the summands in the last line of (3.20) to the lattice edges as weights and then finding $T^{(t+1)}$ via a Maximum Weighted Spanning Tree algorithm. In this way we maximize the similarity, as measured by the entries of $\mathbf{J}$, of neighboring pixels. Figure 3.15 gives an example result of applying this tree approximation procedure to a toy map. Just as in Figure 3.13 the spanning tree, T^*, in Figure 3.15 connects regions of different land cover classes with only one edge! Because the EM procedure developed here is more formal and rigorous we prefer it over using mutual information to determine T^*. One main struggle with the EM, however, is the computation of $\mathbb{P}(\theta_u, \theta_v \mid X, T^{(t)})$ for each edge in L. This would involve integrating out every other node in the tree, $T^{(t)}$. For the time being, we approximate $\mathbb{P}(\theta_u, \theta_v \mid X, T^{(t)})$ with $\mathbb{P}(\theta_u \mid X, T^{(t)})\mathbb{P}(\theta_v \mid X, T^{(t)})$.

3.4.4 EM Tree Classification Results

According to the last line of (3.20), a combination of the entries of $\mathbf{J}$ and the joint posteriors weight the edges of the lattice and allow for the determination of a spanning tree approximation which maximizes these weights and captures the most important spatial information. With missing data imputed we use EM to estimate this spanning tree by letting the pixel labels be latent. After determining T^* using EM, we begin by checking the accuracy of this approach on our original training dataset from the northeast of North America with a confusion matrix in Table 3.7. To be fair, we also present a confusion matrix of the same training data set after imputing missing data in the same way (and with the same parameters), but classified using the independent pixel classification method. Table 3.8 contains these results and perhaps demonstrates the superiority of the independent pixel classification ($\eta = 0$) to the tree classification ($\eta = 1$). Overall, the tree results echo our analysis from earlier (Section 3.1.1) in displaying a reasonably good total accuracy and class-specific difficulties. The graphical model should not be dismissed based on these results alone. For instance, these independent classification results might be better because the likelihood is more informative and the prior is too strong spatially in the wrong

Table 3.7: Confusion matrix for classification of northeast training data using graphical model. The rows represent our classification; the columns represent the true classification.

Class	1	4	5	7	8	9	10	11	12	13	14	17	Total	Precision
1	178	0	58	0	24	13	0	8	0	0	0	0	281	0.63
4	0	131	43	0	9	2	0	1	36	6	71	0	299	0.44
5	13	0	80	4	28	22	0	39	1	102	0	0	289	0.28
7	0	0	0	23	0	0	0	34	0	0	0	0	57	0.4
8	0	0	2	0	1	1	0	3	0	0	0	0	7	0.14
9	0	0	1	3	2	21	0	69	0	0	0	0	96	0.22
10	0	0	0	29	0	12	243	0	81	0	0	0	365	0.67
11	0	0	0	8	1	4	0	23	0	0	0	0	36	0.64
12	0	0	0	21	0	14	83	0	792	32	31	0	973	0.81
13	20	0	0	7	0	4	1	0	0	64	0	0	96	0.67
14	0	0	0	0	1	0	0	0	8	3	22	0	34	0.65
17	0	0	0	0	0	0	0	0	0	0	0	159	159	1
Total	211	131	184	95	66	93	327	177	918	207	124	159	2692	-
Accuracy	0.84	1	0.43	0.24	0.02	0.23	0.74	0.13	0.86	0.31	0.18	1	-	**0.65**

Table 3.8: Confusion matrix for classification of northeast training data using independent classification analysis pipeline. The rows represent our classification; the columns represent the true classification.

Class	1	4	5	7	8	9	10	11	12	13	14	17	Total	Precision
1	162	0	17	0	15	1	0	2	0	0	0	0	197	0.82
4	0	125	32	0	5	0	0	0	11	0	35	0	208	6×10^{-1}
5	22	6	124	0	31	0	0	0	0	11	0	0	194	0.64
7	0	0	0	20	0	0	0	20	0	0	0	0	40	5×10^{-1}
8	6	0	8	0	2	6	0	1	0	0	0	0	23	9×10^{-2}
9	0	0	3	2	6	36	0	70	0	0	0	1	118	0.31
10	0	0	0	21	0	8	173	0	41	0	0	0	243	0.71
11	0	0	0	12	3	14	0	71	0	0	0	0	1×10^{2}	0.71
12	0	0	0	19	1	1	110	0	798	0	24	0	953	0.84
13	21	0	0	21	1	27	44	13	8	196	0	0	331	0.59
14	0	0	0	0	2	0	0	0	60	0	65	0	127	0.51
17	0	0	0	0	0	0	0	0	0	0	0	158	158	1
Total	211	131	184	95	66	93	327	177	918	207	124	159	2692	-
Accuracy	0.77	0.95	0.67	0.21	0.03	0.39	0.53	0.4	0.87	0.95	0.52	0.99	-	**0.72**

direction. That is, it merely appears that the graphical model with $\eta = 0$ performs better than with $\eta = 1$. We should actually calibrate the values of β and η instead of simply experimenting with values, and we discuss a method for doing so later in Section 3.4.5.

Figure 3.16 gives a classification result for Montreal using the hierarchical model described and the centroid estimator. Clear differences can be seen between Figures 3.3 and 3.16. Because we now incorporate spatial information, the classification result is coarser and displays less of a salt-and-pepper look. Of course we can increase the emphasis on spatial homogeneity by increasing the value of η, as in Figure 3.17. On the other hand, we can recover our map in Figure 3.3 by setting $\beta = 1$ and $\eta = 0$.

We must take care when setting the value of η as overly high values can coarsen the classification result too much. For example, in Figure 3.17 observe that some of the urban (red) portions disappear as we move from $\eta = 1$ to $\eta = 10$. In fact, despite the usefulness

of our graphical model in incorporating spatial information, our original independent pixel classification result in Figure 3.3 already represents a coarsening of reality.

Recall that the spatial resolution of our MODIS data is 500 m^2. A potential issue and difficulty in performing land cover classification at this scale comes in the form of *mixed pixels*. That is, many (*mixed*) pixels consist of multiple land cover classes from our IGBP scheme. Consequently, spectrally similar or geographically close land cover classes often get confused at a given pixel. For this reason our collaborators in Geography often emphatically approve of our independent pixel classification results. With this feeling in mind, our first and main goal of using the hierarchical model consists of ridding our classification result of isolated pixel labels within spatially homogeneous regions (i.e. preventing a salt-and-pepper look). This may suggest that an ideal map requires only a relatively small value of η, somewhere between 1 and 5.

While various challenges will inevitably obstruct the classification process, our hierarchical graphical model can beautifully accommodate any kind of prior knowledge about the global distribution ($\mathbf{h}$) of and spatial relationships ($\mathbf{J}$) between land cover classes; as well as the amount of spatial homogeneity (η) to enforce in the classification results. As mentioned above, we can even recover the independent pixel classification result using our hierarchical graphical model with particular values of β and η. Full classification results for North America using the centroid estimator can be found in Figure B.2 in Appendix B.

3.4.5 Hyperparameter Elicitation

The values of β and η in (3.6) specify the strengths of the global and spatial "priors" on the labels, respectively. They can be chosen by the user or calibrated using a separate, pre-classified image. In the case of calibration, we propose the following iterative procedure:

$$\begin{bmatrix} \beta^{(t+1)} \\ \eta^{(t+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(t)} \\ \eta^{(t)} \end{bmatrix} - H^{(t)^{-1}}(x)U^{(t)} \tag{3.21}$$

where $H^{(t)}(x)$ and U^t are the hessian and gradient of (3.6), respectively. The following expressions comprise the entries of H and U:

$$\begin{aligned}
\frac{\partial \mathbb{P}(\boldsymbol{\theta})}{\partial \beta} &= \gamma_h - \frac{\sum_{\tilde{\theta}} \tilde{\gamma}_h \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)}{\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)} \\
\frac{\partial \mathbb{P}(\boldsymbol{\theta})}{\partial \eta} &= \gamma_J - \frac{\sum_{\tilde{\theta}} \tilde{\gamma}_J \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)}{\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)} \\
\frac{\partial^2 \mathbb{P}(\boldsymbol{\theta})}{\partial \beta^2} &= -\frac{[(\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))(\sum_{\tilde{\theta}} \tilde{\gamma}_h^2 \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)) - (\sum_{\tilde{\theta}} \tilde{\gamma}_h \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))^2]}{(\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))^2} \\
\frac{\partial^2 \mathbb{P}(\boldsymbol{\theta})}{\partial \eta^2} &= -\frac{[(\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))(\sum_{\tilde{\theta}} \tilde{\gamma}_J^2 \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)) - (\sum_{\tilde{\theta}} \tilde{\gamma}_J \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))^2]}{(\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))^2} \\
\frac{\partial^2 \mathbb{P}(\boldsymbol{\theta})}{\partial \beta \partial \eta} &= -\frac{[(\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))(\sum_{\tilde{\theta}} \tilde{\gamma}_h \tilde{\gamma}_J \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J)) - (\sum_{\tilde{\theta}} \tilde{\gamma}_h \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))(\sum_{\tilde{\theta}} \tilde{\gamma}_J \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))]}{(\sum_{\tilde{\theta}} \exp(\beta\tilde{\gamma}_h + \eta\tilde{\gamma}_J))^2}
\end{aligned} \tag{3.22}$$

where $\gamma_h = \sum_v \mathbf{h}^\top \mathbf{e}(\theta_v)$ and $\gamma_J = \sum_{(u,v)\in T} \mathbf{e}(\theta_u)^\top \mathbf{J}\mathbf{e}(\theta_v)$. Unfortunately, the expressions in (3.22) contain some computationally intractable quantities. We identify the following computationally feasible quantities of interest for this procedure to find β and η.

$$\begin{aligned}
\mathsf{egh} &= \textstyle\sum_v \mathbb{P}(\theta_v) h_{\theta_v} \\
\mathsf{egh2} &= \textstyle\sum_v \mathbb{P}(\theta_v) h_{\theta_v}^2 \\
\mathsf{egJ} &= \textstyle\sum_{(u,v)\in T} \mathrm{tr}(\mathbb{P}(\theta_u, \theta_v) J) \\
\mathsf{egJ2} &= \textstyle\sum_{(u,v)\in T} \mathrm{tr}(\mathbb{P}(\theta_u, \theta_v)(J \circ J)) \\
\mathsf{egJJ} &= \textstyle\sum_{(u,v)\in T,(s,t)\in T} \mathbb{P}(\theta_u, \theta_v)\mathbb{P}(\theta_s, \theta_t) J_{\theta_u\theta_v} J_{\theta_s\theta_t} \\
\mathsf{egJJe} &= \textstyle\sum_{(u,v)\in T,(v,w)\in T,\, u\neq w} \sum_{\theta_u,\theta_v,\theta_w} \mathbb{P}(\theta_u, \theta_v, \theta_w) J_{\theta_u\theta_v} J_{\theta_v\theta_w} \\
\mathsf{eghh} &= \textstyle\sum_{u,v\in G} \mathbb{P}(\theta_u, \theta_v) h_{\theta_u} h_{\theta_v} \\
\mathsf{eghJ} &= \textstyle\sum_{u\in T,(v,w)\in T, u\neq v, u\neq w} \mathbb{P}(\theta_u, \theta_v, \theta_w) h_{\theta_u} J_{\theta_v\theta_w}
\end{aligned} \tag{3.23}$$

Assuming top-down and bottom-up messages (along the tree, T) have been computed, the implementation in Algorithm 2 should yield the quantities in (3.23) for a general graph, G, which we wish to approximate with T. In our Remote Sensing context we use the lattice, L, for our graph. We pursue the quantities found in (3.23), to use in the following way:

$$U^{(t)} = \begin{bmatrix} \gamma_h^{(t)} - \mathsf{egh} \\ \gamma_J^{(t)} - \mathsf{egJ}^{(t)} \end{bmatrix} \tag{3.24}$$

Algorithm 2: Computation of β and η

Initialize: $\mathsf{egh} \leftarrow 0$; $\mathsf{egh2} \leftarrow 0$; $\mathsf{egJ} \leftarrow 0$; $\mathsf{egJ2} \leftarrow 0$; $\mathsf{eghh} \leftarrow 0$; $\mathsf{egJJe} \leftarrow 0$; $\mathsf{eghJ} \leftarrow 0$;
for $i \leftarrow 1, \ldots, |V|$ **do**
  $v \leftarrow \sigma[i]$; `// Reference node`
  Compute $\mathbb{P}(\theta_v)$;
  $\mathsf{egh} \leftarrow \mathsf{egh} + \sum_{\theta_v} \mathbb{P}(\theta_v) h_{\theta_v}$
  $\mathsf{egh2} \leftarrow \mathsf{egh2} + (h^2)^\top \mathbb{P}(\theta_v)$
  for $j \leftarrow i+1, \ldots, |V|$ **do**
    $w \leftarrow \sigma[j]$;
    FIND $u = \sigma[\max_{k<j}\{(\sigma[k], w) \in T\}]$ `//` $k \geq i$
    SEND $u \rightsquigarrow w$ and compute $\mathbb{P}(\theta_v, \theta_w)$
    if $(v, w) \in T$ **then**
      $\mathsf{egJ} \leftarrow \mathsf{egJ} + \mathrm{tr}[\mathbb{P}(\theta_v, \theta_w) \cdot J]$
      $\mathsf{egJ2} \leftarrow \mathsf{egJ2} + \mathrm{tr}[\mathbb{P}(\theta_v, \theta_w) \cdot (J \circ J)]$
    else
      if $(v, w) \in G$ **then**
        STORE $\mathbb{P}(\theta_v, \theta_w)$
      else
      end
    end
    $\mathsf{eghh} \leftarrow \mathsf{eghh} + h^\top \mathbb{P}(\theta_v, \theta_w) h$
    $S = \{(s, w) \in T : \sigma^{-1}[s] > j\}$
    for $s \in S$ **do**
      Compute $\mathbb{P}(\theta_v, \theta_w, \theta_s)$
      if $(v, w) \in T$ **then**
        $\mathsf{egJJe} \leftarrow \mathsf{egJJe} + \sum_{\theta_v, \theta_w, \theta_s} \mathbb{P}(\theta_v, \theta_w, \theta_s) \cdot J_{\theta_v, \theta_w} \cdot J_{\theta_w, \theta_s}$
      else
      end
      $\mathsf{eghJ} \leftarrow \mathsf{eghJ} + \sum_{\theta_v, \theta_w, \theta_s} \mathbb{P}(\theta_v, \theta_w, \theta_s) \cdot h_{\theta_v} \cdot J_{\theta_w, \theta_s}$
    end
  end
end

$$\begin{array}{lll} \widehat{H}_{11}^{(t)} & = \frac{\partial^2 \mathbb{P}(\boldsymbol{\theta})}{\partial \beta^2} & = -[\mathsf{eghh}^{(t)} + \mathsf{egh2} - \mathsf{egh}^2] \\ \widehat{H}_{22}^{(t)} & = \frac{\partial^2 \mathbb{P}(\boldsymbol{\theta})}{\partial \beta^2} & = -[(\mathsf{egJ2}^{(t)} + \mathsf{egJJe}^{(t)} + \mathsf{egJJ}^{(t)}) - \mathsf{egJ}^{(t)^2}] \\ H_{12}^{(t)} & = H_{21}^{(t)} & = -[\mathsf{eghJ}^{(t)} - (\mathsf{egh})(\mathsf{egJ}^{(t)})] \end{array} \tag{3.25}$$

The notation, $\widehat{H}$, indicates that we will not compute the hessian matrix exactly but an approximation to it. Recall that in (3.20) we actually need to compute $\mathbb{P}(\theta_u, \theta_v \,|\, X, T^{(t)})$ for each edge in L. We currently approximate these with the products of the marginal posteriors, but the procedure in Algorithm 2, for the calibration of β and η in (3.6), will also obtain the joint posteriors we want here. The procedure described in Algorithm 2 could also produce an empirical estimate of η that should help ensure a realistic amount of spatial homogeneity in maps produced using our spanning tree approximation and the centroid estimator.

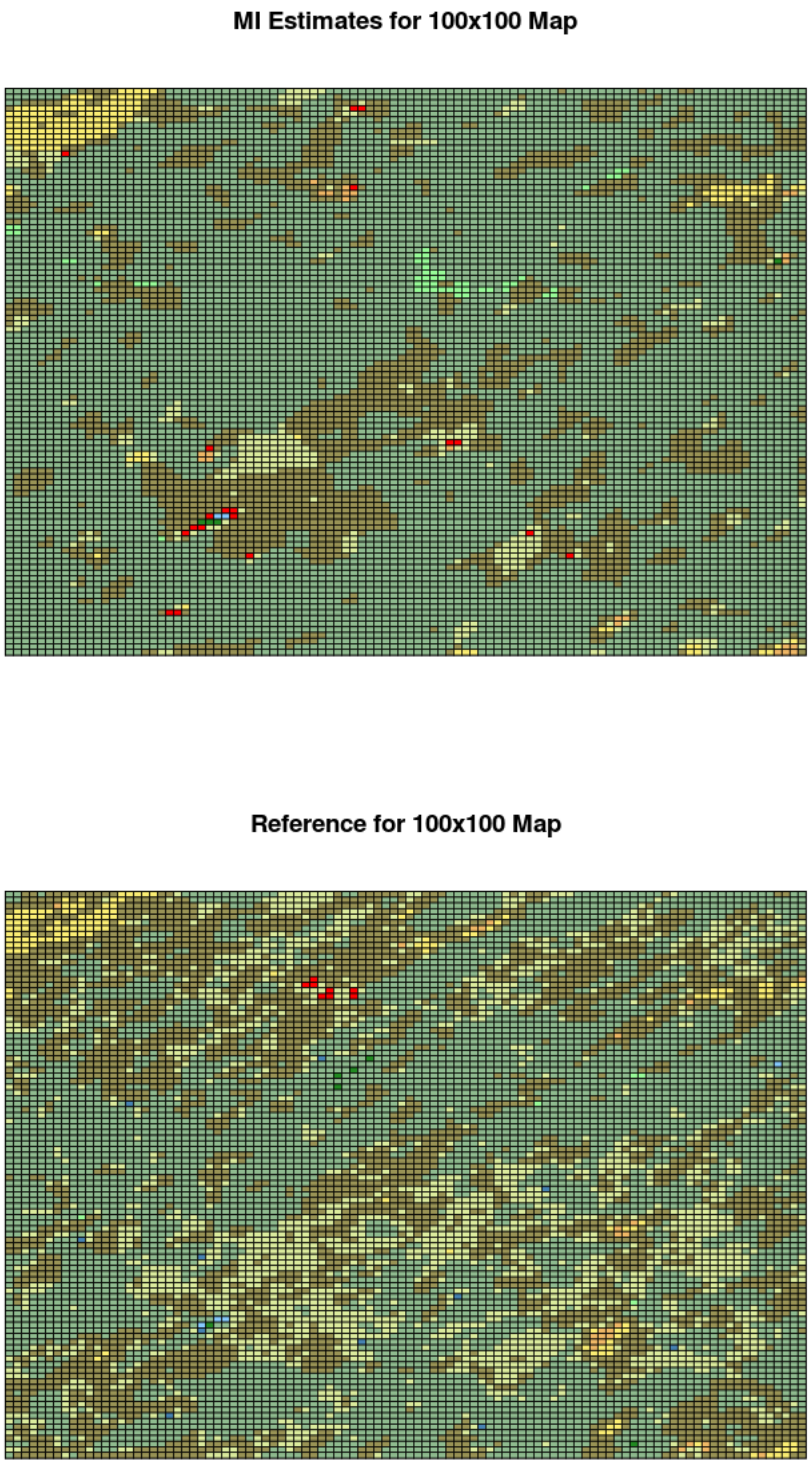


Figure 3.14: Land cover map of a 100 pixel $\times$ 100 pixel region in MODIS tile h10v05, the southeast portion of the United States above Florida, using the centroid estimator based on a T determined with mutual information ($\beta = 1$, $\eta = 1$).

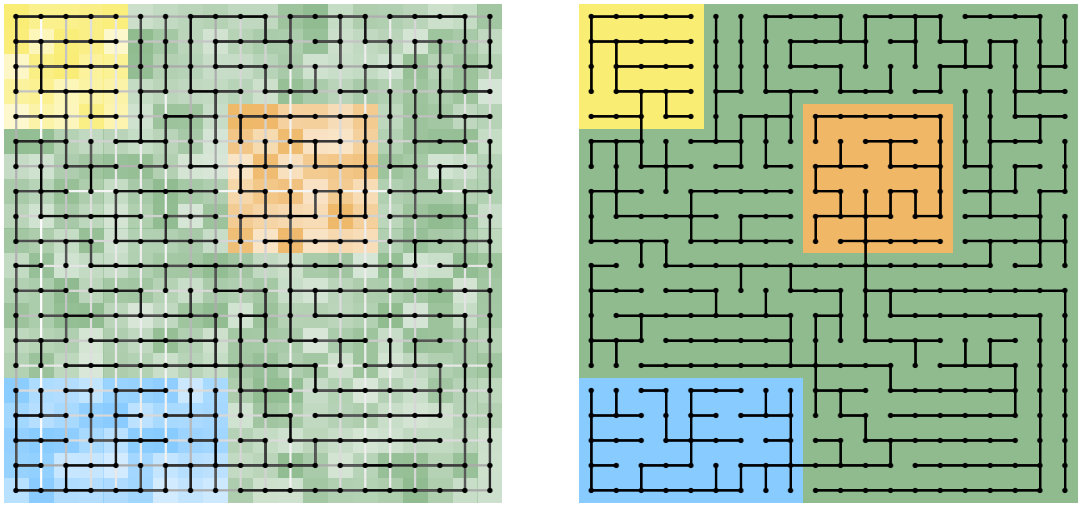


Figure 3.15: Example result of approximating tree, T^*, using EM. Opacity of pixels in left plot indicate strength of the posterior of the true class. Darkness of edges in left plot indicate the strength of edge weights as shown in (3.20). Final solution in right plot.

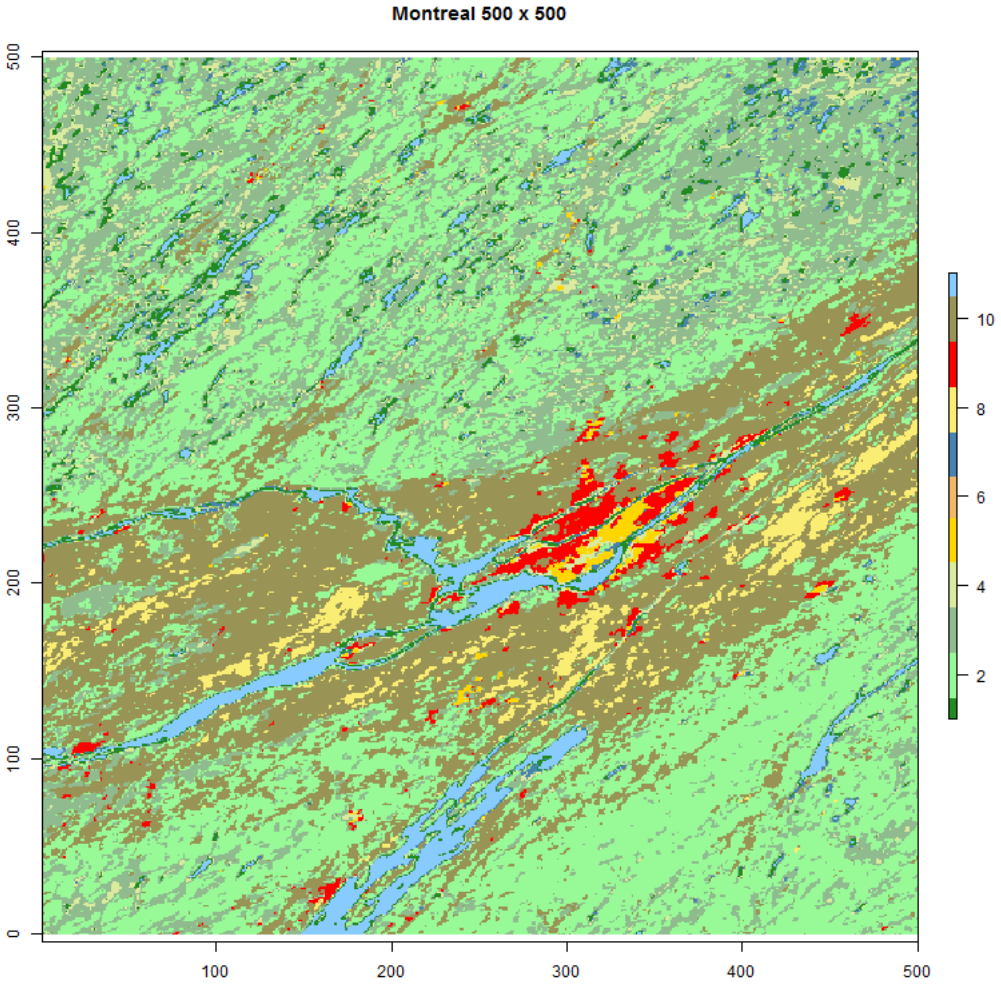


Figure 3.16: Land cover map of Montreal, Canada using training data in the northeast portion of North America with $\beta = 1$ and $\eta = 1$.

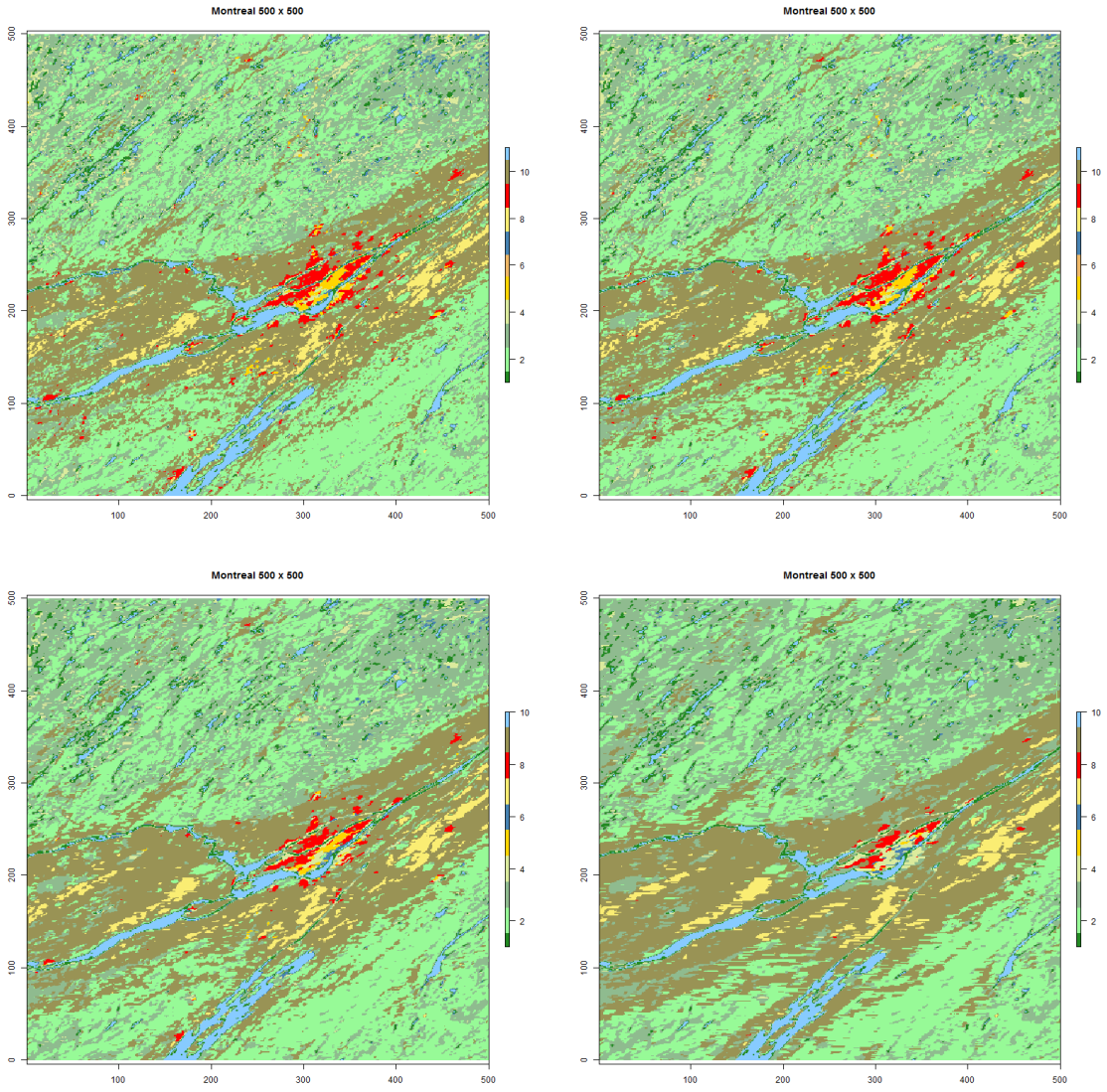


Figure 3.17: Land cover map of a 500 pixel square region surrounding Montreal, Canada (top-left: $\eta = 1$, top-right: $\eta = 2$, bottom-left: $\eta = 5$, bottom-right: $\eta = 10$; $\beta = 1$ for all plots).

Chapter 4

Change Point Detection

4.1 Introduction

In an effort to explore the problem more and work up to using the machinery developed throughout the previous chapters, we consider both univariate and multivariate change point frameworks. In what follows we introduce and assess a few change point detection models tailored to the data characteristics of MODIS time series. That is, exploiting land cover information from STEP training data, we specify a series of models and EM procedures to detect distributional changes in multispectral time series while accounting for missing data. We assess the performance of our final method in a simulation study (Section 4.4) and case study in the Xingu Basin in the Amazon (Section 4.5).

4.2 Univariate Change Point Detection

Each pixel in our original MODIS data consisted of a single-year, multivariate observation of 7 spectral values at 46 time points. Here, however, we possess multivariate observations for $n = 11$ years. We aim to detect changes at the scale of years. That is, we wish to identify the year in which a change occurred and not the time point (within a year) at which it occurred. To start simply we develop a univariate change point detection procedure which uses either band 7 or a vegetation index constructed from the spectral bands, such as EVI, as opposed to all 7 spectral bands. Let X_{iv} denote the value of pixel v in year i, and c_v be

the unknown year of the change point for pixel v. We specify the following likelihood:

$$\begin{aligned} X_{iv} \,|\, i < c_v &\overset{\text{ind}}{\sim} N(\mu_F, \sigma_F^2) \\ X_{iv} \,|\, i \geq c_v &\overset{\text{ind}}{\sim} N(\mu_{cv}, \sigma_{cv}^2) \end{aligned} \tag{4.1}$$

The use of the "F" subscript stems from our case study region of interest consisting of forested pixels that change for some reason (i.e. fire, deforestation, etc.). Similar to before, we face the problem of missing data. Let i_Y and i_Z denote the year indices of non-missing (Y) and missing data (Z), respectively. Then the full data log-likelihood is

$$\begin{aligned} l(Y, Z, \Theta) = &\sum_{i_Y < c} \log P(Y_{i_Y}; \Theta) + \sum_{i_Y \geq c} \log P(Y_{i_Y}; \Theta) \\ &+ \sum_{i_Z < c} \log P(Z_{i_Z}; \Theta) + \sum_{i_Z \geq c} \log P(Z_{i_Z}; \Theta) \end{aligned} \tag{4.2}$$

Appendix C.1 gives derivations of the updates for μ_{cv} and σ_{cv}^2 in an EM procedure constructed to estimate them in the presence of missing data. Initial values should be obtained by taking the mean and variance of the non-missing values (Y). When analyzing pixel v, we run this EM procedure until convergence for each choice of change point year (c_v) including the possibility of no-change. Change detection for pixel v could then be performed by choosing the value of c_v which maximized $Q(\Theta, \Theta^{(t)})$ at convergence. However, because we possess some knowledge about the frequency of change in land cover, we will instead postpone inference until after we enrich our model in a Bayesian way.

4.2.1 Univariate Bayesian CPD

So far we have only been looking at the likelihood of a change. That is, we model the data in terms of a specific change point, c_v, or lack there of. To extend this we would like to incorporate our prior beliefs about observing change points. Specifically, we will define the probability of seeing a certain number of change points; then given this number, we specify the probability that the change point occurs in year c_v. As per our collaborator's expert opinion we start with the following if K_v is the number of change points at pixel v and C_v

is the location of the change:

$$\begin{aligned} P(K_v = 0) = .98, \quad & P(K_v = 1) = .02 \\ P(C_v = c_v \,|\, K_v = 1) \quad & = \tfrac{1}{\binom{n-1}{1}} \end{aligned} \tag{4.3}$$

We assume that there exists at most one change point in a given time series. The former probabilities of observing a change were elicited from our collaborator. The latter probability on the location of the change point represents a non-informative (or uniform) prior belief about what year the change point will occur in (i.e. each year is equally likely). Incorporating this *a priori* information involves a simple amendment to (4.1). We will decide on the change point in the same way we did before. The outcome with the highest value of $Q(\Theta, \Theta^{(t)})$ corresponds to the change point we would call.

4.3 Multivariate Bayesian CPD

The method described above in Section 4.2 can be extended very easily to a multivariate framework. Suppose now we wish to use any subset of the 7 spectral bands we began with:

$$\begin{aligned} X_{iv} \,|\, i < c_v \quad & \overset{\text{ind}}{\sim} N(\mu_F, \Sigma_F) \\ X_{iv} \,|\, i \geq c_v \quad & \overset{\text{ind}}{\sim} N(\mu_{cv}, \Sigma_{cv}) \end{aligned} \tag{4.4}$$

where $\Sigma_F = \sigma_F^2 \Sigma_{s,F} \otimes \Sigma_{t,F}$ and $\Sigma_{cv} = \sigma_{cv}^2 \Sigma_{s,cv} \otimes \Sigma_{t,cv}$. As in (2.12), our annual observations become matrix observations that follow a multivariate normal distribution whose Kronecker covariance breaks into spectral and temporal pieces of variation. We use the methods established in Section 2.2 to estimate the *post-change* parameters of (4.4) in the presence of missing data, for each value of c_v. We identify a change point, again, by the maximizer of $Q(\Theta, \Theta^{(t)})$ after incorporating the same *a priori* information about change points given in (4.3).

Note that the approaches described in Section 4.2 and this one specify a distribution and parameters for a specific pre-change land cover. That is, μ_F and Σ_F are assumed known or estimated from training data in the region of interest. We estimate the post-

change parameters from the data itself at each pixel. Dimensionality can be an issue here since post-change parameter estimation requires a minimum number of observations and we possess only 11 years of data (i.e. 11 observations in the entire time series). This can limit the number of years we consider as potential change points. To circumvent this difficulty we move on to proposing a model which makes explicit use of the IGBP classification scheme in a way that does not require estimation of the same, potentially large parameters from just 11 observations.

4.3.1 Class-to-Class CPD

In this scenario, we want to compare the *pre-change* class to other specific classes instead of some "fitted" *post-change* class as described in the previous sections, 4.2 and 4.3. Suppose we have $\mathscr{C}$ potential post-change classes (which do not include the pre-change class). We estimate parameters for each of the $\mathscr{C}$ classes using our previously established EM procedure, in Section 2.2.2, with independent training data from the training site database in the region surrounding the area we wish to detect change in. With these parameters estimated we specify the following model for pixel v:

$$\begin{aligned}
i < c_v : X_{iv} \,|\, \mu_{0v} &\sim N(\mu_{0v}, \kappa_0 \mathbb{I}) \\
\mu_{0v} &\sim N(\mu_F, \Sigma_F) \\
\\
i \geq c_v : X_{iv} \,|\, \mu_{cv} &\sim N(\mu_{cv}, \kappa_c \mathbb{I}) \\
\mu_{cv} \,|\, W_v = g &\sim N(\mu_g, \Sigma_g) \\
W_v \,|\, \boldsymbol{\alpha} &\sim \mathsf{MN}(1, \boldsymbol{\alpha}) \\
\boldsymbol{\alpha} &\sim \mathsf{Dir}(\boldsymbol{\pi}) \\
C_v = c_v \,|\, K_v = 1 &\propto 1 \\
\mathbb{P}(K_v = 1) = .02, \quad &\mathbb{P}(K_v = 0) = .98
\end{aligned} \tag{4.5}$$

where again the covariances specified here take the form of a Kronecker product. In (4.5), K_v once again represents the number of change points at pixel v and C_v the location of the change point. Because X_{iv} contains missing data, Z_{iv}, we derive an expectation-maximization procedure to estimate C_v. Derivations of the following update equations can

be found in Appendix C.2.

$$\begin{aligned}
\alpha_k^{(t+1)} &= \tfrac{\sum_v \mathbb{P}(W_v=k \,|\, Y_v,\Theta^{(t)})+\pi_k-1}{N-\mathscr{C}+\sum_g \pi_g} \\
\mu_{0v}^{(t+1)} &= \left(\tfrac{(c_v-1)}{\kappa_0}\mathbf{I}+\Sigma_F^{-1}\right)^{-1}\left(\tfrac{1}{\kappa_0}\textstyle\sum_{i<c_v}\mathbb{E}_{Z\,|\,Y,\Theta^{(t)}}[X_{iv}]+\Sigma_F^{-1}\mu_F\right) \\
\mu_{cv}^{(t+1)} &= \left(\tfrac{(n-c_v+1)}{\kappa_c}\mathbf{I}+\textstyle\sum_{g=1}^{\mathscr{C}}\mathbb{P}(W_v=g\,|\,Y_v,\Theta^{(t)})\Sigma_g^{-1}\right)^{-1}\cdot \\
&\quad \left(\tfrac{1}{\kappa_c}\textstyle\sum_{i\geq c_v}\mathbb{E}_{Z\,|\,Y,\Theta^{(t)}}[X_{iv}]+\sum_{g=1}^{\mathscr{C}}\mathbb{P}(W_v=g\,|\,Y_v,\Theta^{(t)})\Sigma_g^{-1}\mu_g\right)
\end{aligned} \tag{4.6}$$

Equation (4.6) makes use of the following posterior computation:

$$\mathbb{P}(W_v=g\,|\,Y_v,\Theta^{(t)})=\frac{\alpha_g^{(t)}\mathbb{P}(Y_v\,|\,W_v=g,\Theta^{(t)})}{\sum_{\tilde{g}}\alpha_{\tilde{g}}^{(t)}\mathbb{P}(Y_v\,|\,W_v=\tilde{g},\Theta^{(t)})}. \tag{4.7}$$

For each pixel, v, we select the value of c_v which maximizes the function, $Q(\Theta,\Theta^{(t)})$, of this EM procedure. Determining c_v in this way requires the computation of Q for every possible change point year and iterating, making this procedure potentially tedious if the region of interest is large. We apply this method in a simulation study as well as a case study in the following two sections.

For this class-to-class model, described above, we need to estimate parameters for the land cover classes prevalent in our region of interest: the Xingu River Basin in the southeastern part of the Amazon. To this end we extract data from the MODIS Training Site Database, located in South America between 0 and 20°S. Table 4.1 gives the class frequencies for this training dataset. Evergreen broadleaf forests (class 2) constitute our pre-change data. Because preliminary analysis revealed that bands 1 and 7 contain the most spectrally distinguishable information and a smaller proportion of noise than the inclusion of all 7 bands, we utilize only these two bands in what follows.

Table 4.1: IGBP classes and their representativeness in the Amazon training dataset.

IGBP	Description	# Pixels
2	Evergreen Broadleaf Forests	1335
4	Deciduous Broadleaf Forests	480
5	Mixed Forests	122
7	Open Shrublands	425
8	Woody Savannas	116
9	Savannas	89
10	Grasslands	403
11	Permanent Wetlands	706
12	Croplands	757
13	Urban and Built-Up Lands	109
14	Cropland/Natural Veg. Mosaics	71
15	Permanent Snow and Ice	97
17	Water Bodies	125

4.4 Change Point Simulation Study

Our change point simulation study makes use of the estimated class parameters to simulate datasets consisting of some pixels with a change and some without. More specifically, a single replication involves simulating 60 pixels which do not contain a change and 60 pixels which do contain a single change point. For each pixel we simulate 10 years (observations) of data. Each annual profile consists of simulating bands 1 and 7 over 19 time points. We simulate *no-change* data entirely from class 2 (Evergreen Broadleaf Forests). Whereas *change* data begins with class 2 and moves to a different class at some point in the time series. A randomly chosen 5% of each observation is assumed to be missing. We vary two things in this study: the post-change class and the change point year. There are 12 land cover classes outside of our designated pre-change class and 60 changed pixels in each replication, so we simulate changing to each of these 12 classes 5 times within each replication. The change point year varies from year 3 to year 7, with each one replicated 20 times for a total of 100 replicates in our simulation study.

Table 4.2 gives a confusion matrix of results aggregated for the entire simulation study.

Recall that we have 10 years of data in each simulated pixel, and so we can identify any of these 10 years as the change year OR decide the pixel did not change. Overall, the accuracies and precisions indicate that our method can perform very well in finding a change between Evergreen Broadleaf Forests and the other land cover classes. The most striking features of this confusion matrix appear in the *No-Change* row and column. We call a noticeable amount of false-positives in years 9 and 10 of these pixel time series. As we shall see in the case study below, our method can pick up on small changes. This is likely the case here in the sense that sufficient variation in the simulated data caused a change to be called.

Tables 4.3 and 4.4 summarize the simulation study results when the post-change classes are Mixed Forests (class 5) and Wetlands (class 11), respectively. These two particular post-change classes contain nearly all of the mistakes made by our method outside of the *no-change* situation described above. Mixed Forests and Wetlands are both mixture classes and thus challenging ones to identify and distinguish, as evidenced by the results in Section 3.1. Furthermore, Mixed Forests can, by definition, contain Evergreen Broadleaf Forests which likely explains why most of the mistakes in Table 4.3 consist of no-change results. Confusion in the other direction also exists upon inspection of the results in Table 4.5. In other words, the mistaken *no-change* pixels in Table 4.2 supposedly changed to classes 5 (Mixed Forests), 8 (Woody Savannas), or 14 (Cropland/Natural Vegetatin Mosaics). These estimated post-change classes can be informally obtained via the arg max of the posteriors in (4.7).

The specification of π in (4.5) provides an advantageous flexibility that we can utilize to inform the model of land cover classes we anticipate seeing after a change has occurred, making our approach particularly well suited for changes in the form of land cover conversion. In any case, changes from Evergreen Broadleaf Forests to *Croplands* or *Urban and Built-Up Lands* constitute the most realistic changes in the data we consider here. In general, though, Evergreen Broadleaf Forests will likely never completely change to a land cover class exhibiting extreme spectral similarities.

Table 4.2: Overall confusion matrix for change point simulation study. Columns correspond to the true year of change. Rows correspond to the year of change we identify. The first row and column correspond to *no change*.

Year	No-Change	1	2	3	4	5	6	7	8	9	10	Total	Precision
No-Change	4919	0	0	0	0	30	55	90	0	0	0	5094	0.97
1	0	0	0	0	0	0	0	0	0	0	0	0	-
2	0	0	0	15	6	1	2	0	0	0	0	24	0
3	4	0	0	1185	12	3	2	0	0	0	0	1206	0.98
4	2	0	0	0	1182	9	6	2	0	0	0	1201	0.98
5	5	0	0	0	0	1157	13	3	0	0	0	1178	0.98
6	16	0	0	0	0	0	1122	14	0	0	0	1152	0.97
7	32	0	0	0	0	0	0	1091	0	0	0	1123	0.97
8	78	0	0	0	0	0	0	0	0	0	0	78	0
9	196	0	0	0	0	0	0	0	0	0	0	196	0
10	748	0	0	0	0	0	0	0	0	0	0	748	0
Total	6000	0	0	1200	1200	1200	1200	1200	0	0	0	12000	-
Accuracy	0.82	-	-	0.99	0.98	0.96	0.94	0.91	-	-	-	-	**0.89**

Table 4.3: Confusion matrix of results when post-change class is Mixed Forests (class 5).

Year	No-Change	1	2	3	4	5	6	7	8	9	10	Total	Precision
No-Change	0	0	0	0	0	30	55	90	0	0	0	175	0
1	0	0	0	0	0	0	0	0	0	0	0	0	-
2	0	0	0	3	2	0	0	0	0	0	0	5	0
3	0	0	0	97	5	0	1	0	0	0	0	103	0.94
4	0	0	0	0	93	2	1	0	0	0	0	96	0.97
5	0	0	0	0	0	68	2	0	0	0	0	70	0.97
6	0	0	0	0	0	0	41	0	0	0	0	41	1
7	0	0	0	0	0	0	0	10	0	0	0	10	1
8	0	0	0	0	0	0	0	0	0	0	0	0	-
9	0	0	0	0	0	0	0	0	0	0	0	0	-
10	0	0	0	0	0	0	0	0	0	0	0	0	-
Total	0	0	0	100	100	100	100	100	0	0	0	500	-
Accuracy	-	-	-	0.97	0.93	0.68	0.41	0.1	-	-	-	-	**0.62**

Table 4.4: Confusion matrix of results when post-change class is Wetlands (class 11).

Year	No-Change	1	2	3	4	5	6	7	8	9	10	Total	Precision
No-Change	0	0	0	0	0	0	0	0	0	0	0	0	-
1	0	0	0	0	0	0	0	0	0	0	0	0	-
2	0	0	0	12	4	1	2	0	0	0	0	19	0
3	0	0	0	88	7	3	1	0	0	0	0	99	0.89
4	0	0	0	0	89	6	5	2	0	0	0	102	0.87
5	0	0	0	0	0	90	10	3	0	0	0	103	0.87
6	0	0	0	0	0	0	82	14	0	0	0	96	0.85
7	0	0	0	0	0	0	0	81	0	0	0	81	1
8	0	0	0	0	0	0	0	0	0	0	0	0	-
9	0	0	0	0	0	0	0	0	0	0	0	0	-
10	0	0	0	0	0	0	0	0	0	0	0	0	-
Total	0	0	0	100	100	100	100	100	0	0	0	500	-
Accuracy	-	-	-	0.88	0.89	0.9	0.82	0.81	-	-	-	-	**0.86**

Table 4.5: Estimated post-change class of simulated no-change pixels.

Class	5	8	14
Count	2	95	982

4.5 Change Point Case Study

We applied the model described in (4.5) to an area ($\sim$134 km^2) in the Xingu River Basin, located in the Southeastern part of the Amazon in the State of Mato Grosso, Brazil. The study region has several distinct types of natural vegetation including moist tropical rainforest, cerrado, and deciduous forest. Despite containing substantial area of protected indigenous lands, large areas of the basin's evergreen broadleaf forests (EBF) have been converted to agricultural lands for soybean production and cattle ranching since 2000.

We characterize the regional land cover classes using a set of training sites in South America located in the Olson 'Tropical and Subtropical Moist Broadleaf Forests' biome between 0 and 20°S [31]. Two spectral bands (band 1 (620-670 nm) and band 7 (2105-2155 nm)), and a temporal subset of 19 observations per year were selected for our analysis in order to exclude the wet season.

To assess our results, we use the PRODES (Monitoring the Brazilian Amazon Gross Deforestation) dataset produced by Brazil's National Institute for Space Research (INPE) [54]. This annual product provides polygon-based maps of deforestation delineated from Landsat data using a combination of methods including linear spectral unmixing, image segmentation, unsupervised classification, and manual interpretation [112]. We derived annual sub-pixel fractions of deforestation for each MODIS pixel by resampling the 30 m data to 500 m spatial resolution corresponding to MODIS pixels. From the fractional deforestation information, the *reference* year of change for each pixel can be determined by applying a threshold.

Applying our proposed class-to-class EM algorithm to a 25 pixel $\times$ 25 pixel region did take some time, but the results clearly demonstrate the quality of our method. Figures 4.1 and 4.2 give side-by-side comparisons of our predictions with distance metric-based results from [51] as well as references determined, as described above, using 20% and 60% change thresholds. Physically, the threshold represents the proportion of the pixel that needs to have changed for the entire pixel to be deemed a changed pixel. The year a pixel reaches

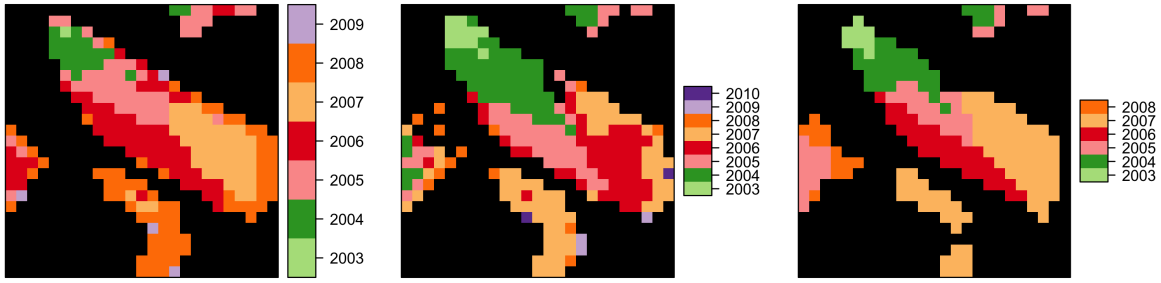


Figure 4.1: Change point reference map for Xingu basin with change threshold of 20% (right), distance metric-based predictions (middle), and our predictions (left).

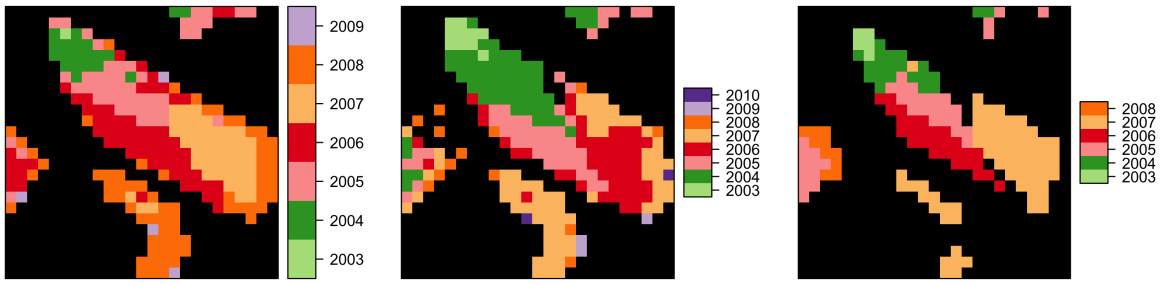


Figure 4.2: Change point reference map for Xingu basin with change threshold of 20% (right), distance metric-based predictions (middle), and our predictions (left).

or surpasses this threshold becomes the change point year called for that pixel.

Our predictions in Figures 4.1 and 4.2 match the references pretty well. We seem to bare closer resemblance to the reference based on a 20% threshold, indicating that our method can pick up on small changes. Table 4.6 summarizes our spatial accuracy at detecting changes. By *spatial accuracy* we just mean accuracy at calling a change or not; and not necessarily calling the correct change year. Very high values in this table further bolster the success of our EM procedure. Not only do our results bare close resemblance to the reference maps, but we seem to display a better match than the distance metric-based results in the middle pane. Indeed, the distance metric-based spatial accuracies are almost

Table 4.6: Our spatial accuracies for change point case study.

Threshold %	Accuracy	Precision
$> 0\%$	0.87	0.93
$> 20\%$	0.96	0.81
$> 40\%$	0.97	0.74
$> 60\%$	0.97	0.64
$> 80\%$	0.99	0.54

Table 4.7: Distance metric-based spatial accuracies for change point case study.

Threshold %	Accuracy	Precision
$> 0\%$	0.88	0.93
$> 20\%$	0.96	0.80
$> 40\%$	0.97	0.73
$> 60\%$	0.98	0.64
$> 80\%$	1	0.54

Table 4.8: Confusion matrix for our change detection in case study region with reference created using a 20% change threshold. Columns are reference year of change and rows are predicted year of change.

Year	No Change	2003	2004	2005	2006	2007	2008	2009	Total	Precision
No Change	365	0	0	0	0	0	9	0	374	0.98
2003	0	1	0	0	0	0	0	0	1	1
2004	1	6	13	0	0	0	0	0	20	0.65
2005	3	1	18	24	0	1	0	0	47	.51
2006	14	0	2	12	32	3	0	0	63	.51
2007	0	0	0	2	2	42	0	0	46	0.91
2008	28	0	0	3	1	35	3	0	70	0.04
2009	2	0	0	0	0	1	1	0	4	0
Total	414	8	33	41	35	81	13	0	625	-
Accuracy	0.88	0.13	0.39	0.59	0.91	0.52	0.23	-	-	**0.77**

identical to ours (Table 4.7). A closer look at the maps (Figures 4.1 and 4.2) and the *temporal accuracies* in Tables 4.8 and 4.9 sheds more light on our ability to detect change and slight temporal superiority to the distance metric-based approach.

Tables 4.8 and 4.9 detail the temporal comparison of change detection results with a reference determined using a 20% change threshold. Our accuracies and precisions vary widely, but retain a very nice overall accuracy of 77% while the distance metric-based

Table 4.9: Confusion matrix for distance metric-based change detection in case study region with reference created using a 20% change threshold. Columns are reference year of change and rows are predicted year of change.

Year	No Change	2003	2004	2005	2006	2007	2008	2009	2010	Total	Precision
No-Change	363	0	0	2	0	1	5	0	0	371	0.98
2003	2	7	1	0	0	0	0	0	0	10	0.7
2004	5	1	32	19	1	0	0	0	0	58	0.55
2005	3	0	0	17	27	0	0	0	0	47	0.36
2006	9	0	0	1	7	33	1	0	0	51	0.14
2007	18	0	0	2	0	46	4	0	0	70	0.66
2008	9	0	0	0	0	1	3	0	0	13	0.23
2009	2	0	0	0	0	1	0	0	0	3	0
2010	2	0	0	0	0	0	0	0	0	2	0
Total	413	8	33	41	35	82	13	0	0	625	-
Accuracy	0.88	0.88	0.97	0.41	0.2	0.56	0.23	-	-	-	**0.76**

method comes in around 76%. The plots in Figure 4.3 make clearer the theme, present in the maps and confusion matrix, that *our* temporal errors are almost exclusively limited to being off by only one year (evidenced by the fact that the pairwise-points on these curves add to around 1). Furthermore, the trends of the curves in Figure 4.3 portray our expectations exactly. That is, our temporal accuracy increases as the change threshold for the reference increases.

An even closer look at a couple of sample pixels from the case study area confirms that temporal accuracy may misrepresent the quality of the method and highlights the low severity of many of these types of errors. For example, in Figure 4.4 the change seems to perhaps have occurred at some point in the middle of 2003. We identify 2004 as the year of the change while the reference labels 2003. This may be an issue of temporal resolution, in only identifying changes at the yearly scale, or perhaps a semantic issue. By *semantic* we mean that our procedure models pre-change data for $i < c_v$ and post-change data for $i \geq c_v$, when perhaps choosing to model pre-change data for $i \leq c_v$ and post-change data for $i > c_v$ would have resulted in a "correct" change point identification.

The second example in Figure 4.5 seems to describe a pixel experiencing a gradual change as opposed to an abrupt change. While the reference calls 2004 at what appears to be the beginning of the change, we call 2006 when the change is close to completion.

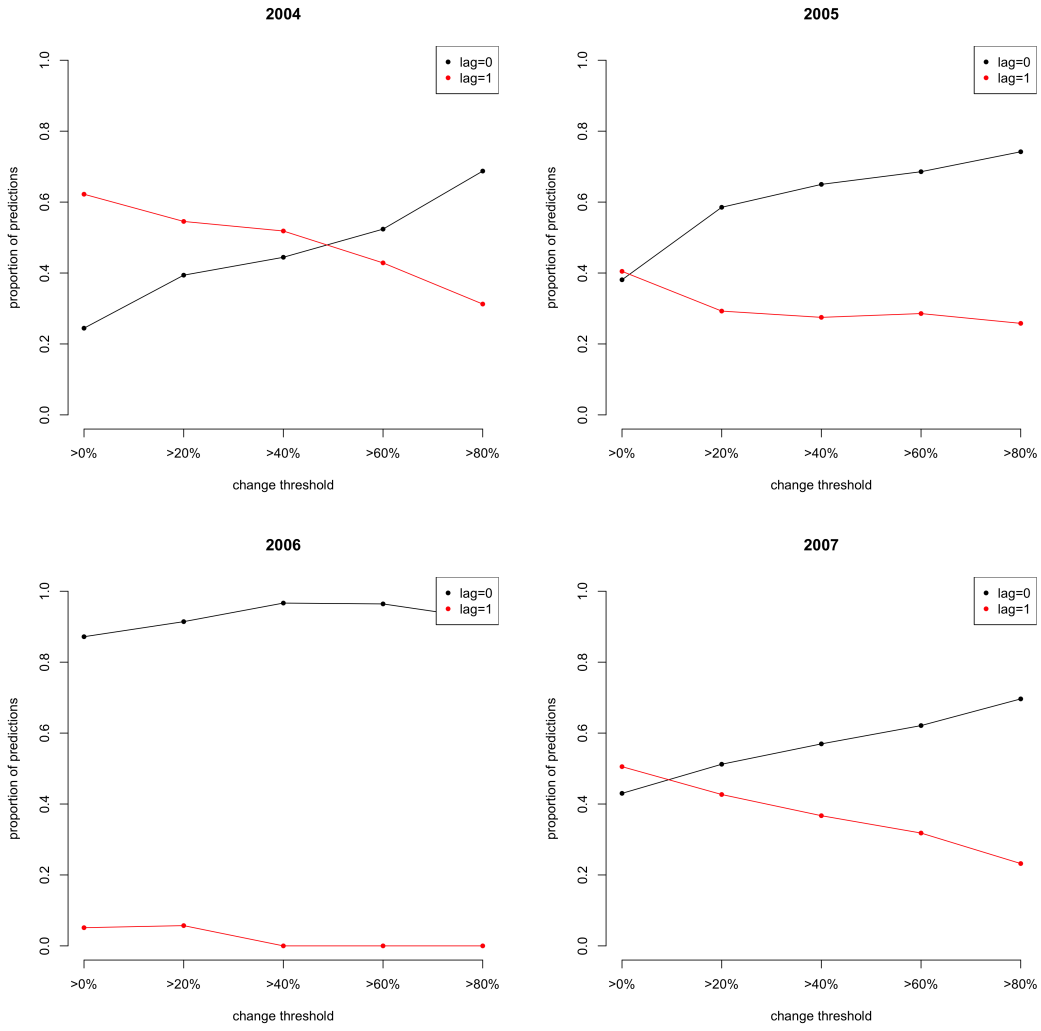


Figure 4.3: These plots display the proportion of pixels in which we identify the correct change year (black) and the proportion of pixels for which our called year is off by 1 (red).

By comparing different land cover classes to class 2, we seek only a distributional change from class 2. When taking the variances into account, it may be difficult to judge a pixel as changed until the pixel's departure from class 2 reaches a certain magnitude.

In any case, the results in Figures 4.4 and 4.5 are more than reasonable and provide evidence that our EM procedure reliably identifies changed pixels and pixels which have not changed. In fact we even prefer our prediction in Figures 4.4 and 4.5 to the reference value,

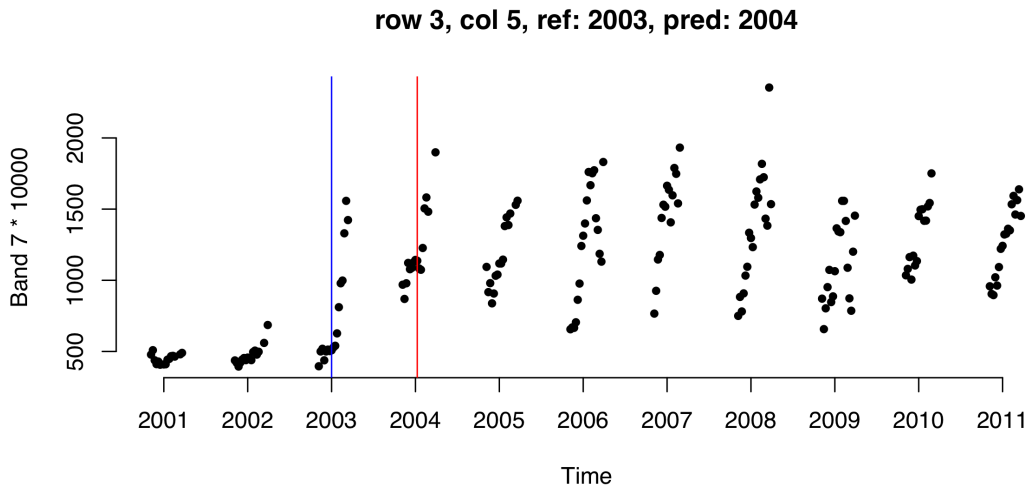


Figure 4.4: Sample time series for pixel 3 in case study dataset. Our prediction is in red and the reference is in blue.

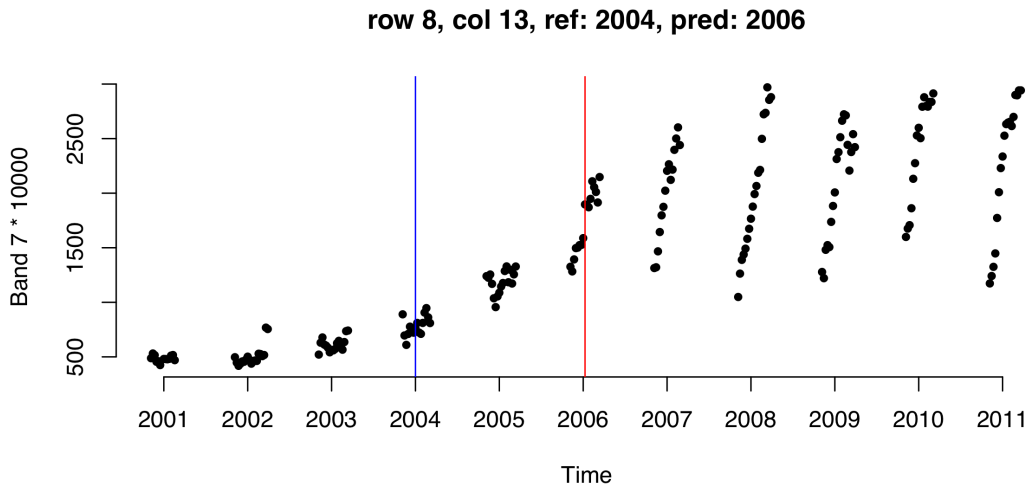


Figure 4.5: Sample time series for pixel 5 in case study dataset. Our prediction is in red and the reference is in blue.

establishing our method as an excellent way to identify change. Though the results may not be compelling enough to call our method more accurate than the distance metric-based approach, our method certainly provides greater interpretability by modeling conversion-type changes and actually allowing for an informal estimate of the post-change land cover.

Finally, our method consists of a principled parameter estimation procedure and robust way to detect change in the presence of missing data.

Chapter 5

Conclusion

In this work we proposed and applied novel approaches to the Remote Sensing tasks of land cover classification and change detection. Most historical approaches to land cover classification assume the data are independent and fail to incorporate valuable spatial information. We began by specifying a matrix normal likelihood for the data, conditional on the land cover class, which partitions the variance into spectral and temporal components. Despite being a parametric assumption, the matrix normal distribution naturally fits the spectral-temporal structure of the data and allows us to vary only particular parameters with land cover class. That is, our matrix normal model involves class-specific means and temporal covariance matrices since important seasonal characteristics vary across land cover classes and warrant retention. A single spectral covariance matrix captures the relationships between spectral bands which we assume does not depend on land cover class.

We develop a new expectation-maximization procedure to estimate the parameters of our matrix normal model in the presence of missing data. Using regionalized training data from the (global) STEP database we obtain estimates which accurately describe the land cover classes in particular areas. We derive another EM algorithm for imputing missing data using these parameter estimates. Despite taking considerable time, the imputation of missing data constitutes an important step in the data processing and classification process. Our formal imputation assumes the pixel labels are unknown and fills in missing values using information from both the class means and covariances.

Once the data have been imputed we convert the spectral-temporal observations into

temporal profiles of three principal component scores. Previous work established the presence of correlation between the seven spectral land bands of MODIS. So, with an eye towards reducing the dimensionality of the data, we target our estimate of the spectral covariance with principal components analysis. This PCA recovered the well-established Tasseled Cap Transformation for MODIS, yielding physically explainable principal components ('Brightness', 'Greenness', and 'Wetness'). With over 90% of the spectral variation explained by the first three principal components we move into a framework which utilizes 3×28 matrix observations (instead of the old 7×28) that contain the most useful spectral-temporal information present in the data.

We developed a collection of tools to assess the quality of training data sites which led to the identification of potentially problematic or confused sites. With the missing values imputed and the dimension reduced we proceeded to replicate an independent pixel classification using a simple, pixel-wise maximum *a posteriori* estimator. In the presence of mixed pixels and low spatial resolution (500 m^2) these simple maps actually achieve satisfying results. However, the persistent salt-and-pepper look, ambiguity of some class definitions, and the presence of within-class variability across different climates required further attention along our initial research trajectory. To account for multi-modality in the data we successfully utilized a hierarchical clustering of climate data to create a new set of refined land cover classes. In pursuit of a Bayesian framework which exploits spatial relationships in its classification procedure, we specified a hierarchical graphical model on the lattice of pixels in an image. After approximating the prior distribution on the lattice in a computationally efficient way, we used the marginal posterior probabilities at each pixel and the centroid estimator to create more spatially homogeneous land cover maps.

We followed our foray into land cover classification with a novel change detection technique that models a change from a pre-defined land cover class to another pre-defined land cover class. That is, we specify a model which assumes that the pre-change data at a given pixel come from a particular IGBP land cover class, and that the post-change data at that pixel come from a different IGBP land cover class. On the heels of our success with land

Table 5.1: New land cover classification scheme based only on land *cover* and excluding land *use*.

CLASS	CLASS NAME
1	Evergreen Needleleaf Forests
2	Evergreen Broadleaf Forests
3	Deciduous Needleleaf Forests
4	Deciduous Broadleaf Forests
5	Mixed Forests
6	Woody Savannah
7	Savannah
8	Dense Shrub
9	Herbaceous Shrub Mix
10	Dense Herbaceous
11	Sparse Shrub
12	Sparse Herbaceous
13	Barren
14	Snow and Ice
15	Water Bodies

cover classification, we use the same matrix normal likelihood and estimate the parameters using the same EM procedure. Because our change point data also contain missing data we successfully identified change points (or lack thereof) using another EM algorithm.

Major advantages of our work include increased interpretability, principled accommodation of missing data, and thorough incorporation of spatial information into our classification method. The avenues of future work extend in many directions. For example, the land cover class definitions according to the International Geosphere-Biosphere Programme (found in Table 1.2) contain some ambiguity and noticeable correlation with each other. The definitions themselves mirror the confusions and points of contention revealed by our analysis in Sections 2.7.4 and 3.1.1. The "Urban and Built-Up," "Cropland," and "Cropland/Natural Vegetation Mosaics" classes describe land *uses* as opposed land *covers*. Consequently, there can exist unnecessarily large amounts of variation in these classes. The "Permanent Wetlands" class is another mixture class which presents almost unavoidable difficulty. Future work may involve a custom land cover classification scheme (LCCS) based on the foundation in [24] which resolves these issues. We propose one such scheme in Table 5. Land cover change detection should, in the future, also incorporate spatial information. Further work will be done into the calibration of β and η as outlined in Section 3.4.5. Finally, adaptation of the centroid estimator to the use of a well constructed

gain matrix should improve the classification results by exploiting the similarities between many of the land cover classes.

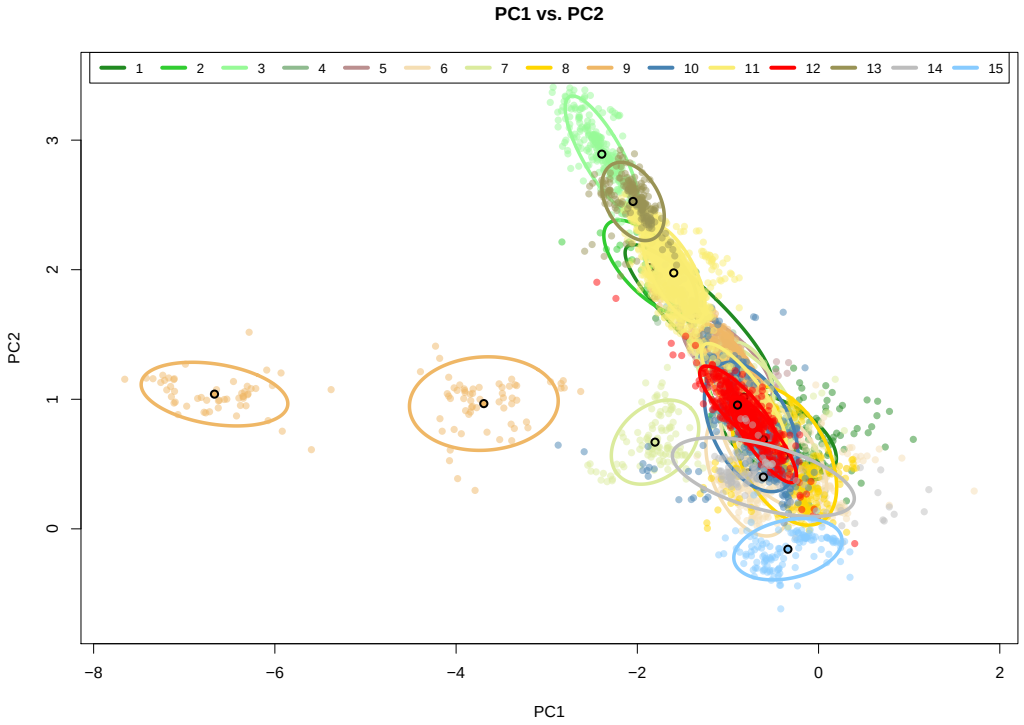


Figure A.1: First two principal components for North America training data after clustering using climate metrics.

Appendix A

Climate Clustering Plots

A.1 Principal Components Plots

Appendix B

North America Maps

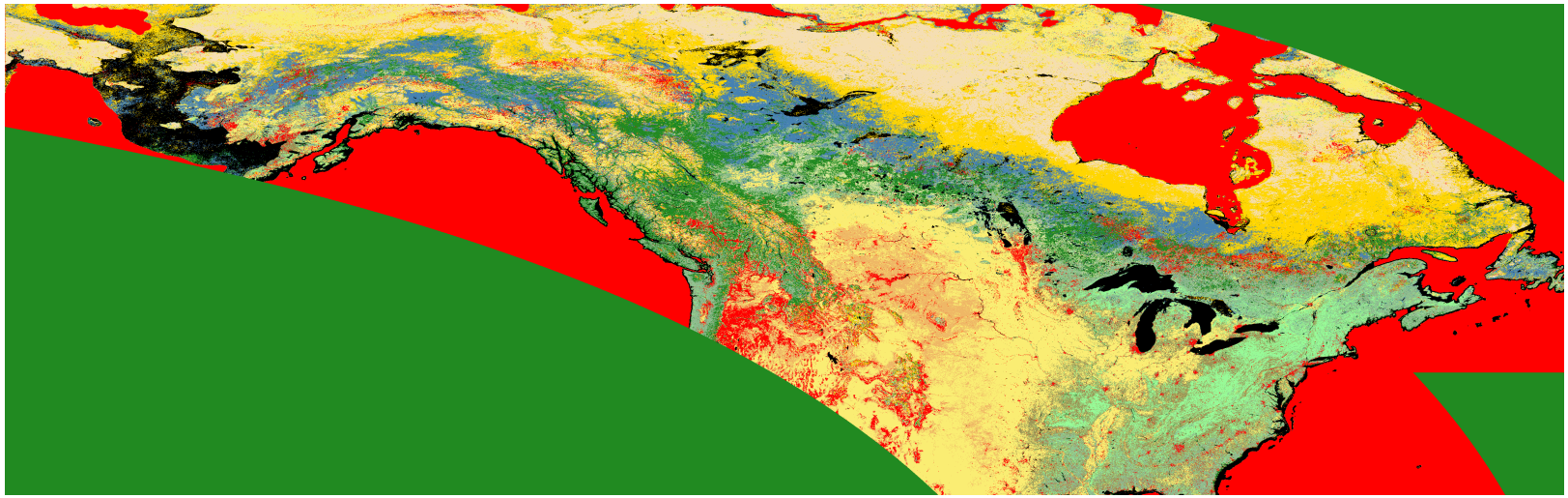


Figure B.1: Map of North America using independent pixel classification and northeast training data.

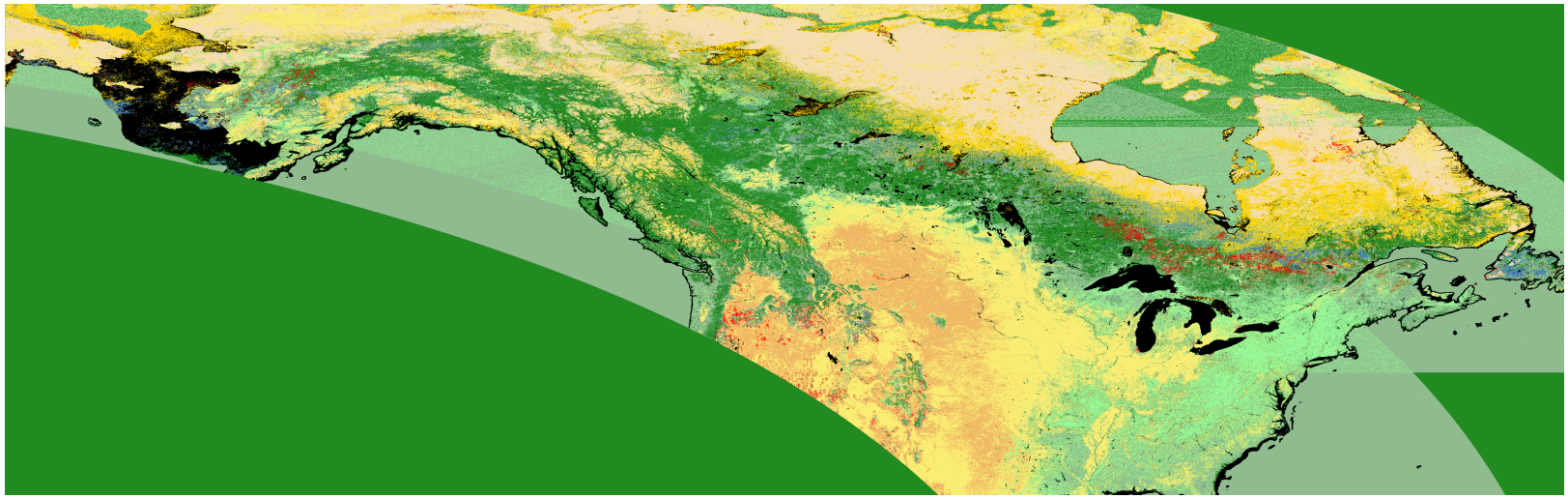


Figure B.2: Map of North America using hierarchical graphical model and northeast training data.

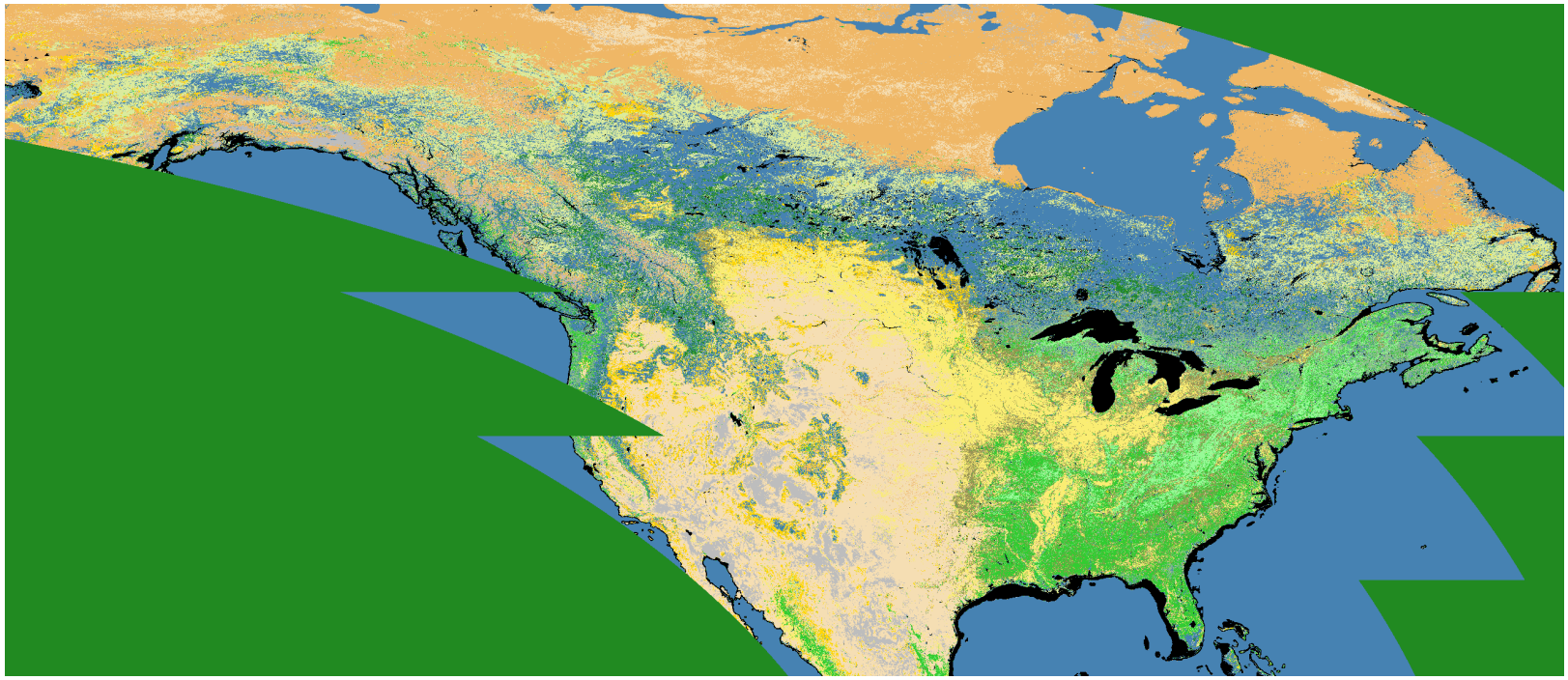


Figure B.3: Map of North America using independent pixel classification, North America training data without the urban class.

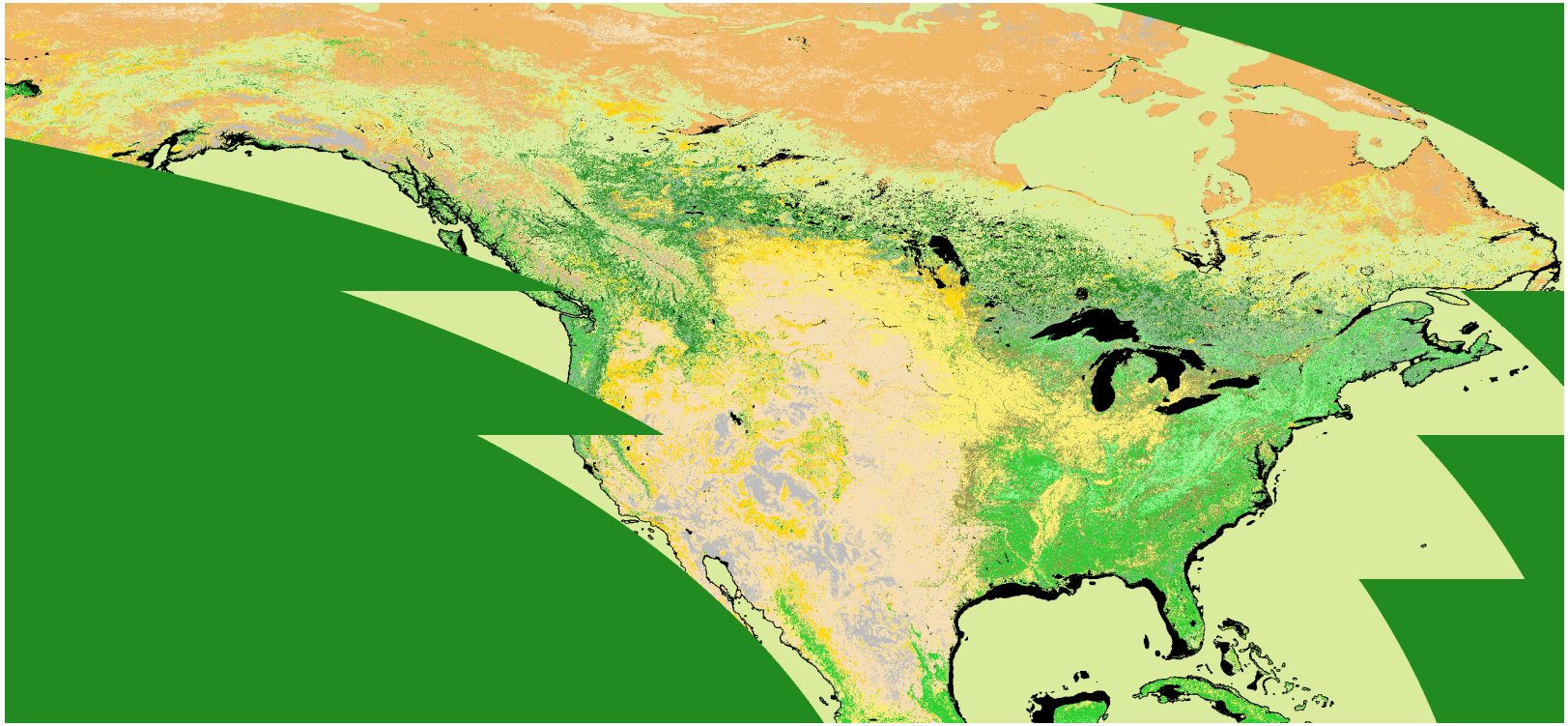


Figure B.4: Map of North America using independent pixel classification, North America training data without urban and wetlands classes

Appendix C

Change Point Derivations

C.1 Univariate Parameter Estimation via EM

So, for the E-step:

$$Q(\Theta, \Theta^{(t)}) = \mathbb{E}_{Z \mid Y;\Theta^{(t)}}[l(Y, Z, \Theta)] = \mathbb{E}_{Z \mid \Theta^{(t)}}[l(Y, Z, \Theta)]$$

We can drop the Y because our yearly observations are assumed to be independent.

$$Q(\Theta, \Theta^{(t)}) = \sum_{v} \Bigg[\sum_{v_{i_Y} < c_v} \log P(Y_{v_{i_Y}}; \Theta) + \sum_{v_{i_Y} \geq c_v} \log P(Y_{v_{i_Y}}; \Theta) + \sum_{v_{i_Z} < c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[\log P(Z_{v_{i_Z}}; \Theta)] + \sum_{v_{i_Z} \geq c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[\log P(Z_{v_{i_Z}}; \Theta)] \Bigg] \tag{C.1}$$

Because we only want to estimate μ_{cv} and σ^2_{cv} we can focus our attention on two of the four terms in (C.1):

$$\begin{aligned} Q(\Theta, \Theta^{(t)}) &= \textstyle\sum_v K + \sum_{v_{i_Y} \geq c_v} \log P(Y_{v_{i_Y}}; \Theta) + \sum_{v_{i_Z} \geq c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[\log P(Z_{v_{i_Z}}; \Theta)] \\ &= \textstyle\sum_v K^* - \frac{n-c+1}{2} \log(\sigma^2_{cv}) - \frac{1}{2\sigma^2_{cv}} \sum_{v_{i_Y} \geq c_v} (Y_{v_{i_Y}} - \mu_{cv})^2 \\ &\quad - \textstyle\frac{1}{2\sigma^2_{cv}} \sum_{v_{i_Z} \geq c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[(Z_{v_{i_Z}} - \mu_{cv})^2] \end{aligned}$$

Let $\tilde{\mu}^{(t)}_{v_{i_Z}} = \mathbb{E}_{Z \mid \Theta^{(t)}}[Z_{v_{i_Z}}]$. Then,

$$
\begin{aligned}
Q(\Theta, \Theta^{(t)}) &= \textstyle\sum_v K^* - \frac{n-c+1}{2}\log(\sigma^2_{cv}) - \frac{1}{2\sigma^2_{cv}}\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu_{cv})^2 \\
&\quad -\frac{1}{2\sigma^2_{cv}}\textstyle\sum_{v_{i_Z} \geq c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[(Z_{v_{i_Z}} - \mu_{cv})^2] \\
&= \textstyle\sum_v K^* - \frac{n-c+1}{2}\log(\sigma^2_{cv}) - \frac{1}{2\sigma^2_{cv}}\sum_{v_{i_Y} \geq c}(Y_{v_{i_Y}} - \mu_{cv})^2 - \\
&\quad \frac{1}{2\sigma^2_{cv}}\textstyle\sum_{v_{i_Z} \geq c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[((Z_{v_{i_Z}} - \tilde{\mu}^{(t)}_{v_{i_Z}})(\tilde{\mu}^{(t)}_{i_Z} - \mu_{cv}))^2] \\
&= \textstyle\sum_v K^* - \frac{n-c+1}{2}\log(\sigma^2_{cv}) - \frac{1}{2\sigma^2_{cv}}\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu_{cv})^2 - \\
&\quad \frac{1}{2\sigma^2_{cv}}\textstyle\sum_{v_{i_Z} \geq c_v} \mathbb{E}_{Z \mid \Theta^{(t)}}[(Z_{v_{i_Z}} - \tilde{\mu}^{(t)}_{v_{i_Z}})^2] + (\tilde{\mu}^{(t)}_{v_{i_Z}} - \mu_{cv})^2 \\
&= \textstyle\sum_v K^* - \frac{n-c+1}{2}\log(\sigma^2_{cv}) - \frac{1}{2\sigma^2_{cv}}\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu_{cv})^2 \\
&\quad -\frac{1}{2\sigma^2_{cv}}\textstyle\sum_{v_{i_Z} \geq c_v} \sigma^{2^{(t)}}_{cv} + (\tilde{\mu}^{(t)}_{v_{i_Z}} - \mu_{cv})^2
\end{aligned}
$$

Since the observed and unobserved data are independent we have $\tilde{\mu}^{(t)}_{v_{i_Z}} = \mu^{(t)}_{cv}$.

$$
\begin{aligned}
Q(\Theta, \Theta^{(t)}) = \sum_v K^* - \frac{n-c+1}{2}\log(\sigma^2_{cv}) \\
- \frac{1}{2\sigma^2_{cv}}\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu_{cv})^2 - \frac{1}{2\sigma^2_{cv}}\sum_{v_{i_Z} \geq c_v} \sigma^{2^{(t)}}_{cv} + (\mu^{(t)}_{cv} - \mu_{cv})^2
\end{aligned}
$$

Now for the M-step we will maximize $Q(\Theta, \Theta^{(t)})$ as a function of μ_{cv} and σ^2_{cv}:

$$
\begin{aligned}
\tfrac{\partial Q}{\partial \mu_{cv}} &= \tfrac{1}{\sigma^2_{cv}}\textstyle\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu_{cv}) + \tfrac{1}{\sigma^2_{cv}}\sum_{v_{i_Z} \geq c_v}(\mu^{(t)}_{cv} - \mu_{cv}) = 0 \\
&\implies \mu^{(t+1)}_{cv} = \tfrac{1}{n-c+1}\left[\textstyle\sum_{v_{i_Y} \geq c_v} Y_{v_{i_Y}} + \sum_{v_{i_Z} \geq c_v} \mu^{(t)}_{cv}\right]
\end{aligned}
$$

$$
\begin{aligned}
\tfrac{\partial Q}{\partial \sigma^2_{cv}} &= -\tfrac{n-c+1}{2\sigma^2_{cv}} + \tfrac{1}{2\sigma^4_{cv}}\textstyle\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu_{cv})^2 + \tfrac{1}{2\sigma^4_{cv}}\sum_{v_{i_Z} \geq c_v} \sigma^{2^{(t)}}_{cv} + (\mu^{(t)}_{cv} - \mu_{cv})^2 = 0 \\
&\implies \sigma^{(t+1)}_{cv} = \tfrac{1}{n-c+1}\left[\textstyle\sum_{v_{i_Y} \geq c_v}(Y_{v_{i_Y}} - \mu^{(t+1)}_{cv})^2 + \sum_{v_{i_Z} \geq c_v} \sigma^{2^{(t)}}_{cv} + (\mu^{(t)}_{cv} - \mu^{(t+1)}_{cv})^2\right]
\end{aligned}
$$

C.2 Derivation of EM updates in Class-to-Class Framework

$\underline{\alpha}$

$$
\begin{aligned}
Q_\alpha &= \mathbb{E}_{Z,W\,|\,Y,\Theta^{(t)}}[\log[\textstyle\prod_v \prod_{g=1}^{\mathscr{C}}(\mathbb{P}(\mu_c\,|\,W_v=g)\mathbb{P}(W_v=g\,|\,\boldsymbol{\alpha}))^{I(W_v=g)}\mathbb{P}(\boldsymbol{\alpha})]] \\
&= \mathbb{E}_{Z,W\,|\,Y,\Theta^{(t)}}[\textstyle\sum_v \sum_{g=1}^{\mathscr{C}} I(W_v=g)[\log\mathbb{P}(\mu_c\,|\,W_v=g) + \log\underbrace{\mathbb{P}(W_v=g\,|\,\boldsymbol{\alpha})}_{\alpha_g}] + \log\mathbb{P}(\boldsymbol{\alpha})] \\
&\simeq \mathbb{E}_{Z,W\,|\,Y,\Theta^{(t)}}[\textstyle\sum_v \sum_{g=1}^{\mathscr{C}} I(W_v=g)\log\alpha_g + \log(\frac{1}{B(\boldsymbol{\pi})}\prod_i \alpha_i^{\pi_i-1})] \\
&= \textstyle\sum_v \sum_{g=1}^{\mathscr{C}} \mathbb{E}_{Z,W\,|\,Y,\Theta^{(t)}}[I(W_v=g)]\log\alpha_g + (\pi_g-1)\log\alpha_g \\
\frac{\partial Q_\alpha}{\partial\alpha_k} &= \tfrac{\sum_v \mathbb{P}(W_v=k\,|\,Y_v,\Theta^{(t)})+\pi_k-1}{\alpha_k} - \tfrac{\sum_v P(W_v=\mathscr{C}\,|\,Y_v,\Theta^{(t)})+\pi_{\mathscr{C}}-1}{1-\sum_{g=1}^{\mathscr{C}-1}\alpha_g}
\end{aligned}
$$

$$
\implies \quad \alpha_k^{(t+1)} = \frac{\sum_v \mathbb{P}(W_v=k\,|\,Y_v,\Theta^{(t)})+\pi_k-1}{\sum_{g=1}^{\mathscr{C}}\sum_v \mathbb{P}(W_v=g\,|\,Y_v,\Theta^{(t)})+\pi_g-1} = \frac{\sum_v \mathbb{P}(W_v=k\,|\,Y_v,\Theta^{(t)})+\pi_k-1}{N-\mathscr{C}+\sum_g\pi_g}
$$

with the following

$$
\mathbb{P}(W_v=g\,|\,Y_v,\Theta^{(t)}) = \frac{\alpha_g^{(t)}\mathbb{P}(Y_v\,|\,W_v=g,\Theta^{(t)})}{\sum_{\tilde{g}}\alpha_{\tilde{g}}^{(t)}\mathbb{P}(Y_v\,|\,W_v=\tilde{g},\Theta^{(t)})}
$$

$\underline{\mu_{0v}}$

$$
\begin{aligned}
Q_{\mu_{0v}} &= \mathbb{E}_{Z,W \mid Y,\Theta^{(t)}} \left[\log[\textstyle\prod_v \prod_{i<c_v} \mathbb{P}(X_{iv} \mid \mu_{0v}) \mathbb{P}(\mu_{0v})] \right] \\
&= \mathbb{E}_{Z,W \mid Y,\Theta^{(t)}} \left[\textstyle\sum_v \sum_{i<c_v} \log \mathbb{P}(X_{iv} \mid \mu_{0v}) + \log \mathbb{P}(\mu_{0v}) \right] \\
&= \mathbb{E}_{Z,W \mid Y,\Theta^{(t)}} \left[\textstyle\sum_v \sum_{i<c_v} (X_{iv} - \mu_{0v})^\top (\frac{1}{\kappa_0}\mathbf{I})(X_{iv} - \mu_{0v}) \right. \\
&\quad \left. + (\mu_{0v} - \mu_F)^\top \Sigma_F^{-1} (\mu_{0v} - \mu_F) \right] \\
\frac{\partial Q_{\mu_{0v}}}{\partial \mu_{0v}} &= \tfrac{-2}{\kappa_0} \textstyle\sum_{i<c_v} \mathbb{E}_{Z \mid Y,\Theta^{(t)}}[(X_{iv} - \mu_{0v})] + 2\Sigma_F^{-1}(\mu_{0v} - \mu_F) \\
\implies \quad \mu_{0v}^{(t+1)} &= \left(\frac{(c_v - 1)}{\kappa_0} \mathbf{I} + \Sigma_F^{-1} \right)^{-1} \left(\frac{1}{\kappa_0} \sum_{i<c_v} \mathbb{E}_{Z \mid Y,\Theta^{(t)}}[X_{iv}] + \Sigma_F^{-1} \mu_F \right)
\end{aligned}
$$

$\underline{\mu_{cv}}$

$$
\begin{aligned}
Q_{\mu_{cv}} &= \mathbb{E}_{Z,W \mid Y,\Theta^{(t)}} \left[\log[\textstyle\prod_v (\prod_{i \geq c_v} \mathbb{P}(X_{iv} \mid \mu_{cv}) \right. \\
&\quad \left. \textstyle\prod_{g=1}^{\mathscr{C}} (\mathbb{P}(\mu_{cv} \mid W_v = g) \mathbb{P}(W_v = g \mid \boldsymbol{\alpha}))^{I(W_v = g)})] \right] \\
&= \mathbb{E}_{Z,W \mid Y,\Theta^{(t)}} [\textstyle\sum_v (\sum_{i \geq c_v} \log \mathbb{P}(X_{iv} \mid \mu_{cv}) \\
&\quad + \textstyle\sum_{g=1}^{\mathscr{C}} I(W_v = g)[\log \mathbb{P}(\mu_{cv} \mid W_v = g) + \log \alpha_g)]] \\
\frac{\partial Q_{\mu_c}}{\partial \mu_{cv}} &= \tfrac{-2}{\kappa_c} \textstyle\sum_{i \geq c_v} \mathbb{E}_{Z \mid Y,\Theta^{(t)}}[X_{iv} - \mu_{cv}] + 2 \textstyle\sum_{g=1}^{\mathscr{C}} \mathbb{E}_{Z,W \mid Y,\Theta^{(t)}}[I(W_v = g)] \Sigma_g^{-1} (\mu_{cv} - \mu_g)
\end{aligned}
$$

$$\implies \quad \mu_{cv}^{(t+1)} = \left(\frac{(n - c_v + 1)}{\kappa_c} \mathbf{I} + \sum_{g=1}^{\mathscr{C}} \mathbb{P}(W_v = g \,|\, Y_v, \Theta^{(t)}) \Sigma_g^{-1} \right)^{-1} \cdot \left(\frac{1}{\kappa_c} \sum_{i \geq c_v} \mathbb{E}_{Z \,|\, Y, \Theta^{(t)}}[X_{iv}] + \sum_{g=1}^{\mathscr{C}} \mathbb{P}(W_v = g \,|\, Y_v, \Theta^{(t)}) \Sigma_g^{-1} \mu_g \right)$$

again with

$$\mathbb{P}(W_v = g \,|\, Y_v, \Theta^{(t)}) = \frac{\alpha_g^{(t)} \mathbb{P}(Y_v \,|\, W_v = g, \Theta^{(t)})}{\sum_{\tilde{g}} \alpha_{\tilde{g}}^{(t)} \mathbb{P}(Y_v \,|\, W_v = \tilde{g}, \Theta^{(t)})}$$

The update for c_v, $c_v^{(t+1)}$, is computed by going through each possible change point year and computing Q with the updates from above. The value of c_v that maximizes Q (and makes Q greater than the previous iteration's max Q) is what c_v should be updated to.

Bibliography

[1] Modis overview.

[2] Abdelgadir A Abuelgasim, Sucharita Gopal, James R Irons, and Alan H Strahler. Classification of asas multiangle and multispectral measurements using artificial neural networks. *Remote Sensing of Environment*, 57(2):79–87, 1996.

[3] Genevera I Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, 2010.

[4] Michèle Basseville and Igor V Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., 1993.

[5] JONA Benediktsson, Philip H Swain, and Okan K Ersoy. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on geoscience and remote sensing*, 28(4):540–552, 1990.

[6] Julien Besag. On the statistical analysis of dirty pictures. *J. R. Stat. Soc.*, 48(3):259–302, 1986. with discussions.

[7] J. B. Boik. Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. *Communications in Statistics - Theory and Methods*, 20:1233–1255, 1991.

[8] Gordon B Bonan, Keith W Oleson, Mariana Vertenstein, Samuel Levis, Xubin Zeng, Yongjiu Dai, Robert E Dickinson, and Zong-Liang Yang. The Land Surface Climatology of the Community Land Model Coupled to the NCAR Community Climate Model*. *Journal of Climate*, 15(22):3123–3149, 2002.

[9] N B Booth and A F M Smith. A Bayesian approach to retrospective identification of change-points. *Journal of Econometrics*, 19(1):7–22, 1982.

[10] Shyam Boriah. *Time series change detection: Algorithms for land cover change*. 2010.

[11] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[12] Luis E Carvalho and Charles E Lawrence. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences*, 105(9):3209–3214, 2008.

[13] N. Rao Chaganty and Dayanand N. Naik. Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference*, 103:421–436, 2002.

[14] C-CT Chen and David A Landgrebe. A spectral feature design system for the hiris/modis era. *Geoscience and Remote Sensing, IEEE Transactions on*, 27(6):681–686, 1989.

[15] C Chow and C Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

[16] Pao-Shin Chu and Xin Zhao. Bayesian Change-Point Analysis of Tropical Cyclone Activity: The Central North Pacific Case. *Journal of Climate*, 17(24):4893–4901, 2004.

[17] Eric P Crist and Richard C Cicone. A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing*, (3):256–263, 1984.

[18] Robert Davis and Peter Holmgren. Fra 2000: Forest cover mapping & monitoring with noaa-avhrr & other coarse spatial resolution sensors. *Forest Resources Assessment Programme*, 2000.

[19] A.P. Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

[20] S De Bruin and BGH Gorte. Probabilistic image classification using geological map units applied to land-cover change detection. *International Journal of Remote Sensing*, 21(12):2389–2402, 2000.

[21] RS DeFries and JRG Townshend. NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, 15(17):3567–3586, 1994.

[22] Ruth DeFries, Matthew Hansen, and John Townshend. Global discrimination of land cover types from metrics derived from AVHRR Pathfinder data. *Remote Sensing of Environment*, 54(3):209–222, 1995.

[23] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[24] A. DiGregorio and L.J.M. Jansen. *Land Cover Classification System: Classification Concepts and User Manual*. Food and Agriculture Organization of the United Nations, 2005.

[25] Pierre Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

[26] J Ronald Eastman and M Filk. Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote Sensing*, 59(6):991–996, 1993.

[27] MB Ek, KE Mitchell, Y Lin, E Rogers, P Grunmann, V Koren, G Gayno, and JD Tarpley. Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research*, 108(D22):8851, 2003.

[28] James B Elsner, Thomas Jagger, and Xufeng Niu. Changes in the rates of North Atlantic major hurricane activity during the 20th century. *Geophysical Research Letters*, 27(12):1743–1746, 2000.

[29] Paul Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing*, 2006.

[30] Paul Fearnhead and Z Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

[31] M. A. Friedl et al. Modis collection 5 global land cover: Algorithm refinements characterization of new datasets. *Remote Sensing of Environment*, 114:168–182, 2009.

[32] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang. MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114:168–182, 2010.

[33] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

[34] Mark A Friedl, Carla E Brodley, and Alan H Strahler. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(2):969–977, 1999.

[35] Mark A Friedl, Douglas K McIver, John CF Hodges, XY Zhang, D Muchoney, Alan H Strahler, Curtis E Woodcock, Sucharita Gopal, Annemarie Schneider, Amanda Cooper, et al. Global land cover mapping from modis: algorithms and early results. *Remote Sensing of Environment*, 83(1):287–302, 2002.

[36] Montserrat Fuentes. Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference*, 136:447–466, 2004.

[37] Pedro Galeano and Daniel Peña. Covariance changes detection in multivariate time series. *Journal of Statistical Planning and Inference*, 137(1):194–211, January 2007.

[38] Andrzej T. Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics - Theory and Methods*, 23(11):3105–3119, 1994.

[39] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[40] Edit Gombay. Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99(3):451–464, March 2008.

[41] James H Goodnight. A tutorial on the SWEEP operator. *The American Statistician*, 33(3):149–158, 1979.

[42] P Goovaerts. Comparison of coik, ik and mik performances for modeling conditional probabilities of categorical variables. In *Geostatistics for the next century*, pages 18–29. Springer, 1994.

[43] Pierre Goovaerts. Geostatistical incorporation of spatial coordinates into supervised classification of hyperspectral data. *Journal of Geographical Systems*, 4(1):99–111, 2002.

[44] Sucharita Gopal and Curtis Woodcock. Remote sensing of forest change using artificial neural networks. *Geoscience and Remote Sensing, IEEE Transactions on*, 34(2):398–404, 1996.

[45] Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[46] MC Hansen, RS DeFries, John RG Townshend, and Rob Sohlberg. Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21(6-7):1331–1364, 2000.

[47] MC Hansen, RS Defries, JRG Townshend, and R Sohlberg. Global land cover classification at 1km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21(6-7):1331–1364, 2000.

[48] George F Hepner, Thomas Logan, Niles Ritter, and Nevin Bryant. Artificial neural network classification using a minimal training set-comparison to conventional supervised classification. 1990.

[49] James Honaker and Gary King. What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581, 2010.

[50] Hong Huang. Classification of hyperspectral remote-sensing images based on sparse manifold learning. *Journal of Applied Remote Sensing*, 7, 2013.

[51] Xiaoman Huang and Mark A Friedl. Distance metric-based forest cover change detection using modis time series. *International Journal of Applied Earth Observation and Geoinformation*, 29:78–92, 2014.

[52] Xiaoman Huang and Mark A Friedl. Distance metric-based forest cover change detection using MODIS time series. *International Journal Of Applied Earth Observation And Geoinformation*, 29:78–92, June 2014.

[53] A Huete, Kamel Didan, Tomoaki Miura, E Patricia Rodriguez, Xiang Gao, and Laerte G Ferreira. Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote sensing of environment*, 83(1):195–213, 2002.

[54] INPE. Project prodes: Monitoring the brazilian amazon forests by satellite, 2012.

[55] X Jia and JA Richards. Managing the spectral-spatial mix in context classification using markov random fields. *Geoscience and Remote Sensing Letters, IEEE*, 5(2):311–314, 2008.

[56] Xiuping Jia and John A Richards. Efficient maximum likelihood classification for imaging spectrometer data sets. *Geoscience and Remote Sensing, IEEE Transactions on*, 32(2):274–281, 1994.

[57] Xiuping Jia and John A Richards. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(1):538–542, 1999.

[58] I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2005.

[59] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[60] Chulhee Lee and David A Landgrebe. Fast likelihood classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 29(4):509–517, 1991.

[61] Tze-San Lee. Change-Point Problems: Bibliography and Review. *Journal of Statistical Theory and Practice*, 4(4):643–662, December 2010.

[62] Roderick JA Little and Donald B Rubin. Statistical analysis with. 2002.

[63] SE Lobser and WB Cohen. MODIS tasselled cap: land cover characteristics expressed through transformed MODIS data. *International Journal of Remote Sensing*, 28(22):5079–5101, 2007.

[64] Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *arXiv preprint arXiv:1201.2577*, 2012.

[65] THOMASR LOVELAND, JAMESW MERCHANT, JESSLYNF BROWN, and DONALDO OHLEN. Development of a land-cover characteristics database for the conterminous u. s. *Photogrammetric engineering and remote sensing*, 57(11):1453–1463, 1991.

[66] TR Loveland, BC Reed, JF Brown, DO Ohlen, Z Zhu, L Yang, and JW Merchant. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, 21(6-7):1303–1330, 2000.

[67] D Lu, P Mausel, E Brondizio, and E Moran. Change detection techniques. *International journal of remote sensing*, 25(12):2365–2401, 2004.

[68] Dengsheng Lu, P Mausel, E Brondizio, and E Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2407, 2004.

[69] Robert Lund and Jaxk Reeves. Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate*, 15(17):2547–2554, 2002.

[70] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

[71] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[72] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[73] Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley-Interscience, 2004.

[74] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics, 1997.

[75] Ulrich Menzefricke. A Bayesian analysis of a change in the precision of a sequence of independent normal random variables at an unknown time point. *Applied Statistics*, 30(2):141–146, 1981.

[76] Zheng Mingguo, C Qiangguo, and Qin Mingzhou. The effect of prior probabilities in the maximum likelihood classification on individual classes: A theoretical reasoning and empirical testing. *Photogramm. Eng. Remote Sens*, 75(9):1109–1116, 2009.

[77] Gabriele Moser, Sebastiano B Serpico, and Jon Atli Benediktsson. Land-cover mapping by markov modeling of spatial–contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101(3):631–651, 2013.

[78] Marwan Jamil Muasher and DAVID A Landgrebe. The kl expansion as an effective feature ordering technique for limited training sample size. *Geoscience and Remote Sensing, IEEE Transactions on*, (4):438–441, 1983.

[79] Douglas Muchoney, Alan Strahler, John Hodges, and Janet LoCastro. The igbp discover confidence sites and the system for terrestrial ecosystem parameterization: tools for validating global land-cover data. *Photogrammetric Engineering and Remote Sensing*, 65(9):1061–1067, 1999.

[80] D. N. Naik and S. Rao. Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics*, 28:91–105, 2001.

[81] Mahesh Pal and Paul M Mather. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote sensing of environment*, 86(4):554–565, 2003.

[82] Christos H. Papadimitriou and Ken Steiglitz. *Combinatorial optimization: algorithms and complexity*. Dover, 1998.

[83] No-Wook Park, Wooil M Moon, Kwang-Hoon Chi, and Byung-Doo Kwon. Multi-sensor data fusion for supervised land-cover classification using bayesian and geostatistical techniques. *Geosciences Journal*, 6(3):193–202, 2002.

[84] L Perreault, J Bernier, B Bobée, and E Parent. Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. *Journal of Hydrology*, 235(3-4):221–241, August 2000.

[85] Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook, 2008.

[86] R.B. Potts. Some generalized order-disorder transformations. In *Proceedings of the Cambridge Philosophical Society*, volume 48, pages 106–109. Cambridge Univ Press, 1952.

[87] John Ross Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[88] Abdullah F Rahman, Danilo Dragoni, Kamel Didan, Armando Barreto-Munoz, and Joseph A Hutabarat. Detecting large scale conversion of mangroves to aquaculture with change point and mixed-pixel analyses of high-fidelity MODIS data. *Remote Sensing of Environment*, 130:96–107, 2013.

[89] Navin Ramankutty, Amato T Evan, Chad Monfreda, and Jonathan A Foley. Farming the planet: 1. geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 22(1), 2008.

[90] Craig Rodarmel and Jie Shan. Principal component analysis for hyperspectral image classification. *Surveying and Land Information Science*, 62(2):115–122, 2002.

[91] S N Rodionov. A brief overview of the regime shift detection methods. In V Velikova and N Chipev, editors, *Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability, UNESCO-ROSTE/BAS Workshop on Regime Shifts*, pages 17–24. UNESCO-ROSTE/BAS Workshop on Regime Shifts, 14-16 June 2005, Varna, Bulgaria, 17-24, 2005.

[92] A. Roy and R. Khattree. On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference*, 134(2):462–485, 2005.

[93] Khattree R. Roy A. Testing the hypothesis of a kroneckar product covariance matrix in multivariate repeated measures data. *Statistical Methodology*, 2, 2005.

[94] Steven W Running and Joseph C Coughlan. A general model of forest ecosystem processes for regional applications I. Hydrologic balance, canopy gas exchange and primary production processes. *Ecological Modelling*, 42(2):125–154, 1988.

[95] C. Schaaf et al. First operational brdf, albedo nadir reflectance products from modis. *Remote Sensing of Environment*, 83(1):135–148, 2002.

[96] Xiaofeng Shao and Xianyang Zhang. Testing for Change Points in Time Series. *Journal Of The American Statistical Association*, 105(491):1228–1240, September 2010.

[97] Mahendran Shitan and Peter J. Brockwell. An asymptotic test for separability of a spatial autoregressive model. *Communications in Statistics - Theory and Methods*, 24(8):2027–2040, 1995.

[98] A Singh. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 1989.

[99] Andrew R Solow and Andrew R Beet. A test for a regime shift. *Fisheries Oceanography*, 14(3):236–240, 2005.

[100] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Statistical change detection for multi-dimensional data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 667–676. ACM, 2007.

[101] M.S. Srivastava and C.G. Khatri. *An Introduction to Multivariate Statistics*. North Holland, New York, USA, 1979.

[102] M.S. Srivastava, T. Nahtman, and D. von Rosen. Models with a Kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics*, 17(4):357–370, 2008.

[103] D A Stephens. Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1):159–178, 1994.

[104] Shannon Sterling and Agnès Ducharne. Comprehensive data set of global land cover change for land surface model applications. *Global Biogeochemical Cycles*, 22(3), 2008.

[105] Thomas A Stone, Peter Schlesinger, Richard A Houghton, and George M Woodwell. A map of the vegetation of South America based on satellite imagery. *Photogrammetric Engineering and Remote Sensing*, 60(5):541–551, 1994.

[106] A. H. Strahler. The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of Environment*, 10:1135–1163, 1980.

[107] H. H. Szu and R. L. Hartley. Nonconvex optimization by fast simulated annealing. 75(11):1538–1540, 1987.

[108] J.R.G. Townsend, C.O. Justice, and V.T. Kalb. Characterization and classification of south american land cover types using satellite data. *International Journal of Remote Sensing*, 8:1189–1207, 1987.

[109] Ruey S Tsay. Outliers, level shifts, and variance changes in time series. *Journal of forecasting*, 7(1):1–20, 1988.

[110] Theodoros Tsiligkaridis and A Hero. Covariance estimation in high dimensions via kronecker product expansions. 2013.

[111] Theodoros Tsiligkaridis, Alfred O HERO, and ZHOU SHUHENG. On convergence of kronecker graphical lasso algorithms. *IEEE transactions on signal processing*, 61(5-8):1743–1755, 2013.

[112] D M Valeriano, E M K Mello, J C Moreira, Yosio E Shimabukuro, V Duarte, I M e Souza, J R dos Santos, C C F Barbosa, and R C M de Souza. Monitoring tropical forest from space: the PRODES digital project. *Proceedings of ISPRS '04*, 2004.

[113] N. B. Venkateswarlu and P.S. V. S. K. Raju. Three stage maximum likelihood classifier. *Pattern Recognition*, 24:1113–1116, 1991.

[114] Jan Verbesselt, R Hyndman, G Newnham, and D Culvenor. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1):106–115, 2010.

[115] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[116] Z. M. Wan, Y. L. Zhang, Q. C. Zhang, and Z. L. Li. Validation of the land-surface temperature products retrieved from terra moderate resolution imaging spectroradiometer data. *Remote Sensing of Environment*, 83:163–180, 2002.

[117] Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with kronecker product structure. *Signal Processing, IEEE Transactions on*, 56(2):478–491, 2008.

Curriculum Vitae

Contact Hunter S. Glanz

Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA 02215, USA

Education

California Polytechnic State University, San Luis Obispo, B.S., Statistics (Cum Laude), September 2005 – June 2009.

California Polytechnic State University, San Luis Obispo, B.S., Mathematics (Cum Laude), September 2005 – June 2009.

Boston University, M.A., Mathematics with emphasis in Statistics, September 2009 – May 2012.

Boston University PhD candidate, Mathematics with emphasis in Statistics, September 2009 – present. Thesis advisor: Luis Carvalho.

Publications

1. Glanz, Hunter. "The View From Here: The Top Job in America." *Math Horizons*, April 2009.
2. Glanz, Hunter. "A True Revolution in Statistics." *AMSTAT NEWS*, December 2013.
3. Glanz, Hunter. "Learning For a Living." *STAT trak*, February 2014.
4. Glanz, H., Carvalho, L., Sulla-Menashe, D., and Friedl, M. "A Parsimonious Model for Land Cover Classification and Characterization of Training Data Using Multitemporal Remotely Sensed Imagery." (Submitted).
5. Glanz, H. and Carvalho, L. "An Expectation-Maximization Algorithm for the Matrix Normal Distribution." (Submitted).
6. Glanz, H. and Carvalho, L. "A Spanning Tree Hierarchical Model for Land Cover Classification." (In preparation).