A Spatial-Temporal Contextual Kernel Method for Generating High-Quality Land-Cover Time Series

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of The Ohio State University

By

Adam Wehmann

Graduate Program in Geography

The Ohio State University

2014

Master's Examination Committee:

Dr. Desheng Liu, Advisor

Dr. Ningchuan Xiao

Dr. Brian Kulis
Abstract

In order to understand the variability, drivers, and effects of the currently unprecedented rate, extent, and intensity of land-cover change, land change science requires remote sensing products that are both highly accurate and spatial-temporally consistent. This need for accuracy is exacerbated from the shift in the discipline from the detection of change between two points in time to the analysis of trajectories of change over time. As the length of temporal record increases, the problem becomes more severe. This follows because the accuracy of change detection is bounded below by the product of the accuracies of the source maps. Without exceedingly high accuracy at individual dates, the accuracy of change detection will be low, as map errors simply and vastly outweigh the occurrence of real change. Land-cover classifiers that can better utilize spatial and temporal information offer the chance to increase the accuracy of change detection and the consistency of classification results. By increasing the spatial and temporal dependence of errors between classification maps, the overall area among maps subject to error may be minimized, producing higher quality land-cover products. Such products enable more accurate and consistent detection, monitoring, and quantification of land-cover change and therefore can have wide-reaching impacts on downstream environmental, ecological, and social research.

To address these problems fundamental to the creation of land-cover products, this thesis seeks to develop a novel contextual classifier for multi-temporal land-cover mapping that fully utilizes
spatial-temporal information to increase the accuracy of change detection, while remaining resistant to future advances in the spatial and spectral characteristics of remote sensor technology. By combining the complementary strengths of two leading techniques for the classification of land cover – the Support Vector Machine and the Markov Random Field – through a novel spatial-temporal contextual kernel method, this goal can be achieved for the betterment of multi-temporal remote sensing. Improvement of the technique through parameter selection and hierarchical classification technology makes it tractable to obtain highly accurate and spatial-temporally consistent multi-temporal land-cover maps automatically and without post-processing. Application of this proposed classifier in a case study of an Appalachian Ohio study area shows the technique significantly improves upon competitive techniques in multi-temporal land-cover mapping. It significantly increases the accuracy of change detection, increases the classification accuracy at individual dates, and reduces temporal inconsistency among land-cover change trajectories. Altogether, by producing higher quality multi-temporal land-cover products, we enable the more accurate and consistent explanation of the places, periods, and types of land-cover change occurring on our Earth and the myriad effects we humans are having on it.
Acknowledgments

Foremost, I thank my advisor, Dr. Desheng Liu, for his encouragement, guidance, and interest in my research project. His support and advice has been invaluable to me and it will be my great pleasure to continue working with him for my Ph.D. I also thank Dr. Ningchuan Xiao and Dr. Brian Kulis who served on my thesis defense committee, provided helpful suggestions on this work, and demonstrated great patience with my writing of it. I also thank Dr. Yoonkyung Lee for her insights into multiclass classification, which proved invaluable to my methods.

I extend my thanks as well as my warmest regards to the members of Dr. Liu’s research group and seminars who offered their support, discussion, and encouragement. In particular, I wish to thank Dr. Shanshan Cai whose work together with Dr. Liu originally inspired my interest in remote sensing. I also thank my officemate and friend Ms. Haixia Liu. I further thank Mr. Xiaolin Xu; Mr. Zhouxin Xi; Dr. Hongshuo Wang; Dr. Shiguo Jiang; Ms. Fang Zhang, who was the first OSU geographer I had the pleasure of knowing; Ms. Weishu Gong; Ms. Qian Muo; and Dr. Wei Su. I also wish to thank Ms. Christine Chen, Mr. Alfonso Fernandez, Mr. Wenkai Guo, and Mr. Olive Wigmore from seminar.

In the past, I was inspired and encouraged to pursue geography by my professors at The University of Akron. I thank Dr. Linda Barrett, Dr. Shanon Donnelly, and Dr. Kevin Kreider. I also thank Mr. Joseph DiBiase who was prescient in discussing geography with me earlier in my
youth.

I thank my roommates Mr. Colin Sinclair, Mr. Joshua Bates, and Mr. Jonathan Wille for their distractions, antics, and support through my master’s study. I thank new friends and old friends: Mr. Dan D’Amico, Mr. Sam Kay, Mr. and Mrs. Suneil and Ishani Vallabh, Mr. William Holmes, Mr. Jon McKay, Mr. Jake Prickett, Mr. Nathan Weller, Ms. Christina Welch, and the lovely Ms. Cheryl Sevilla.

I offer my sincere gratitude and love to my parents for always supporting me and encouraging me to follow my passions. I thank my brothers, particularly Ben (he required that I say that). I thank my step-family as well as my dearest grandmother for believing in me… and also always slipping extra cash into my pockets so that I can eat.

Finally, to dear Shado who was taken from us far too soon…
Vita

2005 .................................................. Western Reserve Academy

2010 .......................................................... B.S. Mathematics, The Ohio State University

2010 .......................................................... B.A. English, The Ohio State University

2012 .......................................................... Graduate Certificate in Geographic Information Sciences, The University of Akron

Fields of Study

Major Field: Geography

Minor Field: Computer Science
# Table of Contents

Abstract ........................................................................................................... ii

Acknowledgments .............................................................................................. iv

Vita ...................................................................................................................... vi

Table of Contents .............................................................................................. vii

List of Tables ...................................................................................................... xi

List of Figures ..................................................................................................... xiii

Chapter 1: Introduction ...................................................................................... 14

1.1 Setting and Problem Addressed .................................................................... 14

1.2 Objective of the Thesis ................................................................................ 17

1.3 Organization of the Thesis .......................................................................... 18

Chapter 2: Review ............................................................................................. 19

2.1 Introduction .................................................................................................. 19

2.2 Remote Sensing Data and its Classification .............................................. 20

2.3 Contextual Classification of Remote Sensing Data .................................. 24

2.3.1 Object-Based Image Analysis ............................................................... 26

2.3.2 Markov Random Field Models ............................................................. 27
2.3.3 Contextual Kernel Methods ................................................................. 30
2.3.4 Other Techniques .............................................................................. 32
2.4 Detection and Mapping of Land-Cover Change ........................................ 33
2.5 Support Vector Machines ........................................................................ 35
  2.5.1 Hard Margin SVM .............................................................................. 35
  2.5.2 Soft Margin SVM ............................................................................... 37
  2.5.3 Non-Linear SVM ................................................................................ 38
  2.5.4 Kernel Methods................................................................................... 39
  2.5.5 Advantages ........................................................................................ 41
  2.5.6 Disadvantages .................................................................................... 41
2.6 Markov Random Fields ........................................................................... 43
  2.6.1 Theory ................................................................................................ 43
  2.6.2 Advantages ........................................................................................ 47
  2.6.3 Disadvantages .................................................................................... 48
2.7 Conclusion ................................................................................................ 48
Chapter 3: Methods ....................................................................................... 50
  3.1 Review of the Markovian Kernel............................................................. 50
  3.2 Generalized Markovian Kernel ............................................................... 52
  3.3 Proposed Spatial-Temporal Kernel ........................................................ 54
  3.4 Proposed Spatial-Temporal Classifier .................................................... 57
3.4.2 Multi-class Classification ......................................................... 58
3.4.3 Hierarchical Classification ........................................................... 62
3.4.4 Parameter Estimation ................................................................. 65
3.4.5 Transition Probabilities ............................................................... 65
3.4.6 Illogical Transition Penalties ......................................................... 68

Chapter 4: Case Study in Appalachian Ohio .............................................. 71

4.1 Data .................................................................................................... 71
  4.1.1 Study Area .................................................................................. 71
  4.1.2 Training and Validation Data ....................................................... 73

4.2 Experiments ....................................................................................... 75

4.3 Results ............................................................................................ 78
  4.3.1 Experiment 1 ............................................................................ 80
  4.3.2 Experiment 2 ............................................................................ 90
  4.3.3 Experiment 3 ............................................................................ 93
  4.3.4 Summary of Results ................................................................. 98

4.4 Discussion ....................................................................................... 99

4.5 Conclusion ..................................................................................... 102

Chapter 5: Conclusion ........................................................................... 103

5.1 Implications of Research ................................................................... 103
5.2 Future Work ................................................................................... 104
List of Tables

Table 1. Summary of supervised classifiers commonly employed in remote sensing. .......... 24

Table 2. Kernels employed with remote sensing data. ......................................................... 40

Table 3. Summary of data acquisition for Appalachian Ohio dataset. ............................... 71

Table 4. Summary of Landsat 5 TM spectral bands............................................................. 73

Table 5. Classification scheme for Appalachian Ohio scenes............................................. 73

Table 6. Summary of reference data for Appalachian Ohio dataset. ................................. 74

Table 7. Summary of illogical land-cover transitions for Appalachian Ohio dataset .......... 74

Table 8. Summary of attributes of classifiers in experiments for Appalachian Ohio dataset. ..... 76

Table 9. Summary of experiments for Appalachian Ohio dataset. ..................................... 77

Table 10. Optimal band selections by land-cover transition for LAT model for Appalachian Ohio dataset. .......................................................... 79

Table 11. Accuracies of change detection in Experiment 1 for Appalachian Ohio dataset. .... 82

Table 12. Overall accuracies in Experiment 1 for Appalachian Ohio dataset...................... 84

Table 13. Average accuracies in Experiment 1 for Appalachian Ohio dataset. .................... 85

Table 14. Illogical land cover transition frequencies in Experiment 1 for Appalachian Ohio dataset. .............................................................................................................. 86
Table 15. Land cover trajectory counts and frequencies in Experiment 1 for Appalachian Ohio dataset. ................................................................. 86

Table 16. Interclassifier significance test results in Experiment 1 for Appalachian Ohio dataset. 89

Table 17. Accuracies of change detection in Experiment 2 for Appalachian Ohio dataset. ........ 90

Table 18. Overall accuracies in Experiment 2 for Appalachian Ohio dataset. .................. 91

Table 19. Average accuracies in Experiment 2 for Appalachian Ohio dataset. .............. 92

Table 20. Interclassifier significance test results in Experiment 2 for Appalachian Ohio dataset. 93

Table 21. Accuracies of change detection in Experiment 3 for Appalachian Ohio dataset. ........ 94

Table 22. Overall accuracies in Experiment 3 for Appalachian Ohio dataset. .................. 95

Table 23. Average accuracies in Experiment 3 for Appalachian Ohio dataset. .............. 95

Table 24. Land cover trajectory counts and frequencies in Experiment 3 for Appalachian Ohio dataset. ................................................................. 96

Table 25. Illogical land cover transition frequencies for GT model in Experiment 3 for Appalachian Ohio dataset. ................................................................. 97

Table 26. Illogical land cover transition frequencies for LAT model in Experiment 3 for Appalachian Ohio dataset. ................................................................. 98
List of Figures

**Figure 1.** Multi-temporal neighborhood system................................................................. 55

**Figure 2.** Schematic for use of contextual information in the proposed Spatial-Temporal Markovian Support Vector Classifier (ST MSVC)........................................................................ 61

**Figure 3.** Schematic for use of contextual information in the proposed Spatial-Temporal Hierarchical Markovian Support Vector Classifier (ST HMSVC) ........................................ 64

**Figure 4.** Subsets of satellite images for the Appalachian Ohio dataset and location of subset within Ross County scene. Spectral data is depicted with 7-5-3 band combination .............. 72

**Figure 5.** Classification hierarchy generated for Appalachian Ohio dataset............................ 79

**Figure 6.** Classification maps for a subset of a) the 2003 date of the Appalachian Ohio scene generated with b) SVM, c) MRF STP, and d) MSVC STP classifiers under Experiment 1. Spectral data is depicted with 7-5-3 band combination................................................................. 81

**Figure 7.** Trajectory transition maps for a subset of the Appalachian Ohio scene counted from the land-cover maps generated with a) SVM, b) MRF STP, and c) MSVC STP classifiers under Experiment 1 .................................................................................................................. 88

**Figure 8.** Computational equivalence between HMSVC and MSVC ........................................ 99
Chapter 1: Introduction

1.1 Setting and Problem Addressed

The land system is undergoing a rate, extent, and intensity of change that is unprecedented in human history. The accurate and consistent detection, monitoring, and quantification of this change are therefore crucial to improving our understandings of and responses to the variability, drivers, and effects of this trend. Land cover plays a key role in regulating carbon and water cycles. It influences both the macro- and micro-level climate system. Its change has wide-reaching impacts on both human and natural systems. Yet, our capability to assess its change is greatly hampered by the insufficiently accurate and spatial-temporally inconsistent land-cover products that may be produced or employed in remote sensing studies (Hansen & Reed, 2000; Giri, Zhu, & Reed, 2005; Jung, Henkel, Herold, & Churkina, 2006; Liang & Gong, 2010; Cai, Liu, Sulla-Menashe, & Friedl, 2014). It is recognized that classification methods that make better use of the spatial and temporal properties of remote sensing data can better discriminate between land-cover types and therefore improve classification accuracies (Liu & Zhou, 2004; Jung et al., 2006). Methods that also use this information to consider the context of land cover in one location with the land cover in nearby locations can reduce the effects of noise in and the propagation of errors among classification results (Burnicki, Brown, & Goovaerts, 2007; Liu & Chun, 2009). Altogether, such contextual land-cover classifiers can provide significantly more accurate detection and quantification of land-cover change than traditional non-contextual classification methods. By doing so, improved land-cover products can be created that accurately
and consistently explain the places, periods, and types of land-cover change occurring on our Earth. This outcome is crucial to improving the myriad uses for land-cover classification in land change science as well as downstream environmental, ecological, and social research.

In remote sensing, it is well-known that classification accuracy at individual dates must be exceedingly high in order to also detect land-cover change accurately. The problem is that, when the level of accuracy is insufficient, the occurrence of change due to error greatly outweighs actual land-cover change (Pontius & Lippitt, 2006). This follows because, when the errors in a set of source maps are completely independent, the product of the accuracies of the source maps is a lower bound for the accuracy of change detection (Congalton & Green, 1999). Likewise, assuming complete dependence of the errors in a set of source maps, the minimum accuracy among the source maps offers an upper bound for this accuracy (Liu & Cai, 2012). Hence, both the spatial-temporal dependence of errors as well as the accuracy of the individual maps used for change detection must be increased in order to improve the overall accuracy of change detection (van Oort, 2005; Burnicki et al., 2007; Burnicki, 2011). Because the accuracy of change detection is bounded below by the product of accuracies at individual dates, the problem of inaccurate change detection is further compounded as the number of maps to be compared increases. This is problematic for land change science because, as the field has matured, its focus has naturally shifted from the detection of change between two dates to the analysis of trajectories of land-cover change over many dates (Liu & Cai, 2012). Such analysis offers a much richer understanding of the patterns of change that are occurring in an area that is more readily linked to known environmental and social processes. Therefore, due to the growing length and complexity of our understanding of land-cover change, it is of the utmost importance that the methods to produce more accurate and spatial-temporally consistent classifications continue to advance.
Concurrent with the growing length and complexity of studies of land-cover change, time also brings with it continued sensor advancement, which fundamentally increases the spectral and spatial resolutions of satellite imagery. It is recognized that the spatial and temporal resolutions of current land-cover products are inadequate for anticipated scientific uses despite the availability of long-term, finer-resolution records, such as the Landsat data archive (National Research Council, 2007). Therefore, to develop higher resolution, multi-temporal land-cover products, classification methods must be advanced to cope with the challenges induced by this increased detail. Here, we call attention to the two main challenges that sensor advancement engenders. First, increasing spectral resolution can contraindicate the use of classification methods that rely on a small number of features relative to training sample sizes. As the number of spectral features increases, it is readily apparent that the volume of the feature space to which they belong increases so fast that data in that space will become sparse. Consequently, conventional methods for discriminating features can lose power, as measures such as distance begin to look similar for all features in these enlarged spaces. Accordingly, more training data is also needed to adequately sample these spaces (Hughes, 1968). Because training data is expected to remain limited due to its high human cost of production, classifiers that are robust against this Curse of Dimensionality are thus required to deal with sensor advancement. Second, increasing resolution can also contraindicate the use of classification methods that rely on distributional assumptions. On one hand, as spectral resolution increases, it is often the goal to discriminate between more detailed land-cover types, such as vegetation species or mineral types. However, the signatures for the spectral responses of these land-cover types can be very similar and their distributions overlap significantly. Classifiers that can handle multi-modal and non-Gaussian data are thus expected to perform better than the ones depending on simplifying assumptions. On the
other hand, as spatial resolution increases, the goal may not be to identify more detailed land-cover types, but instead more precisely delineate their extent, quantify their area, or measure landscape heterogeneity. In this case, the increased spatial resolution reveals more within-class spectral variability of land cover and less between-class variability among land covers, which further challenges the discriminative capabilities of classifiers. Hence, in total, improved classification methods must make better and more efficient use of high dimensional data and employ methods that increase their discriminative capability vis-à-vis increasing spectral and spatial resolutions.

Existing contextual classifiers offer the ability to address these challenges of discriminating land cover and increasing the spatial-temporal consistency of classification results, but, to our knowledge, no spatial-temporal classifier currently exists that does these things while efficiently using high dimensional data or optimally using contextual information. The focus of this thesis is thus the development of a spatial-temporal contextual classifier that improves classification accuracy over the state-of-the-art, increases spatial-temporal consistency among land-cover maps, and is robust against future advancements in sensor technology. Within the realm of land change science, the application of such a classifier can enable the creation of significantly improved land-cover products that aid in understanding the effects of land-cover change.

1.2 Objective of the Thesis

The objective of this thesis is to develop a spatial-temporal contextual classifier that improves classification accuracy over the state-of-the-art, increases spatial-temporal consistency among land-cover maps, and is robust against future advancements in sensor technology for the purpose
of multi-temporal land-cover mapping in remote sensing.

1.3 Organization of the Thesis

This thesis consists of five principal chapters, which are organized as follows:

**Chapter 1** introduces the need and significance of advancing classification methods in remote sensing.

**Chapter 2** reviews the challenges of classifying remote sensing data, the use of contextual classifiers in remote sensing, and provides background on two leading classification techniques that form the basis of the methods advanced in this thesis.

**Chapter 3** presents an improved method for multi-temporal land-cover mapping that integrates spatial-temporal contextual information into a statistically optimal classifier.

**Chapter 4** evaluates the performance of this classifier in a case study of land-cover change detection in an Appalachian Ohio study area in comparison to preceding methods.

**Chapter 5** concludes the thesis by discussing the implications of the work and directions for future research.
Chapter 2: Review

2.1 Introduction

Through remote sensing, we can acquire information about the state of the Earth’s land system from distances, perspective, and scales that are otherwise impractical for human observers. Satellites and airborne sensors take instantaneous, spatially-continuous measurements of the surface in different wavelengths of the electromagnetic spectrum. These measurements provide quantitative information about the reflectance properties of both natural and manmade features, which we can analyze to perform tasks critical to scientific research such as the estimation of biophysical quantities and the mapping of the land surface. As humans, we are able to use our innate abilities of pattern recognition and reasoning to interpret images of the Earth’s surface. However, to do so for the vast areas covered by remote sensors is a time-consuming, labor-intensive, fatigue-sensitive, and thus high-cost task. To use remote sensing data more efficiently, we teach computers to recognize and use patterns in the data to categorize measurements into meaningful units, which we can then summarize, reason about, and employ in modeling efforts. In the context of machine and statistical learning, we call these units classes and the process of categorization classification. Computer classification automates, standardizes, and makes objective and repeatable the categorization process. It is also potentially able to discover and distinguish patterns that are unknown to or not easily assessed by human observers. However, computer classification is not a panacea because its results depend on many factors and no one algorithm is suitable for all data across all domains. However, with knowledge and
understanding of the problems within a domain, algorithms that excel at particular tasks can be designed. It is under this recognition that contextual classification has been advanced in remote sensing.

In the remainder of this section, we will first describe the remote sensing data and the general framework for its classification. Second, we will discuss a broad categorization of contextual classification methods in remote sensing and then review four approaches to using contextual information with applications to multi-temporal land-cover mapping. Third, we will touch on frameworks for performing change detection. Finally, we will detail the mathematical and statistical basis for the techniques upon which the methods advanced in this thesis depend before concluding.

2.2 Remote Sensing Data and its Classification

In remote sensing, satellite and airborne sensors provide measurements of the radiance of the Earth’s surface in a number of different bands of the electromagnetic spectrum. These measurements are sampled in sequence over a portion of the Earth in an effectively instantaneous timeframe to form images. Each pixel in an image is geo-referenced to correspond to a known ground location. Remote sensing imagery is thus said to be spatially-referenced, spatially-extensive, and spatially-continuous. The number of bands for which a sensor collects measurements is known as the spectral resolution of the sensor. Each pixel in a remote sensing image thus associates with a set of measurements equal in size to the number of bands measured by its source sensor. This set forms the spectral signature or feature vector for a pixel. Both sensors and their associated imagery are commonly categorized as visible, infrared, multispectral,
or hyperspectral depending on either the portion of the spectrum or the number of bands for which the sensor collects measurements. The radiometric resolution of a sensor refers to the quantization scheme used to convert measured radiance to digital numbers for storage in computer memory. Sensors that quantize over eight bits, for example, will record measurements as digital numbers in the range of 0 to 255. With greater radiometric resolution, finer differences in energy can be measured. The spatial resolution of a sensor refers to the ground sampling distance between measurements taken by the sensor. This distance effectively determines the area on the ground that a pixel covers. Pixels in an image from the Landsat Thematic Mapper sensor, for instance, are spaced 30 meters apart, meaning that no ground feature smaller than 30 by 30 m can be fully resolved in the image. Altogether, the spectral, radiometric, and spatial resolutions of remote sensing imagery describe the intrinsic properties of data that can be used for classification.

The radiance measured for a location on the Earth’s surface is subject to both noise and variation resulting from physical effects, spectral mixing, and phenology. In order to reliably compare spectral signatures across space and time, physical effects, such as those due to atmospheric, illuminative, and topographic conditions, must be corrected for prior to classification. Typically, radiance is first converted to top of the atmosphere reflectance, which is then used to retrieve ground surface reflectance through the combination of different physical, statistical, or geometric models. Without accounting for these effects, the spectral signature of the same land cover at different points in time or in different areas of an image can vary in complex ways, increasing the difficulty of its classification. Residual noise either left over or possibly induced by these correction procedures can also affect the classification of a pixel, although the effect is generally small when correction has been performed appropriately and competently. Additionally, the
spectral signature for a pixel is subject to both linear and non-linear mixing of the spectral responses of land cover and features at the sub-pixel level. Linear mixing occurs when a pixel is heterogeneous in the land-cover types it contains at the sub-pixel level. The responses of multiple land-cover types then combine proportional to the area they occupy within the pixel. Non-linear mixing occurs when the wavelength of light is additionally altered due to its interaction among and within land-cover types. Hence, the spectral signature of a pixel may be a complex combination of the responses of its constituent land-cover types, which results in variation that also increases the difficulty of its classification. Finally, many natural land cover are subject to phenological variation, resulting in changes in leaf cover, pigmentation, canopy structure, and other attributes over time. The biophysical conditions at a location for which the spectral response of land cover is measured can thus also induce variation in those responses over time. When working with time sequences of remote sensing imagery, this can thus have an effect on the classification result if model parameters or other information is shared across image dates.

With these properties of and variation inherent to remote sensing data in mind, land-cover classification generally seeks to learn the relationships between categories of land cover and the spectral responses of those categories in remote sensing data. Hence, land-cover classification makes use of methods developed in the machine and statistical learning communities. Depending on whether we have a priori information about what relationships should be learned, we call the learning process supervised or unsupervised. In supervised learning, a classification scheme for the image is developed prior to classification and portions of its pixels are labeled to provide examples of each class in the scheme. The job of the classifier is to use these examples to learn how to distinguish the provided classes from one another. Hence, the learned relationships can predict the classification for the remaining unlabeled pixels the image. The provided labeled
examples are called training data. These data are typically collected through manual interpretation of a sample of locations in a higher-resolution satellite image or, in the best case, by truthing the corresponding ground locations by physically visiting them at the same time an image is acquired. The performance of a supervised classifier depends on its generalization capability, the quality and quantity of the training data, and how well the data matches any assumptions of the classification model. In unsupervised learning, in contrast, no information about the correct classification of any of the pixels in an image is provided, so the job of the classifier is to uncover structure among the data that aids in its interpretation. Typically, the data are partitioned into clusters based on the similarity of its spectral responses under some criteria. These clusters then form spectral classes, which may be combined together to form information classes representing land cover to a human analyst. While both learning styles have been used in remote sensing, we will focus on classification using supervised models.

In remote sensing, a number of classifiers have commonly been employed. The earliest examples are the maximum likelihood, parallelepiped, and minimum distance classifiers. More recently, statistical and machine learning classifiers, such as linear discriminant analysis, decision trees, artificial neural networks, and support vector machines have gained favor due to their increased performance on many datasets. Among these classifiers, support vector machines (SVMs) have consistently achieved superior performance compared to the others (Huang, Davis, & Townshend, 2002; Foody & Mathur, 2004). As those referenced authors note, artificial neural networks are the most competitive with SVMs as the two achieve similar levels of classification accuracy in many situations. However, with increasing data dimensionality or decreasing training dataset size, SVMs tend to outperform artificial neural networks. The reason for their superior performance is their remarkable generalization capability, which results from a combination of
Theoretical and practical attributes related to the simultaneous maximization of the geometric margin of a decision surface and minimization of empirical error, sparse representation of training data, and the use of very general feature representations (Vapnik, 1999). Compared with maximum likelihood and linear discriminant analysis classifiers, SVM and the others are non-parametric, which means that they do not assume the data has been generated from any particular distribution. For remote sensing data, which may contain land-cover classes that are overlapping, multi-modal, and non-Gaussian in nature, such classifiers typically perform better than their parametric counterparts because they do not depend on distributional assumptions.

Table 1 below summarizes the preceding information for the five most common supervised classifiers.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Distributional Assumption</th>
<th>Training Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLC</td>
<td>Maximum Likelihood</td>
<td>parametric</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
<td>parametric</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>DTC</td>
<td>Decision Tree</td>
<td>non-parametric</td>
<td>hierarchical variable selection and thresholding</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
<td>non-parametric</td>
<td>back-propagation</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
<td>non-parametric</td>
<td>quadratic programming</td>
</tr>
</tbody>
</table>

2.3 Contextual Classification of Remote Sensing Data

Contextual information refers to the site-specific relationships between a pixel or group of pixels of interest and the rest of the pixels in a scene. Methods for using contextual information can be broadly divided into three categories based on how these relationships are specified. In the first category, these relationships are specified in terms of the texture or structure of image objects.
Texture refers to variations in patterns of image tone within an object or local neighborhood, whereas structure refers to descriptors such as the size or shape of an object or its topological relation to other objects. Quantifications of these elements are typically included as additional features for a pixel or image object during classification. Methods based on Object-Based Image Analysis (OBIA) are typical of this category. In the second category, these relationships are specified in terms of the probability of assignment of a class label to a pixel in one location relative to the configuration of class labels for pixels in surrounding locations. Methods making use of such dependencies therefore require the specification of an initial configuration of class labels, which make them iterative approaches to improving classification results. Methods based on Markov Random Field (MRF) models are typical of this category. In the third category, ancillary information about the relationship of a pixel to some spatially organizing system is used to improve classification results. For example, the membership of a pixel to the set of pixels located in the boundaries of an agricultural field, ecological zone, or soil type may be included as an additional indicating feature for a pixel or image object during classification or may be used to refine classification results during post-processing. Because of the *ad hoc* nature of methods in this category, we will not review it further.

In the following, we will review the two major traditions and an emerging one for using contextual information in remote sensing. We will finish by mentioning some additional techniques that are significant but do not constitute a larger tradition of methods. The three traditions we will review are Object-Based Image Analysis (OBIA) techniques, Markov Random Field (MRF) models, and the emerging Contextual Kernel Methods (CKM).
2.3.1 Object-Based Image Analysis

OBIA techniques segment remote sensing images into homogeneous regions called image objects (Blaschke, 2010). Objects are then attributed with local statistical measures (e.g. mean, standard deviation) derived from information sources that are geographically coincident. Geometric measures (e.g. roundness, compactness, elongatedness) that also characterize these objects are typically attributed as well. The topological relationships between objects derived from segmentations at the same scale or among those derived at different scales can also distinguish image objects. Based on these attributes, image objects can be combined to form new objects and form information classes. Alternatively, feature vectors can be formed for pixels based on the attributes of the objects to which they belong and then used along with training data to perform supervised classification. Altogether, such methods provide a powerful, albeit often ad hoc, technique to contextually classify remote sensing imagery.

In the literature, a number of studies have assessed or applied OBIA approaches for the purposes of change detection or multi-temporal land-cover mapping. Hall and Hay (2003) propose the generation of a multi-scale segmentation in which image objects undergoing change can be identified and delineated based on scale. Blaschke (2005) propose the use of overlay operators on vectorized image objects to perform change detection. Desclee, Bogaert, and Defourny (2006) propose the use of a multi-date segmentation and changes in object statistics to detect land-cover change. Niemeyer, Marpu, and Nussbaum (2008) propose the use of a multivariate alteration detection (MAD) technique to detect change among features extracted from image objects at two dates. Conchedda, Durieux, and Mayaux (2008) propose the use of a multi-date segmentation with a nearest neighbor classifier to group spectral-temporally similar objects to land-cover transition classes between image dates. Schopfer, Lang, and Albrecht (2008) detail a framework
for conducting, evaluating, and comparing the results of vector-based object-change detection. Bontemps et al. (2008) propose a probabilistic framework for detecting land-cover change by deviation of the temporal signature of image objects from a reference unchanged signature. Gamanya, de Maeyer, and de Dapper (2009) apply fuzzy overlay techniques to independent segmentations at two dates to detect land-cover change. Petitjean et al. (2012) propose an unsupervised classification of time series features constructed from image objects derived using independent image segmentations. Although many approaches to performing change detection have been proposed in OBIA literature, studies have largely been conducted using only two image dates with unclear methodology on how to best utilize temporal information.

2.3.2 Markov Random Field Models

Markov Random Fields (MRF) became a practical framework for modeling image context in the 1970s, although the theory began with the Ising model of statistical mechanics (Ising, 1925). A proof of equivalence between the MRF and Gibbs Random Field was formulated by Hammersley and Clifford (1971). Further work was accomplished by British spatial statistician Julian Besag (1974; 1975; 1986), who also introduced the Iterated Conditional Modes (ICM) energy minimization technique. MRF are undirected graphical models, which are part of the larger class of probabilistic graphical models studied in machine and statistical learning. Some statistical details will be provided in Section 2.6 of this thesis.

Most MRF-based classifier make use of the maximum a posteriori (MAP) criterion to establish the optimal class labeling for an image in what is known as the MAP-MRF framework (Li, 2009). Hence, depending on the optimization technique, MRF theory provides an (at least locally) optimal, probabilistic method for assigning labels to pixels in images through an analysis of their
contextual dependencies. Within the context of the ICM optimization algorithm, the solution to the MRF problem involves an initial non-contextual classification of an image and then an iterative update of the class labels of that image based on the minimization of an energy function that accounts for likelihood and prior information about the labeling for each pixel. This information is specified in terms of a local neighborhood system and, over iterations, information from pixels in surrounding neighborhoods is propagated across the image lattice to inform other pixels.

In remote sensing literature, a number of classifiers based on MRF models have been proposed. Here we will focus on classifiers employed for the purpose of multi-temporal land-cover mapping and change detection. Solberg, Taxt, and Jain (1996) formulate MRF models for the classification of multi-sensor and multi-temporal imagery. They use a unidirectional cascade of temporal information to fuse temporal information sources. Meglani and Serpico (2003) improve upon the previous model using a novel bidirectional sharing of temporal information. Liu, Kelly, and Gong (2008) formulate locally-adaptive models for estimating class transition probabilities that can increase classification accuracy in study areas where an assumption of stationarity of the transition probabilities is inappropriate. Liu and Cai (2012) propose a multi-temporal MRF framework that includes an arbitrary number of image dates. They also incorporate a novel temporal exclusion term that penalizes the occurrence of illogical land-cover transitions during classification. Results indicated a reduced risk of classification error propagation, increased accuracies of change detection, and reduced occurrence of illogical land-cover transitions. Although not used in MRF, the use of temporal rules to improve classification results has precedence in Liu and Zhou (2004) and Powell et al. (2008). For the purpose of change detection, Bruzzone and Prieto (2002) propose an unsupervised classifier that applies a MRF
model to the difference image using adaptive semi-parametric class estimates. Results indicated an overall reduction in error compared to traditional thresholding techniques. Melgani and Bazi (2006) propose a spatial MRF model for unsupervised change detection. They apply the model to a set of change maps produced using an ensemble of thresholding algorithms to produce a global change map with more robust accuracy and generally lower error rate than the majority of the algorithms in the ensemble.

A few studies have compared classifiers based on MRF models with other techniques. Liu and Chun (2009) investigate the performance of a MRF approach against non-contextual and filter-based approaches in regards to error propagation during land-cover change detection. MRF prove to have the smallest error propagation rate with spatial autocorrelation of the errors being the most significant factor affecting this rate. Thoonen et al. (2012) compares MRF and OBIA approaches for the classification of a hyperspectral image in a Belgium heathland study area. Based on a set of object metrics, the MRF approach scored better in thematic accuracy, had higher edge complexity, lower fragmentation of land cover, lower over-smoothing, and lower mean edge deviation than the OBIA approaches. Their assessment was that the MRF classification result was much closer to their reference data than the OBIA result and differed less when it differed than the OBIA approach. Overall, the authors concluded that the MRF was better suited for classification of their study area than the OBIA approach. Cai and Liu (2013) compare MRF and OBIA approaches for high- and moderate-resolution imagery. Their results indicate the MRF approach results in higher classification accuracies than the OBIA approach and that, based on over-segmentation, under-segmentation, edge location, and shape error metrics, the two approaches were comparable in preserving feature geometries. However, the OBIA approach had difficulty resolving small object features as well as performing on the lower resolution
imagery due to its dependence on segmentation results. The MRF approach was able to adapt to different sized features and perform well at both resolutions.


2.3.3 Contextual Kernel Methods

Kernel methods are an emerging tradition in remote sensing. Such methods exploit the similarity of data examples in a transformed feature space in order to classify them. Some mathematical details will be provided in Sections 2.5.3 and 2.5.4 of this thesis. Contextual kernel methods make use of contextual information in the formulation of the kernel function.

In remote sensing, much research into contextual kernel methods has originated out of Image and Signal Processing Group at the University of Valencia in Spain. This research has primarily focused on the creation of simple linear and non-linear combinations of spectral, spatial, and temporal information defined through local statistics. Camps-Valls et al. (2006) introduce a number of composite kernels to account for different interactions between spectral and spatial information and evaluate their use in a SVM classifier on a hyperspectral dataset. Results show significant increases in accuracy using spatial information within cross-information and weighted summation composite kernels over the standard RBF kernel using spectral features. Camps-Valls, Marsheva, and Zhou (2007) propose a semi-supervised graph-based composite kernel classifier.
for hyperspectral imagery. They include the previously formulated composite kernels in a graph-based spreading function to transductively classify unlabeled image samples. Results indicated higher classification accuracies with the use of this contextual classifier than for SVMs trained with the equivalent kernels for a range of limited training sample sizes. Camps-Valls et al. (2008) propose a family of supervised classifiers for multi-temporal image classification and change detection. Traditional change detection methods, such as differencing and ratioing, were reformulated in terms of kernels. Kernels were also presented for contextual and multi-source data fusion. Results indicated cross-information composite kernels achieve the highest classification accuracies. They conclude that kernel-based change detection is a viable alternative to traditional change detection methods due to its ability to handle non-linear relationships between temporally-related pixels. The book *Kernel Methods for Remote Sensing Data Analysis* edited by Camps-Valls and Bruzzone (2009) collects the previous articles and other applications of kernel methods in remote sensing.

Contextual information based on morphology and label frequencies have been incorporated into kernels. Fauvel, Chanussot, and Benediktsson (2012) propose a contextual kernel method for the supervised classification of high-resolution and hyperspectral imagery. They define spatial features using morphological neighborhoods for pixels based on their flat zone characteristics. The spatial information from these neighborhoods is integrated with spectral information in a composite kernel. Results indicated that their method was competitive against or better than a range of other contextual approaches. They compared with MRF, morphological profiling, feature extraction and clustering, and multi-scale spatial decomposition approaches. Moser and Serpico (2013) propose a methodology for integrating MRF with SVM through the specification of a novel spatial contextual Markovian kernel. Their method reformulates the decision function
of a MRF as a kernel. They use it for the classification of high-resolution and hyperspectral imagery. Results indicated the proposed approach outperforms SVM, MRF, composite kernel, and graph kernel classifiers on their datasets.

For the purpose of change detection, Volpi et al. 2013 propose an SVM-based supervised change detection technique that incorporates local textural statistics and mathematical morphology. The result is promising, but no comparison to the state-of-the-art was performed. Their work relates to research by Bovolo et al. (2010), Bovolo et al. (2012), and Volpi et al. (2012) on change detection by one-class SVM, spectral change vector analysis, and kernel k-means.

2.3.4 Other Techniques

In addition to the previously reviewed traditions, a few significant approaches have been proposed that are not part of any recognizable tradition.

Melgani (2004) proposes a supervised fuzzy-logic approach to multi-temporal and multi-source contextual classification. After an initial non-contextual classification by artificial neural network, the optimal label for a pixel is assigned in a winner-takes-all strategy based on maximum probabilities defined over each potential label based on the class conditional probability, class relative frequency in a local neighborhood, and class transition probability for a pixel after fuzzy filtering. Results indicated comparable classification accuracy to MRF approaches without the need for iteration or greater classifier complexity in a case study utilizing two image dates.

Boucher, Seto, and Journel (2006) propose the modeling of spatial context using geostatistical
methods and its integration with cascaded temporal information in a maximum likelihood classification scheme for multi-temporal imagery. Indicator kriging is used to measure spatial autocorrelation among land-cover classes. Ground observation data is used to define adaptive neighborhoods of high class confidence subject to an initial image segmentation defining data discontinuities. Information is propagated temporally by a Markov chain. Results indicated significant improvement over a maximum likelihood classification, but the technique was not compared with other contextual methods.

Bruzzone and Carlin (2006) propose an adaptive multi-level feature extraction method for modeling hierarchical spatial context in the supervised classification of high resolution imagery. The spatial context is incorporated as additional features for SVM classification. Results indicated higher classification accuracies compared to a Gaussian pyramid decomposition segmentation algorithm, but no comparison was made to non-segmentation approaches.

Negri, Dutra, and Sant’Anna (2014) propose a technique for displacing locally the position of the separating hyperplane in SVM based on spatial contextual information. They re-project the highest confidence examples to be closest to the hyperplane, employ a repulsion model to shift the hyperplane locally, and then reclassify using the new decision boundary. Results show an increased accuracy over the results obtained with a standard SVM and better preservation of edge pixels than obtained using a majority filter to post-process the classification results.

2.4 Detection and Mapping of Land-Cover Change

Thresholding methods such as image differencing and ratioing, change vector analysis, time
series analysis, and post-classification comparison have been used to detect land-cover change. Time series analysis typically seeks to detect vegetative disturbances through the identification of change-points directly in time series by regression, curve-fitting, or segmentation approaches. Vegetative disturbances are frequently the focus in this area due to their characteristic patterns of change and importance to ecological research. Collectively, these techniques are known as profile-based approaches (e.g. Kennedy, Yang, and Cohen, 2010; Huang et al., 2010). Alternatively, the detection of land-cover change may employ the post-classification comparison framework, which compares land-cover maps at individual dates to identify changes between dates. These classification-based comparisons are advantageous over image differences and ratios, which operate only between two dates, because the collection of transitions between dates form land-cover change trajectories, whose patterns are more readily linked to known processes of change. The shortcoming of this approach is the need for highly accurate classifications at individual dates, which may lack spatial-temporal consistency if not produced while taking into account spatial and temporal context. Some representative studies using post-classification comparison have been conducted for southern Cameroon (Mertens & Lambin, 2000), the northern Ecuadorian Amazon (Mena, 2008), the Ukrainian Carpathians (Kuemmerle et al., 2009), and Appalachian Ohio (Liu & Cai, 2012). Some studies have included the use of temporal rules for assessing the consistency of classification results (Liu & Zhou, 2004) and for incorporating expert knowledge post-classification (Powell et al., 2008; Liu & Cai, 2012). Altogether, such studies can be variously improved through the better use of spatial-temporal contextual information or increased ability for class discrimination.
2.5 Support Vector Machines

Having reviewed some significant traditions to contextual classification, we will now provide background on the leading non-contextual classifier, the SVM, on which we will later base the development of our methods. We introduce the hard and soft margin formulations of this linear classifier, discuss its extension to non-linear classification, situate it as a kernel machine, and finally conclude by discussing its advantages and disadvantages.

2.5.1 Hard Margin SVM

Consider the two-class supervised classification setting. Given a label set \( L \in \{-1,1\} \) and a set of observations \( X \subseteq \mathbb{R}^d \), assume there is a subset \( T \subseteq X \) of these observations of size \( \ell \) provided with labels. Formally, \( T = \{(x_i, y_i) | x_i \in X, y_i \in L\} \) for \( i = 1, \ldots, \ell \). This training dataset \( T \) contains observation-label pairs that we can use to learn the labels for the remaining observations in \( X \).

If the training data is linearly separable, a hyperplane exists that separates all the observations of one class from all of the observations of the other. This hyperplane can be written in the form of a linear discriminant function:

\[
f(x) = w^T x + b = 0
\]  

(2.1)

where \( w \) and \( b \) are the weights and bias that position the hyperplane. As a decision boundary, this hyperplane predicts the class of a new observation using the rule:

\[
y = \text{sign}(f(x)) = \begin{cases} 1 \text{ if } f(x) \geq 0 \\ -1 \text{ if } f(x) < 0 \end{cases}
\]

(2.2)
Without loss of generality, we can scale a separating hyperplane so that:

\[ y_i (w^T x_i + b) \geq 1 \text{ for } i = 1, \ldots, \ell \]  

(2.3)

Then, the distance from the hyperplane to the closest point of either class is \( 1/\|w\| \) and the distance between classes is \( 2/\|w\| \). Hence, to locate a separating hyperplane as far as possible from both classes, i.e. to maximize the margin to each class, we must minimize \( \|w\| \) subject to condition (2.3). This is equivalent to the following optimization problem:

\[
\min_w \frac{1}{2} \|w\|^2 \text{ subject to } y_i (w^T x_i + b) \geq 1 \text{ for } i = 1, \ldots, \ell
\]  

(2.4)

By introducing a Lagrangian multiplier \( \alpha_i \) for each observation in the training data, we obtain the primal form of the function to be minimized:

\[
L(w, b) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{\ell} \alpha_i \{1 - y_i (w^T x_i + b)\} \text{ for } i = 1, \ldots, \ell
\]  

(2.5)

After minimizing this function with respect to \( w \) and \( b \), the goal becomes to maximize its dual form:

\[
Q(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \alpha_i \sum_{j=1}^{\ell} \alpha_j y_i y_j x_i^T x_j
\]

subject to \( \sum_{i=1}^{\ell} \alpha_i y_i = 0 \) and \( \alpha_i \geq 0 \) for \( i = 1, \ldots, \ell \)  

(2.6)

This quadratic optimization problem can be solved using standard techniques. The solution to this problem has the property that \( \alpha_i > 0 \) only for the portion of observations that are near the decision boundary. These observations are known as the support vectors. Using this information, the primal variables \( w \) and \( b \) are calculated as:

\[
w = \sum_{i=1}^{\ell} \alpha_i y_i x_i \text{ and } b = \frac{1}{y_j} - \sum_{i=1}^{\ell} \alpha_i x_i^T x_j \text{ for each observation } j
\]  

(2.7)
Although theoretically the bias should be exactly the same for all observations \( j \), it is often averaged in practice due to numerical imprecision in the calculation of the dual variables satisfying (2.6).

2.5.2 Soft Margin SVM

Remote sensing data is typically linearly inseparable, which means that no hyperplane exists that can separate all the observations from one class from all the observations of the other. When this is the case, the hard margin condition of (2.3) is violated and a soft margin condition must be used instead. This relaxation is accomplished through the introduction of slack variables \( \xi_i \):

\[
y_i (w^T x_i + b) \geq 1 - \xi_i \text{ where } \xi_i \geq 0 \text{ for } i = 1, \ldots, \ell
\]

The effect of these slack variables is to allow a limited number of misclassifications to result from the placement of the separating hyperplane. Under this formulation, the optimization problem to solve is:

\[
\min_{w} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{\ell} \xi_i \text{ subject to } y_i (w^T x_i + b) \geq 1 - \xi_i \text{ for } i = 1, \ldots, \ell
\]

(2.9)

It turns out that, in going from the primal to dual forms for this problem, the slack variables are eliminated and only the final constraint in (2.6) is altered. The constraint \( \alpha_i \geq 0 \) becomes \( C \geq \alpha_i \geq 0 \). In other words, the feasible region for the solution of the dual problem is limited from the positive orthant to a finite area defined by a box constraint.

The parameter \( C \) then controls the size of this area and thus the cost of misclassification. It must be well chosen either by the analyst or by an automated method such as a grid search to maximize
the performance of the classifier.

2.5.3 Non-Linear SVM

Even though the use of a soft margin allows the margin of the separating hyperplane to be maximized in the linearly inseparable case, the accuracy of prediction will not be high if a linear boundary is not an appropriate separator for the data. In such a case, a non-linear decision boundary is preferred. To perform non-linear classification, SVM employ a powerful technique called the kernel trick, which allows a linear decision to be made in a transformed feature space. This linear decision then corresponds to a non-linear decision in the data’s original space.

The kernel trick involves replacing inner products in the SVM classification model with a non-linear mapping of similar form. The goal of the non-linear mapping is to map the data from its original space to a new space where it gains a linear structure. Typically, the space is higher dimensional because, if that space is not densely populated, the problem is more likely to be linearly separable (Cover, 1965). Supposing $\phi: x \in \mathbb{R}^d \mapsto \phi(x) \in F \subseteq \mathbb{R}^D$ is such a map, the soft margin formulation of the dual problem in (2.6) becomes:

$$
Q(\alpha) = \sum_{i=1}^\ell \alpha_i - \frac{1}{2} \sum_{i=1}^\ell \sum_{j=1}^\ell \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)
$$

subject to $\sum_{i=1}^\ell \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$ for $i = 1, ..., \ell$

(2.10)

And so the primal solution has form:

$$
\mathbf{w} = \sum_{i=1}^\ell \alpha_i y_i \phi(x_i) \text{ and } b = \frac{1}{y_j} - \sum_{i=1}^\ell \alpha_i \phi(x_i)^T \phi(x_j)
$$

(2.11)

for each observation $j$
with decision function:

\[ y = \text{sign}(f(x)) = \text{sign}(w^T x + b) = \text{sign}(\sum_{i=1}^{\ell} \alpha_i y_i \phi(x_i)^T \phi(x) + b) \] (2.12)

In practice, there is no need to specify the embedding \( \phi \) explicitly. For prediction, it is sufficient that, given a function \( k: X \times X \rightarrow \mathbb{R} \), the so-called Gram matrix \( [K]_{i,j} = k(x_i, x_j) \) is positive semi-definite (PSD), i.e. that the condition \( z^T K z \geq 0 \) holds for all \( z \). When this holds for all matrices \( K \) computed using \( k \), it is known that some embedding \( \phi \) exists such that the function \( k \) can be decomposed as \( k(x_i, x) = \phi(x_i)^T \phi(x) \) (Shawe-Taylor & Cristianini, 2004). The implication of this is that arbitrary kernel functions can be constructed for use in the SVM based on the knowledge one has of the classification problem.

**2.5.4 Kernel Methods**

Different kernel functions have been employed in remote sensing literature. The most commonly used kernel is the radial basis function (RBF) kernel:

\[ k(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \text{ where } \gamma > 0 \] (2.13)

which is known to map the data into an infinite-dimensional feature space. When \( \gamma \) is parameterized as \( 1/(2\sigma^2) \), this kernel is called the Gaussian RBF kernel. Other kernels used with remote sensing data include the linear, polynomial, and sigmoid kernels.

Table 2 below gives examples of these kernels.
**Table 2. Kernels employed with remote sensing data.**

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Form</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>( k(x_i, x) = x_i^T x )</td>
<td></td>
</tr>
<tr>
<td>RBF</td>
<td>( k(x_i, x) = \exp(-\gamma |x_i - x|^2) )</td>
<td>( \gamma )</td>
</tr>
<tr>
<td>Polynomial</td>
<td>( k(x_i, x) = (\gamma x_i^T x + \delta)^p )</td>
<td>( \gamma, \delta, p )</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>( k(x_i, x) = \tanh(\gamma x_i^T x + \delta) )</td>
<td>( \gamma, \delta )</td>
</tr>
</tbody>
</table>

In general, the choice of kernel determines the kinds of relations that can be learned from the data. Before a SVM can be trained, any parameter included in the kernel function, such as \( \gamma \) in the RBF kernel, must be chosen by the analyst. This choice must be made in addition to the choice of soft margin parameter \( C \). Commonly, a cross-validated grid search is employed to automatically search for the best combination of parameters during training.

Kernels can also be constructed as a combination of existing kernels. If \( k_1 \) and \( k_2 \) are kernels over the same space, \( A \) is a positive semi-definite \( d \times d \) matrix, \( f(\cdot) \) is a real-valued function, \( p(\cdot) \) is a polynomial with positive coefficients, and \( a > 0 \), the kernels listed below are also valid (Shawe-Taylor & Cristianini, 2004):

- \( k(x_i, x) = ak(x_i, x) \)
- \( k(x_i, x) = k_1(x_i, x) + k_2(x_i, x) \)
- \( k(x_i, x) = k_1(x_i, x)k_2(x_i, x) \)
- \( k(x_i, x) = f(x_i)f(x) \)
- \( k(x_i, x) = x_i^T Ax \)
- \( k(x_i, x) = p(k_1(x_i, x)) \)
- \( k(x_i, x) = \exp(k_1(x_i, x)) \)
In particular, the first two properties show that a new kernel can be constructed as a linear combination of existing kernels.

2.5.5 Advantages

Classification using SVMs offers a key few advantages to remote sensing studies. The first is that, as a kernel method, SVMs make efficient use of high-dimensional data – such as that supplied by hyperspectral sensors. Feature reductions are not required prior to classification, although the discarding of redundant or non-informative features can still have a positive effect on classification results. The second is that, again as a kernel method, SVMs allow the incorporation of novel feature mappings that may include contextual, expert, or ancillary information about a remote sensing problem. This information can help better discriminate among land-cover classes by increasing their separability in a transformed space. The third is that, because SVM maximize the geometric margin related to a separating hyperplane, they minimize the upper bound on the expected risk of classification. This theoretical equivalence guarantees their great generalization performance in a range of classification situations. Finally, due to their minimization of structural risk, SVM remain effective even with limited training set sizes, which are commonly encountered in remote sensing.

2.5.6 Disadvantages

One disadvantage of SVMs is that an appropriate multiclass classification strategy must be used to extend them from the native two-class classification.

Two commonly used strategies are One-Against-One (OAO) and One-Against-All (OAA)
approaches, which employ series of two-class subproblems to determine a multi-class solution. In OAO, $k$ classes are divided into $\binom{k}{2} = \frac{1}{2}k(k - 1)$ two-class subproblems. For each observation for each subproblem, an SVM is trained and a prediction is made. Hence, across subproblems, the majority prediction can be chosen as the class for the given observation. This strategy is effective and typically fast despite the quadratic number of subproblems since each problem only uses a small portion of the training data. However, this strategy can be expensive when non-traditional kernels are used for classification because of the quadratic number of kernel matrices that must be computed for classification. In OAA, $k$ classes form $k$ two-class subproblems, where the class of interest is pitted against the rest of the classes. For each observation for each subproblem, a prediction made based on an SVM trained on that subproblem. The class with the largest decision value is chosen as the class for an observation. This strategy can also be effective, but is typically slower despite the linear number of subproblems since each problem includes all the training data. Subproblems are also unbalanced, which can further affect training time and accuracy.

Another disadvantage of SVMs is that they offer no native probabilistic interpretation of classification results. Strategies such as the pairwise coupling of predictions from two-class subproblems must be used to create artificial probability estimates. The basic idea of pairwise coupling is to create multi-class probability estimates by combining all pairwise class probabilities. Methods have been proposed by Refregier and Vallet (1991); Price, Knerr, Personnaz, & Dreyfus (1995); Hastie and Tibshirani (1998); and Wu, Lin, and Weng (2004). The method of Wu, Lin, and Weng (2004) is commonly employed with SVM due to its implementation in the popular LIBSVM software package (Chang & Lin, 2011).
2.6 Markov Random Fields

2.6.1 Theory

Suppose $J = \{i | i = (r_i, c_i)\}$ is a regular lattice with sites $i \in J$ indexed by their rows $r_i$ and columns $c_i$. Let $X = \{x_i \in \mathbb{R}^d | i \in J\}$ be a set of observations at the sites in $J$. Here we will consider $X$ to be a satellite image on the lattice $J$ so that each vector $x_i$ consists of measurements of radiance in the sensor’s instantaneous field of view among different bands of wavelengths in the electromagnetic spectrum.

Define a configuration $Y$ of labels for the set $J$ by $Y = \{x_i = y_i | i \in J\}$ where $y_i \in L$ is a label drawn from a label set $L$, which here is the set of land-cover types known to be present in a satellite image. The configuration $Y$ is called a classification for the image $X$. The label at each site $i$ then corresponds to a realization of a random variable $F_i$ which takes values from $L$. We call the field $F = \{F_i | i \in J\}$ a Markov random field (MRF) if the following two conditions hold:

1. **Positivity**: $P(Y) > 0$ for all possible configurations $Y$

2. **Markovanity**: $P(y_i | y_{\partial i}) = P(y_i | y_{J - i})$ for each $i \in J$ and neighboring sites $\partial i$ for $i$

In a Bayesian framework, the optimum classification, or label configuration, for an image $X$ can be found using the maximum a posteriori (MAP) criterion:

$$Y_{opt} = \arg \max_{Y} P(Y|X)$$

(2.14)

which can be rewritten using Bayes’ theorem in terms of likelihood and the prior probability:

$$Y_{opt} = \arg \max_{Y} P(X|Y)P(Y)$$

(2.15)
Assuming class-conditional independence of observations, the likelihood can be rewritten as

\[ P(X|Y) = \prod_{i\in I} p(x_i|y_i) \]

where \( p(\cdot) \) is the marginal likelihood function at pixel \( i \). This suggests one strategy to obtain the MAP estimate is to maximize of the joint posterior probability over all pixels simultaneously. However, in general, it will not be feasible to do this due the computational complexity of the problem, so an alternate strategy must be used. One common choice is the Iterated Conditional Modes (ICM) algorithm, which maximizes the conditional posterior probability of each pixel sequentially (Besag, 1986). Other energy minimization strategies include simulated annealing, maximizer of posterior marginals, graph cuts, loopy-belief propagation, and tree-reweighted message passing (Szeliski et al., 2006; Tso & Mather, 2009). In remote sensing, ICM is a common choice because of its good performance compared to its computational cost.

Hence, by instead applying the Markovanity condition to (2.14), we obtain:

\[
y_{i,\text{opt}} = \arg \max_{y_i} \left\{ p(y_i|x_i, y_{\partial i}) \right\} = \arg \max_{y_i} \left\{ p(x_i|y_i)p(y_i|y_{\partial i}) \right\}
\]

(2.16)

so that the probability may be maximized for each pixel in turn rather than for all jointly.

By the Hammersley-Clifford theorem, a unique Gibbs random field exists for every MRF as long as the GRF is defined in terms of cliques in a neighborhood system (Hammersley and Clifford, 1971). Cliques are complete subgraphs of nested sizes, which constitute the relationships able to be captured in a local neighborhood. The GRF is a global model with probability distribution function:
\[ P(Y) = \frac{1}{Z} \exp \left( -\frac{1}{T} U(Y) \right) \] (2.17)

where \( U(Y) \) is an energy function, \( T \) is a constant, and \( Z = \sum e^{-\frac{1}{T} U(Y)} \) is the partition function, which is summed over all possible configurations \( Y \). Hence, maximizing the conditional posterior probability \( p(y_i|x_i, y_{\partial i}) \) of each pixel sequentially in (2.16) is equivalent to minimizing the energy \( U(y_i|y_{\partial i}) \) in the following:

\[ p(y_i|y_{\partial i}) = \frac{1}{Z_i} \exp(-U(y_i|y_{\partial i})) \] (2.18)

The energy \( U(y_i|y_{\partial i}) \) is expressed as a sum of clique potentials \( U(y_i|y_{\partial i}) = \sum_{c \in C} V_c(y_i) \) and different models for the potential \( V_c \) may be assumed. Commonly, assuming homogeneity and isotropy of the random field, a spatial energy is derived:

\[ U_{\text{SPATIAL}}(y_i|y_{\partial i}) = -\beta \sum_{j \in \partial i} I(y_i = y_j) \] (2.19)

where \( \beta \) is a parameter controlling the relative contribution of the spatial energy to the total energy and \( I(a = b) \) is an indicator function (equal to 1 when \( a = b \) and 0 otherwise). The parameter \( \beta \) may be estimated by various methods including maximum likelihood (Descombes, Morris, Zerubia, & Berthod, 1999), statistical and numerical procedures (Melgani & Serpico, 2003; Serpico & Moser, 2006), and genetic algorithms (Tso & Mather, 1999).

By considered the temporal neighbors of a pixel in a multi-temporal classification, temporal energy may also be derived:
\[ U_{TEMPORAL}(y_i|y_{\partial i}) = -\beta_1 \sum_{j \in \partial_p(i)} p(y_i|y_j) - \beta_2 \sum_{j \in \partial_f(i)} p(y_j|y_i) \] (2.20)

where \( \partial_p(i) \) indicates the past temporal neighbors of a pixel \( i \) and \( \partial_f(i) \) indicates its future temporal ones. The transition probabilities \( p(y_i|y_j) \) and \( p(y_j|y_i) \) can be estimated empirically for each target map using the temporal sequence of classification maps (Melgani & Serpico, 2003; Liu, Song, Townshend, & Gong, 2008).

To model spectral energy, one strategy is to assume the marginal likelihood \( p(x_i|y_i) \) for a pixel can be modeled under class-conditional independence as multivariate Gaussian where \( p(x_i|k) \sim \mathcal{N}(\mu_k, \Sigma_k) \) for each class \( k \in L \). Then, for example, when spatial energy is considered, the optimal classification of a pixel in terms of its total energy becomes:

\[ y_{i, opt} = \arg \min_k \left\{ -1/2 \ln |\Sigma_k| - 1/2 (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + U(k|y_{\partial i}) \right\} \] (2.21)

However, this model for the spectral energy may produce poor results in situations where classes have very similar spectral properties or the spectrum of a class is not well approximated by a Gaussian distribution. Consequently, a non-parametric approach to estimating the likelihood energy may perform better. When this approach is taken, the \( a \ posteriori \) probabilities \( p(y_i|x_i) \) are estimated instead after applying Bayes’ rule to \( p(x_i|y_i) \). In this case, the energy to be minimized is:

\[ y_{i, opt} = \arg \min_k \left\{ -\ln \frac{p(k|x_i)p(x_i)}{p(k)} + U(k|y_{\partial i}) \right\} \] (2.22)

The \( p(x_i) \) term can be assumed to be constant, as its distribution is modeled from the training
data, which remains fixed. Hence, by the product rule of logarithms, it becomes an additive constant in the total energy function and so it need not be estimated in terms of the minimization.

Both the original $-\ln(p(x_i|k))$ and the \textit{a posteriori} $-\ln(p(k|x_i)/p(k))$ model for the spectral energy is associated with a base classifier, which must be used to specify an initial label configuration from which to derive contextual energies in the first step of ICM. For the former, this configuration can be generated using the MAP estimate and training data. For the latter, the configuration may be generated using other classifiers, including non-probabilistic alternatives such as the support vector machine. In that case, when there is no explicit form for the \textit{a posteriori} probability associated with the classifier, general procedures such as pairwise probability coupling may be used. In either case, a well-modeled class conditional density or highly accurate classifier is the best choice to generate the initial label configuration because it increases the chance of finding at least a good locally optimal energy minimum solution for the MRF.

For a complete treatment of MRF, the reader is referred to Koller and Friedman (2009) or Li (2009).

2.6.2 Advantages

The major advantage of using MRF models for image context is their full characterization of and statistically rigorous use of contextual information at the pixel level. Such use does not require \textit{ad hoc} reasoning about the patterns of contextual influence in a study area because such patterns can be learned through the model itself. Additionally, MRF are flexible enough to incorporate a wide variety of contextual information sources (Solberg \textit{et al.}, 1996; Melgani & Serpico, 2003),
including expert knowledge (Liu & Cai, 2012), adapt to non-stationary change processes (Liu et al., 2008; Zhang et al., 2011), and be applied in a variety of settings (Moser, Serpico, & Benediktsson, 2013). Additionally, MRF are largely automatable, which removes the need for great skill on the part of an analyst as compared to OBIA methods provided an appropriate model has been selected. Altogether, MRF have shown consistently high performance on a range of imagery types and among a variety of situations for the contextual classification task.

2.6.3 Disadvantages

MRF are not without their disadvantages. Chiefly, their classification results are highly dependent on the accuracy of an initial non-contextual classification. Without the specification of a good initial configuration, energy minimization techniques may reach a poor local optimum. Additionally, they require that the likelihood probability \( p(x_i | y_i) \) which may not be well-estimated or able to be estimated by parametric method given the character of remote sensing data. When a non-parametric method is used to estimate this likelihood, procedures such as kernel density estimation or pairwise coupling must be relied on. The artificial probability estimates generated in the latter are not guaranteed to be appropriate, while the former is subject to the choice of bandwidth. Finally, the MRF decision rule is applied linearly in the space of contextual features, but non-linearity or margin-maximization of this decision rule could improve its application.

2.7 Conclusion

In remote sensing literature, pairwise coupled SVM are the top-performing method used to initialize contextual classifiers utilizing MRF models and to provide estimates of the likelihood
energy needed during energy minimization. However, the two techniques remain separated in this classification approach despite their respective and complementary advantages. The integration of the two classifiers would thus offer a powerful method for improving the contextual classification of land cover as well as the products available for the analysis of land-cover change.
Chapter 3: Methods

In this chapter, we first review the spatial contextual kernel method recently proposed in literature that forms the basis for the development of our spatial-temporal kernel. Second, we remark on the generalization of this kernel for use with a common class of MRF energy functions. Third, we detail the extension of this kernel to spatial-temporal context. Fourth, we construct a hierarchical classification algorithm utilizing this kernel to increase the tractability of its use for classification scenarios involving a large number of land-cover classes.

3.1 Review of the Markovian Kernel

We base the development of our spatial-temporal contextual kernel on the Markovian kernel method recently proposed by Moser and Serpico (2013). They suggest the use of this kernel to rigorously integrate the contextual information provided by the MRF approach with the SVM classifier. The following details the construction of this kernel for use in SVM.

Recall that, given a set of observations $X = \{x_i \in \mathbb{R}^d | i = 1, \ldots, n\}$ organized on an image lattice, the optimal configuration of labels $Y = \{y_i | i = 1, \ldots, n\}$ that classifies these observations minimizes the total energy function for an MRF model that is a combination of likelihood and prior energies. This function has form:

$$U_i(y_i | x_i, y_{\neq i}) = -\ln p(x_i | y_i) + \beta \mathcal{E}_i(y_i | y_{\neq i})$$  \hspace{1cm} (3.1)
where $\mathcal{E}_i$ is a local prior energy function and $\beta$ is a positive parameter controlling the weight given to the prior energy relative to the likelihood energy. Typically, to account for spatial context, we employ the Potts (or pairwise multilevel logistic model) $\mathcal{E}_i(y_i|y_{\partial i}) = -\sum_{j \in \partial i} I(y_i = y_j)$ as a smoothness prior (Tso & Mather, 2009).

When $y_i$ is one of two classes, e.g. $\{-1,1\}$, minimization of the total energy using the ICM algorithm can be formulated in terms of the following energy-difference function:

$$\Delta U_i(x_i, y_{\partial i}) = U_i(-1|x_i, y_{\partial i}) - U_i(1|x_i, y_{\partial i})$$ (3.2)

At each iteration of the minimization scheme, the pixel $i$ is then assigned a new label using the following rule based on the updated likelihood and prior energies:

$$\hat{y}_i = \text{sign}(\Delta U_i(x_i, y_{\partial i})) = \begin{cases} 1, & \Delta U_i \geq 0 \\ -1, & \Delta U_i < 0 \end{cases}$$ (3.3)

In Moser and Serpico (2013), it is proved that the energy difference function in (3.2) can be expressed as the following kernel expansion:

$$\Delta U_i(x_i, y_{\partial i}) = \sum_{j=1}^{\ell} \alpha_j y_j K_{MRF}(x_i, \varepsilon_i; x_j, \varepsilon_j) + b$$ (3.4)

where $K_{MRF}$ is a so-called Markovian kernel depending on the spectral information $x_i$ and $x_j$ and corresponding contextual features $\varepsilon_i$ and $\varepsilon_j$:

$$K_{MRF}(x_i, \varepsilon_i; x_j, \varepsilon_j) = K(x_i, x_j) + \gamma \varepsilon_i \varepsilon_j$$ (3.5)

given $\gamma > 0$ and $\varepsilon_i = \mathcal{E}_i(-1|y_{\partial i}) - \mathcal{E}_i(1|y_{\partial i})$
The parameter $\gamma$ controls the influence of the contextual features relative to that of a base kernel $K$. Its role is similar to the $\beta$ parameter used in the previously discussed MRF models. The dual variables $\alpha_j$ and bias $b$ specify a separating hyperplane in the transformed feature space induced by the Markovian kernel. They may be computed by training a SVM utilizing this kernel once the smoothing parameter $\gamma$ and the contextual features $\epsilon_i$ have been provided.

Under the ICM energy minimization scheme, the classification of a pixel at each iteration is thus predicted by the application of a Markovian SVM (MSVM). Because the values of the contextual features $\epsilon_i$ in each iteration depend on the previous iteration’s label configuration, the MSVM must be retrained at each iteration, which can result in the selection of new support vectors and movement of the separating hyperplane based on the influence of the contextual information.

Note that the proof of the kernel expansion, as provided in the appendices of Moser and Serpico (2013), relies on the assumption that the spectral data for each class in the transformed feature space induced by $K$ is Gaussian with different means but the same covariance matrix (or the equivalent functions in the infinite-dimensional case). As the authors note, this assumption is not expected to be restrictive because of the non-linearity of the mapping typically employed for the base kernel as well as the high dimensionality of the resulting feature space.

### 3.2 Generalized Markovian Kernel

Suppose the prior energy function of an MRF model is a linear combination of $\kappa$ local energy functions:
\[
U_i \left( y_i \mid x_i, \{ y_{\partial_k i} \}_{k} \right) = -\ln p(x_i \mid y_i) + \sum_{k=1}^{K} \beta_k \epsilon_{i,k}(y_i \mid y_{\partial_k i})
\]  
(3.6)

where \( \beta_k \) and \( \partial_k \) are, respectively, the parameter and local neighborhood associated with the \( k \)th function \( \mathcal{E}_{i,k} \).

Then, if \( y_i \in \{-1, 1\} \), the energy difference function can be expressed as:

\[
\Delta U_i \left( x_i, \{ y_{\partial_k i} \}_{k} \right) = \Delta U_i^{\text{likelihood}}(x_i) + \sum_{k} \Delta U_{i,k}^{\text{prior}}(y_{\partial_k i})
\]

where \( \Delta U_i^{\text{likelihood}}(x_i) = -\ln p(x_i \mid -1) + \ln p(x_i \mid 1) \)

(3.7)

where \( \Delta U_{i,k}^{\text{prior}}(y_{\partial_k i}) = \beta_k \left[ \epsilon_{i,k}(-1 \mid y_{\partial_k i}) - \epsilon_{i,k}(1 \mid y_{\partial_k i}) \right] \)

Hence, each prior energy difference is associated with a contextual feature:

\[
\epsilon_{i,k} = \epsilon_{i,k}(-1 \mid y_{\partial_k i}) - \epsilon_{i,k}(1 \mid y_{\partial_k i})
\]

(3.8)

Thus, the Generalized Markovian Kernel is the sum of a base kernel and a linear combination of contextual features:

\[
K_{GMRF} \left( x_i, \{ \epsilon_{i,k} \}_{k}; x_j, \{ \epsilon_{j,k} \}_{k} \right) = K(x_i, x_j) + \sum_{k=1}^{K} \gamma_k \epsilon_{i,k} \epsilon_{j,k}
\]

(3.9)

The same arguments as presented for the proof of the spatial kernel in Moser and Serpico (2013) follow for this generalized version. Of particular note is that the transformed feature space induced by the base kernel is now augmented with multiple additional features \( \{ \epsilon_k \}_{k=1}^{K} \).
3.3 Proposed Spatial-Temporal Kernel

We base our spatial-temporal kernel on the MRF model employed in Liu and Cai (2012).

Suppose we have a time series of remote sensing images at \( T \geq 2 \) different dates that are precisely coregistered with each other such that a time series of observations is formed with respect to the image lattice. Denote this set of images as \( \mathbf{X} = \{ X_t | t = 1, \ldots, T \} \) and the set of observations for each image as \( X_t = \{ x_{t,i} \in \mathbb{R}^d | i = 1, \ldots, n \} \).

In this model, the optimal configuration of labels \( Y_t = \{ y_{t,i} | i = 1, \ldots, n \} \) for an image \( X_t \) minimizes the total energy function at each pixel:

\[
U_{t,i} \left( y_{t,i} \big| x_{t,i}, \{ y_{\partial_k(t,i)} \}_k \right) = -\ln p(x_{t,i} | y_{t,i}) + \sum_{k=1}^K \beta_{t,k} \mathcal{E}_{t,i,k} \left( y_{t,i} | y_{\partial_k(t,i)} \right)
\]

with local contributions:

\[
\mathcal{E}_{t,i,1} \left( y_{t,i} | y_{\partial_1(t,i)} \right) = -\sum_{j \in \partial_1(t,i)} I(y_{t,i} = y_j) \quad \text{spatial energy} \quad (3.10)
\]

\[
\mathcal{E}_{t,i,2} \left( y_{t,i} | y_{\partial_2(t,i)} \right) = -\sum_{j \in \partial_2(t,i)} p(y_{t,i} | y_j) \quad \text{past temporal energy} \quad (3.11)
\]

\[
\mathcal{E}_{t,i,3} \left( y_{t,i} | y_{\partial_3(t,i)} \right) = -\sum_{j \in \partial_3(t,i)} p(y_j | y_{t,i}) \quad \text{future temporal energy} \quad (3.12)
\]

\[
\mathcal{E}_{t,i,4} \left( y_{t,i} | y_{\partial_4(t,i)} \right) = \sum_{j \in \partial_4(t,i)} I(y_j \neq y_{t,i}) \quad \text{past temporal penalty} \quad (3.13)
\]

\[
\mathcal{E}_{t,i,5} \left( y_{t,i} | y_{\partial_5(t,i)} \right) = \sum_{j \in \partial_5(t,i)} I(y_{t,i} \neq y_j) \quad \text{future temporal penalty} \quad (3.14)
\]
Under this model, the classification of a pixel at date \( t \) and location \( i \) is determined by a combination of its likelihood energy with the prior energy contributions of the pixels in its spatial-temporal neighborhoods \( \{ \partial_k \}_{k=1}^K \). Consequently, \( \partial_1 \) denotes the indices of the spatial neighbors at date \( t \), \( \partial_2 \) and \( \partial_4 \) the indices of the past temporal neighbors at \( t - 1 \), and \( \partial_3 \) and \( \partial_5 \) the indices of the future temporal neighbors at \( t + 1 \). Given the spatial resolution of our usual data sources, we employ a queen-contiguous neighborhood system in all cases, which includes either eight spatial or nine spatial-temporal neighbors excepting for pixels located on the image edges. Figure 1 summarizes this neighborhood system.

![Multi-temporal neighborhood system](image)

**Figure 1.** Multi-temporal neighborhood system.

Hence, at date \( t \), the function \( I(y_{t,i} = y_j) \) indicates whether the label at pixel \( i \) is the same as the label at its neighbor \( j \). As well, \( p(y_{t,i}|y_j) \) and \( I(y_j \neq y_{t,i}) \) are the transition probability and illogical transition indicator for a pixel \( i \) at date \( t \) given its past temporal neighbor \( j \) at date \( t - 1 \), while \( p(y_j|y_{t,i}) \) and \( I(y_{t,i} \neq y_j) \) are similar for the future temporal neighbor \( j \) at date \( t + 1 \) given a pixel \( i \) at date \( t \). Before this MRF model can be applied, transition probabilities and illogical transition definitions must be supplied.
The local energy functions (3.10) through (3.14) are each associated with a contextual feature arising from the energy difference function in (3.7). These contextual features are listed below:

\[
\begin{align*}
\epsilon_{t,i,1} &= \mathcal{E}_{t,i,1}(-1|y_{\partial_1(t,i)}) - \mathcal{E}_{t,i,1}(1|y_{\partial_1(t,i)}) & \text{spatial} \\
\epsilon_{t,i,2} &= \mathcal{E}_{t,i,2}(-1|y_{\partial_2(t,i)}) - \mathcal{E}_{t,i,2}(1|y_{\partial_2(t,i)}) & \text{past temporal} \\
\epsilon_{t,i,3} &= \mathcal{E}_{t,i,3}(-1|y_{\partial_3(t,i)}) - \mathcal{E}_{t,i,3}(1|y_{\partial_3(t,i)}) & \text{future temporal} \\
\epsilon_{t,i,4} &= \mathcal{E}_{t,i,4}(-1|y_{\partial_4(t,i)}) - \mathcal{E}_{t,i,4}(1|y_{\partial_4(t,i)}) & \text{past temporal penalty} \\
\epsilon_{t,i,5} &= \mathcal{E}_{t,i,5}(-1|y_{\partial_5(t,i)}) - \mathcal{E}_{t,i,5}(1|y_{\partial_5(t,i)}) & \text{future temporal penalty}
\end{align*}
\]

Thus, noting the dependence on \( t \) of all variables, but dropping it from the notation for brevity, the Spatial-Temporal Markovian Kernel associated with this MRF model is:

\[
K_{STP}(x_i, \epsilon_i; x_j, \epsilon_j|\gamma, \theta) = \gamma_0 K(x_i, x_j|\theta) + \gamma_1 \epsilon_{i,1} \epsilon_{j,1} + \gamma_2 \epsilon_{i,2} \epsilon_{j,2} + \gamma_3 \epsilon_{i,3} \epsilon_{j,3} + \gamma_4 \epsilon_{i,4} \epsilon_{j,4} + \gamma_5 \epsilon_{i,5} \epsilon_{j,5}
\]

To classify using this kernel, the \( \gamma \) parameters, the transition probabilities and illogical transition definitions associated with the local energy functions \( \mathcal{E}_{i,2} \) through \( \mathcal{E}_{i,5} \), and the base kernel \( K \) and any parameters \( \theta \) it employs must be supplied. Note that we include a parameter \( \gamma_0 \) among \( \gamma \) to weight the base kernel \( K \) for later computational convenience in determining the optimal combination of base kernel with contextual features.

This model uses a mutual approach to sharing contextual information. Hence, the MRF model and, consequently, the Markovian kernel used for the classification of each image should be
updated for all images simultaneously during each iteration based on the label configurations generated in the previous iteration.

For the comparison of models, we will also define a kernel that does not include temporal penalties:

$$K_{ST}(x_i, e_i; x_j, e_j | y, \theta) = \gamma_0 K(x_i, x_j | \theta) + \gamma_1 e_{i,1} e_{j,1} + \gamma_2 e_{i,2} e_{j,2} + \gamma_3 e_{i,3} e_{j,3}$$  (3.21)

In the remainder of the thesis, we use the moniker STP versus ST to denote the presence or absence of penalties, respectively.

### 3.4 Proposed Spatial-Temporal Classifier

In the following, single superscripts mark quantities computed at the $r^{th}$ iteration, while double superscripts mark quantities computed at the $r^{th}$ iteration in the $p^{th}$ subproblem. However, note that the base kernel parameters $\theta_t$, misclassification costs $C_t$, and kernel parameters $y_t$, like the training set $T_t$, are held fixed over all iterations, but not over subproblems.

#### 3.4.1 Two-class Classification

For each image $X_t \in X$, suppose we have a training dataset $T_t = \{(x, y) | x \in X_t, y \in L\}$ of observation-label pairs such that $T_t \subseteq X_t$ and the label set $L = \{-1, 1\}$ contains two classes. Without loss of generality, in the case that two-class label sets use other class codes, we can recode the labels to $\{-1, 1\}$ to apply the following algorithm.

In the initialization phase, for each image $X_t$, select the optimal combination of base kernel
parameters $\hat{\theta}_t$ and misclassification cost $\hat{C}_t$ that provide the highest training accuracy for an SVM trained with a chosen base kernel $K(\cdot | \hat{\theta}_t)$ over the training data $T_t$. Based on this trained model, generate a non-contextual classification $\hat{Y}_t^0$ for each image and construct contextual features $\{e_{t,i}^0\}_{i=1}^n$ based on that classification. Estimate the kernel parameters $\hat{y}_t$ for the Markovian kernel $K_{STP}(y_t | x_{t,i}, e_{t,i}^0; x_{t,j}, e_{t,j}^0)\hat{\theta}_t$ where $x_{t,i}, x_{t,j} \in T_t$ via an appropriate technique, holding the base kernel parameters $\hat{\theta}_t$ and misclassification cost $\hat{C}_t$ fixed.

In the iterative phase, for each image $X_t$ at each iteration $r \geq 1$, calculate the Markovian kernel $K_{STP}(x_{t,i}, e_{t,i}^{r-1}; x_{t,j}, e_{t,j}^{r-1} | \hat{y}_t, \hat{\theta}_t)$ where $x_{t,i}, x_{t,j} \in T_t$ using the contextual features $\{e_{t,i}^{r-1}\}_{i=1}^n$ from the previous iteration and initial parameters $\hat{y}_t$ and $\hat{\theta}_t$. Train a MSVM using this kernel given the misclassification cost $\hat{C}_t$. Based on this trained model, generate a contextual classification $\hat{Y}_t^r$ for each image. If no stopping criteria have been met, generate updated contextual features $\{e_{t,i}^r\}_{i=1}^n$ based on the classification set $\hat{Y}_r$ and then start a new iteration. Otherwise, stop and output the final set $\hat{Y}_r$ of contextual classifications.

### 3.4.2 Multi-class Classification

In the case that the label set $L$ contains more than two classes, the algorithm must be modified to solve a series of two-class subproblems. As suggested in Moser and Serpico (2013), the OAO multi-class decomposition strategy may be used. Under this strategy, an SVM or MSVM must be trained for each pair of classes in the label set at each date $t$ in each iteration $r$.

For each image $X_t \in X$, suppose we have a training dataset $T_t = \{(x, y) | x \in X_t, y \in L\}$ of observation-label pairs such that $T_t \subseteq X_t$. Given the label set $L$, suppose $P$ is the set of possible
subproblems. Then, the number of subproblems is quadratic in the number of classes, i.e. \( |P| = \binom{L}{2} = \frac{1}{2}L(|L| - 1) \). Subset the training dataset by subproblem such that \( T^p_t = \{(x, y) \in T_t | y \in \pi_p, \pi_p \in P\} \) for each \( p = 1, \ldots, |P| \). For the purpose of classification and constructing contextual features, as in the two-class case, we can recode the subproblems labels given by each set \( \pi_p \) to \{-1, 1\} and apply the following algorithm.

In the initialization phase, for each image \( X_t \) and subproblem \( p \), select the optimal combination of base kernel parameters \( \hat{\theta}^p_t \) and misclassification cost \( \hat{C}^p_t \) that provide the highest training accuracy for an SVM trained with a chosen base kernel \( K(\cdot | \hat{\theta}^p_t) \) over the training data \( T^p_t \).

Based on this trained model, generate a non-contextual classification \( \hat{Y}^{0,p}_t \) for each image and construct contextual features \( \{\epsilon_{t,i}^{0,p}\}_{i=1}^n \) based on that classification. Estimate the kernel parameters \( \hat{\gamma}^p_t \) for the Markovian kernel \( K_{STP}(\gamma_t | x_{t,i}, \epsilon_{t,i}^{0,p}; x_{t,j}, \epsilon_{t,j}^{0,p}, \hat{\theta}^p_t) \) where \( x_{t,i}, x_{t,j} \in T^p_t \) via an appropriate technique, holding the base kernel parameters \( \hat{\theta}^p_t \) and misclassification cost \( \hat{C}^p_t \) fixed.

In the iterative phase, for each image \( X_t \) and subproblem \( p \) in each iteration \( r \geq 1 \), calculate the Markovian kernel \( K_{STP}(x_{t,i}, \epsilon_{t,i}^{r-1,p}; x_{t,j}, \epsilon_{t,j}^{r-1,p} | \hat{\gamma}^p_t, \hat{\theta}^p_t) \) where \( x_{t,i}, x_{t,j} \in T^p_t \) using the contextual features \( \{\epsilon_{t,i}^{r-1,p}\}_{i=1}^n \) from the previous iteration and initial parameters \( \hat{\gamma}^p_t \) and \( \hat{\theta}^p_t \). Train a MSVM using this kernel given the misclassification cost \( \hat{C}^p_t \). Based on this trained model, generate a contextual classification \( \hat{Y}^{r,p}_t \) for each image. If no stopping criteria have been met, generate updated contextual features \( \{\epsilon_{t,i}^{r,p}\}_{i=1}^n \) based on the classification set \( \hat{Y}^{r,p}_t \) and then start a new
iteration. Otherwise, stop and output the final contextual classification $\hat{Y}^r_t$ at each date $t$ by the rule $\hat{y}_{t,i} = \text{majority } \hat{y}_{t,i}^{r,p}$ for $t = 1, ..., T$ and $i = 1, ..., n$ where the values $\hat{y}_{t,i}^{r,p}$ have been recoded from the labels $\{-1, 1\}$ to their original class codes $\pi_p$.

In essence, a two-class MRF model is iteratively applied in each subproblem and coupled with the models of its temporal neighbors through the previous iteration’s subproblem maps. The final classification result is thus the majority vote among the MRF models over the subproblems at each image date.

Figure 2 shows the flow of information in this classifier focused on a target date $t$ over two iterations, corresponding to the non-contextual initialization step and the first contextual step. Edge dates only use the temporal information from their single temporal neighbors.
Figure 2. Schematic for use of contextual information in the proposed Spatial-Temporal Markovian Support Vector Classifier (ST MSVC).
3.4.3 Hierarchical Classification

Because the OAO multi-class strategy is quadratic in the number of classes, the computational burden of the proposed classifier can be substantial when the number of classes is large. Such a case will occur with many of the standard land-cover classification schemes defined by national and international organizations that are commonly employed in remote sensing. A hierarchical approach offers the ability to reduce this burden to linear in the number of classes. It also offers the possibility of constructing two-class subproblems that are well separated in feature space. Such subproblems will have simpler solutions that require fewer support vectors. The time taken to solve these subproblems will be shorter than the time taken to solve more complex ones due to the efficiency of SVM training algorithms. All subproblems in hierarchical classification are also relevant to and meaningful for the classification task compared with those of the one-versus-one strategy.

Here, we make use of a class-based large-margin clustering method by Cevikalp (2010) to generate a binary hierarchical decision tree. The method recursively employs the Normalized Cuts Clustering algorithm of Shi and Malik (2000) to find the optimal cut that maximally divides clusters of classes. The cuts are based on the weighted distances between the convex hulls of all pairs of classes in a kernel-transformed feature space. This makes the method well-suited for building a decision tree of SVMs since the cuts are determined for the data in the space of the SVM solution. The stated utility of the method is in reducing the testing time of an SVM classifier in the presence of a large number of classes, but the great reduction in number of subproblems for which SVMs must be trained also results in a significantly reduced training time for our proposed contextual classifier.
In the context of our proposed classifier, we require that the same hierarchy of classes be used for all images in the time series. Hence, we build this hierarchy on the first image in the series and then train a decision tree of SVMs for each image. In the initialization step, these decision trees are used to generate full classification maps for each subproblem, which in turn are used to construct the first set of contextual features. In the iterative steps, the contextual features of the previous iteration are used in MSVMs to predict updated classification maps for the subproblems at each image date. If no stopping criteria has been met, new features are then generated based on these updated maps and the MSVMs are retrained for the next iteration. Otherwise, the subproblem maps at each image date are hierarchically combined to form the final multiclass classifications.

Figure 3 shows the flow of information in this classifier focused on a target date $t$ over two iterations, corresponding to the non-contextual initialization step and the first contextual step. For diagrammatic simplicity, it is illustrated for a balanced four-class hierarchy, but any hierarchy can take its place. As previously, edge dates only use the temporal information from their single temporal neighbors.
Figure 3. Schematic for use of contextual information in the proposed Spatial-Temporal Hierarchical Markovian Support Vector Classifier (ST HMSVC).
3.4.4 Parameter Estimation

Estimation of the kernel parameters $\gamma_t$ is an important task because it determines the relative influences of the spectral and contextual features. In Moser and Serpico (2013), the authors propose the use of the Ho-Kashyap algorithm to estimate the parameters as the solution to a system of linear equalities. The method uses a minimum squared error procedure, which is proven to converge with linearly separable features and terminates otherwise (Serpico & Moser, 2006). Here, because the parameter estimation problem is formally similar to that of Multiple Kernel Learning, we employ the technique by Varma and Babu (2009). It estimates the kernel parameters through a projected gradient descent procedure subject to L2 regularization of the parameters. We choose it for its good performance with non-linear kernels (Gönen & Alpaydin, 2011) and use of an existing SVM solver.

In contrast to Moser and Serpico (2013), we estimate kernel parameters for each subproblem instead of estimating them jointly across all subproblems. We do so based on the increase in performance of the former over the latter that we found in our preliminary tests.

3.4.5 Transition Probabilities

Following Liu and Cai (2012), we estimate the past temporal $p(y_{t,i}|y_j)$ and future temporal $p(y_j|y_{t,i})$ transition probabilities of (3.11) and (3.12) at each iteration empirically based on the previous iteration’s subproblem maps. We compare the use of the Global Transition (GT) and Locally Adjusted Global Transition (LAT) models formulated in Liu et al. (2008).

Because our proposed classifier is formulated in terms of two-class subproblems, at each iteration $r$, each model is associated with a past temporal GT matrix of the form:
and a future temporal GT matrix:

\[
\mathcal{T}_{\text{future}}(\mathbf{\hat{y}}) = \begin{bmatrix}
    p(\mathbf{\hat{y}}_{t+1} = -1|\mathbf{\hat{y}}_t = -1) & p(\mathbf{\hat{y}}_{t+1} = 1|\mathbf{\hat{y}}_t = -1) \\
    p(\mathbf{\hat{y}}_{t+1} = -1|\mathbf{\hat{y}}_t = 1) & p(\mathbf{\hat{y}}_{t+1} = 1|\mathbf{\hat{y}}_t = 1)
\end{bmatrix}
\]  

where \(-1 \equiv l_j\) and \(-1 \equiv l_k\) for the labels \(l_j, l_k \in \pi_p\) for the given subproblem \(\pi_p\). Global refers to the fact that the probabilities are defined as the proportion of the total number of pixels of a class \(a\) at a date \(t_1\) that transition to a class \(b\) at date \(t_2\) where \(t_1 < t_2\). In other words:

\[
p(\mathbf{y}_{t_2}^P = b|\mathbf{y}_{t_1}^P = a) = \frac{\sum_i I(\mathbf{y}_{t_1}^P = a \text{ and } \mathbf{y}_{t_2}^P = b)}{\sum_i I(\mathbf{y}_{t_1}^P = a)}
\]

where \(i\) refers to the location of pixels on the image lattice for either the classification \(Y_{t_1}^P\) or \(Y_{t_2}^P\) and the sums are over all locations \(i\) in these classifications.

In the LAT model, the global transition probabilities are adjusted based on pixelwise probabilities of change under the following scheme:

\[
\mathcal{T}_{\text{local}}(i) = \begin{bmatrix}
    [\mathcal{T}]_{0,0}(\mathbf{\hat{y}}) \cdot (1-p(i)) & [\mathcal{T}]_{0,1}(\mathbf{\hat{y}}) \cdot p(i) \\
    [\mathcal{T}]_{1,0}(\mathbf{\hat{y}}) \cdot p(i) & [\mathcal{T}]_{1,1}(\mathbf{\hat{y}}) \cdot (1-p(i))
\end{bmatrix}
\]

where \([\mathcal{T}]_{j,k}\) refers to the entry in the \(j^{th}\) row and \(k^{th}\) column of the GT matrix \(\mathcal{T}\) and \(p(i)\) refers to the pixel-wise change probability. Note that the matrix must be row standardized to specify a
valid transition probability model.

By modifying the GT model in this way, the LAT model decreases the likelihood that spectrally similar pixels will change class, while increasing the likelihood those spectrally different pixels will change. This is reasonable because we assume that the temporal dependence between land cover should increase with the use of contextual information. In other words, we assume that the identification of stable land cover is dominated by errors of omission, whereas unstable land cover is dominated by errors of commission.

The pixelwise change probability \( p(i) \) is defined by fitting a logistical function to the difference between spectral bands at two image dates:

\[
p(i) = \frac{1}{1 - \exp \left( \frac{[x_{t_1,i}]_b - [x_{t_2,i}]_b}{b} \right)}
\]

(3.26)

where \( b \) is the band number, indexing the \( b^{th} \) component of the spectral feature vector \( x_{t,i} \) at location \( i \) and date \( t \).

We select the optimal band at each pair of adjacent dates for each transition type by searching for the difference band between dates with the highest standard deviation among the pixels that underwent the given transition in the last iteration. We fix our selections after the initialization phase of the algorithm. Since the subproblems of the hierarchical classifier include transitions from clusters of classes to other clusters (hereafter called superclasses), we take the majority vote among the bands corresponding to the possible individual transitions between these superclasses.
3.4.6 Illogical Transition Penalties

Following Liu and Cai (2012), we specify the past temporal $I(y_j \neq y_{t,i})$ and future temporal $I(y_{t,i} \neq y_j)$ illogical transition indicators of (3.13) and (3.14) in terms of an expert-provided illogical transition matrix. This matrix is comprehensive to all transitions of land cover occurring between adjacent dates and is assumed to be the same for all pairs of adjacent dates. It indicates the penalties associated with transitions that are considered illogical given understanding of the ecological rules at work in a study area or implausible given the time interval between image dates. In general, this matrix acts as a look-up table and has form:

$$
\begin{array}{cccc}
& l_1 & \ldots & l_{|L|} \\
\hline
l_1 & I(l_1 \neq l_2) & \ldots & I(l_1 \neq l_{|L|}) \\
\vdots & \vdots & & \vdots \\
l_{|L|} & I(l_{|L|} \neq l_2) & \ldots & I(l_{|L|} \neq l_{|L|}) \\
\end{array}
$$

where the indicator $I(l_j \neq l_k)$ is equal to 1 if the transition from class $l_j \in L$ at a date $t_1$ to class $l_k \in L$ at a date $t_2$ is illogical and equal to 0 otherwise. Hence, to evaluate the pixelwise indicator $I(y_j \neq y_{t,i})$ for the past temporal local penalty function in (3.13), we look up the penalty in the matrix given by (3.27) associated with transition from the label $y_j$ of a past temporal neighbor at date $t - 1$ to the candidate label $y_{t,i}$ for a pixel at date $t$ and location $i$. We do the similar look up for the future temporal penalty function defined in (3.14).

Because our proposed non-hierarchical classifier is formulated in terms of two-class subproblems, the illogical transition matrix associated with each subproblem is a submatrix. It is specified as:
To
\[
\begin{array}{c|c|c}
 l_j \equiv -1 & l(k) \equiv 1 \\
 I(l_j \neq l_j) & I(l_j \neq l(k)) \\
 I(l(k) \neq l_j) & I(l(k) \neq l(k))
\end{array}
\]
From
\[
\begin{array}{c|c|c}
 l_j \equiv -1 & l(k) \equiv 1 \\
 I(l_j \neq l_j) & I(l_j \neq l(k)) \\
 I(l(k) \neq l_j) & I(l(k) \neq l(k))
\end{array}
\]

where the labels \(l_j, l_k\) belong to the given subproblem \(\pi_p\).

For our proposed hierarchical classifier, this definition generally cannot be obtained as a submatrix of the standard illogical transition matrix due to the occurrence of superclasses among the subproblems. For example, for a three-class problem, the first split in the hierarchy must separate one class from the other two. The cluster of two classes constitutes a superclass for which the look up table of either (3.27) or (3.28) cannot be applied because of the fuzzy membership between individual classes for pixels assigned to this superclass. To apply an illogical penalty in this case, we perform a weighted aggregation of the penalties over individual classes constituting each superclass. The scheme we employ is:

\[
l_{t_1,t_2}(l_j \neq l_k) = \sum_{l_j \in l_j} \sum_{l_k \in l_k} w_{t_1,t_2}(l_j, l_k) \cdot I(l_j \neq l_k) \cdot \frac{1}{|l_j| \cdot |l_k|} \cdot \sum_{l_j \in l_j} \sum_{l_k \in l_k} w_{t_1,t_2}(l_j, l_k)
\]

where \(w_{t_1,t_2}(l_j, l_k) = \frac{1}{n} \sum_i I\left(y_{t_1,i}^{r-1,p(l_j)} = l_j \text{ and } y_{t_2,i}^{r-1,p(l_k)} = l_k\right)\)

where \(l_j\) and \(l_k\) denote a pair of superclasses of size \(|l_j|\) and \(|l_k|\), respectively, and the weight \(w_{t_1,t_2}(l_j, l_k)\) is the proportion of pixels in the change map undergoing a transition from class \(l_j\) in the land cover subproblem map \(Y_{t_1}^{r-1,p(l_j)}\) to class \(l_k\) in the land cover subproblem map \(Y_{t_1}^{r-1,p(l_k)}\). The notation \(p(l_j)\) or \(p(l_k)\) denotes the subproblem for which \(l_j\) or \(l_k\) are separated out as single classes.
Note that the original penalty is recovered in the degenerate case that \( l_j \) and \( l_k \) each consist of only one class.
Chapter 4: Case Study in Appalachian Ohio

4.1 Data

4.1.1 Study Area

We investigate the performance of the proposed classifier for multi-temporal mapping in an Appalachian Ohio study area. The study area is Ross County in southeastern Ohio, which is one of the five counties investigated in Liu and Cai (2012). It is approximately 2,700 square kilometers (1,042.5 square miles) in area. As explained in their paper, this county is dominated by agricultural land use with a smaller coverage of forest compared to the surrounding counties. Some mining activity has also taken place within the 1988 to 2008 time period investigated.

Liu and Cai (2012) obtained satellite images at five year intervals during this period from the Landsat 5 Thematic Mapper (TM) sensor. The images were determined to be accurately co-registered. They were subset to a 1500-by-2000 pixel scene that includes the entire county. Table 3 below lists the acquisition date of each image.

<table>
<thead>
<tr>
<th>Year</th>
<th>Satellite Sensor</th>
<th>Acquisition Date</th>
<th>Path/Row</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>Landsat 5 TM</td>
<td>15 June</td>
<td>19/33</td>
</tr>
<tr>
<td>1993</td>
<td>Landsat 5 TM</td>
<td>9 August</td>
<td>19/33</td>
</tr>
<tr>
<td>1998</td>
<td>Landsat 5 TM</td>
<td>13 July</td>
<td>19/33</td>
</tr>
<tr>
<td>2003</td>
<td>Landsat 5 TM</td>
<td>20 July</td>
<td>19/33</td>
</tr>
<tr>
<td>2008</td>
<td>Landsat 5 TM</td>
<td>17 July</td>
<td>19/33</td>
</tr>
</tbody>
</table>
Figure 4 below depicts subsets of the Ross County scene centered on the city of Chillicothe, Ohio. These subsets will be used later for visual comparison of classification results.

Figure 4. Subsets of satellite images for the Appalachian Ohio dataset and location of subset within Ross County scene. Spectral data is depicted with 7-5-3 band combination.
The Landsat 5 satellite flew from March 1, 1984 until deactivation on June 5, 2013. Its mission was a key part of the NASA Earth Observing System. The TM is a multispectral sensor with seven bands. The spectral and spatial characteristics of each of these bands are listed in Table 4 below.

Table 4. Summary of Landsat 5 TM spectral bands.

<table>
<thead>
<tr>
<th>Band</th>
<th>Name</th>
<th>Wavelengths (μm)</th>
<th>Spatial Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blue-Green</td>
<td>0.45 - 0.52</td>
<td>30 meters</td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>0.52 - 0.60</td>
<td>30 meters</td>
</tr>
<tr>
<td>3</td>
<td>Red</td>
<td>0.63 - 0.69</td>
<td>30 meters</td>
</tr>
<tr>
<td>4</td>
<td>Near Infrared</td>
<td>0.76 - 0.90</td>
<td>30 meters</td>
</tr>
<tr>
<td>5</td>
<td>Middle Infrared 1</td>
<td>1.55 - 1.75</td>
<td>30 meters</td>
</tr>
<tr>
<td>6</td>
<td>Thermal Infrared</td>
<td>10.4 - 12.5</td>
<td>120 meters</td>
</tr>
<tr>
<td>7</td>
<td>Middle Infrared 2</td>
<td>2.08 - 2.35</td>
<td>30 meters</td>
</tr>
</tbody>
</table>

Source: Chander et al., 2004

Data from this sensor is quantized at 8 bits.

4.1.2 Training and Validation Data

Liu and Cai (2012) developed a six class classification scheme that is comprehensive of the land-cover types observed at Landsat resolution in the Ross County study area. The scheme includes a mix of natural and human-constructed land-cover types, which are listed in Table 5 below.

Table 5. Classification scheme for Appalachian Ohio scenes.

<table>
<thead>
<tr>
<th>Name:</th>
<th>Water</th>
<th>Forest</th>
<th>Crop</th>
<th>Urban</th>
<th>Mine</th>
<th>Grass/Shrub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Code:</td>
<td>W</td>
<td>F</td>
<td>C</td>
<td>U</td>
<td>M</td>
<td>G</td>
</tr>
</tbody>
</table>

73
The dataset includes reference land-cover classes collected for a random sample of locations at each of the five image dates. The locations were manually classified by using field surveys, aerial photographs, mine permit maps, and topographic maps to interpret their land cover. For the full details, we refer the reader to the discussion in their paper.

The reference dataset was randomly split into training and validation subsets. The observation counts for each class in these subsets are listed in Table 6 below.

Table 6. Summary of reference data for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Year</th>
<th>Training Data</th>
<th>Validation Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td>F</td>
</tr>
<tr>
<td>1988</td>
<td>131</td>
<td>788</td>
</tr>
<tr>
<td>1993</td>
<td>134</td>
<td>736</td>
</tr>
<tr>
<td>1998</td>
<td>155</td>
<td>820</td>
</tr>
<tr>
<td>2003</td>
<td>155</td>
<td>836</td>
</tr>
<tr>
<td>2008</td>
<td>158</td>
<td>800</td>
</tr>
</tbody>
</table>

Based on their expert knowledge of the study area, Liu and Cai (2012) developed a definition of illogical land-cover transitions for the five-year time interval in this study area. The definition is replicated in the matrix in Table 7 below.
Table 7. Summary of illogical land-cover transitions for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>From</th>
<th>W</th>
<th>F</th>
<th>C</th>
<th>U</th>
<th>M</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The symbol X indicates the transition is illogical.

From this matrix, we see that primarily transition from the urban class and transition to the forest class are penalized. Transitions to and from the mine class are also penalized, reflecting the implausibility of reforestation and urbanization in mined areas as well as the separation of water and mine-related land cover. No restrictions are placed on the transition from forest to other land-cover types or the conversion to or from grass or shrub land.

4.2 Experiments

We compare the performance of four classifiers – the non-contextual SVM, the contextual MRF, and the proposed contextual MSVC and HMSVC – over three experiments.

Table 8 below lists the classifiers and their attributes.
In the first experiment, we compare groups 1 and 2 to assess the performance of the proposed spatial-temporal MSVC against the spatial-temporal MRF previously employed in literature. In the second experiment, we compare groups 2 and 3 to establish the equivalence of the proposed hierarchical and non-hierarchical MSVC for spatial-temporal classification. In the third experiments, we compare within group 3 to determine the effects of variations in the specification of temporal context, including the use of the Global Transition (GT) and Locally-Adjusted Global Transition (LAT) Probability Models. As previously described, different forms of spatial-temporal context are used, including spatial (S), spatial-temporal (ST), spatial-temporal with illogical land-cover transition penalties (STP), and spatial-temporal with weighted illogical land-cover transition penalties (STWP). Weighting or non-weighting refers to the aggregation of penalties for use in the hierarchical classifiers. The non-contextual SVM and HSVM classifiers are compared throughout to establish baseline performances. These experiments are summarized in Table 9 below:

Table 8. Summary of attributes of classifiers in experiments for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Group</th>
<th>Classifier</th>
<th>Hier.</th>
<th>Context</th>
<th>Transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td></td>
<td>Spatial</td>
<td>Temporal</td>
</tr>
<tr>
<td></td>
<td>HSVM</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td>MRF S</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MRF ST</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MRF STP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>#2</td>
<td>MSVC S</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MSVC ST</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MSVC STP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>#3</td>
<td>HMSVC ST</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>HMSVC STP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

In the first experiment, we compare groups 1 and 2 to assess the performance of the proposed spatial-temporal MSVC against the spatial-temporal MRF previously employed in literature. In the second experiment, we compare groups 2 and 3 to establish the equivalence of the proposed hierarchical and non-hierarchical MSVC for spatial-temporal classification. In the third experiments, we compare within group 3 to determine the effects of variations in the specification of temporal context, including the use of the Global Transition (GT) and Locally-Adjusted Global Transition (LAT) Probability Models. As previously described, different forms of spatial-temporal context are used, including spatial (S), spatial-temporal (ST), spatial-temporal with illogical land-cover transition penalties (STP), and spatial-temporal with weighted illogical land-cover transition penalties (STWP). Weighting or non-weighting refers to the aggregation of penalties for use in the hierarchical classifiers. The non-contextual SVM and HSVM classifiers are compared throughout to establish baseline performances. These experiments are summarized in Table 9 below:
Table 9. Summary of experiments for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Purpose</th>
<th>Classifiers Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Evaluate against previous literature</td>
<td>SVM, MRF S/ST/STP, MSVC S/ST/STP</td>
</tr>
<tr>
<td>#2</td>
<td>Evaluate hierarchical versus non-hierarchical</td>
<td>H/SVM, ST H/MSVC</td>
</tr>
<tr>
<td>#3</td>
<td>Evaluate variations in temporal context</td>
<td>HSVM, HMSVC ST/STP GT/LAT</td>
</tr>
</tbody>
</table>

For all experiments that employ SVM classifiers or their variations, we use the radial basis function (RBF) kernel as a base kernel and fix the misclassification cost $C$ to a value of 10. This RBF kernel has form $K(x_j, x) = exp\left(-G\|x_j - x\|^2\right)$ and we choose the bandwidth $G$ through a 3-fold cross-validated search over the parameter range $2^{-10}:1:3$. We employ the LIBSVM software for training our SVM models (Chang & Lin, 2011).

For the MRF models, we select the $\beta$ model parameters by genetic algorithm using the overall classification accuracy on the training set as the evaluation criterion, while selecting for the parameter vector with minimum norm (Tso & Mather, 1999; Dorsey & Mayer, 1995). We use SVM as the base classifier and estimate the likelihood energy by pairwise coupling of subproblem results (Liu & Cai, 2012; Wu et al., 2004).

For the MSVCs, we employ the multiple kernel learning algorithm of Varma and Babu (2009) to select the kernel parameters $\gamma$ under L2 regularization and using the variant Armijo algorithm to determine step size.

For the hierarchical classifiers, we use the method of Cevikalp (2012) to construct the classification hierarchy for each image. As previously mentioned, we construct this hierarchy based on the first image in our time series and then replicate it to the images at other dates. We
modify the algorithms to use both SVM and MSVM classifiers.

We terminate the MRF and MSVC algorithms when a maximum number of contextual iterations have occurred and output the set of classifications from the iteration that yielded the highest accuracy of change detection. We choose 10 iterations to terminate the algorithms due to the exhibited convergence of all techniques within this length of computation.

We establish the significance of the difference between classification results using the two kappa comparison test, which is commonly employed in remote sensing studies. The test is given as:

\[
Z = \frac{|\hat{\kappa}_1 - \hat{\kappa}_2|}{\sqrt{\text{var}(\hat{\kappa}_1) + \text{var}(\hat{\kappa}_2)}}
\]

(4.1)

where \(H_0: (\kappa_1 - \kappa_2) = 0\) and \(H_A: (\kappa_1 - \kappa_2) \neq 0\), rejecting if \(Z \geq Z_{\alpha/2}\). For further details and definition of the kappa coefficient \(\kappa\) and variance, we refer the reader to Congalton and Green (2009).

4.3 Results

Figure 5 below depicts the structure of the decision tree used for the hierarchical methods.
We see that there is a clear division between vegetative and non-vegetative types at the first split. Subsequent splits reasonably separate out the forest and urban classes, which are known to have stand-out spectral characteristics. The hierarchy delays splitting the hardest-to-separate classes until the end. Altogether, this hierarchy replicates well what an analyst might create based on their knowledge of the spectral characteristics of and relationships between these classes.

Table 10 below lists the optimal bands selected to distinguish each transition between adjacent image dates in the LAT model.

<table>
<thead>
<tr>
<th>From</th>
<th>W</th>
<th>F</th>
<th>C</th>
<th>U</th>
<th>M</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>U</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>M</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

We see that the optimal band is not constant between all adjacent pairs of image dates despite the same land-cover transition. We can attribute this to difference in the spectral information between bands at different image dates due to the effects of phenology and other sources of variation in the spectral responses of land cover as well as error among the initial non-contextual land-cover maps.

In the remainder of this section, the three experiments are examined in detail.

4.3.1 Experiment 1

In the first experiment, the goal was to determine the performance of the Markovian kernel methods relative to their source MRF models.

Figure 6 below depicts the source satellite image and classification results at the 2003 image date for a subset of the study area centered over the city of Chillicothe, Ohio.
Figure 6. Classification maps for a subset of a) the 2003 date of the Appalachian Ohio scene generated with b) SVM, c) MRF STP, and d) MSVC STP classifiers under Experiment 1. Spectral data is depicted with 7-5-3 band combination.

From the visual depiction of the classification results, we see the contextual methods (MRF and MSVC) have a desirable smoothing effect on the characteristic salt-and-pepper classification noise common to non-contextual classifiers (SVM). However, the MSVC result reveals finer detail in terms of the edge complexity of land cover and better preserves linear features, such as roads and rivers, than the MRF classifier.
For all image dates, we first examine the accuracy of change detection. Because we do not have paired validation data among the image dates, we follow the method of Liu and Cai (2012) to estimate the pessimistic, optimistic, and average accuracies of change detection. The pessimistic accuracy assumes that there is no temporal dependence of the error structure between image dates. Hence, it offers a lower bound on the accuracy of change detection estimated by the product of the accuracies at individual dates. The optimistic accuracy assumes that there is complete temporal dependence of the error structure between image dates. Hence, it offers an upper bound on the accuracy of change detection estimated as the minimum accuracy among individual dates. The average accuracy is simply the mean of the pessimistic and optimistic accuracies and represents a compromise between the extremes of no and complete spatial dependence. Table 11 below lists the pessimistic, optimistic, and average accuracies of change detection for the compared classifiers.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Pessimistic</th>
<th>Optimistic</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>60.52</td>
<td>94.41</td>
<td>77.46</td>
</tr>
<tr>
<td>S</td>
<td>64.68</td>
<td>95.24</td>
<td>79.96</td>
</tr>
<tr>
<td>ST</td>
<td>71.32</td>
<td>94.97</td>
<td>83.15</td>
</tr>
<tr>
<td>STP</td>
<td>71.29</td>
<td>95.79</td>
<td>83.54</td>
</tr>
<tr>
<td>SVM</td>
<td>61.27</td>
<td>94.83</td>
<td>78.05</td>
</tr>
<tr>
<td>S</td>
<td>66.01</td>
<td>96.64</td>
<td>81.32</td>
</tr>
<tr>
<td>ST</td>
<td>79.23</td>
<td>96.92</td>
<td>88.08</td>
</tr>
<tr>
<td>STP</td>
<td>79.22</td>
<td>97.06</td>
<td>88.14</td>
</tr>
</tbody>
</table>

From these results, we see three trends. The first is that an increasingly sophisticated use of contextual information results in concordant increases in the accuracy of change detection. In particular, we see that the non-contextual SVM provides the lowest average accuracy of change detection.
detection, but that this accuracy improves with the inclusion of spatial (S), spatial-temporal (ST), and illogically penalized spatial-temporal (STP) context in the case of both contextual classifiers. The second trend relates to the magnitude of the increase in accuracy with the inclusion of spatial versus spatial-temporal context. We see that, although the use of spatial context offers marginal increases in the average accuracy of change detection over the use of no context, the use of spatial-temporal context results in a substantial increase in this accuracy. This is expected because spatial-temporal contextual classifiers are designed to increase the spatial and temporal dependence in the error structure of the classifications. The final trend is that the proposed contextual methods (MSVC) substantially improve on the contextual models (MRF) on which they are based. We see that not only are the average accuracies of change detection higher, but also the magnitude of the changes from spatial-temporal to spatial and from spatial-temporal to non-contextual are greater in the case of MSVC than MRF. Altogether, we conclude that the proposed classifier can substantially improve on the previous MRF and SVM methods.

Second, we examine the overall and average accuracies at each image date. To do so, we continue to employ the error matrix approach (Congalton & Green, 1999), which matches reference samples with classification samples to count the occurrence of each pair of classes between the classification scheme of the reference data and that used for the classification. Because these schemes are the same, the result has the form of a matrix with a diagonal indicating the counts of samples that have been correctly classified and off-diagonals indicating the counts of samples that have been incorrectly classified. The overall accuracy is thus the percentage of samples that have been correctly classified, while the average accuracy is the average of individual class accuracies, which are each determined as the percentage of pixels in each reference class that are correctly classified -- the so-called producer’s accuracy for each class. To
aid in the reporting of our results, we also present the mean overall accuracy among image dates and the mean average accuracy among image dates.

Table 12 below reports the overall accuracies among image dates.

<table>
<thead>
<tr>
<th>Year</th>
<th>SVM</th>
<th>MRF</th>
<th>MSVC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>ST</td>
<td>STP</td>
</tr>
<tr>
<td>1988</td>
<td>90.88*</td>
<td>91.95*</td>
<td>91.79*</td>
</tr>
<tr>
<td>1993</td>
<td>87.85*</td>
<td>89.25*</td>
<td>94.86</td>
</tr>
<tr>
<td>1998</td>
<td>89.64*</td>
<td>91.16*</td>
<td>93.65</td>
</tr>
<tr>
<td>2003</td>
<td>89.57*</td>
<td>90.78*</td>
<td>92.25*</td>
</tr>
<tr>
<td>2008</td>
<td>94.41*</td>
<td>95.24</td>
<td>94.97*</td>
</tr>
<tr>
<td>Mean</td>
<td>90.47</td>
<td>91.68</td>
<td>93.47</td>
</tr>
</tbody>
</table>

Table 13 below reports the average accuracies among image dates.

From these results, we see that on average overall accuracies increase among image dates with increasingly sophisticated use of contextual information. In particular, in all cases, the use of spatial-temporal context improves upon the use of spatial context, which in turn improves upon the use of no context at all. Moreover, MSVC outperforms both MRF and SVM in terms of overall accuracy and the margin between contextual and non-contextual classifier is greater with MSVC and SVM than with MRF and SVM. Altogether, MSVC outperforms both MRF and SVM.
Table 13. Average accuracies in Experiment 1 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Year</th>
<th>SVM</th>
<th>MRF</th>
<th>MSVC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S</td>
<td>ST</td>
</tr>
<tr>
<td>1988</td>
<td>66.25</td>
<td>67.27</td>
<td>66.92</td>
</tr>
<tr>
<td>1993</td>
<td>84.25</td>
<td>85.90</td>
<td>87.63</td>
</tr>
<tr>
<td>1998</td>
<td>87.06</td>
<td>89.74</td>
<td>86.14</td>
</tr>
<tr>
<td>2003</td>
<td>79.48</td>
<td>82.00</td>
<td>77.01</td>
</tr>
<tr>
<td>2008</td>
<td>91.59</td>
<td>92.42</td>
<td>89.85</td>
</tr>
<tr>
<td>Mean</td>
<td>81.73</td>
<td>83.47</td>
<td>81.51</td>
</tr>
</tbody>
</table>

From the results, we see that, on average, increasingly sophisticated use of contextual information results in decreased average accuracies for the MRF model, while the trend is the opposite for the MSVC methods – average accuracy increases with the move from no context to spatial context to spatial-temporal context. Among the individual dates, the average accuracy in the year 1988 particularly suffers. Through examination of the error matrices at this date for the classification results from the different classifiers, we note that the reason is that the very few test samples in the mine class are frequently classified as the urban class and that, due to the small test sample size, this has inordinate influence on the average individual class accuracy. Altogether, we again see that MSVC improves upon MRF.

Third, we examine the effects of the different classifiers on the occurrence of illogical change in land-cover change trajectories. In the following, MRF denotes the MRF STP model and MSVC denotes the MSVC STP. Hence, the contextual spatial-temporal models that penalize illogical land-cover transitions are compared to each other and to the non-contextual SVM. Following the format of Liu and Cai (2012), Table 14 below lists the frequency that each type of illogical transition occurs at each date for each classifier.
From these results, we see that, with the exception of the period 1998 → 2003, the total percentage of illogical transitions decreases with the move from the non-contextual SVM to the contextual MRF to the proposed MSVC. Hence, the MSVC is more effective than the previous MRF at removing illogical transitions among the image dates.

We also examine the effect of the three classifiers on the frequency of change and illogical change in terms of the overall number of land cover trajectories. Table 15 below depicts the counts and frequencies for the three types of trajectories.

### Table 14. Illogical land cover transition frequencies in Experiment 1 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th></th>
<th>W→F</th>
<th>W→M</th>
<th>C→F</th>
<th>U→W</th>
<th>U→F</th>
<th>U→C</th>
<th>U→M</th>
<th>M→F</th>
<th>M→U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988 SVM</td>
<td>0.29</td>
<td>0.00</td>
<td>2.37</td>
<td>0.00</td>
<td>0.03</td>
<td>0.92</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>3.64</td>
</tr>
<tr>
<td>1993 MRF</td>
<td>0.30</td>
<td>0.00</td>
<td>1.51</td>
<td>0.00</td>
<td>0.00</td>
<td>1.17</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>3.00</td>
</tr>
<tr>
<td>1993 MSVC</td>
<td>0.31</td>
<td>0.00</td>
<td>1.46</td>
<td>0.01</td>
<td>0.01</td>
<td>0.09</td>
<td>0.03</td>
<td>0.00</td>
<td>0.01</td>
<td>1.93</td>
</tr>
<tr>
<td>1998 SVM</td>
<td>1.26</td>
<td>0.00</td>
<td>4.79</td>
<td>0.04</td>
<td>1.20</td>
<td>0.79</td>
<td>0.03</td>
<td>0.00</td>
<td>0.04</td>
<td>8.16</td>
</tr>
<tr>
<td>1998 MRF</td>
<td>1.90</td>
<td>0.00</td>
<td>1.60</td>
<td>0.01</td>
<td>1.19</td>
<td>0.25</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>4.15</td>
</tr>
<tr>
<td>1998 MSVC</td>
<td>0.49</td>
<td>0.00</td>
<td>1.22</td>
<td>0.06</td>
<td>1.18</td>
<td>0.86</td>
<td>0.07</td>
<td>0.00</td>
<td>0.02</td>
<td>3.90</td>
</tr>
<tr>
<td>2003 SVM</td>
<td>0.17</td>
<td>0.00</td>
<td>3.75</td>
<td>0.02</td>
<td>0.07</td>
<td>1.19</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
<td>5.26</td>
</tr>
<tr>
<td>2003 MRF</td>
<td>0.00</td>
<td>0.00</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.47</td>
</tr>
<tr>
<td>2003 MSVC</td>
<td>0.08</td>
<td>0.00</td>
<td>1.16</td>
<td>0.02</td>
<td>0.02</td>
<td>1.28</td>
<td>0.04</td>
<td>0.00</td>
<td>0.04</td>
<td>2.64</td>
</tr>
<tr>
<td>2008 SVM</td>
<td>0.12</td>
<td>0.00</td>
<td>1.53</td>
<td>0.04</td>
<td>0.06</td>
<td>0.94</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>2.75</td>
</tr>
<tr>
<td>2008 MRF</td>
<td>0.14</td>
<td>0.00</td>
<td>1.32</td>
<td>0.03</td>
<td>0.01</td>
<td>1.14</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
<td>1.68</td>
</tr>
<tr>
<td>2008 MSVC</td>
<td>0.20</td>
<td>0.00</td>
<td>0.64</td>
<td>0.00</td>
<td>0.02</td>
<td>0.30</td>
<td>0.03</td>
<td>0.00</td>
<td>0.01</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Note that MRF = MRF STP and MSVC = MSVC STP.

### Table 15. Land cover trajectory counts and frequencies in Experiment 1 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Illogical Change</th>
<th>Change</th>
<th>No Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
</tr>
<tr>
<td>SVM</td>
<td>566,758</td>
<td>18.89</td>
<td>1,536,035</td>
</tr>
<tr>
<td>MRF</td>
<td>275,322</td>
<td>9.18</td>
<td>1,117,125</td>
</tr>
<tr>
<td>MSVC</td>
<td>278,336</td>
<td>9.28</td>
<td>1,100,035</td>
</tr>
</tbody>
</table>

Note that MRF = MRF STP and MSVC = MSVC STP.
From these results, we see that the contextual classifiers (MRF and MSVC) reduce both the number of pixels undergoing change at any date and the number of pixels undergoing illogical change at any date compared with the non-contextual classifier (SVM). Furthermore, while the MSVC reduces the overall number of pixels undergoing change compared to the MRF model, it actually increases the number of pixels undergoing illogical change. Combined with the previous result, we note that an overall decrease in the occurrence of illogical transitions among all dates does not necessarily require a decrease in the number of trajectories containing illogical transitions.

Fourth, we examine the spatial pattern of trajectories undergoing land-cover transitions and illogical land-cover transitions. Figure 7 depicts the transition maps for each classifier.
Figure 7. Trajectory transition maps for a subset of the Appalachian Ohio scene counted from the land-cover maps generated with a) SVM, b) MRF STP, and c) MSVC STP classifiers under Experiment 1.
From these visual results, we see that the spatial pattern of transitions changes among the
different classifiers. In particular, the contextual classifiers (MRF and MSVC) result in greater
areas undergoing no transition and reductions in the transition counts than the non-contextual
classifier (SVM).

Finally, we examine whether significant differences exists between the classifications results for
pairs of classifiers at each image date. We compare the SVM result produced prior to contextual
classification using MRF with the SVM result produced prior to MSVC. The two differ due to
the necessity of a common parameter setting for the estimation of pairwise coupled likelihood
energies for the MRF model. We also compare the non-contextual SVM with the contextual
MRF, the non-contextual SVM with the contextual MSVC, and the MRF with MSVC. Table 16
below reports the z-scores for the two kappa comparison test.

Table 16. Interclassifier significance test results in Experiment 1 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Year</th>
<th>SVM</th>
<th>SVM vs. MRF</th>
<th>SVM vs. MSVC</th>
<th>MRF vs. MSVC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S</td>
<td>ST</td>
<td>STP</td>
</tr>
<tr>
<td>1988</td>
<td>0.27</td>
<td>0.71</td>
<td>0.52</td>
<td>0.63</td>
</tr>
<tr>
<td>1993</td>
<td>0.40</td>
<td>0.77</td>
<td>4.43*</td>
<td>5.20*</td>
</tr>
<tr>
<td>1998</td>
<td>0.00</td>
<td>1.00</td>
<td>2.77*</td>
<td>2.44*</td>
</tr>
<tr>
<td>2003</td>
<td>0.35</td>
<td>0.80</td>
<td>1.76</td>
<td>1.76</td>
</tr>
<tr>
<td>2008</td>
<td>0.34</td>
<td>0.71</td>
<td>0.45</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Starred quantities are significantly different at α = 0.05 level by two kappa z-test.

From these results, we see that the non-contextual classifications produced at each date by SVM
prior to the use of MRF or MSVC have no significant differences. Hence, our previous
comparisons are unlikely to be invalidated due to any differences that may exist between them.
In general, the spatial-temporal MRF models produce classifications that are significantly different from those produced by SVM only at a few dates, whereas the spatial-temporal MSVC produces classifications that are significantly different at all dates. Furthermore, MSVC produces a number of classifications that are significantly different from the ones produced by MRF. Altogether, this provides more evidence that the MSVC technique improves on the MRF classifier.

4.3.2 Experiment 2

In the second experiment, the goal was to evaluate the use of hierarchical versus non-hierarchical classification in the proposed spatial-temporal classification methods. The expectation is that the hierarchical method, although relying on a potentially much smaller number of subproblems, should rival or exceed the classification performance of the non-hierarchical method in order to be an effective substitute.

First, we examine the accuracy of change detection. As previously, we look at the pessimistic, optimistic, and average accuracies of change detection. Table 17 below records these accuracies for the non-contextual SVM and HSVM classifiers and their spatial-temporal contextual MSVC and HMSVC counterparts.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Pessimistic</th>
<th>Optimistic</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>61.94</td>
<td>88.94</td>
<td>75.44</td>
</tr>
<tr>
<td>HSVM</td>
<td>61.27</td>
<td>94.83</td>
<td>78.05</td>
</tr>
<tr>
<td>ST MSVC</td>
<td>79.23</td>
<td>96.92</td>
<td>88.08</td>
</tr>
<tr>
<td>ST HMSVC</td>
<td>84.47</td>
<td>95.38</td>
<td>89.93</td>
</tr>
</tbody>
</table>
From these results, we see that the average accuracy of change detection for the contextual methods (H/MSVC) is substantially higher than the non-contextual methods (H/SVM). Moreover, both the MSVC and HMSVC methods see a similar magnitude of increase over their SVM and HSVM counterparts. We also note that the HSVM method results in a higher base non-contextual average accuracy of change detection than the SVM method, illustrating that hierarchical classification is potentially well suited for the classification of land cover.

Second, we examine the overall and average accuracies at individual dates for the four classifiers. Again, we base the calculation of these quantities on the error matrix approach. We also include the mean overall and mean average accuracies to aid in interpretation of the results. Table 18 below lists the results for the overall accuracy.

<table>
<thead>
<tr>
<th>Year</th>
<th>SVM</th>
<th>HSVM</th>
<th>ST MSVC</th>
<th>ST HMSVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>91.34*</td>
<td>91.03*</td>
<td>94.38*</td>
<td>98.48</td>
</tr>
<tr>
<td>1993</td>
<td>88.63*</td>
<td>88.94*</td>
<td>95.95*</td>
<td>98.29</td>
</tr>
<tr>
<td>1998</td>
<td>89.64*</td>
<td>90.33*</td>
<td>93.92</td>
<td>95.72</td>
</tr>
<tr>
<td>2003</td>
<td>89.04*</td>
<td>89.44*</td>
<td>96.12</td>
<td>95.59</td>
</tr>
<tr>
<td>2008</td>
<td>94.83*</td>
<td>94.69*</td>
<td>96.92</td>
<td>95.38</td>
</tr>
<tr>
<td>Mean</td>
<td>90.70</td>
<td>90.89</td>
<td>95.46</td>
<td>96.69</td>
</tr>
</tbody>
</table>

Bold quantities indicate best results. Starred quantities are significantly different at $\alpha = 0.05$ level by two kappa z-test.

From these results, we see that the hierarchical and non-hierarchical classifiers offer similar mean levels of overall accuracy. The contextual classifiers (H/MSVC) again outperform their non-contextual counterparts (H/SVM) and that the hierarchical techniques (HSVM/MSVC) on average marginally outperform the non-hierarchical techniques (SVM/MSVC).
Table 19 below reports the average accuracies among image dates.

**Table 19. Average accuracies in Experiment 2 for Appalachian Ohio dataset.**

<table>
<thead>
<tr>
<th>Year</th>
<th>SVM</th>
<th>HSVM</th>
<th>ST MSVC</th>
<th>ST HMSVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>66.22</td>
<td>65.72</td>
<td>72.81</td>
<td>96.86</td>
</tr>
<tr>
<td>1993</td>
<td>80.77</td>
<td>81.13</td>
<td>91.06</td>
<td>94.01</td>
</tr>
<tr>
<td>1998</td>
<td>86.91</td>
<td>87.97</td>
<td>89.56</td>
<td>86.50</td>
</tr>
<tr>
<td>2003</td>
<td>78.45</td>
<td>80.19</td>
<td>88.94</td>
<td>91.78</td>
</tr>
<tr>
<td>2008</td>
<td>93.05</td>
<td>93.21</td>
<td>95.22</td>
<td>93.10</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>81.08</td>
<td>81.64</td>
<td>87.52</td>
<td>92.45</td>
</tr>
</tbody>
</table>

From these results, we see the same general trends as we saw in the overall accuracies. However, we note that the inclusion of context in the hierarchical method (HMSVC) clearly solves the individual class accuracy problem of the 1988 date.

Finally, we examine whether significant differences exist between the classification results for pairs of classifiers at each image date. We compare the non-contextual classifiers (SVM vs. HSVM), the non-hierarchical non-contextual and contextual classifiers (SVM vs. MSVC), the hierarchical non-contextual and contextual classifiers (HSVM vs. HMSVC), and the non-hierarchical and hierarchical contextual classifiers (MSVC vs. HMSVC).

Table 20 below reports the z-scores for the two kappa comparison test among these pairs and dates.
From these results, we see that the non-contextual classifications produced at each date by the SVM and HSVM classifiers have no significant differences. Hence, our previous comparisons are unlikely to be invalidated due to any differences that may exist between them. In combination with our previous results, we see that the use of contextual information significantly improves classification results. Most important, we note that combination of hierarchical and contextual classification can significantly improve on contextual classification alone.

4.3.3 Experiment 3

In the third experiment, the goal was to investigate the effect of variations in temporal context on classifier performance. Seven variations were evaluated for the HMSVC classifier. The use of spatial-temporal context was compared for the case where no illogical transition penalties (ST), an unweighted aggregation of penalties (STP), and a weighted aggregation of penalties (STWP) was used in the presence of a Global Transition (GT) or a Locally-Adjusted Global Transition (LAT) Probability Model.

First, we examine the accuracy of change detection. As previously, we look at the pessimistic, optimistic, and average accuracies of change detection. Table 21 below records these accuracies.
among the variations in temporal context.

**Table 21.** Accuracies of change detection in Experiment 3 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Pessimistic</th>
<th>Optimistic</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSVM</td>
<td>61.94</td>
<td>88.94</td>
<td>75.44</td>
</tr>
<tr>
<td>GT ST</td>
<td>84.47</td>
<td>95.38</td>
<td>89.93</td>
</tr>
<tr>
<td>STP</td>
<td>85.96</td>
<td>95.72</td>
<td>90.84</td>
</tr>
<tr>
<td>STWP</td>
<td>86.34</td>
<td>96.22</td>
<td>91.28</td>
</tr>
<tr>
<td>LAT ST</td>
<td>86.04</td>
<td>95.94</td>
<td>90.99</td>
</tr>
<tr>
<td>STP</td>
<td>86.67</td>
<td>95.80</td>
<td>91.24</td>
</tr>
<tr>
<td>STWP</td>
<td>86.84</td>
<td>96.22</td>
<td>91.53</td>
</tr>
</tbody>
</table>

From these results, we see a trend of increasing accuracy of change detection from ST to STP to STWP. We also see each LAT model outperforming its counterpart GT model. All variations outperform the non-contextual HSVM classifier. However, the differences in accuracy among the variations in temporal context are marginal.

Second, we examine the overall accuracy at individual dates. Table 22 below reports the overall accuracy at each date and the mean overall accuracy across dates for each variation in temporal context.
Table 22. Overall accuracies in Experiment 3 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Year</th>
<th>HSVM</th>
<th>GT ST</th>
<th>STP</th>
<th>STWP</th>
<th>LAT ST</th>
<th>STP</th>
<th>STWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>91.03*</td>
<td>98.48</td>
<td>98.02</td>
<td>97.72</td>
<td>97.26</td>
<td>98.33</td>
<td>97.42</td>
</tr>
<tr>
<td>1993</td>
<td>88.94*</td>
<td>98.29</td>
<td>98.75</td>
<td>97.98</td>
<td>98.13</td>
<td>98.44</td>
<td>98.60</td>
</tr>
<tr>
<td>1998</td>
<td>90.33*</td>
<td>95.72</td>
<td>96.41</td>
<td>97.10</td>
<td>96.27</td>
<td>96.82</td>
<td>96.27</td>
</tr>
<tr>
<td>2003</td>
<td>89.44*</td>
<td>95.59*</td>
<td>95.72*</td>
<td>96.52</td>
<td>97.59</td>
<td>96.52</td>
<td>97.59</td>
</tr>
<tr>
<td>2008</td>
<td>94.69</td>
<td>95.38</td>
<td><strong>96.22</strong></td>
<td><strong>96.22</strong></td>
<td>95.94</td>
<td>95.80</td>
<td><strong>96.22</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>90.89</td>
<td>96.69</td>
<td>97.03</td>
<td>97.11</td>
<td>97.04</td>
<td>97.18</td>
<td>97.22</td>
</tr>
</tbody>
</table>

Bold quantities indicate best results.
Starred quantities are significantly different at α = 0.05 level by two kappa z-test.

In general, we see no significant difference between variations in temporal context with the exception of the 2003 date in which the LAT model outperforms the GT model. In terms of mean overall accuracy, we note a marginal increasing trend from ST to STP to STWP for both models with the LAT model again outperforming the GT model.

We also examine the average accuracies, which we report in Table 23 below along with the mean average accuracies.

Table 23. Average accuracies in Experiment 3 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th>Year</th>
<th>HSVM</th>
<th>GT ST</th>
<th>STP</th>
<th>STWP</th>
<th>LAT ST</th>
<th>STP</th>
<th>STWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>65.72</td>
<td>96.86</td>
<td>88.47</td>
<td>80.14</td>
<td>78.35</td>
<td>96.80</td>
<td>78.64</td>
</tr>
<tr>
<td>1993</td>
<td>81.13</td>
<td>94.01</td>
<td>97.78</td>
<td>93.59</td>
<td>95.26</td>
<td>95.86</td>
<td>95.93</td>
</tr>
<tr>
<td>1998</td>
<td>87.97</td>
<td>86.50</td>
<td>91.41</td>
<td>94.22</td>
<td>91.17</td>
<td>90.60</td>
<td>89.32</td>
</tr>
<tr>
<td>2003</td>
<td>80.19</td>
<td>91.78</td>
<td>93.70</td>
<td>93.59</td>
<td>95.55</td>
<td>94.66</td>
<td>95.55</td>
</tr>
<tr>
<td>2008</td>
<td>93.21</td>
<td>93.10</td>
<td>95.43</td>
<td>95.15</td>
<td>93.07</td>
<td>95.24</td>
<td>95.15</td>
</tr>
<tr>
<td>Mean</td>
<td>81.64</td>
<td>92.45</td>
<td>93.36</td>
<td>91.34</td>
<td>90.68</td>
<td>94.63</td>
<td>90.92</td>
</tr>
</tbody>
</table>
From the results, we do not see a similar ranking among the variations in temporal context as with overall accuracy. Instead, we see that STP outperforms the other variations in terms of mean average accuracy.

Third, we examine the effects of variations in temporal context on the frequency of change and illogical change in terms of the overall number of land cover trajectories. Table 24 below depicts the counts and frequencies for the three types of trajectories.

| Classifier | Variation | Illogical | | Change | | No Change |
|------------|-----------|-----------|-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| HSVM       | GT        | 541,912   | 18.06       | 1,555,017    | 51.83           | 1,444,983       | 48.17           |
|            | ST        | 158,907   | 5.30        | 848,089      | 28.27           | 2,151,911       | 71.73           |
| HSVM       | GT STP    | 155,737   | 5.19        | 834,560      | 27.82           | 2,165,440       | 72.18           |
|            | GT STWP   | 170,660   | 5.69        | 874,374      | 29.15           | 2,125,626       | 72.18           |
| HSVM       | LAT ST    | 163,081   | 5.44        | 866,317      | 28.88           | 2,133,683       | 71.12           |
|            | LAT STP   | 157,843   | 5.26        | 839,341      | 27.98           | 2,160,659       | 72.02           |
|            | LAT STWP  | 157,202   | 5.24        | 868,333      | 28.94           | 2,131,667       | 71.06           |

From these results, we see that the inclusion of unweighted illogical penalties (STP) improves upon the use of no penalties (ST) in terms of the number of trajectories containing illogical transitions. The LAT models also generally have a higher number of illogical trajectory than their GT counterparts. All variations in temporal context substantially reduce the number of trajectories undergoing any type of change compared to the non-contextual model (HSVM).

Finally, we examine the frequency of illogical transitions among the image dates. Table 25 below
Table 25. Illogical land cover transition frequencies for GT model in Experiment 3 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th></th>
<th>W→F</th>
<th>W→M</th>
<th>C→F</th>
<th>U→W</th>
<th>U→F</th>
<th>U→C</th>
<th>U→M</th>
<th>M→F</th>
<th>M→U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>0.25</td>
<td>0.00</td>
<td>1.96</td>
<td>0.01</td>
<td>0.02</td>
<td>0.80</td>
<td>0.01</td>
<td>0.09</td>
<td>0.01</td>
<td>3.16</td>
</tr>
<tr>
<td>ST</td>
<td>0.05</td>
<td>0.04</td>
<td>0.58</td>
<td>0.00</td>
<td>0.00</td>
<td>0.41</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>1.13</td>
</tr>
<tr>
<td>STP</td>
<td>0.79</td>
<td>0.01</td>
<td>0.60</td>
<td>0.02</td>
<td>0.06</td>
<td>0.12</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>1.66</td>
</tr>
<tr>
<td>STWP</td>
<td>0.08</td>
<td>0.01</td>
<td>0.52</td>
<td>0.13</td>
<td>0.01</td>
<td>0.06</td>
<td>0.13</td>
<td>0.00</td>
<td>0.01</td>
<td>0.95</td>
</tr>
<tr>
<td>1993</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>1.21</td>
<td>0.00</td>
<td>4.58</td>
<td>0.04</td>
<td>1.38</td>
<td>0.79</td>
<td>0.04</td>
<td>0.00</td>
<td>0.04</td>
<td>8.08</td>
</tr>
<tr>
<td>ST</td>
<td>0.13</td>
<td>0.03</td>
<td>0.49</td>
<td>0.08</td>
<td>0.03</td>
<td>0.39</td>
<td>0.42</td>
<td>0.01</td>
<td>0.07</td>
<td>1.64</td>
</tr>
<tr>
<td>STP</td>
<td>0.04</td>
<td>0.03</td>
<td>0.48</td>
<td>0.05</td>
<td>0.05</td>
<td>0.44</td>
<td>0.11</td>
<td>0.00</td>
<td>0.05</td>
<td>1.26</td>
</tr>
<tr>
<td>STWP</td>
<td>1.01</td>
<td>0.02</td>
<td>0.54</td>
<td>0.01</td>
<td>0.19</td>
<td>0.23</td>
<td>0.01</td>
<td>0.00</td>
<td>0.06</td>
<td>2.08</td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>0.08</td>
<td>0.00</td>
<td>3.42</td>
<td>0.01</td>
<td>0.05</td>
<td>1.35</td>
<td>0.02</td>
<td>0.00</td>
<td>0.05</td>
<td>4.98</td>
</tr>
<tr>
<td>ST</td>
<td>0.02</td>
<td>0.00</td>
<td>0.72</td>
<td>0.05</td>
<td>0.11</td>
<td>0.42</td>
<td>0.02</td>
<td>0.00</td>
<td>0.05</td>
<td>1.40</td>
</tr>
<tr>
<td>STP</td>
<td>0.02</td>
<td>0.02</td>
<td>0.72</td>
<td>0.02</td>
<td>0.08</td>
<td>0.35</td>
<td>0.04</td>
<td>0.00</td>
<td>0.02</td>
<td>1.26</td>
</tr>
<tr>
<td>STWP</td>
<td>0.03</td>
<td>0.03</td>
<td>0.63</td>
<td>0.01</td>
<td>0.17</td>
<td>0.84</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>1.75</td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>0.37</td>
<td>0.00</td>
<td>1.27</td>
<td>0.08</td>
<td>0.07</td>
<td>0.76</td>
<td>0.04</td>
<td>0.00</td>
<td>0.02</td>
<td>2.61</td>
</tr>
<tr>
<td>ST</td>
<td>0.03</td>
<td>0.01</td>
<td>1.00</td>
<td>0.03</td>
<td>0.05</td>
<td>0.22</td>
<td>0.04</td>
<td>0.00</td>
<td>0.05</td>
<td>1.42</td>
</tr>
<tr>
<td>STP</td>
<td>0.07</td>
<td>0.01</td>
<td>0.81</td>
<td>0.01</td>
<td>0.06</td>
<td>0.26</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>1.25</td>
</tr>
<tr>
<td>STWP</td>
<td>0.07</td>
<td>0.02</td>
<td>0.82</td>
<td>0.01</td>
<td>0.03</td>
<td>0.15</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Table 26 below depicts the frequencies for the locally-adjusted global transition model (LAT).
## Table 26. Illogical land cover transition frequencies for LAT model in Experiment 3 for Appalachian Ohio dataset.

<table>
<thead>
<tr>
<th></th>
<th>W→F</th>
<th>W→M</th>
<th>C→F</th>
<th>U→W</th>
<th>U→F</th>
<th>U→C</th>
<th>U→M</th>
<th>M→F</th>
<th>M→U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998 ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>0.25</td>
<td>0.00</td>
<td>1.96</td>
<td>0.01</td>
<td>0.02</td>
<td>0.80</td>
<td>0.01</td>
<td>0.09</td>
<td>0.01</td>
<td>3.16</td>
</tr>
<tr>
<td>ST</td>
<td>0.12</td>
<td>0.03</td>
<td>0.42</td>
<td>0.03</td>
<td>0.01</td>
<td>0.09</td>
<td>0.07</td>
<td>0.00</td>
<td>0.04</td>
<td>0.80</td>
</tr>
<tr>
<td>STP</td>
<td>0.16</td>
<td>0.01</td>
<td>0.49</td>
<td>0.02</td>
<td>0.02</td>
<td>0.09</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.83</td>
</tr>
<tr>
<td>STWP</td>
<td>0.12</td>
<td>0.03</td>
<td>0.38</td>
<td>0.09</td>
<td>0.01</td>
<td>0.10</td>
<td>0.10</td>
<td>0.00</td>
<td>0.01</td>
<td>0.83</td>
</tr>
<tr>
<td>1993</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998 ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>1.21</td>
<td>0.00</td>
<td>4.58</td>
<td>0.04</td>
<td>1.38</td>
<td>0.79</td>
<td>0.04</td>
<td>0.00</td>
<td>0.04</td>
<td>8.08</td>
</tr>
<tr>
<td>ST</td>
<td>0.71</td>
<td>0.01</td>
<td>0.57</td>
<td>0.06</td>
<td>0.10</td>
<td>0.17</td>
<td>0.23</td>
<td>0.02</td>
<td>0.06</td>
<td>1.93</td>
</tr>
<tr>
<td>STP</td>
<td>0.58</td>
<td>0.03</td>
<td>0.54</td>
<td>0.02</td>
<td>0.23</td>
<td>0.45</td>
<td>0.07</td>
<td>0.00</td>
<td>0.03</td>
<td>1.95</td>
</tr>
<tr>
<td>STWP</td>
<td>0.62</td>
<td>0.07</td>
<td>0.59</td>
<td>0.03</td>
<td>0.20</td>
<td>0.14</td>
<td>0.09</td>
<td>0.02</td>
<td>0.06</td>
<td>1.80</td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003 ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>0.08</td>
<td>0.00</td>
<td>3.42</td>
<td>0.01</td>
<td>0.05</td>
<td>1.35</td>
<td>0.02</td>
<td>0.00</td>
<td>0.05</td>
<td>4.98</td>
</tr>
<tr>
<td>ST</td>
<td>0.01</td>
<td>0.03</td>
<td>0.59</td>
<td>0.03</td>
<td>0.17</td>
<td>0.77</td>
<td>0.05</td>
<td>0.00</td>
<td>0.08</td>
<td>1.73</td>
</tr>
<tr>
<td>STP</td>
<td>0.02</td>
<td>0.01</td>
<td>0.72</td>
<td>0.03</td>
<td>0.10</td>
<td>0.45</td>
<td>0.04</td>
<td>0.00</td>
<td>0.02</td>
<td>1.40</td>
</tr>
<tr>
<td>STWP</td>
<td>0.02</td>
<td>0.03</td>
<td>0.58</td>
<td>0.02</td>
<td>0.16</td>
<td>0.74</td>
<td>0.05</td>
<td>0.00</td>
<td>0.01</td>
<td>1.60</td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008 ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>0.37</td>
<td>0.00</td>
<td>1.27</td>
<td>0.08</td>
<td>0.07</td>
<td>0.76</td>
<td>0.04</td>
<td>0.00</td>
<td>0.02</td>
<td>2.61</td>
</tr>
<tr>
<td>ST</td>
<td>0.13</td>
<td>0.02</td>
<td>0.79</td>
<td>0.03</td>
<td>0.08</td>
<td>0.13</td>
<td>0.05</td>
<td>0.00</td>
<td>0.03</td>
<td>1.26</td>
</tr>
<tr>
<td>STP</td>
<td>0.12</td>
<td>0.02</td>
<td>0.84</td>
<td>0.01</td>
<td>0.06</td>
<td>0.17</td>
<td>0.04</td>
<td>0.00</td>
<td>0.01</td>
<td>1.27</td>
</tr>
<tr>
<td>STWP</td>
<td>0.17</td>
<td>0.02</td>
<td>0.79</td>
<td>0.01</td>
<td>0.04</td>
<td>0.12</td>
<td>0.04</td>
<td>0.00</td>
<td>0.03</td>
<td>1.22</td>
</tr>
</tbody>
</table>

From the results, we note that 18.83%, 5.72%, 5.45%, and 5.45% of all transitions are illogical for the four methods, respectively. Hence, STP and STWP minimize the overall number of illogical transitions under the LAT model.

### 4.3.4 Summary of Results

Altogether, the proposed spatial-temporal kernel method significantly outperforms both the previous spatial-temporal MRF model and the non-contextual SVM classifier on our dataset. Hierarchical classification was demonstrated to rival or exceed the accuracy of non-hierarchical classification. The inclusion of illogical transition penalties increased the accuracy of change detection and further reduced spatial-temporal inconsistency beyond the use of spatial-temporal information alone. The weighted model for aggregation of illogical transition penalties in the hierarchical models marginally outperformed the unweighted model. Locally adjustment of global transition probabilities was also marginally beneficial to classification accuracy. The
proposed classifier thus outperformed the previous MRF models for spatial-temporal contextual classification in remote sensing.

4.4 Discussion

Hierarchical classification results in substantial savings in terms of algorithm performance compared to the OAO multi-class strategy. For the ST MSVC classifier, over the iterative steps, we must train \( R \cdot T \cdot \frac{1}{2} |L|(|L| - 1) \) SVMs, whereas only \( R \cdot T \cdot (|L| - 1) \) for the ST HMSVC classifier where \( R \) is the total number of iterations, \( T \) is the total number of image dates, and \(|L|\) is the number of classes in the label set. Based on this analysis, the computation time should be approximately \( \frac{1}{2} |L| \) times as long for the MSVC than the HMSVC when the number of classes \(|L|\) is held fixed and the complexity of each subproblem is the same. Although in reality this is not the case, substantial savings can still be achieved. In terms of the computational equivalence between the hierarchical and non-hierarchical classifiers, Figure 8 below graphs the equivalence in number of classes between them when the number of subproblems is held fixed.

Figure 8. Computational equivalence between HMSVC and MSVC.
Despite the computational savings it affords and the possibility of marginal increases in overall accuracy, the use of hierarchical classification is not without limitation. It requires that the hierarchy used at one date be replicated among all other image dates due to the necessity of sharing contextual information between dates. This replication may not be appropriate for image time series captured from multiple sensors. Future research might investigate how to couple multi-temporal information arising from different hierarchies at different dates. In the case of multi-source data, an exhaustive enumeration of pairwise subproblems, as given by the OAO strategy, may be ideal. Hence, future research may also investigate algorithm parallelization to reduce the computational burden.

While hierarchical classification was employed to reduce the computation time of the proposed classifier in the presence of a large number of classes, it also has the side effect of improving contextual classification accuracy in the examined case study. As shown for overall accuracy in Table 18 of Experiment 2, hierarchical classification resulted in significant increases in overall accuracy at two image dates, an insignificant increase at one date, and insignificant decreases at the other two dates. All image dates were insignificantly different for non-contextual classification. Hierarchical classification constrains the possible classification of a pixel since, if it is incorrectly classified at an upper level of the hierarchy, it cannot recover the correct classification at lower levels. The choice of hierarchy is thus very important to maintaining accuracy during non-contextual classification. Using the technique to construct the hierarchy employed here, which clusters classes based on their distances in a transformed feature space, we can determine a hierarchy this is likely to minimize this source of error. We confirm from preliminary investigations that an incorrect specification of hierarchy can result in significantly
reduced non-contextual classification accuracy. However, when the same hierarchy was used in contextual classification, we found that a similar level of contextual classification accuracy to a correctly specified hierarchy could be obtained. Thus, the use of contextual information can compensate for an incorrectly specified hierarchy. We believe this is due to the use of Markovian SVMs, which enable the propagation of contextual information over iterations that can redefine the decision boundaries used to separate samples at each level of the hierarchy. The selection of new support vectors based on the influence of contextual information is thus a potent way to compensate for the constraints of hierarchical classification, enabling the possibility of increased classification accuracy when a correctly specified hierarchy is used.

There is no clear relationship between the $\gamma$ parameters used in the Markovian kernel and $\beta$ parameters used in the original MRF model. When a RBF base kernel is employed, the use of the Markovian kernel is equivalent to the concatenation of new features onto an infinite dimensional feature vector induced by the base kernel. The square roots of the kernel weights then multiply their corresponding features in the transformed feature space. We do not guarantee that this is the most effective combination of contextual information with base kernel. Non-linear or data dependent combination of kernels optimized under a multiple kernel learning framework could possibly provide better results, even though such kernels would lose their equivalence with their source MRF energy difference function. Additionally, we do not guarantee the multiple kernel learning technique by Varma and Babu (2009) used here is the most accurate choice. Other methods exist that could offer higher accuracy or greater computational efficiency (Gonen & Alpaydin, 2011). Our use of their algorithm is based on its good performance with respect to its computational cost and the use of an existing SVM solver that we employ for training (Chang & Lin, 2011).
The Markovian kernel reduces the complexity of the classification problem to be solved, as judged by the decrease in the number of support vectors with the inclusion of contextual information. For example, the average decrease in number of support vectors across subproblems for each of the five image dates was 78.79%, 66.30%, 66.17%, 83.89%, and 77.27% for the MSVC STP classifier. The use of contextual information in the kernel therefore offers a significant advantage to discriminating land cover.

4.5 Conclusion

This case study evaluated the proposed classifier against its source classification models, which are leading techniques for the classification of land cover. The proposed classifier demonstrated substantial improvements over these techniques in terms of an increase in the accuracy of change detection, increase in overall accuracy, increase in average accuracy, reduction in the frequency of illogical land-cover transitions, and reduction in the number of land-cover change trajectories undergoing illogical change. Both the accuracy and spatial-temporal consistency of multi-temporal land-cover mapping was improved. Altogether, the proposed classifier fulfilled the objectives of this thesis with respect to its evaluation in the Appalachian Ohio study area.

Further evaluation of the proposed classifier with respect to variations in training set size, realizations of the training data, image types, and land-cover change scenarios in different study areas will be needed to fully understand the classifier’s capabilities and limitations, but the results illustrated here indicate its high suitability for the task of improved multi-temporal land-cover mapping.
Chapter 5: Conclusion

5.1 Implications of Research

In this thesis, a spatial-temporal extension of the Markovian Support Vector Classifier of Moser and Serpico (2013) was proposed. It employed the energy function of the Markov Random Field model formulated in Liu and Cai (2012). It combined the hierarchical classification technique of Cevikalp (2010) and the multiple kernel learning method of Varma and Babu (2009) to perform efficient and fully-automated classification of multi-temporal remote sensing satellite imagery. The result was a substantial improvement in the accuracy of change detection and spatial-temporal consistency of multi-temporal land-cover classification compared with previous methods in an Appalachian Ohio study area.

The robustness of the classifier to sensor advancement is given by its firm theoretical basis for employing contextual information to better discriminate in the face of the increased within-class and decreased between-class spectral variability that is the consequence of improved spatial resolution. Likewise, the use of SVM-based kernel methods ensures robustness to the Curse of Dimensionality and efficient use of high dimensional data without the requirement for feature reduction. Altogether, although no multi-temporal dataset of the kind is currently available to our investigation, these attributes make the classifier well-suited to the classification of hyperspectral and high-resolution remote sensing imagery. In the meantime, the classification of multi-temporal moderate-resolution multispectral imagery was shown to be greatly improved by the
proposed classifier over previous methods for spatial-temporal contextual classification.

5.2 Future Work

While the method of multi-temporal land-cover mapping within the post-classification comparative framework for detecting change has been significantly advanced with this thesis, its organizing framework still depends fundamentally on the classification of land cover at single dates. These classifications are based on training data that typically contains limited contextual information due to common point sampling schemes in homogenous areas and lack of repeat temporal sampling. In the future, data collection methods as well as compensatory classification procedures that can account for these limitations could improve classification results further. As well, the use of the non-parametric SVM classifier in this study offers a significant advantage to discrimination due to its max-margin formulation and robustness to the Hughes phenomena, but also lacks clear methodology for estimating pixel-wise classification confidence. In the future, integration of the Markovian kernel with classifiers that can provide such measures could be beneficial to mapping efforts. Finally, methods that can break with the tradition of classifying land cover at single dates and directly operate on time series of spectral information (without requiring assumptions about the form of land-cover change as in the profile-based approaches reviewed in Section 2.4) offer the opportunity to better utilize the dense time series accumulating in remote sensing archives. The accurate and spatial-temporally consistent classification of such time series will be crucial to realizing the full potential of the land change science for understanding our changing Earth and the myriad effects we humans are having on it.
Chapter 6: References


Heidelberg: Springer.


