Land Cover Change Detection: A Case Study

Shyam Boriah
University of Minnesota
sboriah@cs.umn.edu

Vipin Kumar
University of Minnesota
kumar@cs.umn.edu

Michael Steinbach
University of Minnesota
steinbac@cs.umn.edu

Christopher Potter
NASA Ames Research Center
chris.potter@nasa.gov

Steven Klooster
CSU Monterey Bay
sklooster@gaia.arc.nasa.gov

ABSTRACT

The study of land cover change is an important problem in the Earth Science domain because of its impacts on local climate, radiation balance, biogeochemistry, hydrology, and the diversity and abundance of terrestrial species. Most well-known change detection techniques from statistics, signal processing and control theory are not well-suited for the massive high-dimensional spatio-temporal data sets from Earth Science due to limitations such as high computational complexity and the inability to take advantage of seasonality and spatio-temporal autocorrelation inherent in Earth Science data. In our work, we seek to address these challenges with new change detection techniques that are based on data mining approaches. Specifically, in this paper we have performed a case study for a new change detection technique for the land cover change detection problem. We study land cover change in the state of California, focusing on the San Francisco Bay Area and perform an extended study on the entire state. We also perform a comparative evaluation on forests in the entire state. These results demonstrate the utility of data mining techniques for the land cover change detection problem.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; J.2 [**Computer Applications**]: Physical sciences and engineering—*Earth and atmospheric sciences*

General Terms

Algorithms

1. INTRODUCTION

Remote sensing data consisting of satellite observations of the land surface, biosphere, solid Earth, atmosphere, and oceans, combined with historical climate records and predictions from ecosystem models, offer new opportunities for understanding how the Earth is changing, for determining what factors cause these changes, and for predicting future changes. One important area where remote sensing plays a key role is in the study of land cover change. Specifically, the conversion of natural land cover into human-dominated cover types continues to be a change of global proportions with many unknown environmental consequences. For example, studies [13, 20] have shown that deforestation has significant implications for local weather, and in places such as the Amazon rainforest, cloudiness and rainfall are greater over cleared land than over intact forest. Thus, there is a need in the Earth Science domain to systematically study land cover change in order to understand its impacts on local climate, radiation balance, biogeochemistry, hydrology, and the diversity and abundance of terrestrial species. Land cover conversions include tree harvests in forested regions, urbanization, and agricultural intensification in former woodland and natural grassland areas. These types of conversions also have significant public policy implications due to issues such as water supply management and atmospheric CO_2 output. For example, Charbonneau and Kondolf [9] found that in California between 1984 and 1990, over half of all new irrigated farmland put into production was of lesser quality than prime farmland taken out of production by urbanization.

The land cover change detection problem is essentially one of detecting when the land cover at a given location has been converted from one type to another. Examples include the conversion of forested land to barren land (possibly due to deforestation or a fire), grasslands to golf courses and farmland to housing developments. There are a number of factors that make this a challenging problem including the nature of Earth Science data, which we will discuss later in this paper. Change detection, in general, is an area that has been extensively studied in the fields of statistics [22], signal processing [19] and control theory [24]. However, most techniques from these fields are not well-suited for the massive high-dimensional spatio-temporal data sets from Earth Science. This is due to limitations such as high computational complexity and the inability to take advantage of seasonality and spatio-temporal autocorrelation inherent in Earth Science data.

There are a number of problems in the Earth Science domain that have a data mining requirement due to the unique challenges posed by the types of data encountered. There have been several recent applications of data mining techniques to Earth Science problems [15, 28, 31, 32] using a variety of data types ranging from remote-sensing data to data obtained from climate models. The land cover change detection problem is also one where data mining techniques can have a significant impact. In particular, with the in-

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

creasing spatial and temporal resolutions of the underlying data sets the use of efficient and scalable pattern discovery algorithms is paramount. In our work, we seek to address these challenges with new change detection techniques that are based on novel data mining approaches. Specifically, these techniques will take advantage of some of the inherent characteristics of spatio-temporal data and will be scalable so that they can be applied to increasingly high-resolution Earth Science data sets. The long term goal of our work is to determine where, when, and why natural ecosystem conversions occur, which is a crucial scientific concern.

In this paper we have performed a case study for a new change detection technique for the land cover change detection problem. Specifically, we study land cover change in the state of California, focusing on the San Francisco Bay Area and perform an extended study on the entire state. We also perform a comparative evaluation on forests in the entire state. We examine the results obtained by applying the new technique to the Bay Area and California data sets. For the forests data set, the proposed algorithm is systematically compared with a previously proposed technique from the domain of Earth Science using ground truth information to verify the results of the algorithms. These results demonstrate the utility of data mining techniques for the land cover change detection problem. Finally, we also discuss possible directions for future work in using data mining techniques for the problem of land cover change detection.

1.1 Key Contributions

The key contributions of this paper are as follows:

- We present the land cover change detection problem and provide a discussion of the important challenges from a data mining perspective.
- We present an algorithm for land cover change detection that is simple, efficient and takes advantage of the inherent structure in Earth Science data.
- We systematically evaluate our algorithm by applying it to data sets for the San Francisco Bay Area region as well as the the entire state of California.
- We comparatively evaluate our proposed algorithm and an algorithm from the Earth Science domain and show that our algorithm gives high-quality results.

1.2 Organization of the Paper

The rest of the paper is organized as follows. In Section 2, we describe Earth Science data sets that we have used in this work as well as the data cleaning procedure used. We discuss previous related work in the area of change detection in Section 4. In Section 5 we present the two change detection algorithms evaluated in this paper. Section 6 presents the results of applying the proposed algorithm to two data sets, and a comparative evaluation on a third data set.

2. EARTH SCIENCE DATA

The Earth Science data for our analysis consists of snapshots of measurement values for a vegetation-related variable collected for all land surfaces (see Figure 1). The data observations come from NASA's Earth Observation System (EOS) [2] satellites and the data sets are distributed through the Land Processes Distributed Active Archive Center (LP DAAC) [3].

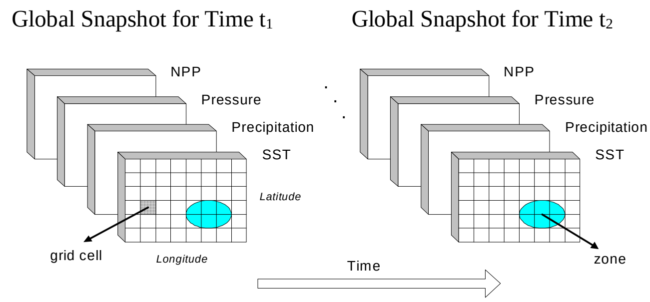

Figure 1: A simplified view of the problem domain.

The specific vegetation-related variable for this analysis was the enhanced vegetation index (EVI) product measured by the moderate resolution imaging spectroradiometer (MODIS). EVI is a vegetation index which essentially serves as a measure of the amount and "greenness" of vegetation at a particular location. It represents the greenness signal (area-averaged canopy photosynthetic capacity), with improved sensitivity in high biomass cover areas. MODIS algorithms have been used to generate the EVI index at 250-meter spatial resolution from February 2000 to the present. In this study, the data coverage is from the time period February 2000—January 2006.

In this preliminary case study, we focused our analysis on the state of California. Specifically, we apply our algorithm to EVI data for the San Francisco Bay Area, which has seen rapid population growth in recent years. We also systematically compare our algorithm with a previously proposed algorithm from the Earth Science domain on forest regions of California, since most land cover changes in forests are due to forest fires and these are easily verified. We preprocessed the data to eliminate poor-quality measurements in order to simplify evaluation. Data cleaning was done by performing the following steps:

1. The MODIS data sets are tagged with a quality assurance (QA) flag which is used to describe atmospheric and sensor conditions when the measurement was taken. We used the QA flag to remove all measurements that were tagged as being of low quality. Another filtering step recommended by Earth Science domain experts was the removal of measurements of EVI above 0.9.

2. We also discarded any locations that contained missing data. Therefore, the data for a location is retained only if the entire time series is available with no missing values and no low quality data.

The final quality-filtered EVI data set for the San Francisco Bay Area contained 180,400 locations (covering a region of 100 miles $\times$ 50 miles), the entire California data set contained over 5 million locations (covering a region of 800 miles $\times$ 200 miles), and the data set for forest locations in the state of California contained 380,285 locations. The length of the time series for all three data sets is 76, corresponding to 6 years and 4 months of monthly data from February 2000 through May 2006.

Source: Google Maps.

Figure 2: This figure shows an example of a change point in the San Francisco Bay Area which corresponds to a new golf course constructed in Oakland, CA. This golf course was built in 2003, which corresponds to the time step at which the time series exhibits a change.

3. THE LAND COVER CHANGE DETECTION PROBLEM

The land cover change detection problem studied in this paper is essentially one of taking a data set of vegetation-related time series and detecting changes by giving each location a change score based on the extent to which it is considered a change point. There are a number of specific challenges associated with Earth Science data that make this a challenging problem. The *spatio-temporal* nature of Earth Science data is especially challenging since traditional data mining techniques do not take advantage of the spatial and temporal autocorrelation present in such data. In particular, change detection becomes challenging since changes in vegetation levels are occurring all the time, i.e., due to the seasonal growth cycles the EVI index is constantly fluctuating. However, these changes in vegetation levels are usually uninteresting for the land cover change detection problem, although they may be useful for tasks such as land cover classification that are outside the scope of this work.

Furthermore, vegetation-related data sets are often of high spatial resolution, which poses computational challenges. Finally, there is the issue of high-dimensionality since long time series are common in Earth Science (and the temporal resolution is increasing).

Figure 2 shows an example of a land cover change pattern that is typically of interest to Earth Scientists. The time series shows an abrupt jump in EVI in 2003. The location of the point corresponds to a new golf course, which was in fact opened in 2003. Changes of this nature can be detected only with high-resolution data.

4. RELATED WORK

The problem of land cover change detection can be tied to previous work in two broad themes: change detection in general (in the fields of statistics, signal processing, etc.), and the specific land cover change problem studied in Earth Science. We will discuss the previous work related to this problem by organizing the discussion into these two themes.

The change detection problem has a very rich literature in the fields of statistics, signal processing and control theory. Although change detection has been studied for a very long time (at least since the 1920s [7]), most activity has occurred in the last several decades due to the growth of information technology. This has led to two developments: the increasing complexity of virtually all activities that use technology, and more detailed monitoring of such activities. This ranges from monitoring of the heart with implanted devices to satellite sensors that are used to detect vegetation levels on the Earth's surface. As a result, there are a variety of change detection problems with differing requirements and a large number of techniques seek to address specific problems. However, most of the previous work in change detection is not suitable for our land cover change detection problem for two reasons: (1) the techniques are computationally expensive, making them infeasible for large scale high-resolution Earth Science data sets, and (2) the techniques are unable to take advantage of the inherent structure present in Earth Science data. The monograph by Brodsky and Darkhovsky [7], and the books by Basseville and Nikiforov [6] and Gustafsson [19] comprehensively discuss the major developments in change detection.

We briefly list a few recent developments in change detection that are relevant in the data mining context. Chen and Gopalakrishnan [11] presented the problem of acoustic change detection where the goal is to segment an audio signal into homogeneous sections based on the speaker and surrounding conditions. Event detection, presented by Guralnik and Srivastava [18], is the problem of recognizing the change of parameters in the underlying model or the change of the model itself at unknown times; given a time series and a set of basis functions, the problem is to find a piecewise segmented model where each segment is represented by a basis function. Therefore, the number of change points, their locations in time, and the basis functions for each segment must all be determined. Ge and Smyth [17] proposed a semi-Markov (Markov model combined with a Bayesian update) model-based technique which uses state switching to detect change points. Their framework models individual segments using regression functions, and each segment is a state in an HMM; a change in states corresponds to a change point. A stream-based approach to change point detection was proposed by Yamanishi and Takeuchi [34]. This technique reduces the problem of change detection in time series into that of outlier detection from time series of moving-averaged scores. Chan and Mahoney [8] proposed a technique called box modeling which divides time series into boxes, and constructs a feature space based on boxes; an anomaly score is assigned based on the relationship between query time series and the boxes.

The problem of land cover change detection has been considered extensively in the Earth Science literature, particularly after the early 1970s when remote-sensing satellites came into use. The survey by Singh [30] gives an overview of the various land cover change detection techniques proposed in the 1970s and 1980s. The primary techniques in use at the time were simple thresholding, a variety of image-based methods (differencing, regression, and ratioing), and principal components analysis. The introduction of satellite instruments such as the Advanced Very High Resolution Radiometer (AVHRR) led to an increase in the quality of multispectral data leading to more sophisticated change detection algorithms being developed. Lu et al. [25] provides a very comprehensive survey of all major change detection algorithms that were proposed until 2003. Another recent survey was provided by Coppin et al. [12]. Although high-resolution data from instruments such as MODIS have been available since 2000, as seen from the recent surveys [25, 12], most change detection techniques were proposed for data of coarser resolutions. A change detection technique developed for MODIS data was proposed by Lunetta et al. [26] in 2006; they performed a case study applying their change detection technique (discussed in detail in Section 5.1) to a region known as the Albemarle-Pamlico Estuary System (this region is along the Virginia-North Carolina border). In this paper, we will compare the performance of our change detection algorithm with the one proposed in [26].

5. CHANGE DETECTION TECHNIQUES

We introduce the following notation in order to describe the land cover change detection algorithms, as well as the properties of the underlying data sets. Let D be a data set with N land locations each of which has a time series of length T. The time series for a location corresponds to T monthly EVI observations at that location. We also define the following notation:

- Y: the number of years of data in the data set. In this paper, we will work with complete years by truncating trailing months. Thus $Y = \frac{T}{12}$.
- n_i: an individual location.
- n_{ij}: a particular month of data for the location n_i.

5.1 Earth Science Techniques for Land Cover Change Detection

Recently, a change detection study that uses MODIS data was performed by Lunetta et al. [26]. This comprehensive study consists of a data cleaning method (based on the discrete Fourier transform), a change detection method, as well as an extensive discussion of the change patterns in the region of interest. They also performed a case study applying their change detection technique to a region known as the Albemarle-Pamlico Estuary System. For our study, the change detection method is of most relevance since it is one of the few change detection methods that has been applied to high-resolution MODIS data. Although Lunetta et al. do not work with EVI data in their paper, they do work with a closely related variable called the Normalized Difference Vegetation Index (NDVI). NDVI, like EVI, is also a measure of the vegetation level at a given land location, the main difference being that EVI is designed to enhance the vegetation signal with improved sensitivity in high biomass (high vegetation density) regions [21].

Since NDVI and EVI are similar variables, we will apply Lunetta et al.'s scheme to EVI data in this paper. Their change detection methodology essentially works with annual sums of NDVI for a given land location. The difference between consecutive years is then computed; this is equivalent to applying first-order differencing [10] to the time series of annual sums. The resulting differences are assumed to follow an approximately normal distribution with $\mu = 0$, which represents that no change has occurred. This is justifiable in that most locations do not exhibit change and therefore on average the change in annual sum between consecutive years is expected to be 0. To detect whether a change has occurred between time t_1 and time t_2, the z-score of the difference of annual sums is computed (the standard deviation is computed with respect to all land locations) and if the z-score is above a threshold, a change is considered to have occurred between t_1 and t_2. The specific steps in Lunetta et al.'s algorithm for land cover change detection are as follows:

1. For each location n_i, the annual EVI sum is computed for each year of data. Let $\{a_{i1}, \ldots, a_{iY}\}$ correspond to this list of annual sums, where $a_{i1} = \sum_{j=1}^{12} n_{ij}$, etc.
2. The difference between the annual sum for a given year and the previous year is computed, i.e., $\{a_{i2} - a_{i1}, a_{i3} - a_{i2}, \ldots, a_{iY} - a_{iY-1}\}$. Let $d_{ij} = a_{ij+1} - a_{ij}$.
3. The z-score is computed for each of the $Y - 1$ values in $\{d_{i1}, d_{i2}, \ldots, d_{iY-1}\}$. This is done for each d_{ij} by subtracting the mean (set to 0) and dividing by the standard deviation ($= st.\ dev.\{d_{1j}, d_{2j}, \ldots, d_{Nj}\}$). Let $\{z_{i1}, z_{i2}, \ldots, z_{iY-1}\}$ correspond to this list of z-scores.
4. If z_{ij} is above a certain threshold, then location n_i is considered to have a change between the consecutive years corresponding to j and $j + 1$.

In this paper we have slightly modified Lunetta et al.'s algorithm in order to apply it to our change detection problem. Essentially, after applying Lunetta et al.'s algorithm to the data set, we are left with a list of z-scores for each location that corresponds to each pair of consecutive years. Since a single change score is assigned to each location in our problem setting, we adapt Lunetta et al.'s algorithm by taking the absolute maximum for each location's list of z-scores to be the change score for the location, i.e.,

$$\textit{change score}(n_i) = \max\{|z_{i1}|, |z_{i2}|, \ldots, |z_{iY-1}|\}.$$

5.2 Recursive Merging Algorithm

We now present our technique (called the recursive merging algorithm) for land cover change detection. We designed this technique based on some key characteristics of the underlying data set: (1) most land locations do not exhibit a change; given the large coverage of land cover data sets, it is fairly obvious that only a small fraction of points will actually exhibit a change, (2) the major mode of behavior in the vegetation signal is seasonality, i.e., the natural seasonal growing cycle is a dominant characteristic of a time series and this intrinsic seasonality should not itself be called a change. The main idea behind the recursive merging algorithm is to exploit seasonality in order to distinguish between points that have had a land cover change and those that have not. In particular, if a given location has not had a land cover change, then we expect the seasonal cycles to look very similar going from one year to the next; if this is not the case, then based on the extent to which the seasons are different one can assign a change score to a land location. The time series for each location is processed as follows:

1. The two most similar consecutive annual cycles are merged, and the distance is stored. Let $\{b_{i1}, \ldots, b_{iY}\}$ correspond to the list of annual cycles, where $b_{i1} = [n_{i1}\ n_{i2}\ \ldots\ n_{i12}]$, etc. Suppose b_{i1} and b_{i2} are the two most similar annual cycles; then, at the end of this step what is left is a list with 1 less element, $\{\frac{b_{i1}+b_{i1}}{2}, b_{i3} \ldots, b_{iY}\}$, along with the distance $s_1 = dist(b_{i1}, b_{i2})$.

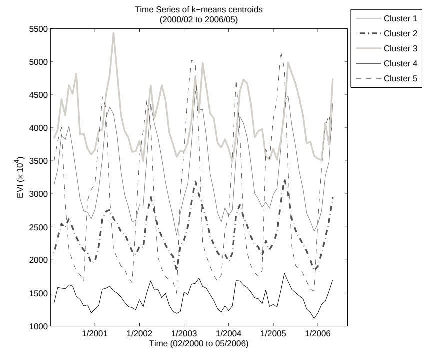

Figure 3: k-means cluster centroids for the Bay Area EVI data set.

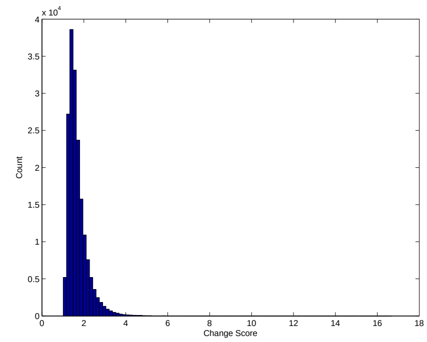

Figure 4: Histogram of change scores produced by the recursive merging algorithm for the Bay Area data set.

2. Step 1 is applied recursively until one annual cycle is left remaining. This results in a list of distances of length $Y-1$, $\{s_1, \ldots, s_{Y-1}\}$. Note that the order in which items are inserted into this list is not related to the order of the annual cycles, it is merely the order in which seasonal cycles were merged.

3. The change score for this location is based on whether any of the observed distances are extreme. This can be quantified by computing the quantity $\frac{\max\{s_1,\ldots,s_{Y-1}\}}{\min\{s_1,\ldots,s_{Y-1}\}}$.

Note that it is possible that $\min\{s_1, \ldots, s_{Y-1}\}$ might be 0, in which case a small value ϵ can be added to this quantity.

Our technique bears some resemblance to the bottom-up segmentation algorithm discussed by Keogh et al. [23]: in the first step, the original time series of length n is replaced with an initial approximation of length $n/2$. Then, the cost of merging each pair of segments is computed and the pair with the lowest cost is merged. This procedure is repeated until a user-supplied threshold for the cost of merging is reached.

The computational complexity of the recursive merging algorithm is $O(NT)$ (the cost of processing a single location is $T-12$ and there are N locations). The storage requirement is $O(N)$.

6. APPLICATION OF CHANGE DETECTION TECHNIQUES TO EVI DATA

In this preliminary study, we focused our analysis on the state of California primarily because our team members have domain expertise in this region. In addition, due to high population growth in recent years, California has experienced many land use changes such as the urbanization of farmland and desert converted to farmland. The state also has many forest fires each year that can lead to temporary changes in the land cover.

6.1 Analysis of the San Francisco Bay Area

We initially performed a clustering of the EVI data for the San Francisco Bay Area region in order to observe the high-level characteristics of the vegetation data. We used the k-means algorithm to cluster the EVI data to produce a minimum number of distinct centroids (in this case five centroids were found to have the optimal SSE). Figure 3 shows the cluster centroids for the Bay Area EVI data. What we observed was that most of the data was predominantly from 5 classes of vegetation thereby implying that for the vast majority of the points in the data set, there is no change in land cover. Additionally, when the clusters are viewed on a map and compared to satellite imagery, the clusters corresponded to the actual vegetation in the region very closely. The land cover type corresponding to each cluster is as follows: Cluster 1 is high seasonal biomass density, moderate interannual variability (shrub cover); cluster 2 is moderate annual biomass density, moderate interannual variability (grass cover); cluster 3 is high annual biomass density, low interannual variability (evergreen tree cover); cluster 4 is low annual biomass density, low interannual variability (urbanized cover); cluster 5 is high seasonal biomass density, high interannual variability (agricultural cover).

We then applied the recursive merging change detection algorithm to the Bay Area EVI data set. The output from the algorithm is a list of change scores, one for each location. Figure 4 shows a histogram of the change scores obtained. We observe that the distribution of the scores conforms to our expectation based on the clustering results that most of the locations in the data set do not exhibit a change. We manually examined the top 31 points which had a change score greater than a threshold of 8. Of these, 22 points were found to correspond to interesting land cover changes and others corresponded to changing crop patterns in farmland. A majority of the points with change scores less that 8 (but greater than 4) were from farmland. We briefly discuss a few of the interesting land cover changes discovered:

1. *Barren land to golf course.* Figure 2 shows the time series and map for a location where a golf course was built. Specifically, this location corresponds to the Metropolitan Golf Links in Oakland, CA which was built in 2003 on a site which was previously a disposal site for dredged material from San Francisco Bay. The time series for the location clearly shows the low level of vegetation at the location prior to 2003, after which the vegetation is relatively uniform, consistent with what is expected at a golf course that is watered throughout the year.

2. *Construction of a housing subdivision.* Figure 5 shows the time series and map for a location where a subdivision is under construction in Hayward, CA on a site which was previously vegetated, with the level of vegetation suggesting that this land may have been grass-

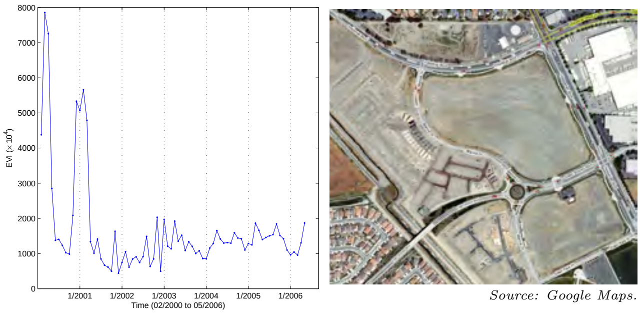

Figure 5: Construction of a subdivision in Hayward.

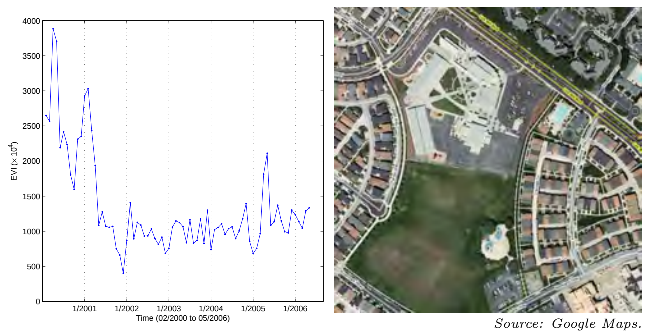

Figure 6: Construction of a shopping center in San Jose.

land prior to the construction. Interestingly, the construction of housing can actually lead to the increase in vegetation at a location such as this one since after the construction period, the area is typically planted with lawns and trees. In such situations, land cover change detection techniques that examine only the beginning and end of a time series may fail to detect this type of change.

3. *Construction of a shopping center.* Figure 6 shows the time series and map for a location where a shopping center has been built on a site which was previously vegetated in San Jose, CA. The year of construction for this shopping center was 2002, which is also when the change in vegetation-level occurs in the time series.

4. *Construction of a golf course.* Figure 7 shows the time series and map for a location where a golf course was built. This is the location of the Golf Club at Boulder Ridge in San Jose, CA. This golf course was constructed in 2001, and this is clearly seen again in the time series for the location.

5. *Construction of a shopping center.* Figure 8 shows the time series and map for a location where a shopping center has been built on a site which was previously vegetated. This location corresponds to the Pacific Commons shopping center in Fremont, CA.

These results show the effectiveness of the recursive merging algorithm on the Bay Area data set. Specifically, in examining the top ranked change locations from the algorithm, we observe that a variety of interesting land cover changes are detected. The ability of the algorithm to discount natural seasonal changes (such as those seen in Figure 3) is particularly appealing.

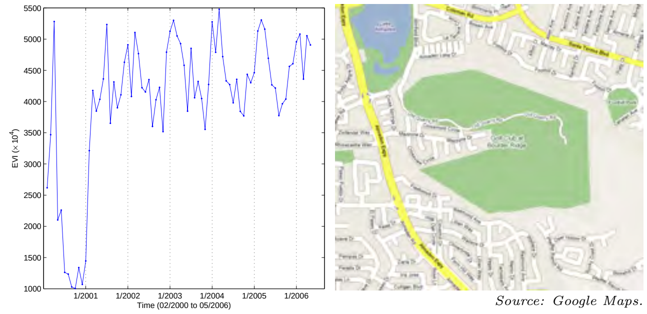

Figure 7: Construction of a golf course in San Jose.

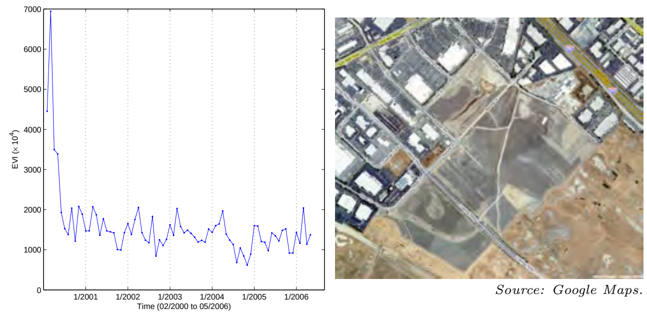

Figure 8: Construction of the Pacific Commons shopping center in Fremont.

6.2 Analysis of the entire state of California

We performed an extended study on the entire state of California, which involves about a 30-fold increase in data over the Bay Area. For this data set we found 2,833 locations that had a change score greater than a threshold of 15, a vast majority of which corresponded to land cover change points. We analyzed about 1,000 of the top change locations for the larger California data set and we found that more types of changes were detected than in the Bay Area data set. In particular, we found numerous instances where the land cover had changed from desert to farmland, forests to barren (due to forest fires), farmland to housing subdivisions, and desert to golf courses.

A few interesting land cover changes that were discovered in the California data set are shown in Figures 9, 10 and 11 (note that the figures show several time series each; these correspond to groups of spatially close locations) and described below:

1. *Desert to farmland.* This is a group of points corresponding to a change from a desert location to farmland in southern California. This type of land cover change is prevalent in California and across the United States [5, 33].

2. *Farmland to subdivision.* This is a location in Sacramento, CA where farmland has been cleared and a housing subdivision is being built. This location is in the north-west outskirts of the city, where the surrounding land currently consists of farms. This is an example of land cover change where urbanization is converting agricultural land into housing.

3. *Desert to golf course.* This is an example of a new golf course being built in Palm Desert, CA. This town

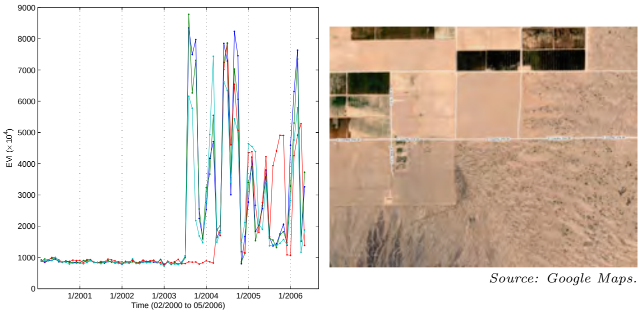

Figure 9: Desert to farmland.

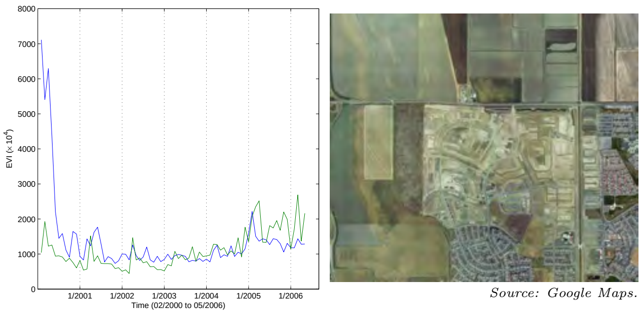

Figure 10: Farmland being converted into a housing subdivision in Sacramento.

has over 100 golf courses, putting intense pressure on the water supply [4] in a region where water is already scarce.

6.3 Forests in California

To provide a comparative evaluation, we also evaluated the recursive merging algorithm on the EVI data set consisting of forests in California in comparison to an algorithm from the Earth Science domain proposed by Lunetta et al. [26]. This specific data set was selected because most land cover changes in forests are due to forest fires, and forest fires can easily be verified using the incidents database available from the California Department of Forestry and Fire Protection [1].

Similar to the Bay Area case study, we apply a given change detection algorithm to the vegetation data set and obtain a change score for each of the 380,285 land locations. The change scores are then sorted and grouped by spatial

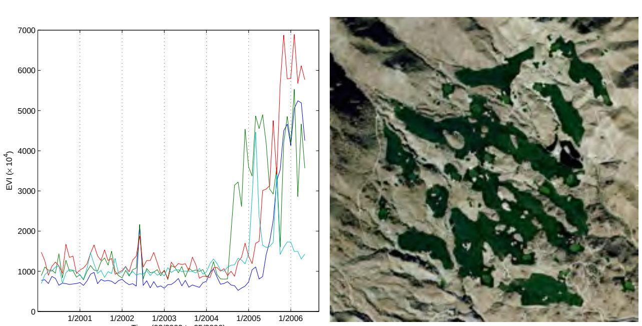

Figure 11: A newly constructed golf course in Palm Desert.

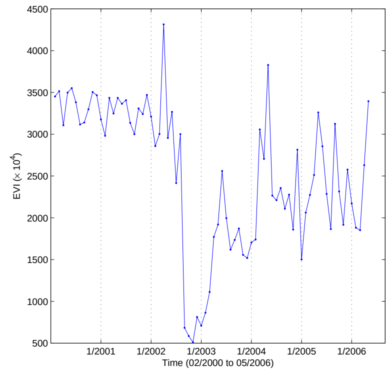

Figure 12: This figure shows an example of a land cover change due to a forest fire.

location to give "events": an event is a group of closely located pixels that exhibit the same change. For this study, the top 15 events (covering about 125 points in each case) were individually examined and the changes were verified using ground truth information. The results of applying the two change detection algorithms are shown in Table 1. We see that for all the events examined, the recursive merging algorithm finds locations with verifiable land cover changes. These include changes due to forest fires, conversion of forests to farmland and loss of forests due to construction or logging. We also examined the top points discovered by the Lunetta et al. algorithm and found that several of those corresponded to forest fires. However, a significant number of locations did not exhibit an apparent land cover change and no change could be discerned by examining the EVI time series.

The dominant land cover change in forests in California is due to forest fires. Therefore, the most common pattern expected is reduction in EVI followed by regrowth (Figure 12). There is also some construction and logging activity that takes place in forests (Figure 13). These change patterns are different from those due to forest fires in that the reduction of EVI tends to be gradual rather than abrupt. Table 2 lists examples of fires detected.

Although we were unable to do a thorough evaluation of highly scored change points except for those covered by the top 15 events, we did examine the top 5% of the total points to get a sense of the nature of land cover changes being detected by our algorithm. We found that among the top 1% of the points (approximately 38,000 points), an overwhelming majority correspond to forest fires.

These results show that our proposed algorithm for change detection gives high quality results for forest data. In particular, we showed that the top results from the recursive merging algorithm are verifiable change points. This is significant since the results of change detection algorithms are often manually examined by expert analysts from the Earth Science domain. Therefore, the analyst's time is used more effectively if the algorithm produces high-quality change points in the top portion of the results.

Table 3 shows the total number of forest fires as well as the number of large fires (those that burn more than 300 acres) in California between 2000 and 2006. Two things that stand out from the Table are: (1) there are thousands of verified fires in California every year, and (2) the large

Land Cover Change Detected	Lunetta et al. Algorithm	Recursive Merging Algorithm
Forest fires	5	12
Conversion to farming	1	2
Construction or logging	0	1
No apparent change	8	0

Table 1: Results of the change detection algorithms. Note that these are events and each event corresponds to a group of pixels. The total number of pixels covered in each case is about 125.

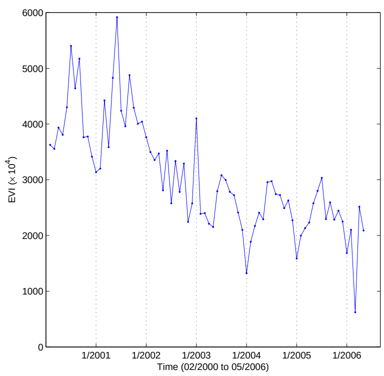

Figure 13: This figure shows an example of a land cover change due to construction or logging.

Year	Forest Fire	Detected by algorithm(s)
September 2002	Curve	Recursive Merging
June 2002	Troy	Recursive Merging
June 2002	Wolf	Lunetta et al., Recursive Merging
July 2002	Pines	Recursive Merging
May 2004	Cachuma	Recursive Merging
Mid-2003	Spanish	Recursive Merging
Late 2003	Grand Prix	Lunetta et al., Recursive Merging
October 2004	Rumsey	Recursive Merging
Mid 2001	Poe	Recursive Merging
$<$ 2000	Kirk Complex	Lunetta et al., Recursive Merging
September 2001	Darby	Recursive Merging
Mid-2004	Geysers	Recursive Merging
Late 2003	Cedar	Lunetta et al.

Table 2: Examples of fires detected by the change detection algorithms.

Year	Total Number of Fires	Large Fires ($>$ 300 acres)
2000	5,177	59
2001	6,223	74
2002	5,759	88
2003	5,961	93
2004	5,574	82
2005	4,908	73
2006	4,805	107

Table 3: Forest fires in California for each year from 2000—2006.

fires are a fraction of the total number of verified fires in the state. From this situation, it can be inferred that some small fires may occur every year that are not seen or verified, especially in sparsely populated regions (i.e., these small fires occur in remote areas where they are not observed but some forest has been destroyed). Change detection algorithms have a significant role to play in detecting such burned regions. There have been a number of studies that have shown the utility of vegetation data for detecting burned regions [27, 14, 29, 16]. Most of these studies were performed in the 1990s on coarse-resolution vegetation data such as those from the AVHRR instrument. The results from our study with forest data suggest that the recursive merging algorithm has the ability to detect forest fires from vegetation data, and thus this approach could potentially be used to monitor fires on a global scale.

7. CONCLUDING REMARKS

In this paper, we have performed a case study showing the application of change detection algorithms for the problem of land cover change detection. We focused on three data sets for the state of California corresponding to the San Francisco Bay Area, the entire state, and forests in the state. We applied two change detection algorithms: the Lunetta et al. algorithm proposed by domain experts from Earth Science and the recursive merging algorithm proposed in this paper. The recursive merging algorithm was applied to the Bay Area data set and the results showed that it discovers high-quality change points. We also applied both algorithms to the forest data set in order to compare their performance and found that the locations given high change scores by the recursive merging algorithm are better-quality change locations. These encouraging results mean that this algorithm is effective for the land cover change detection problem.

However, there are a number of limitations of the current algorithm which require further work. The current algorithm does not make use of spatial information that is present in the data. Since Earth Science data sets exhibit significant spatio-temporal autocorrelation, the use of spatial information can be expected to give superior performance. Some patterns—such as changing crops on farmland—may not be of interest to the analyst, in which case a scheme must be devised to filter such points. Another interesting extension would be to use clustering to discover the dominant land cover patterns in the data, and then characterize the changes in terms of clusters. This approach appears promising since there are a few major classes of data (forest, farmland, shrubland, urban, etc.) and characterizing changes in terms of these major classes increases the information the algorithm provides to the analyst. In this preliminary study, we have not used statistics that are commonly used for land cover change detection [26], such as accuracy, kappa, omission error (false negative rate) and commission error (false positive rate). A more thorough evaluation of the results that incorporates such statistics and examines a much larger set of pixels must be done.

8. ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their comments and suggestions, which improved this paper. We would also like to thank Joseph Knight, Tushar Garg and Fernando Torre for their helpful comments. This work was supported by NASA Grant NNA05CS45G, NSF Grant CNS-0551551, and NSF Grant IIS-0713227. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

9. REFERENCES

[1] California Department of Forestry and Fire Protection Incidents Database. `http://cdfdata.fire.ca.gov/incidents/incidents`.

[2] NASA Earth Observing System. `http://eospso.gsfc.nasa.gov`.

[3] Land Processes Distributed Active Archive Center. `http://edcdaac.usgs.gov`.

[4] In the California desert, they use water like there's no tomorrow—but tomorrow is coming. *U.S. Water News Online*, June 2003.

[5] A. Allen. Environmental planning and management of the peri-urban interface: perspectives on an emerging field. *Environment and Urbanization*, 15(1):135–148, 2003.

[6] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

[7] B. Brodsky and B. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, 1993.

[8] P. K. Chan and M. V. Mahoney. Modeling multiple time series for anomaly detection. In *Proceedings of the 5th IEEE International Conference on Data Mining*, pages 90–97, 2005.

[9] R. Charbonneau and G. Kondolf. Land use change in California, USA: Nonpoint source water quality impacts. *Environmental Management*, 17(4):453–460, 1993.

[10] C. Chatfield. *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, 2004.

[11] S. S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[12] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin. Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, 25(9):1565–1596, 2004.

[13] R. Dickinson and P. Kenneday. Impacts on regional climate of Amazon deforestation. *Geophysical Research Letters*, 19 (19):1947–1950, 1992.

[14] H. Eva and E. F. Lambin. Burnt area mapping in Central Africa using ATSR data. *International Journal of Remote Sensing*, 19(18):3473–3497, 1998.

[15] X. Fern, C. E. Brodley, and M. A. Friedl. Correlation clustering for learning mixtures of canonical correlation models. In *SDM 2005: Proceedings of the 5th SIAM International Conference on Data Mining*, pages 439–448, 2005.

[16] R. H. Fraser, Z. Li, and J. Cihlar. Hotspot and NDVI Differencing Synergy (HANDS): A New Technique for Burned Area Mapping over Boreal Forest. *Remote Sensing of Environment*, 74(3):362–376, 2000.

[17] X. Ge and P. Smyth. Segmental semi-Markov models for change-point detection with applications to semiconductor manufacturing. Technical Report UCI-ICS 00-08, University of California, Irvine, 2000.

[18] V. Guralnik and J. Srivastava. Event detection from time series data. In *KDD '99: Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42, 1999.

[19] F. Gustafsson. *Adaptive Filtering and Change Detection*. John Wiley & Sons, 2000.

[20] A. Henderson-Sellers, R. E. Dickinson, T. B. Durbidge, P. J. Kennedy, K. McGuffie, and A. J. Pitman. Tropical deforestation: Modeling local- to regional-scale climate change. *Journal of Geophysical Research*, 98:7289–7315, 1993.

[21] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1-2):195–213, 2002.

[22] C. Inclán and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427): 913–923, 1994. ISSN 0162-1459.

[23] E. Keogh, S. Chu, D. Hart, and M. Pazzani. Segmenting time series: A survey and novel approach. In *Data mining in Time Series Databases*. World Scientific Publishing Company, 2003.

[24] T. L. Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4): 613–658, 1995.

[25] D. Lu, P. Mausel, E. Brondízio, and E. Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2003.

[26] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sensing of Environment*, 105(2):142–154, 2006.

[27] M. Matson and B. Holben. Satellite detection of tropical burning in Brazil. *International Journal of Remote Sensing*, 8(3):509–516, 1987.

[28] D. Mazzoni, K. Wagstaff, and M. C. Burl. Active learning with irrelevant examples. In *ECML 2006: Proceedings of the 17th European Conference on Machine Learning*, volume 4212 of *Lecture Notes in Computer Science*, pages 695–702. Springer, 2006.

[29] J. Pereira. A comparative evaluation of NOAA/AVHRR vegetation indexes for burned surface detection and mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):217–226, 1999.

[30] A. Singh. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003, 1989.

[31] A. N. Srivastava and J. Stroeve. Onboard detection of snow, ice, clouds and other geophysical processes using kernel methods. In *Proceedings of the ICML 2003 Workshop on Machine Learning Technologies for Autonomous Space Sciences*, 2003.

[32] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 446–455, 2003.

[33] W. C. Sullivan, O. M. Anderson, and S. T. Lovell. Agricultural buffers at the rural-urban fringe: an examination of approval by farmers, residents, and academics in the midwestern United States. *Landscape and Urban Planning*, 69(2-3):299–313, 2004.

[34] K. Yamanishi and J. Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *KDD '02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 676–681, 2002.