Monitoring Global Forest Cover Using Data Mining

VARUN MITHAL,† ASHISH GARG,† SHYAM BORIAH,†
MICHAEL STEINBACH and VIPIN KUMAR‡, University of Minnesota
CHRISTOPHER POTTER and STEVEN KLOOSTER, NASA Ames Research Center
JUAN CARLOS CASTILLA-RUBIO, Planetary Skin Institute

Forests are a critical component of the planet's ecosystem. Unfortunately, there has been significant degradation in forest cover over recent decades as a result of logging, conversion to crop, plantation, and pasture land, or disasters (natural or man made) such as forest fires, floods, and hurricanes. As a result, significant attention is being given to the sustainable use of forests. A key to effective forest management is quantifiable knowledge about changes in forest cover. This requires identification and characterization of changes and the discovery of the relationship between these changes and natural and anthropogenic variables. In this paper, we present our preliminary efforts and achievements in addressing some of these tasks along with the challenges and opportunities that need to be addressed in the future. At a higher level, our goal is to provide an overview of the exciting opportunities and challenges in developing and applying data mining approaches to provide critical information for forest and land use management.

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, Design, Performance

Additional Key Words and Phrases: Computational sustainability, forest cover change, land change, remote sensing

ACM Reference Format:
Mithal, V., Garg, A., Boriah, S., Steinbach, M., Kumar, V., Potter, C., Klooster, S. A., and Castilla-Rubio, J. C. 2011. Monitoring Global Forest Cover Using Data Mining. ACM Trans. Intell. Syst. Technol. V, N, Article A (January 2011), 23 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The world's forests are a critical component of the planet's ecosystem. They play an important role in the composition of the air we breathe, provide ecological diversity, protect the soil, and maintain the hydrological cycle [Bonan 2008]. Timber and other forest products are a primary source of livelihoods for millions around the world. In addition, forests reduce atmospheric greenhouse gases through the process of carbon sequestration. Specifically, recent research [Ollinger et al. 2008] suggests that the role forests play in regulating global climate is larger than previously thought.

Unfortunately, there has been significant degradation of forest cover over recent decades as a result of logging, conversion to cropland or plantation (e.g. Figure 1), and natural or man-made disasters such as forest fires, floods, and hurricanes [Potter et al. 2003]. Although natural disasters play a role, a large fraction of deforestation is due

† These authors contributed equally to this work.
‡ To whom correspondence should be addressed. E-mail: kumar@cs.umn.edu

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

Source: NASA Earth Observatory.

Fig. 1. This figure shows deforestation in the region east of Santa Cruz de la Sierra, Bolivia. The rectilinear, light-colored areas are fields of soybeans cultivated for export.

to direct human activity as a result of increasing economic, social and demographic pressures. Regardless of the cause, the impacts of deforestation can be significant and long lasting. For example, logging can lead to a change in microclimate and regional ecosystem characteristics (see Figure 2) as the original vegetation may never regenerate.

Deforestation and forest degradation account for as much as 20% of anthropogenic greenhouse gas emissions (GHG) and are a significant driver of climate change, which, in turn can impact the health of the global ecosystem [Gullison et al. 2007]. As an example of this complex interplay, Canadian forests may have changed from a sink of atmospheric carbon to a source caused by an increased occurrence in fires that is due to trend of increasing drought frequency and/or warmer winters [Kurz and Apps 2006].

Significant long-term impacts of deforestation and the need for sustainable forest and land use management are well recognized. Long before climate change became a driving issue, conservationists and biologists had been urging increased attention to the sustainable use of forests in general, and tropical forests in particular. Since the important role of forests in addressing the problem of climate change has become more widely recognized, there has been a significant increase in international efforts such as the United Nations Programme on Reducing Emissions from Deforestation and Forest Degradation (UN-REDD) and the inclusion of REDD in climate change discussions at the UN Framework Convention on Climate Change. Importantly, these discussions frequently include proposals to incorporate a financing mechanism in which corporations or countries that are significant emitters of atmospheric carbon offer monetary payments for forest preservation, either through direct transfers or via alternative funding mechanisms [TCG 2010].

A key ingredient for effective forest management, whether for carbon sink management or other purposes, is reliable, quantifiable observations of changes in forest cover. Vast amounts of data from remotely sensed images are now becoming available for detecting changes in forests or more generally, land cover. However, in spite

Source: NASA Earth Observatory.

Fig. 2. Deforestation changes local weather. Cloudiness and rainfall can be greater over cleared land (image right) than over intact forest (left).

of the importance of this problem and the considerable advances made over the last few years in high-resolution satellite data acquisition, data mining, and online mapping tools and services, end users still lack practical, globally scalable tools to help them manage and transform this data into actionable knowledge of changes in forest ecosystems that can be used for decision making and policy planning purposes. Providing this actionable knowledge requires innovations in a number of technical areas: (i) identification of changes in global forest cover, (ii) characterization of those changes, and (iii) discovery of relationships between the number, magnitude, and type of these changes with natural and anthropogenic variables. To realize progress in the above areas, a number of computational challenges in spatio-temporal data mining need to be addressed. Specifically, analysis and discovery approaches need to be cognizant of climate and ecosystem data characteristics such as seasonality, inter-region variability, multi-scale nature, spatio-temporal autocorrelation, high dimensionality and massive data size.

This paper describes our initial efforts, achievements and challenges in addressing some of the above areas. In particular, we present two time series change detection techniques for forest monitoring, illustrative examples of large scale vegetation disturbances, event identification, characterization and relationship mining. Additionally, to provide a practical dimension, carbon risk scoring is used to illustrate a decision making application.

Organization of the Paper: Section 2 describes the earth science datasets that are used in the study and can be used for global forest monitoring. Section 3 discusses the challenges involved in detecting changes in forest cover, limitations of current approaches, and approaches we have proposed for addressing these limitations along with illustrative examples of the vegetation disturbances detected in different parts of the globe. Sections 4 and 5 describe issues and challenges for event identification, characterization of change, and identification of the relationships between the vegetation disturbance events and natural and anthropogenic variables. Carbon risk scoring, an application of the work, is presented in Section 6. Section 7 contains concluding remarks.

2. EARTH SCIENCE DATA

Global remote sensing data sets are available from a variety of instruments at different spatial resolutions as a sequence of global snapshot of measurement values (see Figure 3). In principle, our algorithms can be applied to any geospatial dataset that features regular, repeated observations, consistent image registration and well-defined composite indicators of vegetation. In this study, we have used the Enhanced Vegetation Index (EVI) and FPAR (Fraction of Photosynthetically Active Radiation), two data

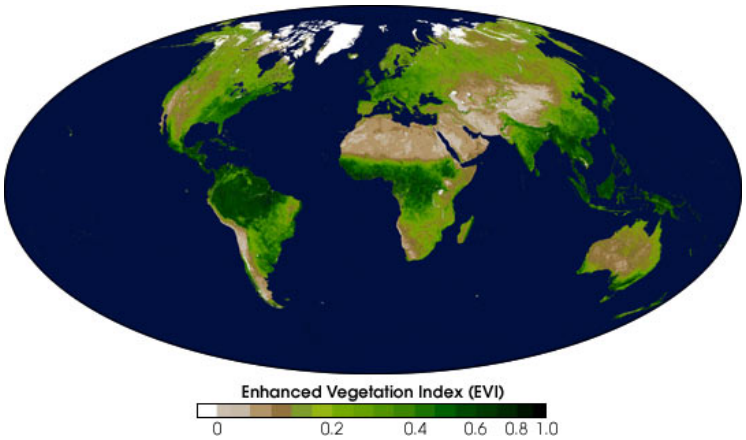


Fig. 3. The above MODIS Enhanced Vegetation Index (EVI) map shows the density of plant growth over the entire globe. Barren areas of rock, sand, or snow tend to have very low values of EVI (white and brown areas). Shrub and grassland tend to have moderate values (light greens), while high values are mostly from temperate and tropical rainforests (dark greens). Source: NASA.

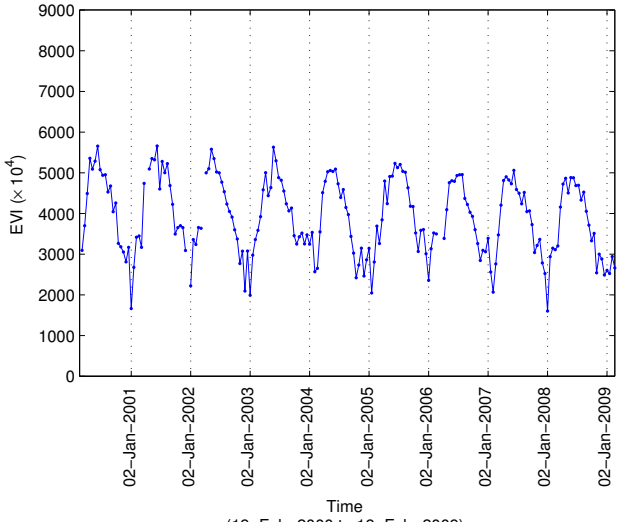
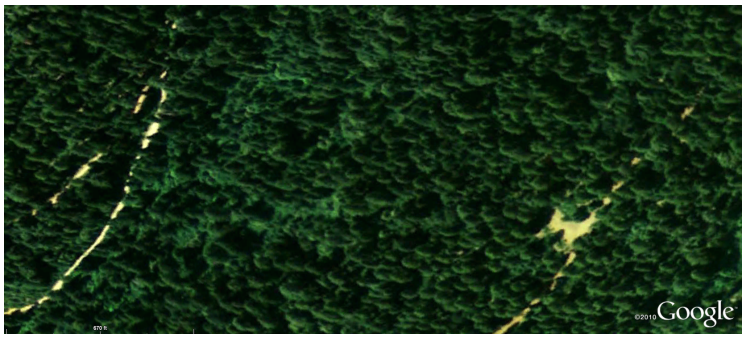


Fig. 4. An EVI time series for a specific location over a period of 109 months from Feb 2000 to Feb 2009. Note that gaps in the time series correspond to missing observations.

products based on measurements taken by the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor onboard NASA's Terra satellite. EVI essentially measures the "greenness" signal (area-averaged canopy photosynthetic capacity) as a proxy for the amount of vegetation at a particular location, while FPAR is a biophysical variable that is a measure of the primary productivity of photosynthesis. MODIS algorithms have been used to generate the EVI index at 16-day intervals at 250-meter spatial resolution and the FPAR at monthly intervals at 4 km and 1 km spatial resolution from February 2000 to the present. We process sequences of global snapshots of these indices to construct a time series for each pixel on the globe (see Figure 3 and Figure 4 for an illustration). Excluding water and permafrost areas, there are approximately



(a) Image of a region in California as it appeared in December 2005.

(b) Image of the same region in May 2009.

Source: Google Earth Imagery.

Fig. 5. Detecting land cover changes using image differencing techniques.

10 million pixels at 4 km resolution, 160 million pixels at 1 km resolution and 2500 million pixels at 250 m resolution.

3. TIME SERIES CHANGE DETECTION APPROACHES TO FOREST DISTURBANCE MONITORING

Due to the importance of the land cover change detection problem, it has received extensive attention from the remote sensing community [Coppin et al. 2004; Lu et al. 2003; Lunetta et al. 2006]. Previous change detection studies have primarily relied on examining differences between two or more satellite images acquired on different dates [Coppin et al. 2004], for example see Figure 5. These approaches have a number of limitations; for example, changes that occur outside the image acquisition windows are not mapped, it is difficult to identify when the changes occurred, and information about ongoing landscape processes cannot be derived. In addition, most of them are inherently unsuited for application at global scale.

An alternative approach is to view the data in terms of a vegetation time series at each location on the globe and identify changes in the time series (essentially provide a change score to each location and time that reflects the extent to which it is considered changed). These techniques do not suffer from the above mentioned limitations of the image based approaches. Furthermore, only time series approaches provide granular information about land cover dynamics that is necessary to quantitatively assess the carbon impact of land cover changes [Ramankutty et al. 2007]. Hence there is an increasing interest in time series based approaches to change detection in vegetation

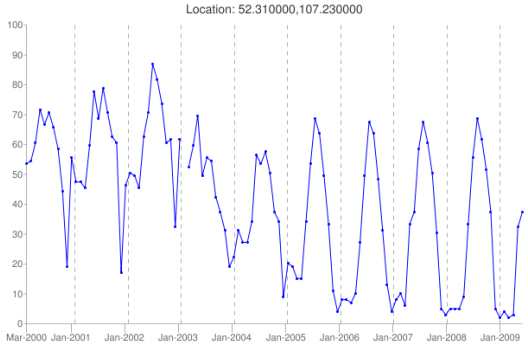


Fig. 6. The FPAR time series shows a conversion from forested area to a different land cover type in southern Siberia.

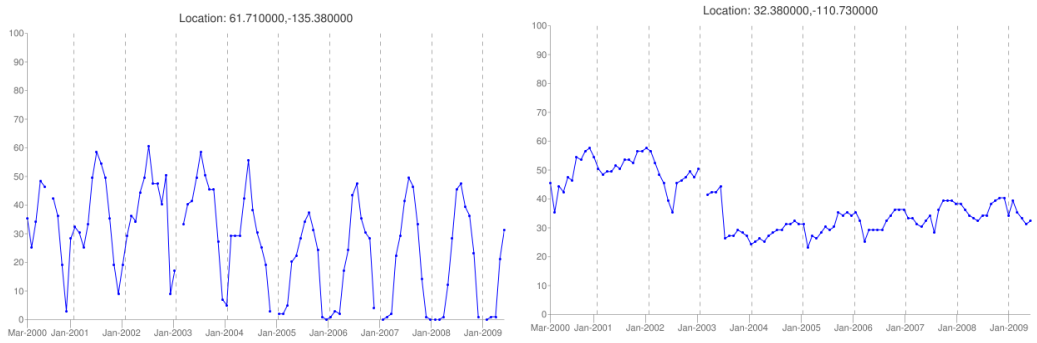
data. For example, Roy et al. [2002] use time series change detection to identify fires globally and to generate the Burned Area Product. This approach is geared to find a specific kind of change i.e. forest fires. As another example, [Lunetta et al. 2006] developed an anomaly detection based approach to identify changes in land cover that relies on the assumption that only a small part of land cover is changed in a spatial neighborhood.

Time series change detection has been studied in a wide variety of domains like statistics [Inclán and Tiao 1994], signal processing [Gustafsson 2000] and control theory [Lai 1995]. However, these techniques are not well-suited for the land cover change detection problem primarily because they are not scalable or are unable to take advantage of the inherent structure present in earth science data. For example, 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. Finding surprises or unexpected patterns in periodic data has been studied in [Keogh et al. 2002]. However, there exists an inherent natural variability and noise in the earth science data because of the local weather and other atmospheric conditions that create additional challenges for the change detection algorithms [Boriah et al. 2010].

In this section we describe two types of approaches for change detection in vegetation index time series. These approaches take as input the vegetation index time series and the annual season length for a location and give as output the change score corresponding to that location. The locations under study can be ranked according to their change score given by the algorithm. The higher ranked locations are those that are most likely to have changed.

3.1. Segmentation based approaches

Segmentation based algorithms operate under the assumption that a given time series can be partitioned into homogeneous segments and boundaries between the segments represent change points. There are two commonly used strategies to segment the time series [Keogh et al. 2001]. A top-down strategy recursively partitions the time series until a stopping criteria is met. A bottom-up strategy on the other hand recursively merges smaller units. Existing techniques for segmentation ignore many key characteristics of the underlying ecosystem data such as seasonality and variability. Here we discuss the recursive merging algorithm [Boriah et al. 2008] that follows a segmentation approach to the time series change detection problem and takes the characteristics of the ecosystem data into account.



(a) The FPAR time series showing vegetation response due to a fire in the Canadian forests.

(b) The FPAR time series corresponding to a forest deforestation in California that did not recover back.

Fig. 7. Land cover changes which have a sudden drop in vegetation response.

The main idea behind the recursive merging algorithm is to exploit seasonality in order to distinguish between points that have undergone a land cover change and those that have not. In particular, if a given location has not undergone 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.

Recursive Merging follows a bottom-up strategy of merging annual segments that are consecutive in time and similar in value. A cost corresponding to each merge is defined as a notion of the distance between the segments. We use the Manhattan distance in our implementation of the algorithm, although other similarity measures such as correlation or the L_2 norm may also be useful. The key idea is that the algorithm will merge similar annual cycles and the final merge would likely correspond to the change (if a change happened) and would have the highest cost of merging. In case the maximum cost of merging is low, it is likely that no change occurred in the time series. It is important to understand that recursive merging is designed to detect changes in which the location follows a particular vegetation model before the change and a different model after the change occurred. In this scenario, the merge with the maximum cost corresponds to the distinction between the two models.

The algorithm described above takes into account the seasonality of the data but not the variability. A high cost of merge in a highly variable time series is perhaps not as reliable an indicator of change as a moderate score in a highly stable time series. In recursive merging algorithms the cost for the initial merges can be used as an indicator of the variability within each model. In a variation of the algorithm, we define the change score as the ratio of the maximum merge cost (corresponding to difference in models) to the minimum merge cost (corresponding to the intra-model variability). Time series with a high natural variability, or time series with noisy data due to inaccurate measurement have a high minimum cost of merging also, thus a smaller change score. Experimental evaluation by Boriah [2010] and Boriah et al. [2010] shows that this step introduces a concept of variability and noise handling in the algorithm and reduces false alarms in change detection.

Segmentation based approaches are useful in detecting land cover type conversions like switching from forests to other land cover types (see example Figure 6), or changes in crop patterns, etc.

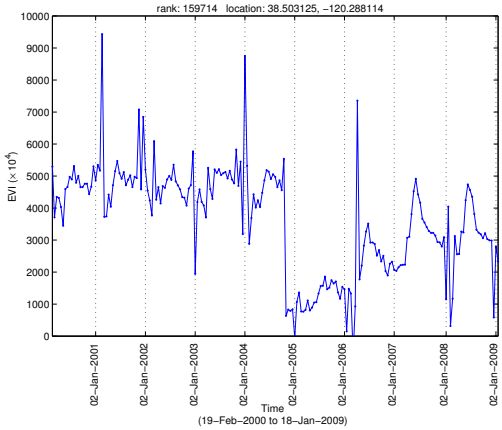


Fig. 8. An EVI time series with noise in observations.

3.2. Predictive Model based approaches

Model-based algorithms construct a model for a portion (usually the beginning) of the time series and use it to predict the future time points. When time steps outside this section are sufficiently different (i.e. it is not plausible these values could be generated by the model), these time steps are considered change points. Model-based algorithms usually have a forgetting factor that determines how much importance is attached to older observations relative to more recent observations.

There are many possibilities for the quantity to be predicted. One possibility is to predict the value of the vegetation index for the next time step (using a model built from previous values). The model can be built using approaches such as ARIMA, ARMA, etc. [Box et al. 2008]. A drawback of such schemes is that it is difficult to obtain high model accuracy due to natural variability in the ecosystem data. A more robust approach is to predict a value that is more stable e.g. the average of a number of observations. In [Boriah 2010], the previous values are used to build a model for FPAR for a 12 month period. When the actual FPAR value for the next 12 months is significantly lower than the predicted value, the corresponding time point is called a change point.

In reality, a score is assigned to each time step in each time series that is a function of the difference between actual and predicted 12 month average FPAR value. In the simplest implementation, at any time point the previous 12 months are used for model construction. This algorithm, called Yearly Delta algorithm, takes into account the seasonality of the vegetation time series since it makes use of a corresponding 12 month period for comparison with average FPAR value of the next 12 month period. A more sophisticated version to handle the variability takes several years of data into consideration, and is found to be more robust in [Boriah 2010]. The scheme described above is designed to find abrupt disturbances in vegetation, due to events like fires, floods, etc. Such events cause the vegetation index to drop abruptly. The vegetation response might recover gradually with time as shown in Figure 7(a) or may not recover for a long time after the disturbance as seen in Figure 7(b).

3.3. Additional Issues

In addition to seasonality and variability, there are several issues that need to be addressed for designing change detection algorithms for forest cover. In this section,

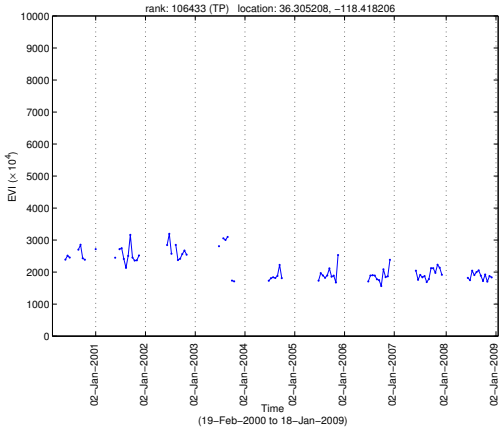


Fig. 9. An EVI time series for which a large fractions of the observations are missing.

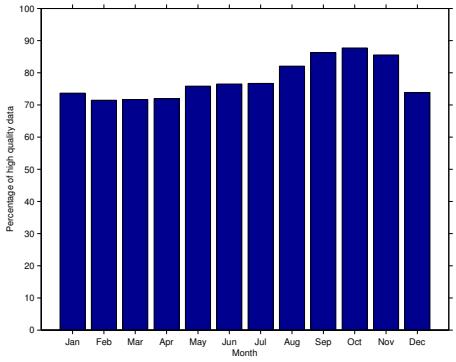


Fig. 10. Percentage of high quality data in California. Data is EVI index for the tile that includes California. Percentages are averages for each month from Feb 2000 to Feb 2009

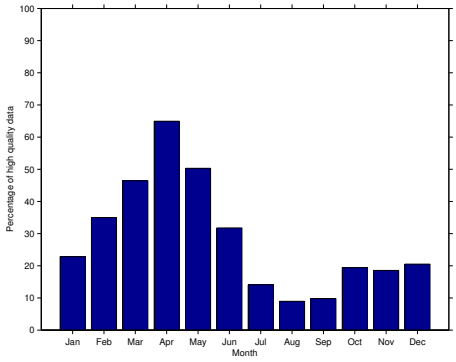


Fig. 11. Percentage of high quality data in Congo. Data is EVI index for the tile that includes the Congo. Percentages are averages for each month from Feb 2000 to Feb 2009

we provide a discussion of a number of specific challenges in global forest monitoring that need to be addressed by data mining approaches.

3.3.1. Data Quality. Earth science data sets are frequently subject to contamination due to clouds, haze, pixel geometry and other factors. Figure 8 shows a vegetation time series that has many noisy values that are visible as sharp spikes. Ideally, the change detection algorithm should not cause false alarms due to presence of these incorrect measurements. The yearly delta algorithm tries to address this issue by calculating a change score over a year. Note that if change is computed on a month-by-month basis, spikes due to noise could cause high change score and thus result in false alarms. The Recursive Merging algorithm addresses this issue by normalizing the change score. The normalization step used in Recursive Merging can help reduce the impact of noise, as a time series with large amount of noise will tend to get a smaller change score.

The observations from remote sensors come with a quality flag that is an indicator of the quality of observation. In case cloud cover or other atmospheric conditions inhibit accurate observation, the quality flag indicates lower quality observations. This quality flag information can be used to filter out data with poor quality values. Figure

9 shows a time series which contains quality data that has been filtered out. Quality issues are particularly severe in the tropics, where cloud cover predominates for many months of the year. For example, Figures 10 and 11 show the percentage of valid data for each month in California and the Congo, respectively. It can be observed that while 70–85% of the data for California is of high quality, there are several months for which only 10–20% of the data is of high quality for the Congo. These figures illustrate that data quality can differ significantly in different regions. However, more significantly, it shows that data for the tropical rainforests we seek to study can have significant quality issues. Figure 12 shows a map of the Congo region with darkness indicating the number of missing values.

One approach is to study only pixels for which there is a complete time series of high quality data; however, this approach leaves very little or no data for study for most parts of the world. Therefore, in order to monitor tropical rainforests, one must have a comprehensive strategy to address data quality issues. Data quality issues can be addressed using several approaches. From a temporal perspective, change detection algorithms could be modified to work with limited data rather than a full time series; i.e. the algorithms are robust to missing values, at the cost of losing some temporal resolution in identifying the time period of change. The two algorithms discussed above can handle missing and noisy values. The yearly delta algorithm while calculating the change score at a given time stamp, ignores the months for which either of the two yearly segments have a missing value. Recursive Merging ignores the months of missing values for calculating distance, and while merging, uses any available data for reconstructing the merged segment. Another approach to address the quality issue is to design effective spatio-temporal interpolation strategies for imputing possibly large amounts of missing data. By modeling the data in a spatio-temporal statistical framework, one can also develop robust interpolation methods that can fill-in missing values using the spatial and temporal neighborhoods with well-understood uncertainty estimates [Banerjee et al. 2008].

3.3.2. Multi-scale Analysis. Given the scale of the data (especially at fine resolution), and the diverse range of changes to be detected, it is necessary to develop techniques for multi-scale analysis that can scale to global data sets and produce high quality results. Specifically, multi-scale analysis techniques can address the following issues: (i) Different types of changes are visible at different spatial resolutions. Some events (especially those that result in slow degradation of vegetation) may be easier to detect at coarse resolution especially if they cover a large spatial extent. (ii) Global analysis using high resolution (e.g. 250m) data is difficult due to the size of the data involved. Data at coarse resolutions can be used to drive a scalable analysis approach. (iii) Due to its very nature, vegetation index data has a high degree of variability. Uncertainty due to this variability needs to be incorporated into the change detection scheme. In particular, coarse resolution data tends to have low variability but may also contain a smaller signal of change.

3.4. Illustrative Examples of Large Scale Vegetation Disturbances across the Globe

In this section we provide illustrative applications of time series based change detection algorithms applied to vegetation index data to detect a variety of changes in the global ecosystem.

3.4.1. Forest Fires. Forest fires burn millions of hectares of the world's forests every year resulting in large-scale economic damage, substantial loss of human and animal life and large amounts of carbon being released into the atmosphere [FAO 2010]. Forest fires can be human induced or due to natural causes such as lightning. In Indonesia, a large number of forest fires are triggered by land clearance for agriculture,

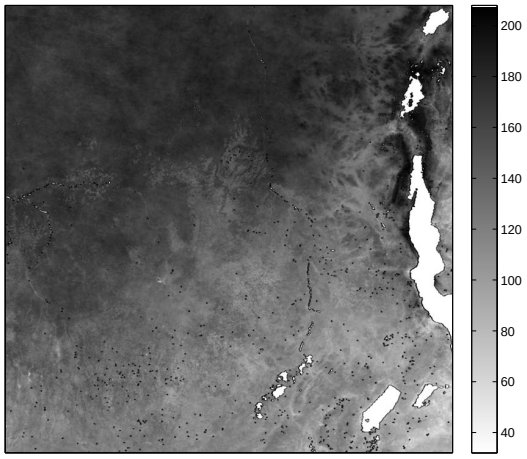


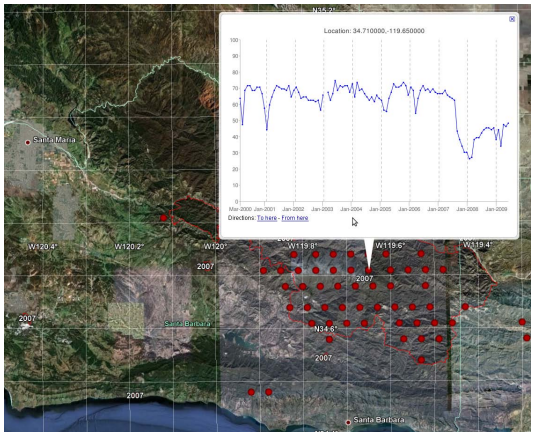
Fig. 12. The Congo region. Darkness indicates number of missing values for each pixel. Data runs from Feb 2000 to Feb 2009.

which sometimes triggers larger fires in adjacent forests. In Canada, the number of natural fires is roughly equal to the number of human induced fires, though natural fires account for 80% of the approximately 2.5M hectares of land area burned annually [Wotton et al. 2010]. Time series-based change detection algorithms can be used both for detecting the occurrence of fires, and to indicate the quantitative loss of vegetation that occurred.

Forest fires often occur on a large scale and validation data in the form of polygons is available in several regions of the world including California, Canada and Greece (validation data sets are generally maintained by government agencies). In the last decade, a large fraction of disturbances to the forests in these areas have been due to these large-scale fires. Fire events detected by our algorithm in these areas are in agreement with independent validation data. Our algorithm is also able to detect the date of occurrence of these events and the quantitative vegetation loss. Figure 13 shows a typical time series for a burnt forest pixel in the Zaca fire in 2007. The vegetation index is high until 2007, when a fire occurs, causing all pixels shown to have an abrupt drop in the vegetation index following 2007. Figure 14 shows the correspondence of our results with the validation data available in Greece. Figure 14(b) also shows the additional information our algorithm can provide in terms of the extent of damage caused by fire. In Figure 14(b) the locations are color coded with the amount of vegetation loss, dark red being maximum and yellow being minimum. The forest fire polygons in Figure 14(a) do not give this information.

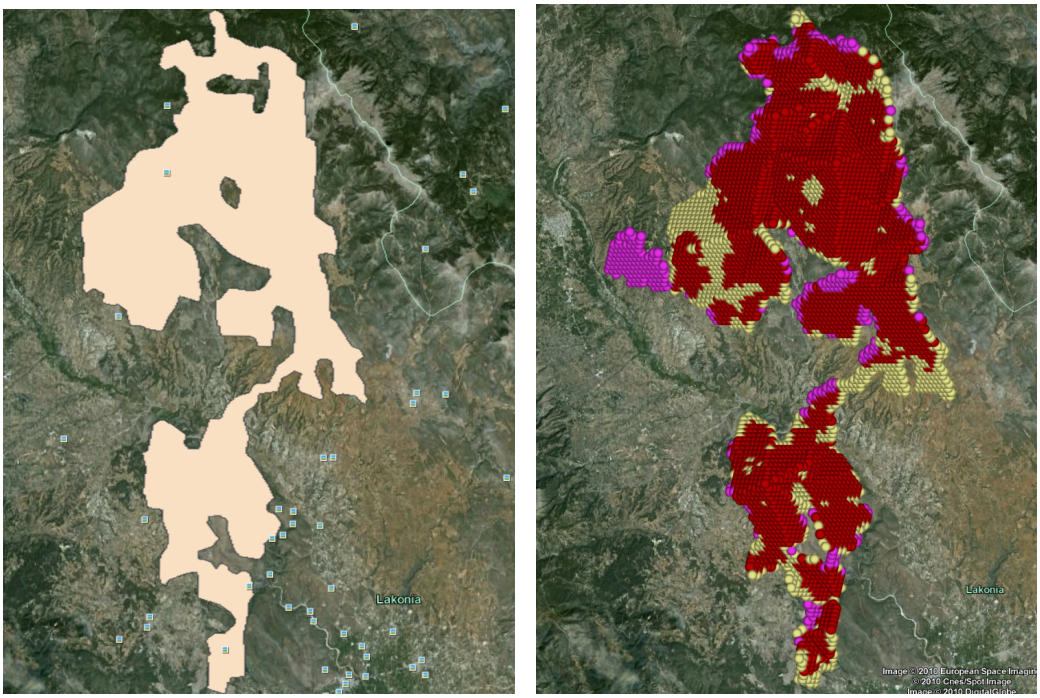
3.4.2. Deforestation. Deforestation by land conversion from forests to agricultural use is driven by many complex socio-economic factors, including macro-economic pressures as land values and commodity prices rise [Fearnside 2008]. Deforestation continues at an alarming rate (approximately 13 million hectares per year in the last decade according to FAO [2010]) and produces such immediate consequences as biodiversity loss, loss of hydrological capacity, and increased net emissions of greenhouse gases [Fearnside 2005]. Below, we present illustrative examples of deforestation events detected by our algorithm in California, Brazil, Borneo and Siberia.

The algorithms found several events in northern California corresponding to logging activities. Figure 15(a) shows the EVI time series for one of the locations in an area



Source: Google Earth imagery.

Fig. 13. Events detected by yearly delta algorithm corresponding to Zaca Fire in Santa Barbara County in California. Also shown is a typical FPAR time series.

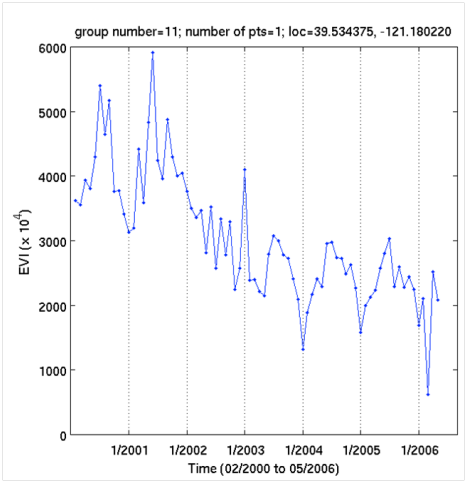


(a) Validation polygon showing the boundary of a forest fire in Greece.

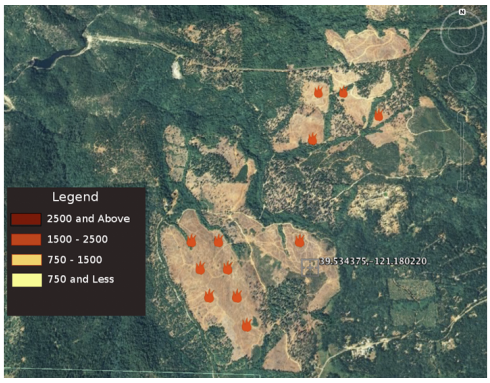
(b) Events overlaid with forest fire polygons.

Source: Google Earth imagery.

Fig. 14. Evaluation of algorithms in Greece.



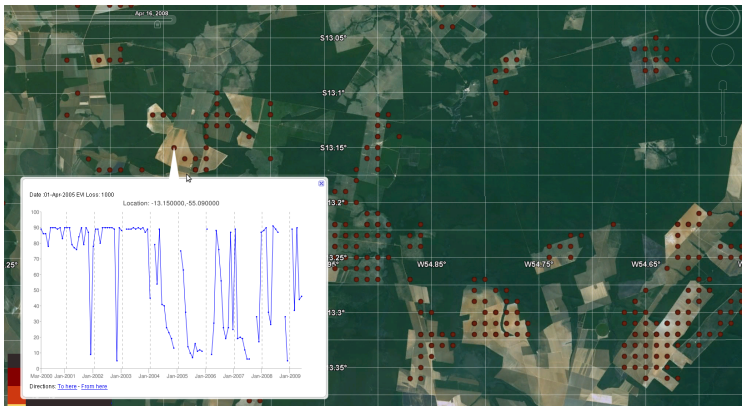
(a) Plot shows the EVI time series for a location that was logged in Northern California.



(b) Logging events in Northern California corresponding to barren land that are visibly cleared. Red dots correspond to change points detected.

Source: Google Earth imagery.

Fig. 15. Logging in northern California



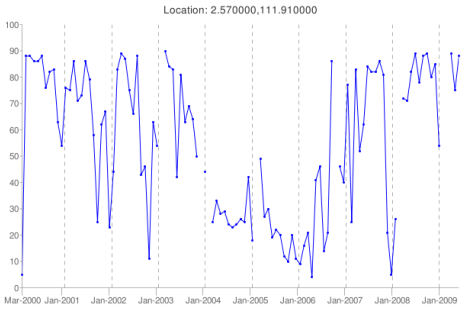
Source: Google Earth imagery.

Fig. 16. Deforestation events detected by our algorithm in Brazil. Plot shows the FPAR time series corresponding to a deforested location. Red dots correspond to change points detected.

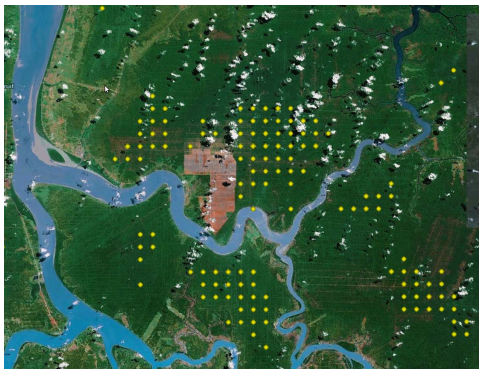
that was logged. Figure 15(b) shows the events we detect corresponding to a patch of barren land that is located in the middle of a forested area.

We also detect numerous locations in Brazil's Amazon basin that were deforested. Most of this activity was found in the Mato Grosso region which is also called the "arc of deforestation." Figure 16 shows the overlay of the locations in Brazil identified as deforested by our algorithm. It can be seen that the disturbance events marked with red dots occur where the imagery shows patches of cleared forests. The FPAR time series shows standing forest until year 2004 after which it is converted to pasture or cropland.

Large forested areas in Borneo have been cleared for palm oil plantations. We detect many of these conversions. Most of the changes detected are near existing plantations,



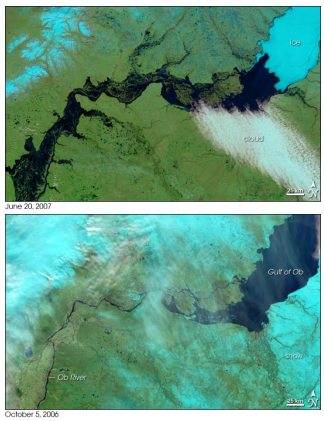
(a) Plot shows the FPAR time series for a location that was converted from forest to palm oil plantation in South East Asia.



(b) Events detected by our algorithm corresponding to forest conversion to plantation in South East Asia. Note these expansions occur in areas surrounding existing plantation. Yellow dots correspond to change points detected.

Fig. 17. Palm oil plantations in South East Asia

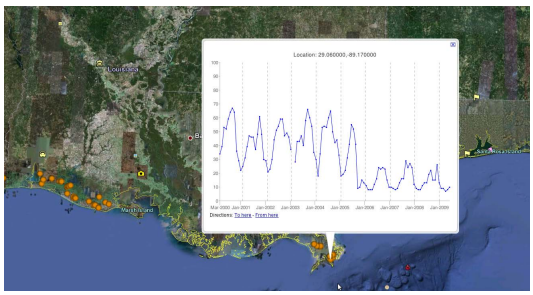
indicating the expansion of plantation area. Figure 17(a) shows the FPAR time series of a pixel of land that was cleared for plantation use in year 2003. Figure 17(b) shows the spatial proximity of the recent change events (which are related to plantation expansions) to the existing plantations. Thus existing maps can be used for characterization of the forest change events detected. Additionally, it indicates the utility of our algorithms in automatically updating the existing maps.



Source: NASA.

Fig. 18. Images of Ob river and the surrounding area. The image on the top is from a period of flooding and the image at the bottom is the same area when it is not flooded.

3.4.3. Natural Disasters. In addition to fires and deforestation, the change detection algorithms also detect events corresponding to natural disasters, such as floods, droughts, earthquakes, and hurricanes, which cause widespread damage to vegetation. We illustrate some of the events detected by our algorithm. The Ob river is one of the largest rivers in Asia and drains into the Kara Sea. In many years, when water from melting snow from the southern latitudes fails to drain into the frozen sea in the



Source: Google Earth imagery.

Fig. 19. Events detected by our algorithms corresponding to the Katrina Hurricane of 2005.

north, large-scale seasonal flooding occurs (shown in Figure 18) causing disturbance of vegetation surrounding the river [Papa et al. 2007].

In August and September of 2005, Hurricanes Katrina and Rita hit the Louisiana coast, flooding 217 square miles of Louisiana's coastal land. The algorithms detected a large number of points corresponding to the areas affected by the Katrina. Figure 19 shows the change events with a typical time series which shows a drop of vegetation in August 2005.

4. EVENT IDENTIFICATION AND CHARACTERIZATION

Land cover changes often span several pixels in a region. Change detection algorithms (most image-based and nearly all time series-based approaches) typically detect disturbance pixels independent of one another, either because of algorithmic design or data constraints.

However, there is a need to group together pixels that correspond to the same change "event," which refers to the physical process (e.g. fire, logging, etc.) that led to the land cover being changed. For example, the time series for several pixels corresponding to a well-known fire are shown in Figure 21. Each time series was found independently by a change detection algorithm, but ideally one should be able to see a unified view of all the time series as is shown in the figure. We refer to the process of summarizing sets of change pixels on the basis of proximity as the spatio-temporal event identification problem. Boriah [2010] provides a discussion of various techniques for this problem.

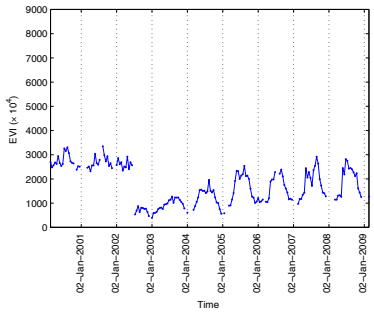
The grouping of a collection of time series as an event can help in the characterization of a detected change (i.e., deducing the fate of the land cover). Characterization is critical for evaluating the impact of land cover change [Foley et al. 2005], and thus is a challenge that must be addressed in its own right. There are two things that need to be characterized (i) the event that occurred, i.e. whether it was a fire, flood, logging, etc. and (ii) the change in vegetation type. For example, after a disturbance forests can be replaced by farmland or plantation or left barren. Deforestation itself can be by mechanical logging and/or slash and burn.

One possible approach for characterizing the clusters of disturbance events is to generate statistics, such as variability, average vegetation loss, recovery time, time interval of change, spatial extent, and prior land cover type, and use these to classify these clusters into categories of interest. Combinations of these cluster properties can be useful for characterization process. As an illustrative example, a forest fire occurs in a small time window and often burns all the neighboring areas and so all the points in the cluster are geographically close and change at the same time. On the other hand, in most cases of logging, the cluster sizes are typically smaller, logging starts at different times for different points in the cluster and the pixel is not logged in a single time

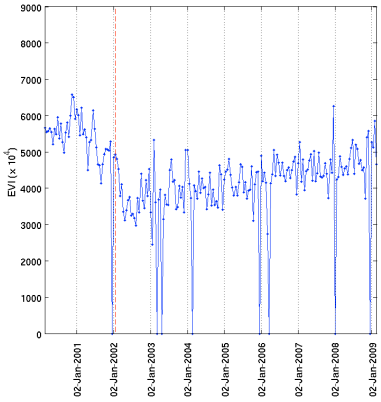
step and so the vegetation response continues to decrease for several weeks. Another distinction is that vegetation response tends to recover much faster for fires than for logging.

Note that changes of interest happen both in space and time which makes change detection challenging. However, these two sources of information naturally complement each other, and can result in a powerful paradigm for characterizing signatures. For example, Figure 20 shows sets of time series corresponding to a forest fire, logging and a drought event, respectively. It can be seen that each of these events has a characteristic signature in the time series. For example, forest fires cause sudden and synchronized declines in vegetation index (e.g., EVI) in several neighboring locations, whereas logging tends to results in a gradual decline in an isolated fashion. Drought and insect damage also cause a gradual decline in vegetation but this happens in a synchronized fashion over relatively large regions.

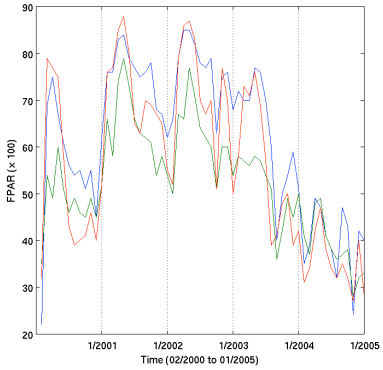
Spatial grouping can be used to improve detection accuracy by detecting marginally changed pixels that may not otherwise be detected. For example, depending on the terrain, pixels in the middle or on the boundary of a fire may be only marginally disturbed. In other cases, such as logging, the activity may be slowly spreading across a region. Therefore, marginally changed time series may indicate ongoing, small-scale changes that are not large enough to exceed thresholds. Another use of spatial grouping is when there is incomplete remote sensing data. If there are a few pixels for which changes can be reliably detected, the remaining disturbed pixels in the region can be detected by examining the spatial neighborhood, potentially overcoming the severe missing and noisy data problem. For example, while most change detection algorithms assign a high score to pixels that are easily identified as changed, other points that have also changed may received a lower rank (or no rank due to missing data). In these instances an event identification algorithm can group together pixels from the same event thereby improving the results of change detection and providing a more complete view of the underlying event characterization.



(a) EVI time series for a forest fire.



(b) EVI time series for a logged location.



(c) FPAR time series for a drought location.

Fig. 20. Time series corresponding to a forest fire, logging and drought event.

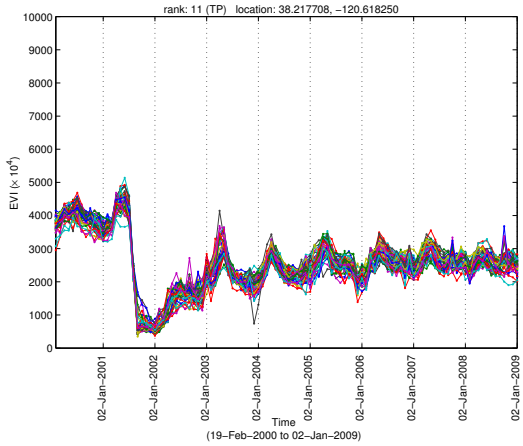


Fig. 21. This collection of time series shows a dramatic drop in vegetation index (EVI) around the summer of 2001.

5. RELATIONSHIP MINING

Understanding the relationship between ecosystem and climate events is of significant importance to climate and earth scientists. For example, Figure 22 shows large scale vegetation disturbances corresponding to a flood of the Ob river. The FPAR time series shows a decrease in vegetation response for the locations in year 2002 when the flood occurred. The precipitation record for the area shows unusually high precipitation during the same year, which explains the severity of the flood and resulting vegetation disturbance. As another example, the anomalous warming of the eastern tropical region of the Pacific (referred to as El Niño phenomenon) has been linked to droughts in Indonesia, which in turn increases fire risk for Indonesian forests. Figure 23 shows a positive relationship between the forest fire events detected by our algorithm and the El Niño climate index.

Automated discovery of such relationships is the goal of relationship mining. Currently, such relationships are typically discovered by the targeted investigations of highly trained scientists of the phenomenon of interest. Although such manual approaches may produce quite noteworthy results, they may miss important relationships, and thus a data driven approach for finding these spatio-temporal and nonlinear relationships can offer dramatic advances in climate science. The automatic discovery of such relationships pose several computational challenges. These relationships are often localized to specific regions and time and may not be visible if an entire data set is analyzed and indeed, may even be reversed (statisticians know this as Simpson's paradox). As an example, it has been observed that El Niño effect on temperate latitudes is most visible during winter. Domain knowledge is used by earth scientists to subset data and otherwise guide the analysis, but intelligent computational techniques should also be capable of mining these patterns in an automated fashion.

One of the promising approaches for discovering relationships between sets of variables is the data mining technique known as association analysis [Tan et al. 2006] which seeks to find patterns that describe the relationships among variables used to characterize a set of objects (transactions). The resulting patterns are either sets of related variables (called itemsets) or rules that relate the occurrence of values in one variable to the occurrence of values of other variables (association rules) for sufficiently large subsets of the data objects (transactions). In the case of spatio-temporal data, this means that the pattern holds for some variables for locations and some time periods.

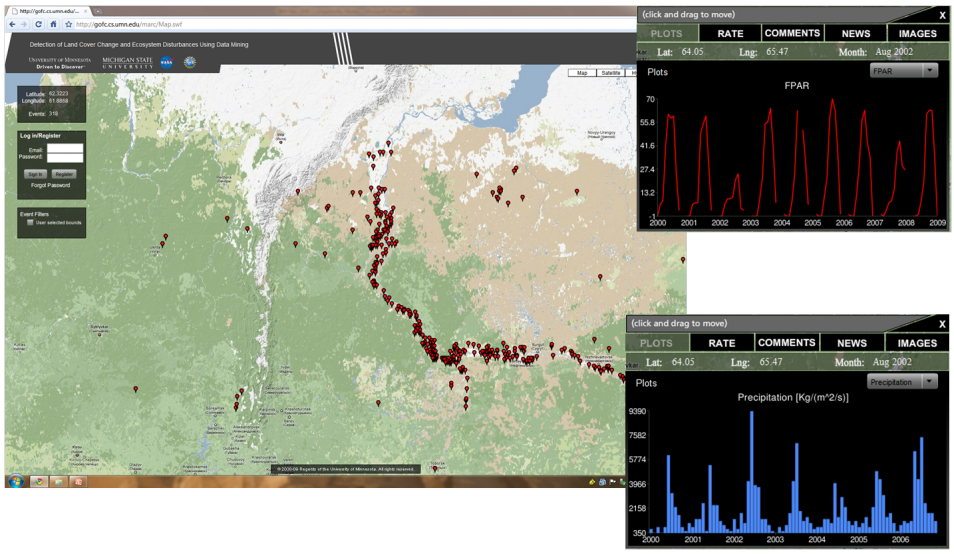


Fig. 22. The figure shows large scale disturbances detected by our algorithm around the Ob river in Russia. The plot in the top right shows the FPAR signal while the histogram in the bottom right shows the precipitation recorded. It can be seen that there is a higher precipitation in the year 2002, which may be a possible cause of the unusually large scale of the flood.

Association analysis has two strengths: (1) the existence of approaches for pruning the pattern search space that allows the search of such patterns to be found in an exhaustive yet efficient manner and (2) the ability to find nonlinear dependencies.

Although association analysis seems conceptually well-suited to finding the right place and time, the use of association analysis for spatio-temporal data poses several challenges since it was originally developed for non-spatial data that consisted of objects described by categorical or binary variables. For example, current techniques for association analysis need to discover association among climate and ecosystem variables that may be separated in space and/or time, e.g., relationship between anomalously high sea surface temperature in the Pacific during winter and the frequency of fires in Indonesian rain forests during summer (Figure 23). Additional challenges include the need to adapt association analysis to continuous instead of binary data, reducing the large number of redundant patterns that can be discovered, and the need to assess the significance of the patterns discovered.

6. CARBON RISK SCORING

For earth scientists and policy makers interested in management of carbon stocks and emissions, the notion of carbon at risk due to forest disturbances and changes is critical. Carbon risk can be modeled in various ways by incorporating a number of factors. The simplest approach to define risk is based on the past history of a location. For example, locations that have had many fires in the past are more likely to have fires in the future. However, such information may not yield good results when the history is short or only a few events have occurred in the past. A more sophisticated approach includes information (past and present) from other group of locations. One way to group locations is based on spatial neighborhood (proximity). To illustrate, if a neighboring location was logged recently, then this should typically increase the risk to the current location.

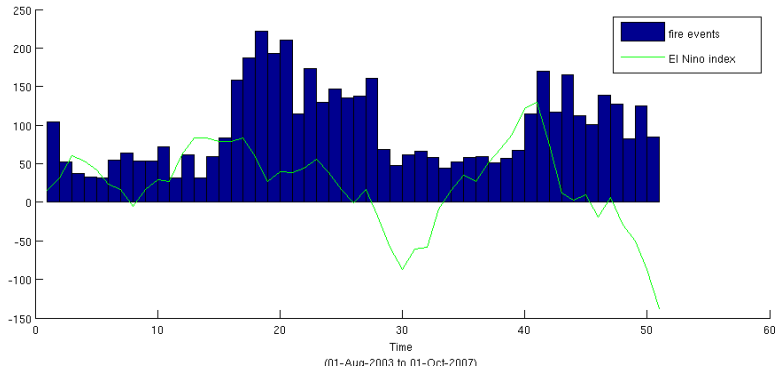


Fig. 23. This figure shows the relationship between the El Niño index (scaled) and the forest disturbances detected in Indonesia. A higher value of the El Niño index is associated with drought conditions that increase combustibility of of forests.

Other possibilities to obtain better risk assessments includes grouping locations according to similar properties, such as vegetation type, location, climate, historical patterns, and range of human activities in the neighborhood. For example, particular locations due to the climate and soil moisture conditions are more prone to fire while others are resistant to it. Additionally, there are various external factors, such as the presence of nearby population centers, rugged terrain, and the age of the vegetation, that can have a significant impact on the risk of the current location. For instance, closeness to a major population center may increase the risk of fire, the availability of good roads increases the risk of logging (legal or illegal), and changes in the climate can cause drought or insect infestations. If the impact (i.e., relationship) of these external factors can be determined, then even better risk assessments can be achieved. However, incorporating these factors into a risk model can be quite challenging due to the complexity and variety of data sets involved [Soler and Verburg 2010].

Although carbon risk-scoring is inherently complex, even simple approaches can provide useful information. In the following, we describe the results of a forest fire risk modeling technique that makes use of information about fire in the spatial neighborhood to estimate risk of fire. We applied our algorithm on 4 km FPAR data to detect fires in Canadian forests for 2001–2009. A total of 15,000 4 km pixels were identified as burned in Canada, of which 11274 were burned in year 2001–2007 and 3702 were in the year 2008–2009. We choose 800 out of these 3702 and compute the distance from each of these 4 km pixels burned in 2008–2009 to the nearest pixel burned in previous years (2001–2007). To test the hypothesis that forests close to the previously burned areas are at a greater risk than other locations, we performed the same computation for 800 randomly located points in the entire Canadian forest. (This covers about 300,000 4 km pixels.) For each case, we computed the cumulative number of data points within a fixed distance from a historic fire. This was repeated 50 times and the median values were computed. Figure 24 shows these results. The blue curve corresponds to the actual fires in 2008–2009, while the red curve corresponds to the randomly selected points. It is clear from the curves that the fire locations in 2008–2009 are generally much closer to the previously burned locations compared with the randomly selected locations. In particular, the plot shows that almost all the fires in 2008–2009 were within 20 pixels (80 km) distance from a previous fire whereas a much smaller fraction of the randomly selected points fall within that distance of previously burned areas.

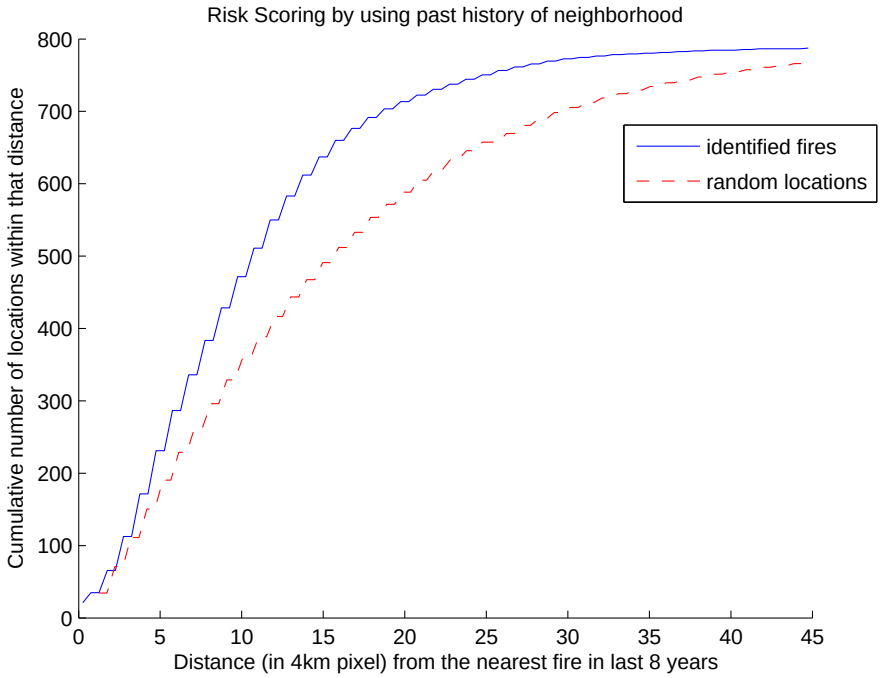


Fig. 24. This figure shows the plot for cumulative number of events versus pixel (location) distance to the nearest historic fire. The blue curve corresponds to actual fires that occurred in the last year, while the red corresponds to an equal number of randomly picked pixels.

7. CONCLUSION

In this paper, we have discussed key issues, approaches and results for the problems of global forest monitoring, relationship mining and risk scoring. Illustrative results for land cover change detection have highlighted the benefits of land cover change detection based on time series approaches: scalability, identification of the time at which the change occurred, quantitative information about the magnitude of the change, and information that can be used to characterize the nature of the change. Verification of the results has shown good agreement with validation data where available. Work on characterizing the nature of the changes and providing carbon risk scoring is underway, although at an earlier stage than the change detection efforts.

Our immediate goals are to produce a global history of changes in forest cover based on 1 km EVI data from the last 10 years, as well as a near real time detection capability. These results will be available in the near future via a cloud-based analytics and visualization platform called ALERTS (Automated Land change Evaluation, Reporting and Tracking System) provided through our research and development collaboration with the Planetary Skin Institute (`www.planetaryskin.org`). Figure 25 shows a preliminary version of such a platform. The set of changes in forest cover will provide a rich source of data for those concerned about health of forest ecosystems. More generally, the success of these efforts will be an exciting realization of the potential of computer science to producing and distributing information that can contribute in a meaningful way to the more sustainable management of forest and other resources.

Like many other topics in computer science, change detection, characterization, and risk scoring pose enduring challenges. Although we have made progress, many op-

Fig. 25. This figure shows the Planetary Skin Institute's ALERTS cloud-based visualization platform.

portunities for enhancements remain for taking into account spatio-temporal context, handling missing and noisy values, and further exploring the space of the different types of changes that can occur. In addition, many categories of changes in land cover or land use cannot be easily detected by using a univariate time series such as those derived from the EVI and FPAR vegetation products. For example, conversion of tropical forests to savannas can be hard to detect in the Amazon forest, as the FPAR value is quite high for both types of vegetation. Such changes can be potentially captured by using multiple variables that represent different parts of the spectrum. Hence there is a great scope for research on change detection algorithms for multivariate time series [Cheng et al. 2009].

Additionally, these efforts also require interdisciplinary collaborations with resource and risk domain experts, and in particular with Earth scientists. Specific projects that focus on particular areas of interest to Earth scientists (e.g., tropical rain forests) will help in refining techniques and make the results more usable. In addition, such focused efforts often make use of additional data (e.g. higher resolution data such as aerial images) that can complement the satellite data on which our results are currently based.

Although the focus of our paper is on forests, there are many challenges involved in the sustainable management of other resources, e.g., water, food, and energy. The techniques or approaches for change detection, characterization, and risk scoring discussed in the context of forests can provide the foundation for detecting changes, characterizing them, and assessing risks in these other areas. However, additional domain specific work will be necessary to handle the unique challenges in each of these areas. We encourage our colleagues in computer science to join us in these efforts to tackle these problems that have a direct impact on society at large.

ACKNOWLEDGMENTS

The research described in this paper was supported by NSF Grant IIS-0713227, NSF Grant IIS-0905581, NASA Grant NNX09AL60G, the University of Minnesota MN Futures Program, and by the Planetary Skin Institute whose objective is to develop contextually sensitive resource and risk management decision support

capabilities of local, regional and global application. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute. We would also like to thank Durga Toshniwal, Ivan Brugere, Divya Alla, Vikrant Krishna, Matt Kappel and Yashu Chamber for their helpful comments.

REFERENCES

BANERJEE, S., GELFAND, A., FINLEY, A., AND SANG, H. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B 70,* 4, 825.

BONAN, G. B. 2008. Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests. *Science 320,* 5882, 1444–1449.

BORIAH, S. 2010. Time series change detection: Algorithms for land cover change. Ph.D. thesis, University of Minnesota.

BORIAH, S., KUMAR, V., STEINBACH, M., POTTER, C., AND KLOOSTER, S. 2008. Land cover change detection: A case study. In *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 857–865.

BORIAH, S., MITHAL, V., GARG, A., KUMAR, V., STEINBACH, M., POTTER, C., AND KLOOSTER, S. A. 2010. A comparative study of algorithms for land cover change. In *Proceedings of the 2010 Conference on Intelligent Data Understanding*. NASA Ames Research Center, 175–188.

BOX, G. E. P., JENKINS, G. M., AND REINSEL, G. C. 2008. *Time Series Analysis: Forecasting and Control*. Wiley.

CHENG, H., TAN, P.-N., POTTER, C., AND KLOOSTER, S. A. 2009. Detection and characterization of anomalies in multivariate time series. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009*. 413–424.

COPPIN, P., JONCKHEERE, I., NACKAERTS, K., MUYS, B., AND LAMBIN, E. 2004. Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing 25,* 9, 1565–1596.

FAO. 2010. Global forest resources assessment (FRA 2010). Food and Agriculture Organization of the United Nations.

FEARNSIDE, P. 2005. Deforestation in Brazilian Amazonia: history, rates, and consequences. *Conservation Biology 19,* 3, 680–688.

FEARNSIDE, P. 2008. Amazon forest maintenance as a source of environmental services. *Anais da Academia Brasileira de Ciências 80*, 101–114.

FOLEY, J. A., DEFRIES, R., ASNER, G. P., BARFORD, C., BONAN, G., CARPENTER, S. R., CHAPIN, F. S., COE, M. T., DAILY, G. C., GIBBS, H. K., HELKOWSKI, J. H., HOLLOWAY, T., HOWARD, E. A., KUCHARIK, C. J., MONFREDA, C., PATZ, J. A., PRENTICE, I. C., RAMANKUTTY, N., AND SNYDER, P. K. 2005. Global Consequences of Land Use. *Science 309,* 5734, 570–574.

GULLISON, R. E., FRUMHOFF, P. C., CANADELL, J. G., FIELD, C. B., NEPSTAD, D. C., HAYHOE, K., AVISSAR, R., CURRAN, L. M., FRIEDLINGSTEIN, P., JONES, C. D., AND NOBRE, C. 2007. Tropical forests and climate policy. *Science 316,* 5827, 985–986.

GUSTAFSSON, F. 2000. *Adaptive Filtering and Change Detection*. John Wiley & Sons.

INCLÁN, C. AND TIAO, G. C. 1994. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association 89,* 427, 913–923.

KEOGH, E., CHU, S., HART, D., AND PAZZANI, M. 2001. An online algorithm for segmenting time series. In *ICDM 2001: Proceedings of the first IEEE International Conference on Data Mining*. 289–296.

KEOGH, E., LONARDI, S., AND CHIU, B. 2002. Finding surprising patterns in a time series database in linear time and space. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 550–556.

KURZ, W. AND APPS, M. 2006. Developing Canada's National Forest Carbon Monitoring, Accounting and Reporting System to Meet the Reporting Requirements of the Kyoto Protocol. *Mitigation and Adaptation Strategies for Global Change 11*, 33–43.

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

LU, D., MAUSEL, P., BRONDÍZIO, E., AND MORAN, E. 2003. Change detection techniques. *International Journal of Remote Sensing 25,* 12, 2365–2401.

LUNETTA, R. S., KNIGHT, J. F., EDIRIWICKREMA, J., LYON, J. G., AND WORTHY, L. D. 2006. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sensing of Environment 105,* 2, 142–154.

OLLINGER, S. V., RICHARDSON, A. D., MARTIN, M. E., HOLLINGER, D. Y., FROLKING, S. E., REICH, P. B., PLOURDE, L. C., KATUL, G. G., MUNGER, J. W., OREN, R., SMITH, M.-L., PAW U, K. T., BOLSTAD,

P. V., COOK, B. D., DAY, M. C., MARTIN, T. A., MONSON, R. K., AND SCHMID, H. P. 2008. Canopy nitrogen, carbon assimilation, and albedo in temperate and boreal forests: Functional relations and potential climate feedbacks. *Proceedings of the National Academy of Sciences 105,* 49, 19336–19341.

PAPA, F., PRIGENT, C., AND ROSSOW, W. B. 2007. Ob' river flood inundations from satellite observations: A relationship with winter snow parameters and river runoff. *Journal of Geophysical Research 112,* D18, 103.

POTTER, C., TAN, P., STEINBACH, M., KLOOSTER, S., KUMAR, V., MYNENI, R., AND GENOVESE, V. 2003. Major disturbance events in terrestrial ecosystems detected using global satellite data sets. *Global Change Biology 9,* 7, 1005–1021.

RAMANKUTTY, N., GIBBS, H. K., ACHARD, F., DEFRIES, R., FOLEY, J. A., AND HOUGHTON, R. A. January 2007. Challenges to estimating carbon emissions from tropical deforestation. *Global Change Biology 13*, 51–66.

ROY, D. P., LEWIS, P. E., AND JUSTICE, C. O. 2002. Burned area mapping using multi-temporal moderate spatial resolution data–a bi-directional reflectance model-based expectation approach. *Remote Sensing of Environment 83,* 1-2, 263–286.

SOLER, L. AND VERBURG, P. 2010. Combining remote sensing and household level data for regional scale analysis of land cover change in the brazilian amazon. *Regional Environment Change*, 1–16.

TAN, P.-N., STEINBACH, M., AND KUMAR, V. 2006. *Introduction to Data Mining*. Addison-Wesley, Boston, MA.

TCG. 2010. Terrestrial Carbon Group Policy Briefs.
`http://www.terrestrialcarbon.org/Publications/PolicyBriefs.aspx`.

WOTTON, B., NOCK, C., AND FLANNIGAN, M. 2010. Forest fire occurrence and climate change in Canada. *International Journal of Wildland Fire 19*, 253–271.