Data Mining for Visual Exploration and Detection of Ecosystem Disturbances

Haibin Cheng, Pang-Ning Tan
Michigan State University
East Lansing,MI,48823
{chenghai,ptan}@msu.edu

Christopher Potter
NASA Ames Research Center
Moffett Field,CA,94035
cpotter@mail.arc.nasa.gov

Steven Klooster
California State University
Monterey Bay,CA,93955
klooster@gaia.arc.nasa.gov

ABSTRACT

Remote sensing data from Earth observation satellites offer unprecedented opportunity for predicting and understanding the behavior of the Earth's ecosystem. However, because of their massive volume, extracting interesting patterns such as ecosystem disturbances from the data is a challenging task. In this paper, we present a case study on the application of data mining to the disturbance event detection problem. We describe two approaches—moving average and random walk—for detecting ecosystem disturbances. We then illustrate how clustering can be used to identify locations with similar incidents of ecosystem disturbance events. Finally, we develop a clustering-based framework to aid the visual exploration and detection of ecosystem disturbances from high resolution vegetation cover data.

Keywords

Clustering, multi-level indexing, spatio-temporal data

1. INTRODUCTION

Ecosystem disturbances, such as wildfires, droughts, herbivorous insect outbreaks, and forest logging, are events that result in a sustained disruption of the ecosystem structure and function. Such events may alter the ecosystem productivity and resource (light and nutrient) availability for organisms on large spatial and temporal scales. The release of carbon dioxide (CO_2) from terrestrial biomass loss during large disturbance events may also contribute to the current rise of CO_2 levels in the atmosphere [21, 24].

Due to their significance and potential implications to climate change, Earth scientists are interested in detecting ecosystem disturbance events at a global scale from historical eco-climatic data. We had previously developed a method [23, 22] to extract such events using 19-year satellite observations of vegetation cover from the Advanced High Resolution Radiometer (AVHRR) array of sensors. The method enables Earth scientists in our team to detect large-scale wildfires exceeding 0.5 Mha of affected land areas throughout the world. Our study also showed that nearly 9 Pg of carbon could have been lost from the terrestrial biosphere to the atmosphere as a result of large-scale ecosystem disturbance over this 19-year period [4].

Despite its successes, our previous method can only detect sustained disturbance events that alter the structure of vegetation cover significantly for a sufficiently large region and result in an overall decline in the average annual vegetation levels. Smaller scale disturbances could have been missed when using the coarse resolution AVHRR data. With the recent availability of high resolution vegetation cover data from the moderate resolution imaging spectroradiometer (MODIS) instrument on board of the NASA's TERRA satellite, smaller scale disturbances can be potentially uncovered. However, because the size of the data has grown by 4 orders of magnitude, applying the disturbance event detection algorithm to such massive data sets is computationally expensive. The problem is further exacerbated by the fact that a disturbance event detection algorithm (or anomaly detection algorithm, in general) requires specification of one or more thresholds to determine whether a detected event should be flagged as a real disturbance. The thresholds are determined by users through a trial and error process during the exploratory data analysis phase. Because of the large amount of data that must be processed, performing exploratory data analysis on the high resolution MODIS data in a real-time fashion is computationally infeasible. The massive size of the data also produces more events for scientists to validate. This necessitates the development of innovative data mining approaches that can assist Earth scientists in real-time exploration of global scale eco-climatic data.

This paper presents a case study on the application of data mining to the disturbance event detection problem. We first describe two approaches—moving average and random walk—for detecting ecosystem disturbances. We then illustrate the use of clustering to group together regions with similar incidents of ecosystem disturbances. This approach provides an unsupervised way to categorize the events and helps reduce the number of events that need to be validated. We also present a multi-level indexing scheme based on clustering to aid the discovery and visual exploration of ecosystem disturbances. We show how the indexing scheme can be used to enable users to quickly focus on regions of interest during the exploratory data analysis phase. While clustering-based methods have been proposed for spatio-temporal indexing and for data reduction purposes [7, 6], none of them are specifically designed for detecting anomalies (such as ecosystem disturbances) in spatio-temporal data. In this work, we evaluate the effectiveness of using clustering-based methods for exploratory analysis of ecosystem disturbances. Our disturbance event detection and clustering algorithms have been integrated into an interactive system developed by our team members. The system enables scientists to explore the data in real-time fashion and to visually inspect clusters of locations where similar types of disturbance events were observed.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GIS '08 Irvine, CA, USA
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

The remainder of this chapter is organized as follows. In Section 2, we describe the vegetation cover data used in this study while Section 3 presents our disturbance event detection algorithms. Our methodology for clustering the disturbance events is presented in Section 4. Section 5 describes the design of a multilevel clustering-based indexing scheme for real-time exploration of high resolution vegetation cover data. Finally, Section 8 concludes the paper.

2. DATA

Ecosystem disturbances can be detected by monitoring changes in the vegetation cover data obtained from satellite observations. FPAR, which is the fraction of photosynthetically active radiation absorbed by vegetation canopies, is a common measure of vegetation cover. FPAR values range from zero (on barren land) to 100% (for the densest plant cover). A high FPAR level suggests a region with dense green leaf cover and is presumably less likely to have been disturbed. The FPAR data used in this study are derived from two sources of satellite measurements—AVHRR and MODIS. Monthly FPAR data from AVHRR sensor aboard the National Oceanic and Atmospheric Administration's (NOAA) polar orbiting satellites is available at 8 km spatial resolution from 1982 to 2000. More recently the National Oceanic and Atmospheric Administration's (NASA) Earth Observing System has provided data with higher spatial and temporal resolution, e.g., 1 km $\times$ 1 km. The MODIS data offer an opportunity to detect changes in vegetation cover at a finer spatial resolution. In this study, we use the monthly FPAR data from MODIS covering the time period between 2000 and 2005.

The data that we work with have been preprocessed and validated by NASA/NOAA. The data comes with explicit quality assurance guarantees, having gone through several stages of validation (see for example [2]), with the final stage being of science quality with well-defined uncertainties. Noisy measurements or data quality issues are not an issue in general, as low quality measurements have been tagged with quality assurance flags.

3. DETECTION OF ECOSYSTEM DISTURBANCES

This section presents two approaches for detecting ecosystem disturbances from FPAR data. The first approach is designed based on hypothesis developed by the domain experts in our team. The second approach uses a graph-based random walk approach to detect more general type of anomalies in FPAR time series.

3.1 Moving Average Disturbance Detection

In this approach, the FPAR time series at each pixel was first detrended using a linear adjustment, which is necessary to minimize the possibility that, in cases where there is gradual but marked increase in monthly FPAR over the 19 year time series, any potential disturbance events occurring relatively near the end of the series are not overlooked. To remove the dominant seasonal oscillations in vegetation phenology observed throughout the globe, our detrended FPAR time series was subsequently deseasonalized by computing the 12-month moving average time series for every pixel location.

The domain experts of our team had hypothesized that a "sustained" disturbance event could be defined as any decline in average annual FPAR levels (at an assigned significance level) that lasts for a temporal threshold value of at least 12 consecutive monthly observations at any specific pixel location [23]. The logic used here is that an actual disturbance involves a sustained decline in FPAR because the structure of the vegetation cover has been severely altered or destroyed during the disturbance event, to a magnitude that

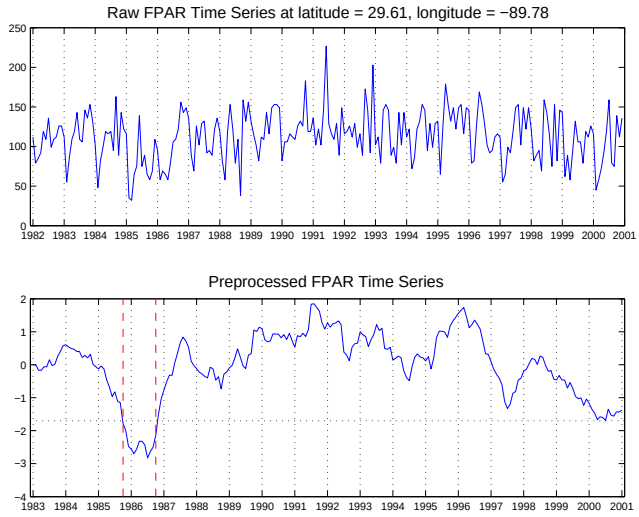


Figure 1: Detection of ecosystem disturbance due to Hurricane Elena.

lowers FPAR significantly for at least one seasonal growing cycle, after which time regrowth and recovery of the former vegetation structure may permit FPAR to increase again. Significant declines in average annual FPAR levels can be defined to be greater than 1.7 standard deviations (SD) below the 19-year average FPAR computed for any specific pixel location. The threshold of 1.7 SD was chosen based on the 95% confidence level in rejecting the null hypothesis using a one-sided statistical t-test.

Figure 1 shows an example of an ecosystem disturbance event recorded at 29.61°N and 89.78°W. The event coincides with the timing of Hurricane Elena (September 1985) while its location is near the vicinity of the landfall points of the storm. In addition to this example, our method successfully detects documented landfall points of other tropical storms including Hurricane Alicia, Gloria, Gilbert, and Hugo. For an extended discussion that includes coverage of droughts, heat waves, cold waves, and blizzards, see [22].

While this approach can successfully detect many large-scale disturbance events, it also has several limitations. First, the timing of the event may vary since it uses 12-month moving average to deseasonalize the time series. Second, it may not be able to detect events in which the vegetation structure of the region recovers in less than one growing seasonal cycle. Figure 2 shows the difficulty of applying this approach to the FPAR time series data located in Chisholm, Alberta where a major wildfire occurs in May 2001.

3.2 Random Walk Based Disturbance Detection

This section presents an alternative approach for detecting ecosystem disturbances based on a graph-based anomaly detection algorithm [17]. It can be used to detect both point-wise anomalies and subsequence anomalies (i.e., anomalies due to unusual segments in a time series). The algorithm initially constructs a kernel matrix $\boldsymbol{K}$, which is a similarity matrix between every pair of points in the time series. The time series and its corresponding kernel matrix can be transformed into a weighted graph representation $\mathcal{G}(V, E)$ where each time point is a node $v \in V$ in the graph and the similarity between two points is represented by a weighted edge $e \in E$ between the corresponding pair of nodes.

We formulate the disturbance event detection problem as a random walk on the weighted graph $\mathcal{G}$. Anomalies (or disturbance

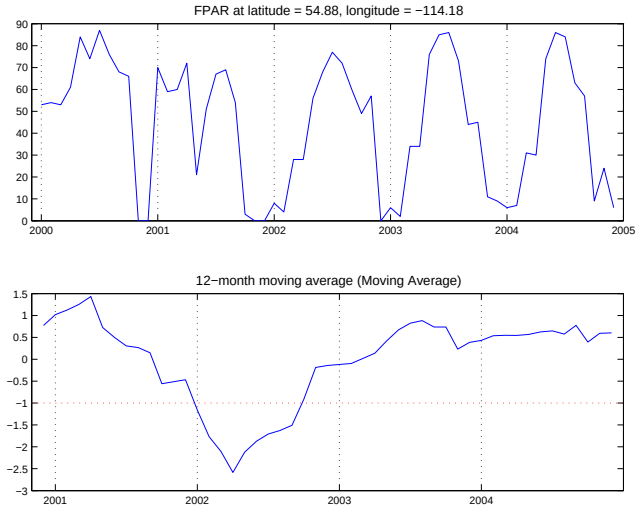


Figure 2: Detection of ecosystem disturbance due to a major wildfire in Chisholm, Alberta using 12-month moving average.

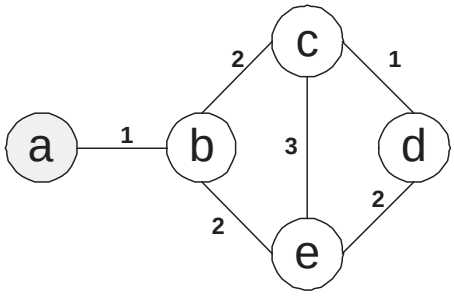


Figure 3: A weighted graph with an anomalous node a.

events) are detected based on nodes that have the lowest connectivity. For example, the node labeled as a in Figure 3 has significantly lower connectivity than other nodes in the weighted graph. This approach can be easily extended to subsequences by mapping each segment of the time series as a node in the weighted graph and using an appropriate similarity measure between segments to compute its kernel matrix.

To estimate the connectivity value of each node, we model the problem as a Markov chain on the state space V with a transition matrix S computed from the kernel matrix $\boldsymbol{K}$ as follows:

$$S(i,j) = \frac{K(i,j)}{\sum_{j=1}^{n} K(i,j)} \quad (1)$$

Equation 1 has two implications. First, a high similarity value between two points i and j implies a high probability of transition from node i to j in the corresponding graph. Second, $\sum_j S(i,j) = 1$, which is an essential property for a Markov chain. The connectivity value of each node in the graph can be estimated from the transition matrix S iteratively using the following equation:

$$\boldsymbol{c} = d/n + (1-d)S\boldsymbol{c}, \quad (2)$$

where d is known as the damping factor. This iterative procedure can be viewed as a random walk on the Markov chain, where given the current node u, there is a probability $1-d$ of visiting one of its adjacent nodes according to the transition matrix S and a probability d of visiting any random node in the graph. This approach is

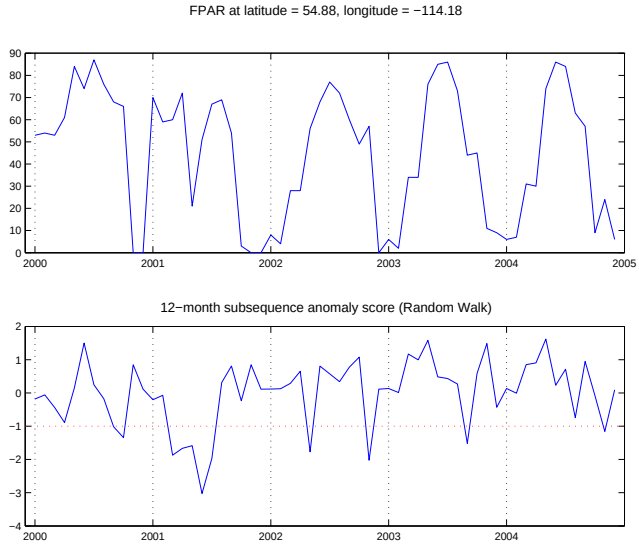


Figure 4: Detection of ecosystem disturbance due to a major wildfire in Chisholm, Alberta using random walk approach.

equivalent to the the formulation used by the PageRank algorithm [20].

Upon convergence, nodes with high connectivity $\boldsymbol{c}$ are considered normal whereas those with low connectivity are declared as anomalous. As an example, Table 1 shows the connectivity values of the nodes given in Figure 3. Node a is successfully detected as an anomaly because it has the lowest connectivity value.

a	b	c	d	e
0.05	0.23	0.72	0.14	0.32

Table 1: Connectivity values obtained by applying random walk on the graph shown in Figure 3

Finally, Figure 11 shows the result of applying our random walk algorithm to detect the Chisholm wildfire event illustrated in Figure 2. Unlike the moving average approach, our algorithm was able to detect the event as a node with the lowest connectivity value. Note that the connectivity values have been standardized by subtracting its mean and dividing by its standard deviation.

3.3 Application of Disturbance Event Detection Algorithm to Large Scale Data

Although the previous approaches can successfully detect many previously documented ecosystem disturbance events in AVHRR data, there are several challenges that must be overcome in order to apply the algorithms to the MODIS data. First, the number of events detected is potentially larger because of the finer spatial resolution of the data. The feasibility of applying the algorithm in a real-time fashion for exploratory data analysis is another issue that must be addressed. Finally, because of its shorter time period (MODIS data is only available since 2000), selecting the appropriate thresholds that maximize the detection rate and minimize the false alarm rate is also a challenge since they are defined based on standardized values of the smoothed times series (for the moving average approach) or the standardized values of their connectivity values (for the random walk approach).

4. CLUSTERING OF ECOSYSTEM DISTURBANCE EVENTS

Clustering is the task of partitioning data into groups of similar objects. In Earth Science studies, clustering has been applied for various applications including land cover classification [26, 27], classification of synoptic-scale circulation patterns [10], and discovery of climate indices [25]. Clustering can also be used to aggregate related (and possibly nearby) data points that have similar characteristics (e.g., climate conditions, landscape variability, etc). Since it does not require any labeled examples, clustering is considered an unsupervised learning task, as opposed to supervised learning tasks such as classification and regression.

In this work, we apply clustering to group together locations that exhibit similar types of disturbance events. The clusters may help users to automatically categorize the different types of disturbance events (e.g., wildfires, insects, deforestation, etc) based on characteristics of the eco-climatic time series during or after the event has occurred. Since there are no labeled examples available, a key challenge is to determine which attributes are appropriate to characterize the different types of disturbance events. Table 2 shows some of the attributes available to describe the characteristics of disturbance events. The user may cluster the data points using any subset of these attributes. Since each attribute can have a different scale, we need to normalize the attribute values before applying the clustering algorithm. Otherwise, the similarity computation between data points will be dominated by attributes that have a large range of possible values. Continuous-valued attribute are normalized by subtracting each attribute with its minimum possible value and dividing it by the range of its possible values. For each discrete attribute, we normalize it by creating a binary attribute for each attribute-value pair. Normalization allows each attribute selected for the clustering task to be treated equally important during similarity computation.

Features	Description
Location	Latitude and longitude of a location
Period	Number of months of consecutively low FPAR values
First month	Month in which disturbance event was detected
FPAR (During)	Minimum and range of FPAR values during period of disturbance event
FPAR (After)	Minimum and range of FPAR values after period of disturbance event
Climate	Precipitation and temperature values at the onset of disturbance event
Land cover	Land cover type at the location

Table 2: Features for clustering disturbance events

4.1 K-means Clustering

While there are many clustering algorithms available, we consider using k-means clustering in this study[1]. K-means is a fast algorithm that scales linearly with the size of the data. K-means is a prototype-based clustering algorithm, in which the clusters are represented by a set of prototypes (specifically, centroids) and data points are assigned to clusters based on their similarity to the cluster centroids.

The k-means algorithm discovers k non-overlapping clusters by identifying k centroids and then assigning each point to the cluster associated with its nearest centroid. Note that a cluster centroid is typically the mean or median of data points in its cluster and

[1]We have also implemented spectral clustering in our disturbance event viewer.

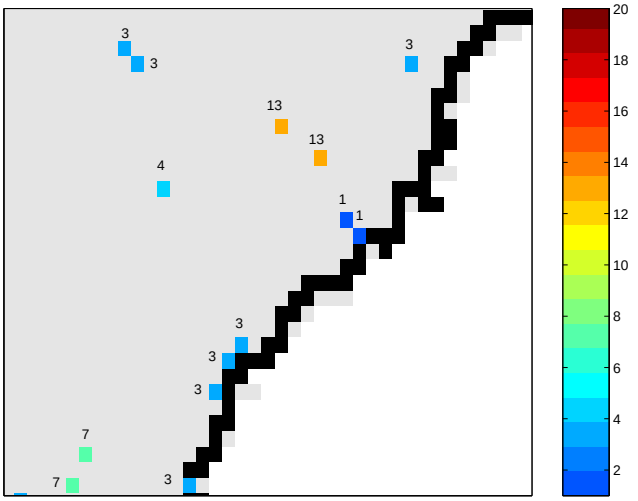


Figure 5: Clusters of ecosystem disturbance events detected in the North Carolina region (each color corresponds to a different cluster).

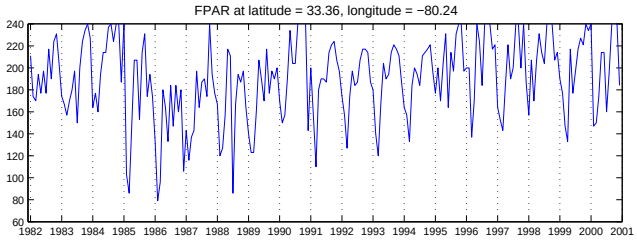
"nearness" is defined by a similarity or distance function. Ideally, the centroids are chosen to minimize the total "error", where the error for each data point is given by a measure of discrepancy between the point and its cluster centroid, e.g., the squared distance. A gradient descent approach for minimizing the squared distance between a data point and its centroid yields the following basic k-means algorithm:

1. Select k points as initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids converge.

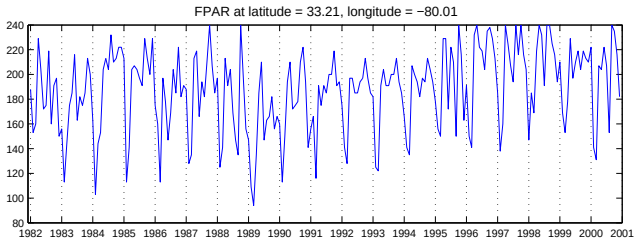
There are a number of variations to the K-means algorithm described above, depending on the method for selecting the initial centroids, the choice of similarity (or distance) measure, and the way the centroid is computed. For this work, we followed the common practice of using the mean as the centroid and selecting the initial centroids randomly.

4.2 Experimental Results

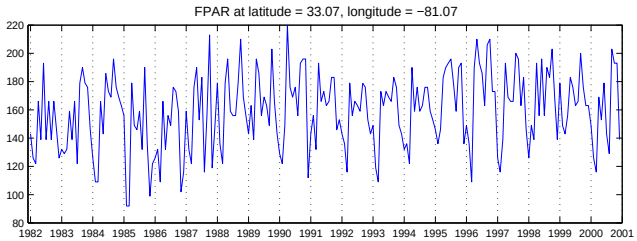
For this experiment, we have performed k-means clustering (with k=20) on all disturbance event locations in North America using precipitation and temperature as data attributes. Figure 5 shows clusters of disturbance events found near the region of North Carolina. Each colored pixel corresponds to a location where disturbance event has been previously detected. We have also labeled each colored pixel according to its cluster ID. Our preliminary results suggest the possibility of distinguishing the different types of ecosystem disturbance events in the region according to their cluster assignment. For example, Figure 6 shows three FPAR time series for one data point assigned to each cluster #1, #13, and #4, respectively. Further investigation revealed that the data points assigned to cluster #1 are associated with disturbance events due to Hurricane Hugo (which occurred in September 1989). Although the disturbance event found for the data point shown in Figure 6(b) also occurs in 1989, it appears earlier in the year, compared to the disturbances found in cluster #1.



(a) FPAR at 32.85°N and 79.84°S (cluster 1)



(b) FPAR at 33.21°N and 80.01°S (cluster 13)



(c) FPAR at 33.07°N and 81.07°S (cluster 4)

Figure 6: FPAR time series at different locations.

5. CLUSTERING FOR EXPLORATORY DATA ANALYSIS

The methods described in the previous section had been successfully applied to the FPAR time series data from AVHRR. With the availability of the higher resolution MODIS data, applying the disturbance event detection algorithm on all the data points can be very expensive. For example, with the 4 km $\times$ 4 km MODIS data, there are more than 1.3 million data points[2] for North America alone. Although the disturbance event detection algorithm scales linearly with the size of the data, it still takes more than a few minutes to process the data for North America. Repeating the analysis using different thresholds will take considerable amount of time, making it infeasible for real-time exploration of the data. The runtime will increase considerably if the analysis is to be performed on the 1 km $\times$ 1 km MODIS data or using more sophisticated (and expensive) anomaly detection algorithms. Techniques must therefore be developed to reduce the amount of processing time.

5.1 Proposed Clustering Framework

In this work, we propose to develop a clustering-based framework to reduce the number of time series that needs to be processed in order to facilitate interactive exploration of the large-scale data. Specifically, our strategy is to build a multilevel index on the MODIS FPAR time series by applying clustering on the time series and choosing representative samples from the clusters at each level of the index hierarchy. Therefore, instead of applying the moving average or random walk disturbance detection algorithm on the entire time series, we would approximate the frequency distribution of disturbance events in each region by applying the algorithm to the selected samples. This helps to reduce the amount of processing time considerably, thus allowing the user to tune the thresholds of their algorithm and to observe the changes in the results in real-time. It will also help the user to focus on a particular region of interest during exploratory data analysis.

Figure 7 shows an example of how the multi-level indexing scheme works. Specifically, we use the multilevel index to monitor the frequency distribution of ecosystem disturbance events in North America[3]. At the top level, the entire continent is partitioned into a 7×10 grid box, where each box contains 256×256 pixel locations. Next, each of the 7×10 boxes is further divided into smaller 4×4 grid boxes, where each of the smaller box now contains 64×64 pixel locations. At the next level, each of these smaller boxes is partitioned into another 4×4 grid boxes, each of which now contains 16×16 pixel locations. A final partitioning of each grid box produces a square region that contains 4×4 pixel locations. A map displayed at this level will show the actual locations where disturbance events have been observed. At all other higher levels, the map displays the frequency distribution of disturbance events in each grid box (region).

For example, in Figure 7, the map at level d displays the frequency distribution of disturbance events for each region in North America. As can be seen from the map, the Baja California peninsula, the midwest region and Canada appears to have highest concentration of disturbance event locations. The user may decide to focus only on these regions and may click on the grid box that corresponds to one of these regions (say California) for further analysis. The map for the California peninsula will now be displayed (level $d-1$). The region is now divided into 4×4 smaller re-

[2]The total number of FPAR time series for the entire world is more than 10.1 million.

[3]We use North America here for illustrative purposes only. We have also created multilevel indices for other continents.

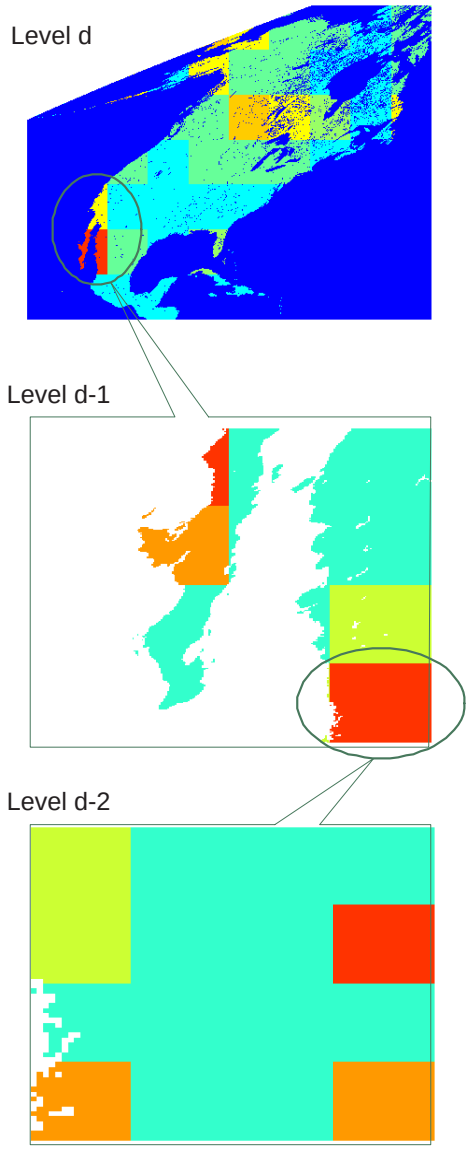


Figure 7: A multilevel approach for visual exploration of ecosystem disturbances.

gions; allowing the user to examine the regions where there is a dense concentration of disturbance events (e.g., the red region in the lower right hand corner of the map). This process is repeated until it reaches the lowest level of the index, where the map displays the actual locations where disturbance events have been detected for that particular region.

5.2 Sampling Representative Time Series from Clusters

The multilevel indexing scheme described above allows the user to explore the data interactively in a hierarchical fashion. Computation time can still be very expensive because the number of time series in each grid box grows quadratically from one level (m) to the next level ($m+1$). For example, there are 256 pixel locations in each grid box at level 1 and 65,536 pixel locations in each grid box at level 3. To reduce the computation time at higher levels of the index, instead of applying the disturbance event detection algorithm to all the time series in the grid box, we apply the algorithm on samples of time series for that region. The key challenge here is to determine which time series should be selected to obtain a representative sample. A good representative sample must provide reasonable coverage of the variability of the FPAR time series in the region. Another issue to consider is the tradeoff between coverage and processing time, which depends on the sample size. If the sample size is large, this will provide a good coverage for the region but at the expense of increasing the processing time. Our experiments on the 4km $\times$ 4km MODIS FPAR data suggest that a sample size containing 100 points per grid box will produce reasonable coverage and processing time.

If N time series must be sampled from a given grid box, we perform clustering on all the time series located in the grid box to obtain k clusters. We then select $\lceil N/k \rceil$ representative time series from each cluster to form the sample for the region. We investigate several sampling approaches in this study. For example, two basic sampling technique is considered:

- **Closest Sampling** would sample time series whose distance is closest to the cluster centroid.
- **Farthest Sampling** would sample time series whose distance is furthest away from the cluster centroid.

Since we are interested in the detection of anomalous events, farthest sampling may be useful to focus on time series that deviate significantly from others in the cluster. However, closest sampling may also be useful to gain a better representative sample of time series in the cluster. To accommodate both types of time series, we propose a third sampling strategy:

- **Mixed Sampling**, which selects a combination of time series that are closest to the centroid and time series that are farthest away from the centroids.

During the construction of the multi-level index, the pixel locations of the sampled time series are stored in the index structure. The samples are retrieved when the grid box is selected by the user during exploratory data analysis. The disturbance event detection algorithm is then applied to the samples and their results will be displayed on the map (as shown in Figure 7.

To evaluate the effect of a sampling strategy, we can compute the sampling effect as follows:

$$\text{Effect} = \frac{1}{d}\sum_{l=1}^{d}\sum_{b_i \in \Omega_l} \frac{N_a(b_i;\theta) - f(b_i;\theta) \times N(b_i;\theta)}{N_a(b_i;\theta)} \quad (3)$$

where d is the number of levels, Ω_l is the set of all grid boxes at level l, θ is the event definition threshold, $N_a(b_i;\theta)$ is the actual number of disturbance events in a grid box, $N(b_i;\theta)$ is the total number of pixel locations in the grid box, and $f(b_i)$ is the fraction of sampled time series in the grid box that produces a disturbance event. Intuitively, Equation 3 measures the difference between the actual number of disturbances (if the detection algorithm has been applied to all the time series) and the number of estimated number of disturbance events for the given region.

5.3 A User-Interactive System for Disturbance Event Detection

Figure 8 shows a snapshot of the interactive system that integrates the disturbance event detection and clustering algorithms developed in this study. The algorithms and user interface are developed using the Matlab programming language. Initially, the map will display the frequency distribution of disturbance events for the

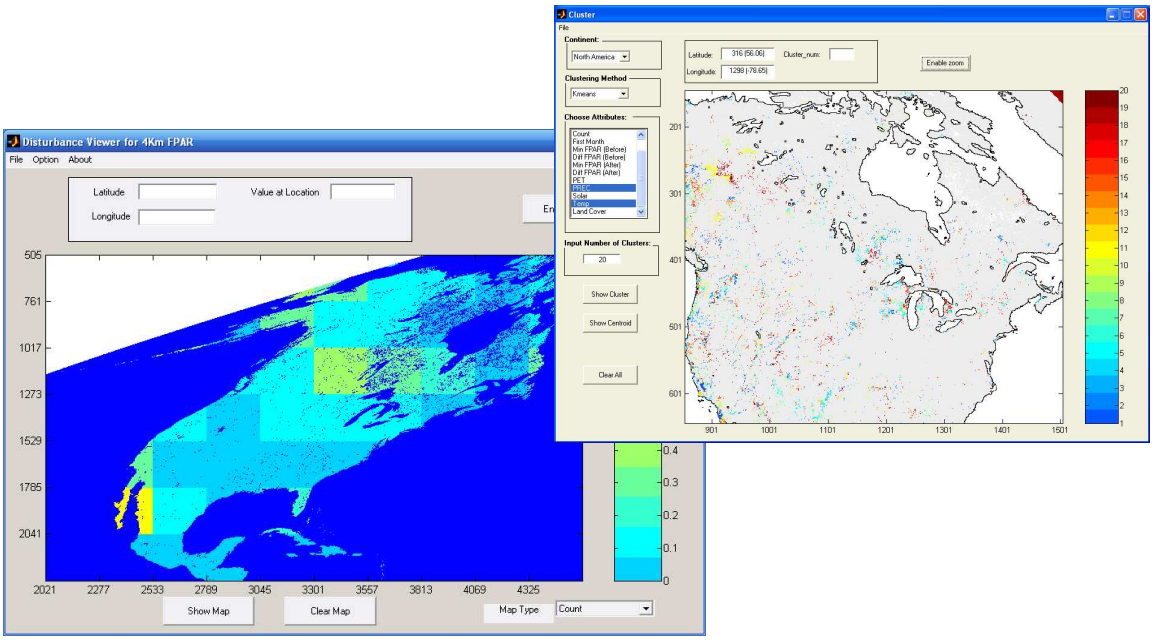


Figure 8: An interactive visual data mining tool for detection and characterization of ecosystem disturbances.

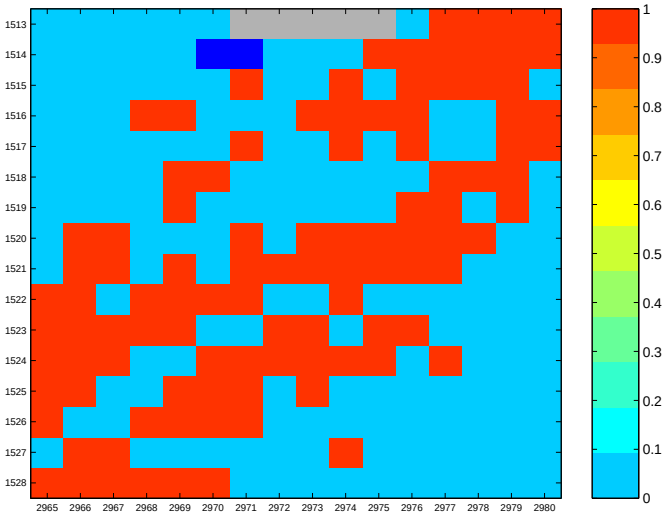


Figure 9: Disturbance locations distributed in the regions of Hayman Fire Event at the highest resolution level.

selected continent (e.g., North America, as shown in Figure 8). The continent is initially divided into a 7×10 grid box. The color of the pixels on the map represents the fraction of locations in each grid box that contain disturbance events. The user may decide to select a grid box that contains a high percentage of disturbance events for further exploration. The viewer will zoom in to the selected grid box and scanned all the subgrids for that region in the next higher resolution level. Also, our viewer enable users to return to the last level by right click the map and choose a different grid at lower resolution level. User can repeat this process until the highest level is reached where each grid represents a true location with latitude and longitude.At the highest resolution level, user can click a single location and the viewer will plot the time series and return anomaly score for each temporal data point.

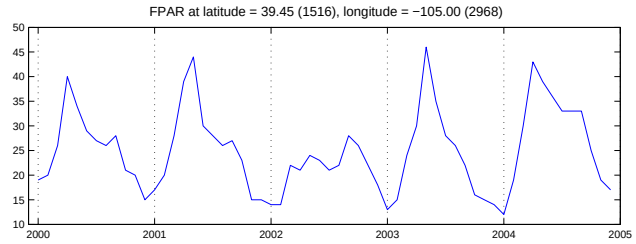
We verify our viewer with several examples of fire events happens between 2000 to 2005. Figure 9 shows a subregion in the highest resolution level, which is close to the Pike-San Isabel National Forest, 30 miles southwest of Denver, Colorado. According to a report published in [1], a forest fire hit this region around June 2002. Figure 10 shows two example locations picked from red color locations in Figure 9. The raw time series and the anomaly score by moving average method is plotted. Continuous low FPAR-Low events which lasts 12 months is successfully detected at those two locations with latitude and longitude $(39.45^\circ N, 105.00^\circ S)$ and $(39.24^\circ N, 104.61^\circ S)$.

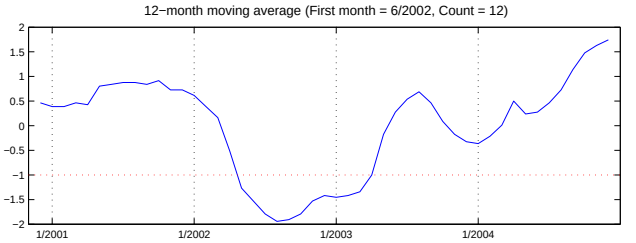
Since all our results are computed on the fly, users can adjust the parameters whenever he wants to. For example, he can change the length of the sliding window or the threshold in moving average method accord to his own purpose as well as the parameters in random walk approach.

6. EXPERIMENTAL EVALUATION

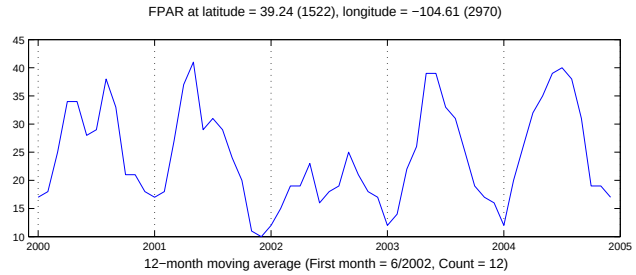
6.1 Multi-Level Indexing Evaluation

Here we would like to evaluate how different aggregation method will affect the result during the exploratory analysis on the high resolution FPAR data. The configuration of the system in this evaluation is based on 3 level indexing and $50, 100, 100$ prototypes for level $1, 2, 3$. We first evaluate how system performs when aggregating data from high resolution level to low resolution level using different definitions of prototypes. Three types of prototypes are used for Kmeans cluster: 1. data points closet to the cluster centroid, 2. data points farthest away from the cluster centroid, 2. mixed with data points farthest away from and closest to the cluster centroid. Metric in formula 3 is used to calculate the sample effect when data is aggregated from high resolution level to lower resolution level. We include the results for these methods as well as the random sampling and uniform sampling in Table 3. The sample effect for random walk and uniform is negative and close to 0, which means





(a) FPAR at 39.45°N and 105.00°S (cluster 3)



(b) FPAR at 39.24°N and 104.61°S (cluster 13)

Figure 10: Examples of FPAR time series for the Hayman Fire Event.

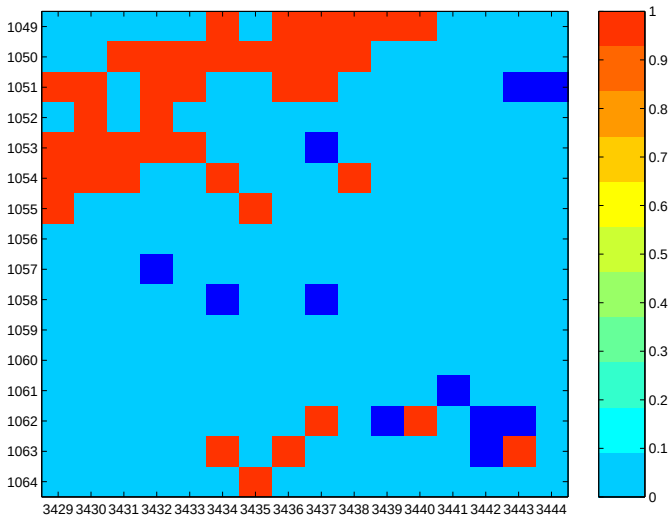


Figure 11: Disturbance locations distributed in the regions of Chisholm Fire Event at the highest resolution level.

the anomaly points are slightly down sampled. It is not good since anomaly points are usually significantly less than normal points. If it is further down sampled, users may not see anything in the lowest resolution. While clustering method gives us more control of the percentage of anomaly points we want to sample. For example, farthest method samples the points least similar to the centroid. As a result, the value of metric3 is the largest. It means much more outliers are sampled than normal points. The sample effect of mixed sample and closest sample are less than farthest since they include points closest to the centroids, which are more likely to be normal points. Similar to imbalance class learning problem [5], we need to oversample the anomaly points and downsample the normal points so that a visible percentage of anomaly points can be presented to the user. Clustering based sampling method seems to work for this purpose. Furthermore, mixed sampling enables us to adjust the percentage of closest and farthest points from the centroid, which makes a more proper way for further investigation.

	1	2	3	avg
Random	0.0088	-0.0107	-0.0754	-0.0258
Uniform	-0.0017	-0.0323	-0.0416	-0.0252
Closest	0.0469	0.1138	0.2457	0.1355
Farthest	0.0816	0.1807	3.4105	1.2243
Mixed	0.0629	0.1365	0.7471	0.3155

Table 3: Sampling effect based on different sampling strategies

6.2 Efficiency Evaluation

Table 4 shows the total runtime for computing all the disturbance events at different levels using the moving average and random walk methods. The number of grid boxes and the number of time series to scan at each level are also recorded. Level 0, which represents the highest resolution where each grid box is a 4km $\times$ 4km location, requires nearly 1.5 hours to scan the entire land points in North America for moving average method. It is even more expensive for random walk method, which requires around 5 hours. Clearly, this is infeasible for exploratory data analysis. After the data is aggregated to level 1 with 50 clusters in each grid box, the time needed to scan all the sampled time series at this level reduces

significantly to only 10 minutes for moving average method. This was further reduced to as few as 10 seconds at the lowest resolution level 3, which makes exploratory analysis feasible.

Level	0	1	2	3
Total Time (MV)	5418.8	661.7	98.4	10.9
Total Time (RW)	17710.0	702.6	110.6	16.0
#Grids	70*4^8	70*4^4	70*4^2	70
#Points	70*4^8	70*$4^4 * 50$	70*$4^2 * 100$	70*100

Table 4: Runtime (in seconds) needed to compute disturbance events at each level for North America

The preceding table shows the amount of time needed to process all the time series at each level. However, during exploratory data analysis, a user selects only one grid box to process as he/she drills down from one level to another. The run time needed to compute disturbance events for each grid box is shown in Table 5. Since all the grid points needed to be scanned at the lowest resolution when the explorer first appears, the time consumed is equal to scan all the data points at level 3, which is about 10 seconds for moving average method. When a region at level 3 is selected, the sample time series for that region will be computed for disturbance event detection. The response time at this level takes only about 1.4 seconds to compute. In turn, it takes another 0.59 seconds to drill down from level 2 to level 1 and 0.3 from level 1 to the highest resolution level 0. The same effect is observed with the random walk method too.

	1-0	2-1	3-2	3
Avg Time(MV)	0.3	0.59	1.4	10.9
Avg Time(RW)	0.98	0.63	1.58	16.0
#Points	4^4	$4^2 * 50$	4^2*100	70*100

Table 5: Response Time For Drilling Down From Low Resolution To High Resolution

7. RELATED WORK

Spatio-temporal data mining is a subject of considerable interest due to its wide range of applications. In recent years, spatio-temporal data mining has been applied to many interesting domains such as clustering spatial locations or time series with similar shapes [14], association analysis to detect strongly correlated events[13], anomaly detection for detecting ecosystem disturbances [23], and trend detection to predict future patterns in a time series [8].

Visualization techniques is a powerful tool for exploratory analysis of spatial-temporal data [9, 28]. For example, [15] presented a 3-D interactive tool for mining association rules. [19] developed a system that enables users to query clusters of time series with similar geometric shapes. These systems have mainly focused on the visualization aspect and do not consider issues such as large scale or high resolution data. Techniques such as wavelet tree [16], R-tree [18] and aRB-tree [3] are some of the well-known indexing techniques for online analytical processing.

Camossi et al. [7] presented a multilevel indexing technique based on spatial and temporal granularity. However, it does not consider the similarity between data points when constructing the index. Their technique is also especially designed for clustering task. Bertolotto et al. [6] developed an approach to use clustering for data reduction purposes, which is similar to our work. However, they did not consider the problem of choosing representative samples for a specific data mining task such as anomaly detection. Denny et al. [11] presented a multi level technique for exploratory data analysis of hot spots. While there are other works on multilevel techniques [12], they are not directly applicable to the disturbance event detection problem.

8. CONCLUSIONS

Data mining plays an important role in the discovery of interesting patterns such as ecosystem disturbances from global scale eco-climatic data. In this paper, we present two algorithms for detecting ecosystem disturbances from global vegetation cover data. We also illustrate two potential applications of clustering: (1) to assist users in categorizing different types of ecosystem disturbance events and (2) to facilitate real-time exploration and analysis of high-resolution eco-climatic data. The disturbance event detection and clustering algorithms are integrated into an interactive system developed by our team (Figure 8). The viewer enables scientists to display the FPAR time series at locations where a disturbance event was detected and to view clusters of locations where similar type of events were observed. For high resolution data, the viewer allows the user to interactively explore the data in a real-time fashion using the clustering-based indexing scheme.

9. ACKNOWLEDGMENTS

This work was supported by NSF IIS Grant #0712987.

10. REFERENCES

[1] Baer report, hayman fire, 5 july 2002. http://www.wilderness.org/Library/Documents/WildfireSummary_Haym

[2] MODIS/TERRA data versioning and maturity. http://daac.gsfc.nasa.gov/MODIS/Terra/data versioning.shtml.

[3] Indexing spatio-temporal data warehouses. In *ICDE '02: Proceedings of the 18th International Conference on Data Engineering*, page 166, Washington, DC, USA, 2002. IEEE Computer Society.

[4] NASA data mining reveals a new history of natural disasters. http://www.nasa.gov/centers/ames/news/releases/2003/03_51AR.html, July 2003. NASA Press Release 03-51AR.

[5] E. A., J. T., and J. N. A multiple resampling method for learning from imbalanced data sets. In *Computational Intelligence*, pages 18–19, 2004.

[6] M. Bertolotto, T. Kechadi, S. D. Martino, and F. Ferrucci. Scalable 2-pass data mining technique for large scale spatio-temporal datasets. In *Knowledge based and Intelligent Information and Engineering Systems*, 2007.

[7] E. Camossi, M. Bertolotto, and T. Kechadi. Mining spatio-temporal data at different levels of detail. In *Lecture Notes in Geoinformation and Cartography*, pages 225–240. Springer Berlin Heidelberg, 2008.

[8] H. Cheng, P.-N. Tan, J. Gao, and J. Scripps. Multi-step ahead time series prediction. In *Proc of the Tenth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006.

[9] M. Costabile and D. Malerba. Special issue on visual data mining. In *Journal of Visual Languages and Computing*, 2003.

[10] M. Crimmins. Synoptic climatology of extreme fire-weather conditions across the southwest united states. *International Journal of Climatology*, 26:1001–1016, 2006.

[11] Denny, G. J. Williams, and P. Christen1. Exploratory hot spot profile analysis using interactive visual drill-down

self-organizing maps. In *Advances in Knowledge Discovery and Data Mining*, pages 536–543. Springer Berlin Heidelberg, 2008.

[12] T. Dusek. Regional income differences in hungary - a multi-level spatio-temporal analysis. ERSA conference papers ersa06p284, European Regional Science Association, Aug. 2006.

[13] eremy Mennis1 and J. W. Liu. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. In *Transactions in Geography Information System*, 2005.

[14] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *Advances in Spatial and Temporal Databases*, pages 364–381, 2005.

[15] M. Kechadi and M. Bertolotto. A visual approach for spatio-temporal data mining. In *IEEE International Conference on Information Reuse and Integration*, 2006.

[16] S.-T. Li, S.-W. Chou, and J.-J. Pan. Multi-resolution spatio-temporal data mining for the study of air pollutant regionalization. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, page 7, 2000.

[17] H. Moonesignhe and P. Tan. Outlier detection using random walks. *18th IEEE International Conference on Tools with Artificial Intelligence*, pages 532–539, January 2006.

[18] M. A. Nascimento and J. R. O. Silva. Towards historical r-trees. In *Proceedings of the ACM symposium on Applied Computing*, pages 235–240, 1998.

[19] S. Olga and D. LIU. Visual interactive clustering and querying of spatio-temporal data. In *International conference on computational science and its applications*, 2005.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. Thepagerank citation ranking: Bringing order to the web. In *Technical report, Stanford University*, 1998.

[21] C. S. Potter. Terrestrial biomass and the effects of deforestation on the global carbon cycle. *BioScience*, 49:769Ű778, 1999.

[22] C. S. Potter, P. Tan, V. Kumar, C. Kucharik, S. Klooster, V. Genovese, W. Cohen, and S. Healey. Recent history of large-scale ecosystem disturbances in north america derived from the avhrr satellite record. *Ecosystems*, 8:808–824, 2005.

[23] C. S. Potter, P. Tan, M. Steinbach, V. Kumar, S. Klooster, R. Myneni, and V. Genovese. Major disturbance events in terrestrial ecosystems detected using global satellite data sets. *Global Change Biology*, 9(7):1005–1021, 2003.

[24] D. Schimel, J. House, K. Hibbard, P. Bousquet, P. Ciais, P. Peylin, M. Apps, D. Baker, A. Bondeau, R. Brasswell, J. Canadell, G. Churkina, W. Cramer, S. Denning, C. Field, P. Friedlingstein, C. Goodale, M. Heimann, R. Houghton, J. Melillo, I. B. Moore, D. Murdiyarso, I. Noble, S. Pacala, C. Prentice, M. Raupach, P. Rayner, B. Scholes, W. Steffen, and C. Wirth. Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, 414:169Ű172, 2001.

[25] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.

[26] J. Tilton. Image segmentation by region growing and spectral clustering with a natural convergence criterion. In *Geoscience and Remote Sensing Symposium Proceedings (IGARSS'98)*, page 1766Ű1768, Seattle, WA, 1998.

[27] N. Vivoy. Automatic classification of time series (acts): A new clustering method for remote sensing time series. *International Journal of Remote Sensing*, 2000.

[28] S. A. Voelz and J. Taylor. Visualizing high-resolution climate data. In *ICCS '01: Proceedings of the International Conference on Computational Sciences-Part I*, pages 212–222, London, UK, 2001. Springer-Verlag.