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
Chris.Potter@nasa.gov

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

ABSTRACT

This paper presents a case study on the application of data mining to the problem of detecting ecosystem disturbances from vegetation cover data obtained from satellite observations. We describe two anomaly detection approaches—moving average and random walk—for detecting such events. We also illustrate how clustering can be used to locate similar incidents of disturbance events. Finally, we present a clustering-based framework to aid the visual exploration of ecosystem disturbances from high resolution 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 [2]. Due to their significance and potential implications to climate change, Earth scientists are interested in detecting disturbance events at a global scale from vegetation cover data obtained from satellite observations. However, applying the detection algorithm to such a massive data set is computationally expensive. The problem is further exacerbated by the fact that the detection algorithm requires specification of one or more thresholds to determine whether a detected event should be flagged as a real disturbance. The thresholds are often determined through a trial and error process during exploratory data analysis. Because of the large amount of data that must be processed, performing exploratory data analysis on the high resolution data in a near real-time fashion is computationally infeasible. The massive size of the data also produces a large number of events for scientists to validate. This necessitates the development of innovative data mining approaches for real-time exploration of the 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 also illustrate the use of clustering to group together regions with similar incidents of disturbance events. This helps to reduce the number of events that need to be validated by the Earth science members of our team. We then 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 enables users to quickly focus on regions of interest during exploratory data analysis. The algorithms developed in this study have been integrated into an interactive system that allows scientists to explore the high resolution data and visually inspect spatial clusters with similar types of disturbance events.

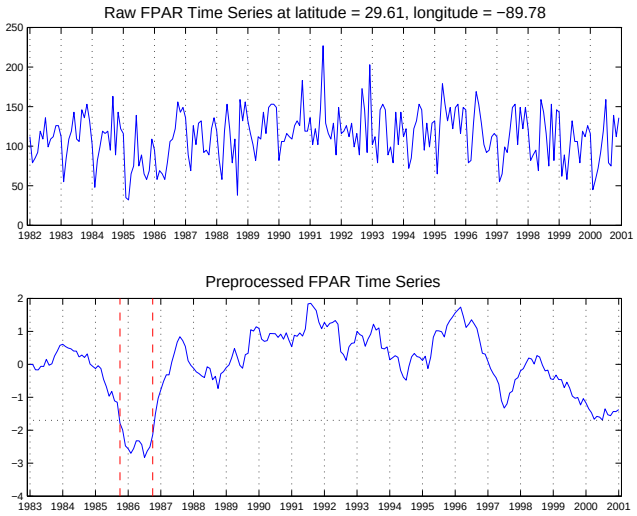


Figure 1: Detection of ecosystem disturbance due to Hurricane Elena (during September 1985) from FPAR time series data.

2. DETECTION OF DISTURBANCE EVENTS

Ecosystem disturbances are detected by monitoring changes in the FPAR (fraction of photosynthetically active radiation) vegetation cover data. A high FPAR level suggests a region with dense green leaf cover and is presumably less likely to have been disturbed. The monthly FPAR data used in this study is obtained from two sources of satellite measurements—AVHRR (which is available at 8 km spatial resolution from 1982 to 2000) and MODIS (which is available at 4 km and 1 km spatial resolution from 2000 to 2006). In this section, we present two approaches for detecting ecosystem disturbances from FPAR data.

First, we introduce the moving average approach to detect time series anomalies. The time series was first detrended using a linear adjustment and deseasonalized by computing its 12-month mov-

ACM GIS '08, November 5-7, 2008. Irvine, CA, USA

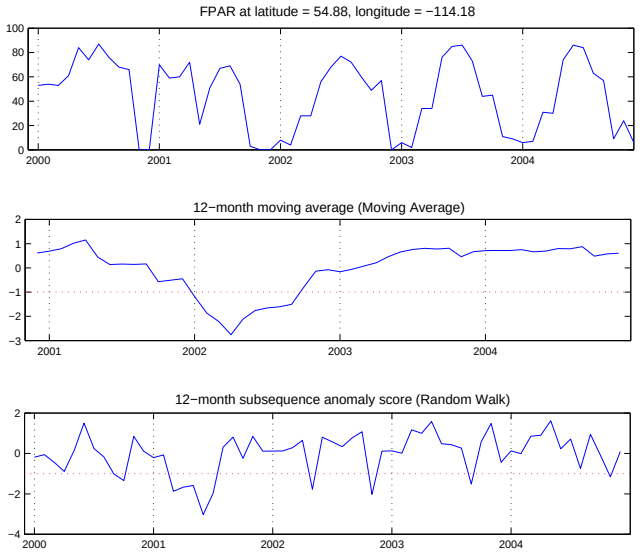


Figure 2: Detection of the Chisholm wildfire disturbance event using moving average (middle) and random walk (bottom) approaches.

ing average. It was previously hypothesized that a "sustained" disturbance event could be defined as any decline in the average annual FPAR levels (at an assigned significance level) that lasts for at least 12 consecutive months [3]. 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. There are several limitations to this approach. First, since it uses 12-month moving average to deseasonalize the time series, the timing of the event may vary depending on the FPAR levels before and after its occurrence. 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 failure of this approach to detect the disturbance event in Chisholm, Alberta where a major wildfire had occurred in May 2001.

The second approach uses a graph-based anomaly detection algorithm [1]. The algorithm constructs a kernel matrix $\boldsymbol{K}$ from the similarity values between every pair of points in the time series. The time series and its corresponding kernel matrix are then transformed into a weighted graph representation $\mathcal{G}(V, E)$ where each node $v \in V$ corresponds to a time point and each weighted edge $e \in E$ represents the similarity between two observations in the time series. The connectivity value of each node is then estimated iteratively using the following equation: $\boldsymbol{c} = d/n + (1-d)S\boldsymbol{c}$, where d is known as the damping factor and S is a transition matrix computed by normalizing the columns of the kernel matrix $\boldsymbol{K}$. This iterative procedure can be viewed as performing a random walk on the graph, where given a node u, there is a probability $1-d$ of transiting to one of its adjacent nodes and a probability d of randomly visiting any other node in the graph. Upon convergence, nodes with high connectivities are considered normal whereas those with low connectivities are declared anomalous. Figure 2 shows the result of applying our random walk algorithm to detect the Chisholm wildfire event. Unlike the moving average approach, our algorithm was able to detect the event as a node with lowest connectivity value. Note that the connectivity values have been standardized by subtracting its mean and dividing by its standard deviation.

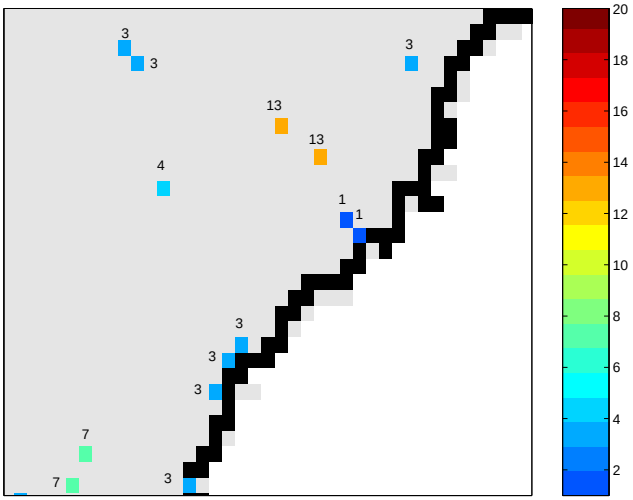


Figure 3: Clusters of disturbance events detected around the North Carolina region.

3. CLUSTERING FOR EVENT CATEGORIZATION

Clustering is the task of partitioning data into groups of similar objects. 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. We have applied k-means clustering (with k=20) to categorize the detected events based on characteristics of the climate (temperature and precipitation) and FPAR time series during or after the event had occurred. Figure 3 shows the clusters found around the North Carolina region. Each colored pixel corresponds to a location where disturbance event has been previously detected. We have also labeled each pixel according to its cluster ID. Our preliminary results suggest the possibility of distinguishing the different types of ecosystem disturbance events according to their cluster assignment. For example, cluster #1 is associated with disturbance events due to Hurricane Hugo.

4. 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 high resolution MODIS data, applying the anomaly detection algorithm to every data point is computationally expensive. For example, with the 4 km × 4 km MODIS data, there are more than 1.3 million data points[1] for North America alone. Although the algorithm scales linearly with the size of the data, it still takes more than several minutes to process the data for the entire North America. Repeating the analysis using different thresholds takes considerable amount of time, making it infeasible for real-time data exploration. The runtime will increase considerably if the analysis is to be performed on the 1 km × 1 km MODIS data or using more sophisticated anomaly detection algorithms. Techniques must therefore be developed to reduce the processing time.

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

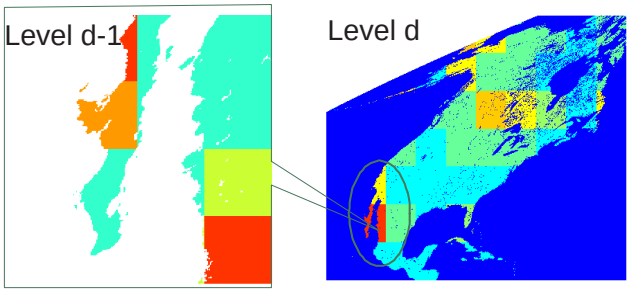


Figure 4: A multilevel approach for visual exploration of ecosystem disturbances.

4.1 Clustering-based Multi-level Indexing

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. Our strategy is to build a multilevel index by applying k-means clustering on the FPAR time series and choosing representative samples from the clusters at each level of the index hierarchy. This strategy allows us to apply the moving average or random walk disturbance event detection algorithm on the sampled time series instead of the entire time series. In turn, this reduces the amount of processing time considerably, thus allowing the user to tune the thresholds of their algorithm and observe the changes in the results in near real-time.

Figure 4 shows an example of how the multi-level indexing scheme works. Specifically, we use the index to monitor the frequency distribution of ecosystem disturbance events in North America. 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 4, the map at level d displays the frequency distribution of disturbance events for each region in North America. The map shows that the Baja California peninsula, the midwest region, and Canada appears to have highest concentration of disturbance events. 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 regions; 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.

4.2 Sampling 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 (d) to the next level ($d+1$). 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 computation time, we sample the time series in each grid box so that the disturbance event detection algorithm is applied only to the sampled time series. The key challenge here is to obtain a representative sample for each region.

Suppose N time series must be sampled from each grid box. We apply k-means clustering to the time series 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 two sampling approaches: (1) **Closest Sampling**, which selects the time series that are closest to the cluster centroid, and (2) **Farthest Sampling**, which selects the time series that are 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. During the construction of the multi-level index, the pixel locations of the sampled time series are stored in an 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 4).

To evaluate the effectiveness of the sampling strategy, we compute the fraction of disturbance events missed due to sampling:

$$\text{Loss} = \frac{1}{d} \sum_{l=1}^{d} \sum_{b_i \in \Omega_l} \frac{N_a(b_i;\theta) - N_s(b_i;\theta)}{N_a(b_i;\theta)} \quad (1)$$

where d is the number of levels, Ω_l is the set of grid boxes at level l, θ is the event definition threshold, $N_a(b_i;\theta)$ is the actual number of disturbance events in the grid box, and $N_s(b_i;\theta)$ is the corresponding number of disturbance events in the sampled time series.

4.3 Experimental Evaluation

In this experiment, we evaluate the effectiveness and efficiency of our clustering-based multilevel indexing scheme. A 4-level index is constructed from the MODIS FPAR data, with 50, 100, and 100 samples chosen for each grid box at levels 1, 2, and 3, respectively. We compare the closest and farthest sampling strategies against two baseline approaches—random sampling of time series from each cluster and uniform sampling across the region. Table 1 shows that the loss for random and uniform sampling is more than 77%, which is considerably higher than farthest sampling. This result suggests that, by sampling the time series that are considerably different than the cluster centroid, we are more likely to sample time series that contain anomalies (i.e., disturbance events).

	1	2	3	avg
Random	0.7679	0.7319	0.8356	0.7784
Uniform	0.7699	0.7196	0.8213	0.7703
Closest	0.7595	0.7008	0.7994	0.7532
Farthest	0.7527	0.6956	0.4243	0.6242

Table 1: Comparison of sampling loss for different strategies

Table 2 shows the runtime for computing the disturbance events for all grid boxes at each level using the moving average (**MV**) and random walk (**RW**) methods. The number of grid boxes and the number of time series to scan at each level are also recorded. Level 0 requires nearly 1.5 hours to process all the land points in North America using the 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 aggregating the data to level 1 (with 50 clusters in each grid box),

the time needed to process the sampled time series reduces significantly to 10 minutes for the moving average method. This was further reduced to as few as 10 seconds at level 3.

Level	0	1	2	3
Total Time (MV)	5418.8	661.7	98.4	10.9
Total Time (RW)	17710.0	1839.6	254.6	56.0
#Grids (70×)	4^8	4^4	4^2	1
#Points (70×)	4^8	$4^4 \times 50$	$4^2 \times 100$	100

Table 2: Runtime (in seconds) needed to compute disturbance events at each level for North America

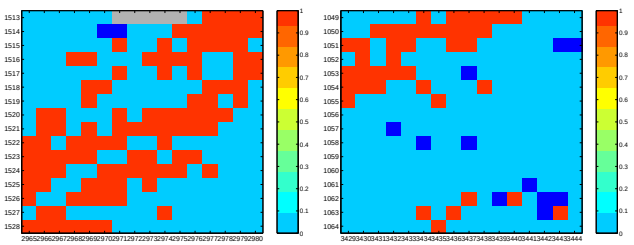
The preceding table shows the time needed to process the sampled time series for all grid boxes at each level. However, during exploratory data analysis, a user selects only one grid box to explore. The time needed to compute disturbance events for each grid box is given in Table 3. At level 3, since the sampled time series for all grid boxes must be processed, the runtime is equal to that given in Table 2. When the user selects one of the regions to explore, the response time is only 1.4 seconds. It takes another 0.59 seconds to drill down from level 2 to level 1 and 0.3 seconds from level 1 to level 0. A similar pattern is observed for the random walk method.

	1-0	2-1	3-2	3
Avg Time(MV)	0.3	0.59	1.4	10.9
Avg Time(RW)	3.24	10.04	18.1	56.0
#Points	4^4	$4^2 \times 50$	$4^2 \times 100$	70×100

Table 3: Response Time For Drilling Down From Low Resolution To High Resolution

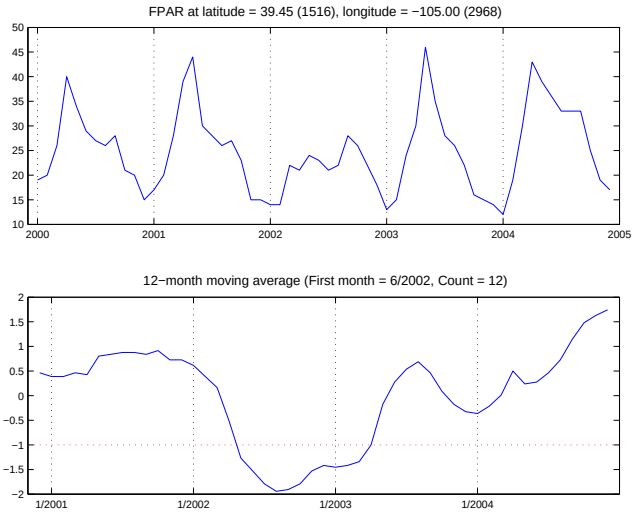
5. USER-INTERACTIVE SYSTEM

Here we introduce the interactive system that integrates the algorithms developed in this study. Initially, the map will display the frequency distribution of disturbance events for the selected continent. 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. The process is repeated until the highest level is reached, where the user can click on any pixel location to display the FPAR time series and the corresponding anomaly score at each time point.



(a) Hayman Fire (b) Chisholm Fire

Figure 5: Disturbance locations distributed in the regions of Hayman Fire Event at the highest resolution level.



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

Figure 6: FPAR time series for the Hayman Fire Event.

We verify the effectiveness of our algorithms using several documented major wildfires occurring between 2000 to 2005. Figure 5(a) shows the map for the region around the Pike-San Isabel National Forest, where a major wildfire was reported in June 2002 [4]. Figure 6 shows the FPAR time series for one of the locations in this region. As can be seen from the figure, the Hayman wildfire event was successfully detected using our approach.

6. CONCLUSIONS

This paper presents a case study on the application of data mining to the disturbance event detection problem. We describe two anomaly detection algorithms for finding ecosystem disturbances and illustrate two applications of clustering: (1) to assist users in categorizing different types of disturbance events and (2) to facilitate real-time exploration of high-resolution vegetation cover data. The algorithms have been integrated into an interactive viewer that enables scientists to display the FPAR time series at locations where disturbance events have been detected. For high resolution data, the viewer allows the user to explore the data in near real-time fashion using a multilevel indexing scheme.

7. ACKNOWLEDGMENTS

This work was supported by NSF IIS Grant #0712987.

8. REFERENCES

[1] H. Moonesinghe and P. Tan. Outlier detection using random walks. In *Proc. 18th IEEE Int'l Conf. on Tools with Artificial Intelligence*, pages 532–539, January 2006.

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

[3] 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.

[4] B. Report. Hayman fire. `http://www.wilderness.org/Library/Documents/WildfireSummary\_Hayman.cfm`.