GEOSPATIOTEMPORAL DATA MINING IN AN EARLY WARNING SYSTEM FOR FOREST THREATS IN THE UNITED STATES

F. M. Hoffman[*], *R. T. Mills*[†], *J. Kumar*[‡], *S. S. Vulli*[§]

Computer Science & Mathematics Division
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831

W. W. Hargrove[¶]

Eastern Forest Environmental
Threat Assessment Center
200 WT Weaver Blvd.
Asheville, NC 28804-3454

ABSTRACT

We investigate the potential of geospatiotemporal data mining of multi-year land surface phenology data (250 m Normalized Difference Vegetation Index (NDVI) values derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) in this study) for the conterminous United States as part of an early warning system to identify threats to forest ecosystems. Cluster analysis of this massive data set, using high-performance computing, provides a basis for several possible approaches to defining the bounds of "normal" phenological patterns, indicating healthy vegetation in a given geographic location. We demonstrate the applicability of such an approach, using it to identify areas in Colorado, USA, where an ongoing mountain pine beetle outbreak has caused significant tree mortality.

***Index Terms*—** Remote sensing, data mining, phenology, cluster analysis, high-performance computing

1. THE FOREST INCIDENCE RECOGNITION AND STATE TRACKING (FIRST) SYSTEM

Early identification of forested areas under attack from insects or disease can enable timely response to protect forest ecosystems from long-term or irreversible damage. Unfortunately, given the sheer size of the United States and limited resources of agencies such as the USDA Forest Service to conduct aerial surveys and ground-based inspections, many threats go unnoticed until a great deal of damage has already been done. To improve threat detection, the USDA Forest Service, in partnership with Oak Ridge National Laboratory and the NASA Stennis Space Center, is developing The Forest Incidence Recognition and State Tracking (FIRST) early warning system. FIRST will detect and monitor threats to forests and wildlands in the conterminous United States (CONUS) as part of a two tier system: An early warning system that monitors continental-scale areas at a moderate resolution using remote sensing data to spatially direct and focus efforts of the second tier, consisting of higher resolution monitoring through airborne overflights—called Aerial Detection Survey (ADS) sketch-mapping—and ground-based inspections. Tier 2 is largely in operation today, but the strategic direction provided by the FIRST system in Tier 1 will improve the efficiency and utility of these costly and labor-intensive surveys.

[*] forrest@climatemodeling.org
[†] rmills@ornl.gov
[‡] jkumar@climatemodeling.org
[§] shivakar@climatemodeling.org
[¶] hnw@geobabble.org

The goals of the FIRST early warning system are to provide a single, unified system for change detection from remotely sensed vegetation properties through time over the domain of the conterminous United States at about 250 m nominal resolution—obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors on board NASA's Terra and Aqua satellites—at frequent intervals, on the order of one week. The system must be automated, requiring unsupervised data mining methods, and provide results as close to real-time as possible. It must "learn" or improve its prognostic ability utilizing a library of previous experiences, including both true and false warnings with attribution to causes for the former. FIRST will utilize data on soils, topography, climatology, and weather events, as well as satellite-derived vegetation parameters. Because of the huge data volumes involved, even at this moderate resolution, FIRST must employ highly scalable data mining and statistical algorithms that operate on very large data sets using moderate- to large-sized clusters and supercomputers. A schematic overview of the FIRST system, as envisioned, is shown in Figure 1.

One of the most important types of data that FIRST will utilize is satellite-derived data indicating the annual temporal patterns of variation in vegetation greenness, i.e., the ecosystem phenology.

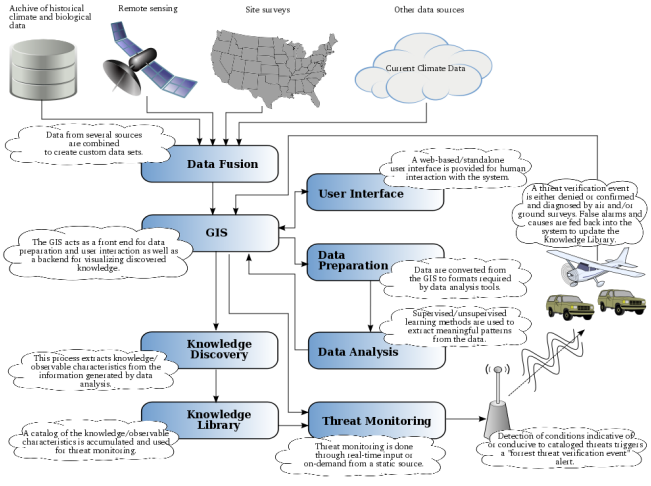

Fig. 1. A conceptual overview of the Forest Incidence Recognition and State Tracking (FIRST) System.

This paper explores the application of data mining techniques to analyze seasonal changes in Normalized Difference Vegetation Index (NDVI) derived from MODIS coverage of the CONUS. The utility of these data has already been demonstrated in [1], in which the authors used raster map arithmetic approaches, such as comparing maximum NDVI from mid-summer against maximum NDVI over a six-year baseline, to detect potential forest disturbances. Some of these disturbances could represent threats to the long-term health and functioning of forest ecosystems. A difficulty with using such approaches is identification of appropriate parameters (maximum NDVI, 20% "spring" NDVI, etc.) to use, since the appropriate choice of parameters may vary by region and/or type of forest disturbance. Here, we experiment with approaches that do not depend on choosing particular parameters; instead, using high-performance computing, we apply geospatiotemporal data mining techniques to perform unsupervised classification based on multiple years of NDVI history for the entire CONUS. These classifications use the full volume of available NDVI data (rather than only a small subset of selected parameters) to construct a potential basis for determining the "normal" seasonal and inter-seasonal variation expected at a geographic location and to detect deviations from the norm that merit additional Tier 2 scrutiny.

2. GEOSPATIOTEMPORAL DATA MINING

Hargrove and Hoffman have developed and applied a scalable multivariate statistical procedure that can be used to define a set of categorical, multivariate classes or states that are useful for describing and tracking the behavior of ecosystem properties through time within a multi-dimensional phase or state space [2, 3, 4]. Referred to as geospatiotemporal data mining, the procedure employs a standard k-means cluster analysis with enhancements that reduce the number of required comparisons, dramatically accelerating iterative convergence, and dynamically optimizing centroid placement within iterations to avoid member-less or empty clusters. This enhanced k-means cluster analysis algorithm has been implemented and tested on large, high performance computing platforms, enabling the analysis of very large, high-resolution remotely sensed data. This geospatiotemporal data mining method was previously applied to remotely sensed hyperspectral imagery for detection of brine scars [5] and to monthly climate and NDVI data from 17 years of 8 km Advanced Very High Resolution Radiometer (AVHRR) images for land surface phenology [6], suggesting that this method could be a key component of an early warning system for detecting forest threats.

Cluster analysis is sensitive to the initial centroids chosen, and traditional methods for intelligently determining initial centroids are intractable for very large data sets. For these data, initial centroids are determined using the method described by Bradley and Fayyad [7], involving two phases of subsampling and cluster trials to produce high quality, highly representative initial centroids. In Phase 1 of Bradley's procedure, the large data set is subsampled many times, N_s, and each subsample is clustered, using k randomly-selected observations from the subsample as initial centroids each time. In Phase 2, the N_s groups of k centroids that result from these smaller clustering trials are combined to produce a single new data set that substitutes for the original data. This data set is then clustered N_s times, using as initial centroids each of the N_s groups of k centroids. A pseudo-F statistic is computed for each of these Phase 2 trials, and the group of centroids that result from the winning trial (i.e., the one with the largest pseudo-F score) is chosen to be the set of initial centroids for clustering the entire, large data set.

For identifying forest disturbances that may represent threats to forest health, the FIRST system will employ remotely sensed land surface phenology along with data about soils and climate. NDVI, a remote sensing product commonly used for vegetation monitoring, can potentially identify reductions in greenness of vegetation due to drought, insect or pathogen invasion, storm damage, and forest regrowth. Deviations from "normal" phenological patterns in NDVI are often the first indication of changes in forest disturbance and recovery. An initial cluster analysis of six years (2003–2008) of the annual cycle of NDVI was performed for the entire CONUS, producing annual maps of phenological ecoregions or "phenoregions" [6]. The NDVI data were derived from the MODIS MOD 13 NDVI product and pre-processed at NASA's Stennis Space Center. The processed NDVI data for the CONUS have a 250 m spatial resolution and are available every 16 days. At this spatial and temporal resolution, each map contains more than 148M cells, and 22 maps are created for each year. Hence, this cluster analysis was performed on 116B NDVI values arranged as annual NDVI traces, providing 22 state space dimensions, for each grid cell (148M records) for each of the six years. The resulting input data set, stored in single-precision binary format, is 77 GB in size. Output from the cluster analysis consists of six maps, one for each year (2003–2008), in which each cell is classified into one of k phenostates, which are defined for the annual NDVI traces across all six years. As a result, the time evolution of phenostates assignment, or phenostate, for every cell in the map indicates a change in the phenological behavior and ecosystem productivity observed at that location due to natural or anthropogenic disturbance, forest regrowth, or ecosystem responses to interannual changes in climate. A map of 50 phenoregions for the CONUS for the year 2008 derived from geospatiotemporal cluster analysis from these data is shown in Figure 2. Comparison of the curre nt phenostate with the nominal behavior of healthy vegetation indicated by the h istorical phenoregion assignment at every location in the CONUS could allow a national early warning system to identify locations where the vegetation appears to deviate from its usual phenological behavior [1].

3. DETECTING ANOMALIES WITH GEOSPATIOTEMPORAL CLUSTERING

A direct approach to detecting anomalies with geospatiotemporal clustering would be to examine the current phenostate compared to historical phenostates at a given map cell, and then flag the present state of a cell as "abnormal" if the cell has very infrequently or never occupied this state in the past. This approach, however, depends on having chosen an appropriate number of clusters, k. If k is too large, then the normal seasonal variation in NDVI will likely result in a different phenostate assignment each year, leading to many "false positives" comission errors, even though the different phenostates may, in fact, be very similar. Because the normal seasonal pattern of NDVI varies regionally and by biome, selecting an appropriate value of k for the entire CONUS may not be possible. This simple method cannot take into account the fact that a newly observed phenostate may, in fact, be very similar to previously observed states at that map cell. An alternative approach for change detection is to create maps of the "transition distance" between years, plotting at each map cell the Euclidean distance between the new and old phenostate centroids; this distance gives a relative multivariate measure of how different the observed phenology is between the two years.

For example, the map in Figure 3 depicts the transition distance between phenostate transitions between the year 2003 and the year 2008 in Colorado, USA. A mountain pine beetle (MPB) outbreak, which began before 2003 and is still ongoing, has caused

significant mortality in Ponderosa and lodgepole pines in Colorado and Wyoming. Areas of high transition distance in the mountains (central and western portions of the state) correspond closely to areas of MPB activity noted by aerial sketch-map surveys, shown as black-outlined polygons in the figure. Given the inexact nature of such these surveys, the spatial correspondence between the largest phenostate transitions and the sketch-map polygons is high. The transition distance map shown in Figure 3 may provide a more comprehensive assessment of MPB damage then the sketch-maps. This 2003–2008 transition distance map depicts the cumulative damage by MPB over the entire time period while year-to-year transition maps for this period (not shown) allow one to chart the yearly progression of the MPB outbreak.

4. CONCLUSIONS

Initial results from the geospatiotemporal cluster analysis of annual phenology patterns from MODIS NDVI confirm its utility for unsupervised change detection in remote sensing data and suggest that it may be successfully implemented as a key component in the FIRST early warning system, which is designed to detect forest threats from natural and anthropogenic disturbances at a continental scale. The enhanced k-means clustering algorithm, which can run on computing platforms ranging from small cluster computers to the largest and fastest supercomputers in the world, enables the analysis of very large, high resolution remote sensing data such as these. While determining what constitutes a "normal" phenological pattern for any given location is challenging—due to interannual climate variability, a spatially varying climate change trend, and the relatively short record of MODIS NDVI observations—significant mortality events, like the progressive damage from MPB, are already easily detectable by simply computing relative transitions between blocks of successive years of phenostates. Moreover, as anomalies are detected and tracked through time, a library of phenostates transitions attributed to pests or pathogens for individual biomes can be built up, allowing the system to hypothesize about the causes of future disturbances detected in functionally similar biomes.

5. ACKNOWLEDGMENTS

The authors wish to thank Joseph P. Spruce at the NASA Stennis Space Center for providing quality controlled NDVI maps generated from the MODIS MOD 13 product. This research was sponsored by the U.S. Department of Agriculture Forest Service, Eastern Forest Environmental Threat Assessment Center. This research used resources of the National Center for Computational Science at Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

6. REFERENCES

[1] William W. Hargrove, Joseph P. Spruce, Gerald E. Gasser, and Forrest M. Hoffman, "Toward a national early warning system for forest disturbances using remotely sensed phenology," *Photogrammetric Engineering & Remote Sensing*, vol. 75, no. 10, pp. 1150–1156, Oct. 2009.

[2] Forrest M. Hoffman, William W. Hargrove, Richard T. Mills, Salil Mahajan, David J. Erickson, and Robert J. Oglesby, "Multivariate Spatio-Temporal Clustering (MSTC) as a data mining tool for environmental applications," in *Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Society (iEMSs 2008)*, Miquel Sànchez-Marrè, Javier Béjar, Joaquim Comas, Andrea E. Rizzoli, and Giorgio Guariso, Eds., Barcelona, Catalonia, Spain, July 2008.

[3] Forrest M. Hoffman, William W. Hargrove, David J. Erickson, and Robert J. Oglesby, "Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models," *Earth Interact.*, vol. 9, no. 10, pp. 1–27, Aug. 2005, doi:10.1175/EI110.1.

[4] William W. Hargrove and Forrest M. Hoffman, "Potential of multivariate quantitative methods for delineation and visualization of ecoregions," *Environ. Manage.*, vol. 34, no. 5, pp. s39–s60, 2004, doi:10.1007/s00267-003-1084-0.

[5] Forrest M. Hoffman, "Analysis of reflected spectral signatures and detection of geophysical disturbance using hyperspectral imagery," M.S. thesis, Department of Physics and Astronomy, University of Tennessee, Knoxville, Nov. 2004.

[6] Michael A. White, Forrest Hoffman, William W. Hargrove, and Ramakrishna R. Nemani, "A global framework for monitoring phenological responses to climate change," *Geophys. Res. Lett.*, vol. 32, no. 4, Feb. 2005, doi:10.1029/2004GL021961.

[7] Paul S. Bradley and Usama M. Fayyad, "Refining initial points for k-means clustering," in *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1998, pp. 91–99, Morgan Kaufmann Publishers Inc.

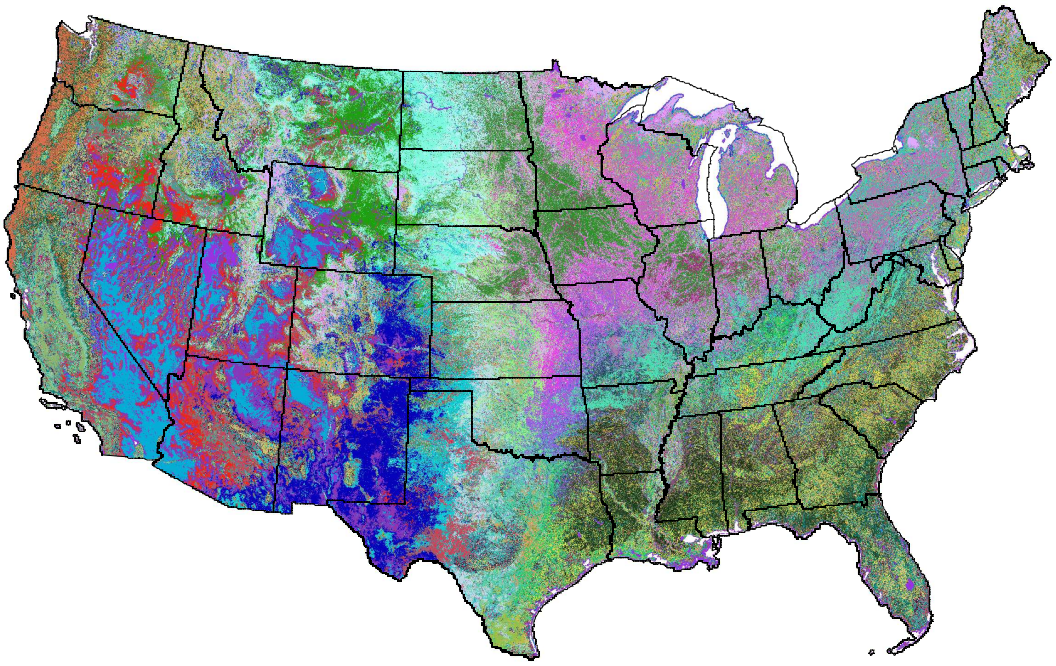

Fig. 2. The 2008 map of 50 phenoregions defined for the CONUS derived from geospatiotemporal cluster analysis of MODIS NDVI data.

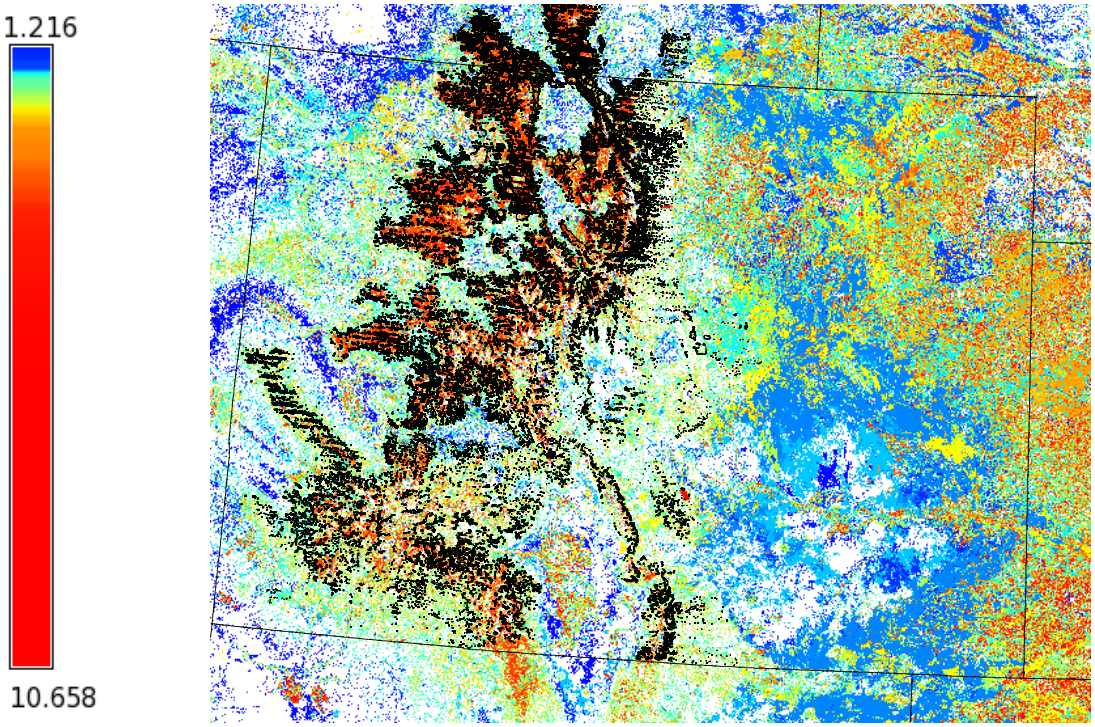

Fig. 3. The map of relative state-space transition distances for phenoregions between 2003–2008 for Colorado, USA.