Geospatiotemporal Data Mining of Remotely Sensed Phenology for Unsupervised Forest Threat Detection

Richard T. Mills$^{\alpha}$**, Forrest M. Hoffman**$^{\alpha}$**, Jitendra Kumar**$^{\alpha}$**, Srinivasa S. Vulli**$^{\alpha}$**, William W. Hargrove**$^{\beta}$**, Joseph Spruce**$^{\gamma}$

$^{\alpha}$Oak Ridge National Laboratory, $^{\beta}$USDA Forest Service, $^{\gamma}$John C. Stennis Space Center

Abstract

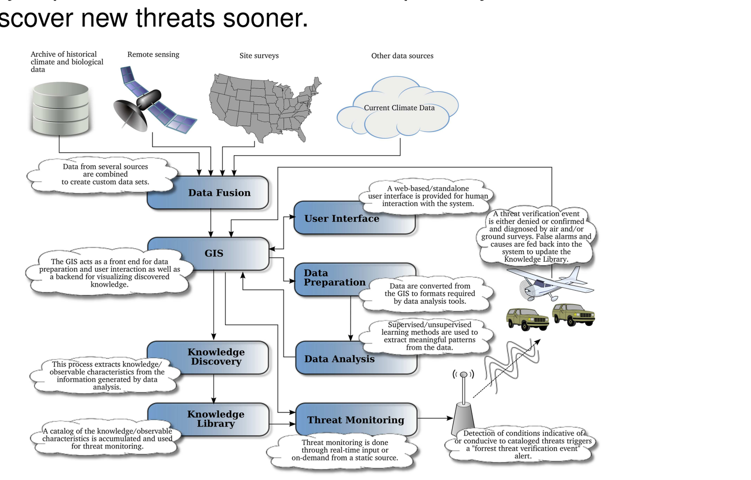
Hargrove and Hoffman have previously developed and applied a scalable geospatiotemporal data mining approach to define a set of categorical, multivariate classes or states for describing and tracking the behavior of ecosystem properties through time within a multi-dimensional phase or state space. The method employs a standard k-means cluster analysis with enhancements that reduce the number of required comparisons, dramatically accelerating iterative convergence. In support of efforts by the USDA Forest Service to develop a National Early Warning System for Forest Disturbances, we have applied this geospatiotemporal cluster analysis procedure to annual phenology patterns derived from Moderate Resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index (NDVI) for unsupervised change detection. We will present initial results from the analysis of seven years of 250-m MODIS NDVI data for the conterminous United States. 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, these results demonstrate the utility of the method for detecting significant mortality events, like the progressive damage from mountain pine beetle, and suggest that the technique may be successfully implemented as a key component in an early warning system for identifying forest threats from natural and anthropogenic disturbances at a continental scale.

1. Introduction

The USDA Forest Service, NASA Stennis Space Center, and DOE Oak Ridge National Laboratory are creating a system to monitor threats to U.S. forests and wildlands at two different scales:

- Tier 1: Strategic — An Early Warning System (EWS) that routinely monitors wide areas at coarser resolution, repeated frequently — a *change detection system* to produce alerts or warnings for particular locations may be of interest
- Tier 2: Tactical — Finer resolution airborne overflights and ground inspections of areas of potential interest — *Aerial Detection Survey (ADS)* monitoring to determine if such warnings become alarms

Tier 2 is largely in place, but Tier 1 is needed to optimally direct its labor-intensive efforts and discover new threats sooner.



2. Normalized Difference Vegetation Index (NDVI) from MODIS

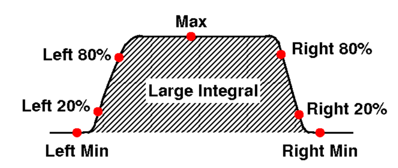
- NDVI exploits the strong differences in plant reflectance between red and near-infrared wavelengths to provide a measure of "greenness" from remote sensing measurements.

$$\text{NDVI} = \frac{(\sigma_{\text{nir}} - \sigma_{\text{red}})}{(\sigma_{\text{nir}} + \sigma_{\text{red}})} \quad (1)$$

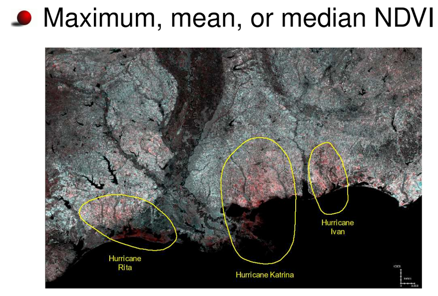
- These spectral reflectances are ratios of reflected over incoming radiation, $\sigma = I_r/I_i$, hence they take on values between 0.0 and 1.0. As a result, NDVI varies between -1.0 and $+1.0$.
- Dense vegetation cover is 0.3–0.8, soils are about 0.1–0.2, surface water is near 0.0, and clouds and snow are negative.
- The Moderate Resolution Imaging Spectroradiometer (MODIS) is a key instrument aboard the Terra (EOS AM, N→S) and Aqua (EOS PM, S→N) satellites.
- Both view the entire surface of Earth every 1 to 2 days, acquiring data in 36 spectral bands.
- The MOD 13 product provides Gridded Vegetation Indices (NDVI and EVI) to characterize vegetated surfaces.
- Available are 6 produces at varying spatial (250 m, 1 km, 0.05°) and temporal (16-day, monthly) resolutions.
- The Terra and Aqua products are staggered in time so that a new product is available every 8 days.
- Results shown here are derived from the 16-day Terra MODIS product at 250 m resolution, processed by NASA Stennis Space Center.

3. Phenology

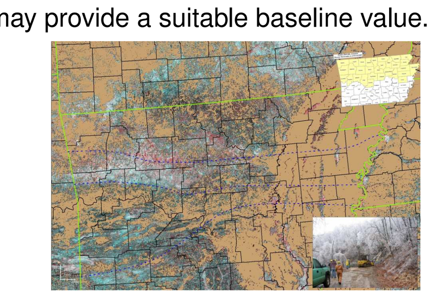
- Phenology is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate.
- FIRST is interested in deviations from the "normal" seasonal cycle of vegetation growth and senescence.
- NASA Stennis Space Center has developed a new set of National Phenology Datasets based on MODIS.
- Outlier/noise removal and temporal smoothing are performed, followed by curve-fitting and estimation of descriptive curve parameters.
- An idealized seasonal NDVI curve is fit through data for each MODIS cell, and seven parameters are extracted.



- Each parameter results in two maps: one for the NDVI value and one for the time of the event.
- Cumulative NDVI shows the annual "greening" of the U.S.
- The Large Integral is strongly correlated with annual gross primary production (GPP) of the conterminous U.S. (CONUS).
- To detect vegetation disturbances, the current NDVI measurement is compared with the normal, expected baseline for the same location.
- Substantial decreases from the baseline represent potential disturbances.
- Any increases over the baseline may represent vegetation recovery.
- Maximum, mean, or median NDVI may provide a suitable baseline value.



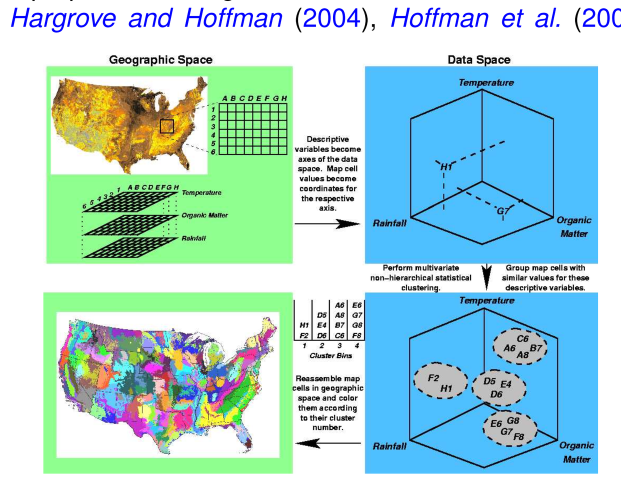
Hurricane damage. Computed by assigning 2006 20% left value to green & blue, and 20% left from 2004 to red (*Hargrove et al.*, 2009). Red depicts areas of reduced greenness, primarily east of storm tracks and in marshes.



Damage from ice storms in the Arkansas Ozarks. Computed by assigning 2009 max NDVI for June 10–July 15 into blue & green, and 2001–2006 max NDVI for June 10–July 27 into red. Storm resulted in 35,000 without power and 18 fatalities.

4. Characterizing Phenology via Geospatiotemporal Data Mining

- Map arithmetic on selected parameters is good for studying the impact of known disturbances, but what is desired is an automated, unsupervised change detection system.
- Here, we utilize high performance computing (HPC) for the entire body of the very large, high resolution NDVI data history.
- We build on a geospatiotemporal data mining approach by Hoffman and Hargrove 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 *Hargrove and Hoffman* (2004), *Hoffman et al.* (2005, 2008).



- Employs 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 memberless or empty clusters.
- The enhanced 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.
- Here, we performed cluster analysis 116B NDVI values (250 m spatial resolution, 16 day intervals, years 2003–2009, with 2007 omitted due to data processing errors yet to be corrected) 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, in which each cell is classified into one of k phenostates, which are defined for the annual NDVI traces across all six years.
- 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.

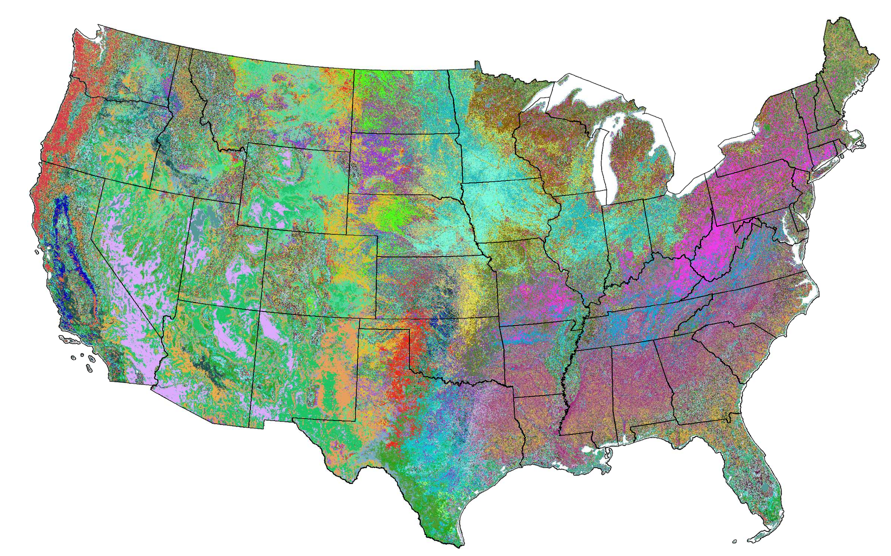


Figure 1: 50 Phenoregions for Year 2009

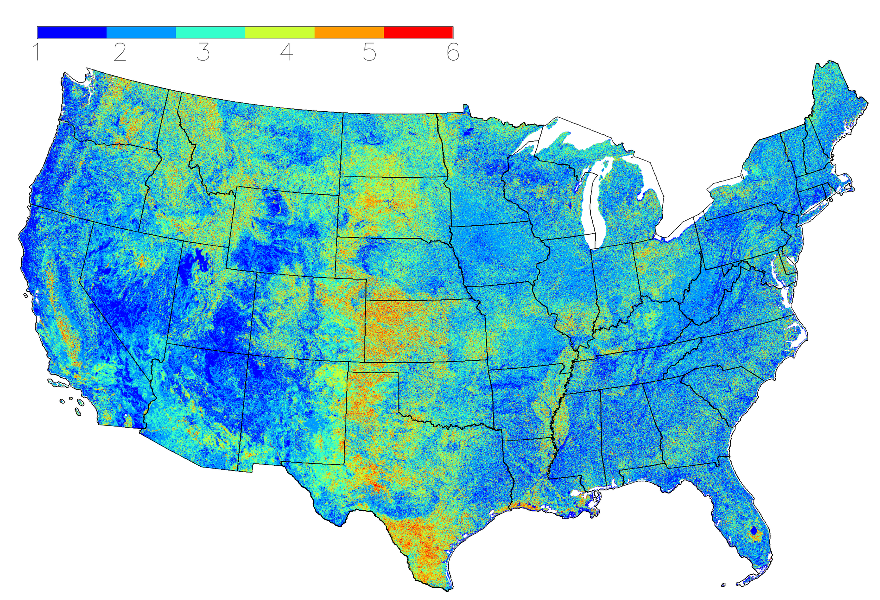


Figure 2: Map of Cluster Persistence for 2003–2009, excluding 2007

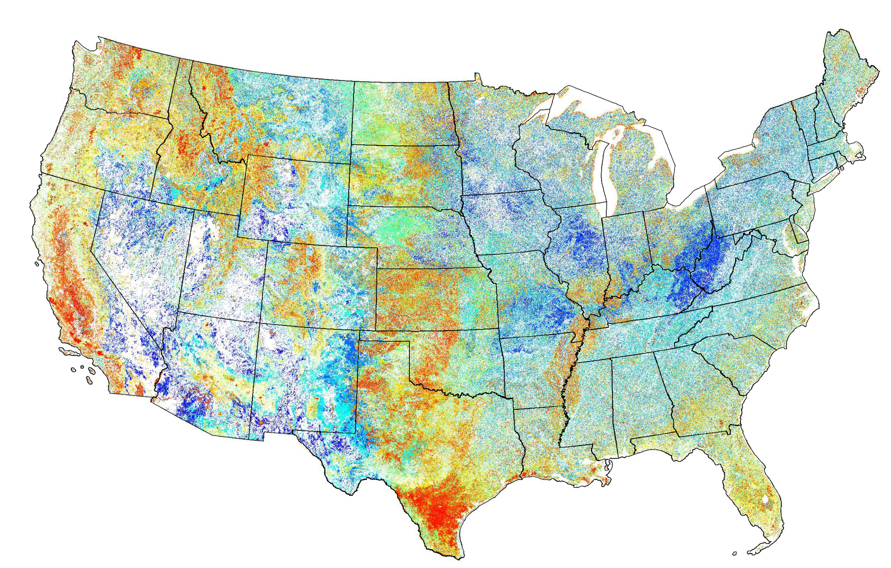


Figure 3: Cluster Transition Distances for (2009 – 2003)

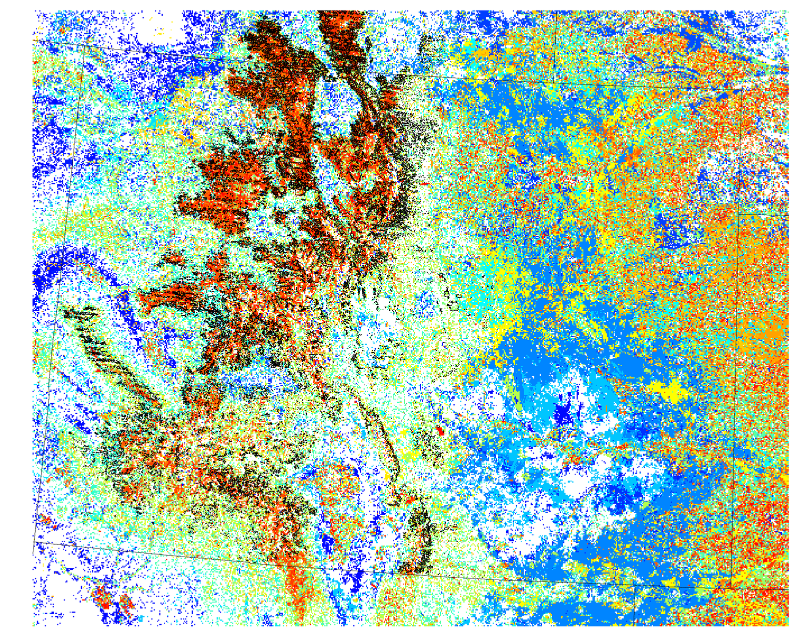


Figure 4: Mountain Pine Beetle in Colorado for (2008 – 2003)

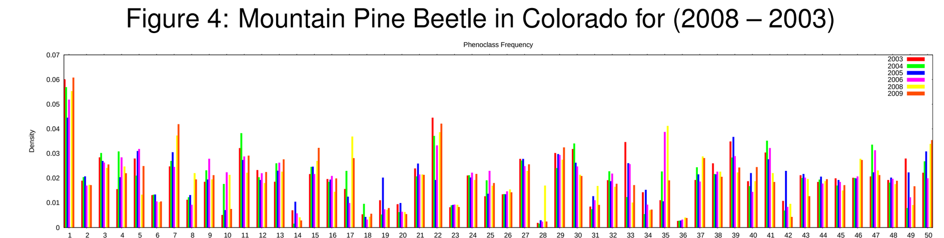


Figure 5: Histograms of cluster occupation by year for 2003–2009, excluding 2007

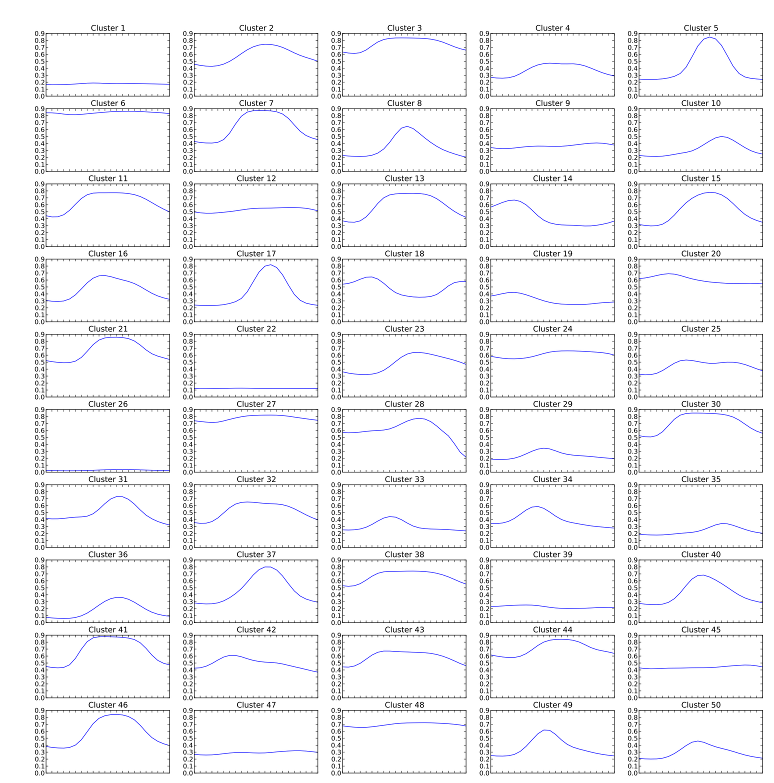


Figure 6: Plots of the 50 phenostates from the 2003–2009 NDVI clustering

5. Detecting Anomalies with Geospatiotemporal Clustering

- A straightforward anomaly detection approach 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" commission 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.
- Figure 4 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 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.

6. Conclusions and Future Work

- Initial results of geospatiotemporal cluster analysis of phenology from MODIS NDVI are promising, suggesting such analysis will be a key component in the FIRST early warning system.
- The enhanced, accelerated k-means clustering algorithm enables the analysis of very large, high resolution remote sensing data.
- Determining "normal" phenological patterns is difficult—due to interannual climate variability, spatially variable climate change trend, and relatively short satellite record—mortality events, like progressive Mountain Pine Beetle damage, are easily detected.
- The next step is to establish biome-specific thresholds based on interannual variability, obtain validation from ADS and ground surveys, and track and accumulate both loss and new growth for carbon accounting.
- Future work will build a library of phenostate transitions attributed to pests or pathogens for individual biomes, allowing the system to hypothesize about causes of future disturbances detected.

References

Hargrove, W. W., and F. M. Hoffman (2004), Potential of multivariate quantitative methods for delineation and visualization of ecoregions, *Environ. Manage.*, *34*(S1), S39–S60, doi: 10.1007/s00267-003-1084-0.

Hargrove, W. W., J. P. Spruce, G. E. Gasser, and F. M. Hoffman (2009), Toward a national early warning system for forest disturbances using remotely sensed phenology, *Photogramm. Eng. Rem. Sens.*, *75*(10), 1150–1156.

Hoffman, F. M., W. W. Hargrove, D. J. Erickson, and R. J. Oglesby (2005), Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models, *Earth Interact.*, *9*(10), 1–27, doi:10.1175/EI110.1.

Hoffman, F. M., W. W. Hargrove, R. T. Mills, S. Mahajan, D. J. Erickson, and R. J. Oglesby (2008), 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)*, edited by M. Sànchez-Marrè, J. Béjar, J. Comas, A. E. Rizzoli, and G. Guariso, Barcelona, Catalonia, Spain.

For more information, see
http://www.geobabble.org/FIRST