Cluster Analysis-Based Approaches for Geospatiotemporal Data Mining of Massive Data Sets for Identification of Forest Threats

Richard T. Mills†, Forrest M. Hoffman†, Jitendra Kumar†, and William W. Hargrove‡

†Oak Ridge National Laboratory, Computational Earth Sciences Group and
‡USDA Forest Service, Eastern Forest Environmental Threat Assessment Center

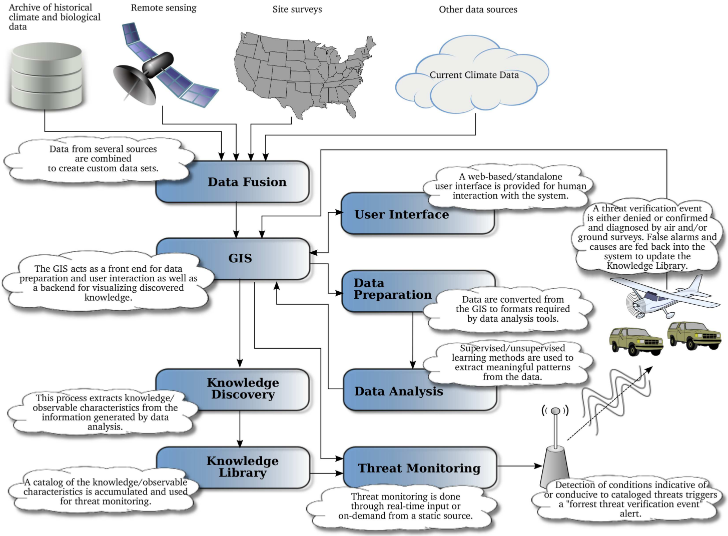
June 2, 2011
International Conference on Computational Science (ICCS 2011)
Nanyang Technological University, Singapore

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.

Overview of the Forest Incidence Recognition and State Tracking (FIRST) System



Normalized Difference Vegetation Index (NDVI)

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

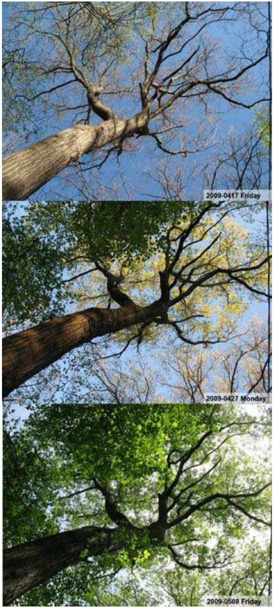
$$\mathrm{NDVI} = \frac{(\sigma_{\mathrm{nir}} - \sigma_{\mathrm{red}})}{(\sigma_{\mathrm{nir}} + \sigma_{\mathrm{red}})} \tag{1}$$

- 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.

MODIS MOD13 NDVI Product

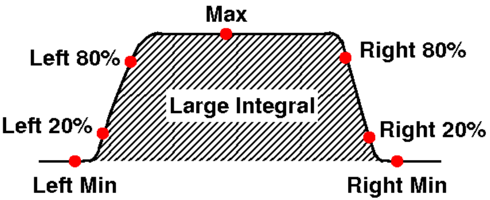
- The Moderate Resolution Imaging Spectroradiometer (MODIS) is a key instrument aboard the Terra (EOS AM, N$\rightarrow$S) and Aqua (EOS PM, S$\rightarrow$N) satellites.
- Both view the entire surface of Earth every 1 to 2 days, acquiring data in 36 spectral bands.
- The MOD 13 product provides Gridded 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.

- 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.



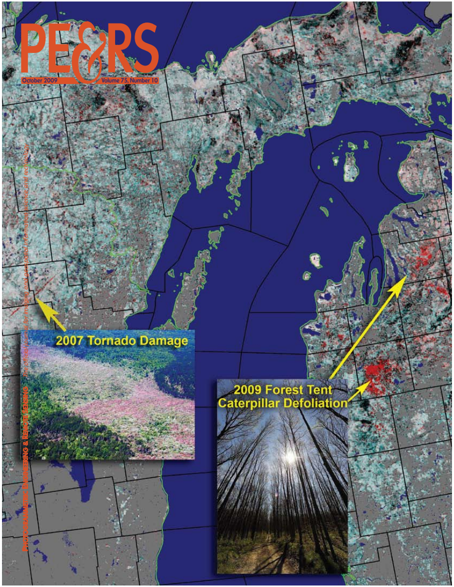
Up-looking photos of a scarlet oak showing the timing of leaf emergence in the spring (Hargrove et al., 2009).

Idealized Phenology Curve



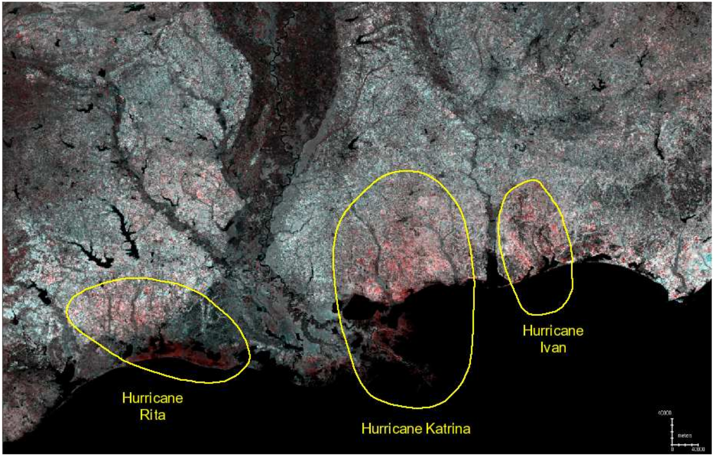
- An idealized seasonal NDVI curve is fit through data for each MODIS cell, and seven parameters are extracted.
- Each parameter results in two maps: one for the NDVI value and one for the time of the event.
- Cumulative NDVI shows the annual "greening" of the U.S.
- The Large Integral is strongly correlated 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.



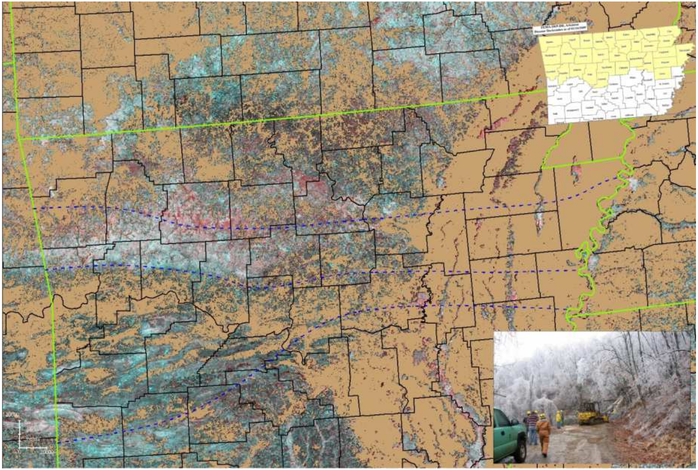
June 10–23, 2009, NDVI is loaded into blue and green; maximum NDVI from 2001–2006 is loaded into red (Hargrove et al., 2009).

Three Hurricanes



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.

Arkansas Ozarks Ice Storm, Jan. 26–29, 2009

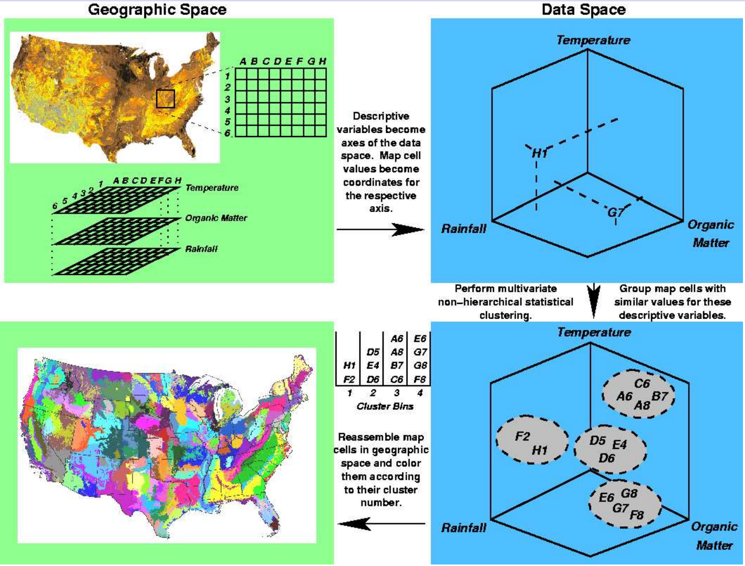


Computed by assigning 2009 max NDVI for June 10–July 15 into blue & green, and 2001–2006 max NDVI for June 10–July 27 into red. Storm resulted in 35,000 without power and 18 fatalities.

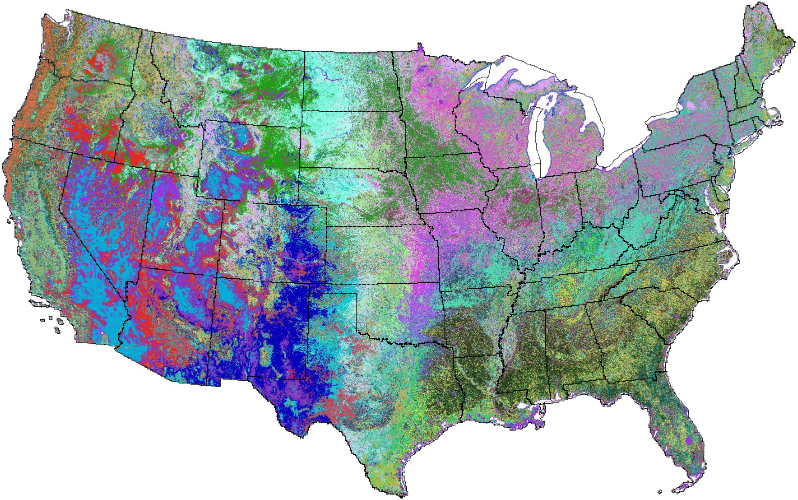
Data Mining for Change Detection

- 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.
- A data mining approach, utilizing high performance computing (HPC) for the entire history of the high resolution NDVI data, may provide a basis for determining "expected" or "normal" phenological variability.
- Hoffman and Hargrove previously employed a highly scalable *k*-means algorithm to automatically detect brine scars from hyperspectral remote sensing data (Hoffman, 2004) and for land surface phenology from monthly climatology and 17 years of 8 km NDVI from AVHRR (White et al., 2005).
- For only the current MODIS NDVI data for six years (2003–2008), 22 maps per year, at 250 m over the CONUS, single-precision data exceed 77 GB, requiring HPC resources.

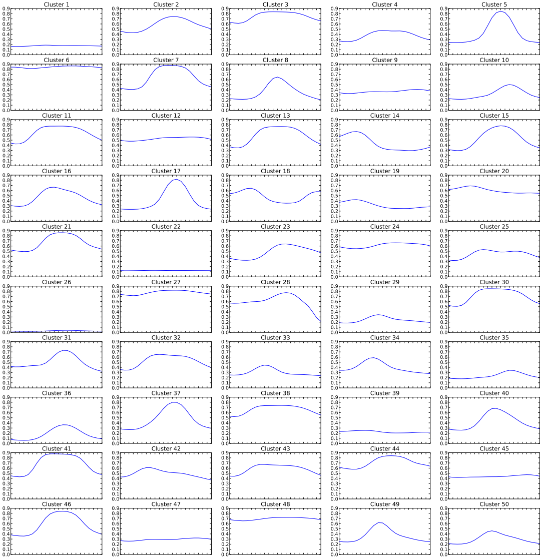
Geospatiotemporal Data Mining



50 Phenoregions for Year 2008 (Clustering 2003–2008)



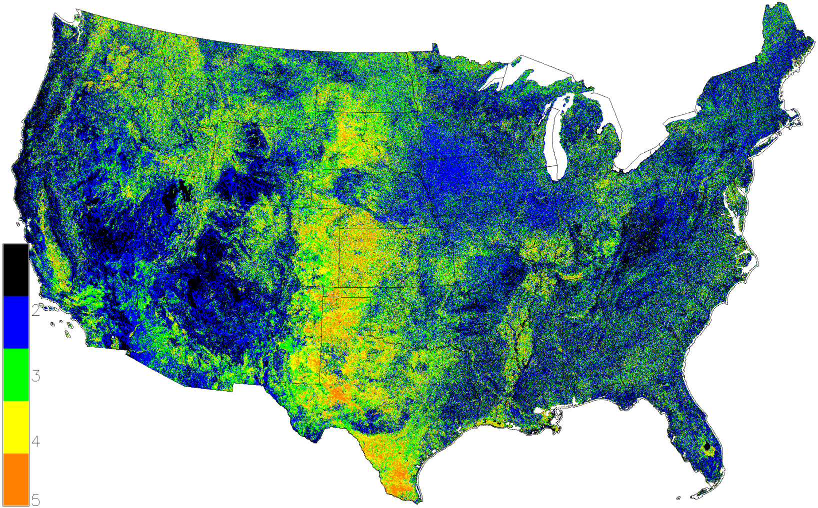
50 Phenoregion Prototypes



Clustering MODIS NDVI Data

- Cluster analysis yields six maps, one for each year, that classify each cell into one of the k phenoclasses. Here $k = 50$.
- The time evolution of phenoclass assignment, or phenostate, of each cell indicates a trajectory of change in the phenological behavior observed at that location due to natural or anthropogenic disturbance and ecosystem responses to interannual climate variability and long term climate trends.
- Comparison of the current phenostate with the nominal historical phenostate for each cell forms the basis for an early warning system for forest threats.
- Frequency of phenostate occupation for each map cell across all years provides insights into the phenological persistence or variability at every location in the CONUS.

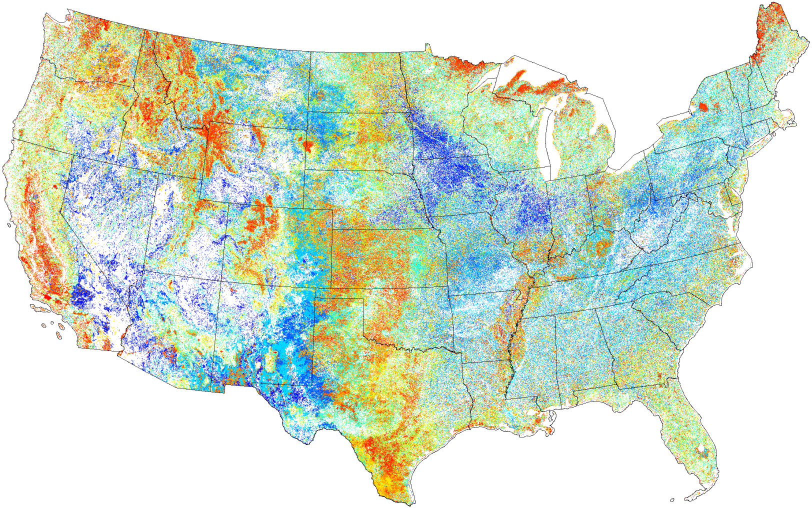
Cluster Persistence Map (2003–2008)



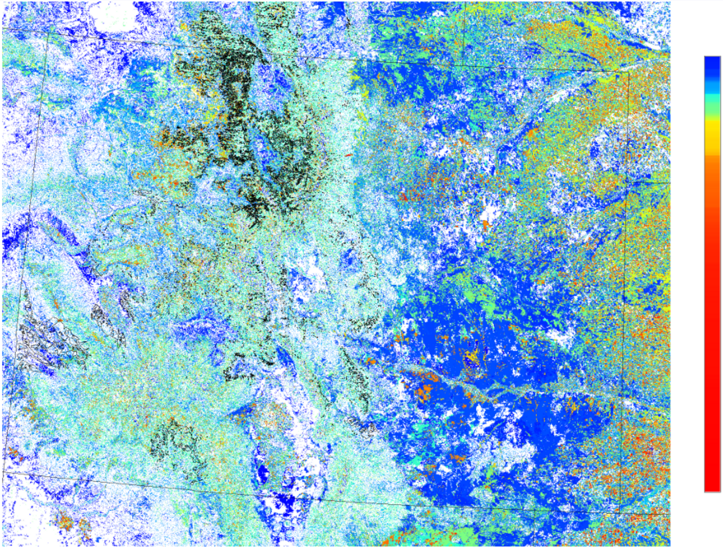
Euclidean Transition Distance as an Indicator of Change

- Cluster persistence is strongly dependent on the choice of k; if k is too large, normal interannual variability may result in a different phenostate assignment each year; if k is too small, important phenological change may be missed.
- A preferable alternative may be to use a larger value of k and to create maps of Euclidean transition distance between phenostate assignments.
- The transition distance between phenostates provides a relative measure of the strength of the observed change in phenological behavior between any two years.
- A large transition distance at any location indicates a significant change in the annual phenological cycle between the initial and final year.

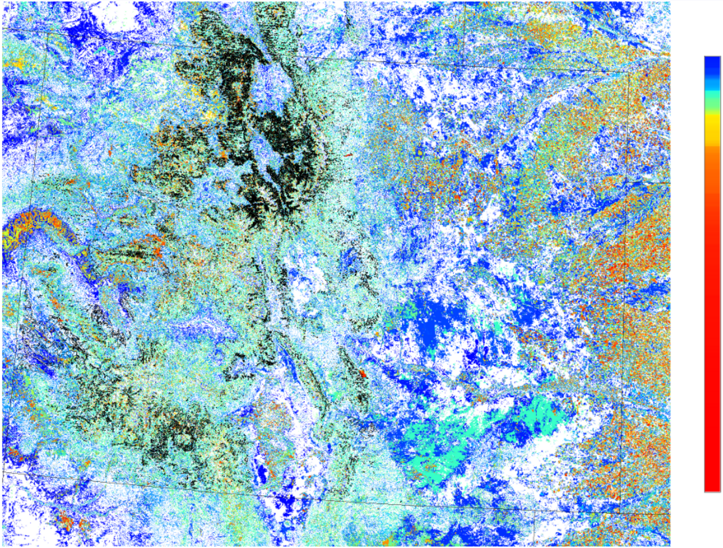
Cluster Transition Distances for (2008 – 2003)



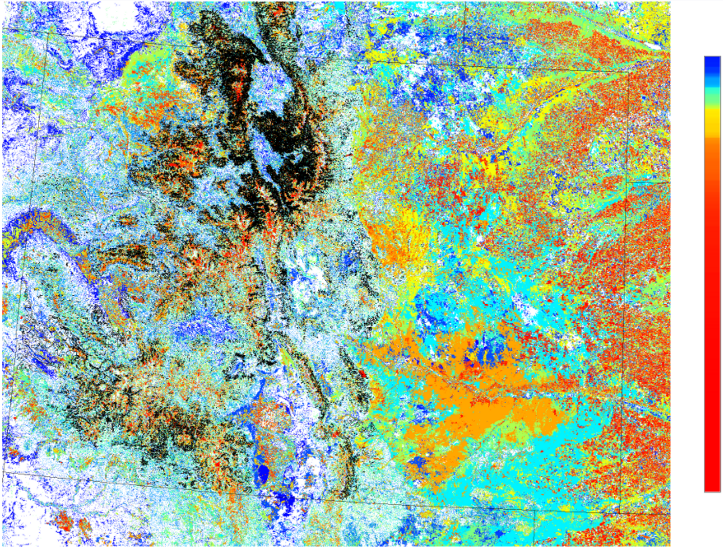
Mountain Pine Beetle in Colorado for (2004 – 2003)



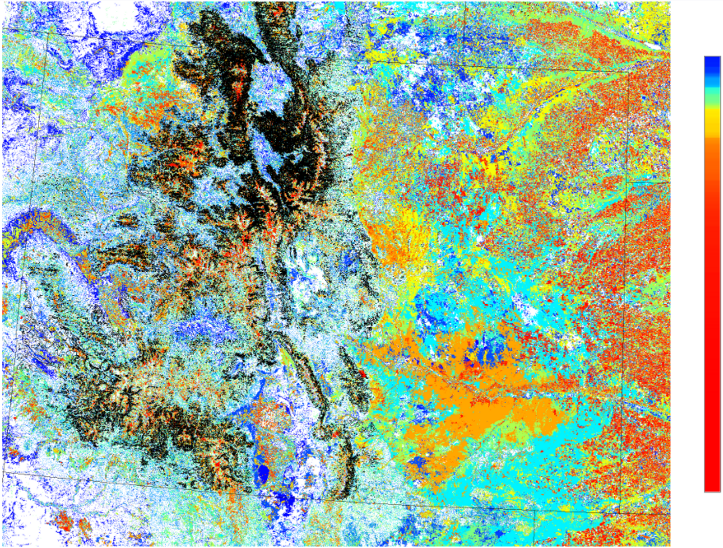
Mountain Pine Beetle in Colorado for (2005 – 2003)



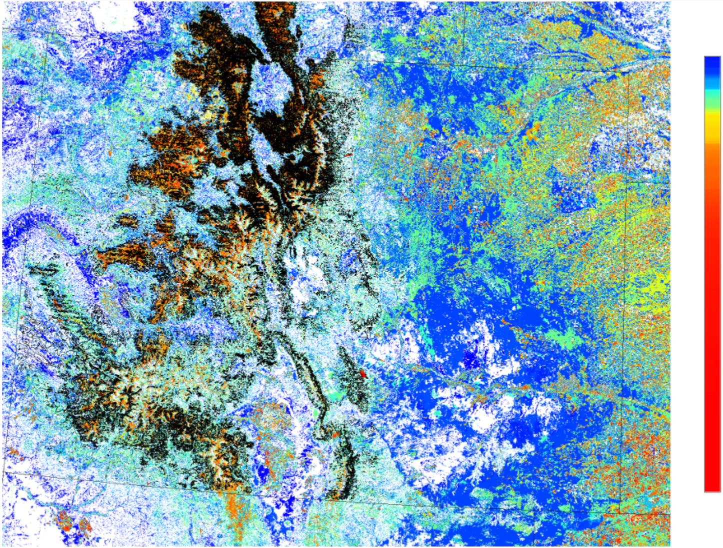
Mountain Pine Beetle in Colorado for (2006 – 2003)



Mountain Pine Beetle in Colorado for (2007 – 2003)



Mountain Pine Beetle in Colorado for (2008 – 2003)



Conclusions and Future Work

- Initial results of geospatiotemporal cluster analysis of phenology from MODIS NDVI are promising, suggesting such analysis could be a key component in an 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 trends, and a relatively short satellite record.
- However, significant 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.

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 MOD13 Collection 5 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.

References

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. *Photogramm. Eng. Rem. Sens.*, 75(10): 1150–1156, October 2009.

Forrest M. Hoffman. Analysis of reflected spectral signatures and detection of geophysical disturbance using hyperspectral imagery. Master's thesis, Department of Physics and Astronomy, University of Tennessee, Knoxville, November 2004.

Michael A. White, Forrest M. Hoffman, William W. Hargrove, and Ramakrishna R. Nemani. A global framework for monitoring phenological responses to climate change. *Geophys. Res. Lett.*, 32(4):L04705, February 2005. doi:10.1029/2004GL021961.