GEOSPATIO-TEMPORAL DATA MINING IN AN EARLY WARNING SYSTEM FOR FOREST THREATS IN THE UNITED STATES

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

Computer Science & Mathematics Division
Oak Ridge National Laboratory
1 Bethel Valley Rd.
Oak Ridge, TN 37831

W. W. Hargrove[§]

Southern Research Station
USDA Forest Service
200 WT Weaver Blvd.
Asheville, NC 28804-3454

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

The Forest Incident Recognition and State Tracking (FIRST) System is designed to be part of a two tier system for detecting and monitoring threats to forests and wildlands in the conterminous United States. When fully deployed, FIRST will serve as an early warning system that monitors continental-scale areas at a moderate resolution using remote sensing data. Employing data mining techniques for feature extraction and change detection, FIRST will detect disturbances and changes in forest health, and will produce alerts or warnings for locations of interest that will help direct the second tier of the larger system. Tier 2 consists of higher resolution monitoring and attribution of forest and wildland disturbances through airborne overflights—called Aerial Detection Survey (ADS) sketch-mapping—and ground-based inspections of areas of interest identified by FIRST. These surveys will be used to determine if a warning issued by FIRST should be elevated to an alarm. Tier 2 is largely in operation already. However, by directing the surveys in Tier 2, the FIRST System in Tier 1 will improve the efficiency and utility of these costly and labor-intensive surveys.

The goals of the FIRST early warning system are to provide a single, unified system for change detection from remotely sensed vegetation properties over the domain of the conterminous United States at about 500 m resolution—nominally from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors on board the 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 datasets using moderate- to large-sized clusters and supercomputers.

A schematic overview of the FIRST system is shown in Figure 1.

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

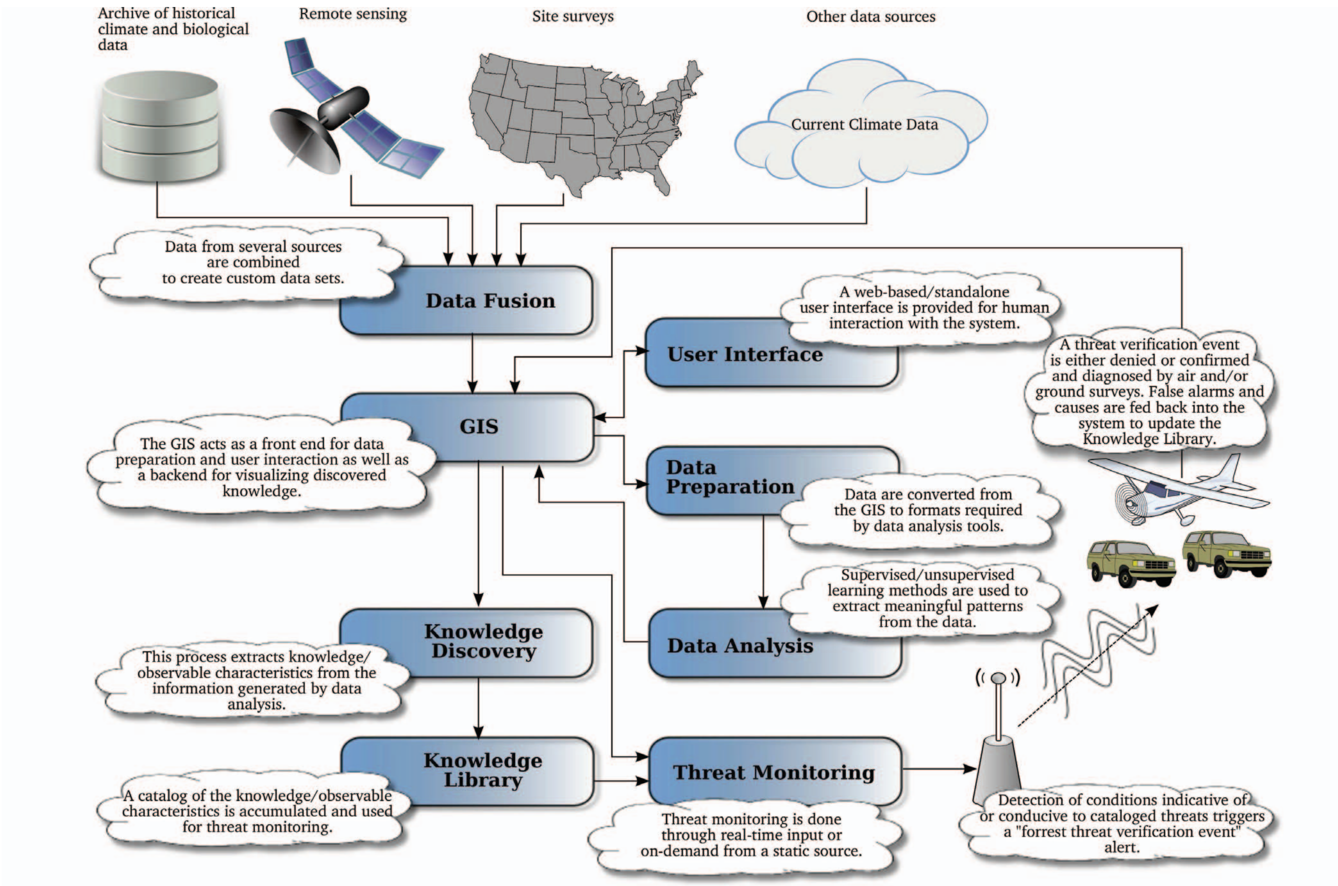

Fig. 1. An overview of the Forest Incident Recognition and State Tracking (FIRST) System.

2. GEOSPATIO-TEMPORAL 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 states that are useful for describing and tracking the behavior of a forested ecosystem through time within a multi-dimensional phase or state space [1, 2]. States derived in this fashion are exhaustive, since they span all occupied portions of the state space, and exclusive, since the forest at any given location and time can occupy exactly one such state. The sequential trajectory of states at any particular location describes and tracks the ecological behavior of that forested ecosystem through time. The statistical procedure can be used to produce historical ecological states narrowly or broadly defined by dividing the occupied state space into many or few states respectively. As with all fuzzy statistical methods, new/current conditions are more likely to match their historical states when the space is divided coarsely into a few broad groups than when finely divided into a large number of groups. By defining multiple sets of states, ranging from narrowly to broadly defined, a more continuous measurement of signal strength will be provided for detecting and assessing threats for forest health. Alarms may then be tuned to a threat-dependent sensitivity.

Key to identifying forest threats will be the inclusion of remotely sensed land surface phenology along with data about soils and climate. Deviations from "normal" phenological temporal development is often the first indication of changes in forest health, including disturbance and recovery. Seasonal changes in Normalized Difference Vegetation Index (NDVI), a self-standardizing measure of land surface greenness derived from red and near infrared wavelengths, has proven useful in tracking land surface phenology and identifying changes in vegetation due to drought, insect or pathogen invasion, storm damage, and forest regrowth. Hargrove and Hoffman used the aforementioned statistical clustering procedure running on a supercomputer to delineate 500 global phenological ecoregions, or phenoregions, based on similarities in monthly climate and NDVI data from 17 years of 8 km AVHRR images [3]. By looking for minimum potential human impacts, a subset of global phenoregions was identified, within which any observed phenological changes might be most directly attributable to shifts in climate. Recently, an initial cluster analysis of cumulative phenology derived from five years of MODIS NDVI has produced phenoregions strongly correlated with ecosystem productivity. Because such large unsupervised classifications through time result in an historical trajectory couched in common terms for comparison, they should be useful for defining the bounds of "normal" behavior at every location in the conterminous U.S. Provided with such a history for comparison, a national early warning system will identify locations that appear to deviate from their usual behavior [4].

3. REFERENCES

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

[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] 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.

[4] 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.