Scalable Machine-Learning Approaches for Analysis of Large Phenological Datasets

Richard Tran Mills

April 4, 2016

Notice and Disclaimers

Optimization notice

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Acknowledgements

The material presented here is the result of collaborative work with several people:

- Forrest M. Hoffman, ORNL
- Jitendra Kumar, ORNL
- William W. Hargrove, USDA Forest Service
- Steven P. Norman, USDA Forest Service
- Joseph P. Spruce, formerly NASA Stennis Space Center

Introduction/Outline

- Scalable approaches to two kinds of unsupervised machine learning:
  - Clustering: Accelerated k-means
  - Matrix decomposition: SVD/PCA
- These techniques have found extensive application in analysis of MODIS NDVI-based vegetation phenology products of the ForWarn project (`http://forwarn.forestthreats.org/`).
- Other talks in this session will demonstrate application of k-means to vegetation phenology.
- We'll look at using PCA to visualize trends and detect anomalies in vegetation phenology from satellite NDVI data.

Data-mining with ForWarn NDVI phenology products

- For each year and each grid cell in the CONUS, construct an observation vector of 46 NDVI values representing the seasonal NDVI trace for that year/location.
- All observation vectors are combined into a data matrix with 46 columns and hundreds of millions of rows (each year corresponds to 146.4 million rows).
- Data are standardized and then clustered via k-means.
- Cluster assignments are mapped back to each map cell and year from which each observation came, yielding one map per year in which each cell is classified into one of k clusters or "phenoclasses".
- These can be viewed as forming a dictionary of prototypical annual NDVI traces.
- Later in the talk, we explore how PCA/SVD approaches can be used to look for trends and anomalies.

Cluster centroids/annual NDVI curve "prototypes"

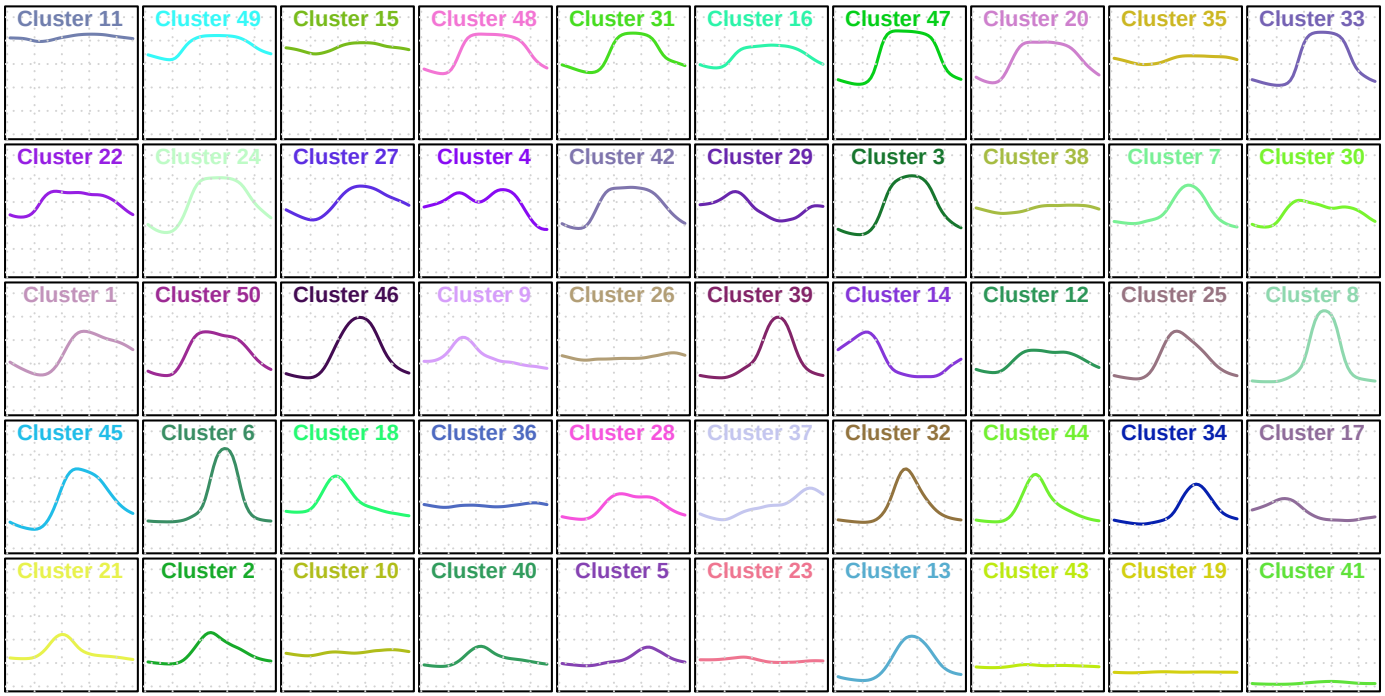
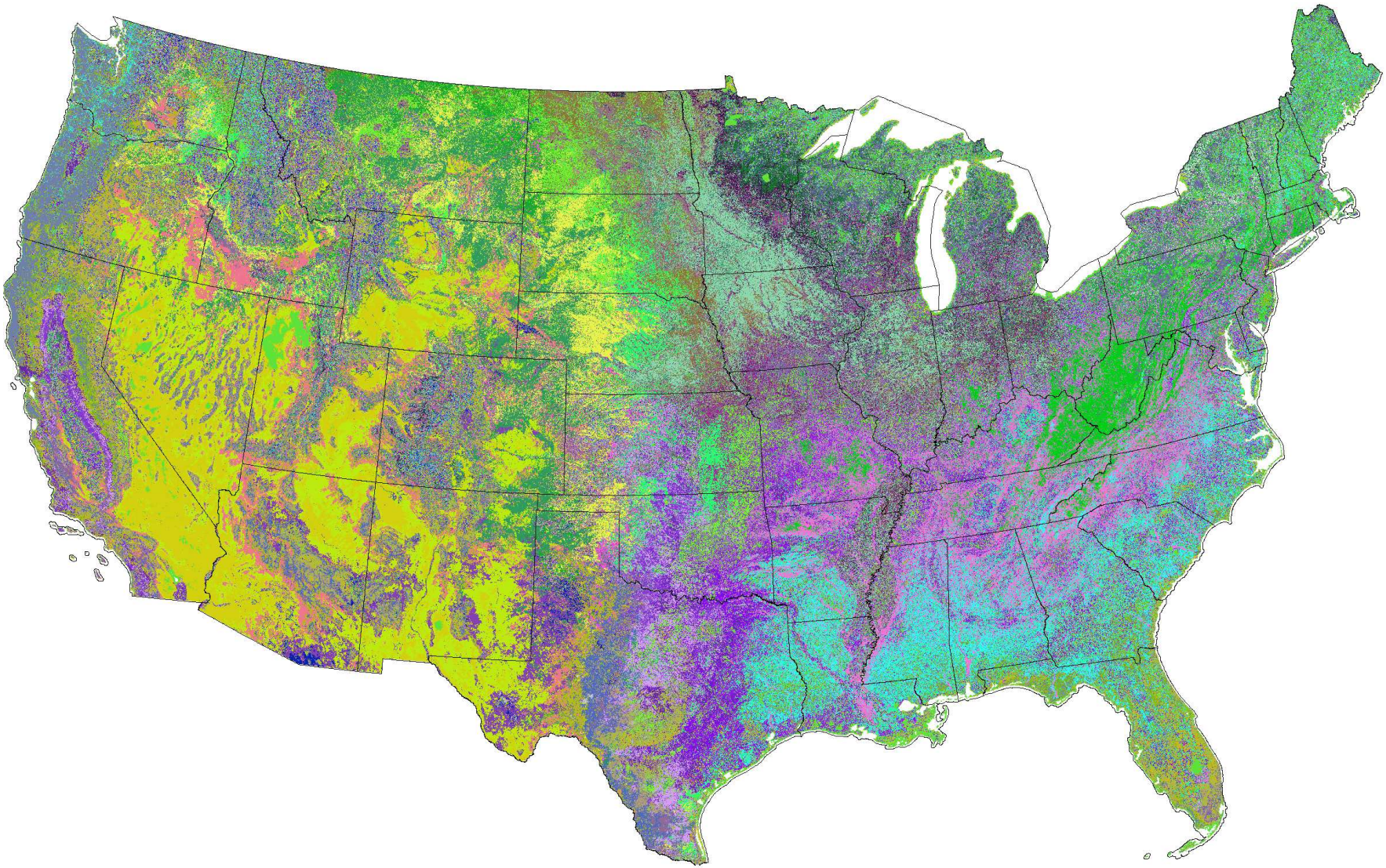


Figure: Fifty centroids (corresponding to "phenoregion" prototypes) from a $k = 50$ clustering. The colors of the centroid plot correspond to the map colors on the next slide.

Phenoregions map, $k = 50$

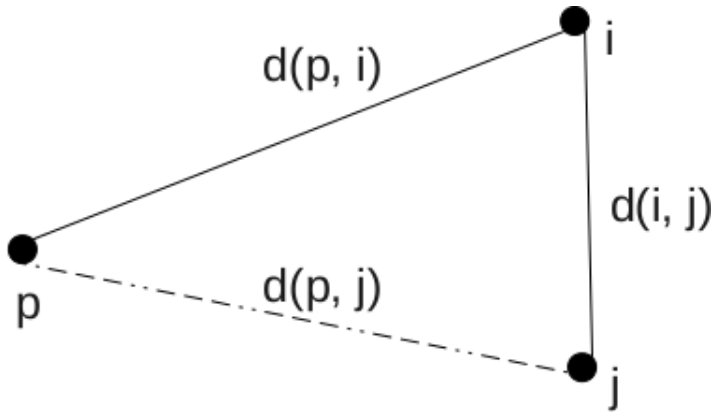


Parallel k-means clustering

- We have two implementations of accelerated k-means clustering, following two parallel programming models
  - A master-worker (MW) model: Central master assigns "aliquots" of work to workers. Facilitates dynamic load balancing but has memory and performance scalability limits due to single, central process.
  - Fully distributed (FD): All processes use static distribution of work. Very scalable, but no dynamic load balancing.
- We improve cluster quality by moving or "warping" clusters that become empty to locations in data space where points that are farthest from their current cluster centroids reside.
- We "accelerate" the k-means process using two techniques described by Phillips (doi:10.1109/IGARSS.2002.1026202):

Accelerated k-means clustering

- Use the triangle inequality to eliminate unnecessary point-to-centroid distance computations based on the previous cluster assignments and the new inter-centroid distances.
- Reduce evaluation overhead by sorting inter-centroid distances so that new candidate centroids c_j are evaluated in order of their distance from the former centroid c_i. Once the critical distance $2d(p, c_i)$ is surpassed, no additional evaluations are needed, as the nearest centroid is known from a previous evaluation.

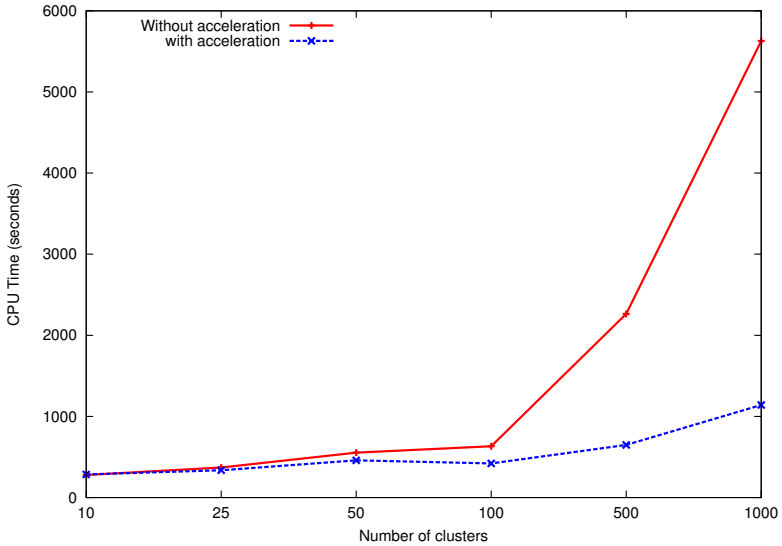


$$d(i,j) \leq d(p,i) + d(p,j)$$
$$d(i,j) - d(p,i) \leq d(p,j)$$
if $d(i,j) \geq 2d(p,i)$:
$$d(p,j) \geq d(p,i)$$
without calculating the distance $d(p,j)$

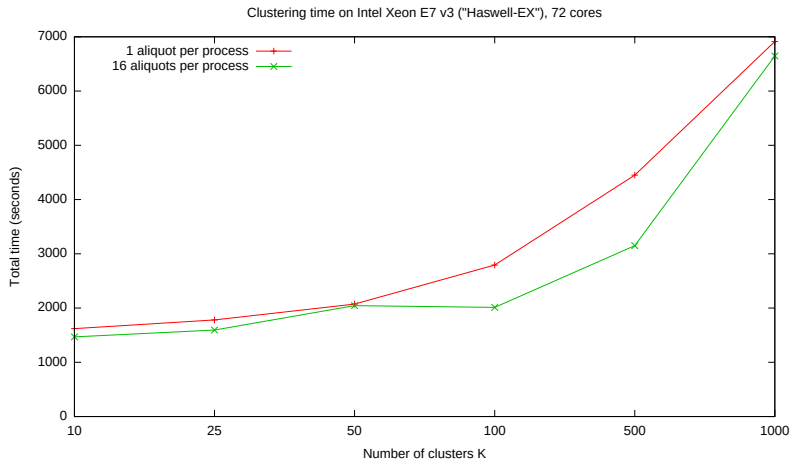
Parallel Performance with Accelerated k-means

- In 2011, we would use $\sim$1024 AMD Opteron cores on a machine like Jaguar, the Cray XT5 at ORNL, for our analyses.
- In 2016, we can do larger analyses on a single compute node of Intel's Endeavor cluster with Intel® Xeon® E7-8890 v3 ("Haswell-EX") processors.
  - AVX2 instruction set: 256-bit (8 single precision `floats`) vector registers with dual-issue fused multiply-add
  - Four 18 core (36 thread) CPUs; over 500 GB DRAM

Parallel Performance with Accelerated k-means



(a) Jaguar



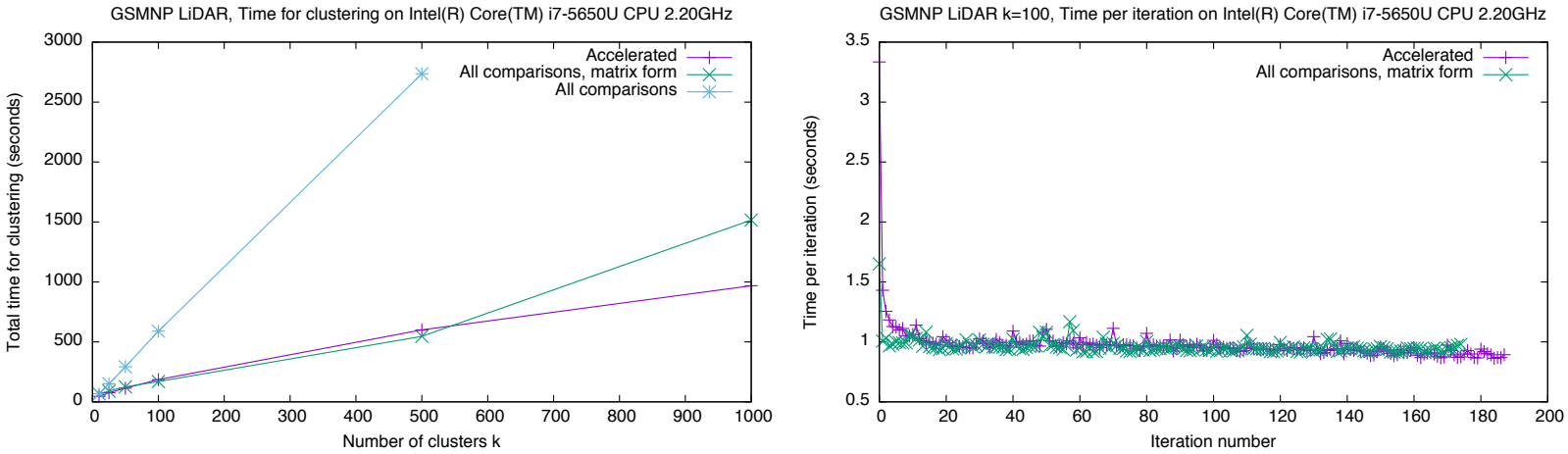
(b) Endeavor

Figure: Times to cluster different versions of the 2000–2009 ForWarn phenology data set on (a) 1024 cores of the Jaguar Cray XT5, ca. 2011 at ORNL and (b) a single 72-core "Haswell-EX" node on Intel's Endeavor cluster. The data set used on Jaguar is the 16 day product, while the one on Endeavor is the 8 day product and is therefore twice as large (251 GB in single precision). Figure (b) illustrates the benefit of using smaller aliquots to enable dynamic load balancing in the master-worker clustering code.

Improving computational intensity

- It is possible to achieve greater computational intensity of the observation–centroid distance calculations by expressing the calculation in matrix form:
  - For observation vector x_i and centroid vector z_j, the squared distance between them is $D_{ij} = \|x_i - z_j\|^2$.
  - Via binomial expansion, $D_{ij} = \|x_i\|^2 + \|z_j\|^2 - 2x_i \cdot z_j$
  - The matrix of squared distances can thus be expressed as $D = \overline{x}\mathbf{1}^\intercal + \mathbf{1}\overline{z}^\intercal - 2X^\intercal Z$, where X and Z are matrices of observations and centroids, respectively, stored in columns, $\overline{x}$ and $\overline{z}$ are vectors of the sum of squares of the columns of X and Z, and $\mathbf{1}$ is a vector of all 1s.
- The above expression for D can be calculated in terms of a level-3 BLAS operation (xGEMM), followed by two rank-one updates (xGER, a level-2 operation).
- Level 2 and 3 BLAS operations admit very computationally efficient implementations, and libraries such as Intel® MKL provide highly optimized versions.

Performance with matrix-form distance calculation



(a) (b)

Figure: Timings for clustering as GSMNP LiDAR dataset using a single worker process on an Intel® Core™ i7-5650U CPU operating at 2.20GHz. (a) Total timings for k-means clustering using the acceleration techniques; doing all distance comparisons but forming the distance matrix using BLAS operations provided by Intel® MKL; and doing all distance comparisons without the benefit of the matrix formulation and BLAS. (b) Timings per iteration for k=100 when using the acceleration technique compared to the matrix formulation for the distance calculations. In early iterations, where many distance comparisons are required, the matrix formulation offers better performance.

Combining the best of both methods

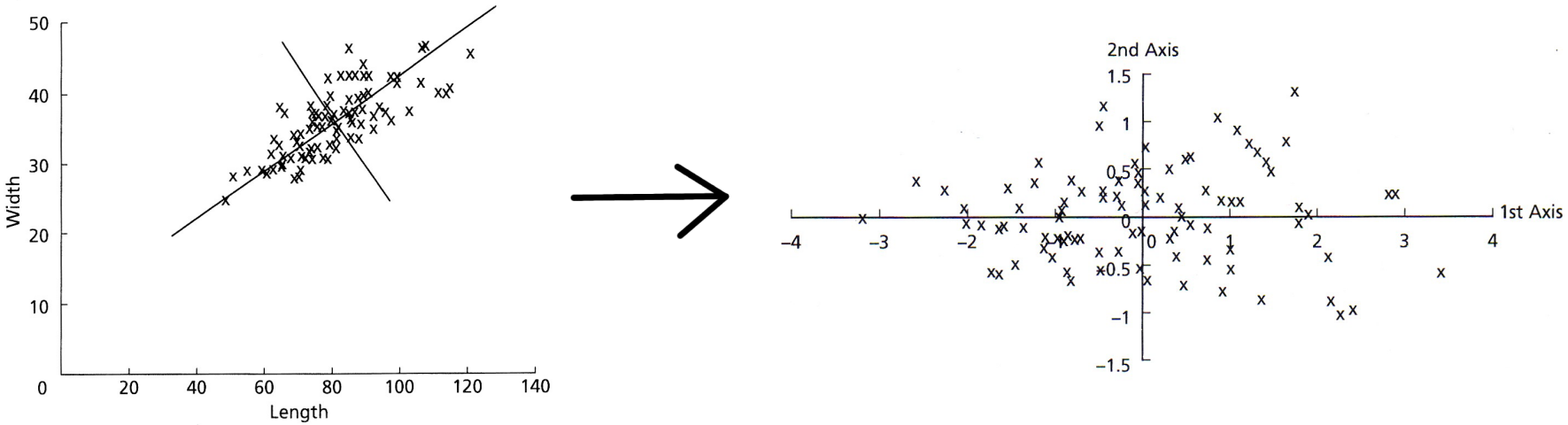
- Using the matrix formulation for the distance calculations is dramatically faster than the straightforward loop over vector distance calculations when many distance comparisons must be made.
- The "acceleration" technique greatly speeds clustering because it reduces the number of distance comparisons that can be made.
- The best performance should be achieved by combining both approaches.
- Straightforward: Use the matrix formulation for distance comparisons in early k-means iterations (when all or most comparisons are necessary), then switch to the "accelerated" approach.
- More complicated: Use the matrix formulation for distance calculations inside accelerated algorithm. Requires online data rearrangement so that only a subset of observation–centroid distances are calculated.

PCA/SVD approaches for forest threat detection

- Our clustering-based approaches can flag a wide range of disturbances, particularly those involving high mortality events such as fire, storms, or mountain pine beetle outbreaks.
- Slower-acting agents, such as hemlock woolly adelgid, that cause a gradual decline in forest health are more difficult.
- Also, the annual phenology of some areas is highly influenced by interannual climate variability: grasslands, for instance, experience rapid greenup after precipitation and do not have smooth annual cycles.
- These areas tend to display a large transition distance from year to year even when there is essentially no real change in the vegetation health.
- To remedy these shortcomings, we have been exploring the use of principal components analysis (PCA) (or the related SVD) as a complementary approach.

A complementary approach: Principal component analysis

Principal Components Analysis (PCA) determines, for a p-dimensional data set, an orthogonal set of p new axes (linear combinations of the original p variables) such that the first axis explains the greatest variance, the second explains the next most variance, and so on.



- Commonly used to determine dominant patterns in data

Varimax-rotated loadings for top 3 components

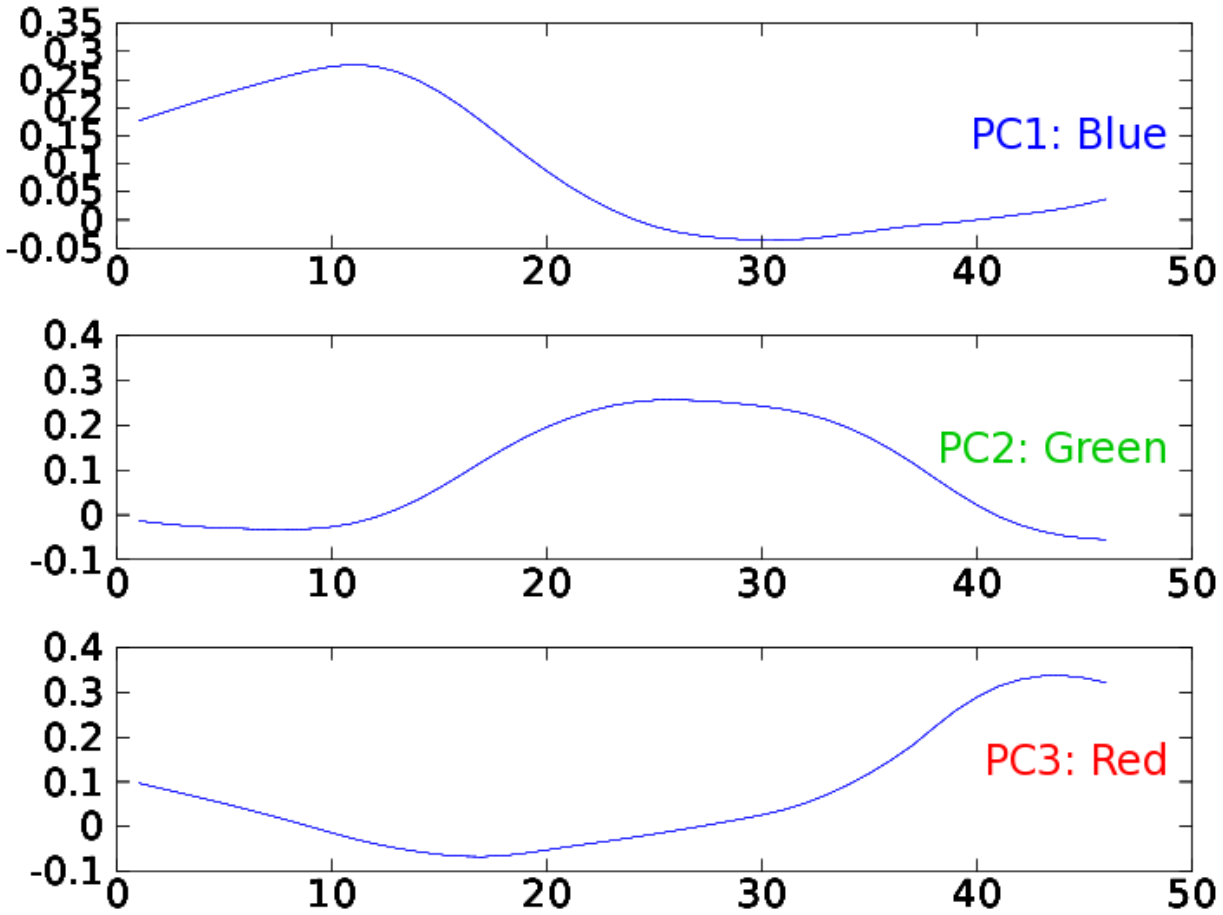
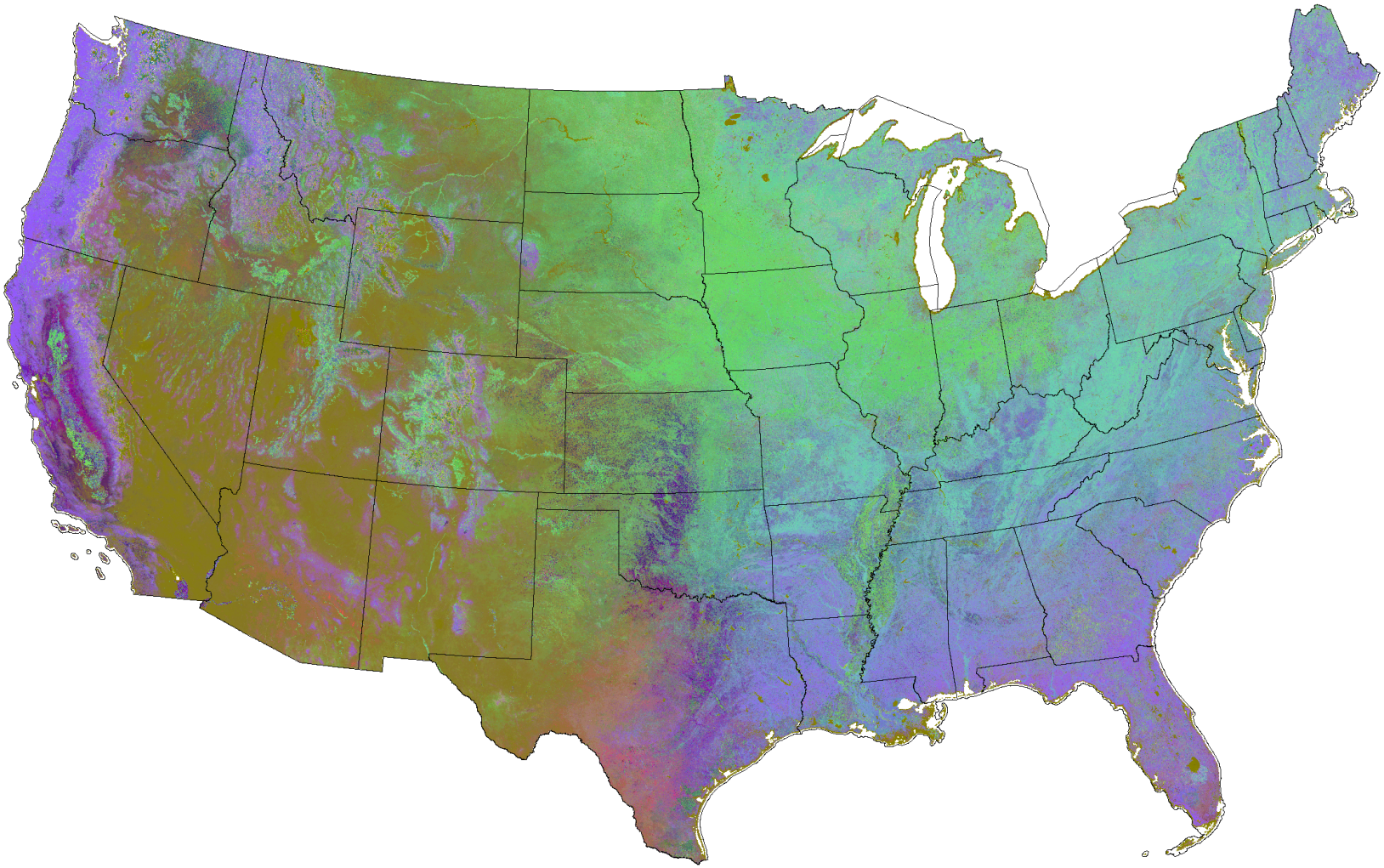


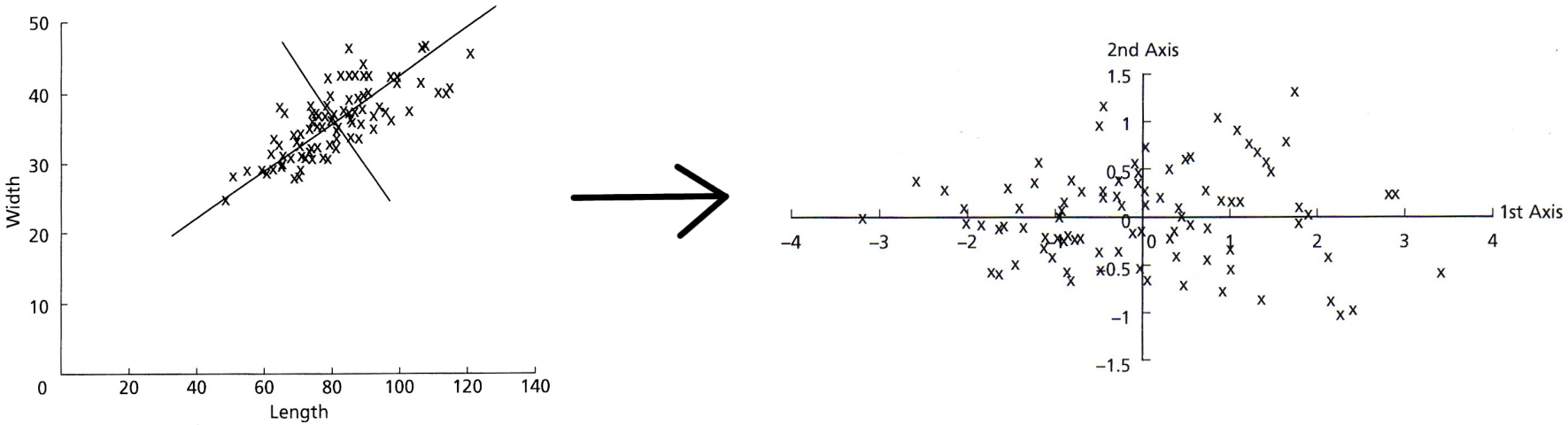
Figure: The loadings (coefficients in the linear combination of the 46 original variables) along the three varimax-rotated principal axes. The x-axis corresponds to the eight-day NDVI-acquisition windows and loadings are

$k = 1000$ map for year 2000, similarity colored



A complementary approach: Principal component analysis

Principal Components Analysis (PCA) determines, for a p-dimensional data set, an orthogonal set of p new axes (linear combinations of the original p variables) such that the first axis explains the greatest variance, the second explains the next most variance, and so on.



- Commonly used to determine dominant patterns in data
- But can also be used to determine the anomalous patterns: Observations that score strongly on low order components do not

Parallel Principal Components Analysis Tool

- We have developed a prototype parallel tool to perform PCA.
- Rather than explicitly forming the covariance matrix, computes thin SVD of the adjusted data matrix.
- Uses the Lawson-Hanson-Chan factorization to exploit the "tall and skinny" $(m >> n)$ nature of our matrices: $(m >> n)$
  - Form reduced factorization $\mathbf{A} = \mathbf{QR}$ (via parallel PLAPACK routine)
  - Gather the matrix $\mathbf{R}$ to process 0.
  - Process 0 calls LAPACK DGESVD to compute the SVD $\mathbf{R} = \mathbf{USV}^T$.
  - Optionally, back transform $\mathbf{Q}$ to get $\mathbf{Q} \leftarrow \mathbf{QU}$.
  - Final SVD is: $\mathbf{A} = \mathbf{QSV}^T$
- A serial bottleneck exists where the SVD of $\mathbf{R}$ is computed, but this matrix is so small (only 46×46 for our NDVI data set) that this serial portion is essentially negligible.

Detecting anomalous observations with PCA

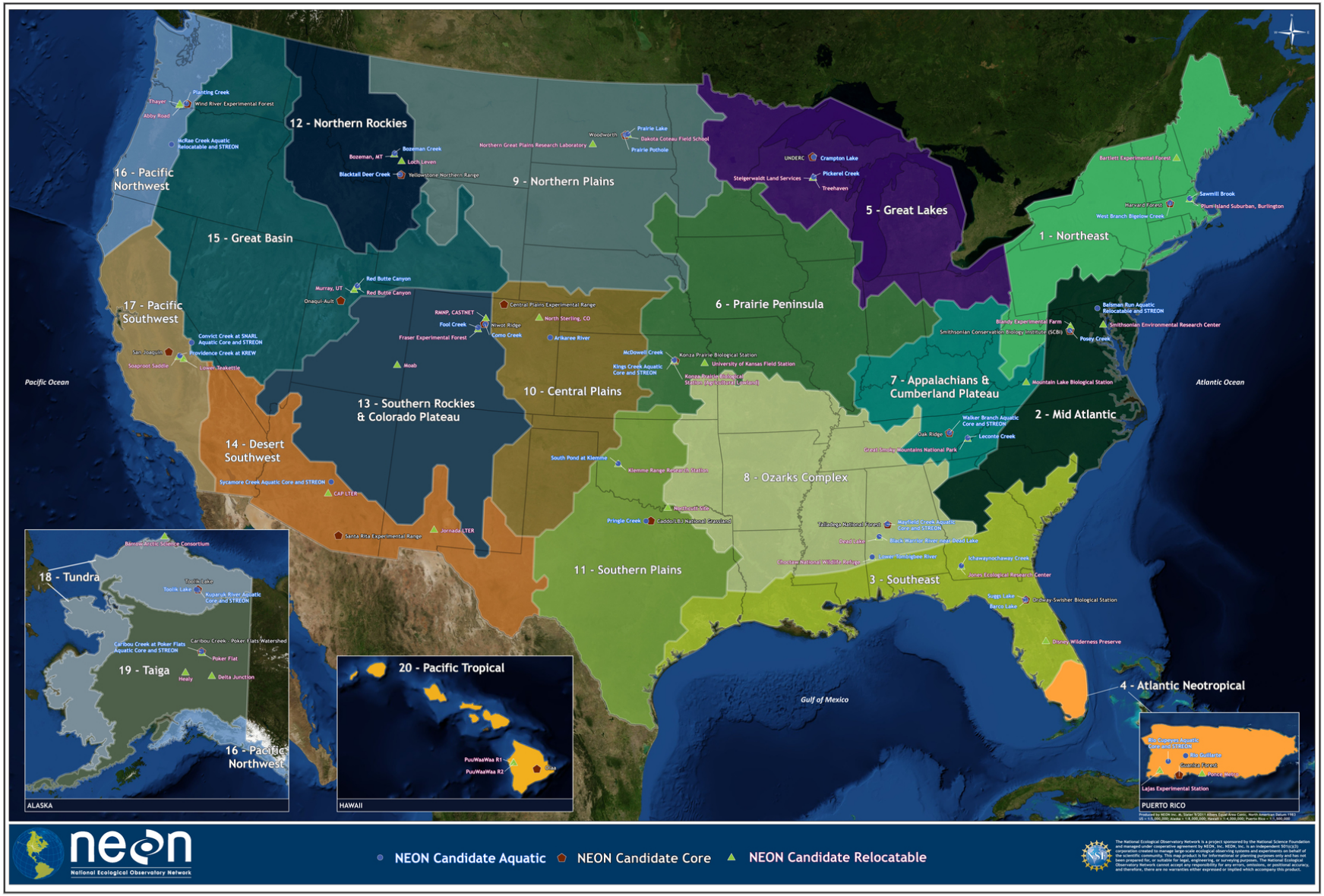
- Can identify anomalies two complementary ways:
- Look at sum of scores onto r lowest-order components:

  $\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$ greater than some outlier threshold
- Look at squared prediction error: How well an observation can be represented in subspace of q highest order components?
  - Idea: decompose into modeled and residual parts: $x = \hat{x} + \tilde{x}$
  - $P = \begin{bmatrix} v_1 & v_2 & \dots & v_q \end{bmatrix}$
  - $\hat{x} = PP^T x = Cx$ and $\tilde{x} = (I - PP^T)x = \tilde{C}x$
  - Abnormal if SPE $= \|\tilde{x}\|^2 = \left\|\tilde{C}x\right\|^2$ exceeds threshold
- Can also do cross-comparison: Construct subspace from one data set, then see how well observations from another can be represented in that space.

Detecting anomalies within single year, single domain

- These approaches will flag any observations that are somehow "unusual" *for the collection of data from which the principal components have been calculated*.
- Some judgement required: choice of NDVI data subset used in the PCA calculation will affect what constitutes a "normal" or "abnormal" observation.
- E.g., Extremely low NDVI may appear normal when using PCA based on national dataset due to presence of areas like the Mohave; appears anomalous when using PCA based only on humid Southeast.
- Here we use PCAs computed over single years and within a spatial domain conforming to the eco-climatic domains established by the National Ecological Observatory Network.

NEON Domains



Detecting anomalies within single year, single domain

- In all examples, PC vectors 10–46 are used as the basis for the "abnormal" space, which explains 5–10% of the variance.
- In all of examples, certain features that are not disturbances but possess very anomalous NDVI traces (e.g., bodies of water) show up very strongly.

Colorado and Southern Wyoming, 2008

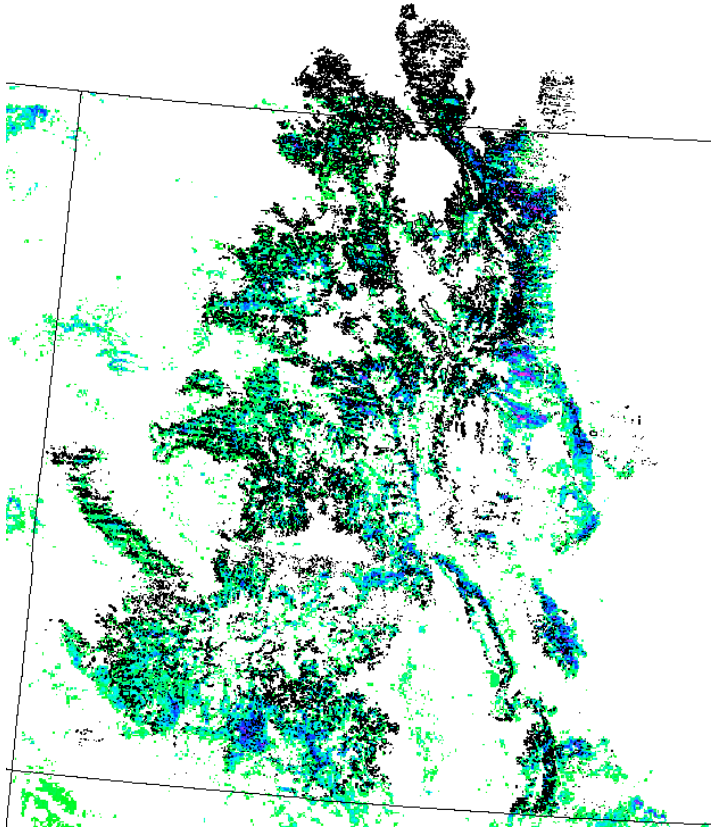


Figure: Portion of the Southern Rockies–Colorado Plateau NEON Domain for year 2008, showing map cells scoring in the 85th percentile. Black polygons show damaged areas noted in aerial detection surveys; extensive damage due to mountain pine beetle and sudden aspen decline are evident.

Vicinity of Louisiana Coast: Hurricane-induced disturbance

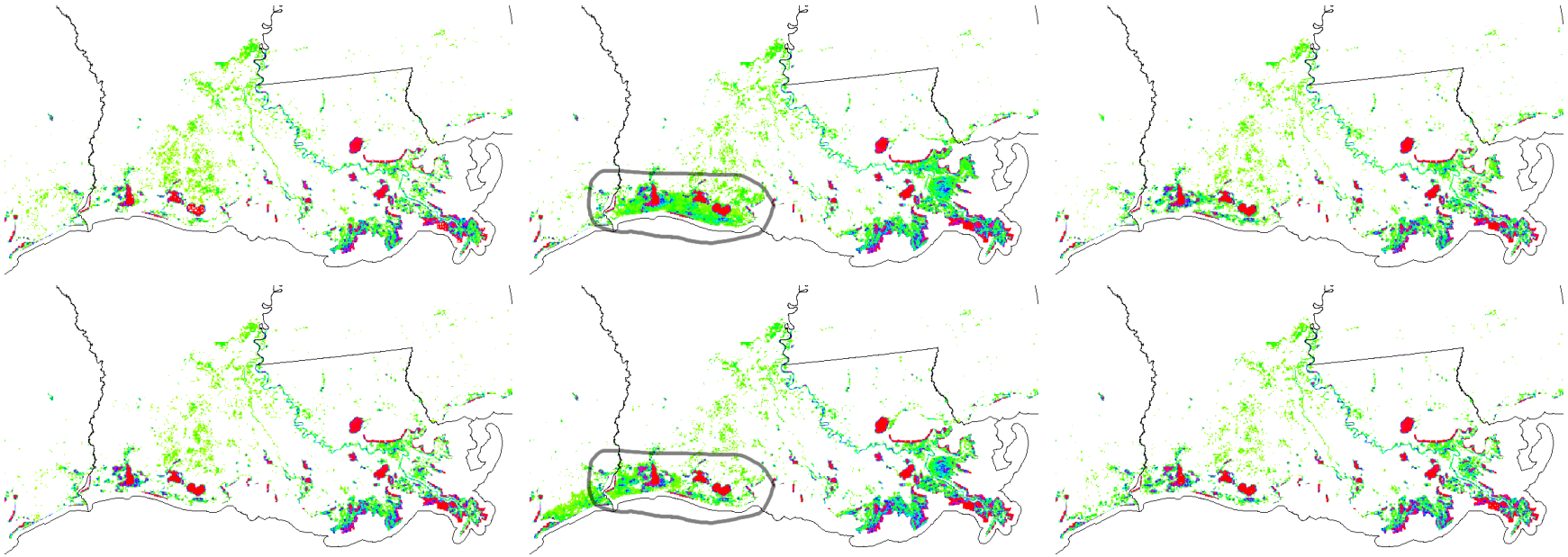


Figure: Portions of the PCA-based anomaly maps (map cells scoring in the 90th percentile are shown) for the Southeast NEON Domain for years 2004–2009, showing the area in the vicinity of the Louisiana coast. From left to right, the top row shows years 2004, 2005, and 2006, respectively, and the bottom row years 2007, 2008, and 2009. The affected regions are circled in the 2005 and 2008 maps. The prominent red features are water bodies.

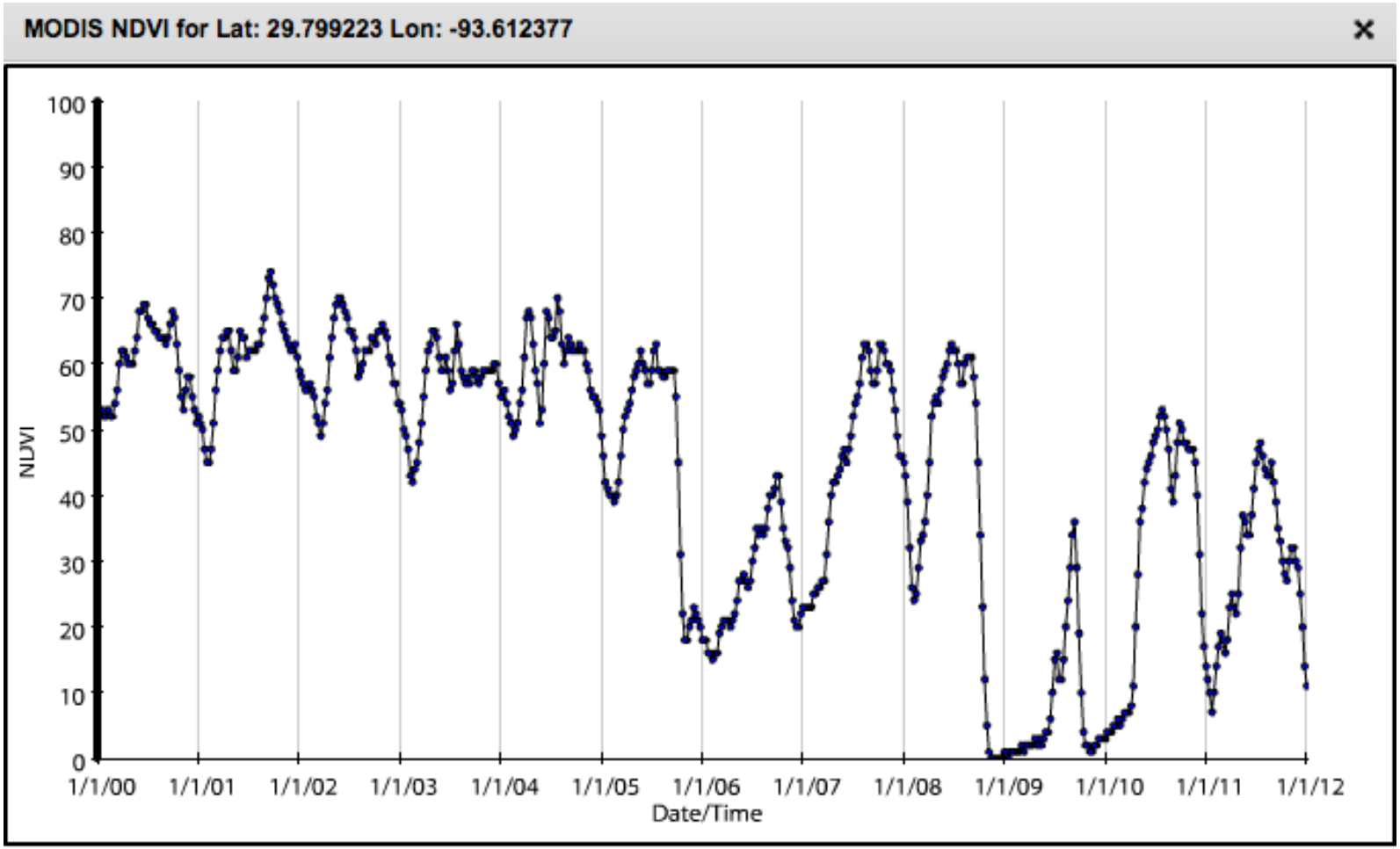


Figure: NDVI trajectory as viewed via the Forest Change Assessment Viewer for a location (close to the center of the circled region in the previous figure) near the coast in southwestern Louisiana showing apparent hurricane-induced mortality from events in 2005 and 2008.

Southern Appalachians: Hemlock decline

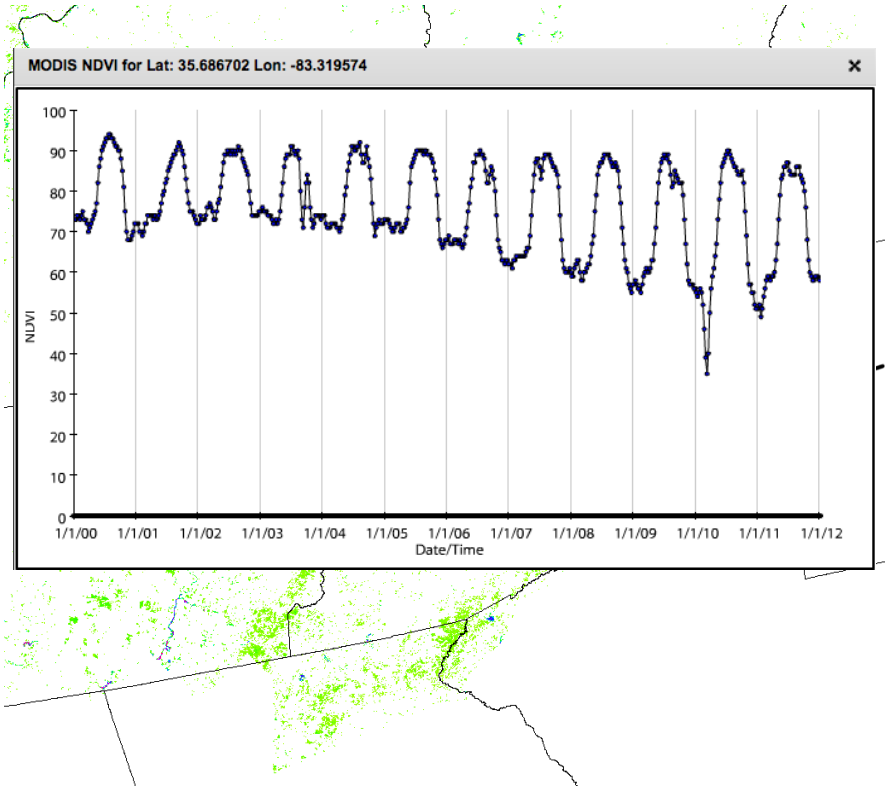


Figure: At left, a portion of the PCA-based anomaly map (map cells scoring in the 90th percentile are shown) for the Southern Appalachians/Cumberland Plateau NEON Domain for year 2010. The arrow indicates a location thought to be affected by hemlock woolly adelgid, and the corresponding NDVI trajectory is shown at right.

Conclusions

- We have developed highly scalable tools for k-means and PCA/SVD of geospatial data sets.
- Analyses that required O(1000) compute nodes can now be done on a handful of (or a single) modern compute node.
- Technologies such as NVRAM will soon lead to several-fold increase in the amount of data that can be processed in memory on a single node.
- Combined with the dramatic increase of parallelism within a node, new possibilities for analyses fusing several types of observational data sets, at unprecedented resolutions, will emerge. For example:
  - Combined analysis of all the MODIS vegetative phenology record with global fine-scale meteorological reanalysis (and possibly other ancillary data layers) to enable attribution of vegetation changes to climate or other events.
  - Use all archived IPCC simulations to project changes to distribution of eco-phenoregions (identified by the historical analysis) for different climate change scenarios.
- My message to the community: Think big!

