Parallel *k*-Means Clustering for Quantitative Ecoregion Delineation Using Large Data Sets

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

[†]Computer Science and Mathematics Division, Oak Ridge National Laboratory
[‡]Eastern Forest Environmental Threat Assessment Center, USDA Forest Service

Thursday June 02, 2011
International Conference on Computational Science (ICCS) 2011, Singapore

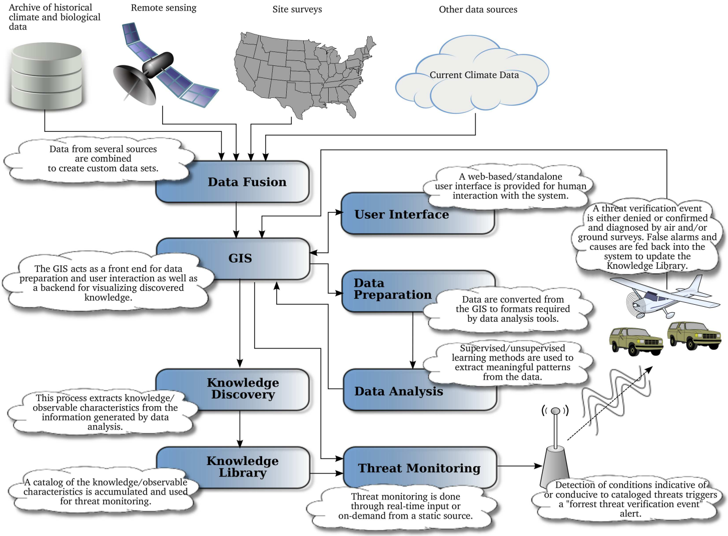
Outline

- Introduction: Delineation of ecoregions
- Computational challenges: Spatio-temporal scales of data and data set size
- Design: Parallel k-means algorithm and enhancements
- Performance: Parallel performance and scaling
- Application: Forest threat detection using MODIS NDVI products

Introduction

- Ecoregions are geographical regions of generally similar combination of ecologically relevant conditions like temperature, precipitation and soil characteristics.
- Understanding and delineation of ecoregions are useful for predicting suitable species range, stratification of ecological samples, and to help prioritize habitat preservation and remediation efforts.
- In the case of threatened or endangered species, a well-executed ecoregion classification can be used to identify and locate the extent of suitable habitat for the purposes of preserving or improving it.
- Large amount of data sets are available from satelite, airborne and ground based remote sensing; GCM model outputs
- Data mining tools can be used to extract knowledge from these data sets

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.

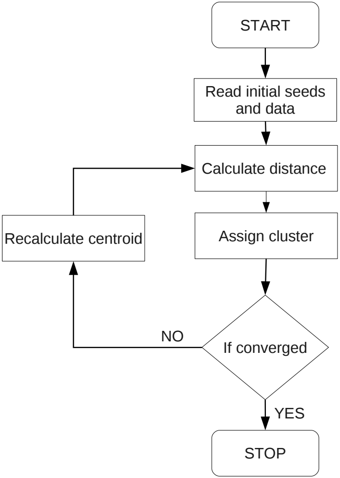
$$\mathrm{NDVI} = \frac{(\sigma_{\mathrm{nir}} - \sigma_{\mathrm{red}})}{(\sigma_{\mathrm{nir}} + \sigma_{\mathrm{red}})} \tag{1}$$

- 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.

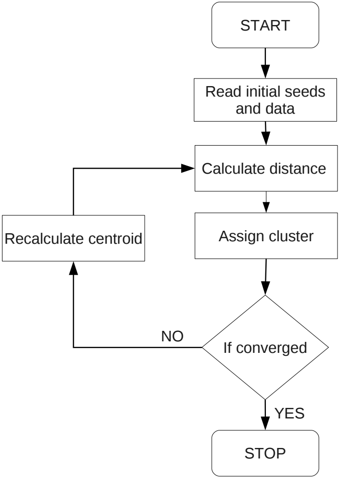
Data Mining for Change Detection

- Changes in forest states are captured by the remote sensing.
- Difficult to use map arithmetic, since the appropriate choice of parameters may vary by region and/or type of forest disturbance.
- An automated, unsupervised change detection system is desired.
- We apply geospatiotemporal data mining techniques to perform unsupervised classification
- Further analysis of clustering outputs for change detection
- Identify unexpected changes in forest states.

k-means cluster algorithm



k-means cluster algorithm

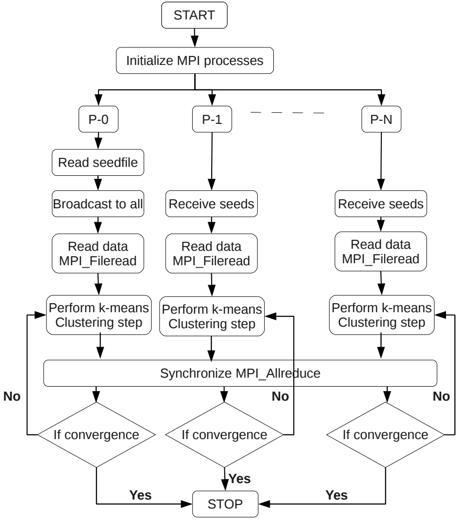


- Serial algorithm
- Requires enough memory to hold all the data
- Not adequate for the large data sets of our interest

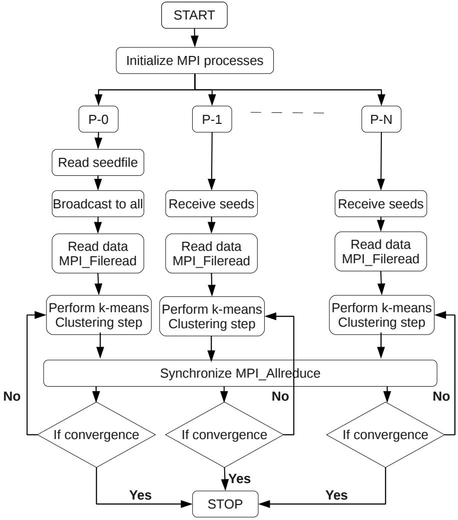
Clustering the MODIS NDVI data

- Data from MODIS: Continental US at 250m resolutions, 16 days
- The ~22B NDVI values in the data set are arranged as annual NDVI traces of 22 values, for each grid cell (~146.4M records) in each of the seven yearly maps (2003-2009),
- The entire set of NDVI traces for all years and map cells is combined into one 84 GB (single precision binary) data set of 22-dimensional "observation" vectors that are analyzed via the *k*-means algorithm.
- After applying *k*-means, cluster assignments are mapped back to the map cell and year from which each observation came, yielding seven maps in which each cell is classified into one of *k* phenoclasses
- The phenoclasses form a "dictionary" of representative or prototype annual NDVI traces (the cluster centroids) derived from the full spatiotemporal extent of the observations in the input data set.

Parallel *k*-means cluster algorithm



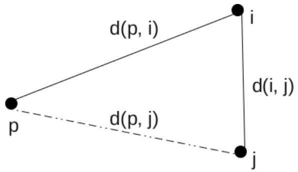
Parallel *k*-means cluster algorithm



- Masterless parallel algorithm
- Data partitioned acrosss distributed memory processors
- Triangular inequality based acceleration
- Warping to handle any null clusters
- Suitable for very large data sets

Enhancements to k-means algorithm

Triangular inequality based acceleration (Phillips 2002):

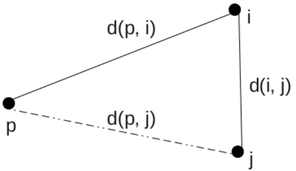


$d(i,j) \leq d(p,i) + d(p,j)$
$d(i,j) - d(p,i) \leq d(p,j)$
if $d(i,j) \geq 2d(p,i)$:
$d(p,j) \geq d(p,i)$
without calculating the distance $d(p,j)$

- Calculate inter-centroidal distances
- Sort the inter-centroidal distances

Enhancements to k-means algorithm

Triangular inequality based acceleration (Phillips 2002):



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

- Calculate inter-centroidal distances
- Sort the inter-centroidal distances

Warping to handle null clusters:

- Avoid empty clusters
- Move "worst of the worst" point to the empty cluster
- Update cluster sizes and recalculate centroid

Phillips, S. J. (2002) "Acceleration of K-Means and Related Clustering Algorithms", ALENEX '02: Revised Papers from the 4th International Workshop on Algorithm Engineering and Experiments, Springer-Verlag, 2002, 166-177

Data sets and resources used

Summary of data sets used

Dataset	No. of dimensions	No. of records	Dataset size
fullUS	25	7,801,710	745 MB
AmeriFlux	30	7,856,224	900 MB
Phenology	22	1,024,767,667	84 GB

Data sets and resources used

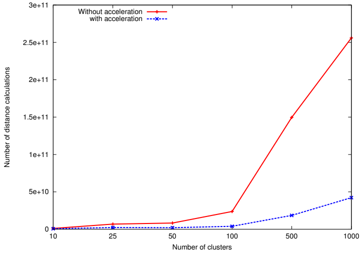
Summary of data sets used

Dataset	No. of dimensions	No. of records	Dataset size
fullUS	25	7,801,710	745 MB
AmeriFlux	30	7,856,224	900 MB
Phenology	22	1,024,767,667	84 GB

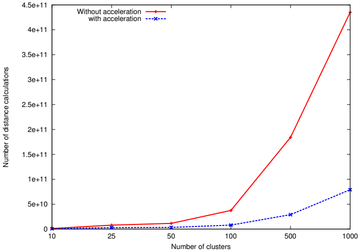
Jaguar Cray XT5 (ORNL):

- 18,688 compute nodes
  - Dual hex-core AMD Opteron 2435 (istanbul) processors 2.6GHz
  - 16GB DDR2-800 memory
- Seastar 2+ router
- Parallel lustre filesystem
- Peak performance: 2.3 petaflops/s

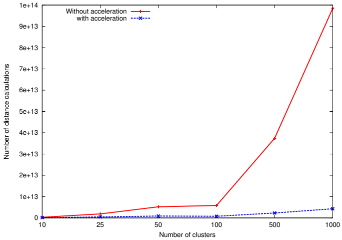
Effect of acceleration: Scaling with increasing k and n: No. of distance calculations



fullUS data set

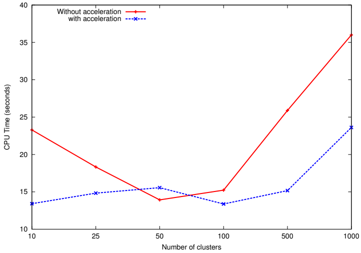


Ameriflux data set

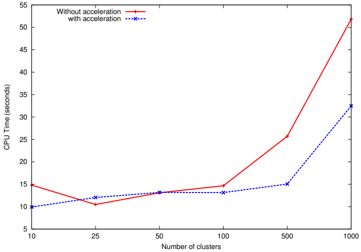


Phenology data set

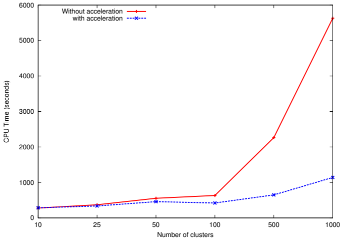
Effect of acceleration: Scaling with increasing k and n: CPU time



fullUS data set

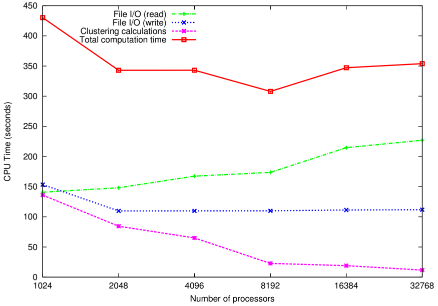


Ameriflux data set

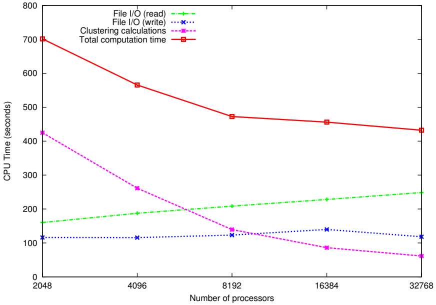


Phenology data set

Strong scaling test: Phenology data set

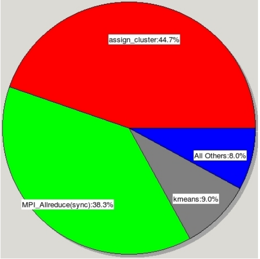


No. of clusters (k) = 50

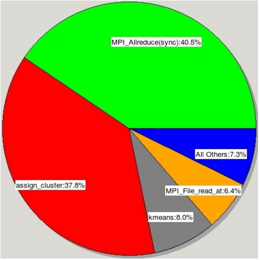


No. of clusters (k) = 1000

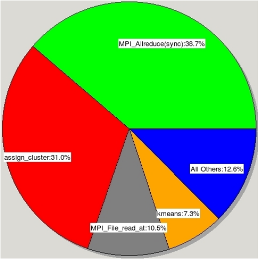
CrayPat summary: Phenology data set, 1000 clusters



Num. procs = 256

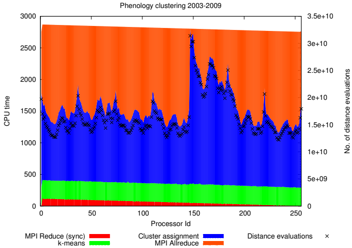


Num. procs = 512

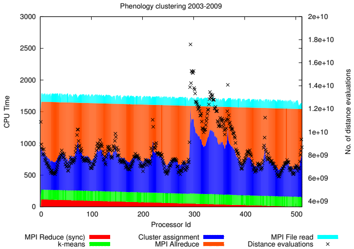


Num. procs = 1024

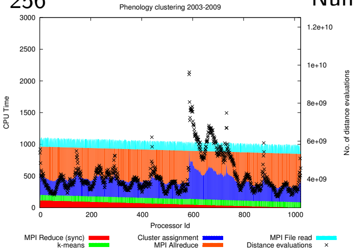
Performance results: Phenology data set, 1000 clusters



Num. procs = 256

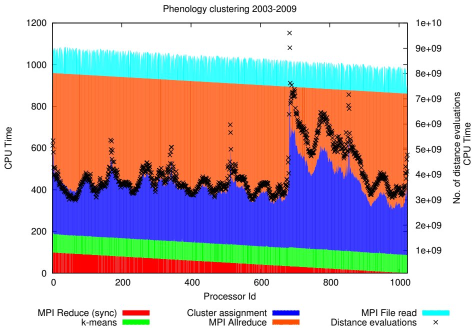


Num. procs = 512

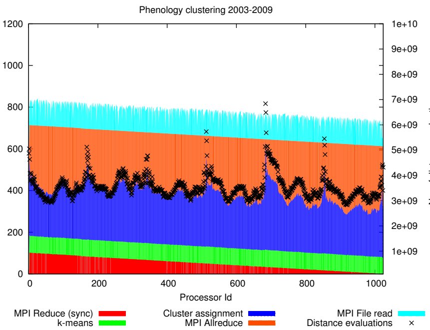


Num. procs = 1024

Performance results: Phenology data set, 1000 clusters, 1024 procs

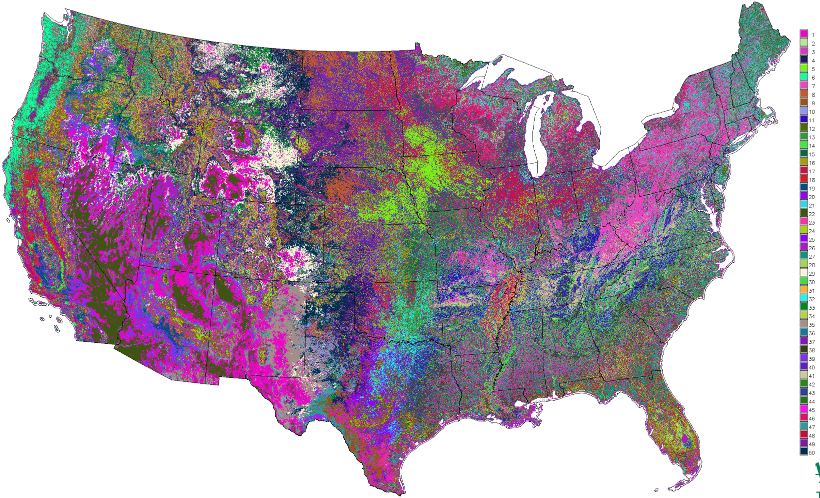


Phenology 2003-2008



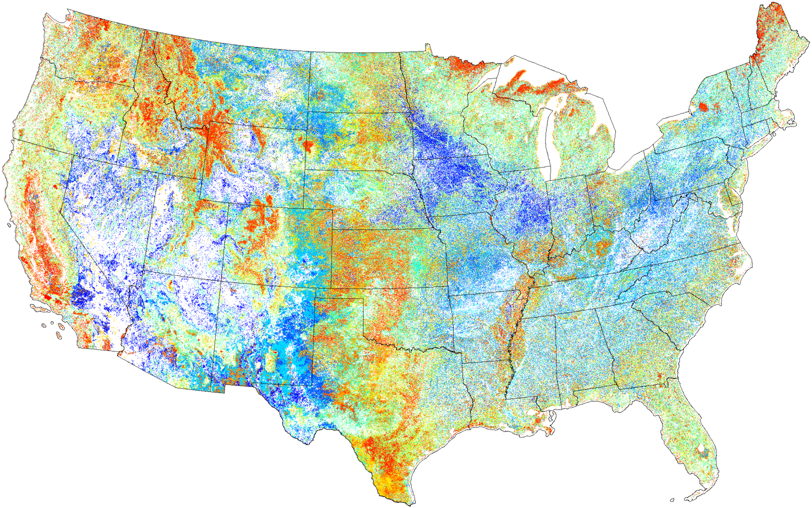
Phenology 2003-2009 (without 2007)

50 Phenoregions for Year 2009 (Clustering 2003-2009)

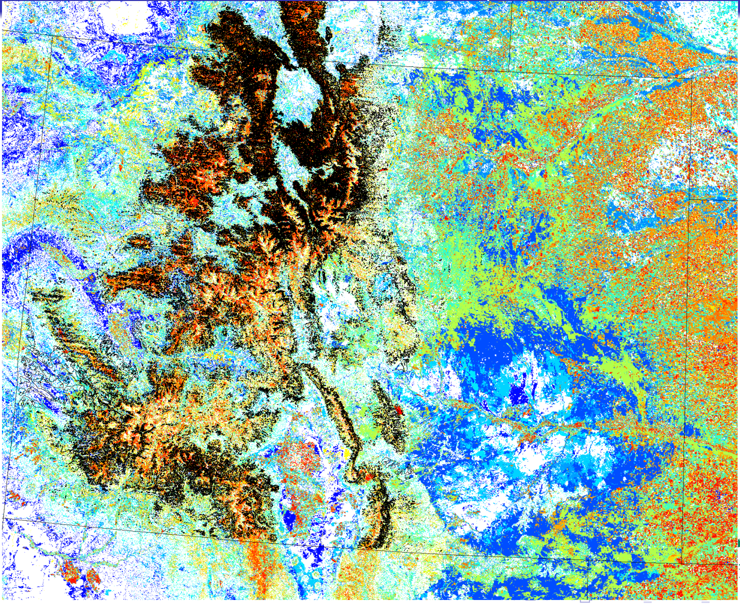


National Laboratory

Transition distance map (2003-2008)



Mountain Pine Beetle in Colorado for (2008 – 2003)



Oak Ridge National Laboratory

Conclusions

- Parallel k-means cluster analysis tool enables the analysis of very large earth sciences data dets
- Enhancements for improved performance of the algorithm
- Scalable design for large data sets
- Good parallel performance and scaling achieved on state-of-the-art supercomputers
- Promising results for geospatiotemporal cluster analysis of phenology from MODIS NDVI
- Successfully applied for forest threat detection; global climate model data comparison (CMIP)

Future Work

- Two-phase I/O for improved parallel I/O performance
- Improved load balancing: block cyclic distribution of data, dynamic load balance
- Support for fuzzy and hierarchical clustering
- Cluster analysis of updated NDVI data sets: 2000-2010(part), every 8 days (200 GB data)
- Cluster analysis for comparison of global climate model results for CMIP5

Acknowledgments

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.

Thank you!

Questions?

Mills, Hoffman, Kumar and Hargrove: "Cluster Analysis-Based Approaches for Geospatiotemporal Data Mining of Massive Data Sets for Identification of Forest Threats ", Session 21b, 2:30PM

Image source: http://blogs.denverpost.com/thespot/files/2010/02/barkbeetle.jpg