Discovery of Patterns in the Global Climate System using Data Mining

Vipin Kumar

University of Minnesota

kumar@cs.umn.edu
www.cs.umn.edu/~kumar

Collaborators:

Chris Potter
NASA Ames
Steve Klooster
California State University, Monterey Bay

Michael Steinbach, Shyam Boriah
University of Minnesota
Pang-Ning Tan
Michigan State University

Research funded by ISET-NOAA, NSF and NASA

Discovery of Climate Patterns from Global Data Sets

Science Goal: Understand global scale patterns in biosphere processes

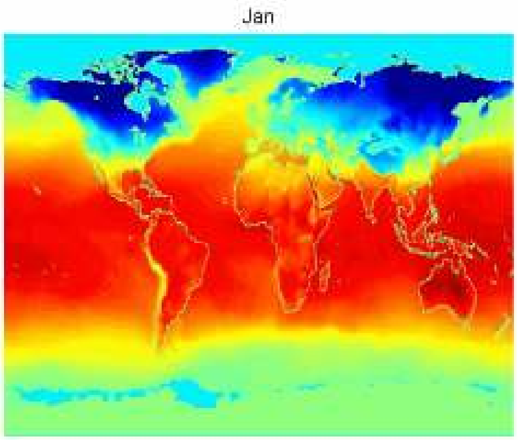
Earth Science Questions:

- When and where do ecosystem disturbances occur?
- What is the scale and location of human-induced land cover change and its impact?
- How are ocean, atmosphere and land processes coupled?

Data sources:

- Weather observation stations
- High-resolution EOS satellites

  1982-2000 AVHRR at 2.5° x 2.5° resolution,
  2000-present MODIS at 250m x 250m resolution
- Model-based data from forecast and other models
- Data sets created by data fusion



Monthly Average Temperature

Earth Observing System

Computer Science Challenges

- Spatio-temporal nature of data
  - Traditional data mining techniques do not take advantage of spatial and temporal autocorrelation.
- Scalability
  - Size of Earth Science data sets has increased 6 orders of magnitude in 20 years, and continues to grow with higher resolution data.
  - Grid cells have gone from a resolution of 2.5° x 2.5° (10K points for the globe) to 250m x 250m (15M points for just California; about 10 billion for the globe)
- High-dimensionality
  - Long time series are common in Earth Science

Detection of Ecosystem Disturbances

Goal: Detection of sudden changes in greenness over extensive land areas due to ecosystem disturbances.

- **Physical**: hurricanes, fires, floods, droughts, ice storms
- **Biogenic**: insects, mammals, pathogens
- **Anthropogenic**: logging, drainage of wetlands, chemical pollution

Motivation: To obtain deeper insight into interplay among natural disasters, human activity and the rise of CO2.

- Satellite observations can reveal completely new pictures of ecological changes and disasters.
- Ecosystem disturbances can contribute to the current rise of CO2 in the atmosphere, with global climate implications.
- In some remote locations, disturbances may have gone undetected.

Haze from forest fires over the Indonesian island of Borneo (October 5, 2006). Over 8 million hectares of forest and farmland burned during August 2006.

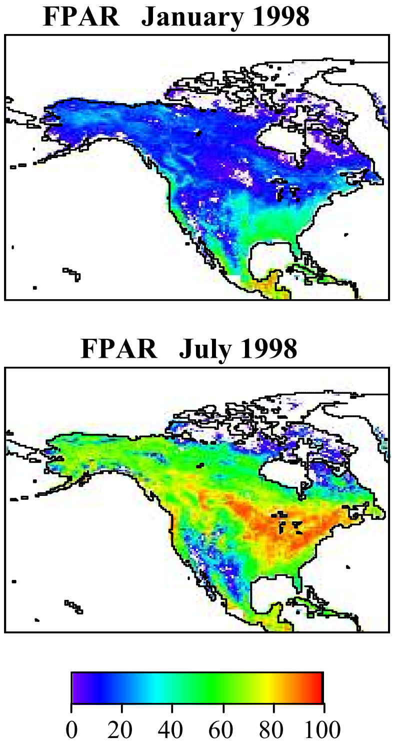
Image Source: NASA

Detection of Ecosystem Disturbances

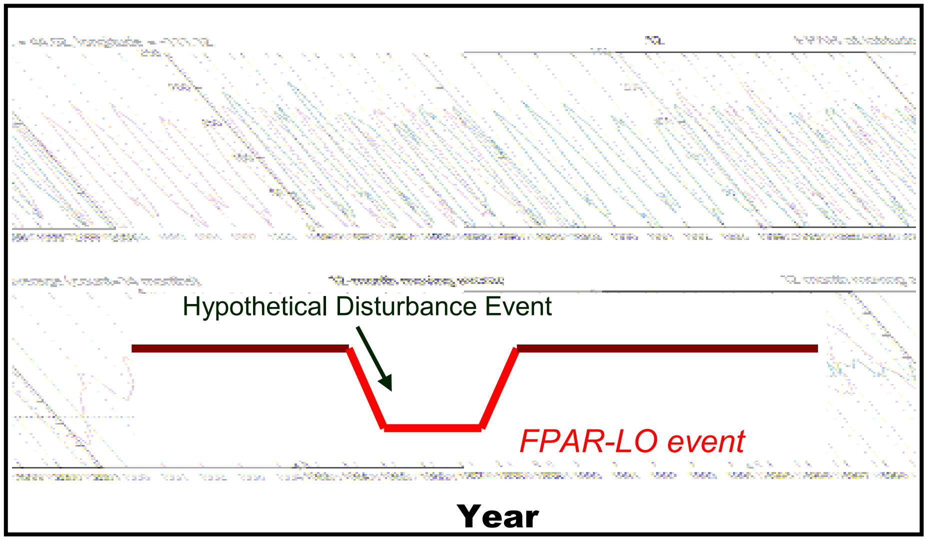
Hypothesis: significant and sustained decline in vegetation FPAR observed from satellites represents a disturbance event

- Can be verified from independent records of such disturbances.

FPAR: Fraction absorbed of Photosynthetically Active Radiation by vegetation canopies


FPAR January 1998

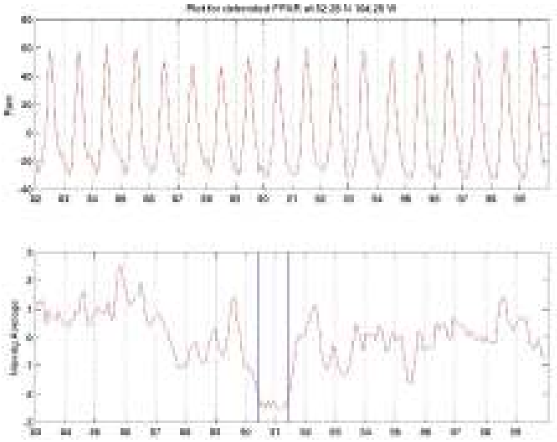
FPAR July 1998



Potter, et al., "Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets", Global Change Biology, 9(7), 1005-1021, 2003.

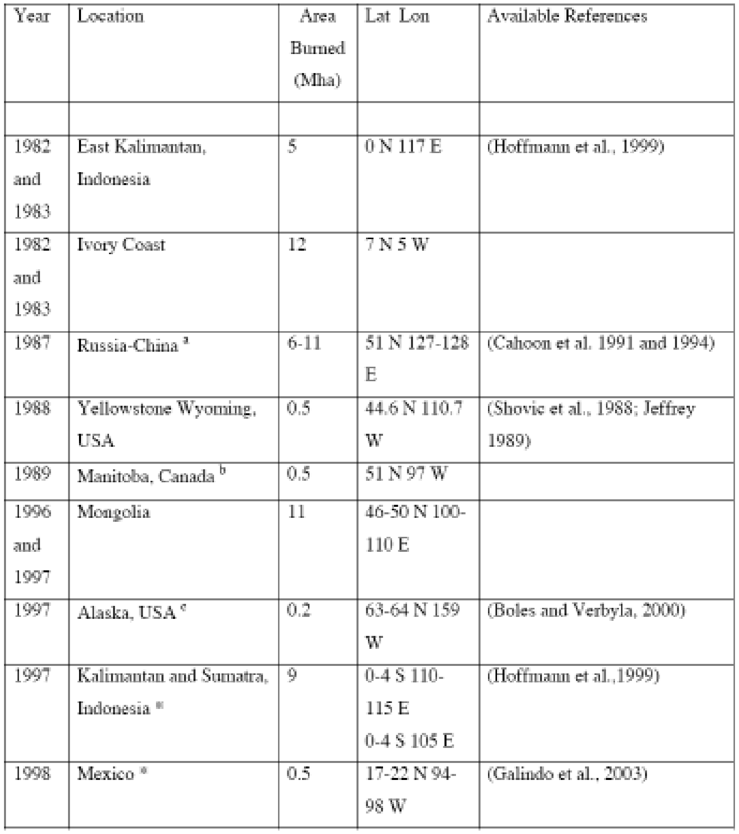
Verification of Disturbances: Fires

Yellowstone Fires 1988



Manitoba, Canada, 1989

List of well-documented wildfires that burned areas covering several Mha in a single year or vegetation growing season.

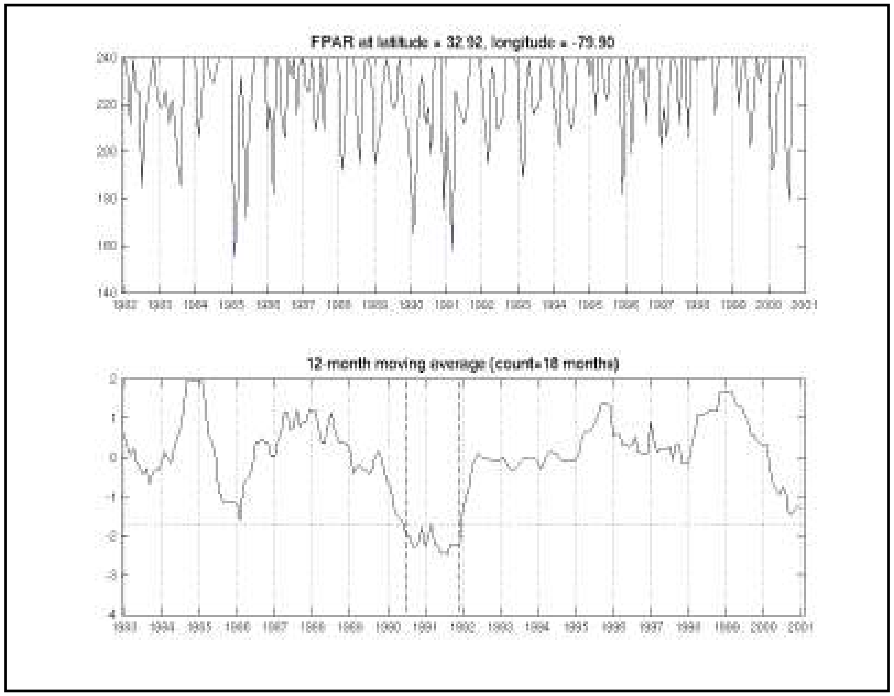


Year	Location	Area Burned (Mha)	Lat Lon	Available References
1982 and 1983	East Kalimantan, Indonesia	5	0 N 117 E	(Hoffmann et al., 1999)
1982 and 1983	Ivory Coast	12	7 N 5 W	
1987	Russia-China [a]	6-11	51 N 127-128 E	(Cahoon et al. 1991 and 1994)
1988	Yellowstone Wyoming, USA	0.5	44.6 N 110.7 W	(Shovic et al., 1988; Jeffrey 1989)
1989	Manitoba, Canada [b]	0.5	51 N 97 W	
1996 and 1997	Mongolia	11	46-50 N 100-110 E	
1997	Alaska, USA [c]	0.2	63-64 N 159 W	(Boles and Verbyla, 2000)
1997	Kalimantan and Sumatra, Indonesia [a]	9	0-4 S 110-115 E 0-4 S 105 E	(Hoffmann et al.,1999)
1998	Mexico [a]	0.5	17-22 N 94-98 W	(Galindo et al., 2003)

For each confirmed wildfire event listed in the table, our disturbance detection method confirms a FPAR-LO event at (or near) the SD >= 1.7 level lasting >12 consecutive months associated with the reported time period of actual fire activity.

Verification of Disturbances: Hurricanes

Hurricane Hugo 1989

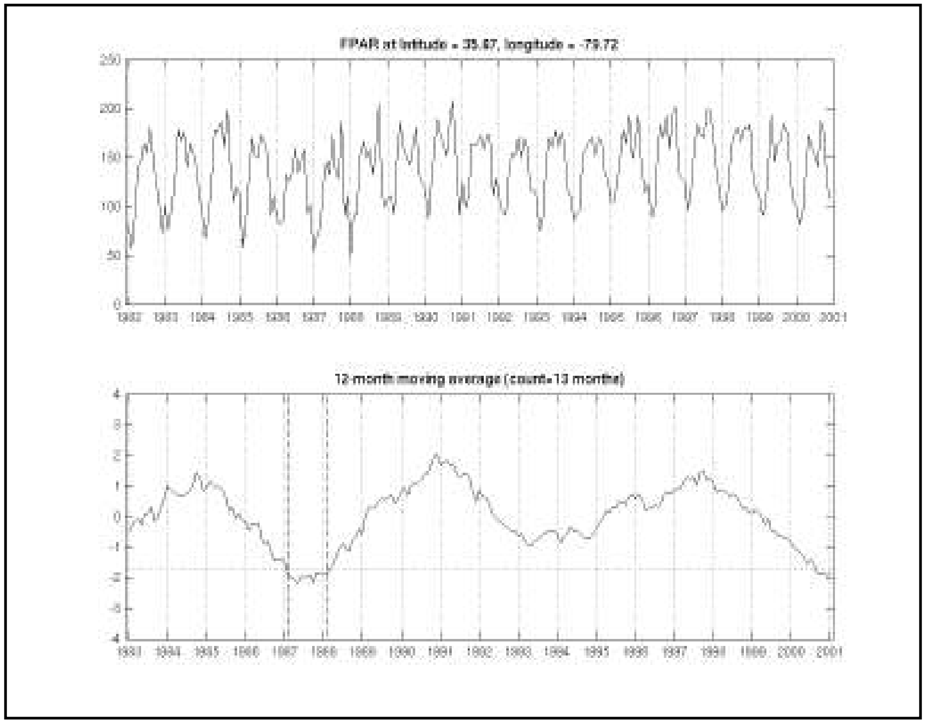


Hurricanes of the 1980s Detected as FPAR-LO Events

Year	Hurricane	Category	Landfall Location	Landfall Lat/Lon
1983	Alicia	3	SE Texas, USA	28.9 N 95.0 W
1985	Gloria	3	East Coast, USA	35.5 N 75.5 W
1985	Elena	3	Mississippi, USA	30.2 N 88.8W
1988	Gilbert	3	East Coast, Mexico	20.4 N 86.5 N, 23.9 N 97.0 W
1989	Hugo	4	North Carolina, USA	33.5 N 80.3 W

Verification of Disturbances: Droughts

Southern Drought 1986



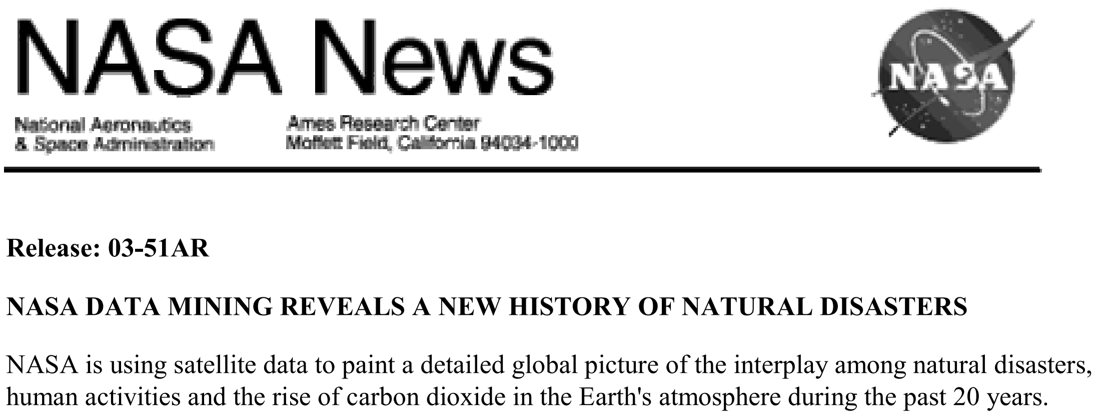
Major Droughts Detected as FPAR-LO Events

Year	Drought	Most Heavily Impacted Regional Locations
1986	Southern USA	Georgia, Carolinas, California
1988	Central USA	Midwest and Northeast states
1989	Northern Plains	Colorado
1993	SE USA	Alabama, Georgia, Carolinas, Tennessee, Virginia
1998	Southern USA	Texas, Oklahoma, Carolinas, Georgia, Florida

Detection of Ecosystem Disturbances

Outcomes: Estimated that 9 billion metric tons of carbon may have moved from the Earth’s soil and surface life forms into the atmosphere in 18 years beginning in 1982 due to wildfires and other disturbances.

- Fossil fuel emission of CO2 to the atmosphere each year was about 7 billion metric tons in 1990.



NASA News

National Aeronautics
& Space Administration

Ames Research Center
Moffett Field, California 94034-1000

Release: 03-51AR

NASA DATA MINING REVEALS A NEW HISTORY OF NATURAL DISASTERS

NASA is using satellite data to paint a detailed global picture of the interplay among natural disasters, human activities and the rise of carbon dioxide in the Earth's atmosphere during the past 20 years.

http://www.nasa.gov/centers/ames/news/releases/2003/03_51AR.html

<u>Uniqueness of study:</u>

- global in scope
- covered more than a decade of analysis
- encompass all potential categories of major ecosystem disturbance – physical, biogenic, and anthropogenic

Land Cover Change Detection

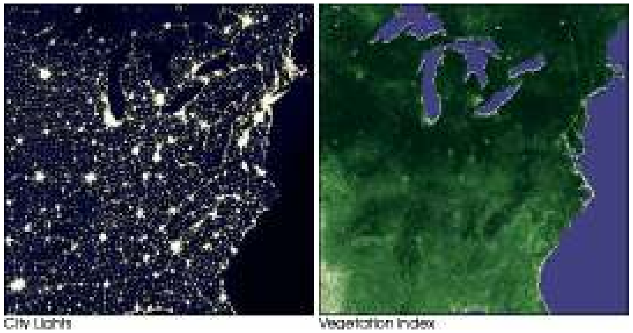
Goal: Determine **where**, **when** and why natural ecosystem conversions occur

- E.g. Deforestation, Urbanization, Agricultural intensification

Motivation:

- Characteristics of the land cover impacts Local climate, Radiation balance, Biogeochemistry, Hydrology, Diversity/abundance of terrestrial species
- Conversion of natural land cover can have undesirable environmental consequences

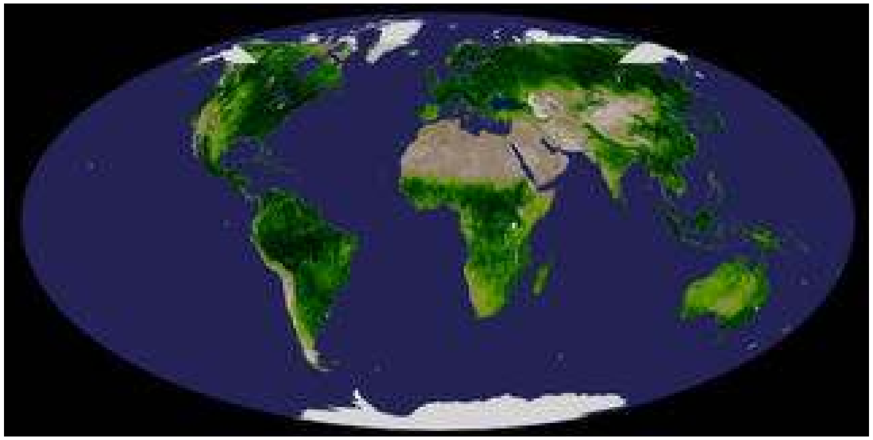
Deforestation changes local weather. Cloudiness and rainfall can be greater over cleared land (image right) than over intact forest (left).



Urbanization tends to reduce vegetation density.

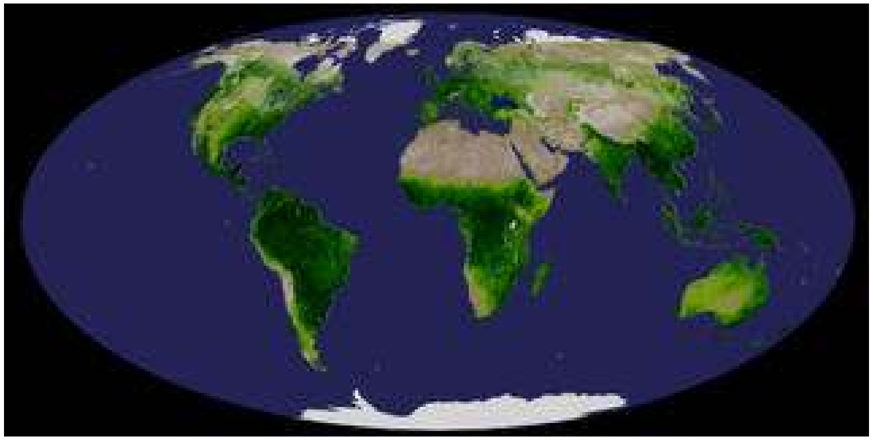
Source: NASA Earth Observatory

Data: Enhanced Vegetation Index



Global EVI in Summer, 2000.

- Enhanced Vegetation Index (EVI) represents the "greenness" signal (area-averaged canopy photosynthetic capacity), with improved sensitivity in high biomass cover areas.
- MODIS algorithms have been used to generate the Enhanced Vegetation Index (EVI) at 250-meter spatial resolution from Feb 2000 to the present

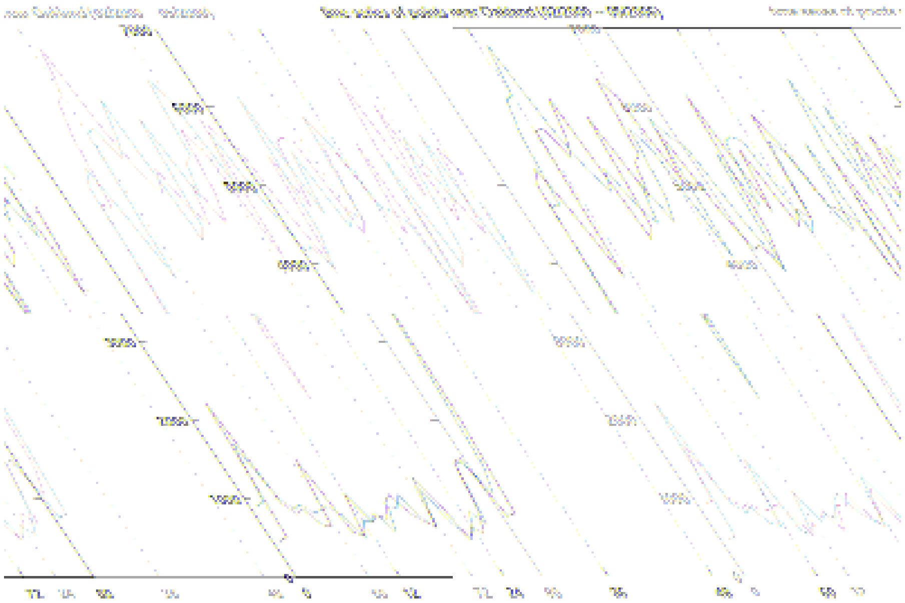


Global EVI in Winter, 2001.

NASA's Terra satellite platform launched in 1999 has the Moderate Resolution Imaging Spectroradiometer (MODIS)

Image Source: NASA/Goddard Space Flight Center Scientific Visualization Studio

Examples of Change Points



- The two time series show an abrupt jump in EVI in 2003; a land cover change pattern we are looking for.
- The location of the points correspond to a new golf course, which was in fact opened in 2003.

- Changes of this nature can be detected only with high-resolution data.

Traditional Change Detection Techniques

- Fisher algorithm
- CUSUM (Cumulative Sum Control Charts)
- HMM-based approaches
- Kalman Filter

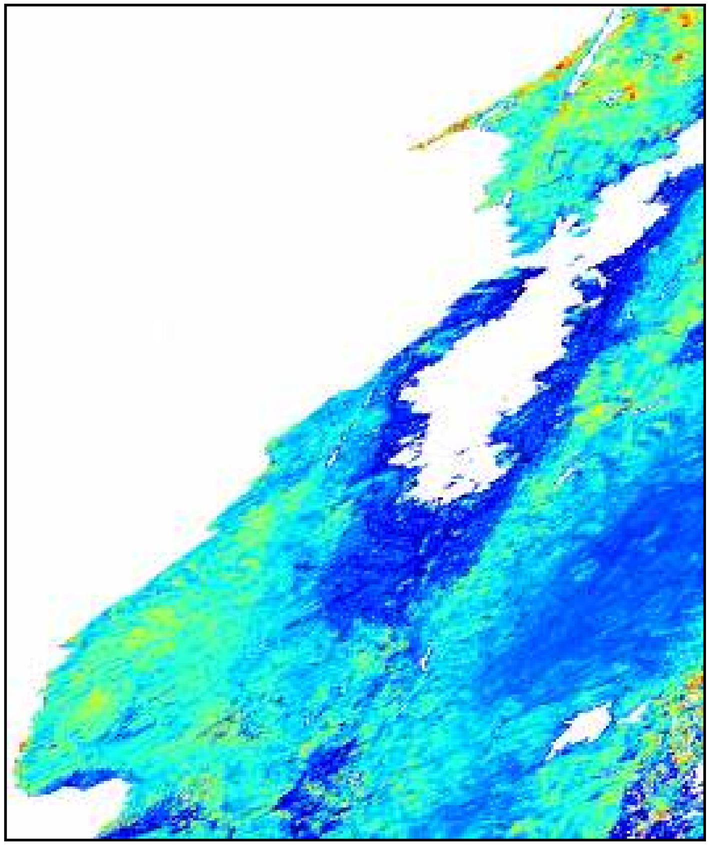
Limitations:

- Most techniques do not scale to massive datasets
- Do not make use of seasonality of Earth Science data and/or intra-season variability
- Spatial and temporal autocorrelation are not exploited

Focus of the Study: Northern California

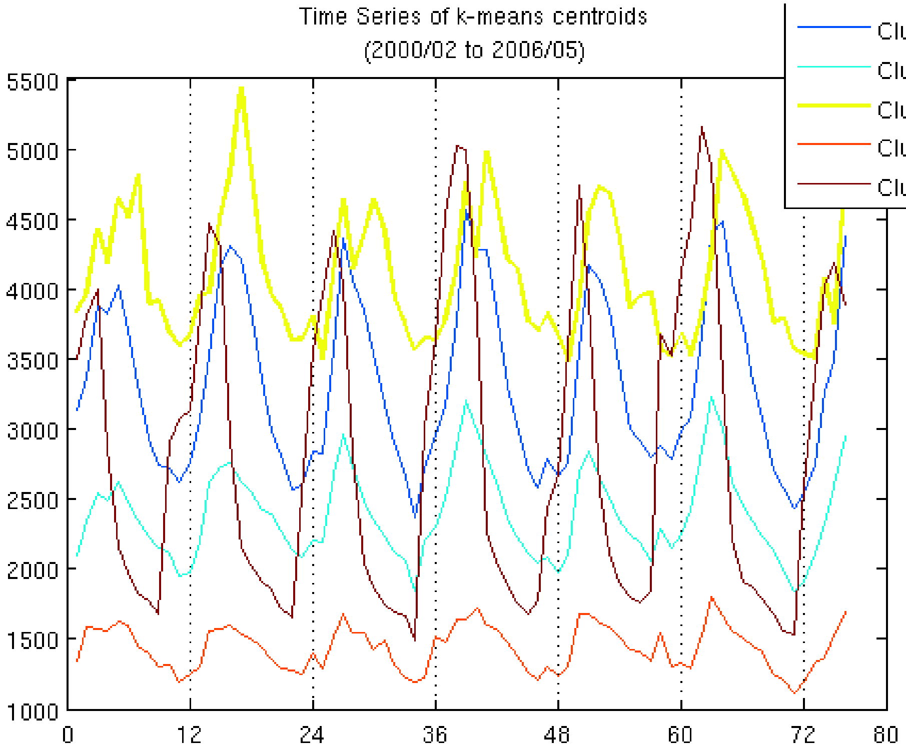
California has experienced rapid population growth and changing economic activities

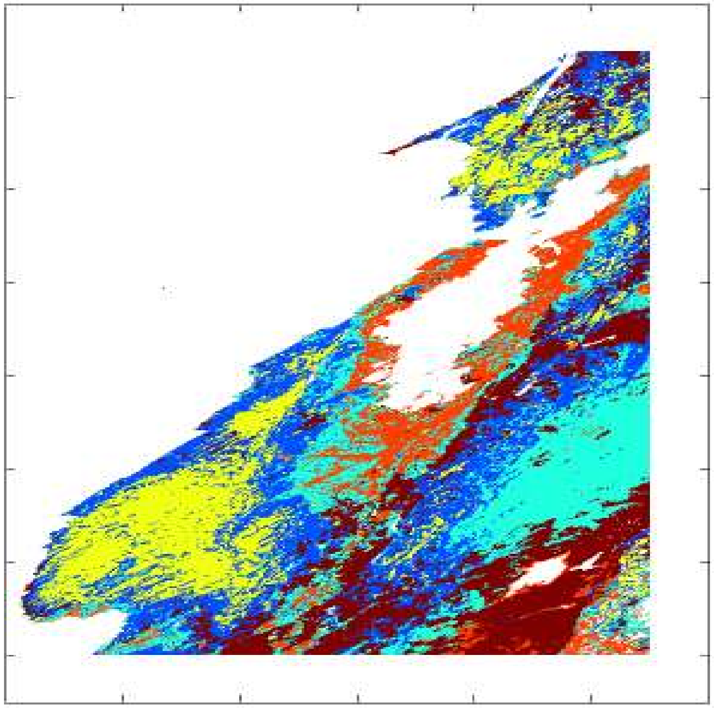
- population increased by 75% between 1970 and 2005
- over half of all new irrigated farmland put into production was of lesser quality than prime farmland taken out of production by urbanization
- San Francisco Bay area selected for analysis



EVI in Northern California for February 2002

High-level view of land cover





Cluster 1 - High seasonal biomass density, moderate interannual variability (**shrub** cover)

Cluster 2 - Moderate annual biomass density, moderate interannual variability (**grass** cover)

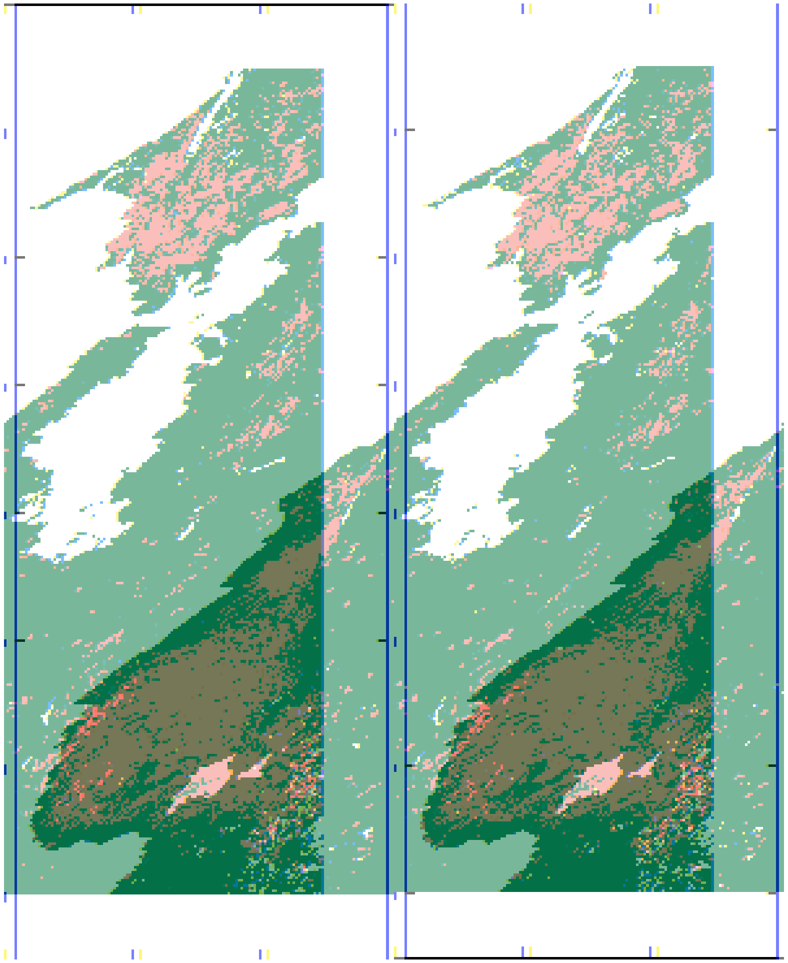
Cluster 3 - High annual biomass density, low interannual variability (**evergreen tree** cover)

Cluster 4 - Low annual biomass density, low interannual variability (**urbanized** cover)

Cluster 5 - High seasonal biomass density, high interannual variability (**agricultural** cover)

Cluster 3

Cluster 3 - High annual biomass density, low interannual variability (**evergreen tree** cover)



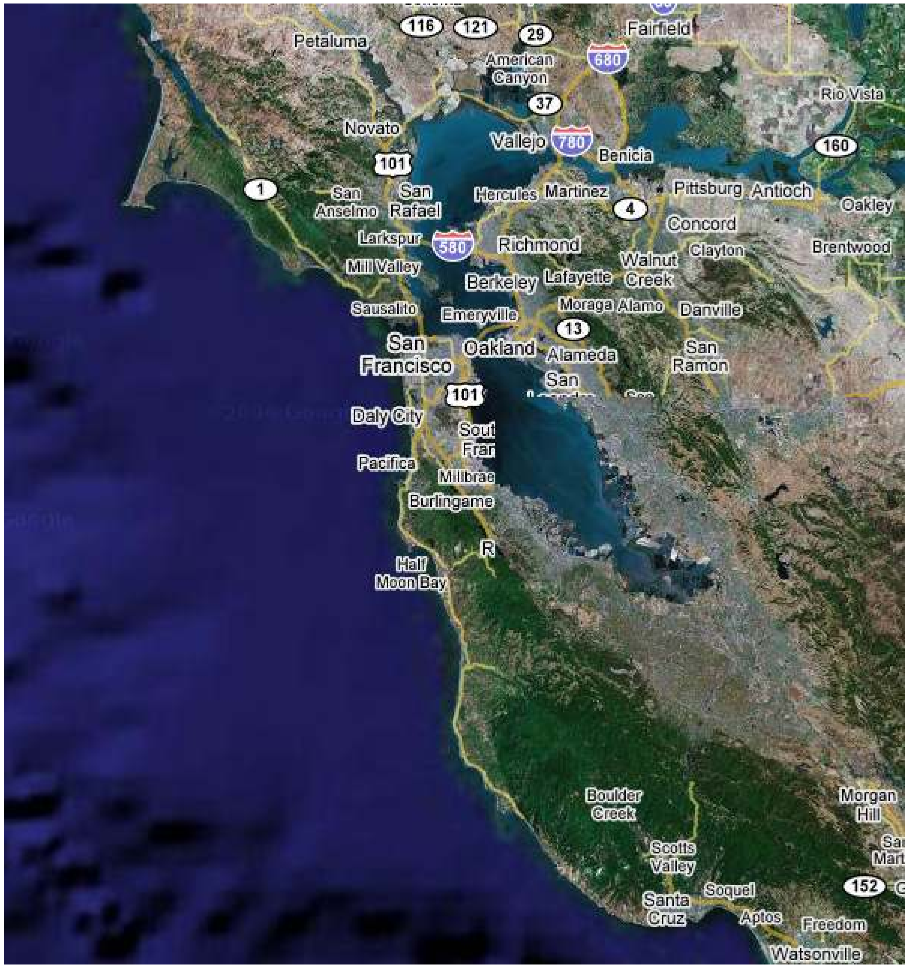
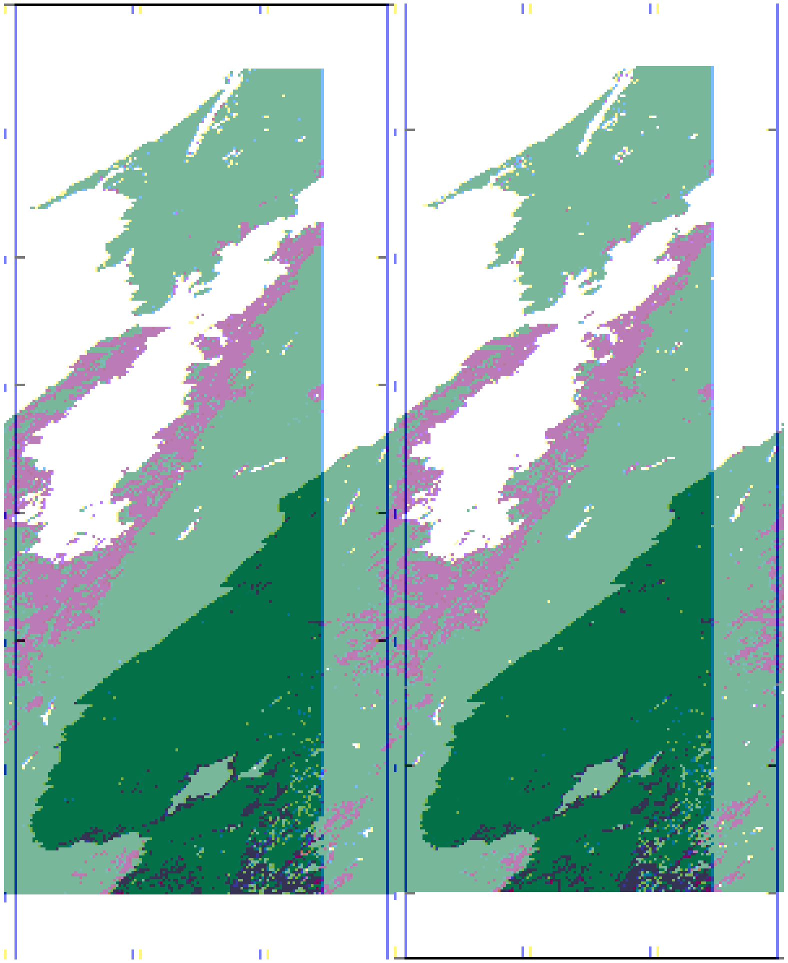


Image source: Google Maps

Cluster 4

Cluster 4 - Low annual biomass density, low interannual variability (**urbanized** cover)



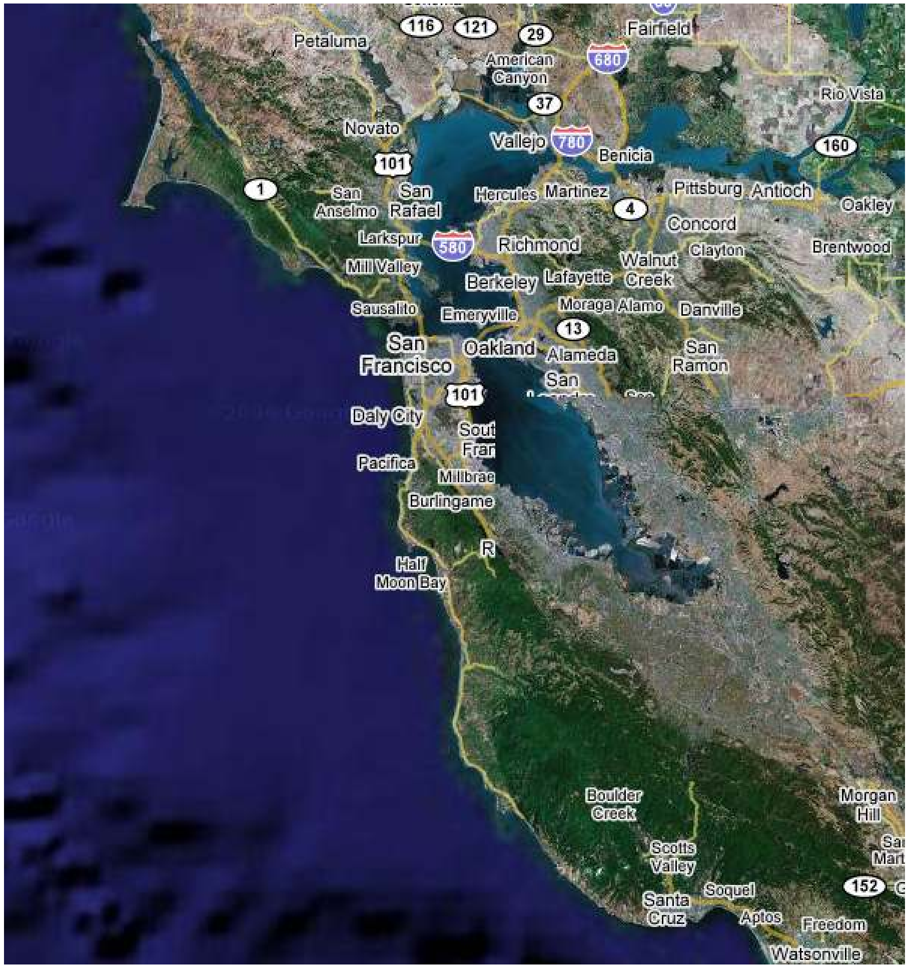
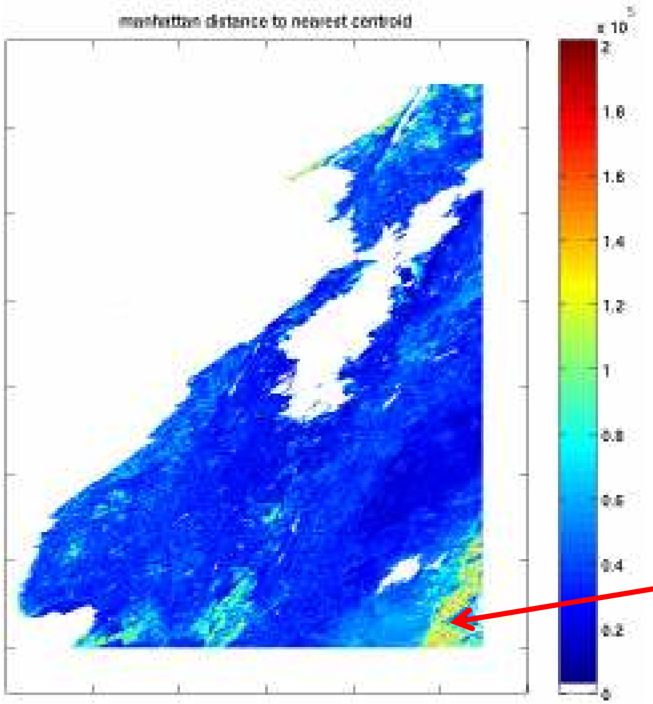
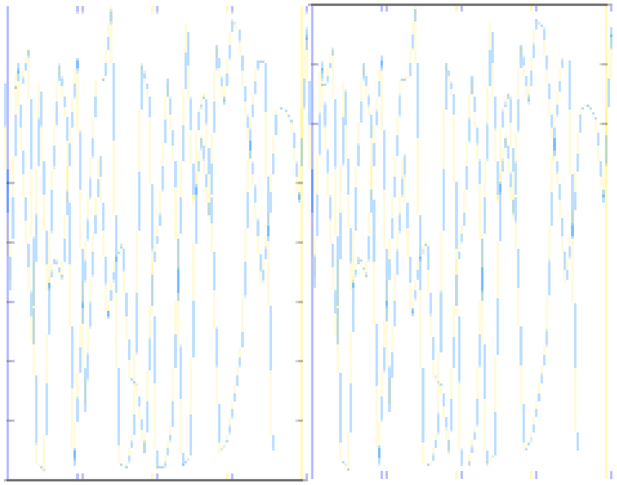


Image source: Google Maps

Distance to Centroid Scheme





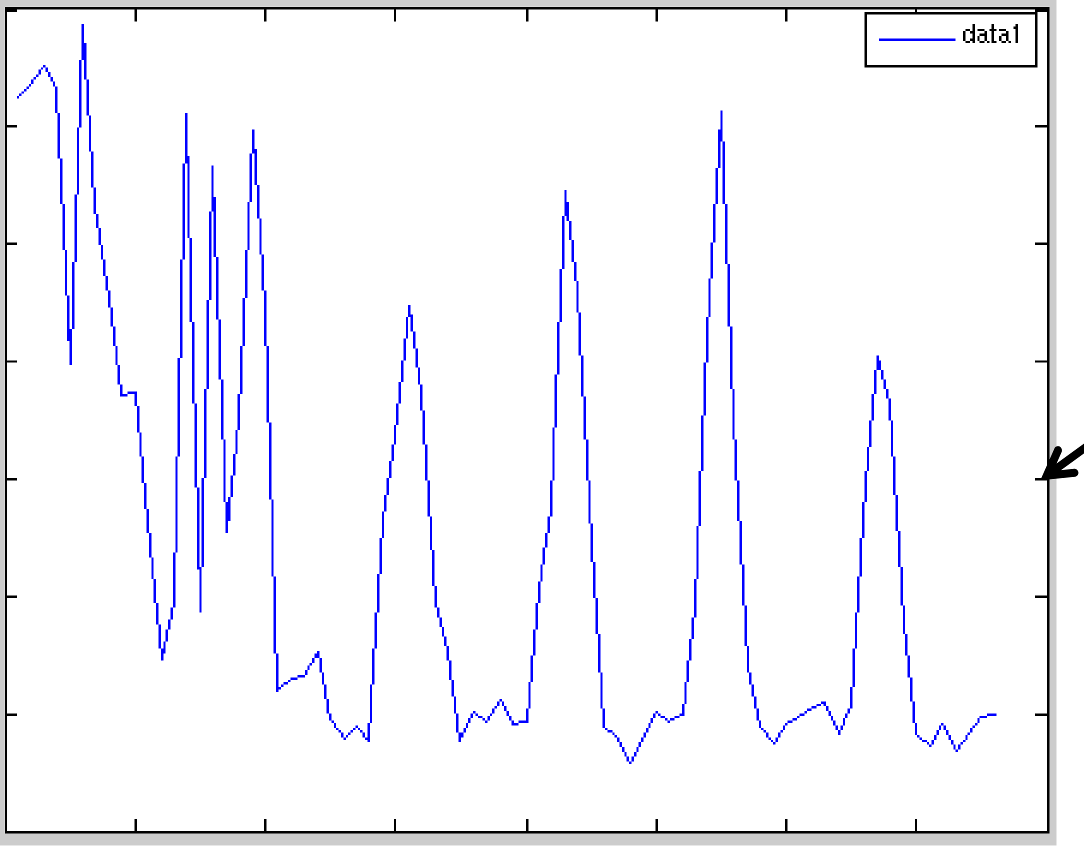
Typical Outlier: Time Series (above) and farm in Central California corresponding to the outlier (below)

Image source: Google Maps

Limitations:

- Sensitive to intra-annual variability in EVI
- Depends on cluster structure
- Treats outliers and change points as the same

Yearly averages scheme



- In this scheme, we look at the differences between yearly averages
- If there is a "jump" in the yearly differences, we consider the time series to have a change point

This time series was given the top score with the above scheme, while more interesting changes were not given high scores.

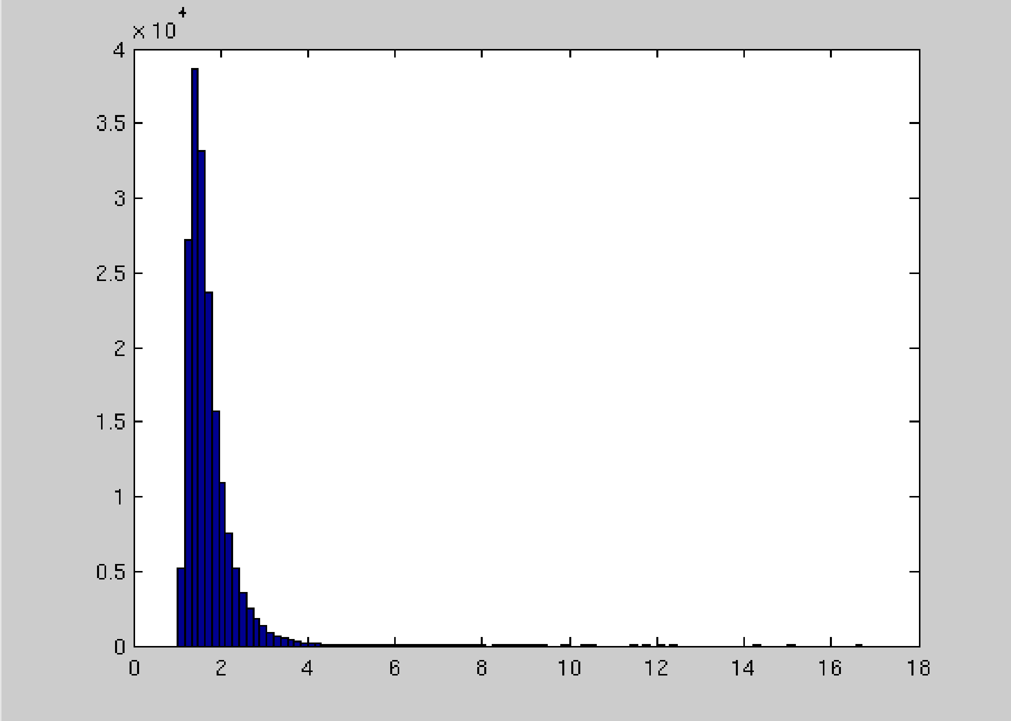
Limitations:

- Interesting points are missed because changes that occur within a year are averaged out, leading to a gradual increase in differences
- Can only detect changes in the mean

A new change detection technique

- The idea behind this technique was to exploit the major mode of behavior (seasonality) in order to detect changes.
- The time series for each location is processed as follows:
  1. The two most similar seasons are merged, and the distance/similarity is stored.
  2. Step 1 is applied recursively until one season is left.
  3. The change score for this location is based on whether any of the observed distances are extreme (e.g. ratio of maximum distance/minimum distance).

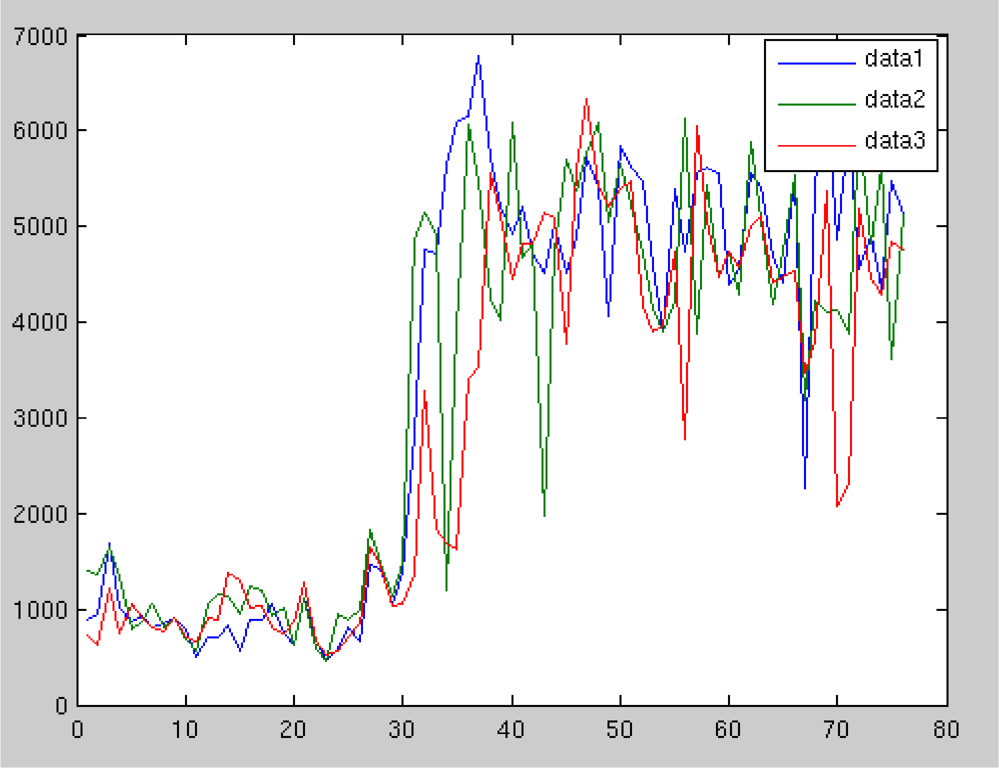
Results: Histogram of Scores



Histogram of all scores

- There are about 180K points in total.
- 900 have score > 4
- 31 points have score > 8. Of these 22 points were found to correspond to interesting land-use changes. Others corresponded to farm land.

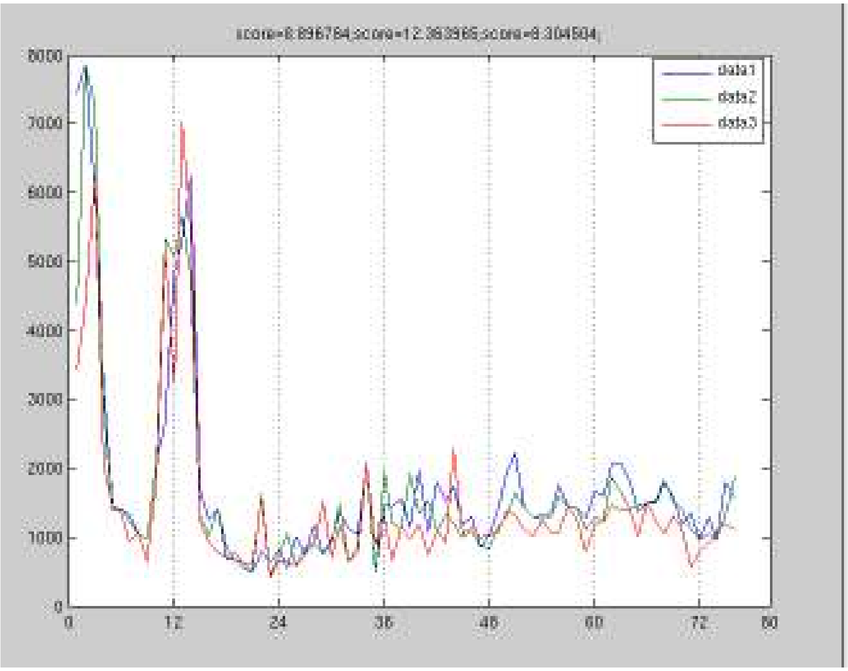
Results: Top 3 scores

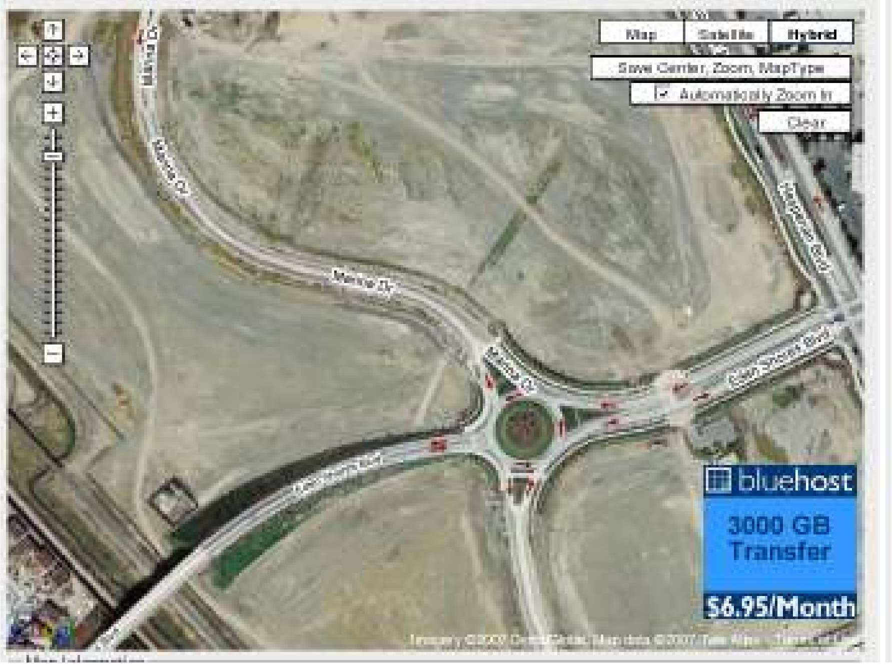


McCormick St
Webster St
Hester St
61
Doolittle Dr
Adams Ave
98th Ave
Airport Dr
Ron Cowan Pkwy
Metropolitan Golf Links
Eden Rd
David St

The top 3 points correspond to a golf course in Oakland. This golf course was built in 2003, which corresponds to the time step at which the time series exhibit a change.

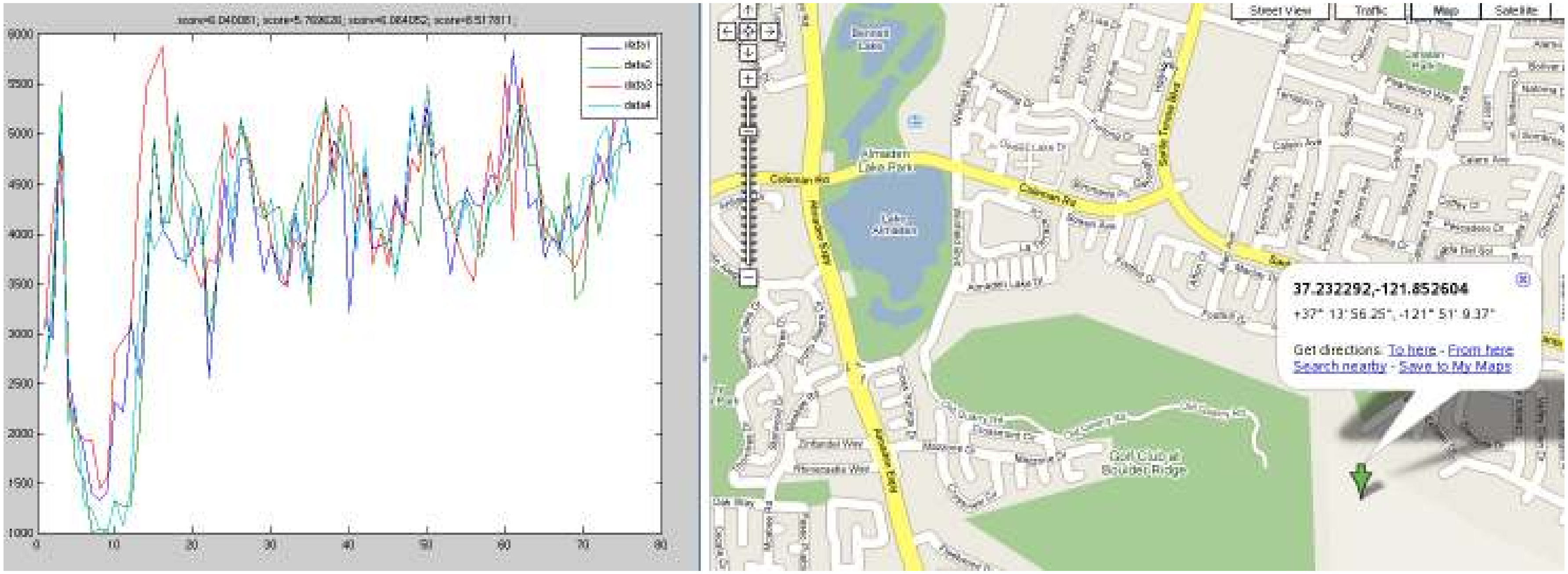
Results: More points





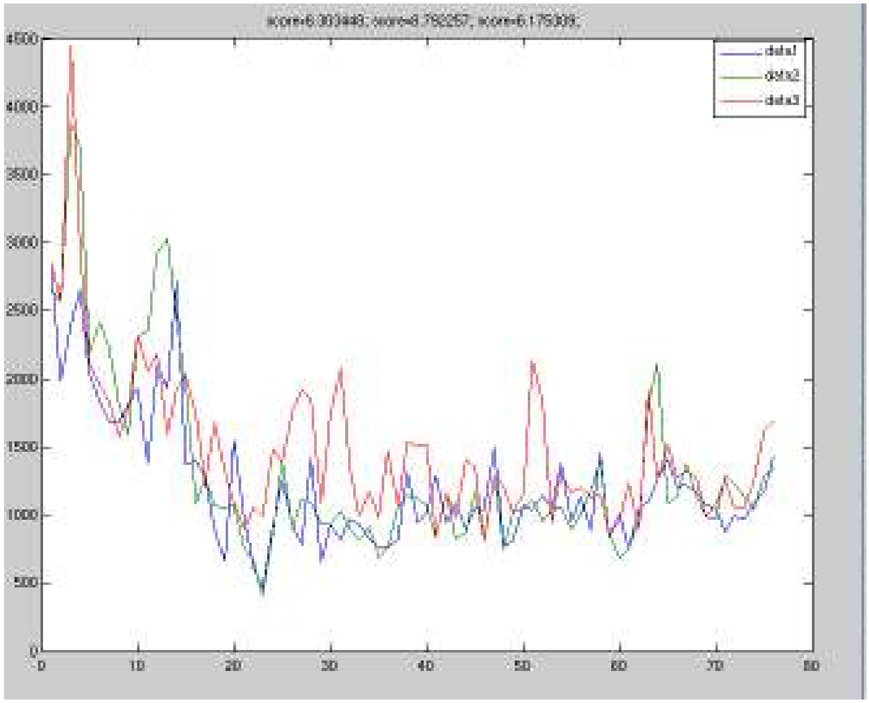
These 3 time series correspond to a subdivision under construction.

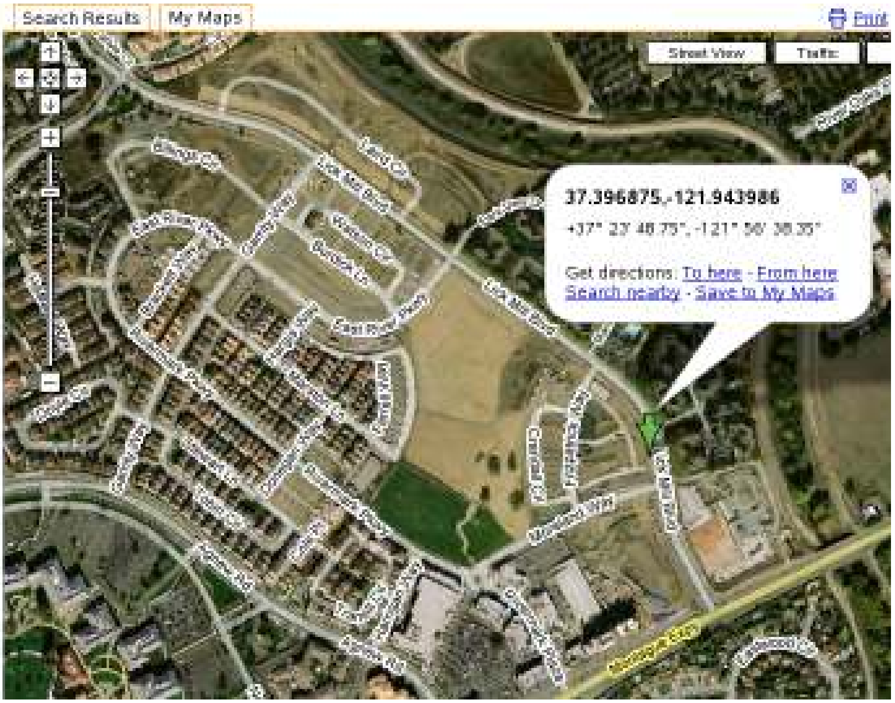
Results: More points



Golf Course (built in 2001, corresponding to change in time series)

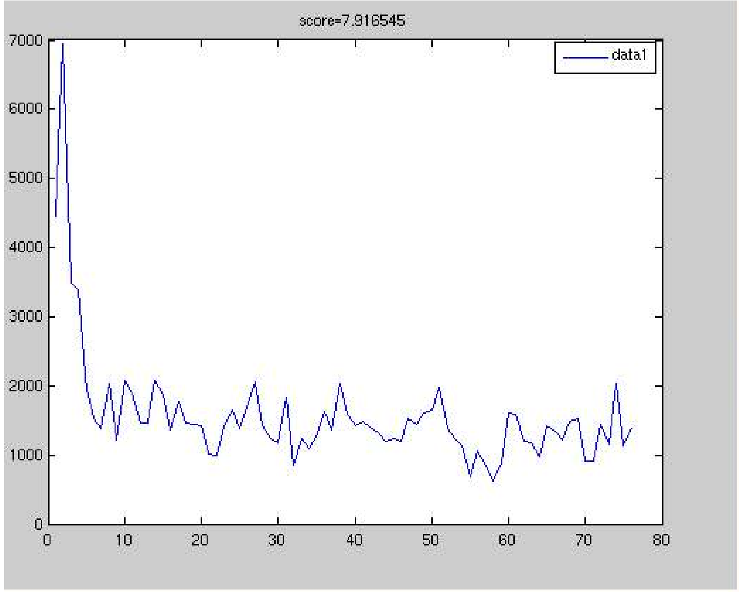
Results: More points

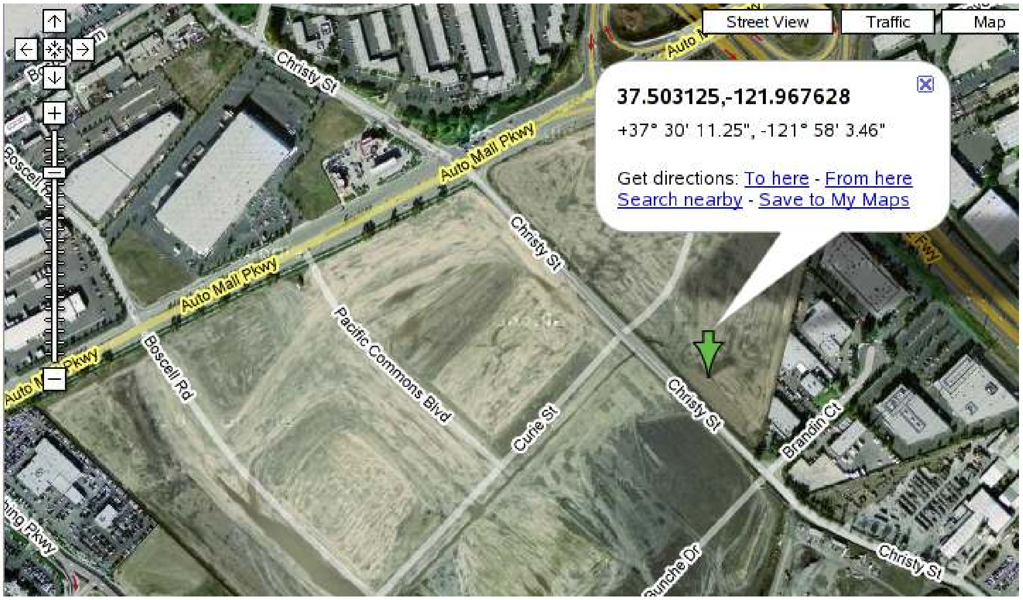




Subdivision built in 2002

Results: More points





37.503125,-121.967628

+37° 30' 11.25", -121° 58' 3.46"

Get directions: To here - From here
Search nearby - Save to My Maps

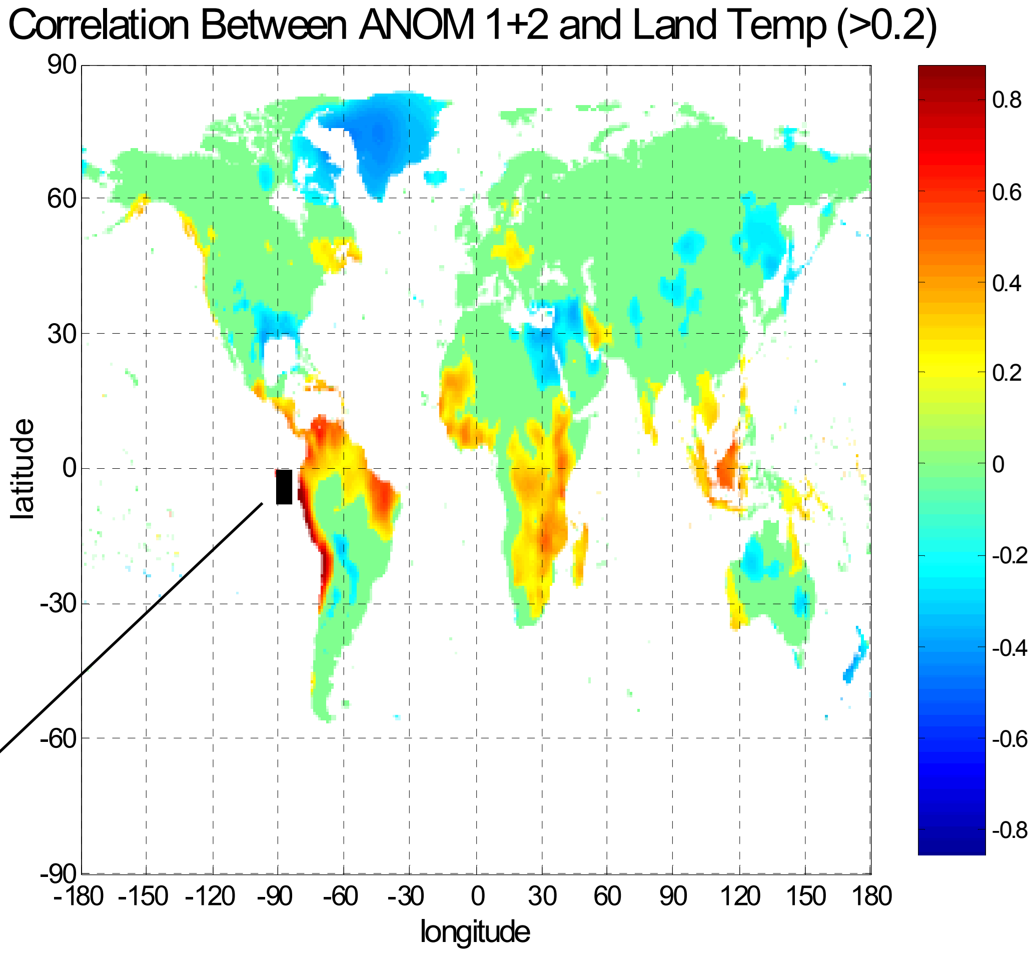
Construction of Pacific Commons shopping area in Fremont, CA

Change Point Challenges

- Spatio-temporal autocorrelation
  - Traditional techniques for change point detection were not developed for spatio-temporal data
- Scalability
  - The data is at 250m resolution (and may become even finer in the future).
  - It is important for any algorithm to be scalable, if it is used with this data
- Characterizing changes
  - Techniques are more useful when changes are characterized in relation to other points
  - This greatly enhances the ability of the domain scientist to explain **why** the change occurred

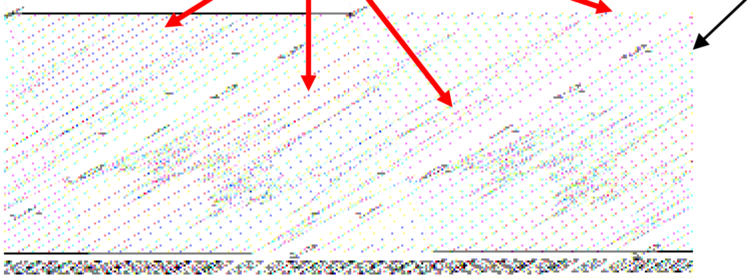
Climate Indices: Connecting the Ocean/Atmosphere and the Land

- A climate index is a time series of sea surface temperature or sea level pressure
- Climate indices capture teleconnections
  - The simultaneous variation in climate and related processes over widely separated points on the Earth



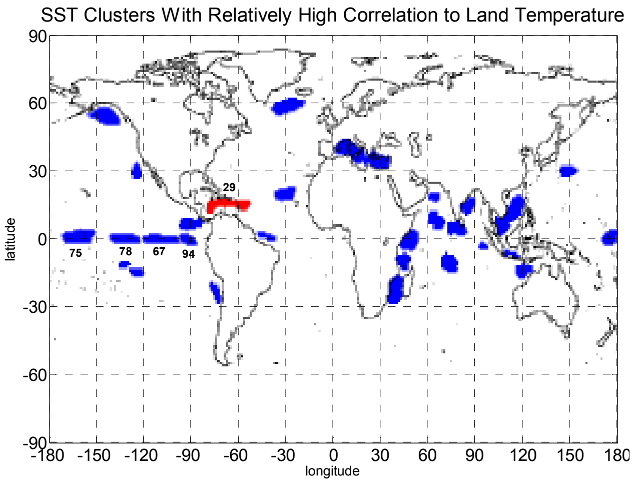
Correlation Between ANOM 1+2 and Land Temp (>0.2)

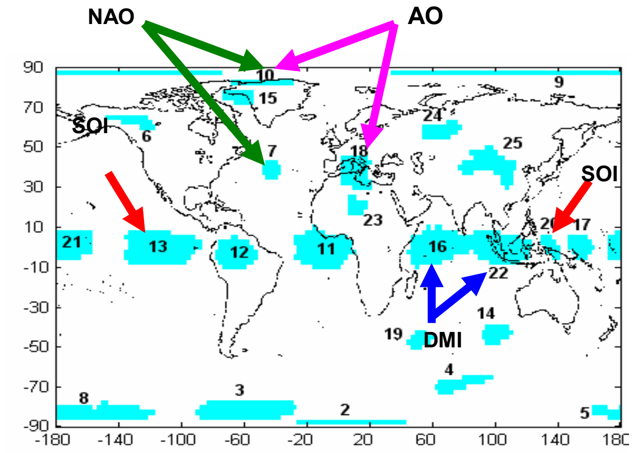
El Nino Events



Nino 1+2 Index

Discovery of Climate Indices Using Clustering





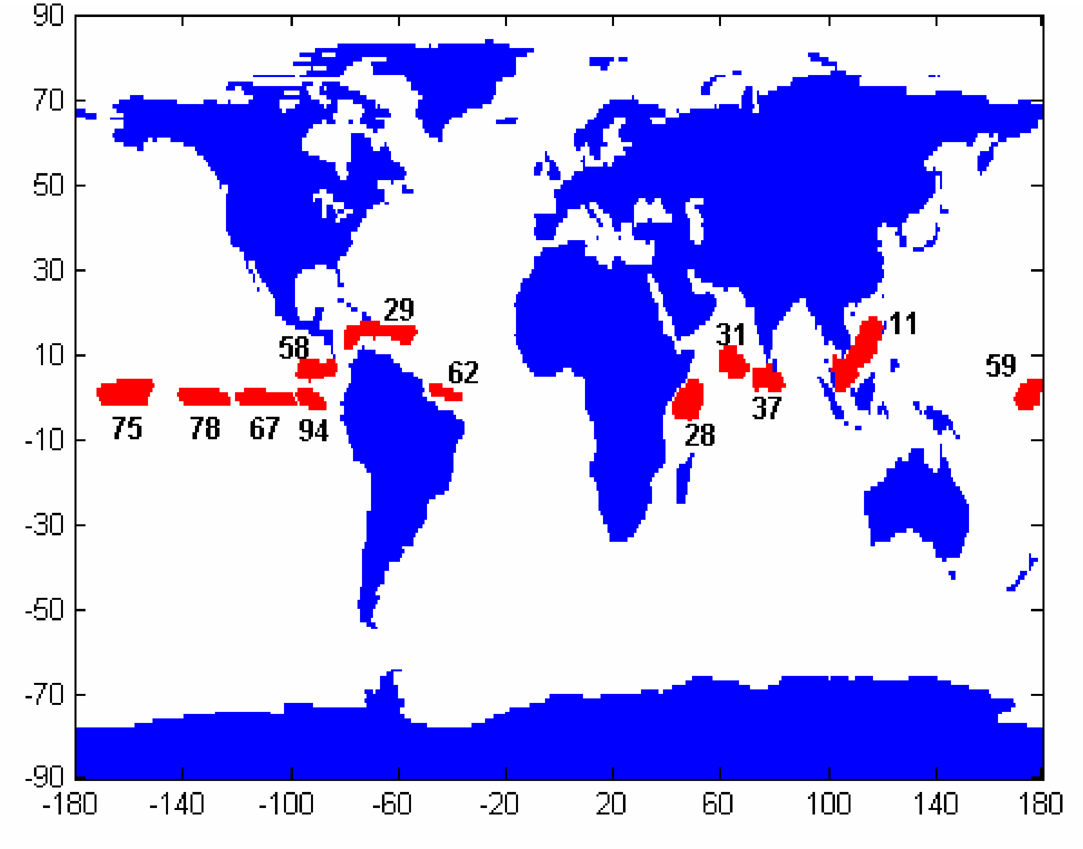
- Clustering provides an alternative approach for finding candidate indices.
  - Clusters represent ocean regions with relatively homogeneous behavior.
  - The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential climate indices.
- Clusters are found using the Shared Nearest Neighbor (SNN) method that eliminates “noise” points and tends to find regions of “uniform density”.
- Clusters are filtered to eliminate those with low impact on land points

Result: A cluster-based approach for discovering climate indices provides better physical interpretation than those based on the SVD/EOF paradigm, and provide candidate indices with better predictive power than known indices for some land areas.

SST Clusters that Reproduce Known Indices

# grid points: 67K Land, 40K Ocean **Current data size range: 20 – 400 MB**

Monthly data over a range of 17 to 50 years

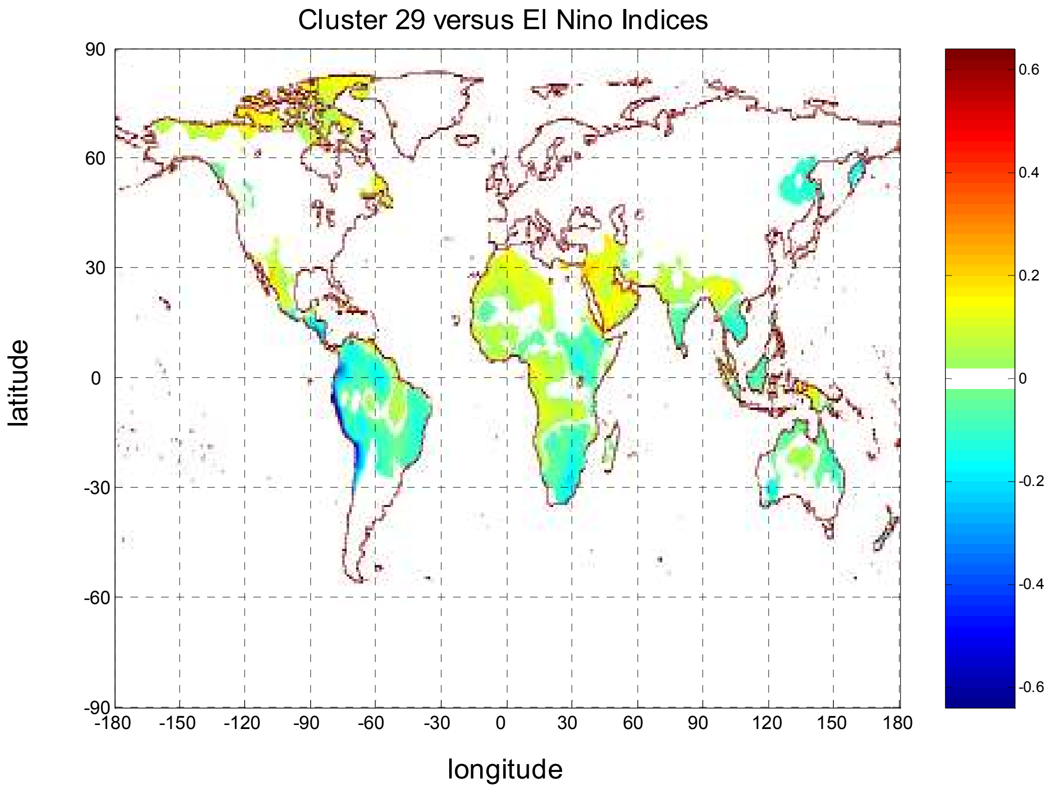


Clusters of SST that have high impact on land temperature

Cluster	Nino Index	Correlation
94	NINO 1+2	0.9225
67	NINO 3	0.9462
78	NINO 3.4	0.9196
75	NINO 4	0.9165

Some SST clusters reproduce well-known climate indices for El Niño.

SST Cluster Moderately Correlated to Known Indices



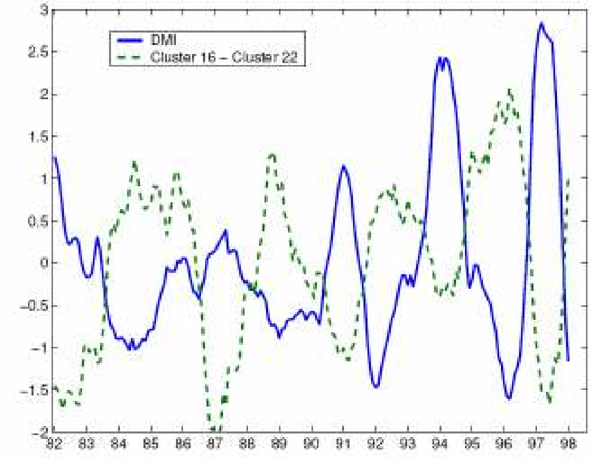
The figure shows the difference in correlation to land temperature between cluster 29 and the El Nino indices. Areas in yellow indicate where cluster 29 has higher correlation.

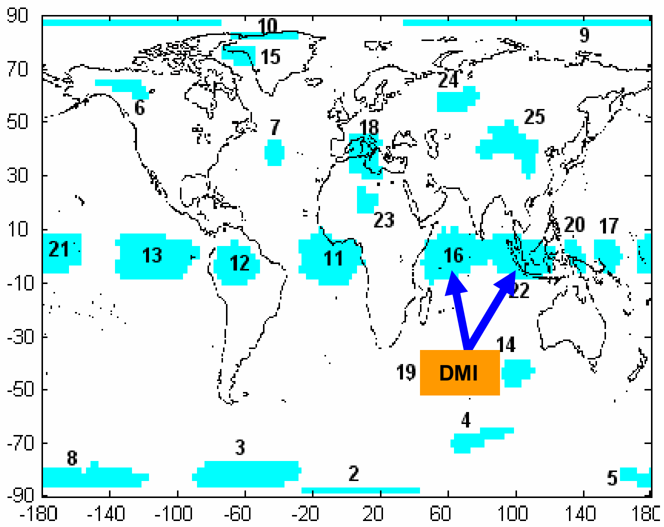
Some SST clusters are significantly different than known indices, but provide better correlation with land climate variables than known indices for many parts of the globe.

Finding New Patterns: Indian Monsoon Dipole Mode Index

- Recently a new index, the Indian Ocean Dipole Mode index (DMI), has been discovered*.
- DMI is defined as the difference in SST anomaly between the region 5S-5N, 55E-75E and the region 0-10S, 85E-95E.
- DMI and is an indicator of a weak monsoon over the Indian subcontinent and heavy rainfall over East Africa.
- We can reproduce this index as a difference of pressure indices of clusters 16 and 22.

Plot of cluster 16 – cluster 22 versus the Indian Ocean Dipole Mode index. (Indices smoothed using 12 month moving average.)



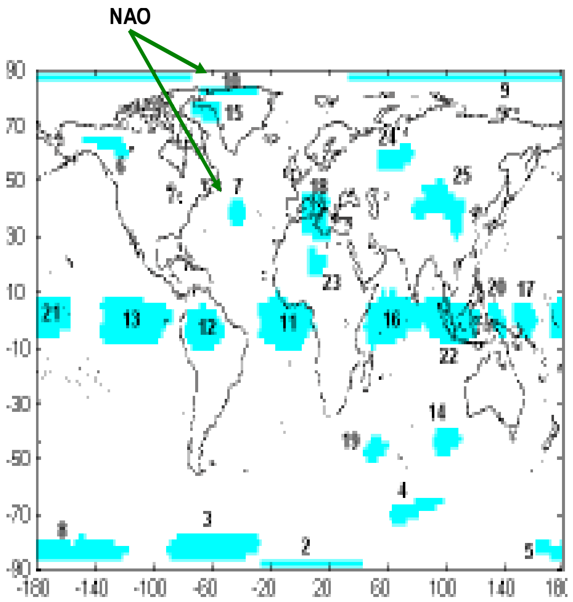


* N. H. Saji, B. N. Goswami, P. N. Vinayachandran and T. Yamagata, "A dipole mode in the tropical Indian Ocean," Nature 401, 360-363 (23 September 1999).

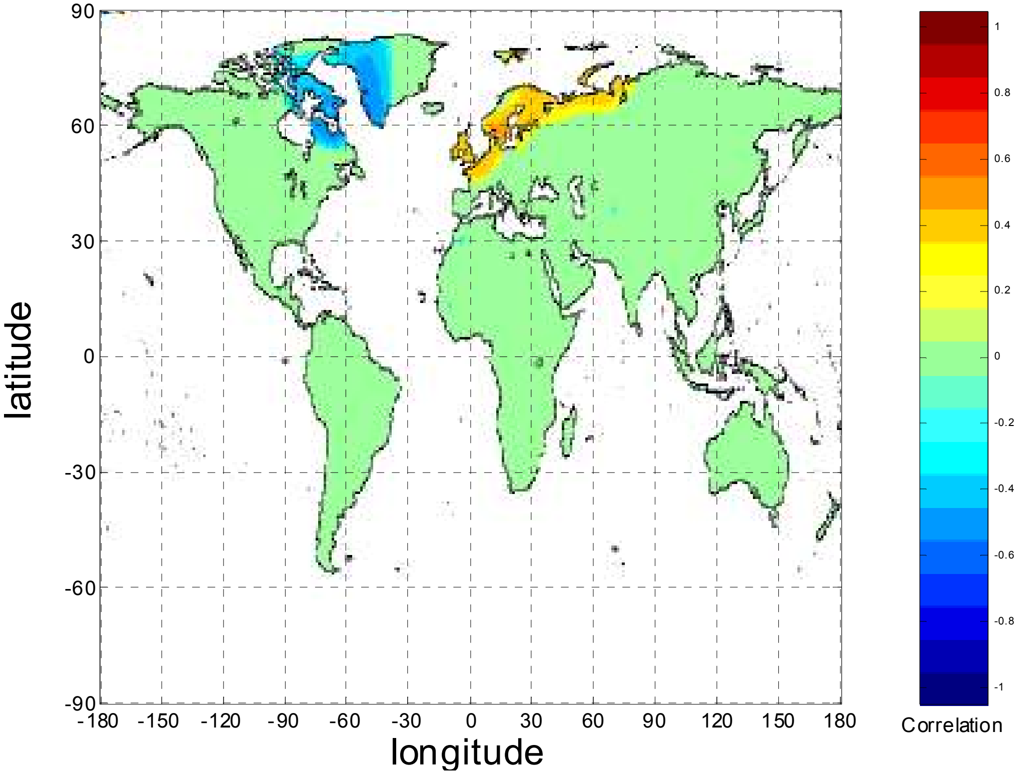
Clustering Challenges

Moving Clusters in Space and Time

- Most well-known indices based on data collected at fixed land stations.
- NAO computed as the normalized difference between SLP at a pair of land stations in the Arctic and the subtropical Atlantic regions of the North Atlantic Ocean



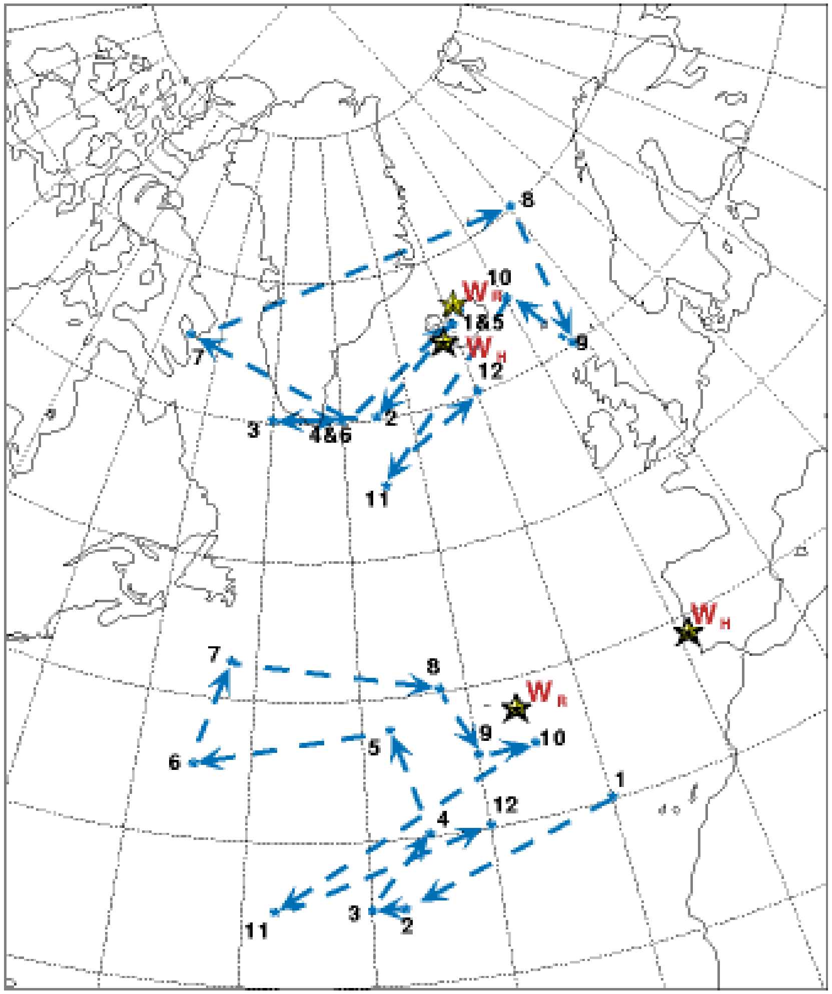
Correlation Between NAO and Land Temperature (>0.3)



Moving Clusters in Space and Time

- However, underlying phenomenon may not occur at exact location of the land station. e.g. NAO

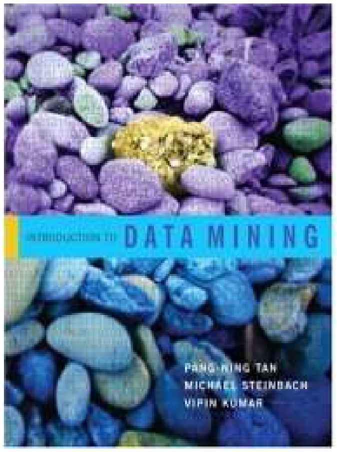
- **Challenge**: Given sensor readings for SLP at different points in the ocean, how to identify clusters of low/high pressure points that may move with space and time.

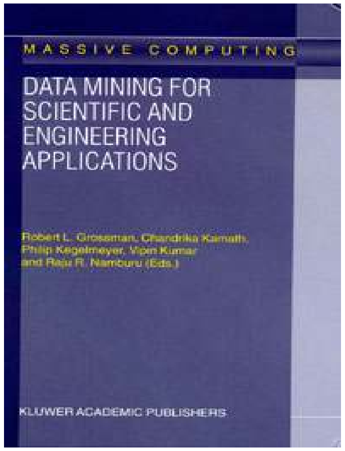


Source: Portis et al, Seasonality of the NAO, AGU Chapman Conference, 2000.

Bibliography

- Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison-Wesley April 2006
- Introduction to Parallel Computing, (2nd Edition) by A. Grama, A. Gupta, G. Karypis, and Vipin Kumar. Addison-Wesley, 2003
- Data Mining for Scientific and Engineering Applications, edited by R. Grossman, C. Kamath, W. P. Kegelmeyer, V. Kumar, and R. Namburu, Kluwer Academic Publishers, 2001
- J. Han, R. B. Altman, V. Kumar, H. Mannila, and D. Pregibon, "Emerging Scientific Applications in Data Mining", Communications of the ACM
  Volume 45, Number 8, pp 54-58, August 2002




MASSIVE COMPUTING
DATA MINING FOR SCIENTIFIC AND ENGINEERING APPLICATIONS
Robert L. Grossman, Chandrika Kamath, Philip Kegelmeyer, Vipin Kumar and Raju R. Namburu (Eds.)
KLUWER ACADEMIC PUBLISHERS

Bibliography: Journal Publications

- C. Potter, S. Klooster, P. Tan, M. Steinbach, V. Kumar and V. Genovese, "Variability in terrestrial carbon sinks over two decades: Part 2 — Eurasia", Global and Planetary Change, Volume 49, Issues 3-4, December 2005, Pages 177-186.
- C. Potter, P. Tan, V. Kumar, C. Kucharik, S. Klooster, V. Genovese, W. Cohen, S. Healey. "Recent History of Large-Scale Ecosystem Disturbances in North America Derived from the AVHRR Satellite Record", Ecosystems, 8(7), 808-824. 2004.
- Potter, C., Tan, P., Steinbach, M., Klooster, S., Kumar, V., Myneni, R., Genovese, V., 2003. Major disturbance events in terrestrial ecosystems detected using global satellite data sets. *Global Change Biology,* July, 2003.
- Potter, C., Klooster, S. A., Myneni, R., Genovese, V., Tan, P., Kumar,V. 2003. Continental scale comparisons of terrestrial carbon sinks estimated from satellite data and ecosystem modeling 1982-98. *Global and Planetary Change.*
- Potter, C., Klooster, S. A., Steinbach, M., Tan, P., Kumar, V., Shekhar, S., Nemani, R., Myneni, R., 2003. Global teleconnections of climate to terrestrial carbon flux*. Geophys J. Res.- Atmospheres*.
- Potter, C., Klooster, S., Steinbach, M., Tan, P., Kumar, V., Myneni, R., Genovese, V., 2003. Variability in Terrestrial Carbon Sinks Over Two Decades: Part 1 – North America. *Geophysical Research Letters.*
- Potter, C. Klooster, S., Steinbach, M., Tan, P., Kumar, V., Shekhar, S. and C. Carvalho, 2002. Understanding Global Teleconnections of Climate to Regional Model Estimates of Amazon Ecosystem Carbon Fluxes. *Global Change Biology*.
- Potter, C., Zhang, P., Shekhar, S., Kumar, V., Klooster, S., and Genovese, V., 2002. Understanding the Controls of Historical River Discharge Data on Largest River Basins.

Bibliography: Conference/Workshop Publications

- Michael Steinbach, Pang-Ning Tan, Shyam Boriah, Vipin Kumar, Steven Klooster, and Christopher Potter, "The Application of Clustering to Earth Science Data: Progress and Challenges", Proceedings of the 2nd NASA Data Mining Workshop, May 2006.
- Vipin Kumar, Michael Steinbach, Pusheng Zhang, Shashi Shekhar, Pang-Ning Tan, Christopher Potter, and Steven Klooster, "Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining", NASA Earth Science Technology Conference 2004.
- Steinbach, M., Tan, P. Kumar, V., Potter, C. and Klooster, S., 2003. Discovery of Climate Indices Using Clustering, KDD 2003, Washington, D.C., August 24-27, 2003.
- Zhang, P., Huang, Y., Shekhar, S., and Kumar, V., 2003. Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries , Proc. of the 8th Intl. Symp. on Spatial and Temporal Databases (SSTD '03)
- Zhang, P., Huang, Y., Shekhar, S., and Kumar, V., 2003. Correlation Analysis of Spatial Time Series Datasets: A Filter-And-Refine Approach, Proc. of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '03)
- Ertoz, L., Steinbach, M., and Kumar, V., 2003. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, Proc. of Third SIAM International Conference on Data Mining.
- Tan, P., Steinbach, M., Kumar, V., Potter, C., Klooster, S., and Torregrosa, A., 2001. Finding Spatio-Temporal Patterns in Earth Science Data, KDD 2001 Workshop on Temporal Data Mining, San Francisco
- Kumar, V., Steinbach, M., Tan, P., Klooster, S., Potter, C., and Torregrosa, A., 2001. Mining Scientific Data: Discovery of Patterns in the Global Climate System, Proc. of the 2001 Joint Statistical Meeting, Atlanta