High Performance Data Mining - Application for Discovery of Patterns in the Global Climate System

Vipin Kumar

University of Minnesota

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

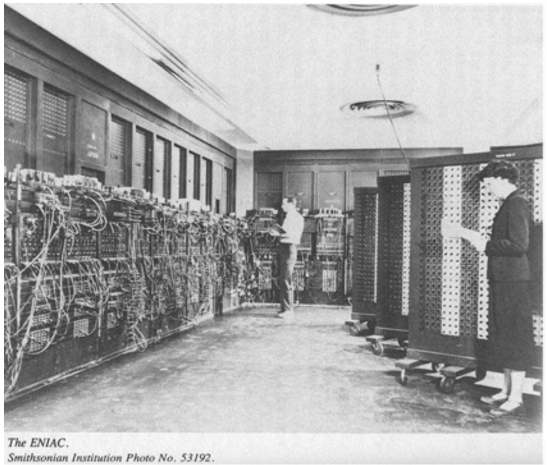
Collaborators and Group Members:

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

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

Research funded by ISET-NOAA, NSF and NASA

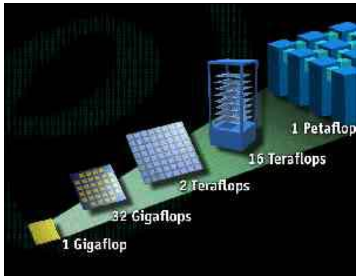
Progress in HPC - past 6 decades



ENIACS – 1945

- 100 K Hz
- 5 K Additions/second
- 357 Multiplications/second

- CPU power increasing by a factor of 30-100 every decade
- Multi-Giga Hz, multi-Gigabyte, multi-core CPUs are commodity
- Teraflops computers are common
- Petaflops scale computing within reach



IBM Blue Gene/L

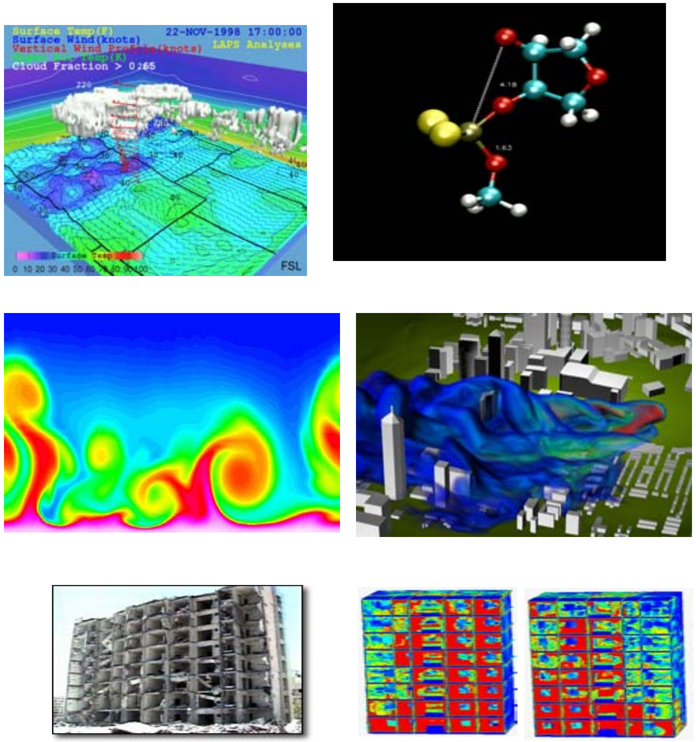
EKA (HP Cluster Platform 3000BL) - Computational Research Laboratories

Jaguar - Cray XT4/XT3 - Oak Ridge National Laboratory

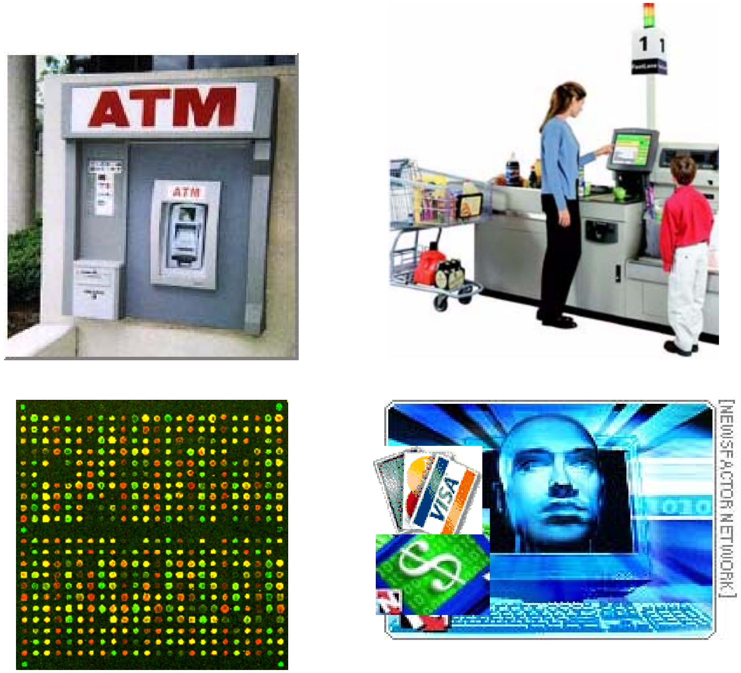
Applications Drive the Technology

“I think there is world market for maybe 5 computers”

- Thomas Watson Sr.



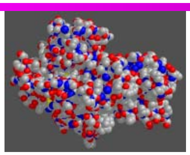
Scientific Computing

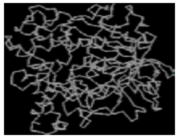


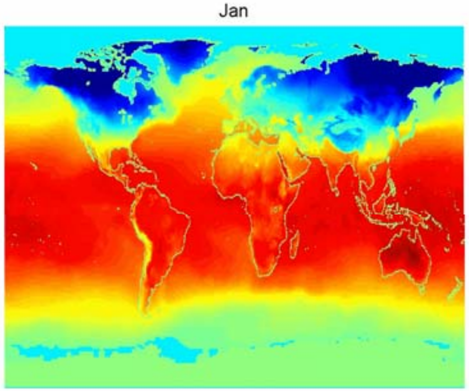
Data Driven Computing

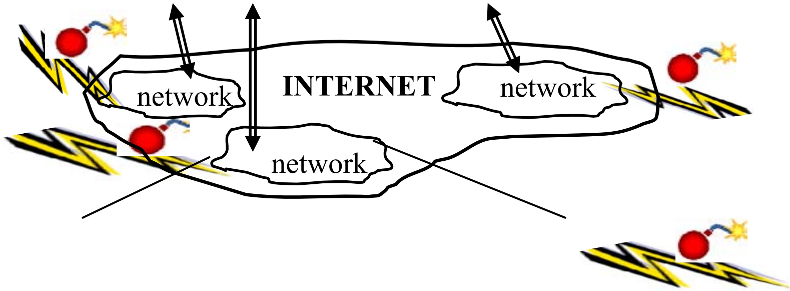
Data Mining **- A Driver for High-Performance Computing**

- Lots of data being collected in commercial and scientific world
- Strong competitive pressure to extract and use the information from the data
- Scaling of data mining to large data requires HPC
- Data and/or computational resources needed
  for analysis are often distributed
- Sometimes the choice is distributed data mining or no data mining
  - Ownership, privacy, security issues



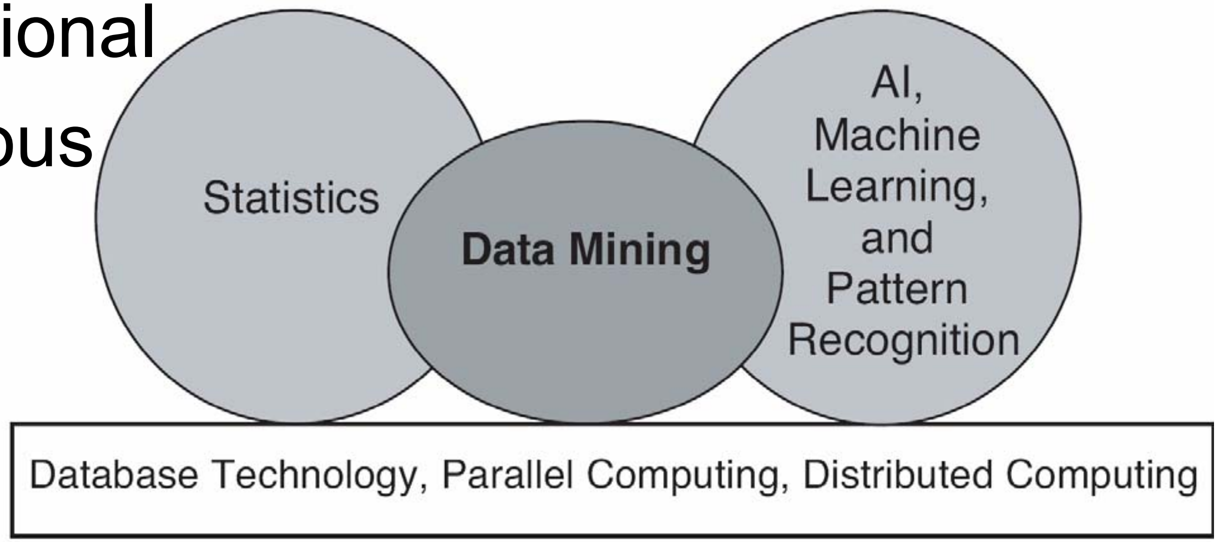




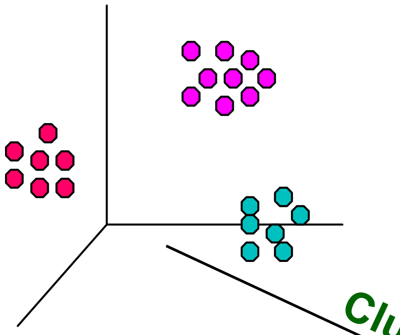


Origins of Data Mining

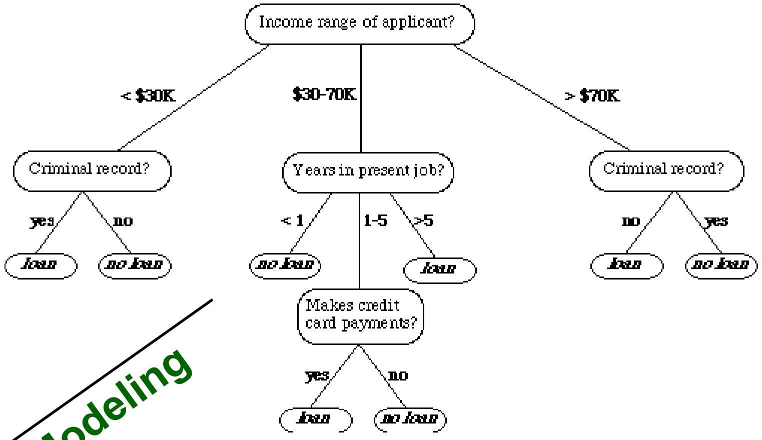
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed


Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

Data Mining Tasks



Clustering



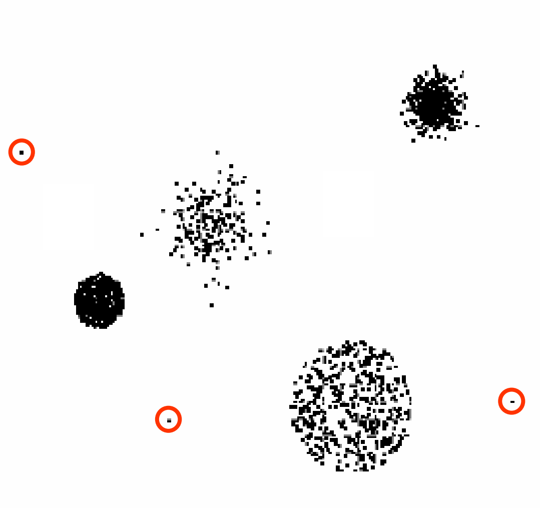
Predictive Modeling

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

Anomaly Detection

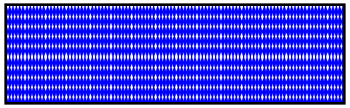


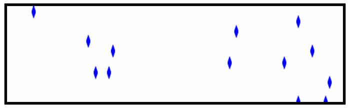
Basic Operations in many Data Mining Kernels

- Counting
  - Given a set of data records, count types of different categories to build a contingency table Count the occurrence of a set of items in a set of transactions
- Distance/Similarity Computations
  - Given a set of data records, perform distance/similarity computations
- Linear Algebra operations
  - SVD, PCA, etc

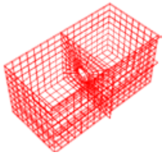
General Issues and Challenges in Parallel Data Mining

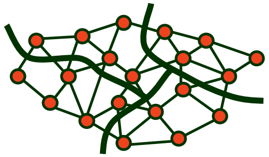
- Dense vs. Sparse



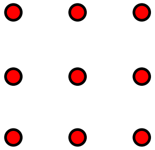


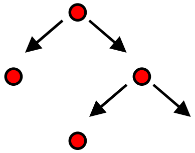
- Structured versus Unstructured





- Static vs. Dynamic

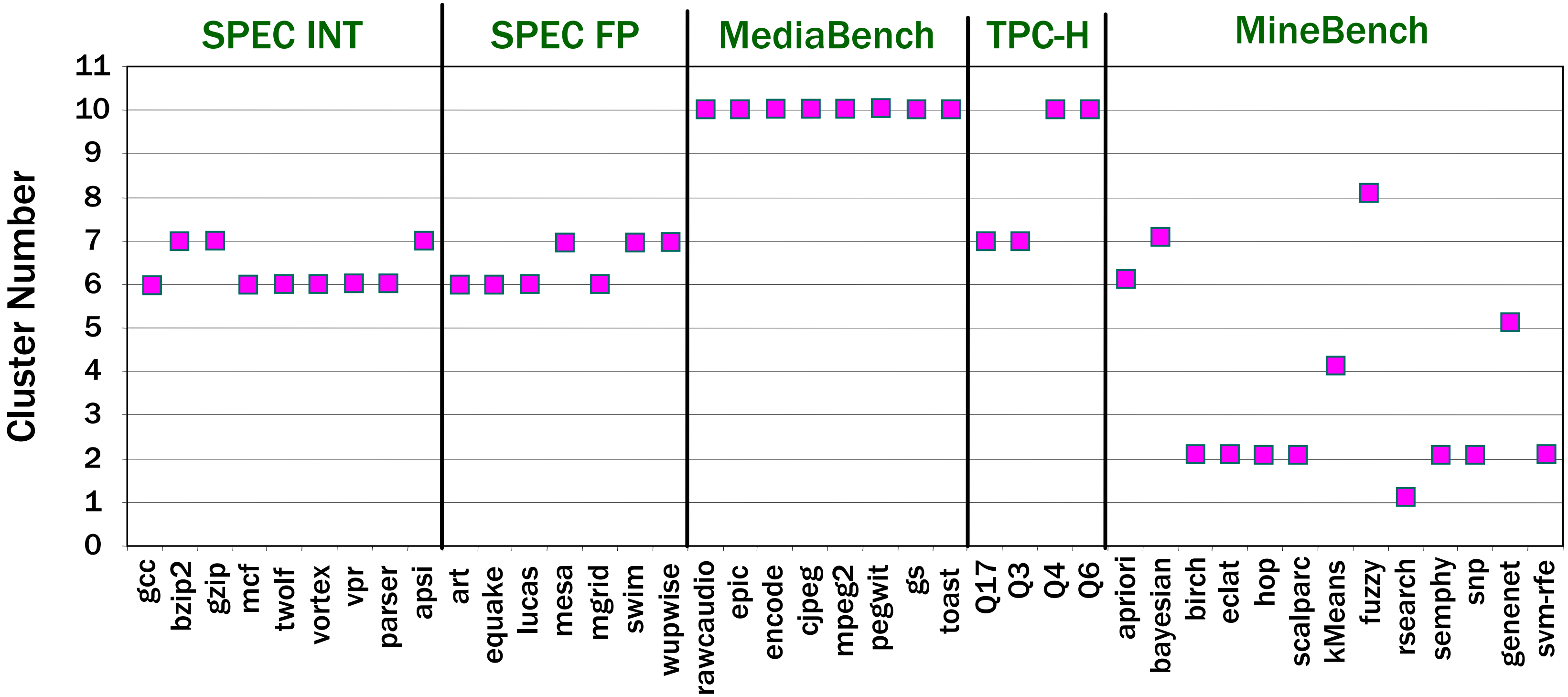




- Many data mining computations tend to be unstructured, sparse and dynamic
  - Data is often too large to fit in main memory
  - Spatial locality is critical
  - Many efficient DM algorithms require fast access to large hash tables

Are Data Mining applications similar to other workloads?

- Performance metrics of several benchmarks gathered from Vtune
  - Cache miss ratios, Bus usage, Page faults etc.
- Benchmark applications were grouped using Kohenen clustering to spot trends:



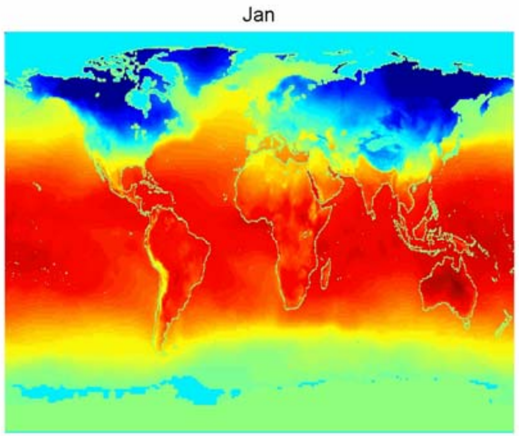
Reference: [Pisharath J., Zambreno J., Ozisikyilmaz B., Choudhary A., 2006]

Discovery of Climate Patterns from Global Data Sets

Science Goal: Understand global scale patterns in biosphere processes

Earth Science Questions:

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



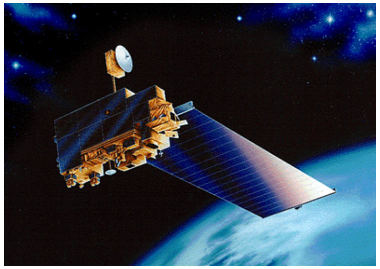
Monthly Average Temperature

Data sources:

- Weather observation stations
- High-resolution EOS satellites

  1982-2000 AVHRR at 1° x 1° resolution (~115kmx115km)
  2000-present MODIS at 250m x 250m resolution

- Model-based data from forecast and other models
- Data sets created by data fusion



Earth Observing System

Data Mining 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 large scale **ecological disasters** that cause sudden changes in greenness over extensive land areas

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

.

- Ecosystem disturbances can contribute to the current rise of CO2 in the atmosphere, with global climate implications
- In many remote locations, disturbances go undetected
- Satellite observations can help detect these disasters and help estimate their impact on the environment

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

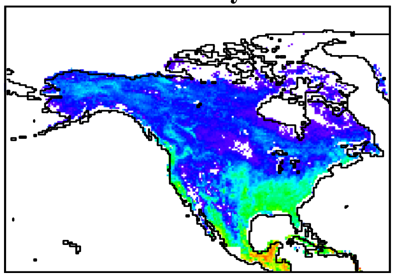
Detection of Ecosystem Disturbances

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

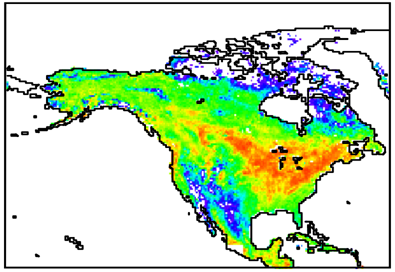
- Can be verified from independent records of such disturbances.

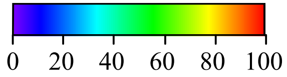
FPAR: Fraction absorbed of Photosynthetically Active Radiation by vegetation canopies

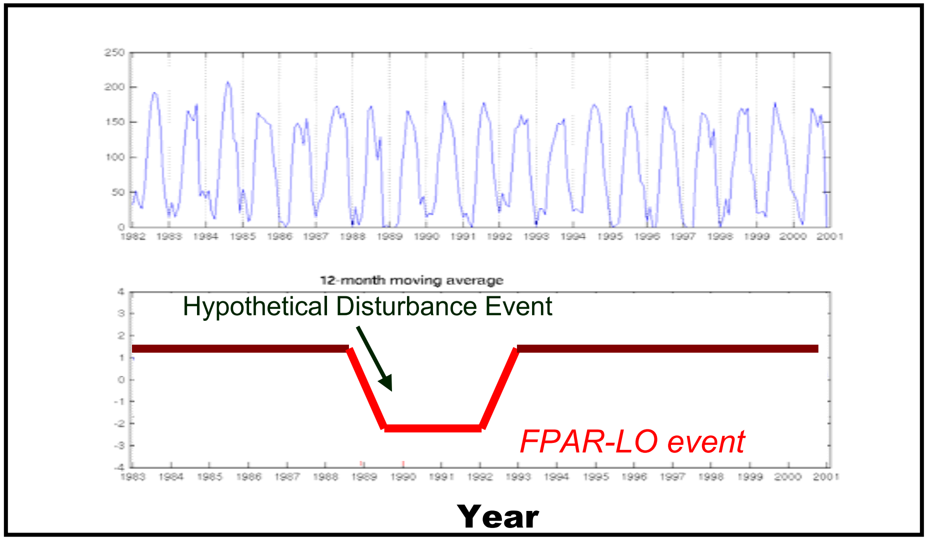
FPAR January 1998



FPAR July 1998



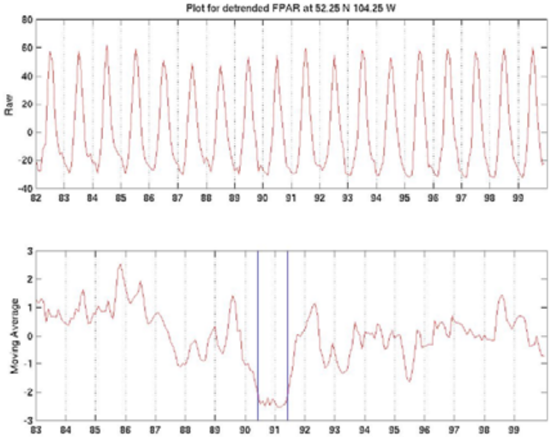




Potter, et al., "Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets", Global Change Biology, 9(7), 1005-1021, 2003.

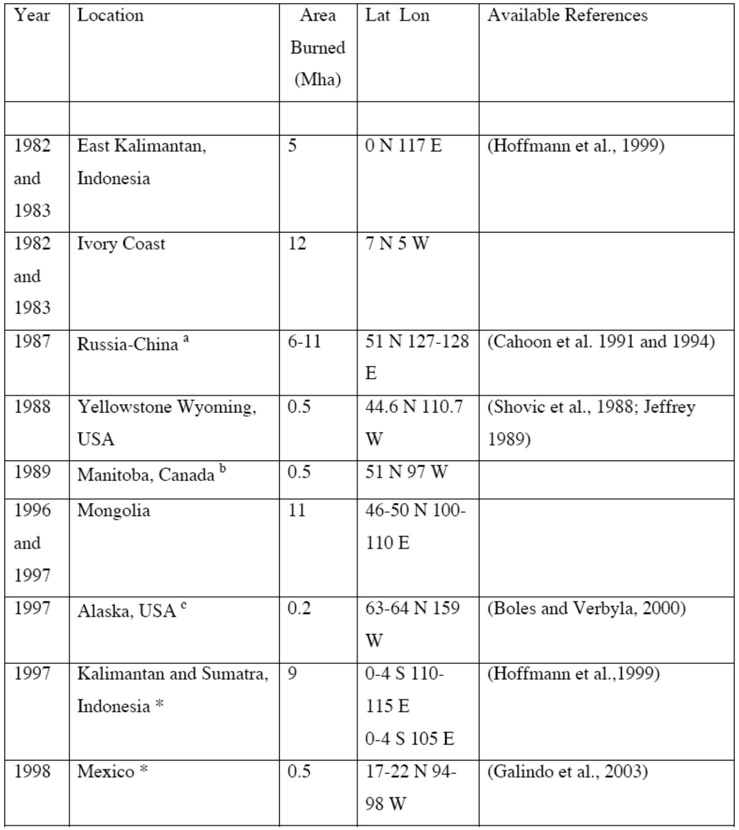
Verification of Disturbances: Fires

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.

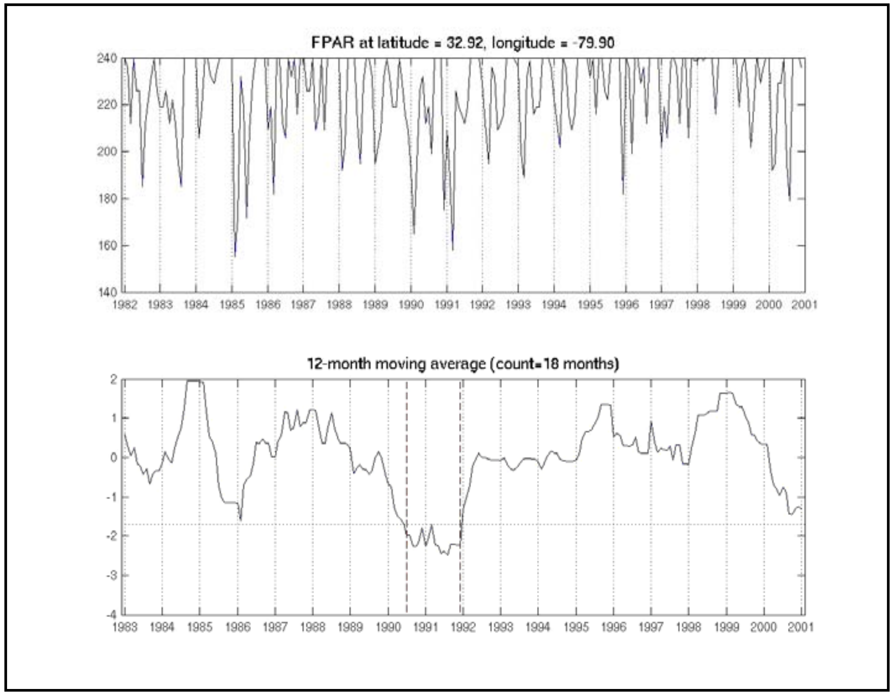


Year	Location	Area Burned (Mha)	Lat Lon	Available References
1982 and 1983	East Kalimantan, Indonesia	5	0 N 117 E	(Hoffmann et al., 1999)
1982 and 1983	Ivory Coast	12	7 N 5 W	
1987	Russia-China [a]	6-11	51 N 127-128 E	(Cahoon et al. 1991 and 1994)
1988	Yellowstone Wyoming, USA	0.5	44.6 N 110.7 W	(Shovic 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 *	9	0-4 S 110-115 E 0-4 S 105 E	(Hoffmann et al.,1999)
1998	Mexico *	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

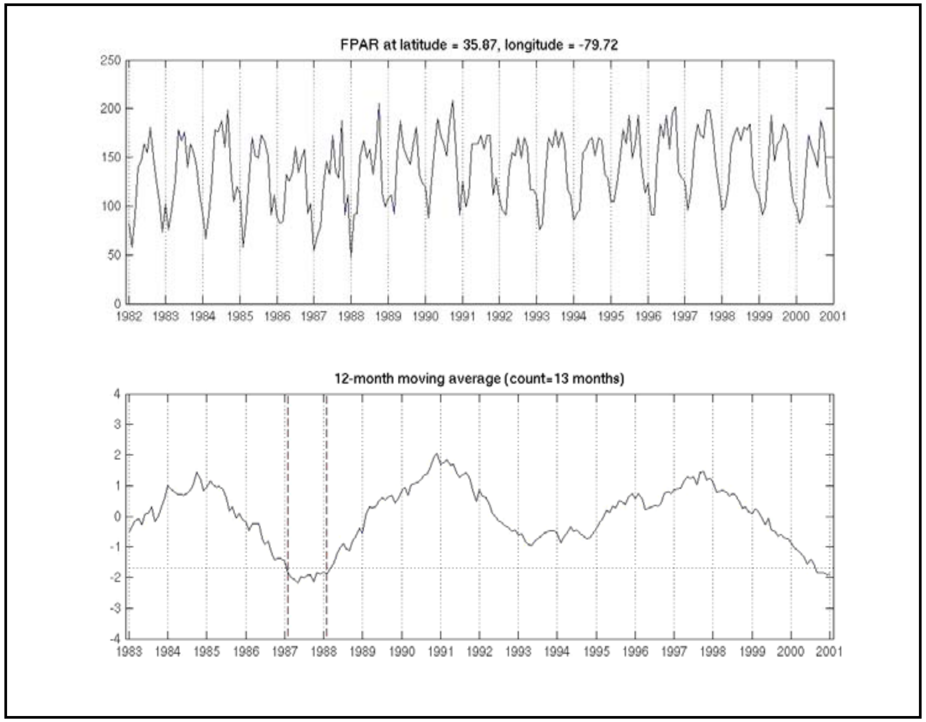


Hurricanes of the 1980s Detected as FPAR-LO Events

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 USA Drought 1986



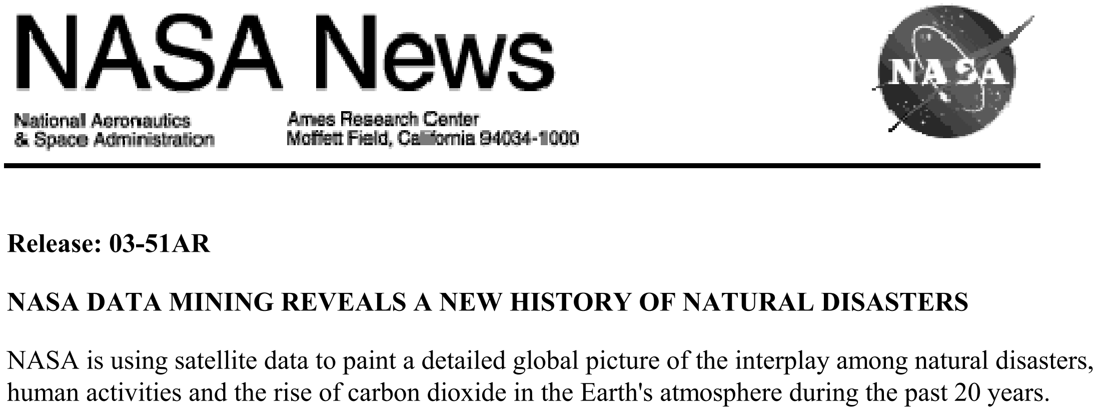
Major Droughts Detected as FPAR-LO Events

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

Study Results

Estimated 9 billion metric tons of carbon 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.

- For comparison, 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

Uniqueness of study:

- 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 land cover changes occur, and their impact on the environment

- E.g. Deforestation, Urbanization, Agricultural intensification

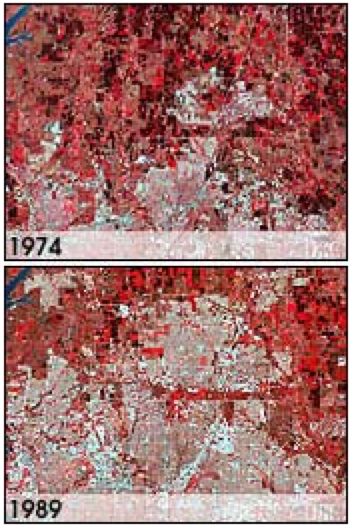
Motivation:

- 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).

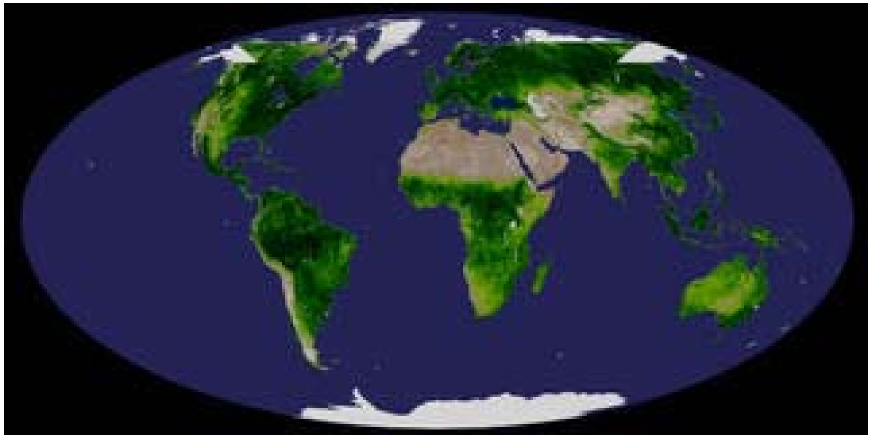
Urbanization. Between 1982 and 1992 19,000 sq. miles (equivalent to the area of half of Ohio) of rural cropland and wilderness were developed in the U.S.

The image on the right shows the expansion of Plano (near Dallas) between 1974 and 1989.



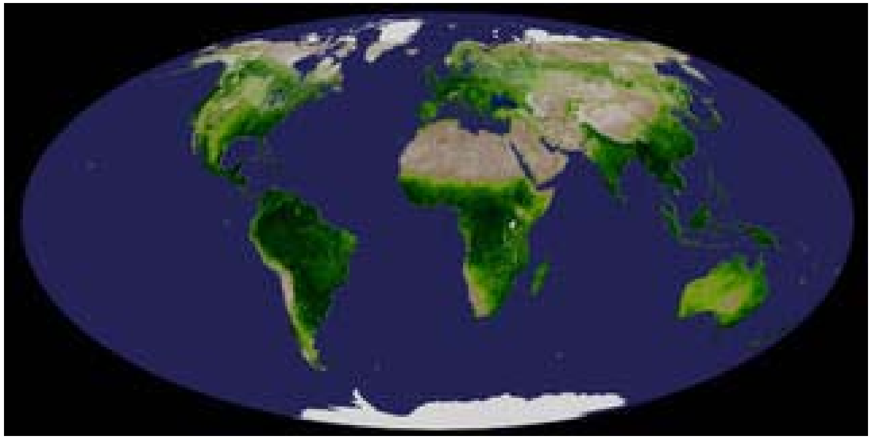
Source: NASA Earth Observatory

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

EVI Data for the Goa Area

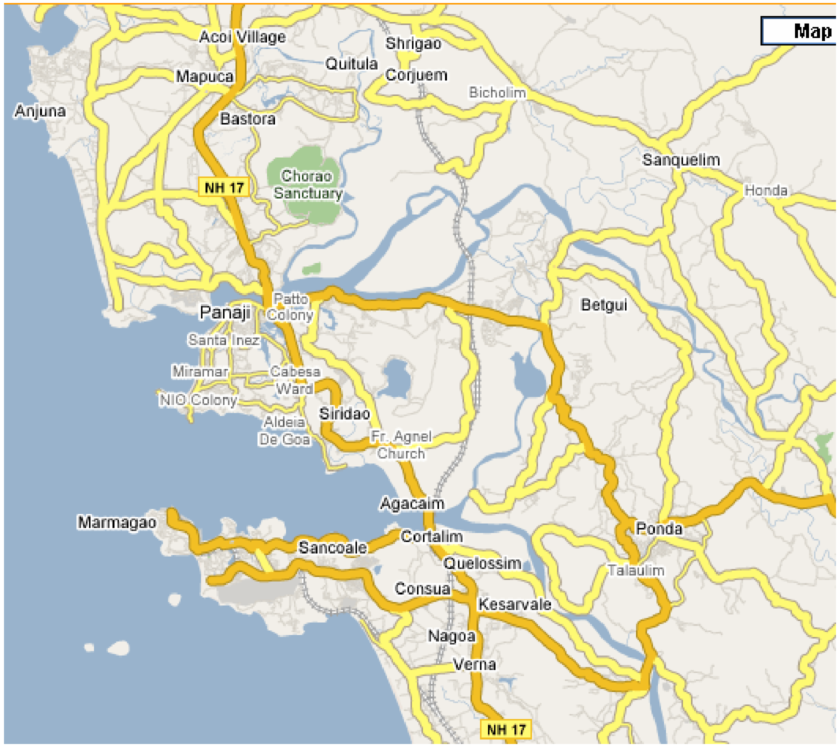
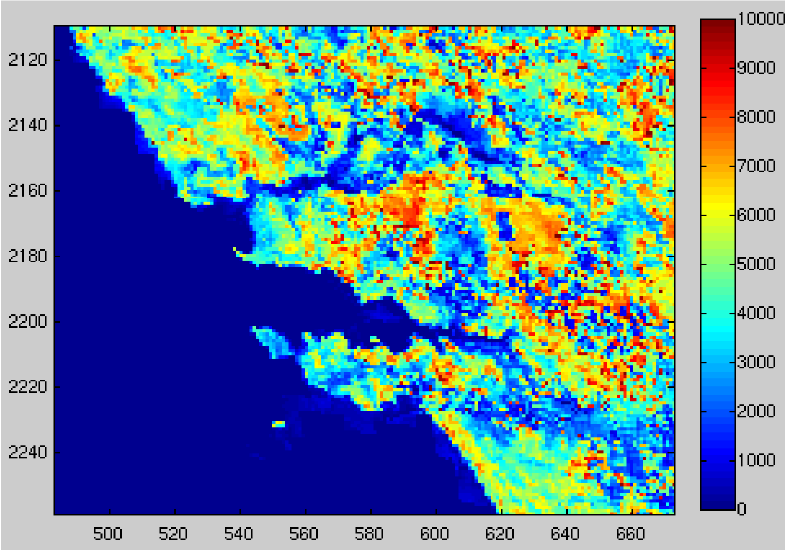
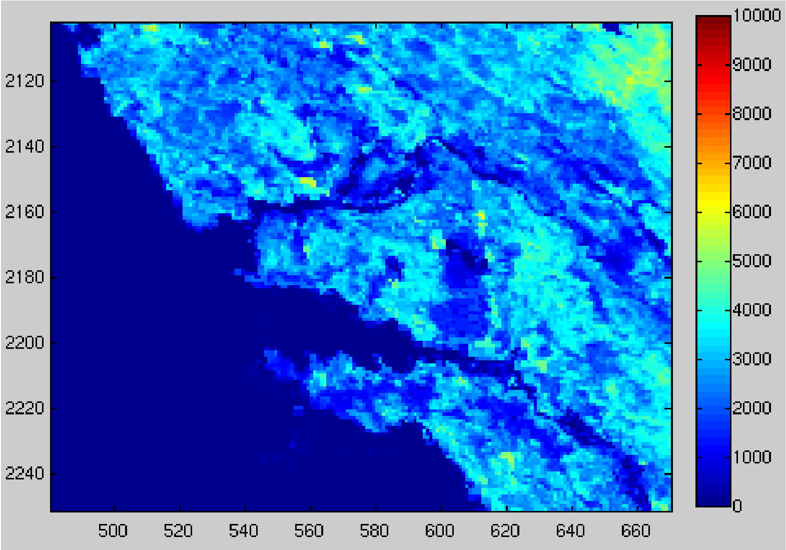


Image from Google Maps



June

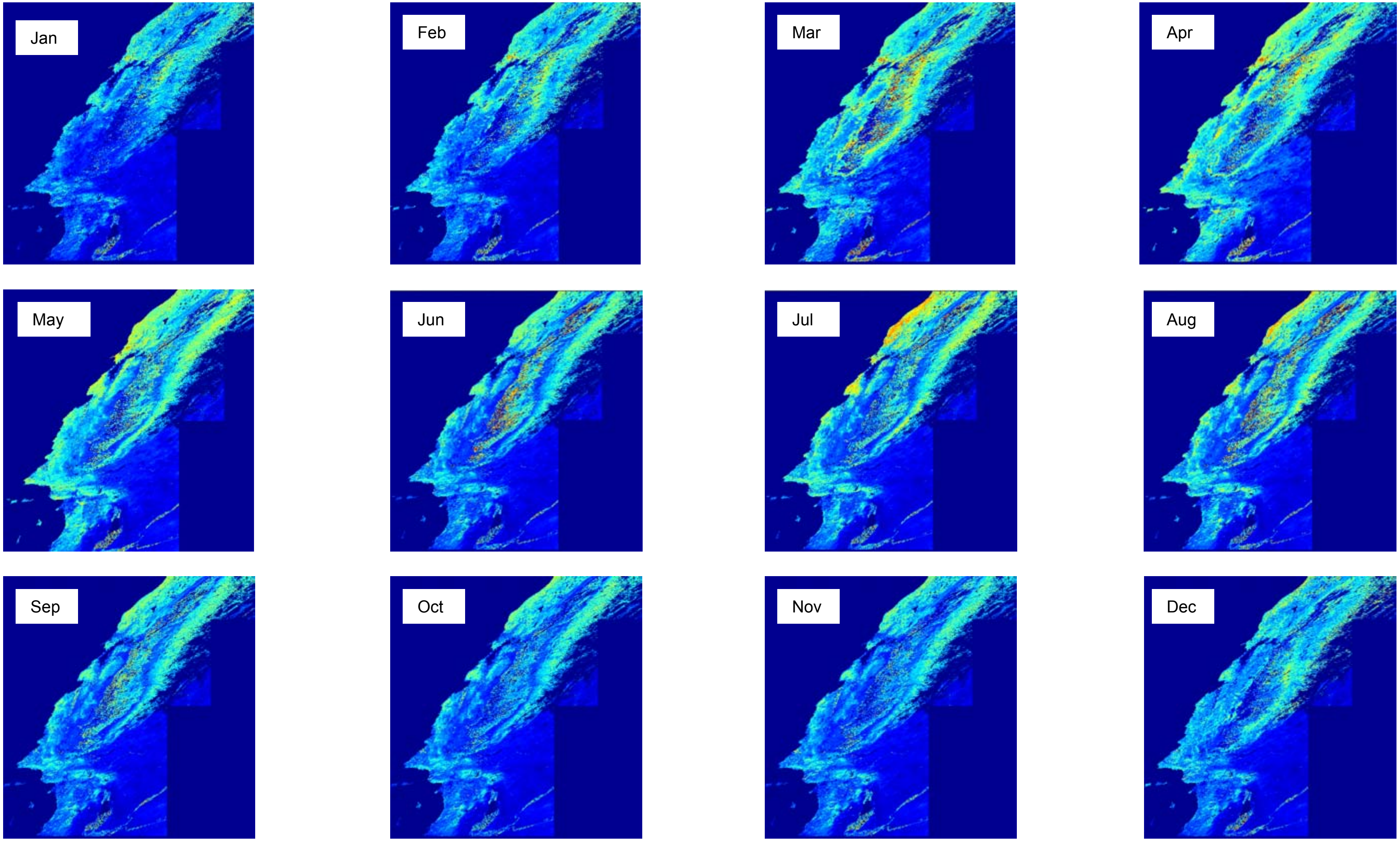
2000



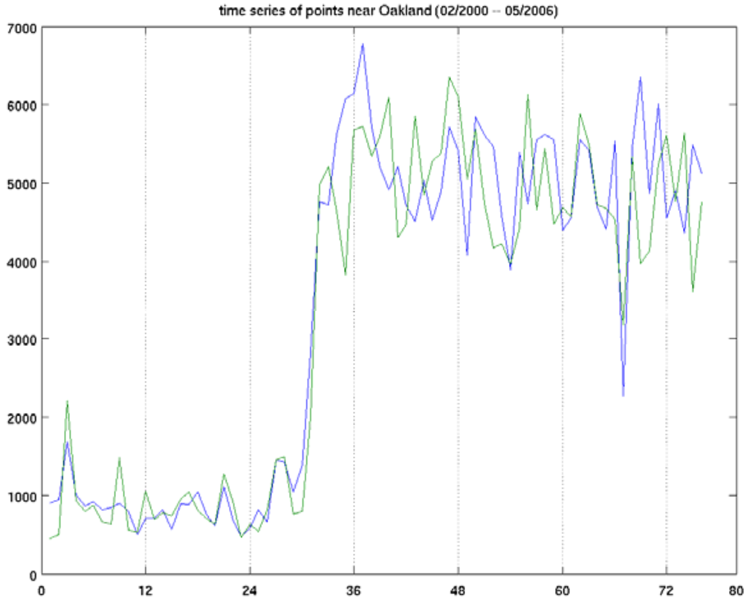
Jan

2001

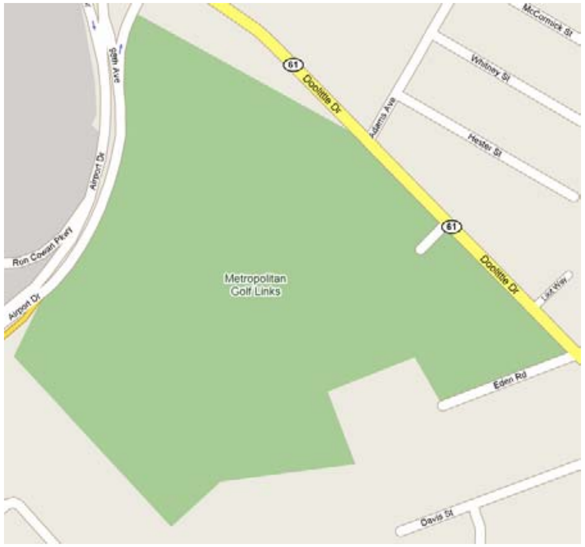
EVI in California from Jan-Dec 2001



Example of a Land Cover Change



- 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

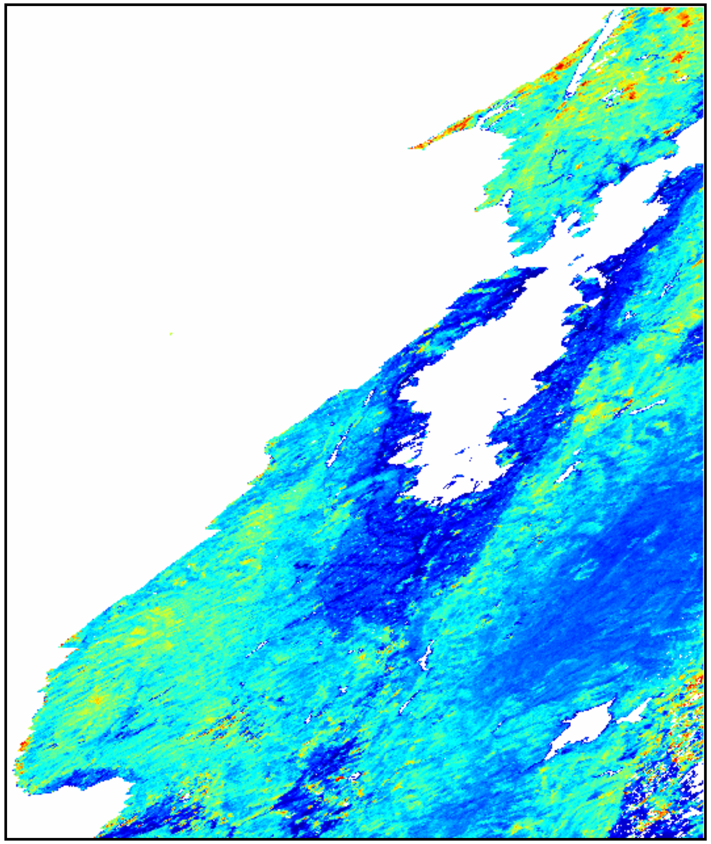
Limitations:

- 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: 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



EVI in Northern California for February 2002

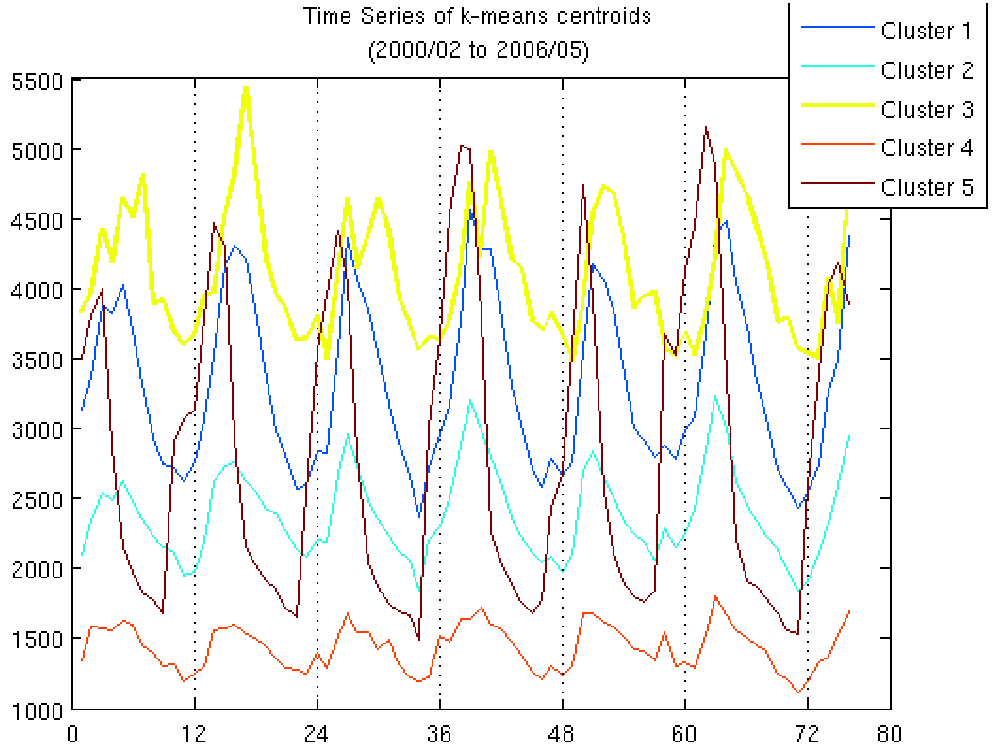
Study 1

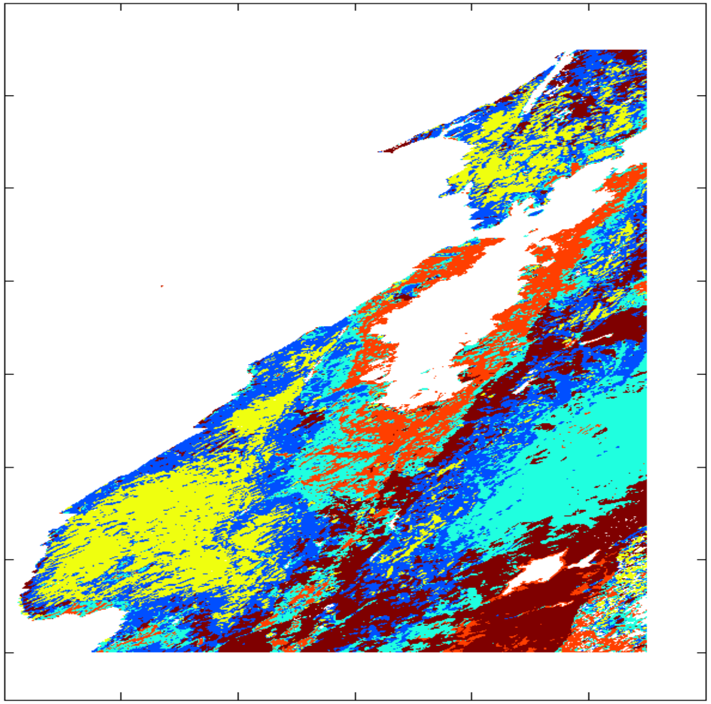
- San Francisco Bay area (180K points, about 100 miles x 50 miles)

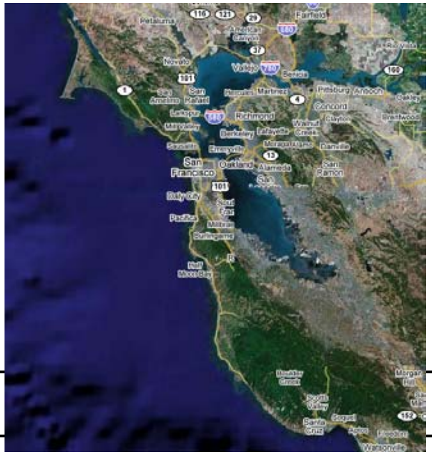
Study 2

- Entire state of California (5M points, about 800 miles x 200 miles)

High-level view of land cover: SF Bay area







Cluster 1 - **shrub** cover

Cluster 2 - **grass** cover

Cluster 3 - **evergreen tree** cover

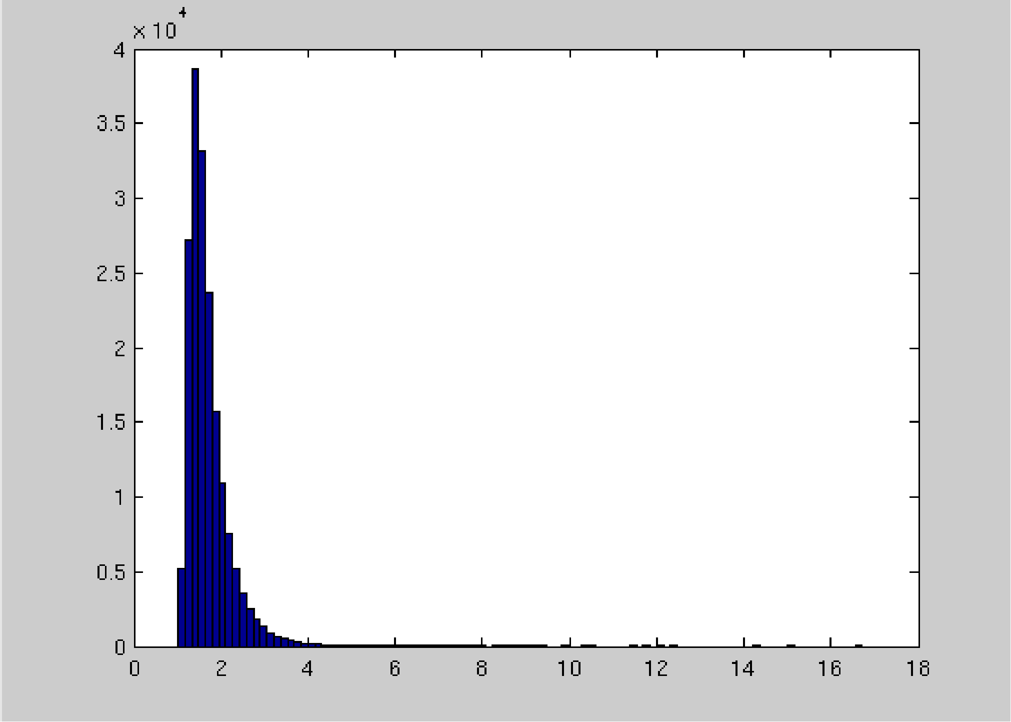
Cluster 4 - **urbanized** cover

Cluster 5 - **agricultural** cover

A new change detection technique

- **Key Idea**: exploit the major mode of behavior (seasonality) 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).

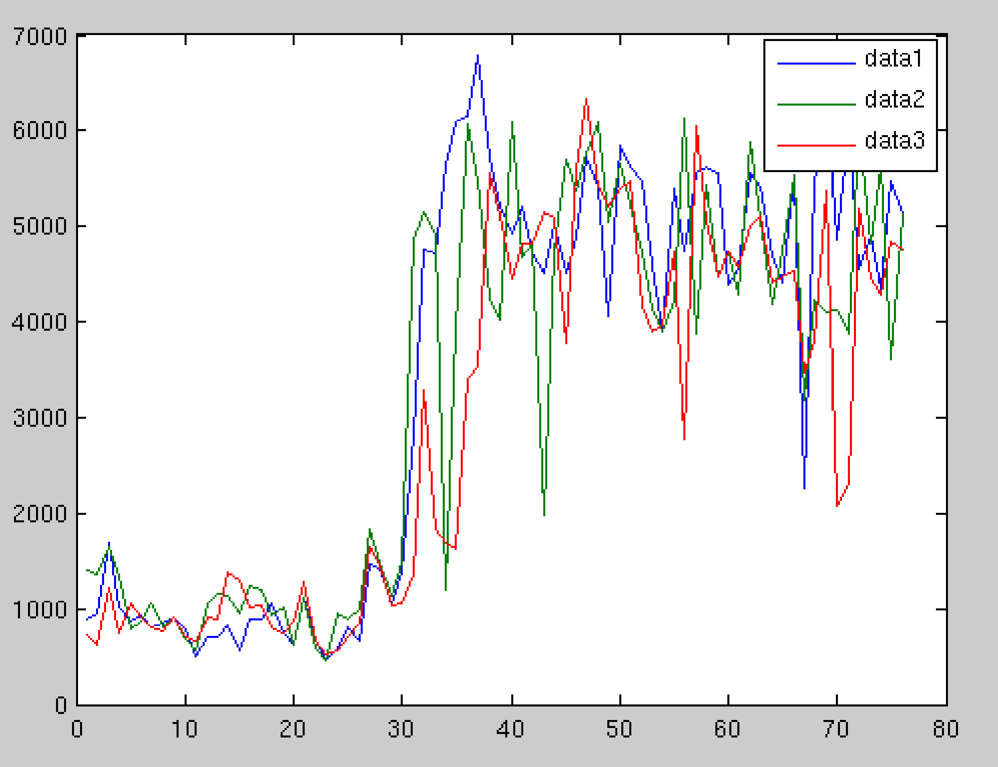
Results: Histogram of Scores

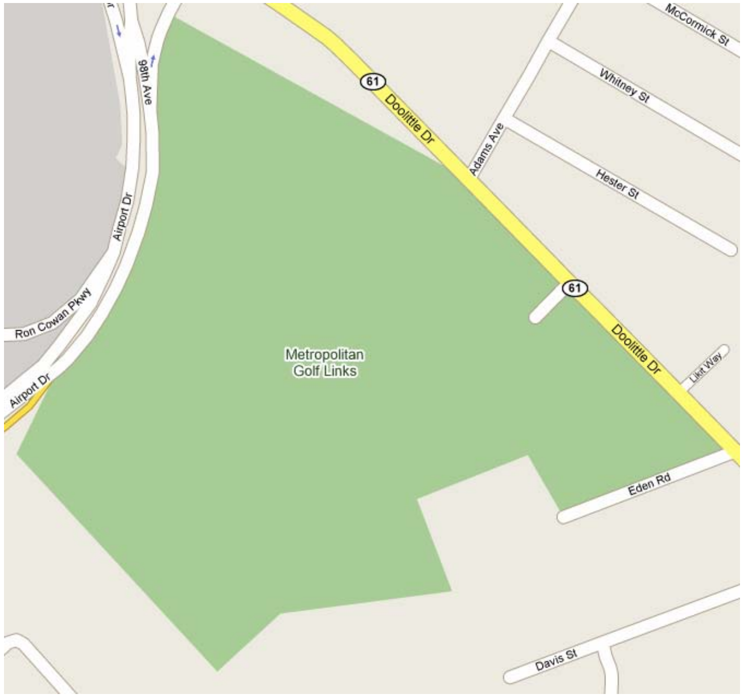


Histogram of all scores

- 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 farmland with changing harvest cycles.

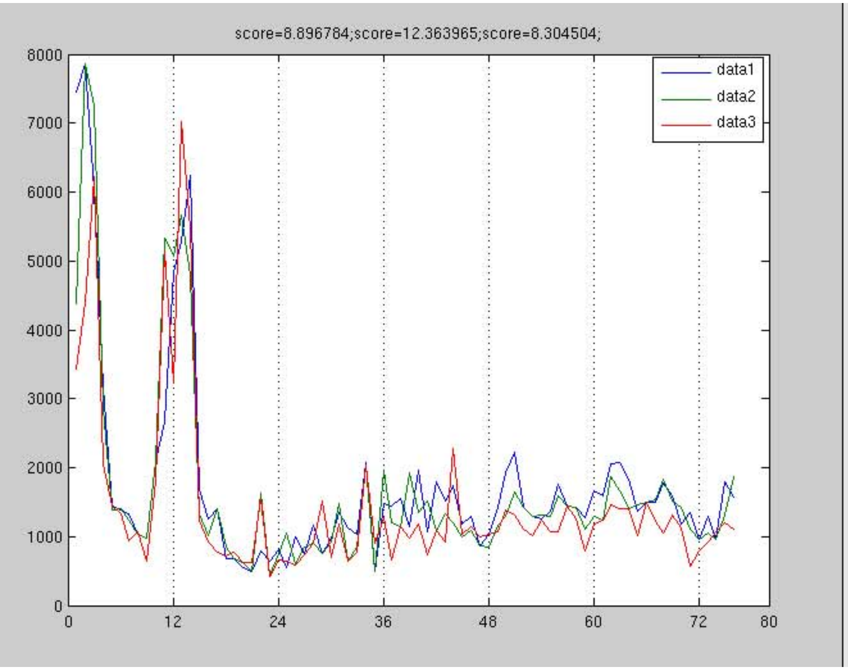
Top 3 scores: New Golf Course in Oakland

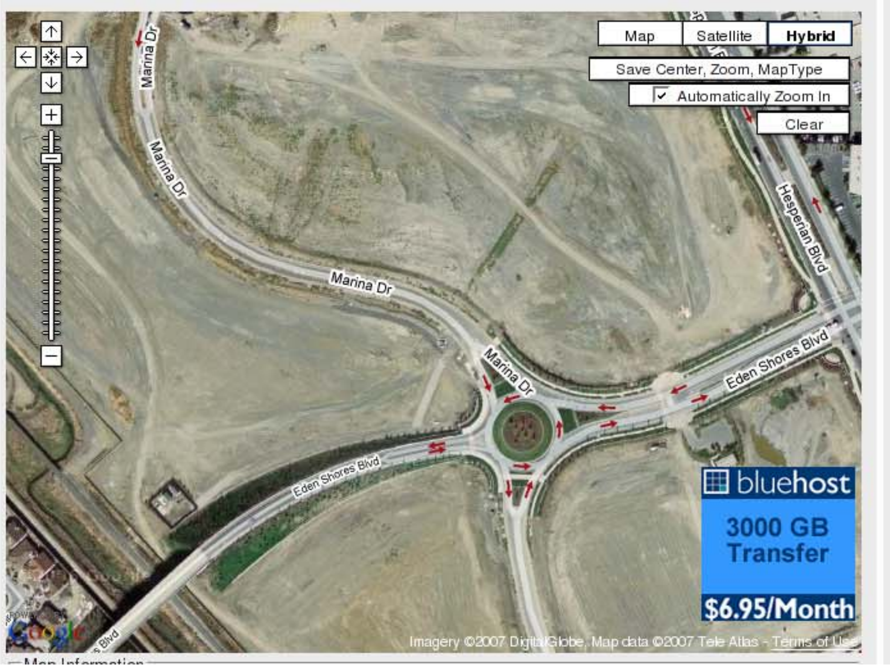




The top 3 points correspond to a golf course in Oakland. This golf course was built in 2003, which corresponds to the time step at which the time series exhibit a change.

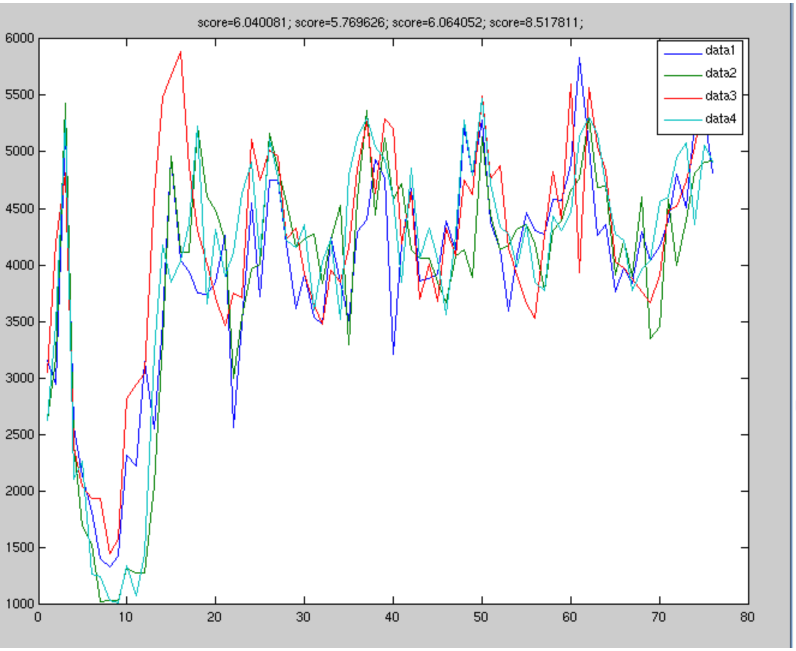
Results: Subdivision under construction in Hayward, CA

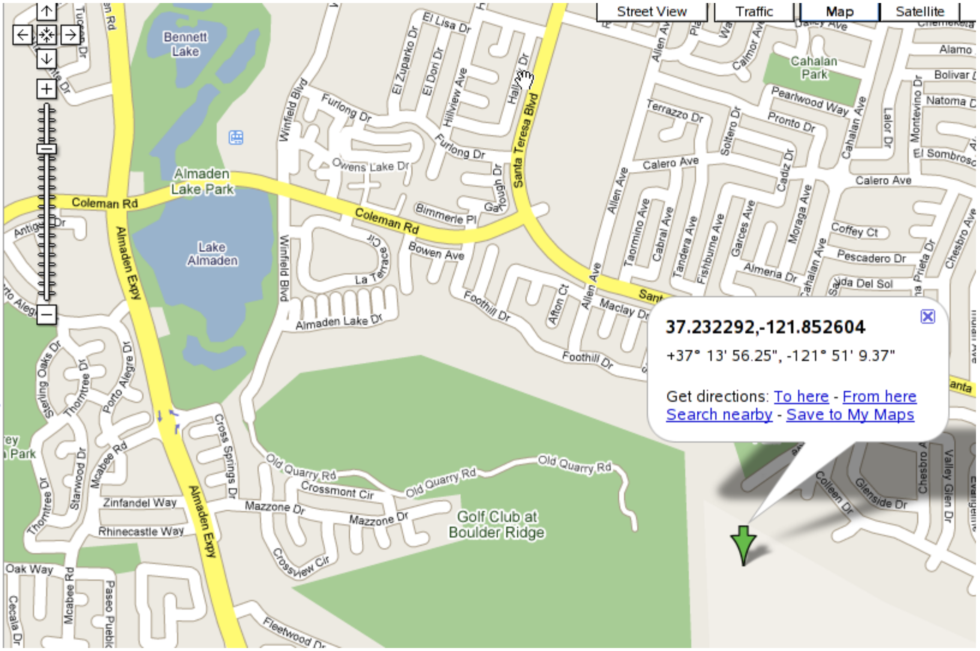




These 3 time series correspond to a subdivision under construction.

Results: New golf course in San Jose





37.232292,-121.852604

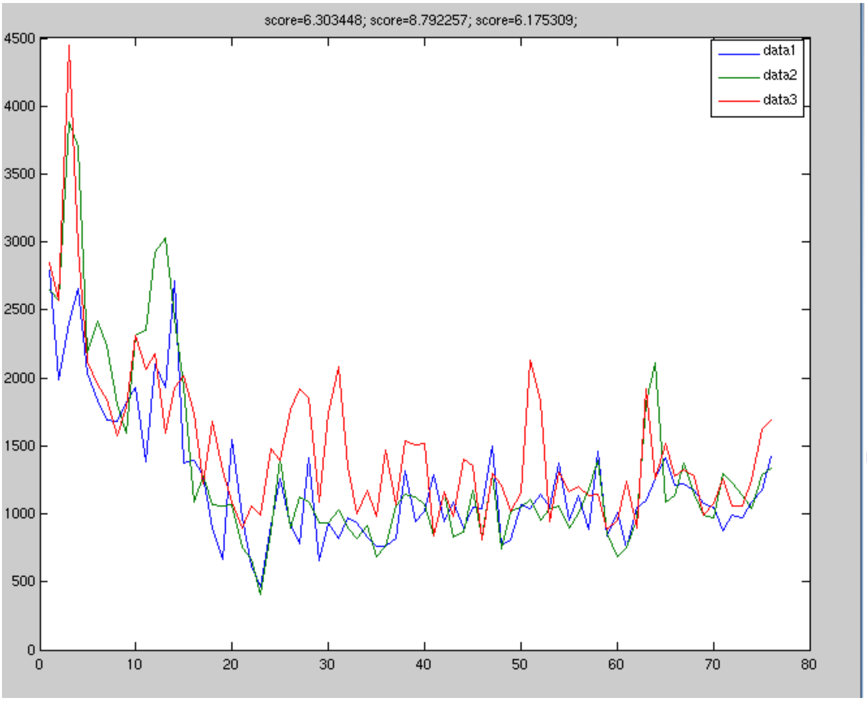
+37° 13' 56.25", -121° 51' 9.37"

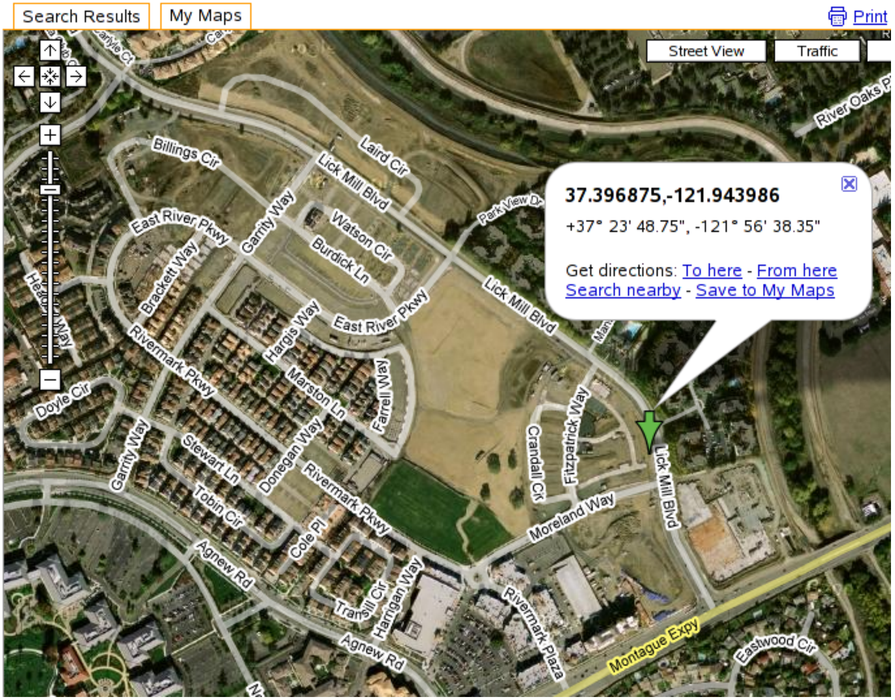
Get directions: To here - From here

Search nearby - Save to My Maps

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

Results: New subdivision in Santa Clara





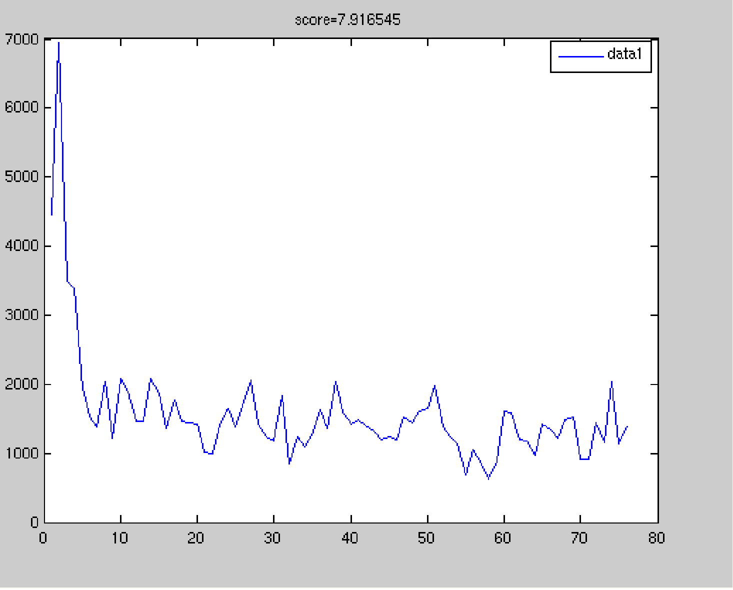
37.396875,-121.943986

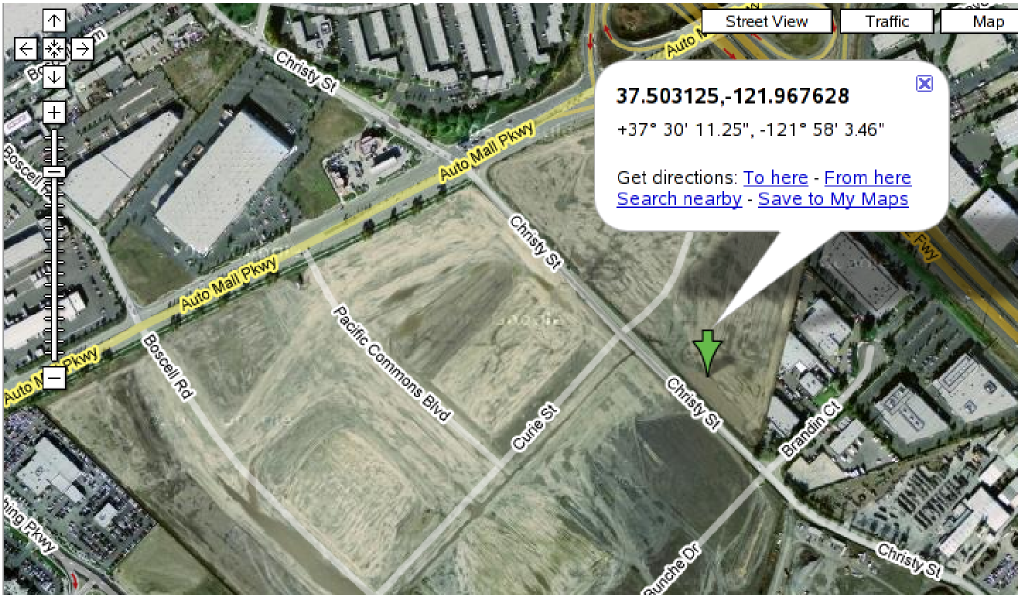
+37° 23' 48.75", -121° 56' 38.35"

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

Subdivision built in 2002

Results: New Shopping Area in Fremont



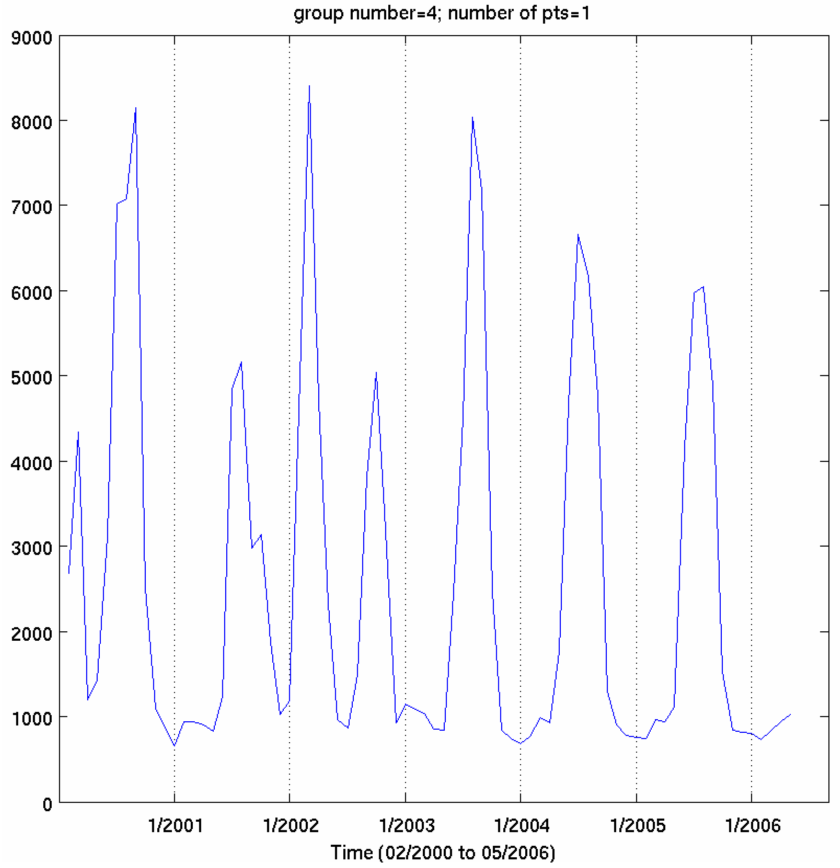

37.503125,-121.967628

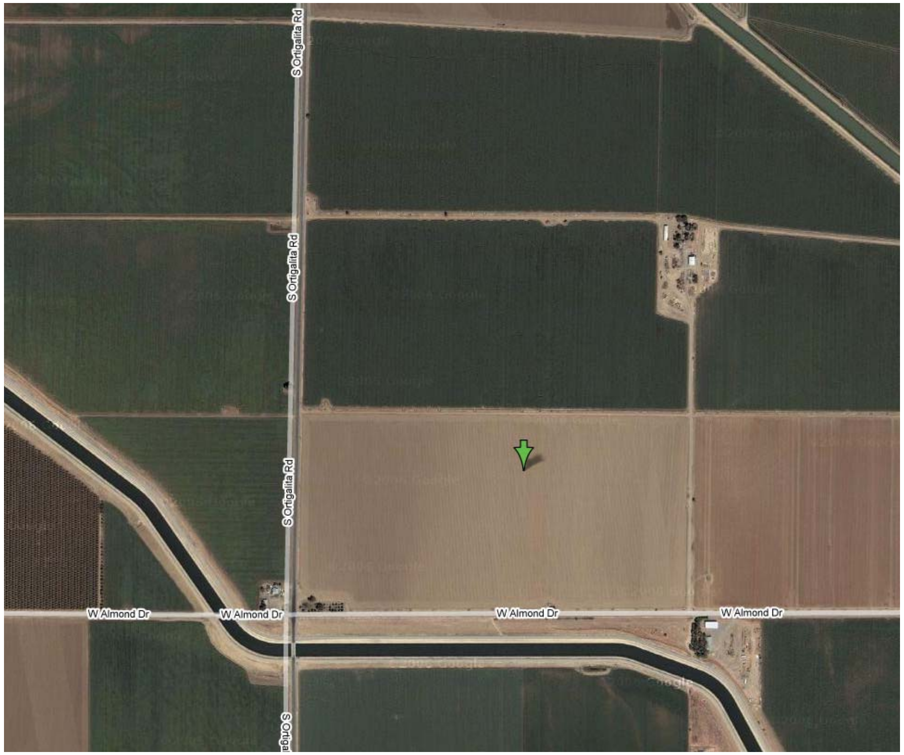
+37° 30' 11.25", -121° 58' 3.46"

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

Construction of Pacific Commons shopping area in Fremont, CA

Farmland





Farmland points such as this one mislead the algorithm. The changing crop patterns appear to be changes, but are not really changes of the type we are looking for. We will need to refine our technique to handle points such as this one.

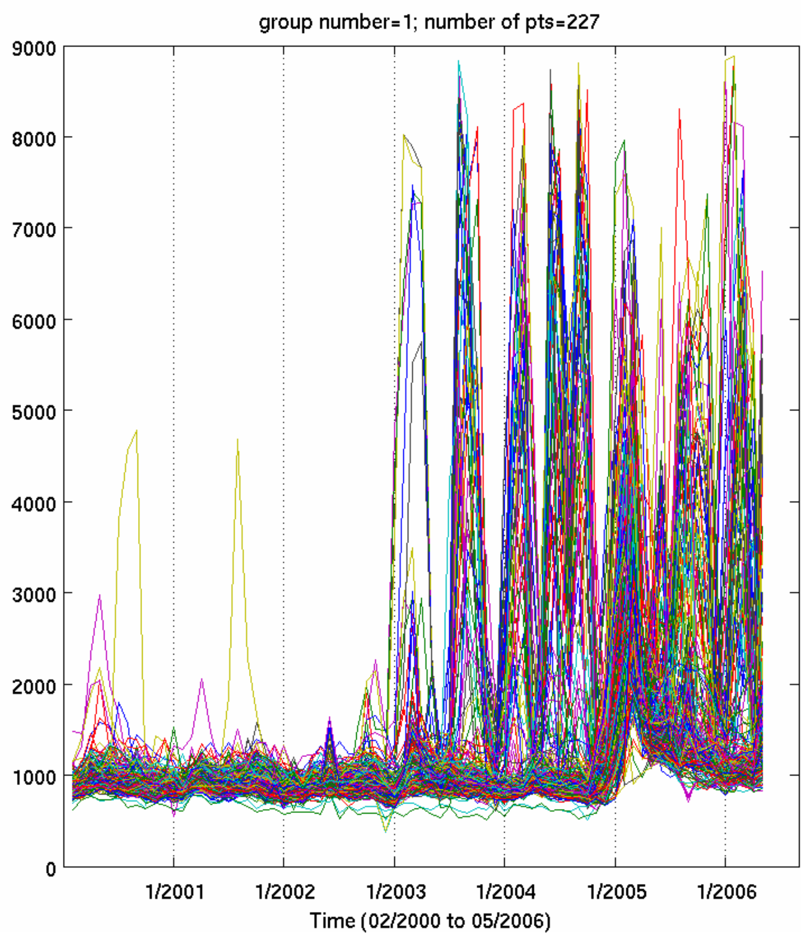
Study 2: Entire California

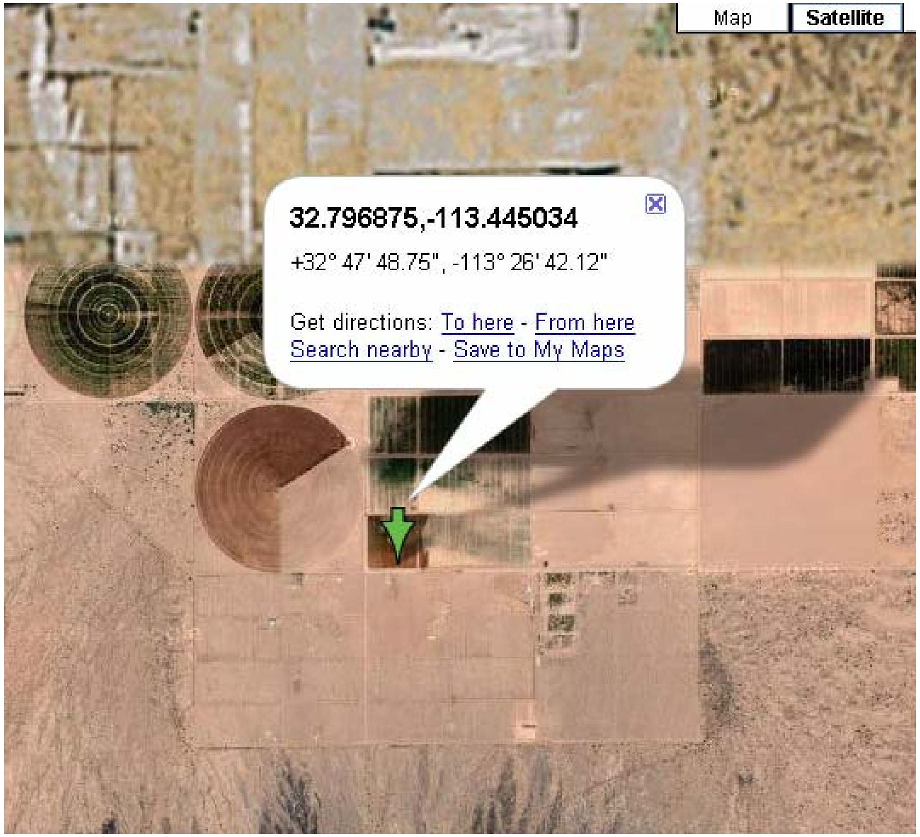
- Data has 5,165,205 locations, an increase of about 30-fold over the Bay Area data
- After applying our algorithm, 2,833 locations with change points are detected at a high threshold
- The larger data has more types of changes:
  - Desert to farmland
  - Desert to golf courses
  - Farmland to housing subdivisions

  ..

  ..

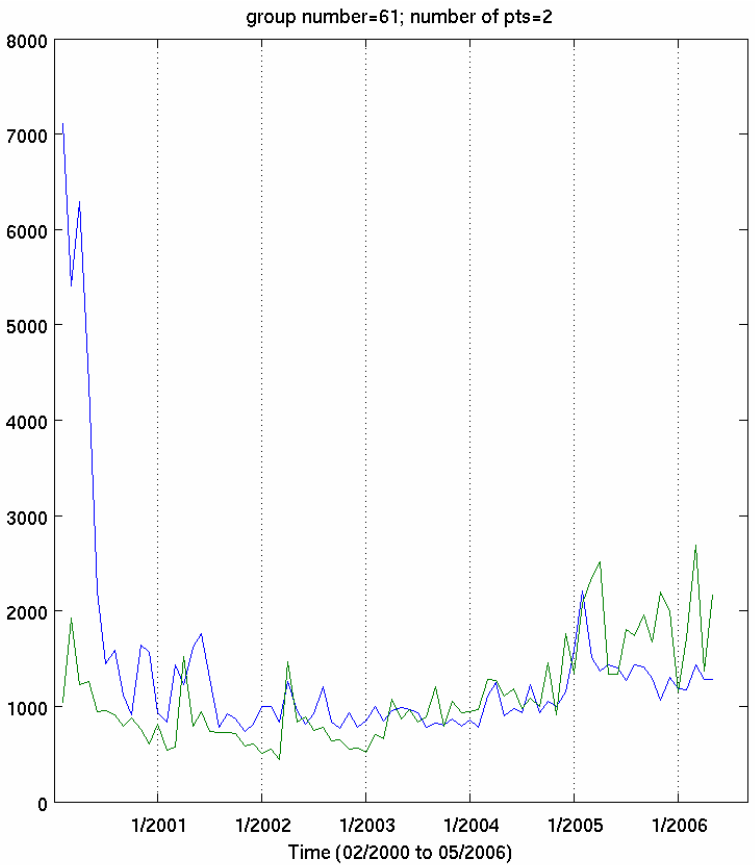
Example: Conversion to farmland

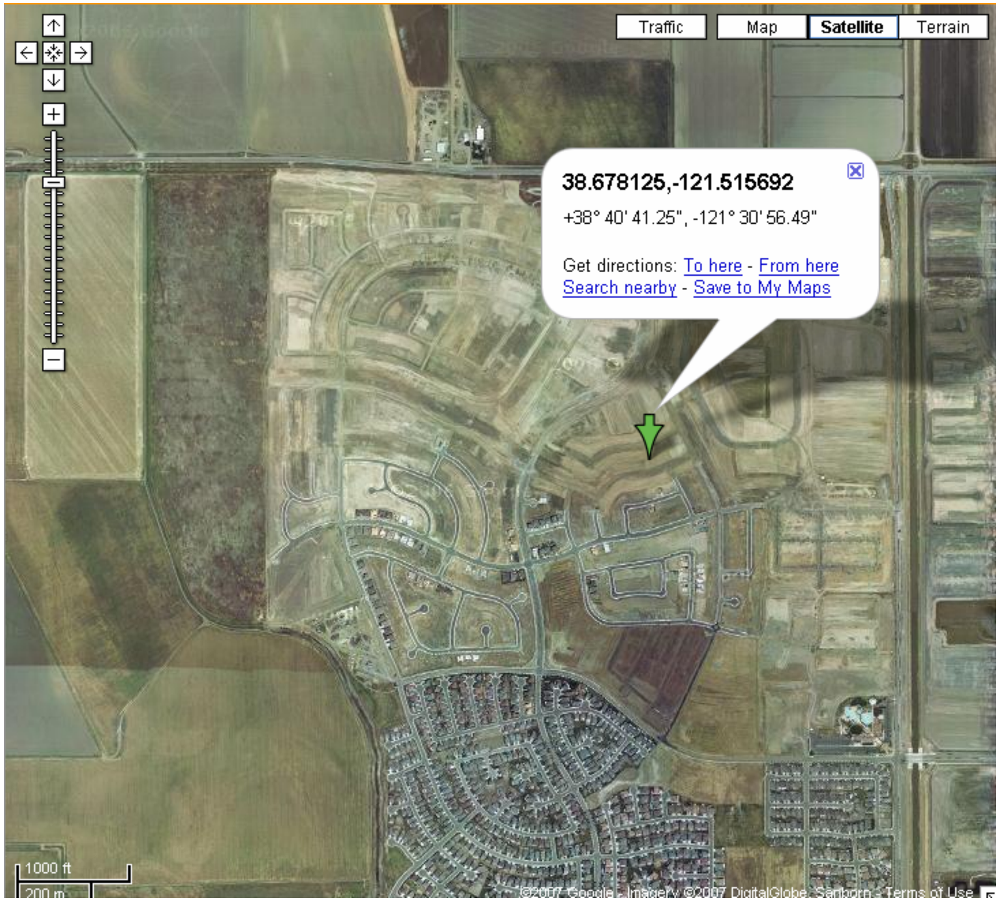



Map
Satellite
32.796875,-113.445034
+32° 47' 48.75", -113° 26' 42.12"
Get directions: To here - From here
Search nearby - Save to My Maps

This is a group of points that were all detected by our algorithm and the points are spatially located close to each other

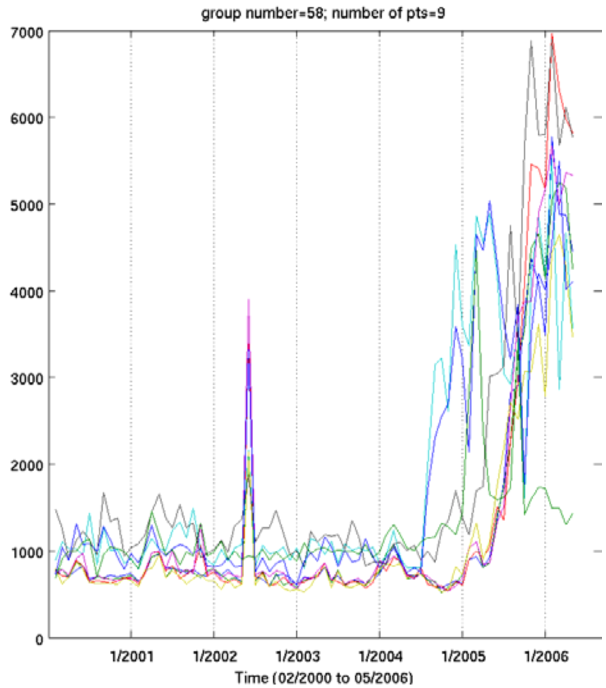
Example: Farmland to subdivision

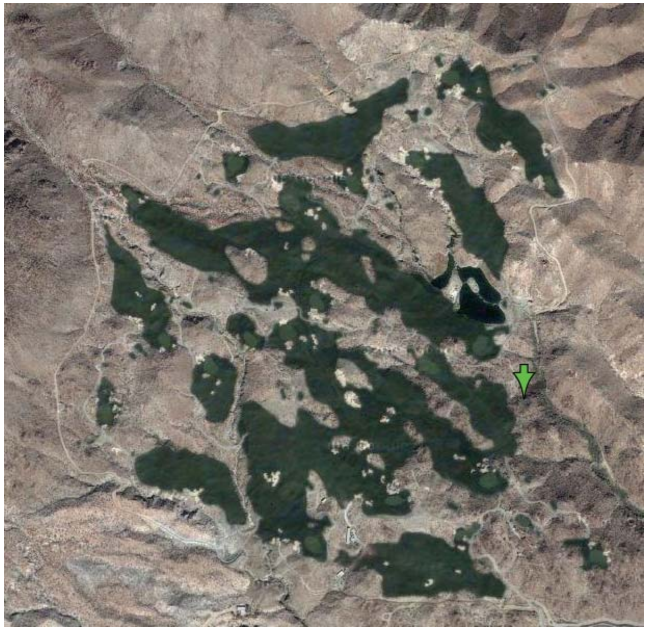



Traffic
Map
Satellite
Terrain
38.678125,-121.515692
+38° 40' 41.25", -121° 30' 56.49"
Get directions: To here - From here
Search nearby - Save to My Maps
1000 ft
200 m
©2007 Google - Imagery ©2007 DigitalGlobe, Sanborn - Terms of Use

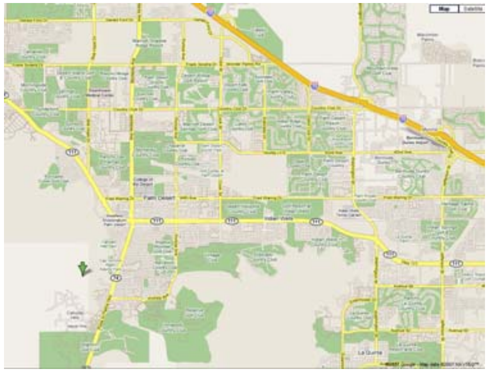
This is a location in Sacramento where farm land has been cleared and a subdivision is being built.

Bunch of Golf Courses in SE California Desert





- This is an example of a new golf course being built in Palm Desert, CA
- This town has over 100 golf courses, putting intense pressure on the water supply



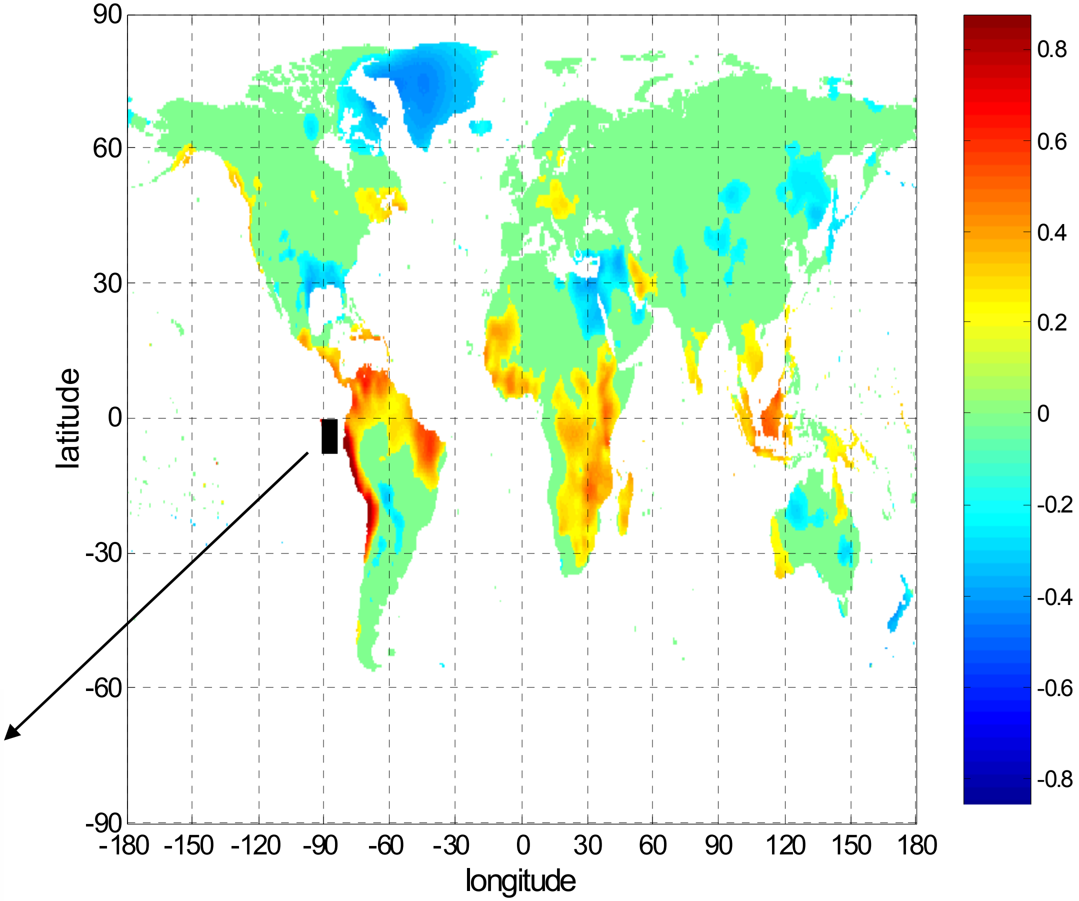
Land Cover Change Detection: Challenges

- Scalability
  - The data is at 250m resolution (and may become even finer in the future).
  - High resolution allows detection of even small localized changes but increases computation time
  - Scalablity is critical, especially if the analysis is done on a global scale
- 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

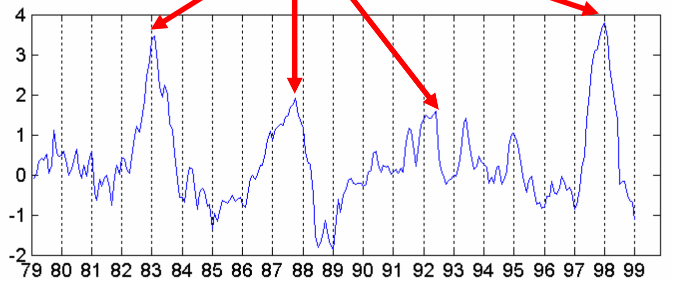
Climate Indices: Connecting the Ocean/Atmosphere and the Land

- 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)



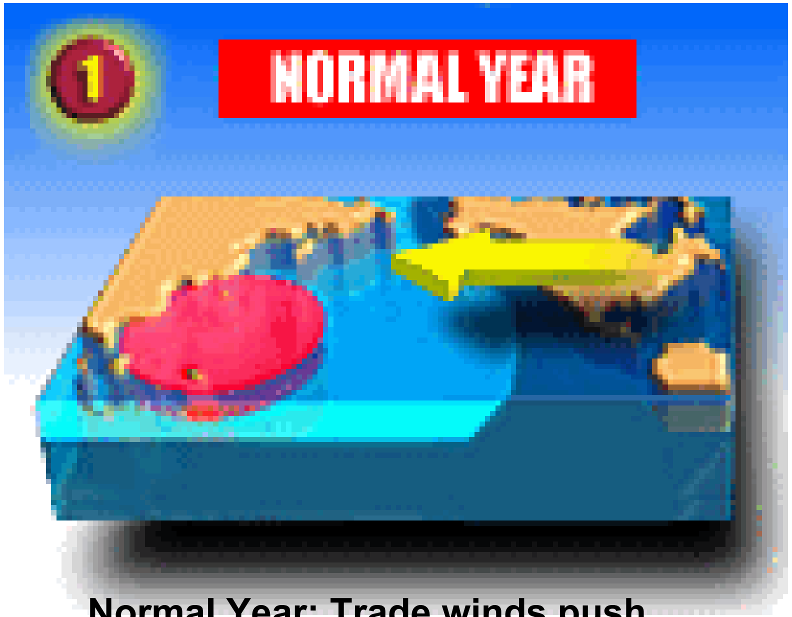
El Nino Events



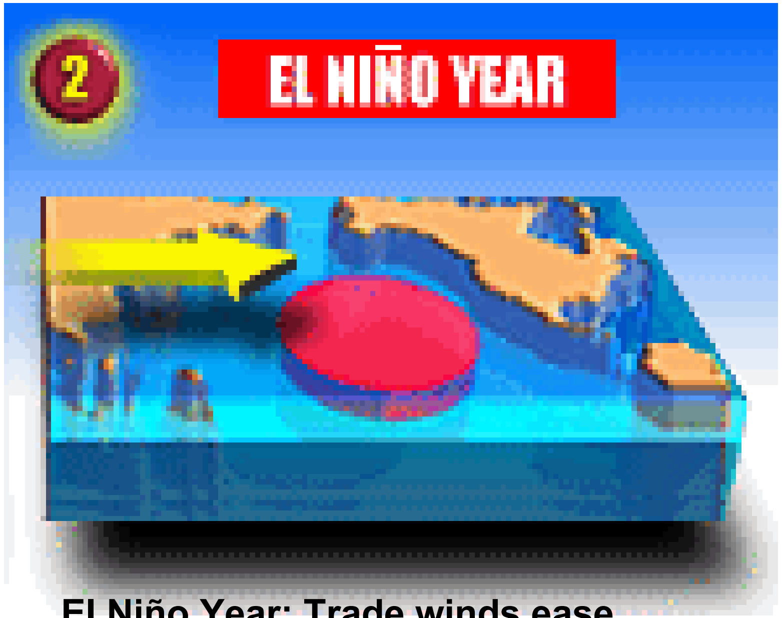
Nino 1+2 Index

The El Niño Climate Phenomenon

- El Niño is the anomalous warming of the eastern tropical region of the Pacific.



Normal Year: Trade winds push warm ocean water west, cool water rises in its place



El Niño Year: Trade winds ease, switch direction, warmest water moves east.

Effects: Drought in Australia, warmer winter in North America, flooding in coastal Peru, increased rainfall in East Africa

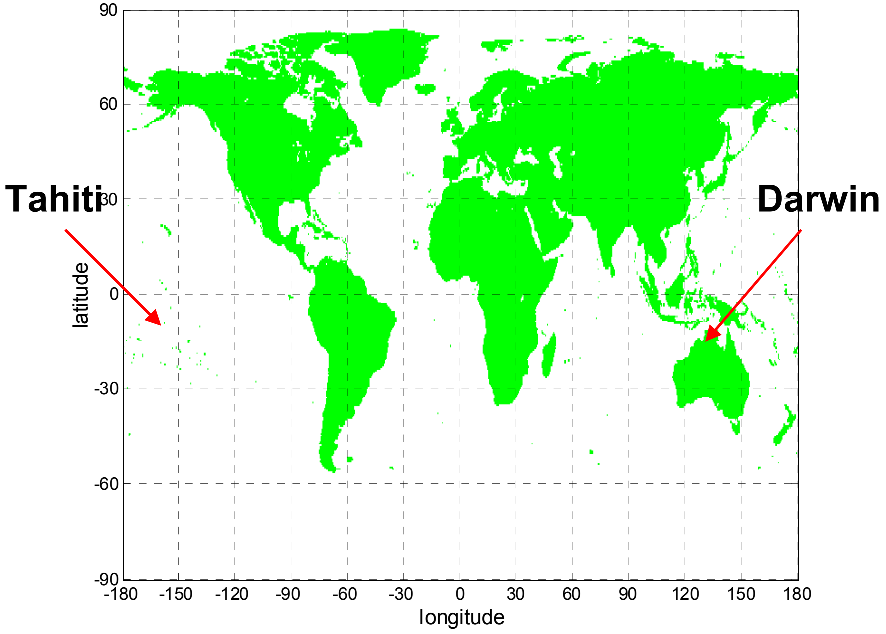
Graphic: http://www.usatoday.com/weather/tg/wetnino/wetnino.htm

A Pressure Based El Niño Index: SOI

- The Southern Oscillation Index (SOI) is also associated with El Niño.

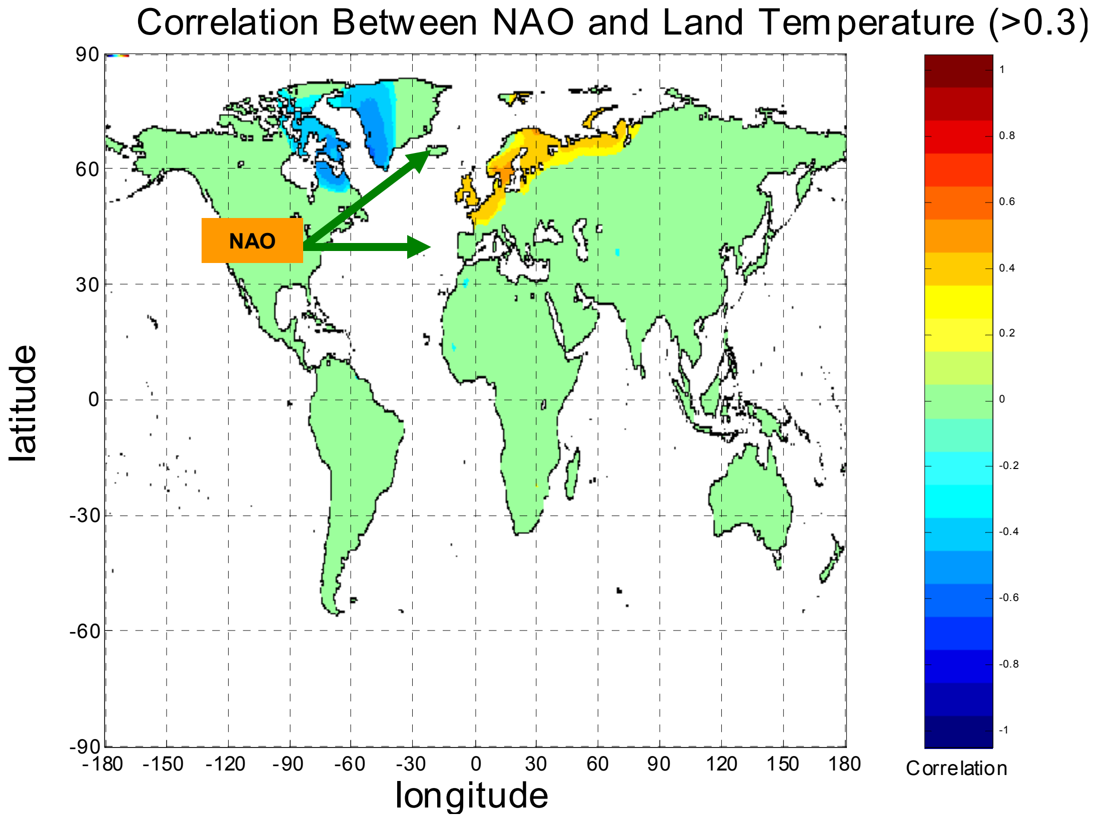
- Defined as the normalized pressure differences between Tahiti and Darwin Australia.

- Both temperature and pressure based indices capture the same El Niño climate phenomenon.



NAO (North Atlantic Oscillation)

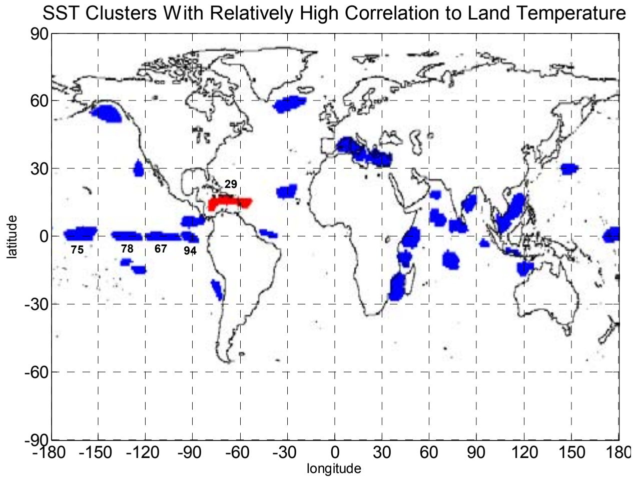
- NAO computed as the normalized difference between SLP at a pair of land stations in the Arctic and the subtropical Atlantic regions of the North Atlantic Ocean

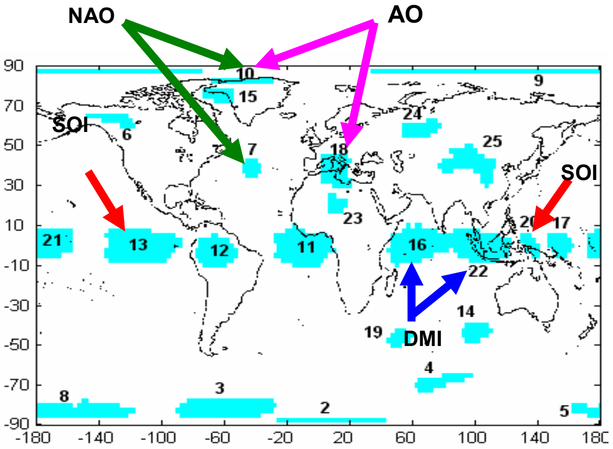


List of Well Known Climate Indices

Index	**Description**
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the _first principal component of SLP poleward of 20° N
PDO	**Pacific Decadel Oscillation:** Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
QBO	**Quasi-Biennial Oscillation Index:** Measures the regular variation of zonal (i.e. east-west) strato-spheric winds above the equator
CTI	**Cold Tongue Index:** Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6° N-6° S, 180° -90° W)
WP	**Western Pacific:** Represents a low-frequency temporal function of the 'zonal dipole' SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
NINO3	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
NINO4	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

Discovery of Climate Indices Using Clustering





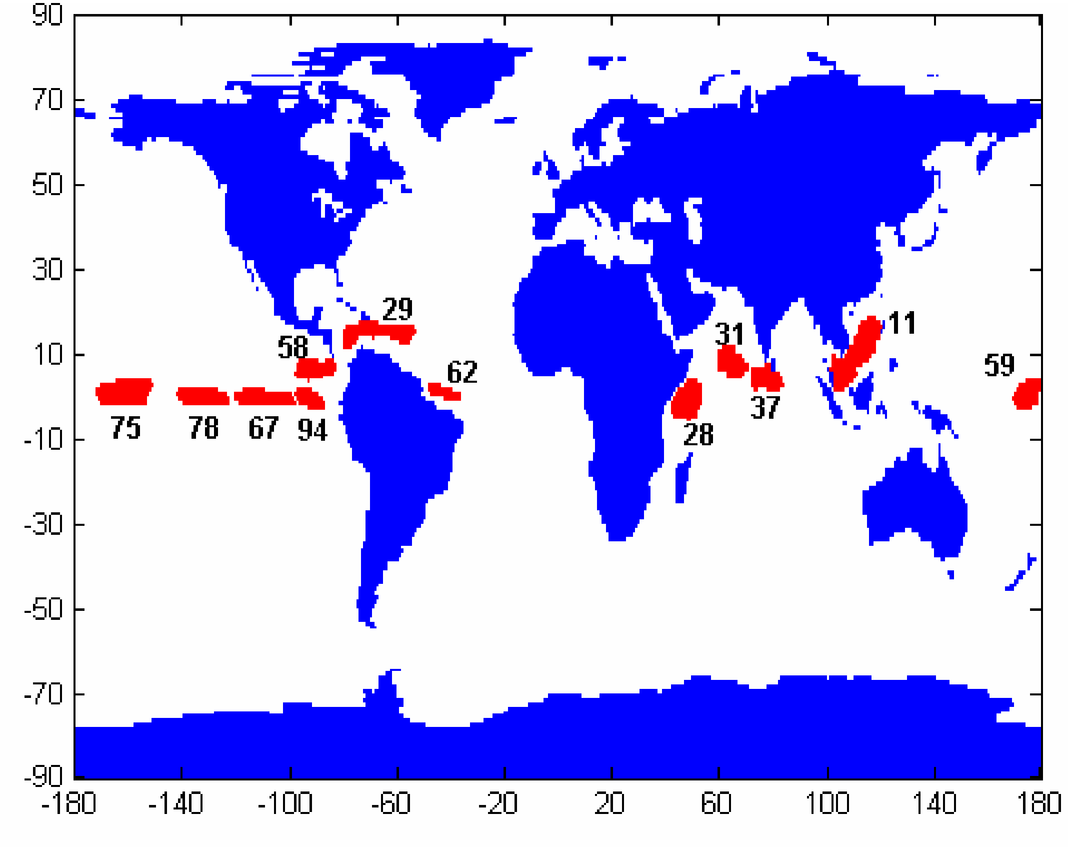
- Clustering provides an alternative approach for finding candidate indices.
  - Clusters represent ocean regions with relatively homogeneous behavior.
  - The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential climate indices.
- 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

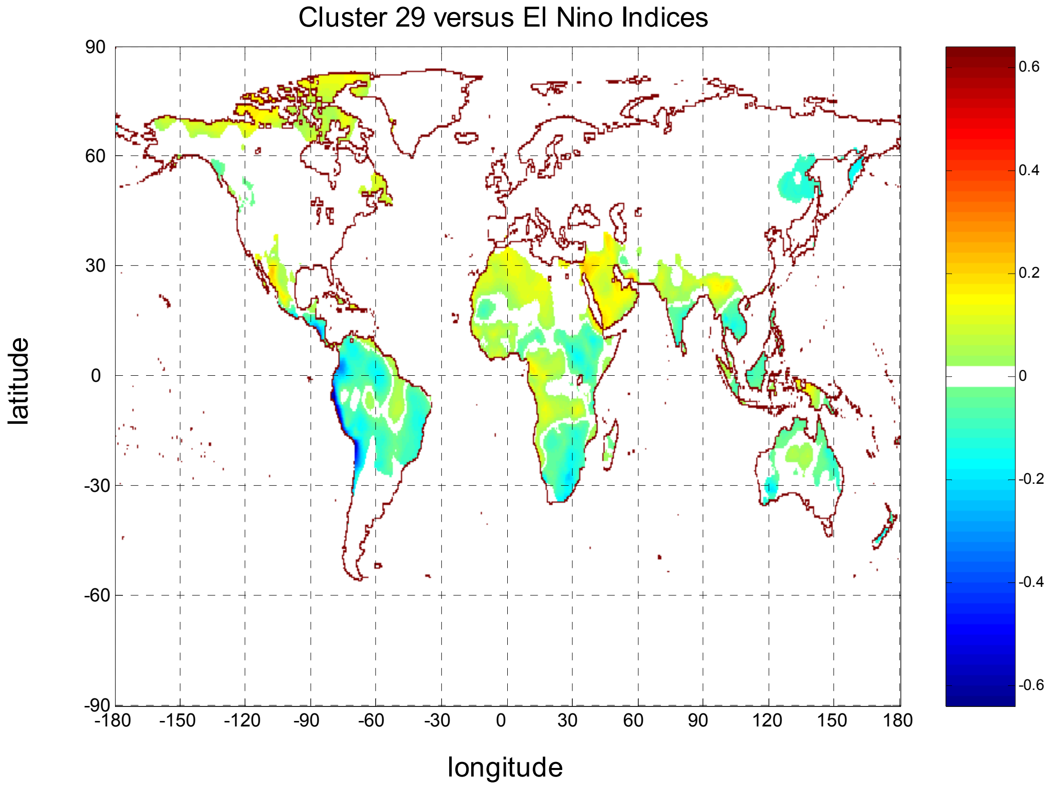


Cluster	Nino Index	Correlation
94	NINO 1+2	0.9225
67	NINO 3	0.9462
78	NINO 3.4	0.9196
75	NINO 4	0.9165

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

Clusters of SST that have high impact on land temperature

SST Cluster Moderately Correlated to Known Indices



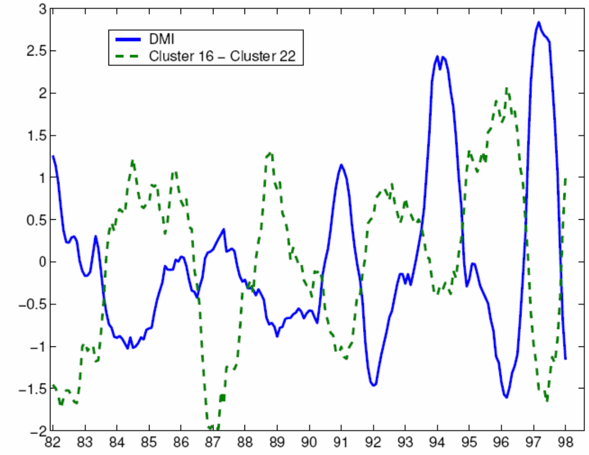
The figure shows the difference in correlation to land temperature between cluster 29 and the El Nino indices. Areas in yellow indicate where cluster 29 has higher correlation.

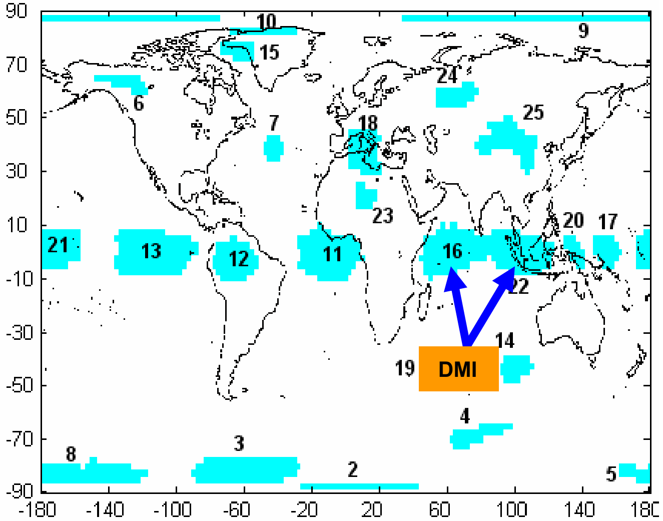
Some SST clusters are significantly different than known indices, but provide better correlation with land climate variables than known indices for many parts of the globe.

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.)

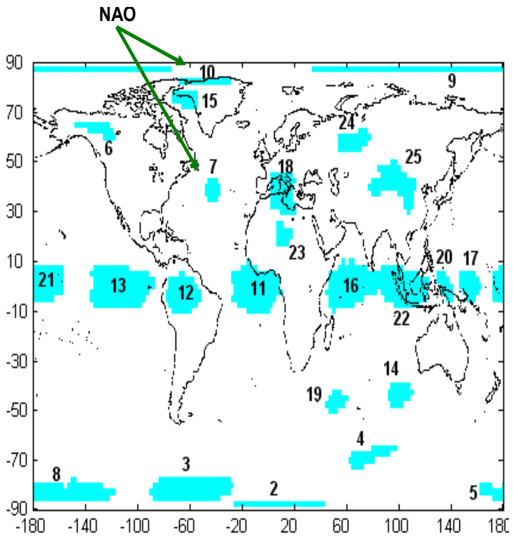


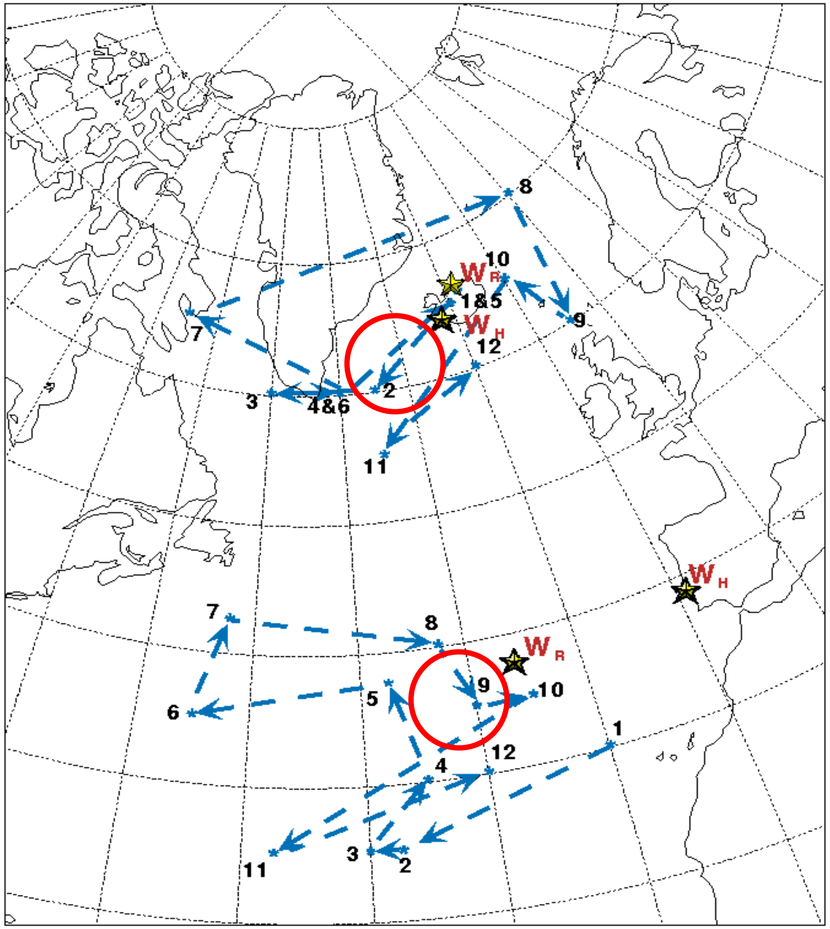


* N. H. Saji, B. N. Goswami, P. N. Vinayachandran and T. Yamagata, "A dipole mode in the tropical Indian Ocean," Nature 401, 360-363 (23 September 1999).

Dynamic Climate Indices

- 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
- 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.

Summary

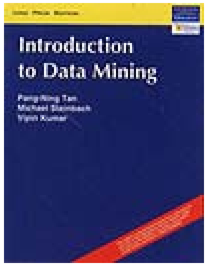
- Data driven applications such as data mining are increasingly driving the state-of-the-art in HPC.

- High Performance Data Mining is making significant contributions in areas such as **climate analysis**, biology, health sciences, scientific/engineering simulations.

- Tremendous scope for future work
  - New and better algorithms
  - Parallel/Distributed formulations

Bibliography (www.cs.umn.edu/~kumar)

- 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



- 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*.
- 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.
- 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