National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

A Probabilistic Approach to Mining Massive Earth Science Data Sets

Amy Braverman and Eric Fetzer

Jet Propulsion Laboratory,
California Institute of Technology

Mail Stop 126-347
4800 Oak Grove Drive
Pasadena, CA 91109-8099

How can we quantitatively characterize large scale (i.e. Level 3) data structure in massive Earth science data sets?

- Traditional approach: gridded daily/monthly averages and standard deviations.
  - Means and standard deviations characterize the data well only if the data are gaussian.
  - Means and standard deviations do not characterize joint relationships among variables.

- New approach: replace averages and standard deviations with multivariate distribution estimates.

- Key idea #1: *empirical probability distributions derived from data are the signatures of physical processes generating those data.*

- Key idea #2: understand and characterize large scale data structure by seeking physical explanations for observed evolution of multivariate probability distributions over space, time and resolution.

How do we produce multivariate distribution estimates?

- Entropy-constrained vector quantization (ECVQ; Chou, Lookabaugh and Gray, 1989) modified for use as a data summarization algorithm.

- ECVQ can be seen as a clustering algorithm similar to K-means. Different loss function:

$$L = \frac{1}{N} \sum_{n=1}^{N} \left[\|x_n - y(x_n)\|^2 + \lambda \left(-\log \frac{N_{y(x_n)}}{N} \right) \right]$$

x_n = multivariate data point

$y(x_n)$ = centroid of cluster to which data point is assigned

$N_{y(x_n)}$ = number of data points assigned to cluster with centroid $y(x_n)$

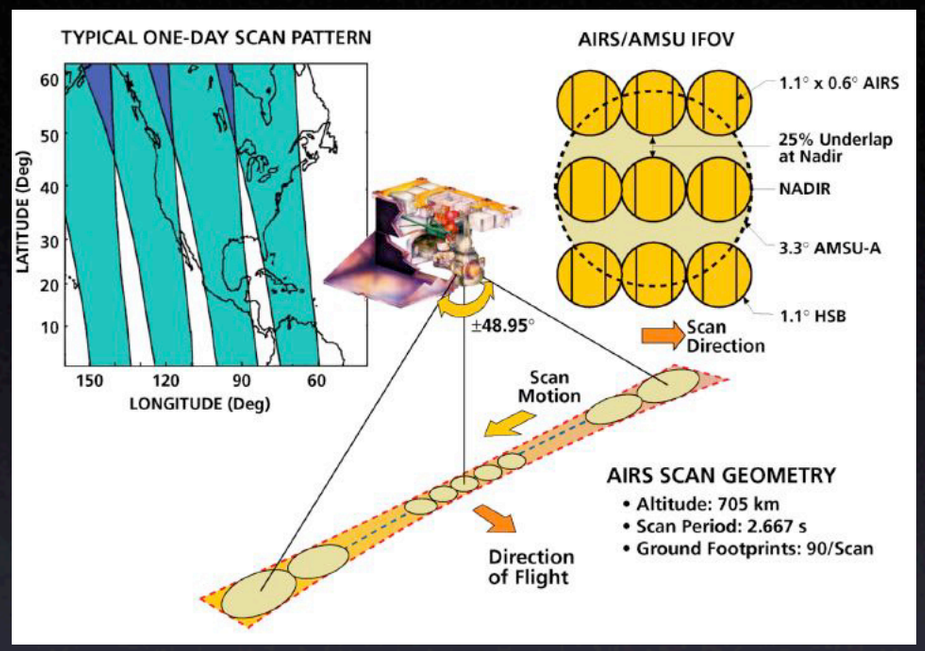
- Result: only as many clusters as necessary to describe the data, up to a maximum of K. (K-means always uses all K clusters.) Information-theoretic complexity of the data determines how many clusters.

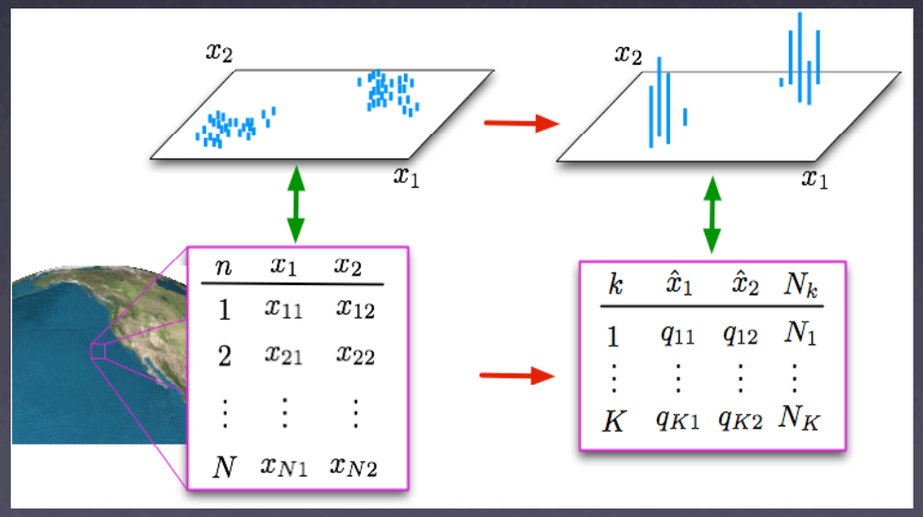
- Strategy: apply ECVQ clustering to all grid cells separately. Produces a set of cluster centroids and weights for each grid cell.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

AIRS data, January 2003





- 240 granules per day; each granule is 90 x 45 footprints (Level 2).
- One 35-dimensional data vector per footprint:
  - atmospheric temperature at 11 vertical levels (indices 1-11)
  - atmospheric humidity at 11 vertical levels (indices 12-22)
  - cloud fraction at 10 vertical levels (excludes lowest; indices 23-32)
  - land fraction of footprint (index 33)
  - day/night flag (index 34)
  - quality flag (index 35).
- For each 5 x 5 degree grid cell, input to ECVQ all data vectors with footprints inside the cell.
- Output for each 5 x 5 cell: a set of representative vectors and their associated weights and distortions. (Distortion is the average squared distance between original and representative vectors.)
- File size: about 37 MB.

Now what?

- Does it work?
  - Are the summary distributions good reflections of the original data?
    - Cluster distortions tell you the answer quantitatively.
    - For science: case studies seem to bear out validity for January 2003.
- What can we do with it?:
  - Compute (custom Level 3 data products):
    - Distribution implied by a summary is a coarsened representation of the original data distribution.
      - Let X be a random vector with probability mass function f_X, the pmf of the original data. Let $Y = q(X)$ be a the random vector having pmf f_Y given by the summary. Then if g is a "nice" continuous function,

$$E\|g(X) - g(Y)\|^2 \approx E\left[(X-Y)'\dot{g}(Y)\dot{g}(Y)'(X-Y)\right]$$

E is the statistical expectation operator

  - Simulate from it: summary can be thought of as a statistical mixture of multivariate gaussians.
  - ***Mine it!***
    - Infinite variety of quantities to characterize differences between distributions- choice depends on scientific objectives. For agnostic data mining, we need one that has intrinsic meaning apart from unknown, future scientific objectives.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Mining

- A distance between distributions:

$$\pi_1 = \{(y_{1k_1}, \pi_{1k_1})\}_{k_1=1}^{K_1} \qquad \pi_2 = \{(y_{2k_2}, \pi_{2k_2})\}_{k_2=1}^{K_2}$$

$$\Delta(\pi_1, \pi_2) = \min_{p_{12}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \|y_{1k_1} - y_{2k_2}\|^2 p_{12}(y_{1k_1}, y_{2k_2})$$

π 's are fixed; fill in p 's such that:

(1) constraints are satisfied
(2) Δ is minimized

column constraints

$$\pi_{11} = p_{11} + p_{21} + p_{31}$$

$$\pi_{12} = p_{12} + p_{22} + p_{32}$$

$$\pi_{13} = p_{13} + p_{23} + p_{33}$$

$$\pi_{14} = p_{14} + p_{24} + p_{34}$$

$$\pi_{21} = p_{11} + p_{12} + p_{13} + p_{14}$$

$$\pi_{22} = p_{21} + p_{22} + p_{23} + p_{24}$$

$$\pi_{23} = p_{31} + p_{32} + p_{33} + p_{34}$$

row constraints

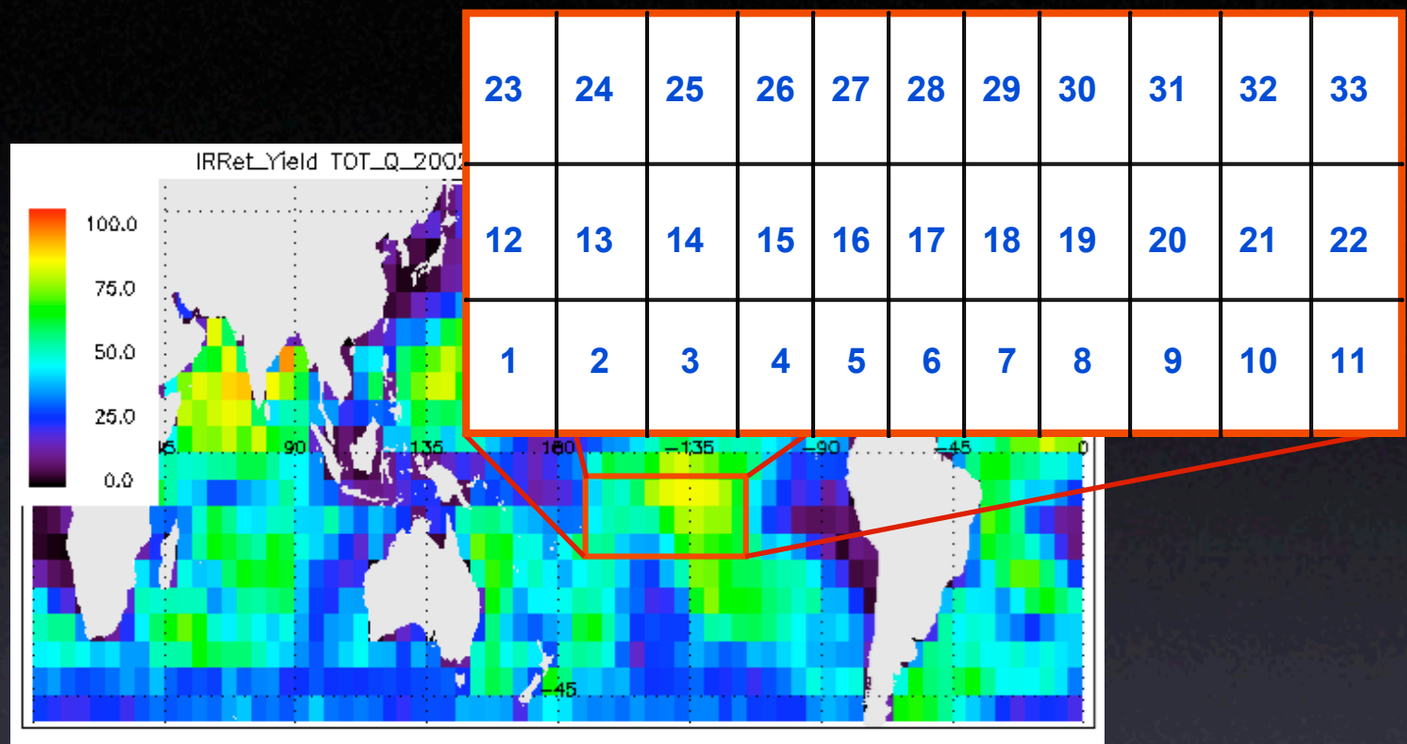
		y_{11}	y_{12}	y_{13}	y_{14}
		π_{11}	π_{12}	π_{13}	π_{14}
y_{21}	π_{21}	p_{11}	p_{12}	p_{13}	p_{14}
y_{22}	π_{22}	p_{21}	p_{22}	p_{23}	p_{24}
y_{23}	π_{23}	p_{31}	p_{32}	p_{33}	p_{34}

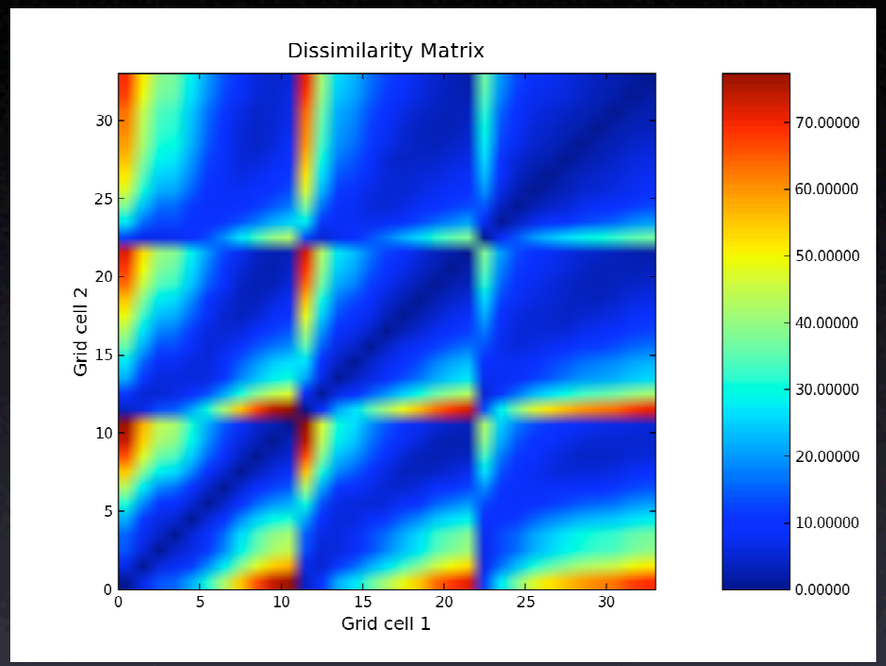
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Example

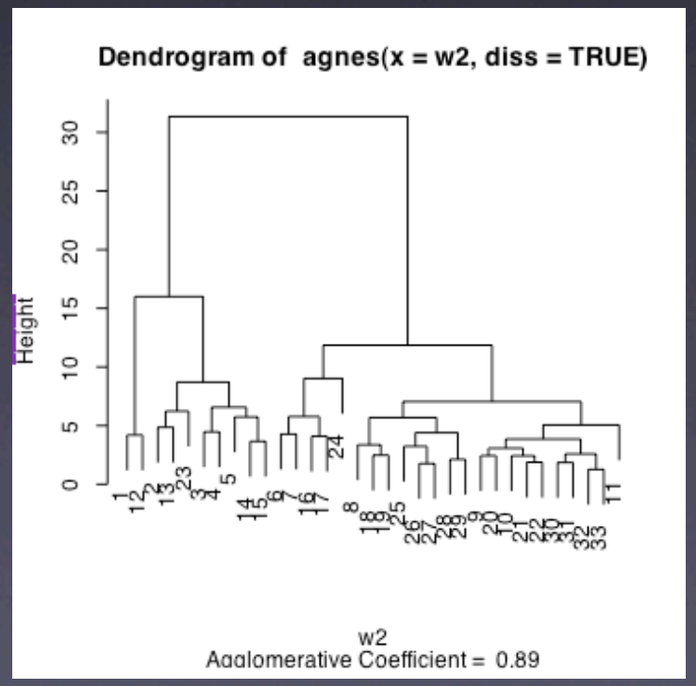
- Use Δ to characterize relationships among the 33 grid cells in the area shown below.

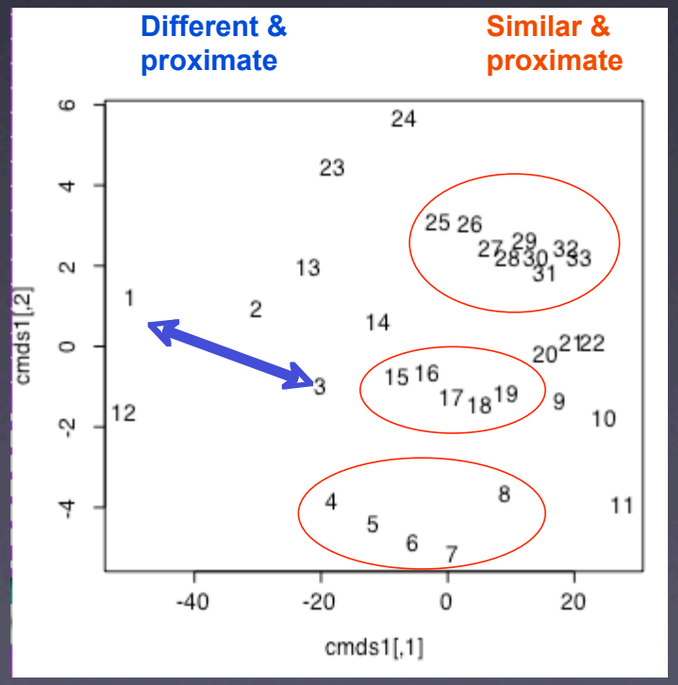




33 x 33 symmetric matrix of Δ 's

Hierarchical cluster analysis dendogram





Multidimensional scaling plot

Summary

- Problem addressed: finding and quantitatively characterizing previously unknown structure in large, satellite data sets.
- Approach: summarize data in a way that preserves statistical properties, but also reduces size and complexity. Use probabilistic metric to quantify degree of difference between summaries.
- Use the metric as a basis for standard and non-standard data mining methods.
- Other uses:
  - Compare data from different sources (e.g. observations to model output).
  - Study time series of summaries.
  - Study effects of changes in resolution.
- Contact information:
  - email: Amy.Braverman@jpl.nasa.gov
  - AIRS website: http://www-airs.jpl.nasa.gov