Reducing Size and Complexity of Very Large Remote Sensing Data Sets

Amy Braverman

Earth and Space Sciences Division
Jet Propulsion Laboratory,
California Institute of Technology

Outline

- Motivation.
- Strategy.
- Method.
- Example.
- Science Analysis.
- Conclusions.

Motivation

- Earth Observing System satellites (Terra, Aqua, Aura) to return vast quantities of data.

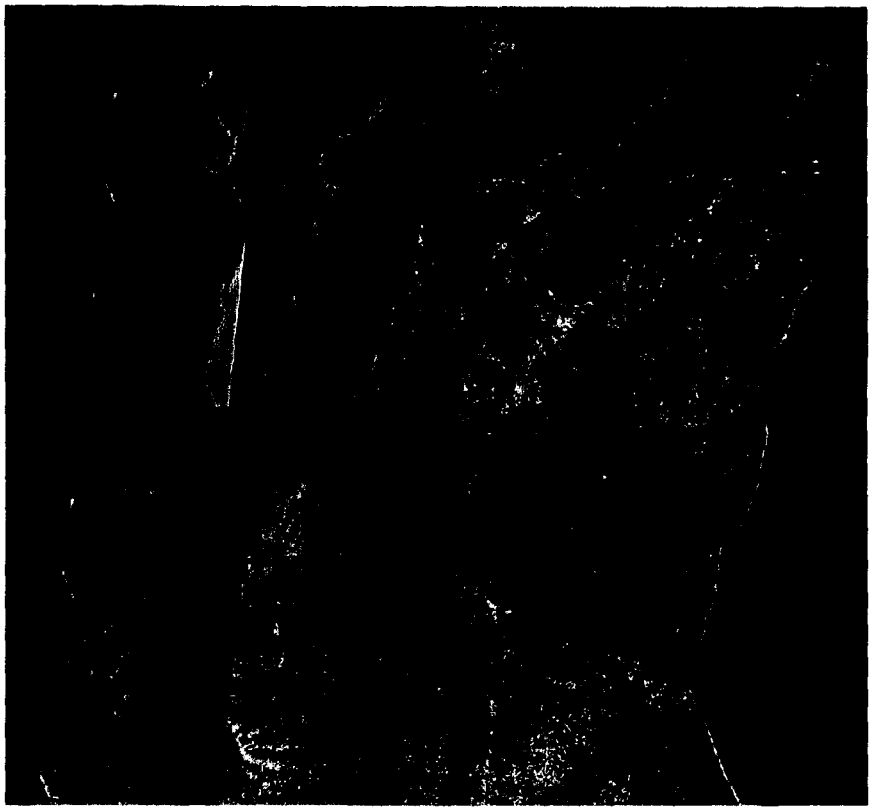
- New type of data provides a global view of local phenomena.

- Data to be made available to the public for analysis, but large segment of the user community can't handle the volume.

Motivation

Multi-angle Imaging SpectroRadiometer (MISR) aboard Terra produces about 1 TB per month of radiance and geophysical data.

Motivation

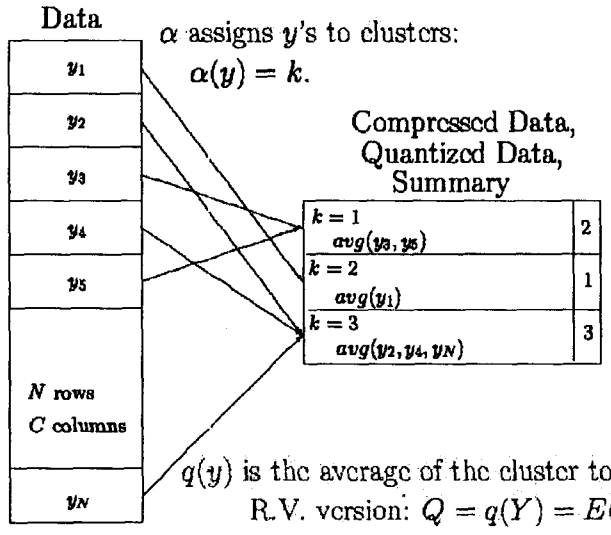


MISR RGB middle-Atlantic states images, March 24, 2000.

Strategy

- Partition data on a monthly, global, one degree by one degree grid.
- Replace original data in each cell with a set of representatives and associated weights called a summary.
- Discrete distribution so defined should be close to the original empirical distribution, but also parsimonious.
- Similar to data squashing (DuMouchel, 2001), quantization (Cover and Thomas, 1991).

Method



α assigns y's to clusters:
$\alpha(y) = k$.

$\beta(k)$ produces cluster means:
$\beta(k) = \frac{1}{N(k)} \sum_{n=1}^{N} y_n 1[\alpha(y_n) = k]$.

$N(k)$ produces cluster counts:
$N(k) = \sum_{n=1}^{N} 1[\alpha(y_n) = k]$.

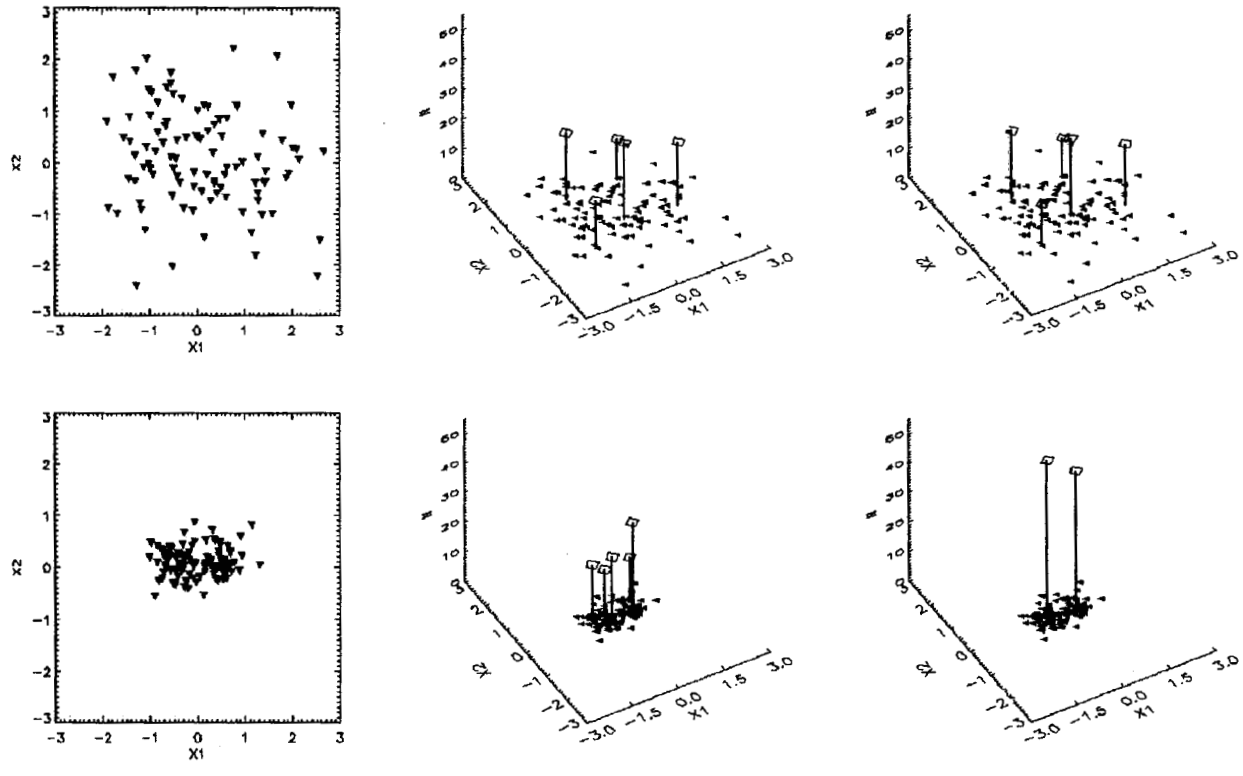
$q(y)$ is the average of the cluster to which y belongs: $q(y) = \beta[\alpha(y)]$.
R.V. version: $Q = q(Y) = E(Y|Q)$. $\Rightarrow E(Q) = E(Y)$,
$Var(Q) \leq Var(Y)$,
$Cov(Q, Y) = Cov(Q)$.

Two figures of merit for q, or equivalently, α:

$\Delta(q) = \frac{1}{N} \sum_{n=1}^{N} \|y_n - q(y_n)\|^2 = trCov(Y - Q)$.

$h(q) = -\sum_{k=1}^{K} \frac{N(k)}{N} \log \frac{N(k)}{N}$.

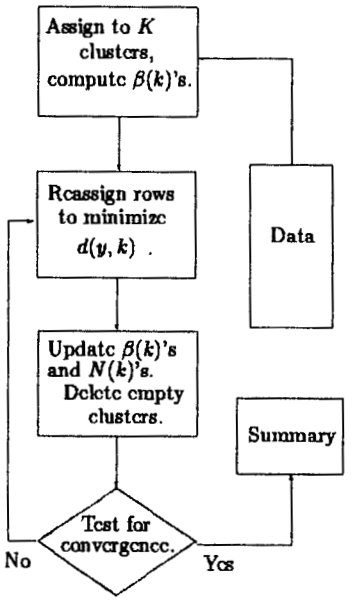
Method



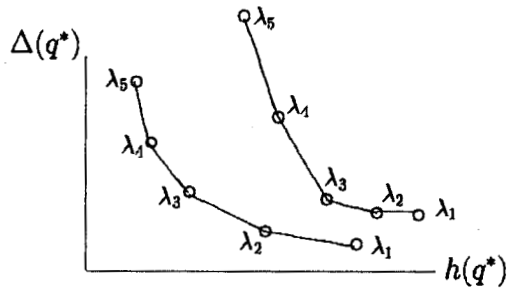
- Want fewer, more massive clusters when that will do $\Longleftrightarrow$ want distortions as similar as possible.

Method

ECVQ Algorithm:



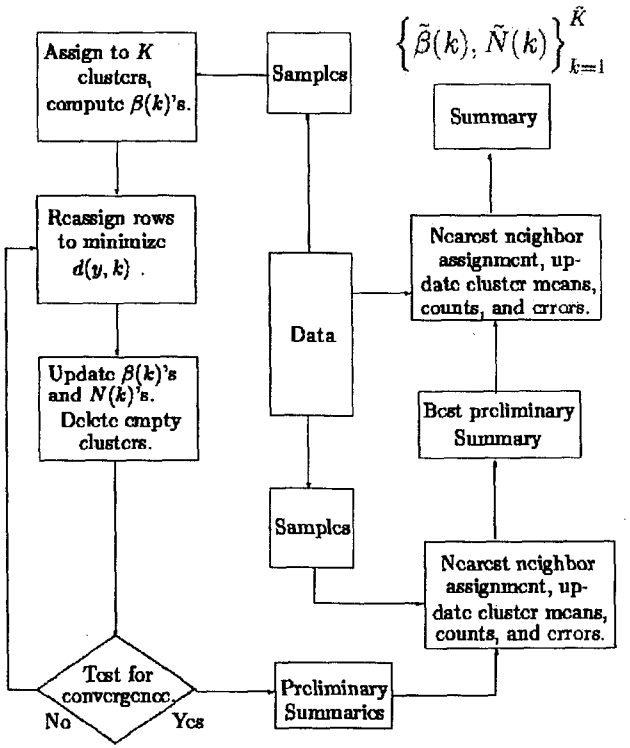
- Minimize $L_\lambda = \frac{1}{N}\sum_{n=1}^{N} d(y,k)$ by choice of α:
  $$d(y,k) = \|y - \beta(k)\|^2 + \lambda\left[-\log\frac{N(k)}{N}\right].$$
- Original application: estimate distortion-rate functions of information sources.



- Computationally intensive, depends on initial random assignment, not distortion-minimizing.

Method

ECVQ Modification 2:



- Minimizes error.
- Iteration and multiple scans on samples only.
- Best preliminary summary is the one with smallest $\hat{\Delta}$.
- $\tilde{\Delta}$ is a goodness-of-fit measure.
- Average of $\hat{\Delta}$'s, $\bar{\Delta}$, is a process performance measure which accounts for sampling variation.

Example

- MISR aerosol data over southern Africa, August-September 2000.

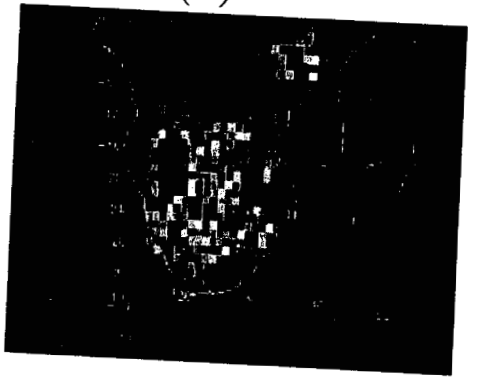
- y_n is a six-dimensional observation (row of data): $y_n = (\tau_n, \chi_{n1}, \chi_{n2}, \chi_{n3}, \chi_{n4}, \chi_{n5})$ with a latitude and longitude, representing a $1.1km^2$ region.

- τ_n is optical depth. χ_{nj}'s measures how close the vector of observed MISR radiances is from that predicted by aerosol model j.

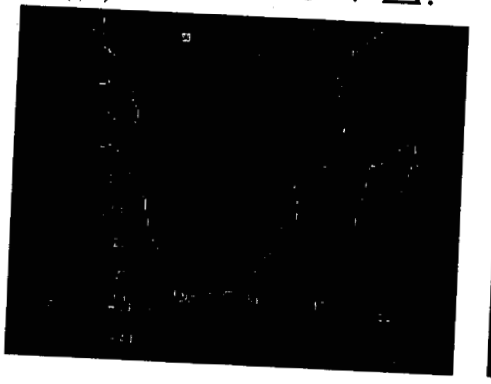
- Original data: 6,304,861 observations. $\sum_{lat,lon} \tilde{K}$=9,322 observations.

Example

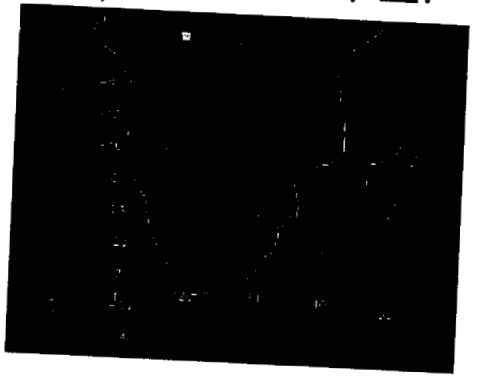
(a) N.



(b) Relative $\sqrt{\tilde{\Delta}}$.



(c) Relative $\sqrt{\bar{\Delta}}$.



(d) $\tilde{K}$.

(e) $\tilde{h}$.

Key:

0	(a)	$\geq 20,000$
0	(b)	≥ 0.05
0	(c)	≥ 0.05
0	(d)	40
0	(e)	≥ 4.5

Science Analysis

Is there a relationship between optical depth and heterogeneity of the model-fit χ's?

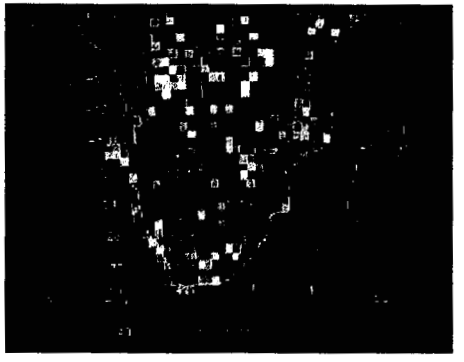
- Measure heterogeneity by $w_n = \frac{1}{5}\sum_{j=1}^{5}(\chi_{nj} - \bar{\chi}_n)^2$.
- Compare: $\rho(\tau, w)$ computed using original data to $\hat{\rho}(\tau, w)$ computed from summaries:

$$\rho(\tau, w) = \frac{\sum_{n=1}^{N}(\tau_n - \bar{\tau})(w_n - \bar{w})}{\sqrt{\sum_{n=1}^{N}(\tau_n - \bar{\tau})^2}\sqrt{\sum_{n=1}^{N}(w_n - \bar{w})^2}},$$

$$\hat{\rho}(\tau, w) = \frac{\sum_{k=1}^{K} N(k)(\hat{\tau}_k - \bar{\tau})(\hat{w}_k - \bar{w})}{\sqrt{\sum_{k=1}^{K} N(k)(\hat{\tau}_k - \bar{\tau})^2}\sqrt{\sum_{k=1}^{K} N(k)(\hat{w}_k - \bar{w})^2}}.$$

Example

(a) $\rho(\tau, w)$.



(b) $\hat{\rho}(\tau, w)$.

Key:

- Quality depends on distortion and on the form of the transformation.

Conclusion

- Computational efficiency is an issue; depends in part on how clustered the data are.
- Improvements: two-stage version; FastECVQ.
- Algorithm parameter settings require experimentation.
- Provides nonparametric, descriptive summary of the data.
- Sensitivity tests required to guage effect of users' transformations.