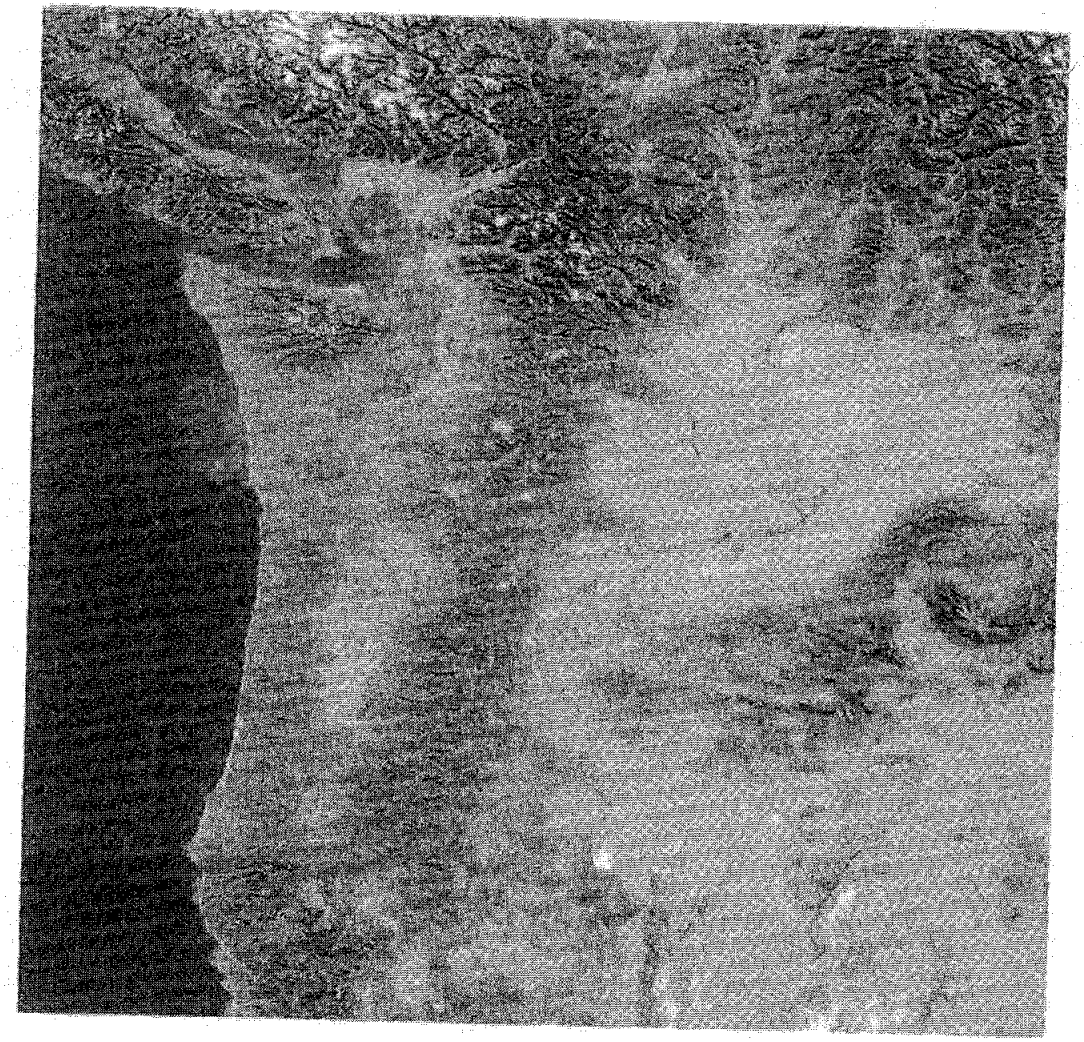
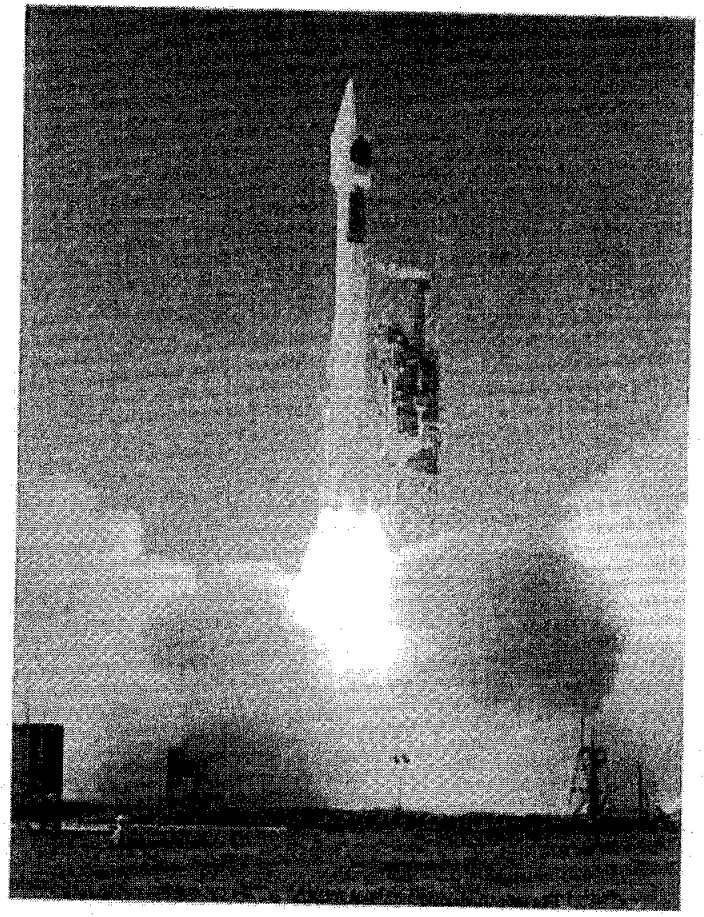
Reducing Size and Complexity of Very Large Remote Sensing Data Sets



Amy Braverman
Jet Propulsion Laboratory
California Institute of Technology
Mail Stop 169-237
4800 Oak Grove Drive
Pasadena, CA 91109-8099
Amy.Braverman@jpl.nasa.gov

Outline

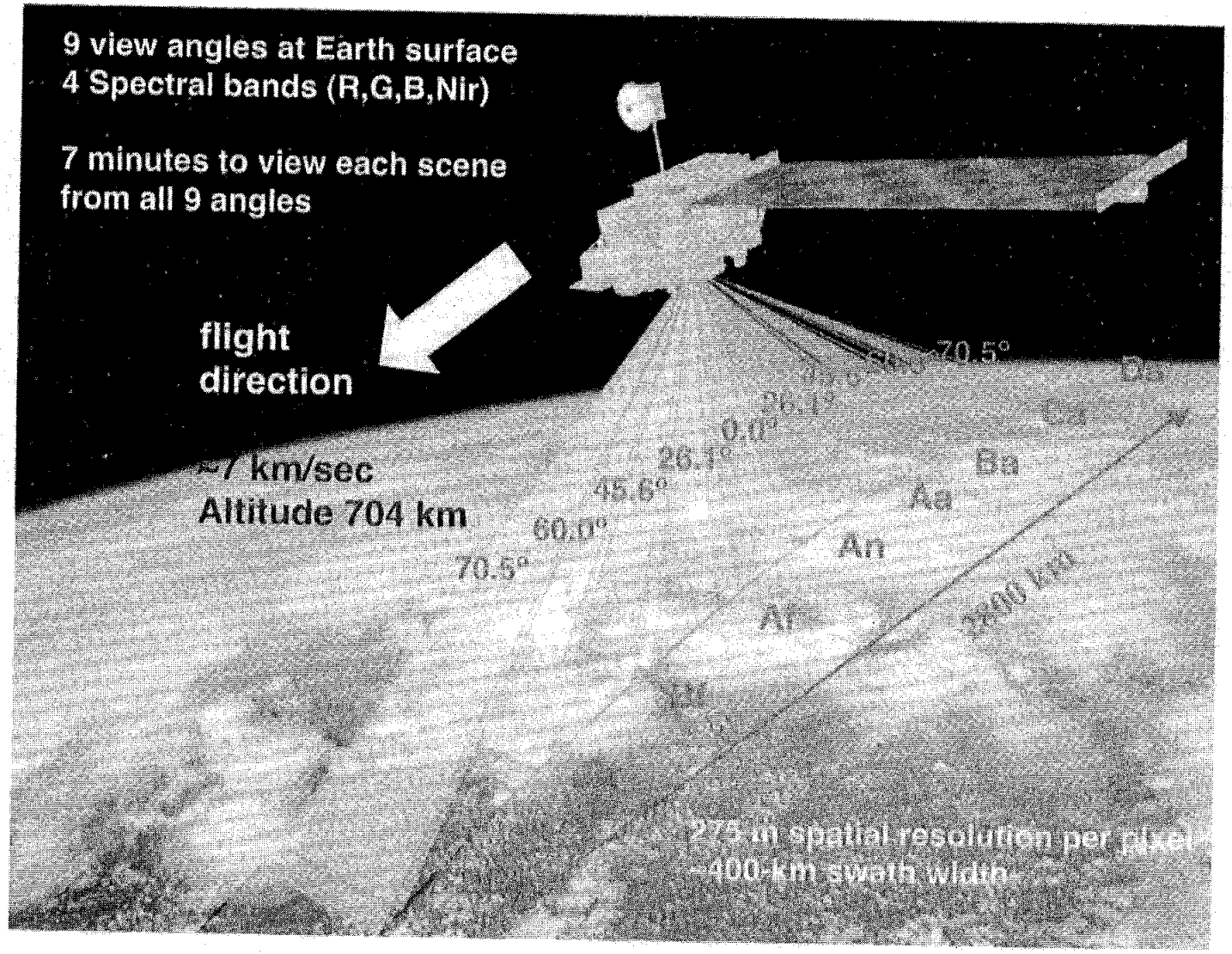
1. Motivation
2. Approach
3. Quantization
4. Rate-distortion Theory
5. Assuring Comparability and Parsimony
6. Compressibility and Compression
7. Algorithm
8. Example Analysis
9. Summary



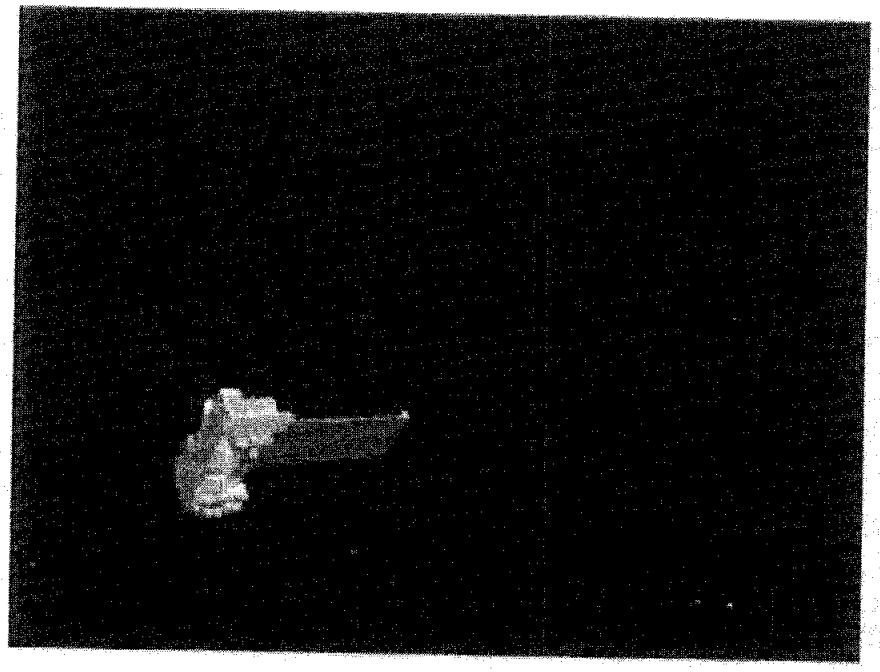
Motivation

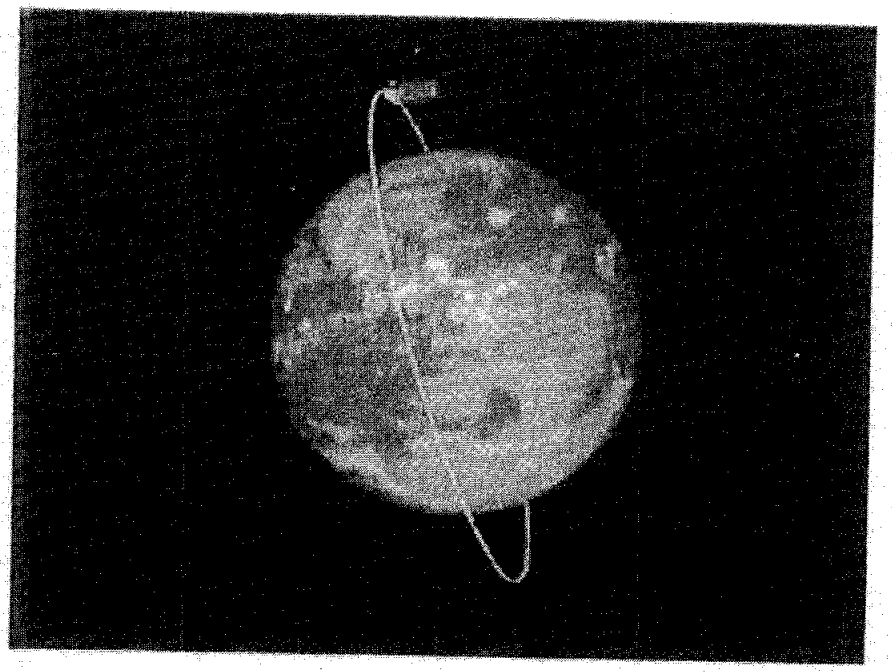
- Earth Observing System satellites return "massive" data volume.
- Traditional approach to data exploration: produce maps of one degree averages and standard deviations for each parameter of interest.
- Good news: this is easy, practical, and everybody understands it.
- Bad news: the method throws away almost all of the distributional information in the data including covariance and higher-order statistics.
- **Need to produce lower volume proxy data sets suitable for exploratory data analysis; should be faithful to original data and parsimonious.**

Multi-angle Imaging SpectroRadiometer



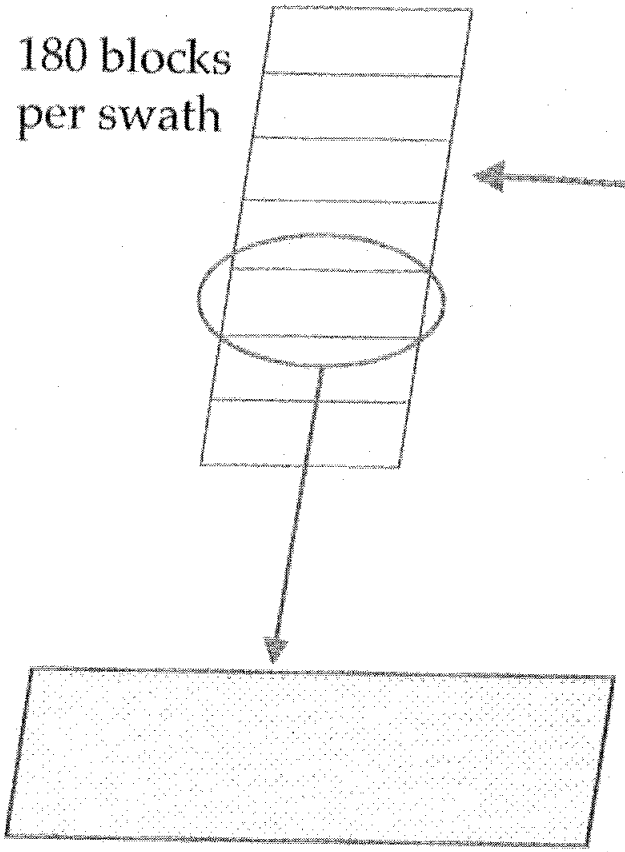
Multi-angle Imaging SpectroRadiometer

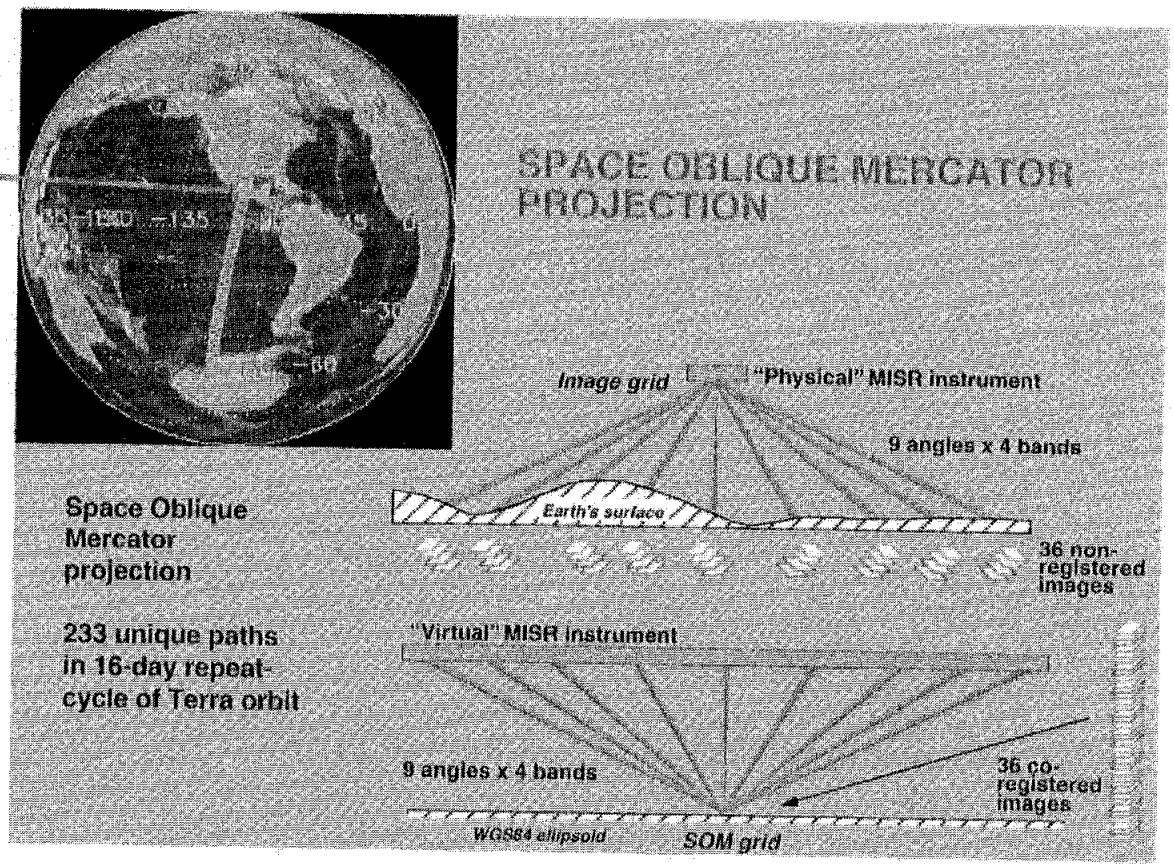




Multi-angle Imaging SpectroRadiometer

180 blocks
per swath



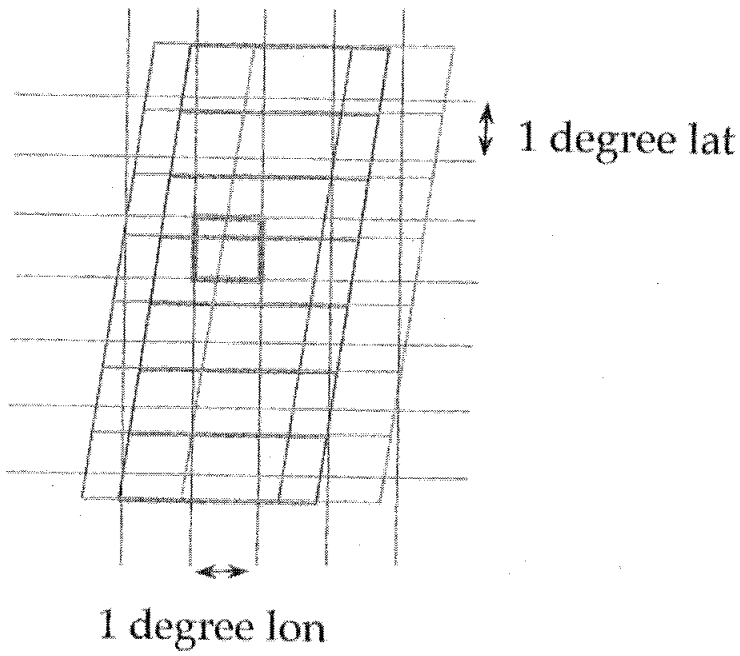


1 block=128 × 512 1.1 km pixels
512 × 2048 275 m pixels

36 measurements
per pixel

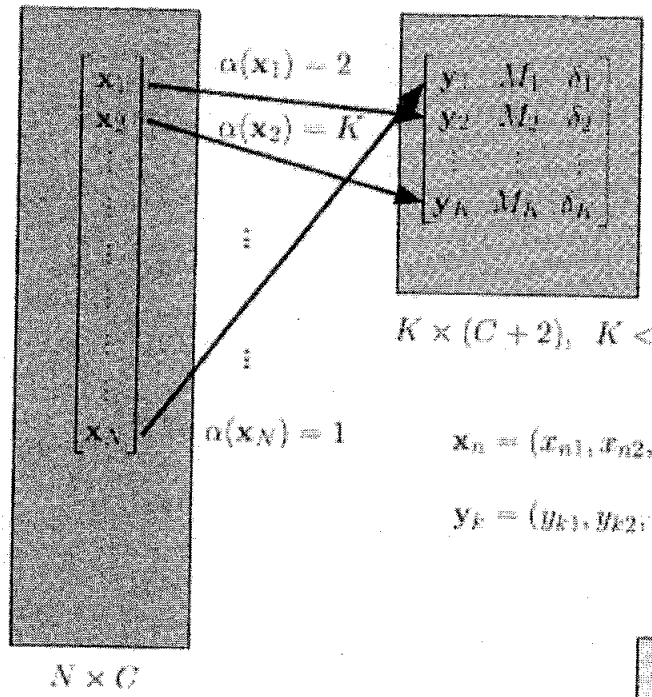
Approach

- New approach: produce an estimate of the joint (empirical) probability distribution of variables of interest within each one degree grid cell.



- Use a clustering algorithm to partition data into groups, represent each group by its centroid and (normalized) membership count.
- Collection of all 180 x 360 = 64,800 grid cell distribution estimates is a proxy for the original data.
- Similar in philosophy to data squashing (Dumouchel, 2001).

Quantization



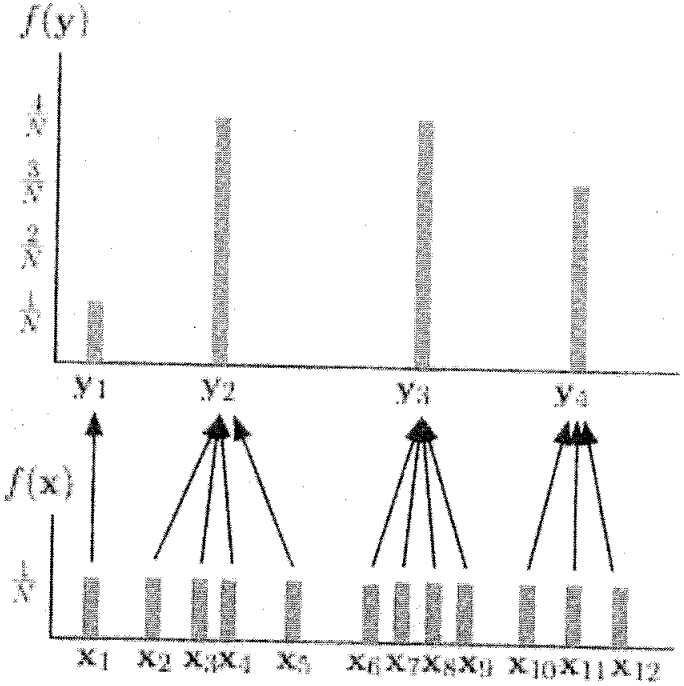
$$\mathbf{y}_k = \frac{1}{M_k} \sum_{n=1}^{N} \mathbf{x}_n 1[\alpha(\mathbf{x}_n) = k]$$

$$M_k = \sum_{n=1}^{N} 1[\alpha(\mathbf{x}) = k]$$

$$\delta_k = \frac{1}{M_k} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{y}_k\|^2 1[\alpha(\mathbf{x}_n) = k]$$

$$\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nC})'$$

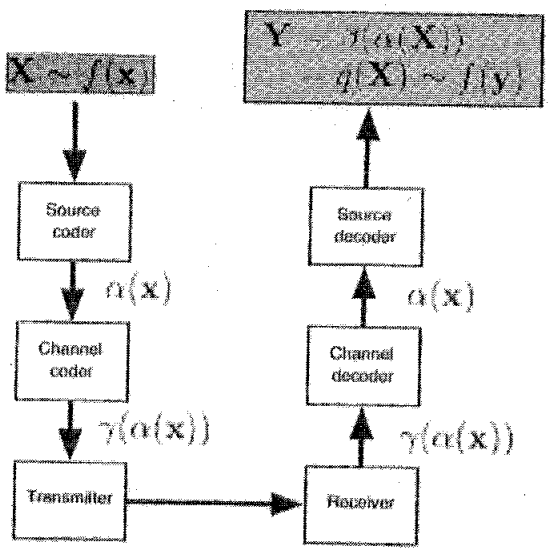
$$\mathbf{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kC})'$$



$$\mathbf{Y} = q(\mathbf{X}) = E(\mathbf{X} \mid \mathbf{Y})$$

$$\delta(\mathbf{X}, \mathbf{Y}) = E\|\mathbf{X} - \mathbf{Y}\|^2$$

$$h(\mathbf{Y}) = -E(\log f(\mathbf{Y}))$$



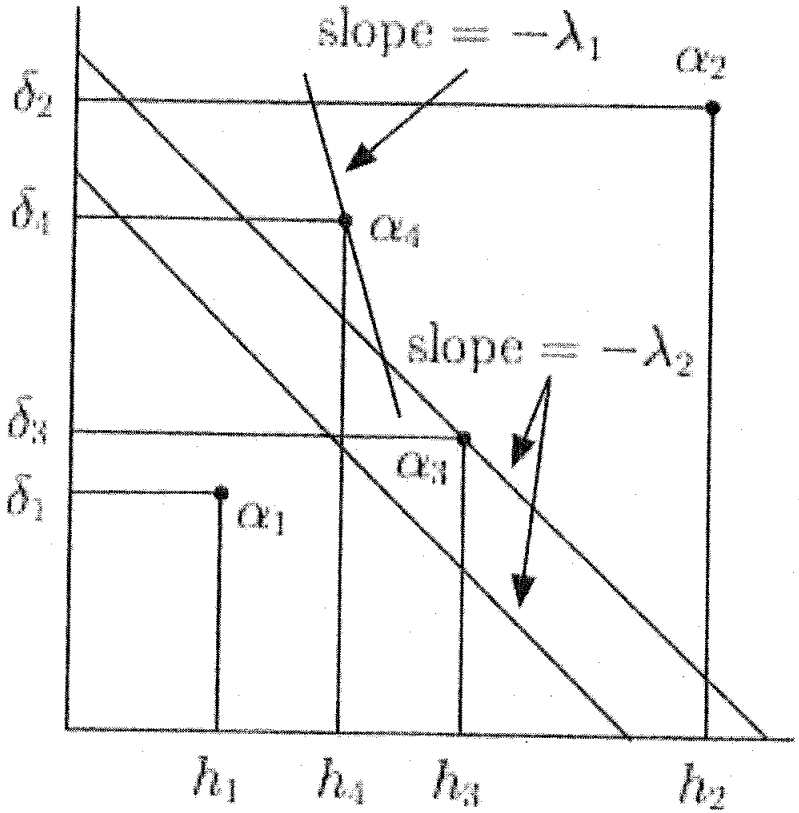
$\alpha(\mathbf{x})$ is the source coder:

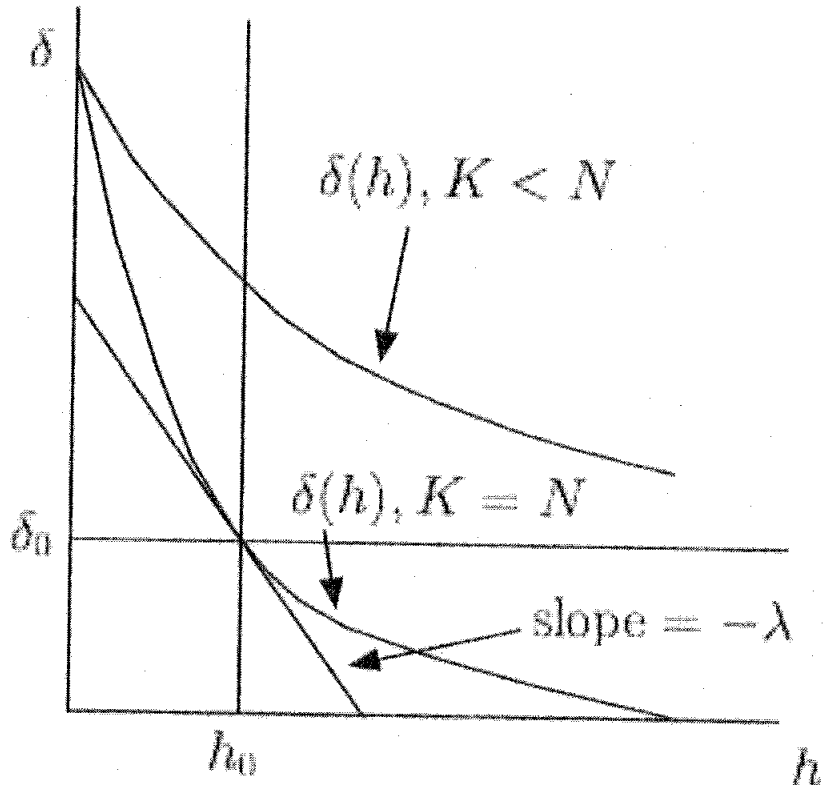
$$\alpha(\mathbf{x}) = argmin_k \{d(\mathbf{x}, k)\}$$

$\beta(k)$ is the decoder: $\beta(k) = \mathbf{y}_k$

$\gamma(k)$ is the channel coder: $|\gamma(k)| = -\log \frac{M_k}{N}$

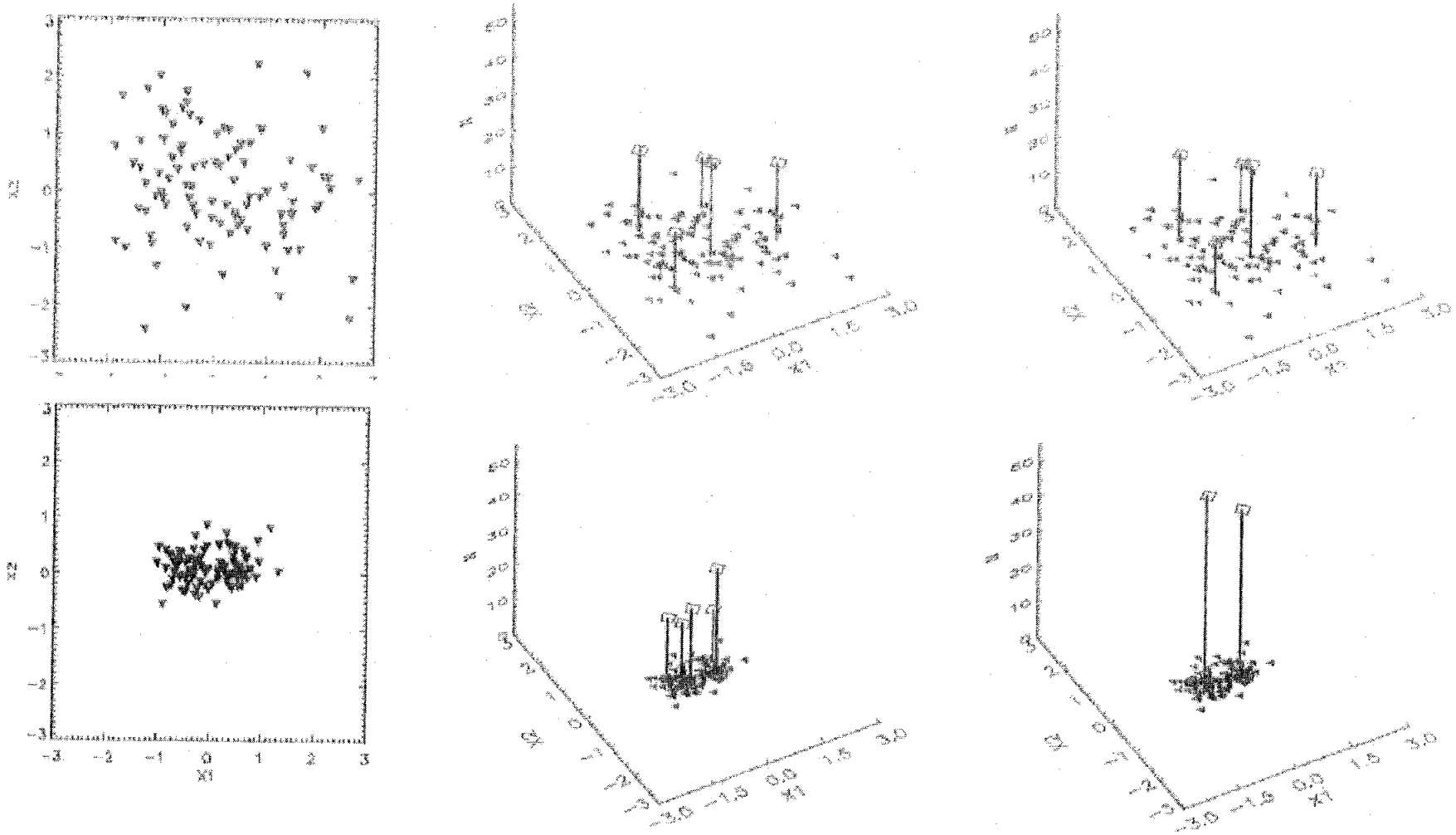
Rate-distortion Theory





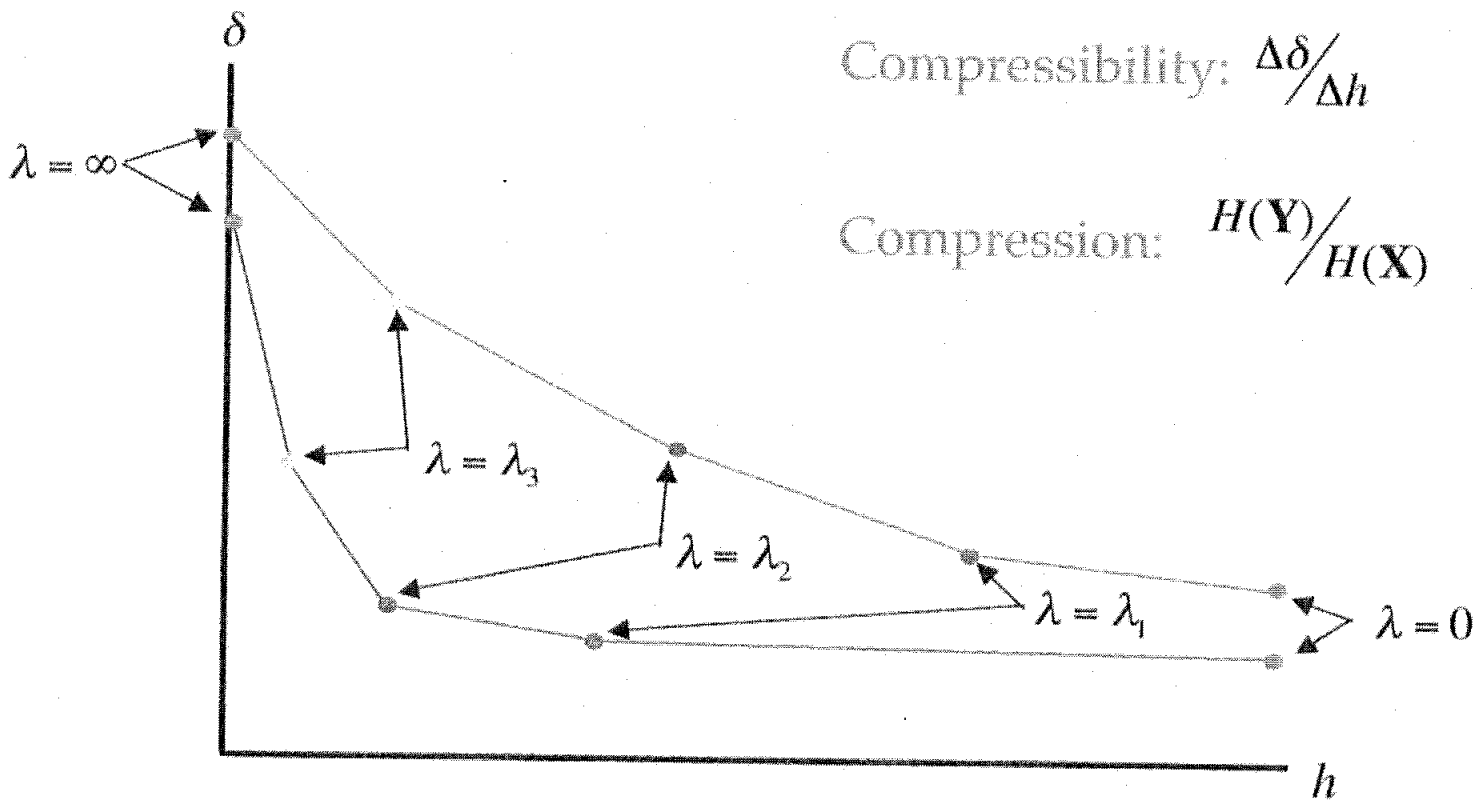
$$\delta(h) = \min_{\{\alpha : H(\mathbf{Y}) \le h\}} \delta(\mathbf{X}, \mathbf{Y})$$

Assuring Comparability and Parsimony



Choose λ to equalize δ .

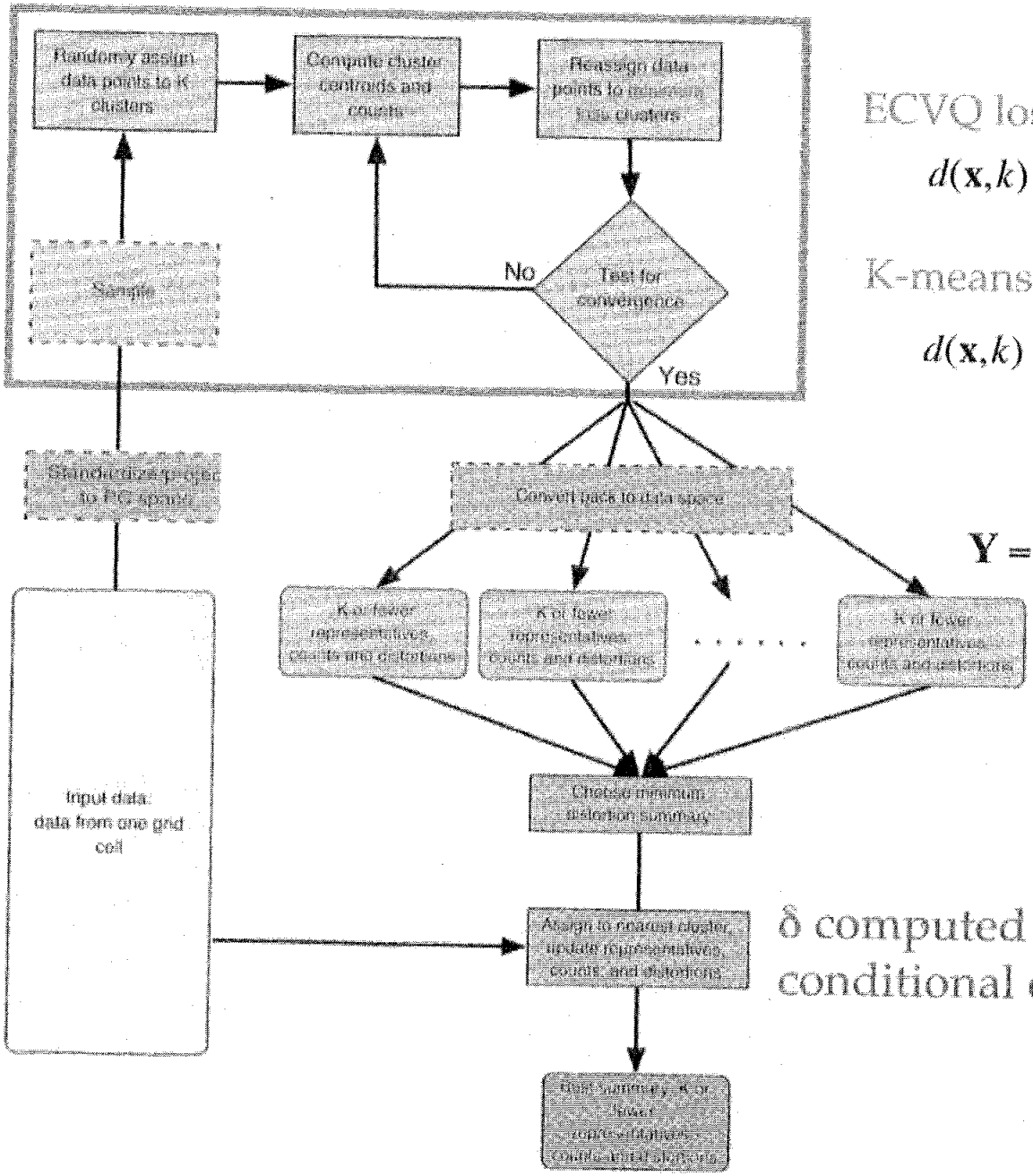
Compressibility and Compression



Compressibility: $\Delta\delta / \Delta h$

Compression: $H(\mathbf{Y}) / H(\mathbf{X})$

Algorithm



ECVQ loss:

$$d(\mathbf{x},k) = \|\mathbf{x} - \mathbf{y}_k\|^2 + \lambda\left(-\log \tfrac{N_k}{N}\right)$$

K-means loss:

$$d(\mathbf{x},k) = \|\mathbf{x} - \mathbf{y}_k\|^2$$

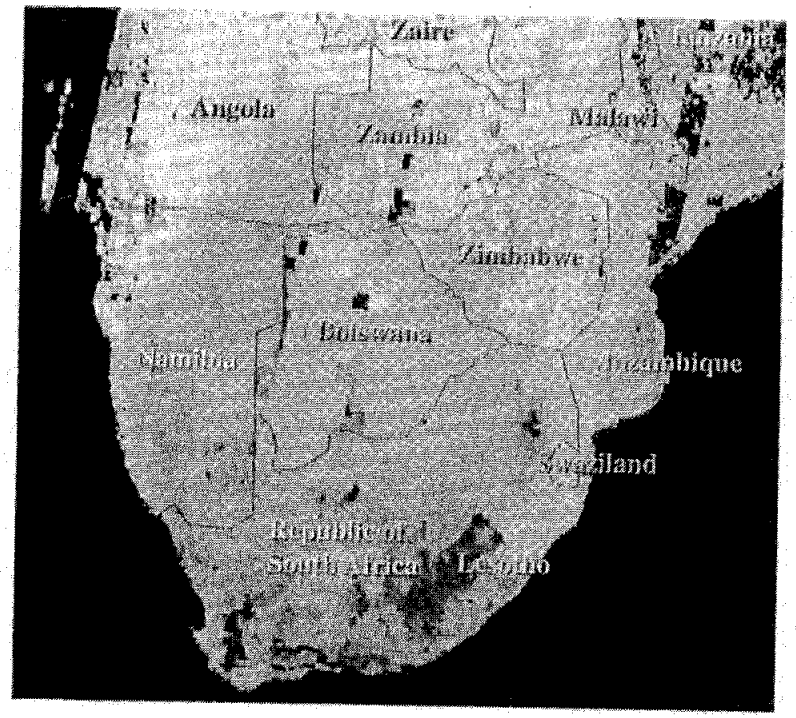
$$\mathbf{Y} = q(\mathbf{X},R) = \sum_{r=1}^{L} q(\mathbf{X},r)1[R=r]$$

δ computed here is conditional on quantizer

MISR Science Analysis

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

$\mathbf{x}_n = (x_{n1}, x_{n2}, x_{n3}, x_{n4}, x_{n5}, x_{n6})'$ where:



x_{n1} = optical depth

x_{n2} = model 1 goodness-of-fit

x_{n3} = model 2 goodness-of-fit

x_{n4} = model 3 goodness-of-fit

x_{n5} = model 4 goodness-of-fit

x_{n6} = model 5 goodness-of-fit

Each observation represents a 17.6 km region with known latitude and longitude.

MISR Science Analysis

- Is optical depth correlated with model discrimination?
- Model discrimination in each grid cell:

$$\overline{w} = \frac{1}{N}\sum_{n=1}^{N} w_n \qquad w_n = \frac{1}{5}\sum_{i=2}^{6}(x_{ni} - \overline{x}_n)^2 \qquad \overline{x}_n = \frac{1}{5}\sum_{i=2}^{6} x_{ni}$$

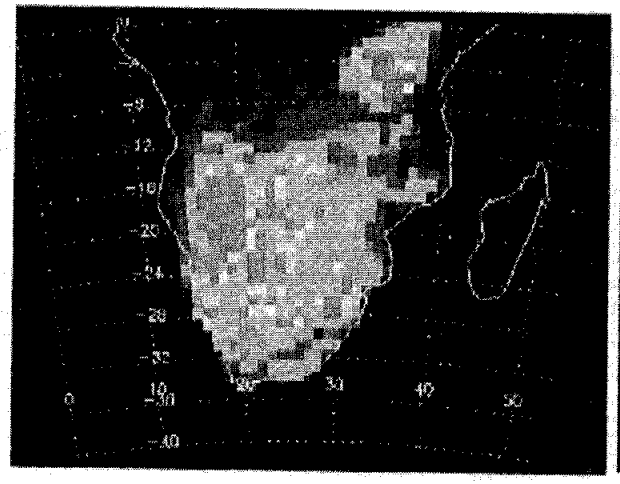
- Correlation:

$$\rho = \frac{\sum_{n=1}^{N}(x_{n1} - \overline{\tau})(w_n - \overline{w})}{\sqrt{\sum_{n=1}^{N}(x_{n1} - \overline{\tau})^2}\sqrt{\sum_{n=1}^{N}(w_n - \overline{w})^2}} \qquad \overline{\tau} = \frac{1}{N}\sum_{n=1}^{N} x_{n1}$$

- Estimate:

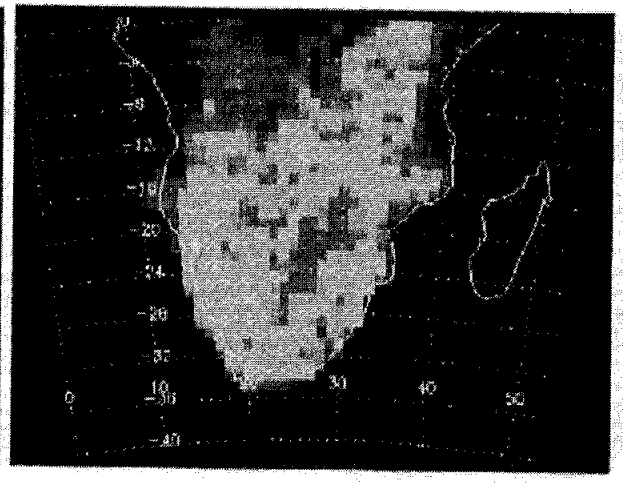
$$\hat{\rho} = \frac{\sum_{k=1}^{K}(y_{k1} - \overline{\tau})(\hat{w}_k - \overline{w})}{\sqrt{\sum_{k=1}^{K} M_k (y_{k1} - \overline{\tau})^2}\sqrt{\sum_{k=1}^{K} M_k (\hat{w}_k - \overline{w})^2}} \qquad \hat{w}_k = \sum_{i=2}^{6}(y_{ki} - \overline{y}_k)^2$$

MISR Science Analysis



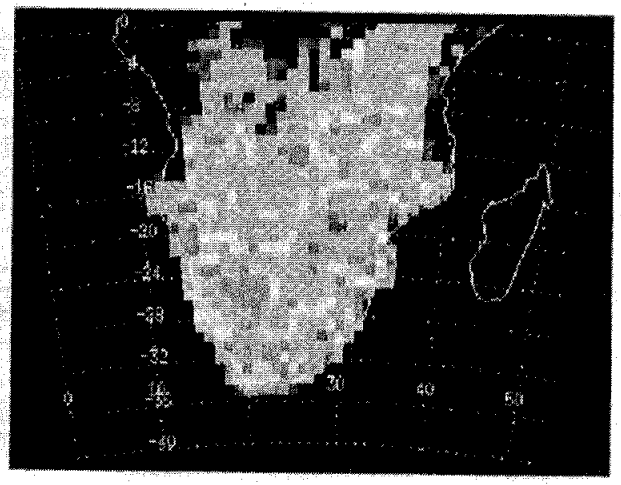
0 ≥ 20,000

N



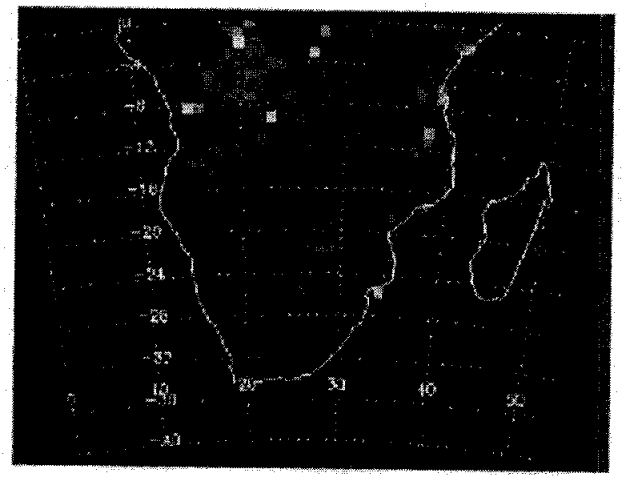
0 40

$\tilde{K}$



0 ≥ 4.5

$H(\mathbf{Y})$

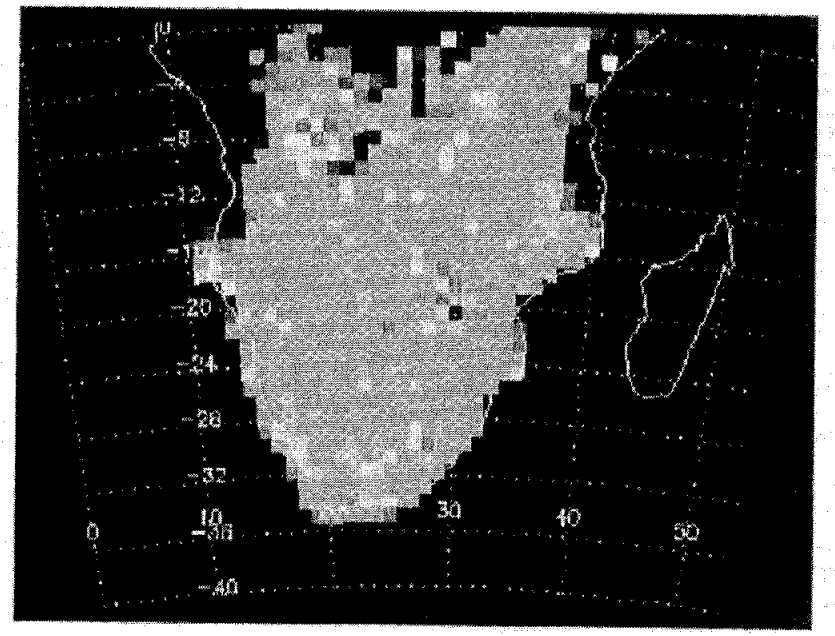


0 ≥ 0.05

$\sqrt{\delta^{rel}}$

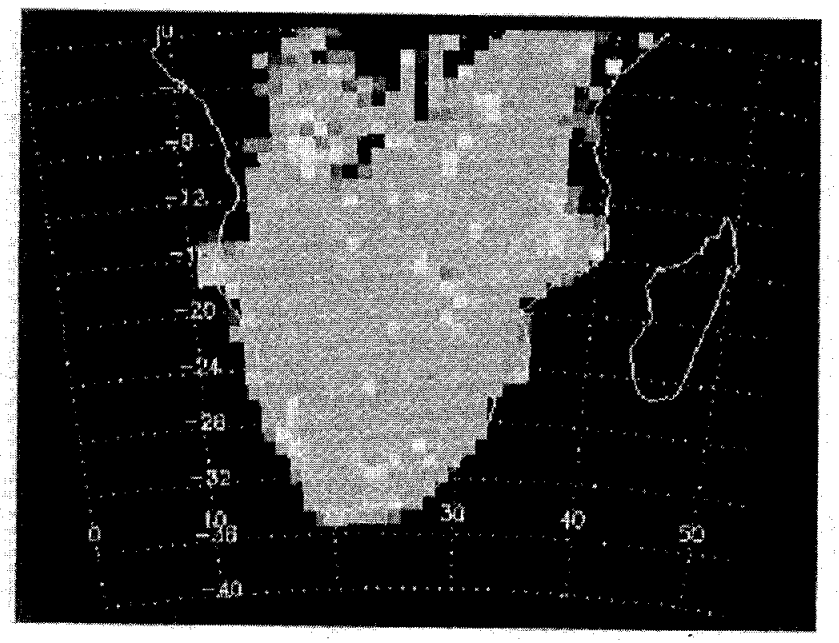
$$\delta^{rel} = \delta \Big/ \sum_{k=1}^{\tilde{K}} \frac{M_k}{N} \|\mathbf{y}_k\|^2$$

MISR Science Analysis



−1 +1

$\hat{\rho}$



−1 +1

ρ

- Accuracy depends on δ *and* form of the transformation, g:

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

Summary

- Accept coarser spatial (temporal) resolution to achieve replication and estimate empirical distributions.
- Calculations based on summaries approximate calculations based on original data. Mean squared errors are known.
- We are *not* making inferences. No spatial statistical models at sub-grid cell resolution. Spatial models possible at grid cell scale.
- Meant for exploratory data analysis of global data sets. Not a replacement for original data.
- Computational demands are significant. Use sampling or two stage version if necessary.
- MISR web site: http://www-misr.jpl.nasa.gov