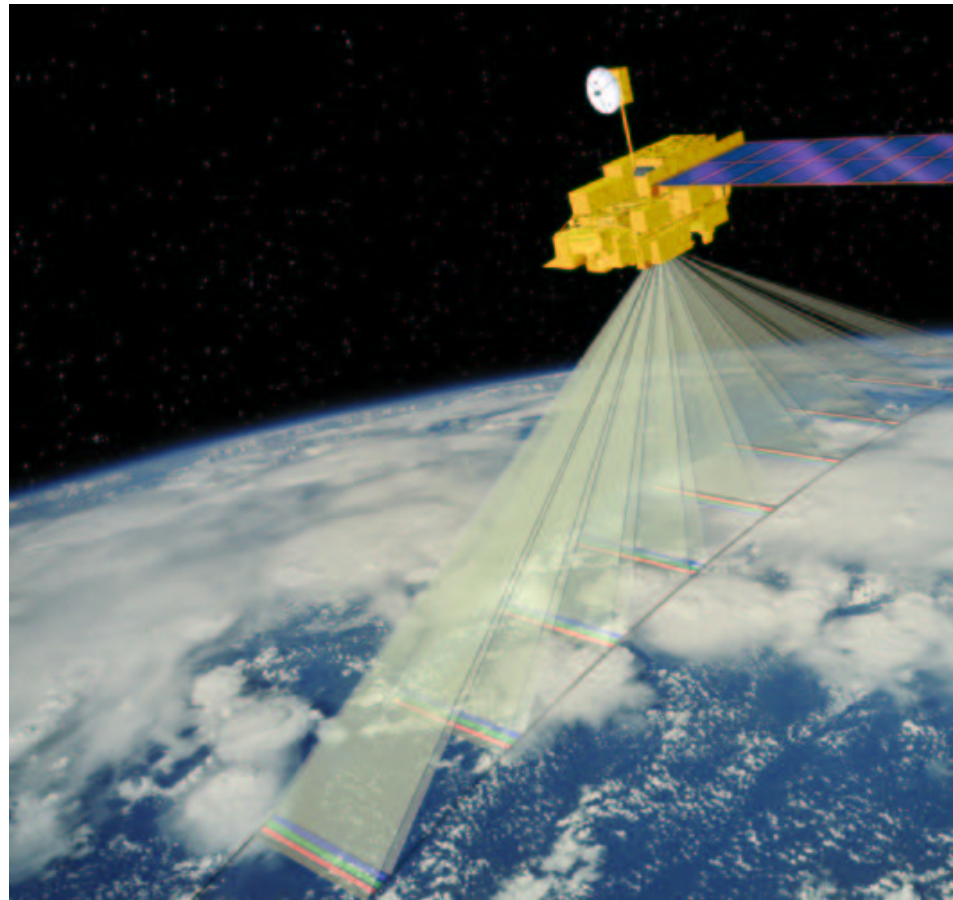




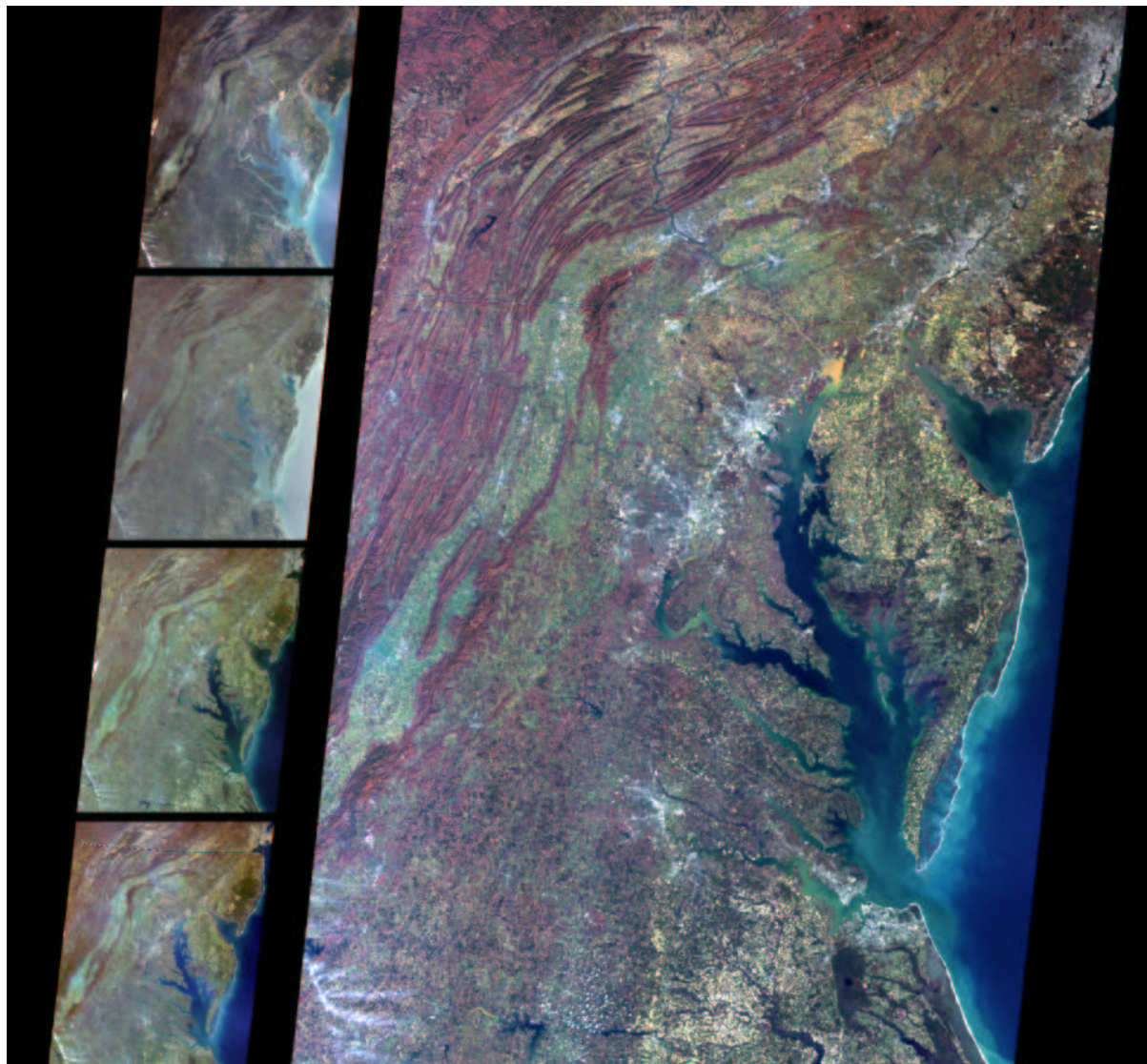
Reducing Size and Complexity of Remote Sensing Data Sets

Amy Braverman
*Jet Propulsion Laboratory,
California Institute of Technology*

- ▶ 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.

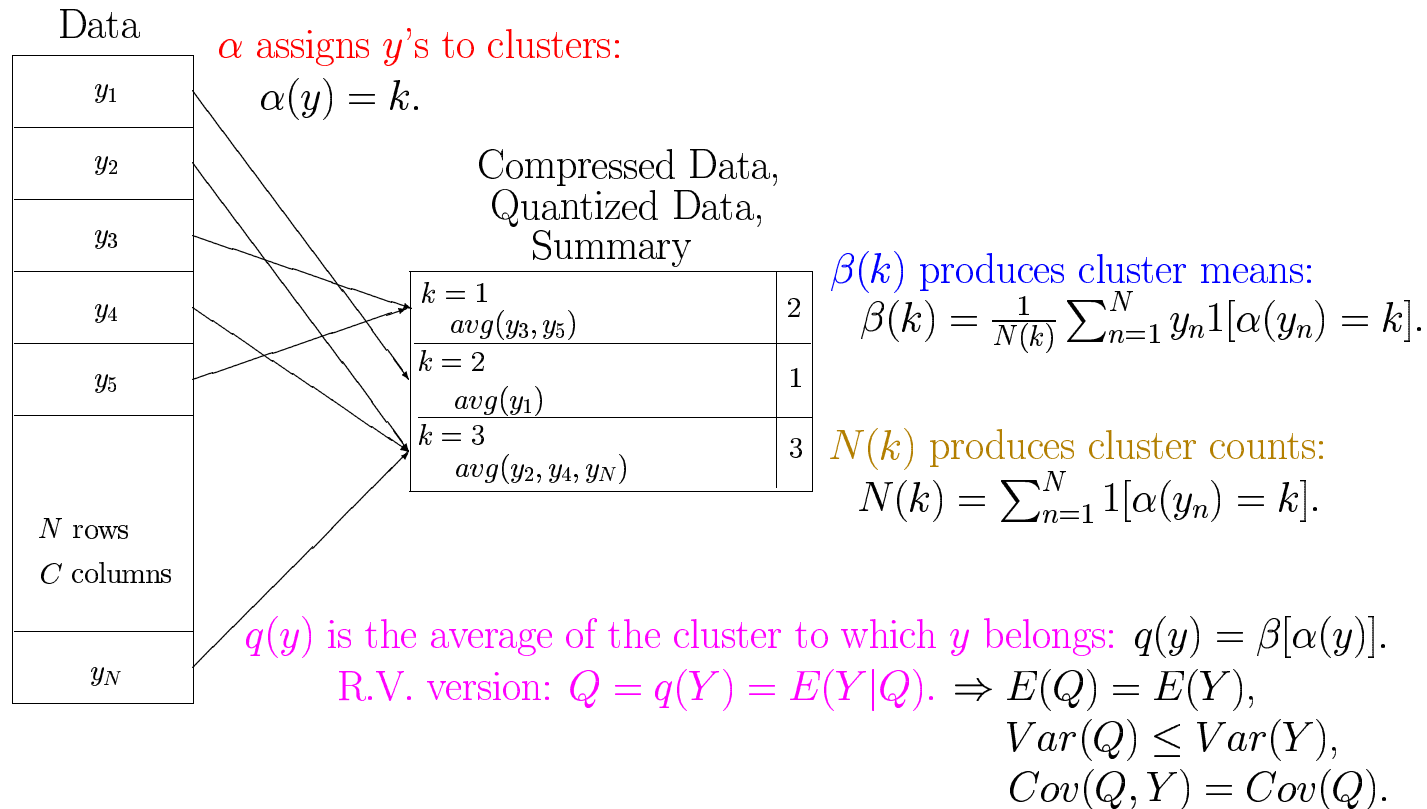


- ▶ **Multi-angle Imaging SpectroRadiometer (MISR) aboard Terra produces about 2 TB per month of radiance and geophysical data.**



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

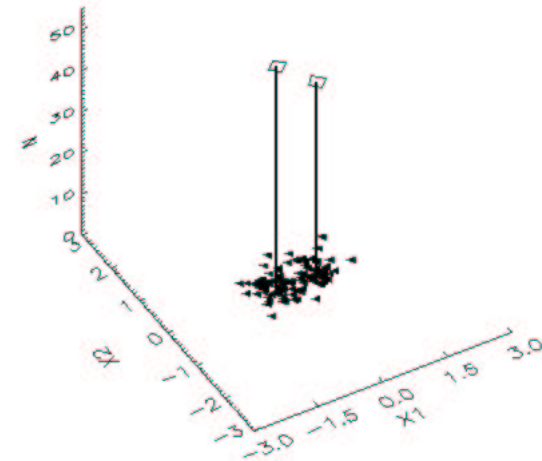
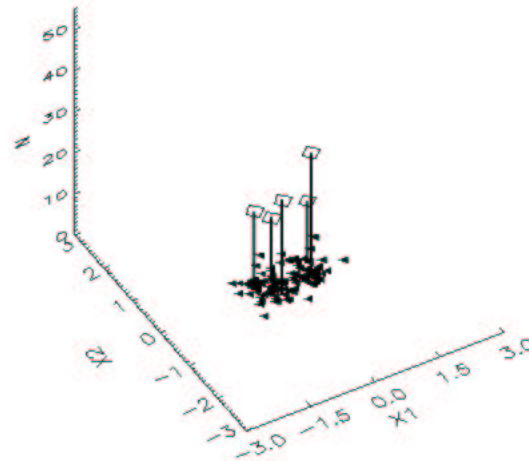
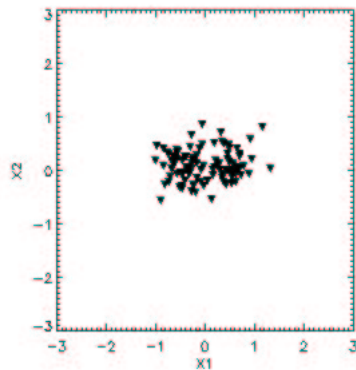
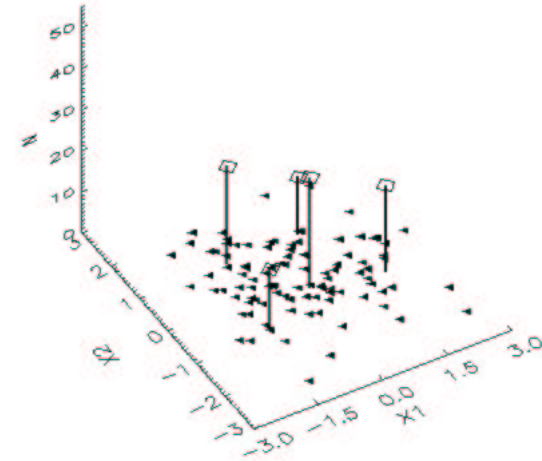
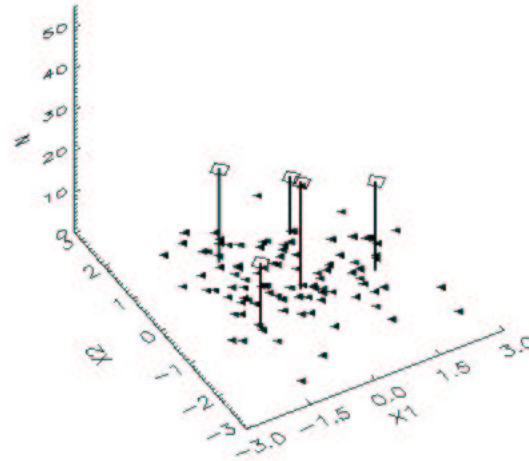
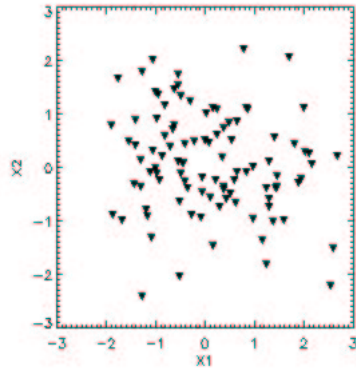
- ▶ Partition data on a monthly, global, one degree by one degree grid.
- ▶ Replace original data in each cell with a set of representatives and associated weights called a summary.
- ▶ Discrete distribution so defined should be close to the original empirical distribution, but also parsimonious.
- ▶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



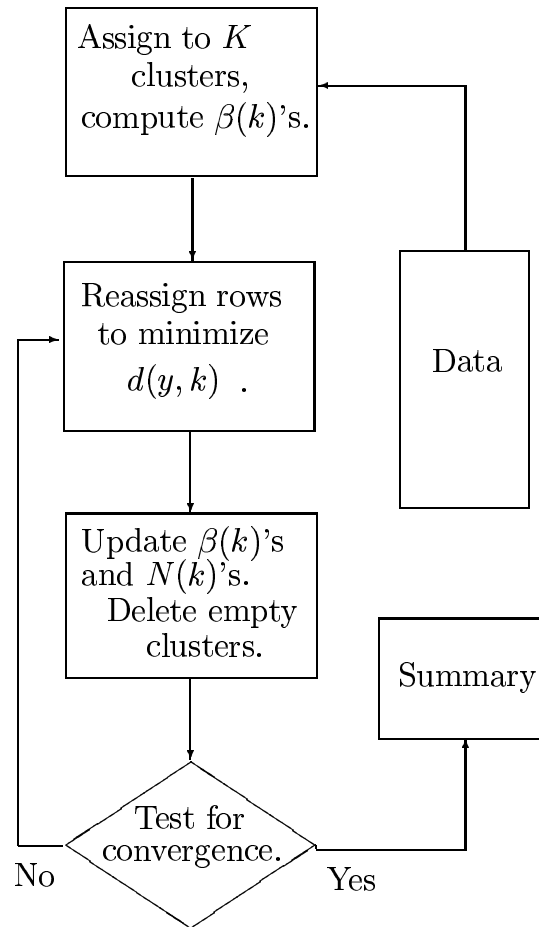
Two figures of merit for q , or equivalently, α :

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

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



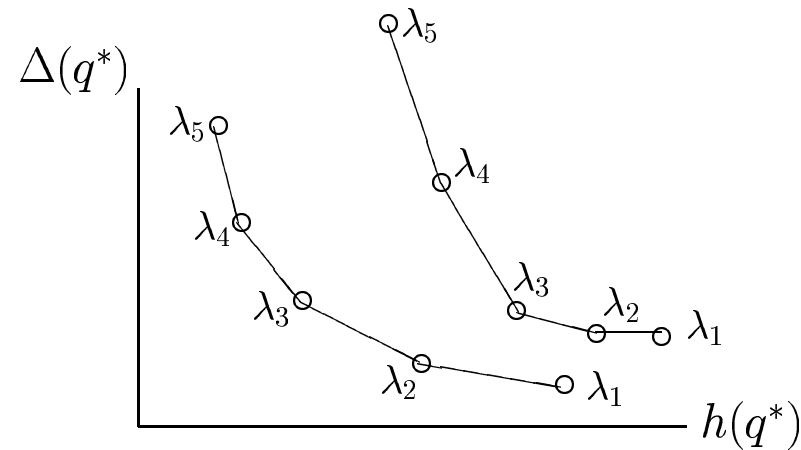
Want fewer clusters when that will do \iff want distortions as similar as possible.



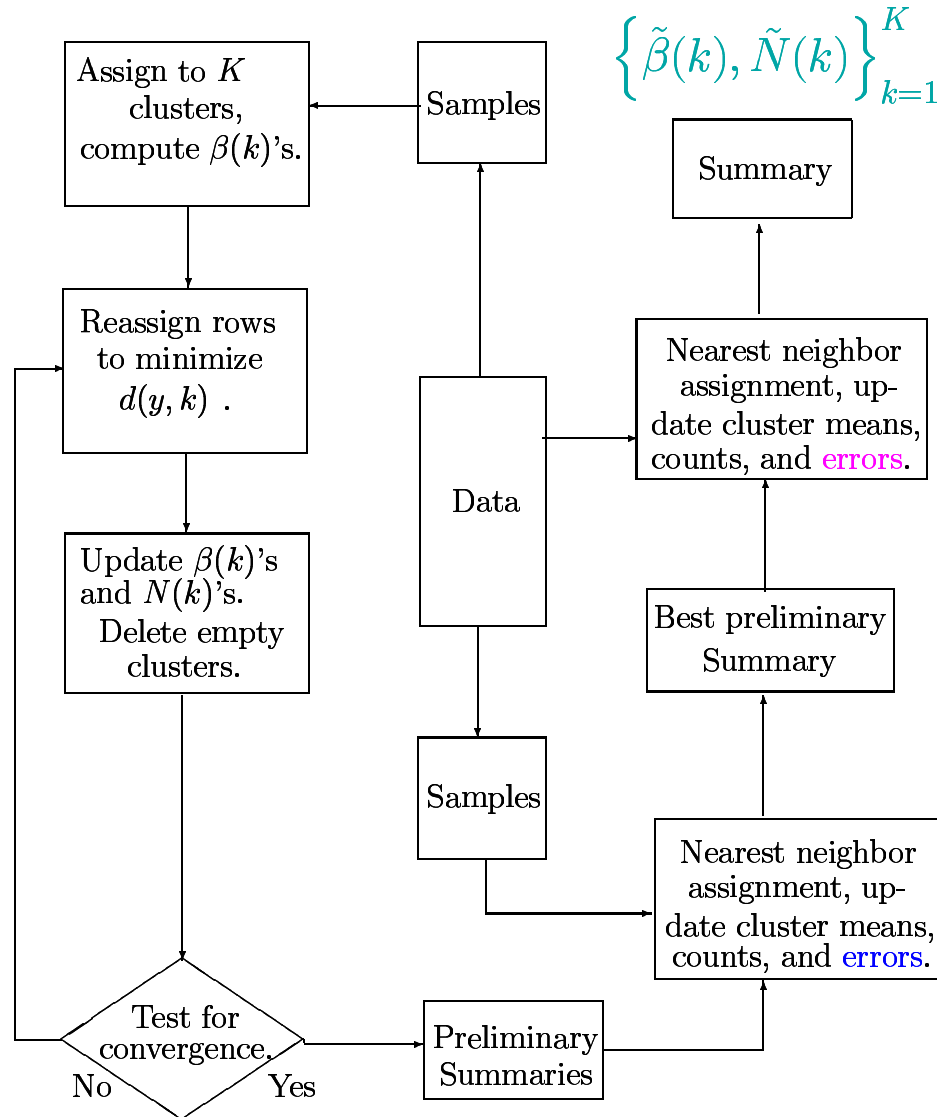
- ▶ Minimize $L_\lambda = \frac{1}{N} \sum_{n=1}^N d(y, k)$ by choice of α :

$$d(y, k) = \|y - \beta(k)\|^2 + \lambda \left[-\log \frac{N(k)}{N} \right].$$

- ▶ Original application: estimate distortion-rate functions of information sources.



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



- ▶ 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.

- ▶ 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 17.6km^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.

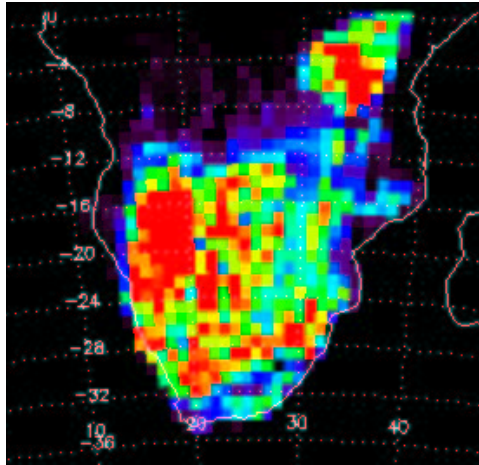
Is there a relationship between optical depth and heterogeneity of the model-fit χ 's?

- ▶ 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:

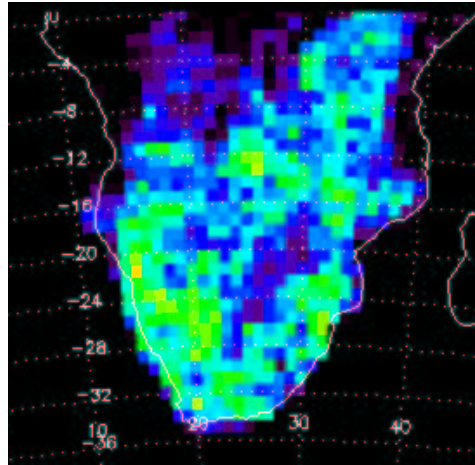
$$\rho(\tau, w) = \frac{\sum_{n=1}^N (\tau_n - \bar{\tau})(w_n - \bar{w})}{\sqrt{\sum_{n=1}^N (\tau_n - \bar{\tau})^2} \sqrt{\sum_{n=1}^N (w_n - \bar{w})^2}},$$

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

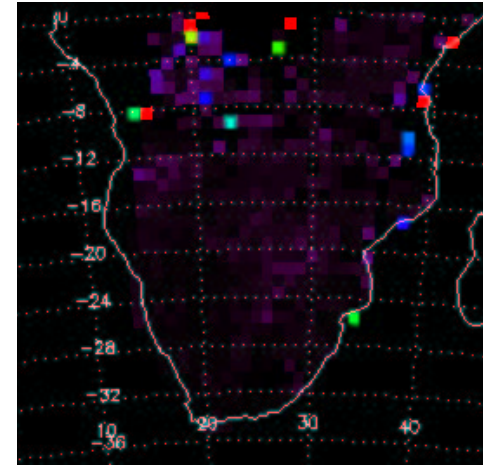
(a) N .



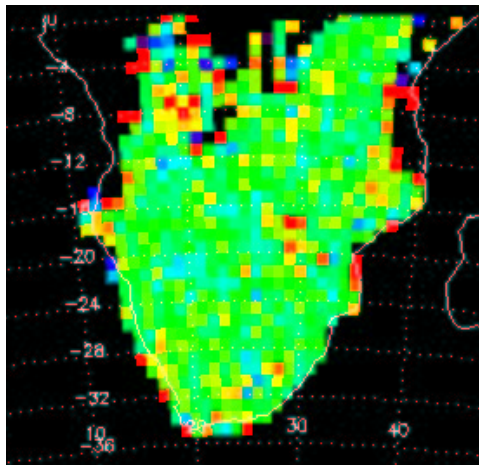
(b) \tilde{K} .



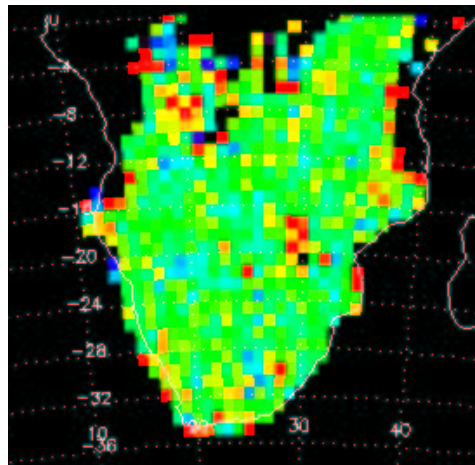
(c) Relative $\sqrt{\tilde{\Delta}}$.



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



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



Key:

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



- ▶ Effectiveness and computational efficiency depend on inherent clustering of the data.
- ▶ Algorithmic improvements: “Fast ECVQ”, summarize in stages.
- ▶ Parameter settings still a bit ad hoc.
- ▶ Provides nonparametric, descriptive summary of the data, not inferential statistics.