Compressing Massive Geophysical Data Sets Using Quantization

Amy Braverman
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109-8099
email: amy@jord.jpl.nasa.gov

Abstract

This paper sets forth a method for compressing massive geophysical data sets. A statistical model for relationships between compressed and uncompressed data is developed and used to evaluate compressors found by an iterative clustering method based on the Entropy-constrained Vector Quantization (ECVQ) algorithm of Chou, Lookabaugh and Gray (1989). The method arbitrates between error induced by compression and level of data reduction. Error includes a component that accounts for uncertainty due to multiple local minima of the ECVQ loss function. Dataset compressibility is identified as an important characteristic to consider when setting the parameter that ultimately determines the balance between error and data reduction. The procedure is demonstrated using a well known data set from the motivating application, Earth science.

1 Introduction

In December 1999 NASA launched its first Earth Observing System (EOS) satellite, *Terra*, into polar orbit. *Terra* carries five instruments designed to study various aspects of Earth's climate systems over the next six years, and will produce vast quantities of high resolution geophysical data. This paper describes a strategy for summarizing this type of data to preserve its high resolution features. The strategy is being developed for one instrument aboard *Terra*, the Multi-angle Imaging SpectroRadiometer (*MISR*). Since *MISR* data are not yet available, this exercise uses a subset of another well known geophysical data set for demonstration purposes. The test data come from the International Satellite Cloud Climatology Project, ISCCP. Both *MISR* and ISCCP are typical of geophysical data sets produced from remote sensing instruments. To provide context and a sense of scale for EOS data, the next section describes *MISR*'s data stream. ISCCP data are described in Section 6.

2 The *MISR* Data Stream

MISR is a set of nine cameras mounted underneath *Terra*, looking down at Earth at nine different angles ($-70.5°$, $-60.0°$, $-45.6°$, and $-26.1°$ aft; $0°$ nadir; $+70.5°$,$+60.0°$, $+45.6°$, and $+26.1°$ forward) along the direction of flight. Each camera has four line arrays of 1504 pixel across the field of view (east-west) perpendicular the flight track (north-south). Each line array is sensitive to one of four wavelengths: NIR, red, green and blue (446, 558, 672 and 866 nanometers), and each pixel views a square region on the ground 275 meters on a side. Thus, one orbital swath on the daylight side of Earth tiles the view into disjoint, contiguous 275 meter spatial regions, and produces 36 radiance measurements for each one. The instrument does not take data as the satellite travels up the night side, so sequential orbits are separated. After 16 days 233 unique but overlapping orbits have completed covering the whole Earth, and every 234th orbit covers the same ground track as the first.

MISR data processing takes this radiance data through several steps. First, data are geometrically and radiometrically calibrated to create the so-called "Level 1" product. There is a seven minute lag between the forward and aft-most views of the same scene. Geometric rectification aligns the observations to produce 36 measurements (nine angles by four wavelengths) associated with the latitude and longitude of each pixel center. There will be about about 30 terabytes per year of Level 1 *MISR* data. Second, Level 2 data are created by converting these 36-vectors into geophysical quantities through complex science algorithms. For example, measurements taken within a 17.6 kilometer area are used to derive aerosol type and amount by matching observed radiances with those predicted by various physical models. Other quantities such as cloud height, wind direction and speed, and fraction of photosynthetically active absorbed radiation are derived at other spatial resolutions (typically 1.1, 2.2, and 35.2 kilometers). This second stage of processing reduces data volume by reducing spatial res-

olution, but increases data volume because many more than 36 geophysical variables are derived. *MISR* will generate about three terabytes a year of Level 2 data.

The third processing step creates monthly summaries of Level 2 derived geophysical data by partitioning the observations according to their membership in cells of a 1° latitude by 1° longitude spatial grid. In the past this Level 3 product has been constructed by simply reporting means, standard deviations and sometimes other simple descriptors of the data in each cell.

Level 3 is essential in view of the massive size of Level 2. NASA will publish *MISR* data for use by researchers at universities and other institutions, many of whom can't work with Level 2 in its entirety. These users must either work with low resolution Level 3 data, or subsets of Level 2. Both alternatives pose difficulties. Low resolution summaries based on means and standard deviations do not capture high resolution multivariate relationships. Subsets do not take advantage of global coverage. Moreover, before examining the data, analysts can only choose subsets based on location and time. There is no way to know which areas and which times contain phenomena of interest without exploring the data first.

The purpose of the method discussed here is to provide a low volume, low resolution road map of the voluminous, high resolution Level 2 data without destroying high resolution relationships. Briefly, the strategy is (each month) to partition Level 2 into 1° latitude by 1° longitude spatial cells, and summarize each cell with a set of representative points and their associated frequencies. Each representative point stands for some number of original observations, that number being given by frequency. The combination of representatives and counts is a compressed version, or summary, of the original data. The algorithm used to find the clusters and their representatives is a modification of the entropy-constrained vector quantization algorithm (ECVQ) of Chou, Lookabaugh and Gray (1989).

3 ECVQ

ECVQ is an iterative algorithm that groups data into a collection of disjoint clusters so as to minimize the loss function

$$L_\lambda = \sum_{n=1}^{N} \|y_n - q(y_n)\|^2 + \lambda \left[-\log \frac{f(y_n)}{N} \right], \qquad (1)$$

where y_n is the nth row of an $N \times C$ data matrix (representing one spatial cell in one month, for example). $q(y_n)$ is the representative of the group to which y_n is assigned, $f(y_n)$ is the number of data points (rows) assigned to the same group as y_n, and λ is a fixed constant. $-\log(N(k)/N)$ is a positive number and varies inversely with $N(k)$. Thus, if $\|y_n - \beta(k_1)\|^2 = \|y_n - \beta(k_2)\|^2$, y_n would be assigned to cluster k_2 if $N(k_2)$ is larger than $N(k_1)$. When $\lambda = 0$, L_λ is euclidian distance, and ECVQ is equivalent to the batch version of the K-means clustering procedure (MacQueen, 1967).

The function $q(\cdot)$, called a quantizer, can be written as the composition of two functions, $q(y_n) = \beta(\alpha(y_n))$. α, called the encoder, takes a C-dimensional data point and returns an integer providing the index of the cluster to which y_n is assigned. β, the decoder, takes the index, k, and returns the representative for cluster k. In this application $\beta(k)$ will always be the mean vector of cluster k. Also, define $N(k)$ as the number of y_n assigned to cluster k:

$$\begin{aligned} N(k) &= \sum_{n=1}^{N} 1[\alpha(y_n) = k], \\ \beta(k) &= \frac{1}{N(k)} \sum_{n=1}^{N} y_n 1[\alpha(y_n) = k]. \end{aligned}$$

Here is a brief description of the ECVQ algorithm:

1. Fix the maximum number of clusters allowed, K, and the compression parameter, λ.
2. Arbitrarily assign the y_n's to the K clusters by specifying initial values for $\alpha(y_n)$. Compute means and frequencies of these clusters, and denote them $\beta(k)$ and $N(k)$ respectively, for $k = 1, 2, \ldots, K$.
3. Reassign each y_n to the cluster with the smallest loss:
$$\begin{aligned} \alpha(y_n) = argmin_k \Big\{ &\|y_n - \beta(\alpha(y_n))\|^2 \\ &+ \lambda \left[-\log \frac{N(\alpha(y_n))}{N} \right] \Big\}. \end{aligned}$$
4. Update $\beta(k)$ and $N(k)$ for all k.
5. Eliminate any clusters for which $N(k) = 0$.
6. Repeat steps (3), (4) and (5) until convergence.

The ECVQ solution has the property that the $\beta(k)$'s are the means of the y_n's they represent. This is a feature that will be important in Section 4.

The algorithm is guaranteed to converge in a finite number of steps, but not necessarily to either a local or global minimum. However the solution improves on the starting point, and provides a sensible summary of the y_n's in the sense described by MacQueen: "The point of view taken in this application is *not* to find some unique,

definitive grouping, but rather to simply aid the investigator in obtaining qualitative and quantitative understanding of large amounts of ... data by providing him with reasonably good similarity groups." (MacQueen, 1967, page 288.)

To apply ECVQ to subsets created by partitioning large or massive geophysical data sets, two modifications are made. First, a sample of M rows is chosen. ECVQ is applied to it, and an initial set of representatives, $\{\beta^*(k)\}_{k=1}^{K^*}$, obtained. The second modification is to pass through the y_n's again assigning them to their nearest $\beta^*(k)$. Empty clusters are deleted, and representatives and counts updated to reflect the second pass. In other words, a preliminary set of representatives is determined from a sample, then the entire data set is clustered using them. The ultimate set of clusters and counts thus reflects *all* the data. The first step of this procedure is called the design step, and the second is called the binning step. The final subset summary is $\left\{\tilde{\beta}(k), \tilde{N}(k)\right\}_{k=1}^{\tilde{K}}$.

Since the design step is carried out on a sample, the final summary is subject to sampling variation. The next section describes a statistical framework in which to assess that variability along with the quality and parsimony of summaries.

4 A Statistical Model for Compressed Data

Consider a randomly drawn row from the $N \times C$ data matrix representing one subset of a partitioned massive geophysical data set. Let that draw be the random vector Y. Y has the empirical distribution of the subset; $P(Y = y_n) = 1/N$. Now let $Q = q(Y)$ for some quantizer function q obtained from ECVQ. Q is a deterministic function of Y with the property that $Q = E(Y|Q)$. This property is called *self-consistency* of Q for Y by Tarpey and Flury (1996).

Self-consistency imparts several important properties on Q as an estimate of Y:

$$E(t'Q) = E(t'Y), \tag{2}$$

$$Var(t'Q) \leq Var(t'Y), \tag{3}$$

$$Cov(Y - Q) = Cov(Y) - Cov(Q), \tag{4}$$

$$E\|g(Y) - g(Q)\|^2 \approx \sum_{ij} E\left[\dot{g}_{(i)}\dot{g}_{(j)}E(Y_{(i)}Y_{(j)}|Q) - Q_{(i)}Q_{(j)}\right], \tag{5}$$

where the subscript (i) indicates the ith component of Y or Q, and $\dot{g}_i = \partial g(Q)/\partial Q_{(i)}$. In (2) and (3) t is a $C \times 1$ vector. (2) and (3) show that linear functions of Q are unbiased estimates of the same functions of Y, and have lower variance. (4) follows from $Cov(Y,Q) = E(YQ') - [E(Y)][E(Q)]' = E\left[E(YQ'|Q)\right] - [E(Q)][E(Q)]' = E\left[E(Y|Q)Q'\right] - [E(Q)][E(Q)]' = E(QQ') - [E(Q)][E(Q)]' = Cov(Q)$, and shows that the covariance of the error in Q as an estimate of Y is the difference in the covariance matrices of Q and Y. Finally, the mean squared error between an arbitrary continuous function of Y and its estimate computed from Q is approximated by (5).

Two figures of merit are used to judge the quality of q. Distortion, $\delta(Y,Q)$, measures mean squared error of Q as an estimate of Y:

$$\delta(Y,Q) = E\|Y - Q\|^2 = trCov(Y - Q). \tag{6}$$

Data reduction, $\Delta(Y,Q)$, is the difference in entropies between Y and Q. The entropy of a discrete random variable such as Q is

$$h(Q) \quad = \quad -\sum_{\xi} P(Q = \xi)\log P(Q = \xi)$$

where ξ indexes realizations of Q. Therefore, data reduction is

$$\begin{aligned}\Delta(Y,Q) &= h(Y) - h(Q) \\ &= \log N - \left[-\frac{1}{N}\sum_n \log\frac{N(\alpha(y_n))}{N}\right] \\ &= E[\log N(\alpha(Y))], \end{aligned} \tag{7}$$

where $N(\alpha(y_n))$ is the population of the cluster to which y_n is assigned by the quantizer's encoder.

Generally, quantizers with high data reduction tend to have large distortion and vica-versa. Consider two extremes. The identity quantizer (which can be obtained from ECVQ by setting $K = N$ and $\lambda = 0$) assigns each data point to its own cluster, and therefore leaves the data unchanged. This produces no distortion, but no data reduction either. $N(k) = 1$ for each of the N clusters, $\Delta(Y,Q) = \log 1 = 0$. At the other extreme, (as when ECVQ is run with $K = 1$) all data points are assigned to a single cluster for which the representative is the subset centroid. Distortion is $trCov(Y)$, the maximum possible value when the decoder is the mean function. Data reduction is also at its maximum, $\log N$. Finding a good intermediate solution means finding a good compromise between data reduction and distortion. Under ECVQ the compromise depends on the parameter λ. More will be said about this in Section 6.

Now suppose that R quantizer functions are available, each with the self-consistency property, and one of them

is chosen at random to quantize Y. This models the situation in which a random training set is used in the design step.

$$Q_* \quad = \quad \sum_r q_r(Y) 1[\rho = r],$$

where ρ is an integer valued random variable specifying which quantizer is selected, and q_r is the rth quantizer function. Q_* is called a random quantizer, and inherits some but not all properties of Q. Linear functions of Q_* are unbiased for corresponding functions of Y, and $Cov(Y - Q_*) = Cov(Y) - Cov(Q_*)$ even though Q_* is *not* self-consistent for Y. The approximation (5) holds with Q_* substituted for Q.

Distortion and data reduction of Q_* are analogous to those of Q except that expectations in (6) and (7) are with respect to both Y and ρ. Y and ρ are independent, but Q_* is jointly distributed with both of them. To evaluate quantizers produced by ECVQ from training sets, it is necessary to consider variation over training sets, as embodied by variation of ρ. For distortion, write

$$\begin{aligned} \delta(Y, Q_*) &= E\left[E\|Y - Q_*\|^2 | \rho\right] \qquad (8) \\ &= trCov(Y - Q_*) \\ &= trE[Cov(Q_*|Y)] + \\ &\quad trE\left\{[Y - E(Q_*|Y)][Y - E(Q_*|Y)]'\right\}. \qquad (9) \end{aligned}$$

(8) provides the basis for estimating $\delta(Y, Q_*)$ from the simulation discussed in the next section. (9) shows that $\delta(Y, Q_*)$ can be decomposed in a different way; into components related to bias and variance. The bias component is the second term on the right side of (9), and gives a measure of the average squared amount by which a randomly quantized random draw from the data deviates from its true value. The first term on the right side of (9) is a measure of instability in quantizer selection. Though this decomposition is useful, estimation of bias and instability is not pursued here. Data reduction for a randomly chosen quantizer is

$$\Delta(Y, Q_*) = h(Y) - h(Q_*) = \log N - h(Q_*). \qquad (10)$$

5 Estimating Random Quantizer Performance

When ECVQ is applied to a sample and the result used to cluster the parent data, it is important to characterize distortion and data reduction in a way that accounts for sampling variation. Suppose S random summaries are generated for the same spatio-temporal subset. Let δ_s be the distortion of the sth summary obtained. Let h_s be the entropy of the distribution of the summary obtained on the sth trial. Define

$$\begin{aligned} \delta_s &= \frac{1}{N}\sum_{n=1}^{N} \|y_n - \beta(\tilde{\alpha}_s(y_n))\|^2, \\ h_s &= -\frac{1}{N}\sum_{n=1}^{N} \log \frac{N(\tilde{\alpha}_s(y_n))}{N}, \end{aligned}$$

where $\tilde{\alpha}_s$ is the final encoder obtained on the sth trial. An estimate of $\delta(Y, Q_*)$ is

$$\bar{\delta} \quad = \quad \frac{1}{S}\sum_{s=1}^{S} \delta_s.$$

However

$$\bar{h} \quad = \quad \frac{1}{S}\sum_{s=1}^{S} h_s$$

does not estimate $h(Q_*)$; it estimates conditional entropy, $h(Q_*|\rho)$. Conditional entropy is defined as

$$\begin{aligned} h(Q_*|\rho) &= \sum_r P(\rho = r) h(Q_*|\rho = r), \\ h(Q_*|\rho = r) &= -\sum_{\xi_r} P(Q_* = \xi_r) \log P(Q_* = \xi_r) \end{aligned}$$

(Cover and Thomas, 1991, page 16). Since $h(Q_*|\rho) \leq h(Q_*)$,

$$\hat{\Delta} \quad = \quad \log N - \bar{h}$$

is an estimated upper bound for $\Delta(Y, Q_*)$, not a point estimate.

Now, armed with S summaries of the same data it would be foolish not to use the best one. The summary with the minimum δ_s will be used to summarize the data even though the estimate of $\delta(Y, Q_*)$ applies to a process in which a summary is chosen at random. Let s^* be the index of the summary with the smallest value of δ_s. Then $\delta_{s^*} \leq \delta_s$, and $E(\delta_{s^*}) \leq E(\delta_s)$. Since $\bar{\delta}$ estimates $E(\delta_s)$, $\bar{\delta}$ is an estimated upper bound on the distortion for the process which selects the best summary.

6 Application to ISCCP

To illustrate estimation of distortion and data reduction, and how their estimates relate to the choice of λ, the method in Section 3 is applied to a test data set. The test data come from the International Satellite Cloud Climatology Project (ISCCP). ISCCP has collected data on cloud type and amount from a multitude of satellites

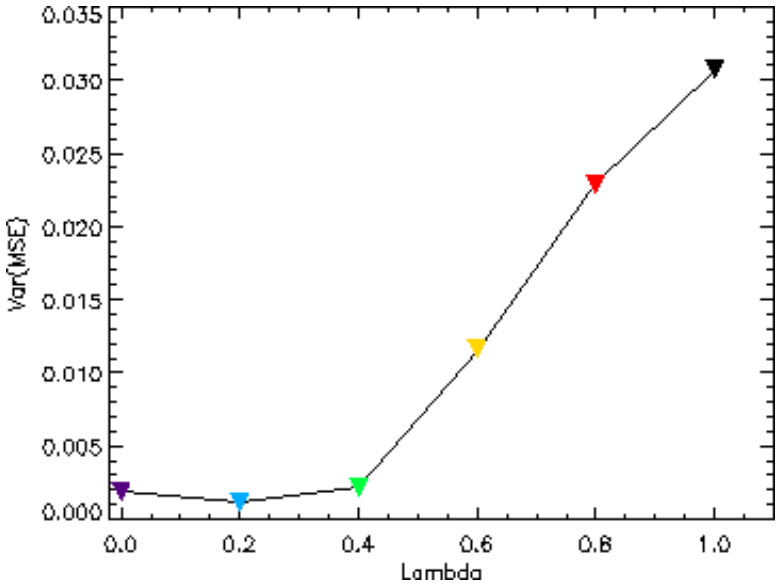

Figure 1: Variance of $\bar{\delta}$ over subsets.

beginning in 1983. Raw data are visible and infrared radiances measured for pixels 25 kilometers on a side. Each measurement has a latitude and longitude (measured to the nearest tenth of a degree) and a date and time. Time is measured in three hour intervals (00, 03, 06, etc. GMT). Radiances are combined with other information (e.g. surface topography) and run through science algorithms to derive the ISCCP equivalent of *MISR*'s Level 2: cloud top pressure (PC) and optical thickness (TAU). Optical thickness is a unitless measure of attenuation of energy through a column of atmosphere. Cloud top pressure is measured in millibars. For cloudy pixels cloud top pressure and optical thickness determine cloud type via the decision matrix in Table 1.

Cloud Top Pressure (MB)	*Cloud Optical Thickness*		
	0 to 3.6	3.6 to 23	23 to 379
50 to 440 (High)	Cirrus	Cirro-stratus	Deep Convection
440 to 680 (Middle)	Alto-cumulus	Alto-stratus	Nimbo-stratus
680 to 1000 (Low)	Cumulus	Strato-cumulus	Stratus

Table 1: ISCCP radiometric cloud classification.

Test data are daytime PC and TAU measurements from July 1991, in pixels identified by ISCCP data processing as cloudy, and located in the northern half of the western hemisphere (latitude 0° to 90°, longitude 0° to 180°) . This yields a relatively small data set having about 5.2 million observations; small enough that true data set quantities can be calculated and compared to their analogues obtained from compressed data.

In this exercise test data are partitioned into subsets on a 10° by 10° grid. For example, all measurements with latitudes between zero and 9.9 (inclusive) and longitudes between 20 and 29.9 (say) belong to the subset for the cell with southwest corner $(0, 20)$. (Cells are identified by the coordinates of their southwest corners.) There are roughly 30,000 observations per cell. This is comparable to the population of a 1° by 1°*MISR* cell, though *MISR* data will have higher dimensionality and more cells.

For each of the 162 ISCCP subsets, a nine element vector of cloud type proportions, p, is obtained from the raw data. Each element is the proportion of cloudy pixels in the subset determined to be of one of the nine cloud types. Then the raw data are compressed, and an estimate of the cloud type proportion vector constructed from the summary. This is a simple example of the type of transformation often applied to geophysical data. Cloud type is a non-linear, discontinuous function of PC and TAU. The hope is that the same function applied to compressed data yields a reasonable estimate

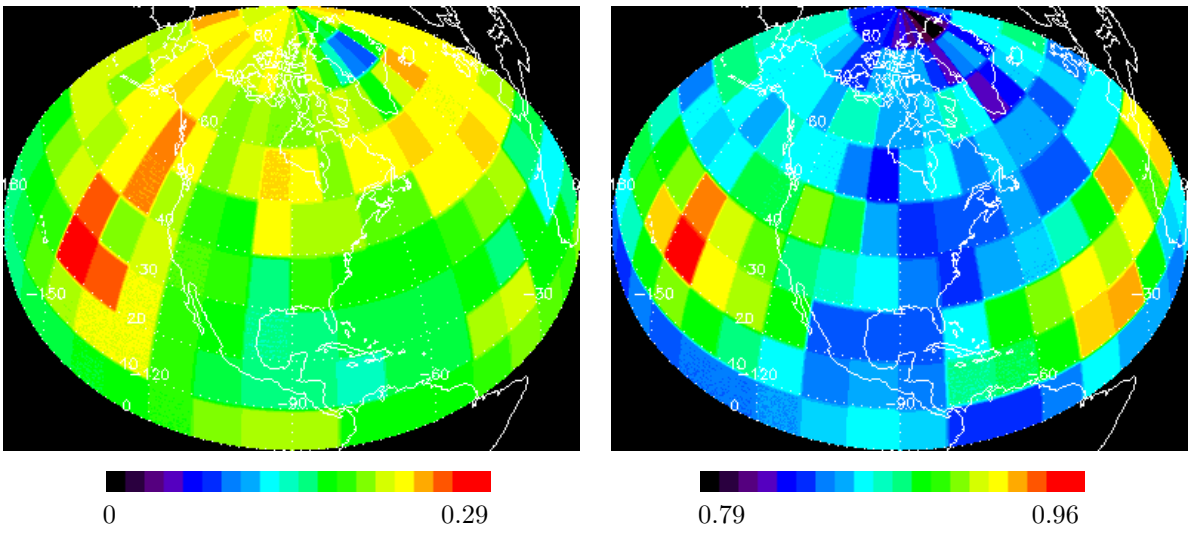

Figure 2: Relative $\bar{\delta}$ (left) and relative $\hat{\Delta}$ (right), by subset, for data compressed at $\lambda = 0.2$.

of the "true" value computed from raw data.

ECVQ is applied to all 162 subsets independently using training sets of size $M = 500$, $K = 10$ for the initial number of clusters, and six different values of λ: $0, 0.2, 0.4, 0.6, 0.8$, and 1.0. This is done $S = 300$ times. Before starting, subsets were standardized using the means and standard deviations of all 5.2 million data points.

Figure 1 is a plot of the variances of the $\bar{\delta}$'s over subsets as a function of λ. The figure shows the variance is minimized at $\lambda = 0.2$. Minimum variance in $\bar{\delta}$ over subsets is grounds for choosing $\lambda = 0.2$ because at this value subset summaries are of similar quality, and so differences between summaries reflect differences in the data, not differences in how well summaries match their parent subsets.

Figure 2 shows estimates of $\bar{\delta}$ and $\hat{\Delta}$ for $\lambda = 0.2$ in the form of maps. $\bar{\delta}$ for a given subset is expressed relative the average data vector norm in that subset. The left panel shows $N\bar{\delta}/\sum_{n=1}^{N} \|y_n\|$. $\hat{\Delta}$ is expressed relative to raw subset complexity. The right panel shows $[\log N - \hat{\Delta}]/\log N$. In most cases relative distortion is in the 15 percent range. Relative data reduction is high, ranging from 79 to 96 percent of raw data complexity.

Generally, subsets with high data reduction also have high distortion, but there are exceptions. For example, regions over the north Atlantic Ocean and Great Britain show relatively high distortion and relatively *low* data reduction, suggesting that data in these subsets are complex and difficult to compress. In contrast, the region with southwest corner (20,-160) has relatively high data reduction and relatively low distortion. Setting λ is like setting the contrast level in an image: it should be sensitive to the "dynamic range" of data values across subsets. The fact that $\lambda = 0.2$ reveals compressibility differentials suggests it is appropriately sensitive, and thus represents a good compromise between distortion and data reduction.

The minimum distortion summary of each subset derived at $\lambda = 0.2$ is adopted. Cluster representatives were destandardized for purposes of computing the estimate of p. Section 4 showed properties of linear functions of compressed data as estimates of those same functions of raw data, and how well mean squared errors for nonlinear functions could be estimated. However, equation (5) cannot be applied to discontinuous functions like the one that converts cloud top pressure and optical thickness into cloud type.

To get an idea how well cloud type proportions are estimated from compressed data, Figure 3 compares cloud

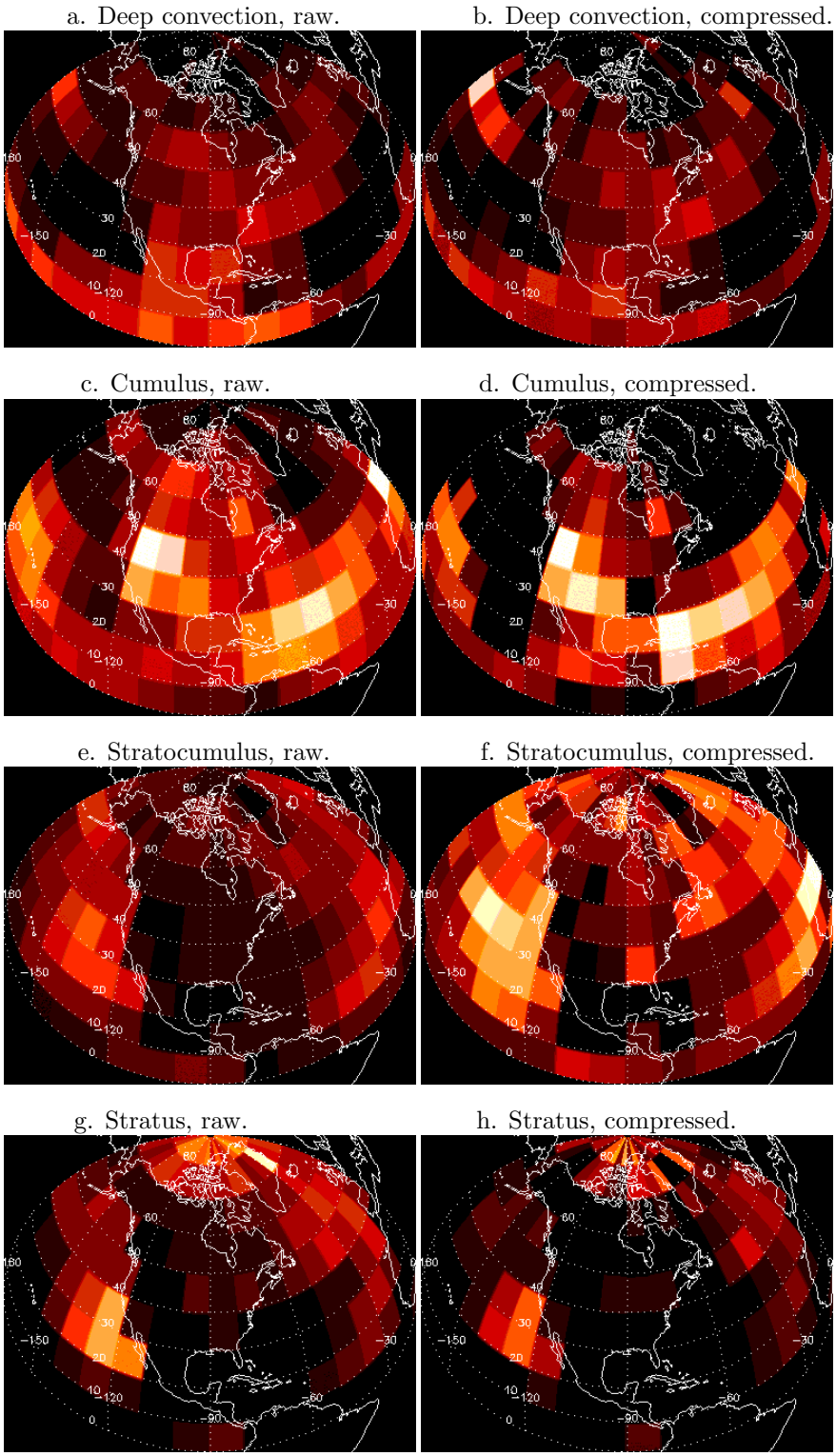

Figure 3: Cloud type proportions computed from raw data and data compressed at $\lambda = 0.2$

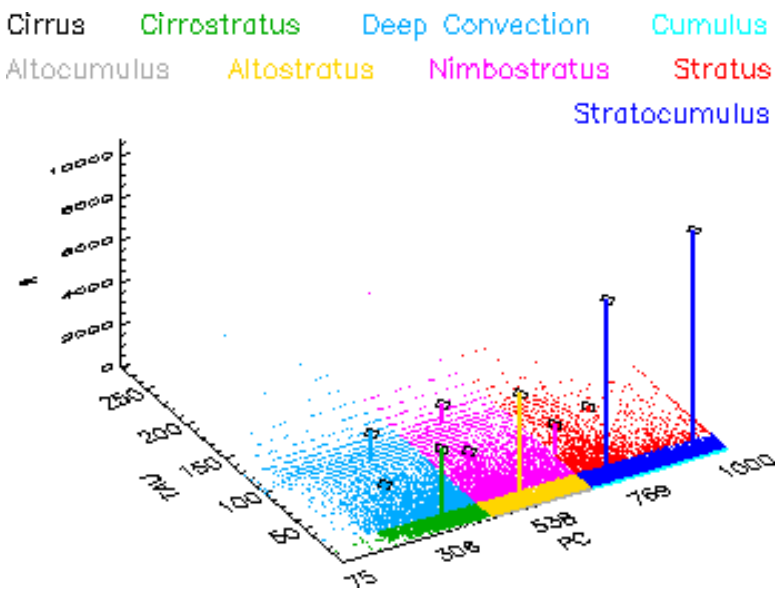

Figure 4: Raw and compressed data for cell $(40, -70)$.

type proportions computed from raw and compressed data for four cloud types representing the best and worst performance among the nine types. (Colors bars are not shown in the figure. The ranges of values depicted from dark to light are as follows. Deep convection: zero to 0.28. Cumulus: zero to 0.60. Stratocumulus: zero to 0.88. Stratus: zero to 0.27.) Overall, quantized maps show less spatial continuity than their uncompressed counterparts. Visually, stratocumulus is the worst, overestimating its proportion in a many areas. Even so, "hot spots" are identifiable.

The artifically large ten degree grid at least partially accounts for discrepancies in Figure 3. Two artifacts are to blame. First, large grid cells will necessarily create a blockier visual effect than smaller ones. Second, and more important, ten degree cells are too heterogenous with respect to cloud top pressure and optical thickness. Figure 4 shows why. The floor of Figure 4 is a scatterplot of cloud top pressure and optical thickness for one 10° spatial region, the region over Nova Scotia (40,-70). Positions of the cluster means are shown by locations of the spikes, and heights of spikes show cluster populations. Cloud types are shown in different colors for both raw data and cluster representatives according to the key.

Sharp and somewhat arbitrary boundaries between cloud types are not reflected in any cluster structure because the data points blanket *PC-TAU* space. Spikes located just inside the region corresponding to one cloud type are frequently the best representative for points in regions belonging to other types. This is less often the case on a finer grid, where one or two cloud types tend to dominate a spatial region and there is more clustering in observation space (see Braverman, 1999).

7 Summary and Conclusions

This paper discusses the context and logic behind use of a lossy data compression algorithm to create compressed versions of large or massive geophysical data sets. The proposed method requires partitioning the data into subsets, and applying the algorithm separately to each one. A parameter specifying the trade-off between data reduction and induced error must be specified. Data reduction and error are measured by expectations over random quantizer functions used to summarize the data. Estimators for these expectations are motivated by a statistical model for the relationship between compressed (summarized) and raw data. A criterion for choosing the parameter is discussed.

In its present form this method would probably have difficulty with the *MISR* data stream. A major impediment is the binning step discussed in Section 3, which requires multiple passes through the data. A remedial modification would be to use samples obtained in the

design step to estimate mean squared error. λ could be selected using the estimate, as could the minimum mean squared error quantizer. Only two passes through a full subset would then be necessary: one to obtain a sample and the other to (finally) bin the data. Other computational savings are possible by reducing sample size and numbers of clusters allowed.

Another potential computational difficulty arises from high dimensionality. This exercise used bivariate data, but *MISR* and other massive geophysical data sets have many more than two variables. Data could be projected into principal component subspaces for purposes of the design step, then cluster representatives retransformed back to observation space before binning. These and other modifications are being investigated.

These modifications will come at some cost in terms of error. How much depends strongly on the data. In fact, success of the method in general depends on the data. When they are clustered, ECVQ can be expected to do well; it is a penalized clustering algorithm. When the data are not clustered, summaries may be less convincing because some subsets are just hard to describe. The figures of merit calibrate quality of ECVQ summaries, and users of compressed data must be mindful of quality issues. It is up to the user to decide when the ECVQ road map of Level 2 is adequate to their needs.

Acknowledgments

Many thanks to the *MISR* instrument team at NASA's Jet Propulsion Laboratory, especially Principal Investigator Dave Diner, and team members Ralph Kahn, Earl Hansen and Mike Smyth. The author would also like to thank Don Ylvisaker, Jim MacQueen, Noel Cressie, and Larry DiGirolamo for their helpful comments and willingness to discuss the issues presented here.

References

Braverman, Amy Joan (1999), *A Rate-distortion Approach to Massive Data Set Analysis*, Ph.D. dissertation, Department of Statistics, University of California, Los Angeles.

Chou, P.A., Lookabaugh, T., and Gray, R.M. (1989), "Entropy-constrained Vector Quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 31-42.

Cover, Thomas A. and Thomas, Joy T. (1991), *Elements of Information Theory*, Wiley, New York.

MacQueen, James B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-296.

Tarpey, Thaddeus and Flury, Bernard (1996), "Self-Consistency: A Fundamental Concept in Statistics," *Statistical Science*, **11**, 3, 229-243.