Compression Algorithm for Infrared Hyperspectral Sounder Data.

Irina Gladkova [1], **Leonid Roytman** [1], **Mitch Golberg** [2]

[1] City College of New York, 138th Street and Convent Avenue, New York, NY
[2] NOAA/NESDIS, Washington, DC
(questions can be sent to gladkova@cs.ccny.cuny.edu)

This research is sponsored by NOAA/NESDIS under Roger Heymann (OSD), Tim Schmit (ORA) HES sensor Compression Group leads.

The research is undertaken by NOAA/NESDIS, for its GOES-R Earth observation satellite series, to be launched in the 2013 time frame, to enable greater distribution of its science data both within the U.S. and internationally.

In this paper we will present a new lossless algorithm for compression of the signals from NOAA's environmental satellites. The project's aim is the design, analysis, and implementation of compression techniques that are suitable for the next-generation NOAA/NESDIS Geostationary Operational Environmental Satellite (GOES) instruments. We are using current spacecraft to simulate data from the upcoming GOES-R instrument and focusing on Aqua Spacecraft's AIRS instrument in our case study.

The AIRS is a high resolution instrument which measures infrared radiances at 2378 frequencies ranging from 3.74-15.4 μm. The AIRS takes 90 measurements as it scans 48.95 degrees perpendicular to the satellite's orbit every 2.667 seconds. We use Level 1A digital counts data granules, which represent 6 minutes (or 135 scans) of measurements. Therefore, our data set consists of a $90 \times 135 \times 2378$ cube of integers ranging from 12-14 bits.

It should noted that noise in the channels introduce added complexity in compression. Therefore, in practice, we utilize only 2108 out of 2378 channels picked by NOAA/NASA for their favorable characteristics.

Our compression algorithm consists of the following steps:

1. Channel Partitioning
2. Adaptive Clustering
3. Projection onto principal directions
4. Entropy coding of the residuals

During the first stage, we partition the data into three units. The partitioning takes into account that range of the digital counts varies (12, 13 and 14 bits) with respect to the channel index, and hence we will separately process each of these ranges. After this subdivision, each of the resulting $90 \times 135 \times K_n$ granules will be processed independently, and the compression ratio is calculated with respect to the effective range of each part.

The purpose of the second step of our algorithm is to transform the data so that its distribution is as close to normal as possible. It is obtained via adaptive clustering procedure that we are developing specifically for the considered sounder data. This clustering algorithm will be presented in the full paper. Here we present the motivation for such an approach and outline its role in the lossless compression procedure.

Step 3 of our algorithm is the Karhunen-Loève transform that is known to be optimal (in some sense) in the presence of a normal distribution of the data. This is a well established technique for dimensionality reduction and a set of theorems about optimal properties of this transform can be found in numerous texts on multivariate analysis.

After the third step, we have $90 \times 135 \times K_n$ granules of residuals that are approximately normally distributed but have a lower entropy (due to properties of the Karhunen-Loève transform). In our computations, the entropy on average decreased from 9.5 to 3.2 during step 3. Therefore, the lower bound [4] on the number of bits per residual entry is 3.2 bits. We build our Huffman codebook [3] based on a normal distribution with variance computed from the residuals, so that the actual codebook doesn't have to be transmitted.

To justify the need for clustering, we start with an observation about the nature of the correlation of data points between the consecutive satellite images that form a granule. We start with a trivial example, when the granule consists of two 90×135 images, i.e. $K_n = 2$. As we progress from one image to the next, as illustrated by Figure 1 which display the satellite images for channel indexes 653 and 654 respectively, we see that they, like many other images in the granule, are visually similar.

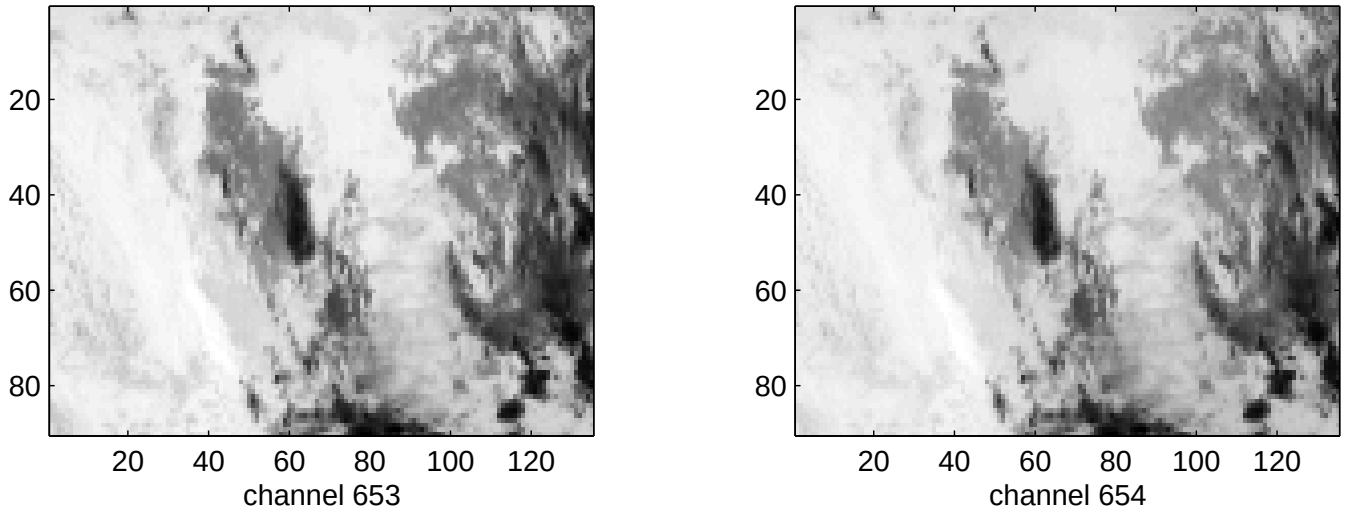

Figure 1: Example of two consecutive images from granule 126.

To illustrate this relationship, one can plot the gray level values of pairs of consecutive images as shown in Figure 2. Each dot represents a pixel in the image, with the horizontal coordinate being the gray value of the first image and the vertical coordinate being the gray level value of the second image. As they are highly correlated, the scatter plot in the Figure 2 shows points along the diagonal. In the best case, one would find that the points are normally distributed about a line segment. In this case, projection on the principle directions will yield the best compression ratio. Most of the relations between consecutive images are distributed tightly about a line segment. This explains the usage of Principal Component Analysis in data retrieval extensively used by NOAA scientists. Nevertheless, there are occasions (as illustrated below) when the distribution has more than one component.

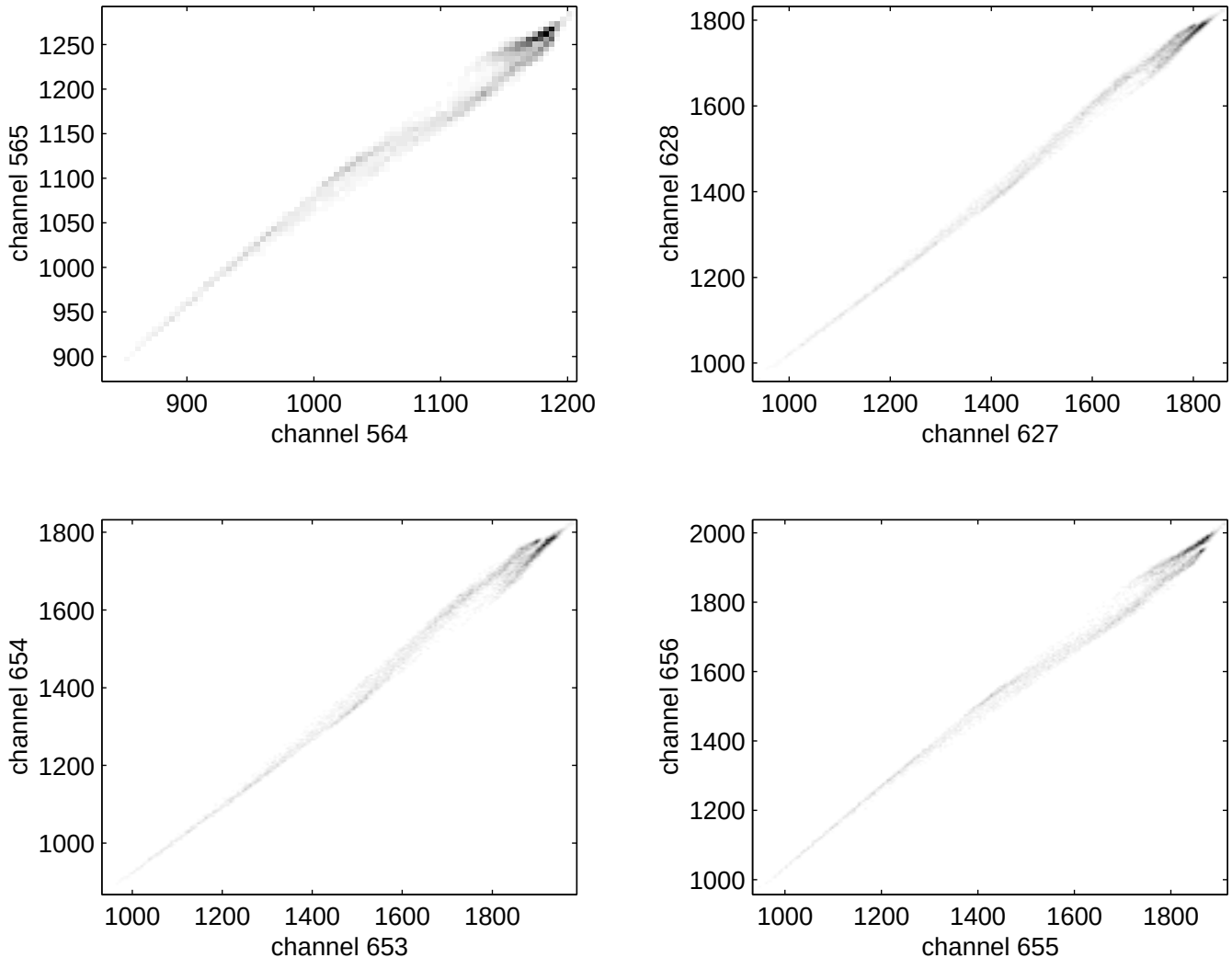

Figure 2: Examples of scatter plots of consecutive images.

As can be noted from the scatter plots in figures 2 through 4, the points of the actual data have more than one concentration (cluster), and different clusters have different mean and principal projection directions.

The scatter plot of n consecutive images would result in n-dimensional cloud of points which also has a distribution featuring multiple clusters. For visual clarity, our illustrations are 2 dimensional projections of the n-dimensional configuration. In this paper, we will introduce an algorithm that identifies clusters based on the distances to the principal direction. A detailed description of this new algorithm will be presented in the full paper.

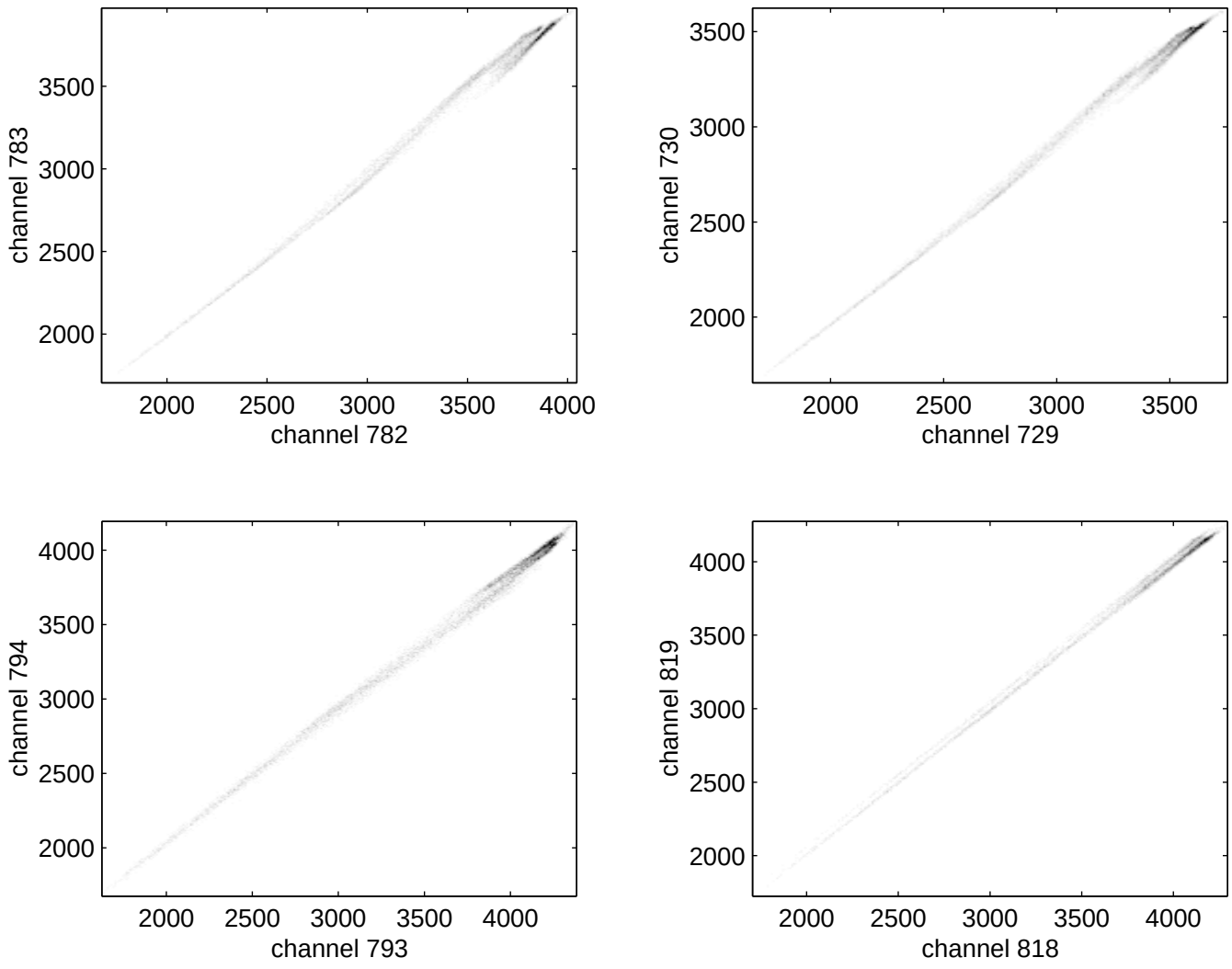

Figure 3: Examples of scatter plots of consecutive images.

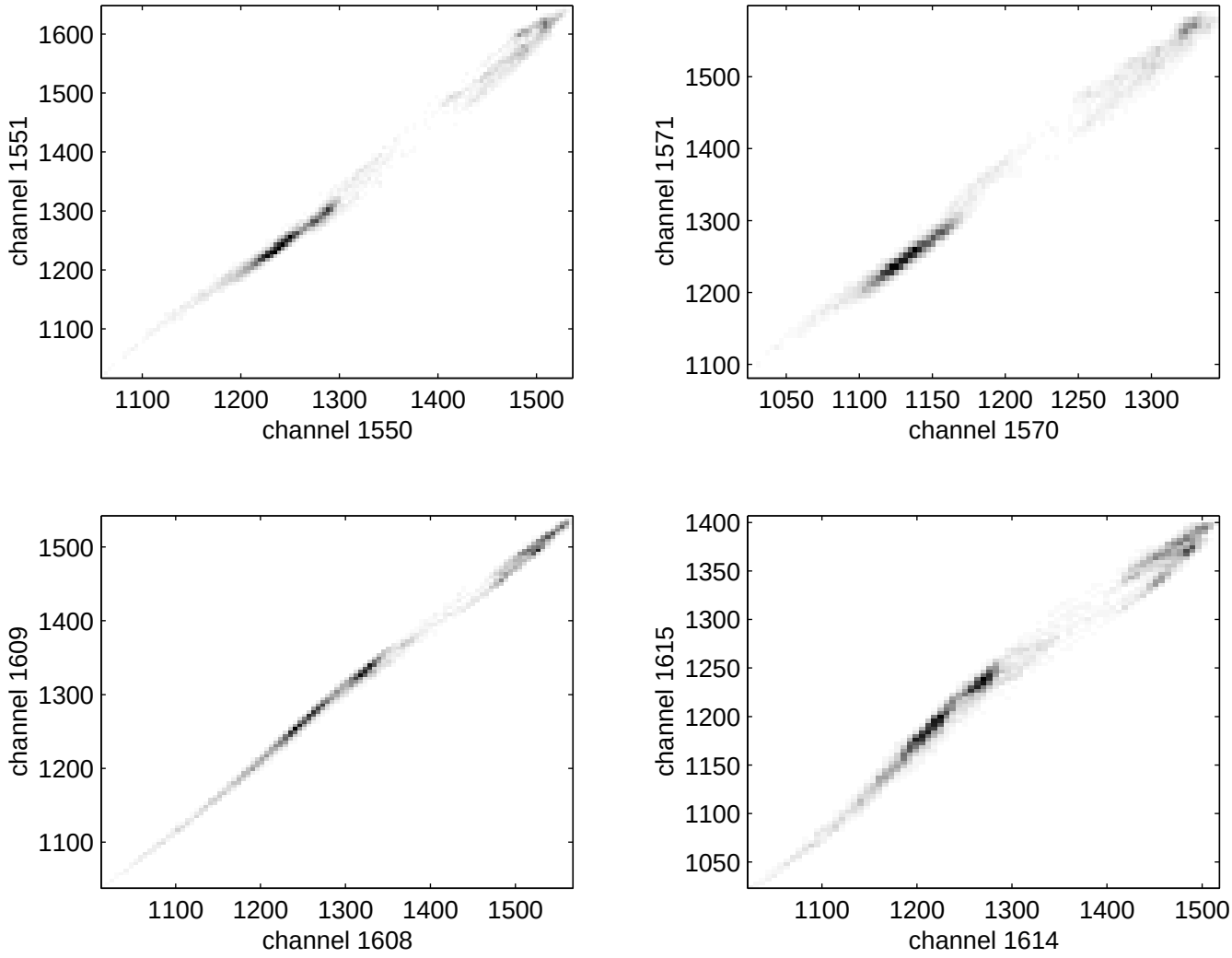

Figure 3: Examples of scatter plots of consecutive images.

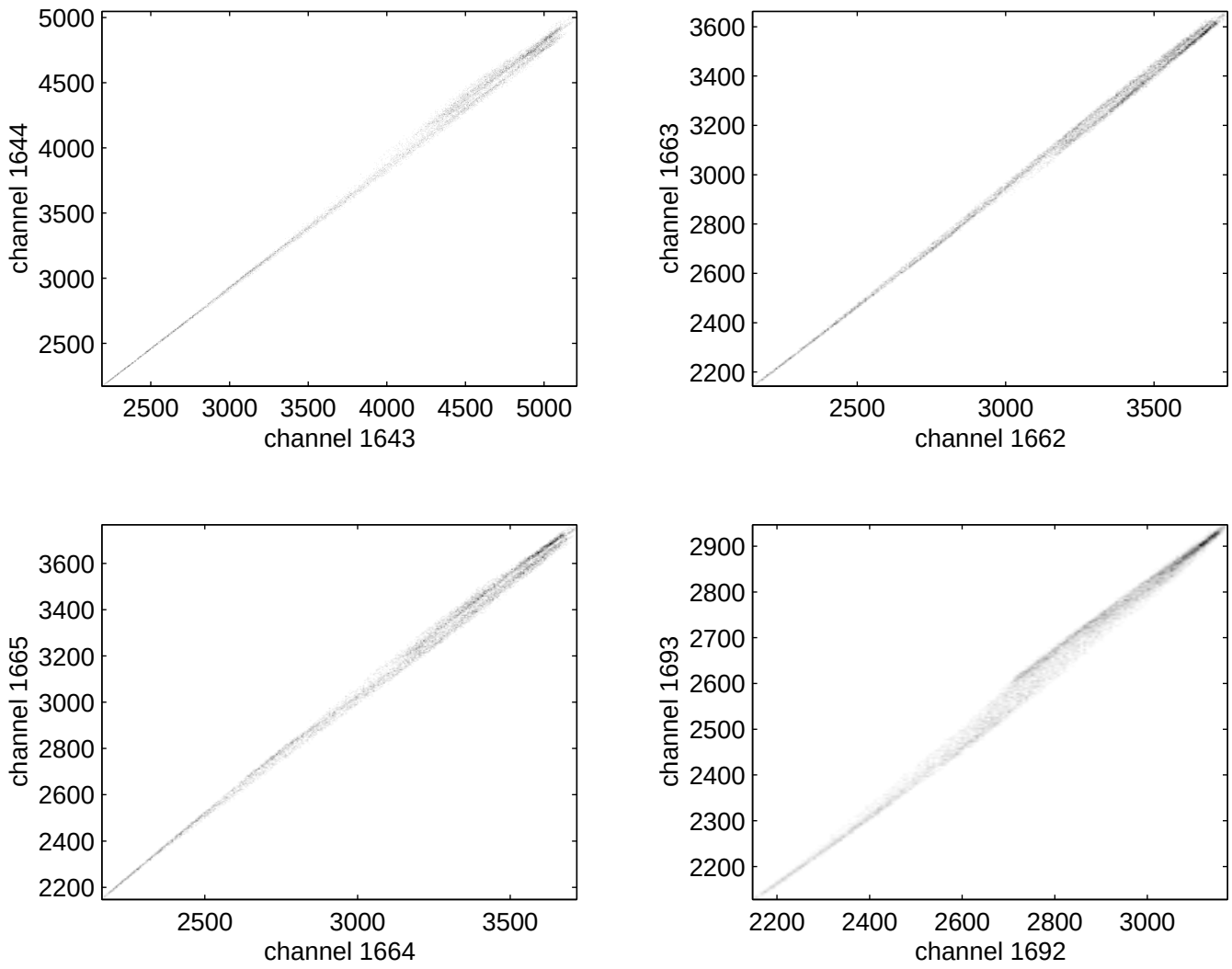

Figure 4: Examples of scatter plots of consecutive images.

As a result of clustering, the granule of residuals has a zero mean and is almost normally distributed. As was mentioned earlier, the purpose of the clustering stage is to transform the data so that its distribution is as close to normal as possible. The cost of the clustering is an increase in memory utilization (due primarily to record-keeping data structures needed for the decompression procedure). There are several theorems [2] concerning the optimality of the Karhunen-Loève transform provided that the data is normally distributed. Therefore, if we know that the additional memory utilization introduced by clustering is negligible then the overall transformation is nearly optimal.

We have observed that the ratio of memory needed for step 2, with respect to the total memory occupied by the original granule, varied in the 10 considered granules but on average was approximately 0.03 % of the memory used by the original granule. This negligible memory increase justifies the use of a known optimal approach.

In this abstract we give some results from output produced by our new algorithm, currently in an advanced state of development, for achieving a high degree of compression of sounder data. Initially, we have focused our algorithm on 1502 pristine channels of data, but we are currently extending the algorithm to obtain good compression of images with striping effects, in such a manner as to permit future data recovery from these channels.

In the following table, we give preliminary resulting ratios of lossless compression for our 10 test granules. One should note that these ratios are higher than those presented in our previous work [1], where we have not used adaptive clustering in the second stage of the compression algorithm.

Granule	Location	Ratio (treating all data as 14 bit)	Ratio (based on effective range)
9	Pacific Ocean, Daytime	3.6412	3.3314
16	Europe, Nighttime	3.6261	3.3291
60	Asia, Daytime	3.5185	3.2315
82	North America, Nighttime	3.7833	3.3651
120	Antarctica, Nighttime	3.5897	3.2831
126	Africa, Daytime	3.5097	3.2259
129	Arctic, Daytime	3.6345	3.3711
151	Australia, Nighttime	3.5001	3.2162
182	Asia, Nighttime	3.4892	3.1771
193	North America, Daytime	3.5106	3.2278

References

[1] I. Gladkova, L.Roytman, M. Goldberg, J. Weber, Compression of AIRS data using Empirical Mode Decomposition, *Proc. of SPIE*, vol. 5548, pp. 88-98, 2004

[2] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998

[3] D. Huffman, A method for the construction of minimum redundancy codes. *Proc. of IRE*, vol. 40, pp.1098-1101, 1952

[4] C.E. Shannon, Communications in the presence of noise, In *Proc. Of the IRE*, vol. 37, pp.10-21, Jan.1949