The Application of Principal Component Analysis (PCA) to AIRS Data Compression

Lihang Zhou[2], Mitchell D. Goldberg[1], Walter W. Wolf[2], Larry Mcmillin[1], Chris Barnet[1]
[1]NOAA/NESDIS/ORA, Camp Springs, MD 20746;
[2]QSS Group, Inc, Lanham, MD 20706

1. INTRODUCTION

Why Data Compression?

- Data compression is a topic of much importance to data archives, in regards to the access and distribution of the new generation of high spectral resolution infrared sounders.

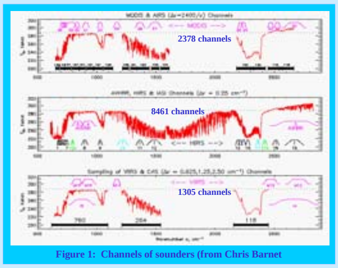


Figure 1: Channels of sounders (from Chris Barnet

- NOAA/NESDIS is processing and distributing AIRS data and products in near real-time. This offers us a great opportunity to use real AIRS observations for a test-bed for data compression studies of the future operational hyper spectral sounding instruments.

How should we do it?

- Instead of using some general-purpose compression technique, we want to develop our application in a way that can exploit the special characteristics of hyper spectral data, such as:

- ***Redundant information:***

Observations from hyper spectral instruments are not independent. Principal Components Analysis (PCA) is the most economical way to extract the independent information and reduce the dimension of the data.

Residue Range	Band 1	Band 2	Band 3
<0.1	30.30%	67.50%	52.07%
<0.4 &>=0.1	50.30%	31.70%	42.18%
<0.8 &>=0.4	15.89%	0.79%	5.30%
<1.6 &>=0.8	3.36%	0.01%	0.45%
<3.2 & >=1.6	0.1499%	0.0000%	0.0075%
> =3.2K	0.0001%	0.0000%	0.0000%

- ***Small reconstruction residues:***

The differences between the PCA reconstructed brightness temperature and the original observation are mostly very small numbers.

- ***Stable distribution:***

The range and distribution of the residues vary little from granule to granule…

In this poster we present our studies of the applications of PCA to AIRS data compression, and our experiments for achieving the near loss-less compression.

2. APPROACH

Implementation of PCA

- The method to be used to generate and apply the eigenvectors is described in detail in Goldberg et al.

- For data compression, we use the dependent eigenvector to compute the PCS, to achieve higher compression ratios.

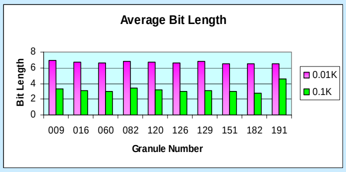
- Since timing is one of the biggest concerns when running operational algorithms, we divided the complete AIRS spectrum into three bands to speed up the computation, as show in table 2.

Table 2: Calculation Time for Three Spectral Bands:

Band #	Chan.#	Freq.	Cal. Time
1	688	650-920	2m56.87s
2	689	920-1400	2m59.19s
3	719	1400-2650	3m39.11s

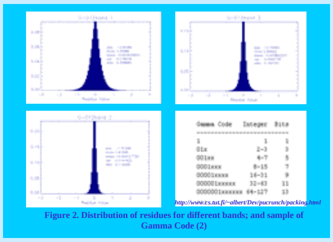
Encoding Residues

- We used Huffman coding to code the residuals. Below is a chart that shows the average bit lengths needed for keeping the brightness temperature accuracy to 0.01K, and to 0.1K, respectively.



- As one can see from table 1 and figure 2, the residuals are mostly very small numbers, and the distribution is stable.

- Elias Gamma Coding, which is like a static Huffman Coding, seems to be a more efficient way to encode the residues.



```
Gamma Code      Integer  Bits
==========================
1                    1     1
01x                2-3     3
001xx              4-7     5
0001xxx           8-15     7
00001xxxx        16-31     9
000001xxxxx      32-63    11
0000001xxxxxx   64-127    13
```

http://www.cs.tut.fi/~albert/Dev/pucrunch/packing.html

Figure 2. Distribution of residues for different bands; and sample of Gamma Code (2)

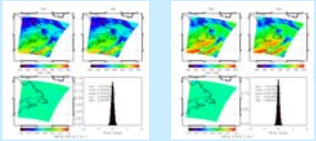
- Although Huffman code is appropriate, an even simpler and more efficient compression technique to use is Elias Gamma Coding, which does not need to store the tree for decoding.

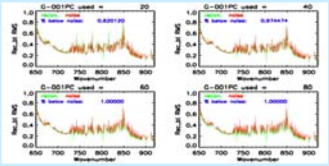
3. RESULTS

Lossy Compression

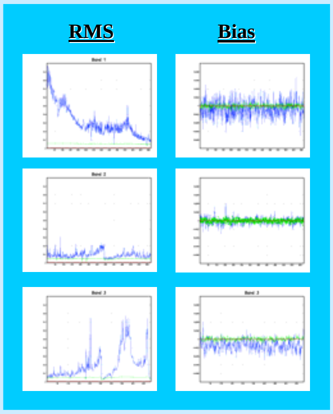
- Figures in this sector display the RMS error and other statistical information, when using different numbers of PCS to reconstruct brightness temperatures, and then comparing them with the original observations.

- Such information can be served as statistical metadata for archiving.





Near Lossless Compression



- Figures above show the BIAS and RMS of reconstructed brightness temperature (Blue curve), and reconstructed plus the residues (0.1K-Green curve; 0.01K-red curve), compared with the observed data.

- For 0.1K accuracy, the RMS are mostly less than 0.05K, the Bias are less than 0.002K. For 0.01K, both RMS and Bias appear to be close to zero.

File Type	File Size (MB)
L1B	121
Eigenvector	1/1/0.7
PC Scores (40pcs)	1.9/1.9/1.9
Encoded (0.1K)	4.0/2.0/2.3
Encoded (0.01)	9.7/5.9/5.6

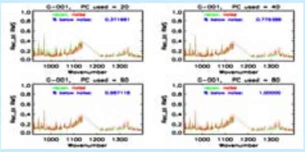
- Above table gives the file size for original and compressed files.

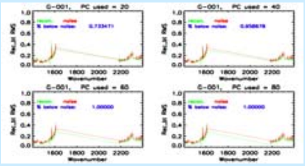
4. Summary

Where We Are

Compression Factors:

- *Lossy compression*: 40 PCS: ~11
- *Near lossless:* 40 PCS + residual:
  - Store brightness temperature to 0.01K, ~4
  - Store brightness temperature to 0.1K, ~8





Future Work

- To develop a data compression system using the approach presented in this poster to be used for archiving the compressed AIRS level 1B data and the corresponding metadata.

- To develop web capability to make the compressed 1B data accessible by the interested users.

- Continued optimizing of the compression and residue encoding technique.

REFERENCES:

- *Goldberg et. al., 2003: AIRS near-real-time products and algorithms in support of operational numerical weather prediction, IEEE Trans. Geosci. Remote Sensing, Vol. **41**, pp 379-389,*
- *Larry Mcmillin, AIRS Data Compression, AIRS STM, Spring, 2004.*
- *Huang, H-L and P. Antonelli, Application of principal component ananlysis to high-resolution infrared measurement compression and retrieval, J. Appl. Meteo., 40, 365-388, 2001.*

- *Comments/feedback: please email to: Lihang.zhou@noaa.gov*