Challenges in data compression for current and future imagers and hyperspectral sounders

Nigel Atkinson

Met Office, Fitzroy Road, Exeter, EX1 3PB, UK

Abstract

For current high-volume datasets such as IASI, near-real-time dissemination of the full-spectrum presents a challenge, but is nevertheless achievable within current bandwidth constraints, e.g. on EUMETCast. However, this may not be the case for data from future instruments.

In this paper we will look at the principles for efficient data compression. It is shown that appropriate normalisation of the spectrum (usually by the instrument noise) is a key to efficient compression – with implications for the design of level 1 data formats. EUMETSAT's compact VIIRS format is a good example of a format that has been designed with compression and dissemination in mind. Format design should be considered as a key part of the mission planning process – i.e. efficient dissemination formats should be available well before launch.

Principal Component (PC) compression is well known in the IASI community, and has the potential to separate signal and noise – though the technique has not so far been widely exploited in operations. Furthermore, PC compression is EUMETSAT's baseline for near-real-time dissemination of MTG-IRS data. Simulations of MTG-IRS are shown, which illustrate some possible tradeoffs between data volume and reconstruction accuracy.

1. Introduction

With each generation of satellite instruments, the data volumes have increased dramatically. To some extent, this has been compensated for by decreased transmission costs (e.g. use of DVB-S2 for satellite re-broadcasting of the data) and cheaper mass storage. Nevertheless, transmission and storage costs are significant. Table 1 gives some examples of current and future sensors that have high-volume datasets. Note that IASI (sounder) data volume is much less than VIIRS (imager) in absolute terms, but it is necessary also to consider the use of the data: for IASI, NWP centres require global coverage, whereas VIIRS is typically used only for limited regions. Global transmission of full VIIRS (i.e. all channels at full spatial resolution), via a mechanism such as EUMETCast, would be impossible, whereas it is achievable (just) for IASI on two Metop satellites.

For several of these datasets, dissemination of the full, uncompressed level 1 dataset is not cost-effective. Therefore a range of compression methods need to be studied. A significant challenge in the coming decade will be to accommodate MTG-IRS, the hyperspectral sounder on Meteosat Third Generation (MTG), scheduled to launch in 2021.

Table 1: Data volumes for some current and future sensors (level 1 datasets)

Sensor	Data volume GB/day	Comment
IASI	16 (BUFR)	Global coverage for NWP
CrIS	8 (BUFR)	Global coverage for NWP
VIIRS	800	User typically requires local area, from direct readout. Archiving at source (NOAA)
IASI-NG	>32 ?	Twice the spectral resolution of IASI, and lower noise
MTG-IRS	700 (uncompressed)	Full disk required, for NWP, also archiving
MTG-FCI	?	Imager on MTG
MetImage	?	On EPS-SG

In this paper we define:

- Lossless: Reconstruct the input exactly – to machine precision
- Near-lossless: Reconstruct the input, with a defined maximum error. For example, digitisation of a real signal y, with digitisation step δy, produces a maximum error $\delta y/2$ and a root mean square error $\delta y/\sqrt{(12)}$.
- Lossy: aim to preserve the "useful" information, while discarding unwanted "noise". Tradeoff will vary for different applications.

This paper will discuss principles of effective dataset design, using VIIRS as an example, showing how simple improvements to the dataset can substantially decrease data volume. Next, MTG-IRS will be discussed, showing a range of tradeoffs that could be considered for various types of near-lossless and lossy compression schemes.

Lossless compression is not discussed in detail in this paper, but is under investigation elsewhere. For example, Huang and Huang (2007) show high-performance compression algorithms applied to AIRS, GIFTS and AVIRIS. These algorithms achieve significant improvements over standard JPEG-LS or JPEG2000 algorithms, but less than a factor 2. Lossy or near-lossless algorithms will certainly be needed for dissemination of MTG-IRS.

2. Design of the VIIRS level 1 datasets

The VIIRS Sensor Data Record format (SDR) is in hdf5 format. The dataset appears to have been designed as an intermediate storage format that retains all the sensor information, but not as a format suitable for dissemination. It has a number of deficiencies from the data compression point of view:

1. For several channels, the radiances are stored as floating point numbers. Thus the precision of the storage is far in excess of the precision that would be justified by the sensor measurements

```
GROUP "VIIRS-M13-SDR_All" {
   DATASET "BrightnessTemperature" {
     DATATYPE  H5T_IEEE_F32LE
     DATASPACE  SIMPLE { ( 768, 3200 ) / ( H5S_UNLIMITED, H5S_UNLIMITED ) }
    }
```

2. Latitude and longitude are given as 32-bit floating point numbers *for every image spot*. Although convenient to the end user, this is wasteful because the scan pattern is highly predictable. Also, the user needs to download a large geolocation dataset even if only interested in one channel (relevant to data downloads from NOAA/CLASS).
3. Both radiance and brightness temperature are given (wasteful, as it is easy to convert from one to the other).

Taking a typical 10-minute overpass, and considering only M-band channels (740m spatial resolution), we get:

- 2.4 GB uncompressed
- 1.3 GB with gzip

For their EARS-VIIRS service (Soerensen et al., 2013), EUMETSAT have devised a new format that uses scaled integer radiances for all channels, and a tie-point scheme for geolocation – only one geolocation point is needed for every 16×16 group of samples (see Figure 1).

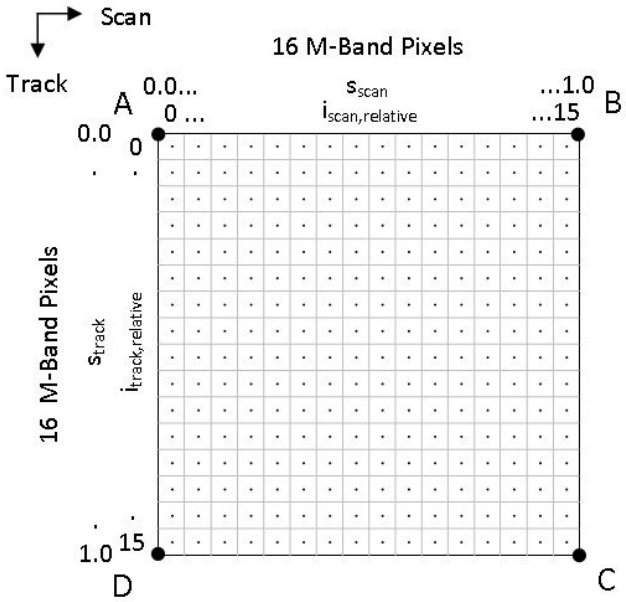

Figure 1: Tie point scheme for EARS-VIIRS

The M-band data volume can be reduced to

- 0.43 GB with gzip
- 0.38 GB with bzip2 – a 6-fold reduction compared with the original SDR

EUMETSAT have provided software (Java) that can quickly convert between the NOAA SDR format and the EUMETSAT compressed format.

This approach is a good illustration of the gains that can be made when datasets are designed with dissemination in mind – not as an afterthought.

3. Principal Component (PC) compression

In 2004, Lee and Bedford proposed a compression scheme for IASI that was based on PC scores and quantised residuals. The approach is to compute PC scores (relative to a set of eignevectors computed off-line from large training set), then the differences (residuals) between the original radiances and the radiances reconstructed from the PC scores.

1. Noise-normalise the radiance vector, $\mathbf{y}$:

$$\mathbf{y} = \frac{\mathbf{y}}{\mathbf{n}} - \mathbf{y_0} \tag{1}$$

   where $\mathbf{n}$ is the noise normalisation vector[1] and $\mathbf{y_0}$ is a mean radiance.

2. Compute the integer PC score vector, $\boldsymbol{s}$:

$$\mathbf{s} = \mathrm{NINT}\left(\frac{\mathbf{E}^{\mathrm{T}}\mathbf{y}}{f_s}\right) \tag{2}$$

   where $\mathbf{E}$ is the eigenvector matrix, f_s is a scaling factor (typically 0.5) and NINT signifies nearest integer.

3. Compute the integer residual:

$$\Delta\mathbf{y} = \mathrm{NINT}\left(\frac{\mathbf{y} - f_s\mathbf{E}\mathbf{s}}{f_r}\right) \tag{3}$$

   where f_r is a scaling factor for the residuals. f_r can be the same as f_s, but does not have to be.

4. Huffman encode the PC score and residual (or use a standard tool such as bzip2 or gzip)

Steps 1 and 2 of this scheme are routinely carried out in the EUMETSAT EPS ground segment (e.g. level 2 retrievals use reconstructed radiances) but to date there has been no requirement to disseminate residuals, since EUMETCast bandwidth has expanded significantly since Metop dissemination was first envisaged, such that full spectra can be disseminated.

There are two significant aspects to this scheme for compression. Firstly, the separation of signal (mainly in the PC score) and noise (mainly in the residual) gives an advantage. Secondly, the quantisation of the residual.

An experiment was carried out on an IASI level 1C dataset with various compression options, and repeated using the equivalent level 1B dataset. Results are summarised in Table 2. The table entry "AAPP quantised" refers to noise-normalised radiances that have been quantised with a scaling factor f = 0.5, then stored as 16-bit numbers.

We see that the quantisation has the largest effect on overall data volume; the separation of PC score and residual is secondary, but still significant. In the bottom row, the volume for level 1B is 50% higher than for level 1C; it is probable that the use of a diagonal normalisation matrix, $\mathbf{n}$, in the level 1C case is causing information to be lost.

[1] Lee was working with 1B (self-apodised) radiances, so a noise vector is appropriate. For 1C radiances (i.e. apodised) it is better to use the square root of the full noise covariance matrix, but it is common practice to use the diagonal as an approximation, as the computation is simpler.

Table 2: Compression experiment with IASI

Compression method	Volume relative to AAPP (L1C)	Volume relative to AAPP (L1B)
L1C BUFR	0.68	-
AAPP format + bz2	0.63	0.67
AAPP quantised + bz2	0.37	0.36
AAPP quantised + diff from mean + bz2	0.28	0.34
PC scores + quantised residuals (Lee)	0.20	0.30

4. MTG-IRS simulation

The characteristics of IASI and MTG-IRS are compared in Table 3. Information on MTG is taken from EUMETSAT (2007).

Table 3: Comparison of IASI and MTG-IRS

	IASI	**MTG-IRS**
Spectral sampling	0.25 cm^{-1}	0.25 cm^{-1}
Samples per spectrum	8461	1808 (2 bands)
Spatial sampling at nadir	25 km	4 km
Samples per hour	54000	8.0×10^6
Data volume for radiances, assuming 16-bit words	0.92 GB/h	28 GB/h

IASI spectra were used to simulate MTG-IRS. First, a transformation matrix, $\mathbf{M}$, was computed, based on the deapodisation function used for IASI (Lee, 2004). This matrix can be used to transform IASI spectra ($\mathbf{y}$) to MTG-IRS equivalents ($\mathbf{y}'$) and also to transform the 8461×8461 EUMETSAT level 1C covariance matrix, $\mathbf{C}$ (T. Hultberg, 2013, pers. comm.).

$$\mathbf{y}' = \mathbf{M}\,\mathbf{y}$$

$$\mathbf{C}' = \mathbf{M}\,\mathbf{C}\,\mathbf{M}^{\mathrm{T}}$$

PC eigenvectors were computed from the transformed covariance matrix, and were used to compute PC scores and residuals of the spectra. 300 PCs were used (150 in each band), on the grounds of (i) this being the EUMETSAT baseline, and (ii) a plot of eigenvalues suggested this was an appropriate cutoff.

We note from Section 3 that there are two parameters f_s and f_r that control the quantisation. We choose to set f_s to the constant value of 0.5 (since that controls the PC scores which are of most value to users), but to investigate the effect of varying f_r., i.e. to vary the precision with which we represent the residuals.

Bit volumes are shown in Figure 2, while the reconstruction errors with respect to the raw radiance are shown in Figure 3. Each plot shows three different compression options: (i) using the noise-normalised radiance and converting to integer, using different quantisation levels, (ii) PC score and quantised residual, and (iii) using the PC score

only, with no residual. Bzip2 compression was used as the final step (gzip was also tried, but was not as effective).

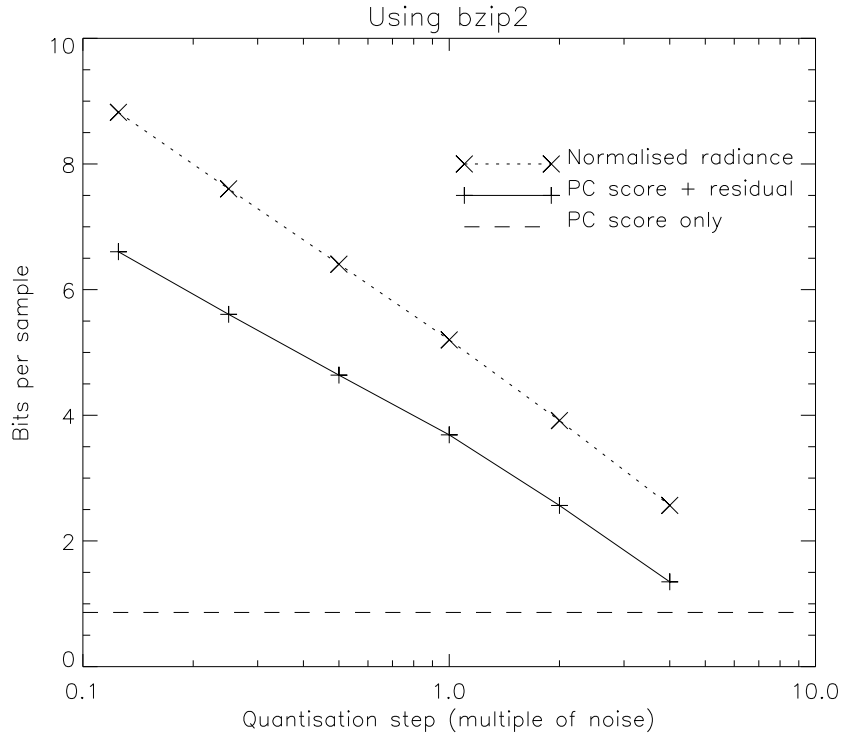

Figure 2: Bit rate for different compression options (all with bzip2). The uncompressed level is 16 bits per sample.

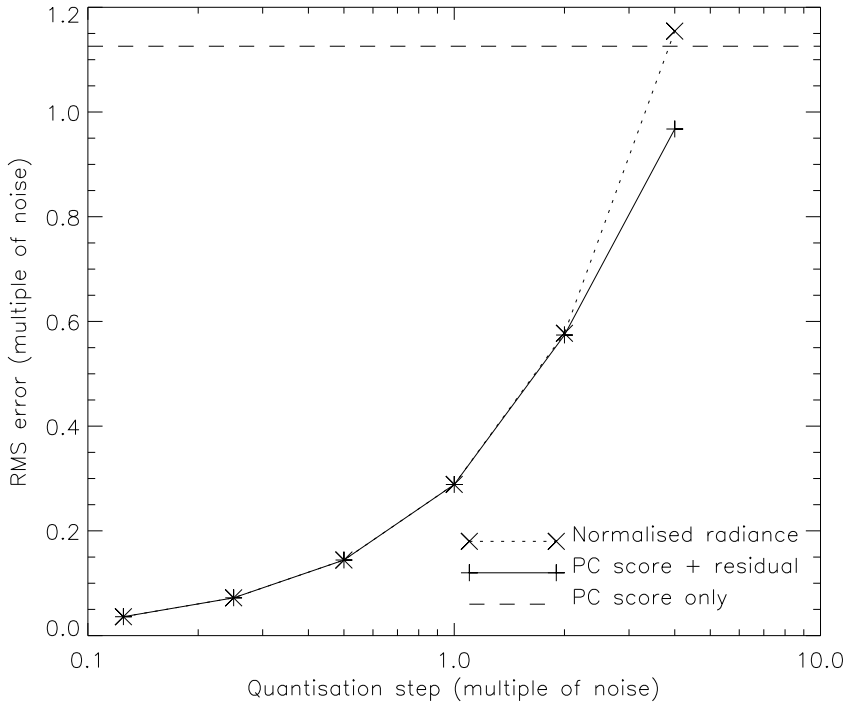

Figure 3: Reconstruction error for the different compression options

Note that the "reconstruction error" is defined as the difference between the reconstructed spectra and the original measured (noisy) spectra. It could be argued that

the PC-score-only reconstruction is closest to the "truth", as most of the instrument noise has been filtered out. The other compression schemes are closer to the measured radiance.

We see that PC score + residual does give better compression than simple quantisation of the radiances – but the latter has the advantage of simplicity, and may be a viable candidate for operational use, e.g. for archiving purposes.

5. Conclusions

For MTG-IRS, the baseline for near-real-time dissemination is PC scores. This should give a rate of about 1.6 GB/h, a factor 18 lower than the uncompressed volume (or about the same as two IASIs). This is in line with the expected EUMETCast capacity in the time-frame of MTG.

With PC scores and residuals there are possibilities for trade-off. Coarsely-quantised residuals might be useful for detecting outlier events. But the data volume (for MTG-IRS) would be a factor 3–6 higher than with PC scores alone. It might be useful to make residuals available off-line.

Datasets should be designed with dissemination and compression in mind, not as an afterthought. For example, use of suitably scaled integers; avoidance of unnecessary duplication; etc. EUMETSAT's EARS-VIIRS service is an illustration of good practice.

References

Atkinson, N.C., 2013, *Technical lossless / near lossless data compression*, presentation, ECMWF / EUMETSAT NWP-SAF workshop on efficient representation of hyperspectral infrared satellite observations, http://old.ecmwf.int/newsevents/meetings/workshops/2013/NWP-SAF_satellite_observations/presentations/index.html

EUMETSAT, 2007, *MTG Mission Requirements Document*, v2C, available at http://www.eumetsat.int (Future satellites, MTG, MTG Resources).

Huang, B. and Huang, H.-L., 2007, *Current Status of Lossless Compression of Ultraspectral Sounder and Hyperspectral Imager Data*, poster, Joint 2007 EUMETSAT Meteorological Satellite & 5th AMS Satellite Meteorology and Oceanography Conference, Amsterdam, http://www.ssec.wisc.edu/meetings/jointsatmet2007/

Lee, A.C.L. and S. Bedford, 2004, *Support Study on IASI Level 1c Data Compression Final Report*, EUMETSAT contract EUM/CO/03/1155/PS

Soerensen, A., Rojo, E., Heinemann, T., Burla, M. and Dieterle, S., 2013, *EARS-ATMS, EARS-CrIS and EARS-VIIRS: Three New Regional Services for Suomi NPP*, presentation, CSPP/IMAPP Users' Group Meeting, Univ. of Wisconsin-Madison, 21-23 May, http://www.ssec.wisc.edu/meetings/cspp/program.html