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

Nigel Atkinson (Met Office)
or ...

How to store your satellite data without needing too many of these...
Background

- This talk includes material that I presented at the ECMWF / EUMETSAT NWP-SAF workshop on efficient representation of hyperspectral infrared satellite observations in November 2013. Slides available on the ECMWF web site.
- and subsequent short study on options for MTG-IRS.
Contents

• Introduction

• Compression of VIIRS
  ➢ example of how the dataset design can be improved

• Compression for IASI
  ➢ to illustrate the various possibilities

• Simulation of MTG-IRS – based on IASI data

• Conclusions
Large level 1 datasets include

- **Current sensors:**
  - IASI – 16 GB/day (global, BUFR)
  - CrIS – 8 GB/day (global, BUFR)
  - VIIRS – 8 GB *per 12-min overpass* (SDR, day)
    - 5 GB at night
    - ~ 800 GB/day
  - not all disseminated, but stored in NOAA/CLASS

- **Future sensors:**
  - IASI-NG – twice the spectral resolution of IASI
  - METimage
  - MTG-FCI
  - MTG-IRS – 700 GB/day (uncompressed)
    - Dissemination required, for NWP
  - Also archiving at EUMETSAT
Types of data compression

- **Lossless**
  - Reconstruct the input exactly – to machine precision

- **Near-lossless**
  - Can reconstruct the input with a defined maximum error
  - Error typically a defined (small) fraction of instrument noise
  - Example: digitisation error (or quantisation error)
    - Max error = $\delta y / 2$
    - RMS error = $\delta y / \sqrt{12}$
  - Will discuss PC scores plus residuals

- **Lossy**
  - e.g. jpeg for images
  - e.g. PC scores for hyperspectral sounders
Standard compression tools
- useful for lossless and lossy compression

- File compression tools
  - `bzip2`, `gzip`, `xz`, `lzma`, `compress`, etc.

- Image compression tools
  - `jpeg`, `jpeg-ls`, `jpeg2000`, `szip`, etc.

- They work best with files containing *integers*, not *floats*

- **BUFR** – widely used in operational context
  - Look at all occurrences of an element within a message
  - Express as a minimum, an increment width and a set of increments (in a reduced number of bits) to be added to that minimum.

*BUFR effectively accounts for data with a limited range of values, but does not do any entropy (Huffman) encoding*

  - Sometimes worthwhile to apply file compression after BUFR encoding
A badly designed format?

VIIRS SDR (NOAA)

• Some channels represented by 32-bit floats

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

• Latitude and longitude given as 32-bit floats for every image spot
  ➢ so you need a large geolocation file even if only interested in 1 channel
  ➢ A limitation when downloading data from NOAA/CLASS

• Both radiance and BT/reflectance are given (unnecessary)

• Typical 10-minute pass:
  ➢ 2.4GB uncompressed – M-band channels only (740m spatial)
  ➢ 1.3GB with gzip

• Too large to put this on EUMETCast!
EUMETSAT’s compressed VIIRS format

- Tie points for geolocation
- 16-bit integers for all radiances
- Remove duplication (no BTs)
- 0.93 GB uncompressed
- 0.43 GB with gzip
- 0.38 GB with bzip2
  - 6-fold reduction compared with input file
- Can be accommodated on EUMETCast
- EUMETSAT have provided conversion software (java) – works well

© Crown copyright  Met Office
EARS-VIIRS service

- Regional VIIRS
- Lannion now
- Other stations in early 2014
1. Full spectra (only BUFR compression)
   - nominally lossless (but apodisation has introduced losses compared with original self-apodised spectra)
   - Operational on EUMETCast
   - could reduce volume via noise-normalisation + quantisation

2. Channel subset
   - Operational on EUMETCast and GTS (366 channels = 4%)

3. PC scores only
   - lossy (but most of the “loss” is noise)
   - Operational on EUMETCast (not many users?)
   - 4% of BUFR volume

4. PC scores plus quantised residuals
   - near-lossless
   - not used operationally (no need at present, because full spectra are available)
   - 44% of BUFR volume (see later slide)
PC scores and quantised residuals
(proposed by Tony Lee and Steve Bedford, 2004)

1. Noise-normalised radiance:
   \[ y = \frac{r}{n} - y_0 \]

2. PC Score (integer):
   \[ s = \text{NINT} \left( \frac{E^T y}{f_s} \right) \]
   Typically \( f_s = 0.5 \)

3. Residual (integer):
   \[ \Delta y = \text{NINT} \left( \frac{y - f_s E s}{f_r} \right) \]
   Typically \( f_r = 0.5 \)
   gives 1% noise increase

4. Huffman encode and disseminate

Eigenvectors, \( E \) are computed off-line
As an example, I used a direct-readout pass received at Exeter,

IASI_xxx_1C_M02_20130422084559Z_20130422085406Z_V_T_20130422085721Z
size: 166.695 MB, number of scans: 61 (= 488 sec), number of spectra: 7320

First, look at the standard formats. Reference is AAPP l1c format which uses 16 bit integers for the spectra. Size = 135.4 MB (spectra comprise 91.5% of this)

<table>
<thead>
<tr>
<th>Format</th>
<th>Volume w.r.t. AAPP l1c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native PFS</td>
<td>1.23</td>
</tr>
<tr>
<td>PFS + gzip</td>
<td>0.91 (11 sec)</td>
</tr>
<tr>
<td>PFS + xz</td>
<td>0.73 (85 sec)</td>
</tr>
<tr>
<td>PFS + bz2</td>
<td>0.72 (25 sec)</td>
</tr>
<tr>
<td>BUFR</td>
<td>0.68</td>
</tr>
<tr>
<td>AAPP l1c + bz2</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Larger than AAPP due to IIS etc.
### Experiments with IASI data (2)

Comparison of l1b (self apodised) and l1c (gaussian apodised)

<table>
<thead>
<tr>
<th>Format (all with bz2, except BUFR)</th>
<th>Volume w.r.t. AAPP (1b)</th>
<th>Volume w.r.t. AAPP (1c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUFR</td>
<td></td>
<td>0.68</td>
</tr>
<tr>
<td>AAPP l1b/l1c</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>Quantised at 0.5xNEΔT</td>
<td>0.36</td>
<td>0.37</td>
</tr>
<tr>
<td>Quantised at 0.5xNEΔT + diff from mean</td>
<td>0.35</td>
<td>0.28</td>
</tr>
<tr>
<td>PCs + 16-bit residuals</td>
<td>0.55</td>
<td>0.47</td>
</tr>
<tr>
<td>PCs + residuals quantised at 0.5xNEΔT</td>
<td>0.30</td>
<td>0.20</td>
</tr>
<tr>
<td>PCs + residuals quantised at 0.125xNEΔT</td>
<td></td>
<td>0.31</td>
</tr>
<tr>
<td>300 PC scores only</td>
<td>0.018</td>
<td>0.029</td>
</tr>
</tbody>
</table>

1b contains more information than 1c?

A large part of the benefit comes from quantisation.
Implications for IASI-NG

- IASI-NG will have doubled spectral resolution compared with IASI, and lower noise
- Design of 1c format needs care, to avoid loss of information (e.g. apodisation should be reversible)
- Likely that compressed data volume will be more than double that of IASI
A look at MTG-IRS

<table>
<thead>
<tr>
<th></th>
<th>IASI</th>
<th>MTG-IRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral sampling</td>
<td>0.25 cm(^{-1})</td>
<td>0.625 cm(^{-1})</td>
</tr>
<tr>
<td>Samples per spectrum</td>
<td>8461</td>
<td>1808 (2 bands)</td>
</tr>
<tr>
<td>Spatial sampling at nadir</td>
<td>25 km</td>
<td>4 km</td>
</tr>
<tr>
<td>Spectra per hour</td>
<td>54000</td>
<td>8.0 (\times) 10^6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(full disk every hour)</td>
</tr>
<tr>
<td>Samples per hour</td>
<td>4.6 (\times) 10^8</td>
<td>1.4 (\times) 10^{10}</td>
</tr>
<tr>
<td>Data volume for radiances assuming 16-bit words</td>
<td>0.92 GB/h</td>
<td>28 GB/h</td>
</tr>
<tr>
<td>Operational BUFR (2 IASI instruments)</td>
<td>1.3 GB/h</td>
<td></td>
</tr>
</tbody>
</table>

Factor 30 higher than IASI

- Lossy compression is essential for NRT dissemination of MTG-IRS (baseline is PC scores)
- Some users still want access to the full spectrum - don’t trust the PCs!
- Requirement to archive the full spectrum centrally
Simulation of MTG-IRS

- Used EUMETSAT’s IASI covariance matrix
- Transformed to IRS spectral resolution, unapodised
- Computed eigenvalues and eigenvectors for each band
- Applied them to transformed IASI spectra
- Looked at near-lossless compression options
IRS data volumes

- Tradeoff between accuracy and data volume
Reconstruction error

PC score: error appears large, but this is mostly elimination of instrument noise.

Error increase: 1%
Conclusions

• For MTG-IRS
  - Baseline dissemination is PC scores only
    - 1.6GB/h (factor 18 lower than uncompressed)
    - just over two IASIs
  - Possibility of tradeoff with PC scores + residuals
    - 3 to 6-fold increase compared with PC scores only
    - quantised residuals could be made available off-line

• Design your datasets with compression in mind from the start – not as an afterthought

• Some key decisions are being made now for next-generation instruments
Thank you for listening!

Questions?