ESA EO Data Management & Operations Framework (EOF) Workshop 2024

Copernicus Data Compression

Serge RIAZANOFF – Grégory MAZABRAUD
Kévin GROSS – Alexis MARTIN--COMTE

11/04/2024

Rationale and scope of the Copernicus Data Compression Activity

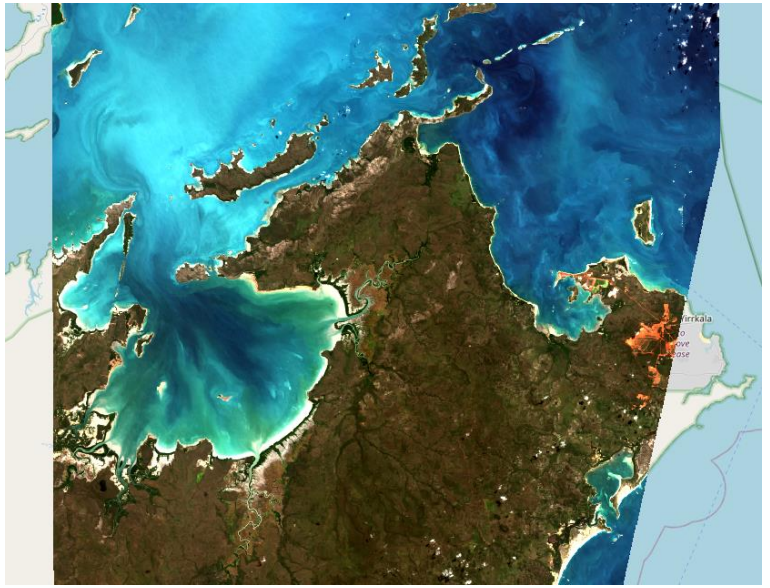
Challenges of Data Exponential growth – Due to high costs and hardware limitation, in particular the transfer rate from storage devices, it gets more and more complicated to manage the exponential increase of Earth Observation & Modelling data volume (Several tens of petabytes for Copernicus data).

Presentation – Studied datasets – Sentinel-2

A huge amount of data!

Sentinel-2 in few numbers

- 2 satellites, Sentinel-2A in 2015, Sentinel-2B in 2017,
- 10 days cycle per satellite,
- 13 spectral bands: 4 bands at 10 m, 6 bands at 20 m, 3 bands at 60 m.
- swath of 290 km.



Sentinel-2 product features

- UTM projection on MGRS grid (tiles of 100x100 km²),
- One JPEG2000 lossless file per band per product,
- 16 bits unsigned integer digital number.
- **~600 MiB** per product (~**630 MB**),
- **~110 000** products per cycle (S2A + S2B),
- **~40** millions of products at the end of 2022,
- **~20 PiB** (**~22 PB**) of S2 data at the end of 2022.

Lossy compression

Studied aspects

Trade-off between compression factor and access time – Partitioning, input data is cut and packed into blocks. The size of the blocks varies depending on the constraints of the following step or constraints during decoding. Big block size improve compression factor, but small block size improve subset decoding and multithreading decoding,

Variety of studied compressions – Many transformations such as the Hadamard transform, the **Discrete Cosine Transform** (**DCT**) or the **Discrete Wavelet Transform** (**DWT**). The DCT is popular and is used in several image compression algorithms such as JPEG, WebP and **JPEG-XL**, or video compression algorithms, such as MPEG-2 (DVD) and MPEG-4 (Blu-ray and streaming).
There are also new approaches based on machine learning, **HiFiC** for example,

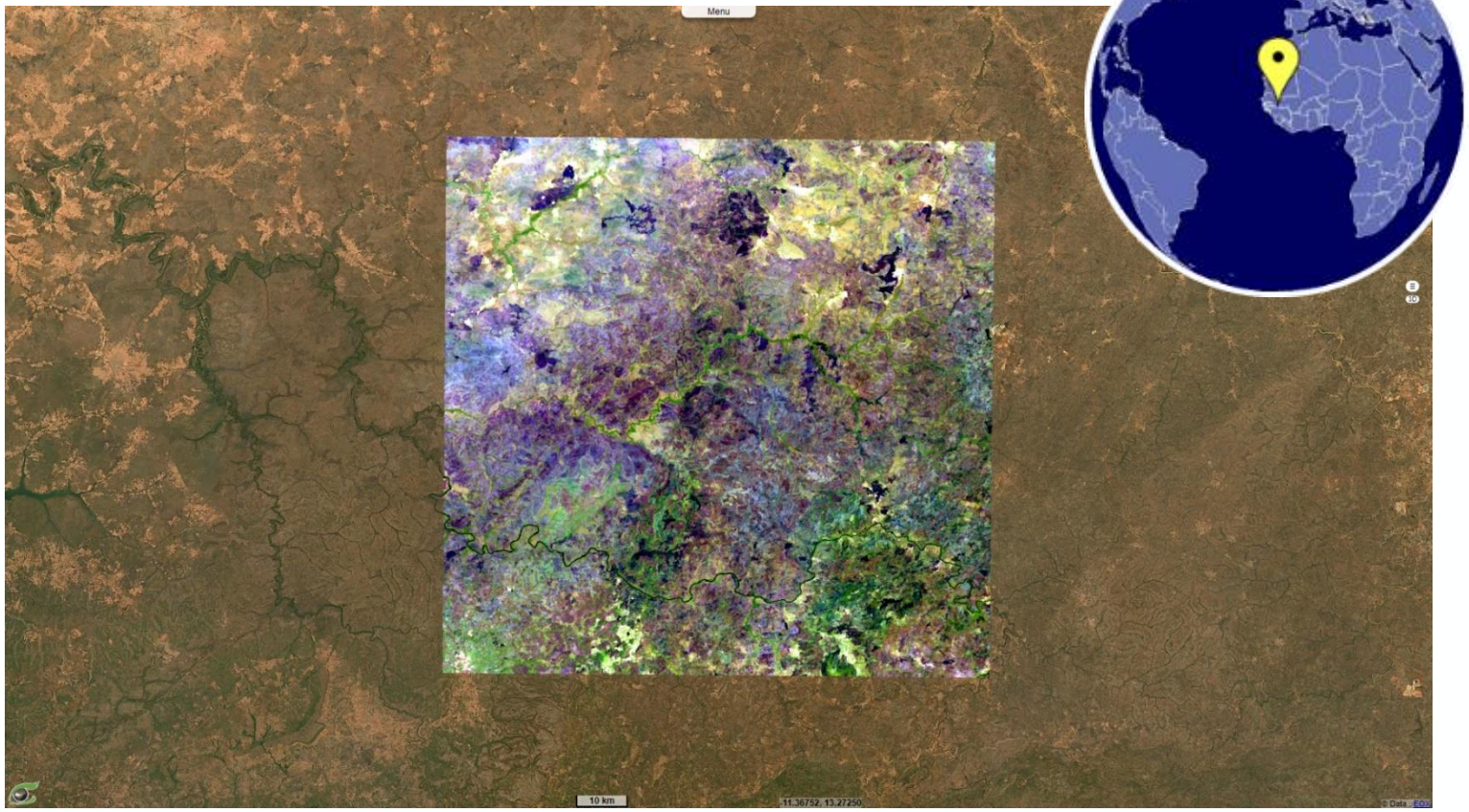
Visualisation of the relationship between temporal domain and frequential domain of a function, based on Fourier transform which is close to the DCT. (Wikipedia)

Quantization – Lossy operation, such as a reduction of the accuracy of the transformed data,

Post compression processing – One or more lossless compression steps, such as data reorganization, run-length encoding and entropy encoding.

Input data – Time series

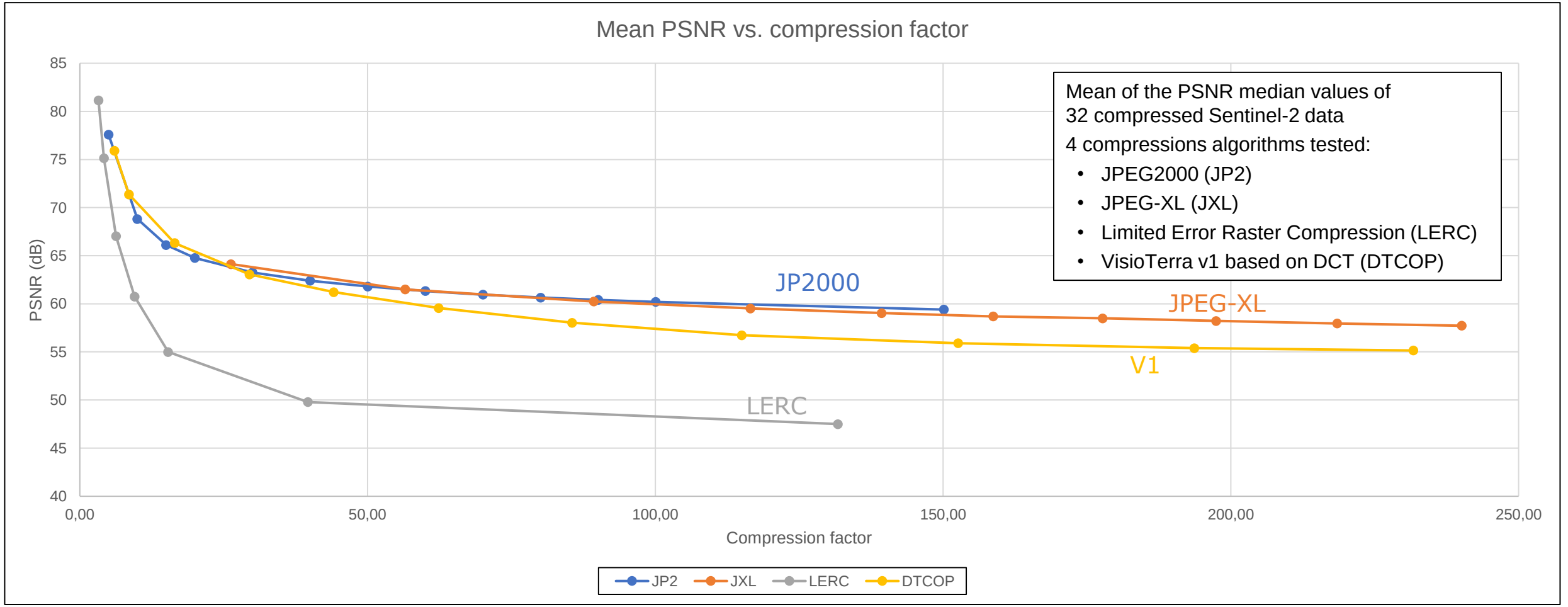
- Sentinel-2 L1C / tile **T28PGV**
- **32 dates** between 2016-01 / 2018-03
- Practically no clouds
- Tested with:
  - **JPEG200**
  - **JPEG-XL**
  - **LERC**
  - VisioTerra v1 based on DCT (**DTCOP**)



- Hyperlook : https://visioterra.org/VtWeb/hyperlook/ba9034465fab49f9937ecd44cd61924a

Generic quality estimators – PSNR results

V1 results close to standard high compression



Thematic quality estimators

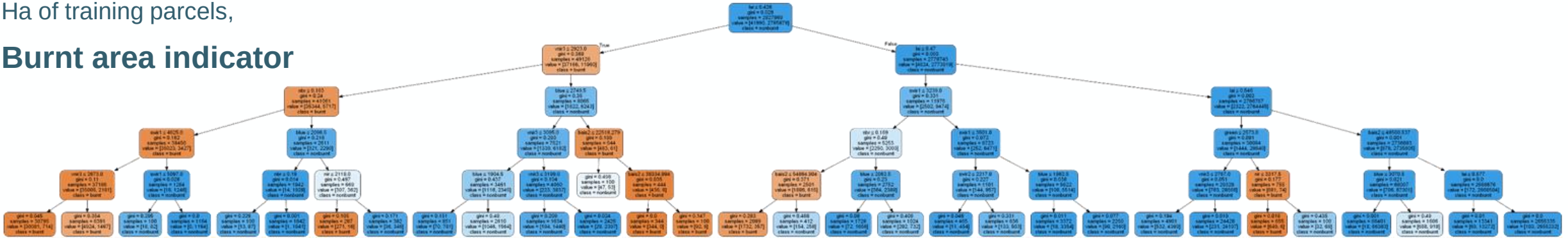
Thematical point of view (useful for end users)

Fire indicator

$$BAIS2 = \left(1 - \sqrt{\frac{B06 * B07 * B8A}{B4}}\right) * \left(\frac{B12 - B8A}{\sqrt{B12 + B8A}} + 1\right)$$

Fire indicator is based on the BAIS2 (Burned Area Index for Sentinel-2) and using a threshold defined through classification on more than 28.000 Ha of training parcels,

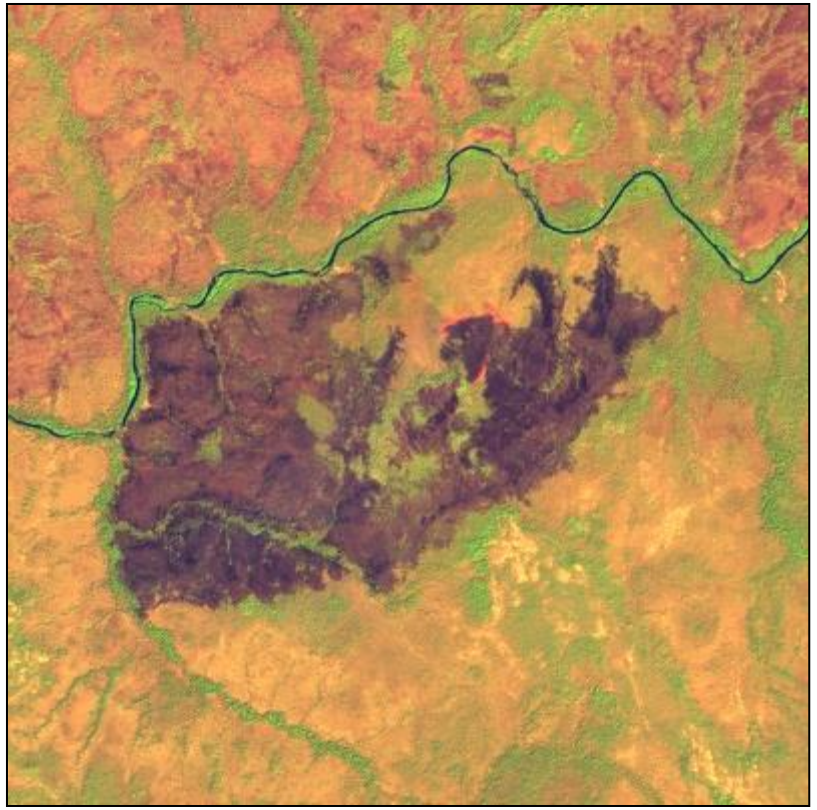
Burnt area indicator



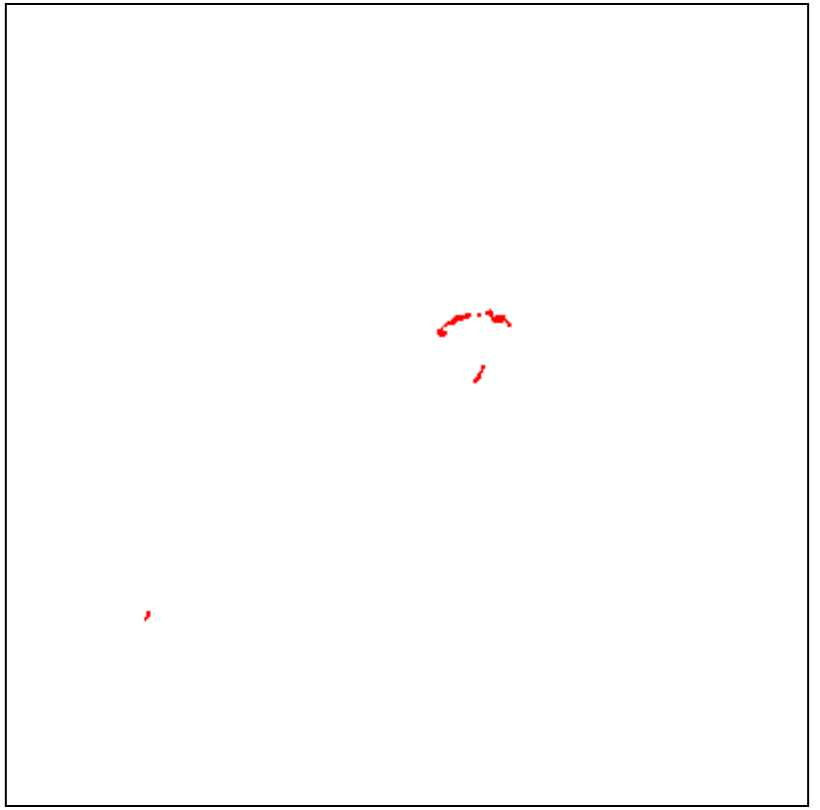
This indicator is based on decision tree produced by scikit-learn using the same training set as for the fire indicator with 12 features (B02, B03, B04, B05, B06, B07, B08, B11, B12, NBR, LAI et BAIS2).

	Fire	Burnt area
Overall accuracy	99,9965 %	99,7618 %
Kappa coefficient	0,9752	0,9161

Thematic quality estimators – Fire and burnt area



11,8,2 Agriculture

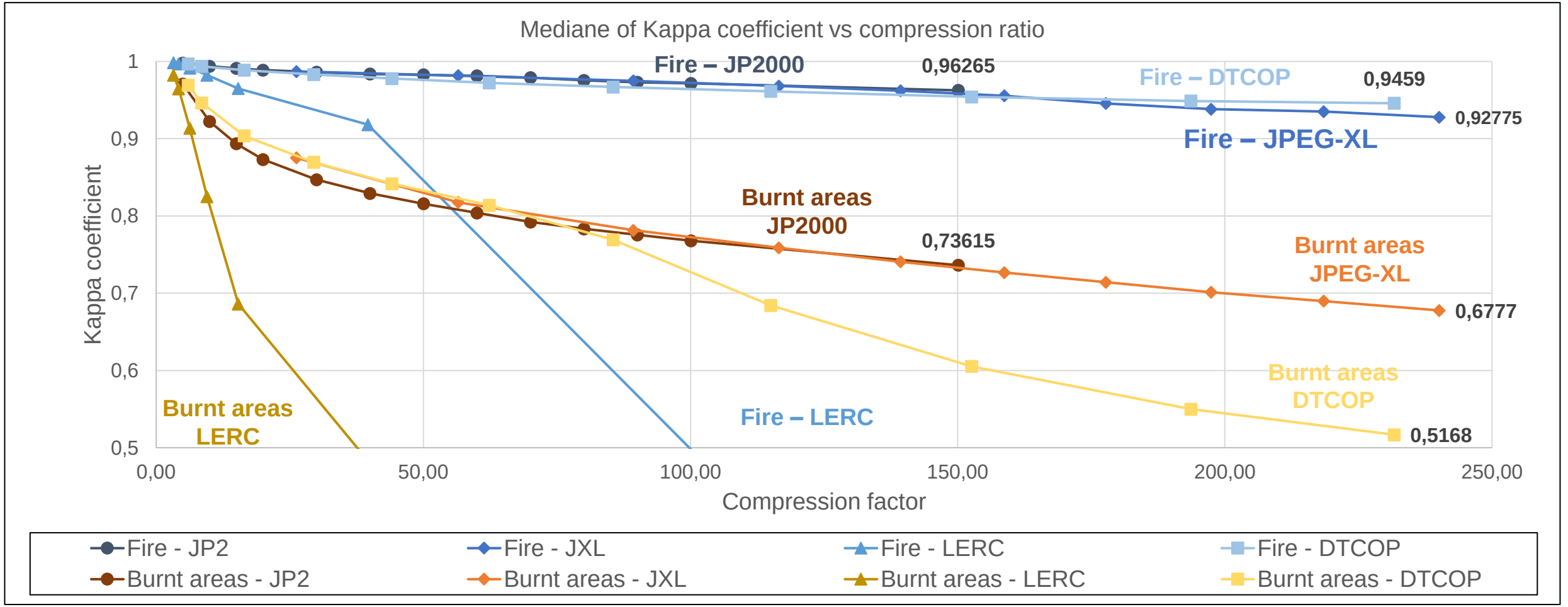


Fire indicator

Burnt areas indicator

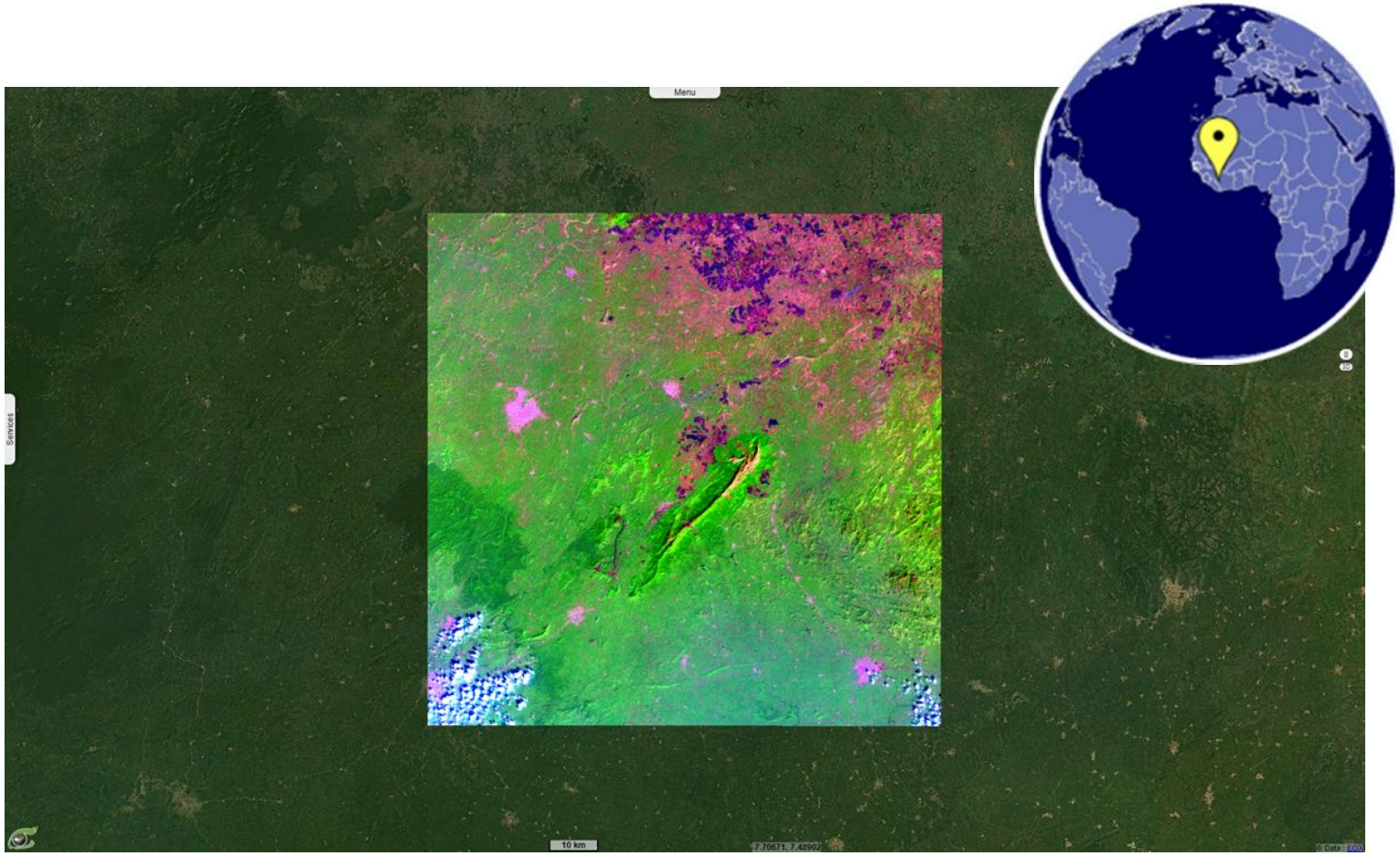
Thematic quality estimators – Fire and burnt area results

V1 results close to standard high compression



Input data – Classification

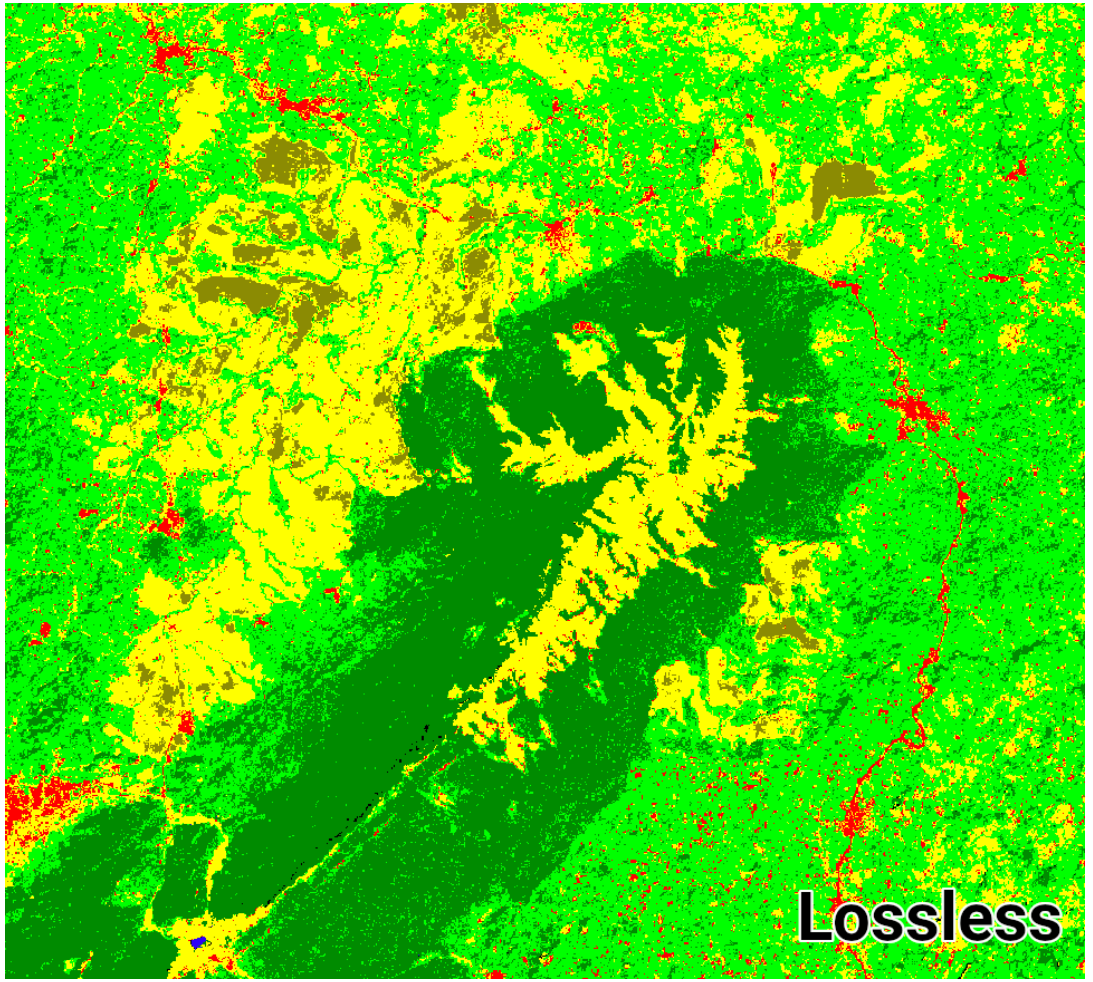
- Sentinel-2 L1C / tile **T29NNJ**
- Acquisition date: **2020/01/11 10:54:21**
- Practically no clouds
- Tested with:
  - **JPEG2000**
  - **JPEG-XL**
  - **LERC**



- Hyperlook : https://visioterra.org/VtWeb/hyperlook/652dff0c787141978afa562e18e4ffd4

Thematic quality estimators – Classification results

Lossy compression "cleans" classification results

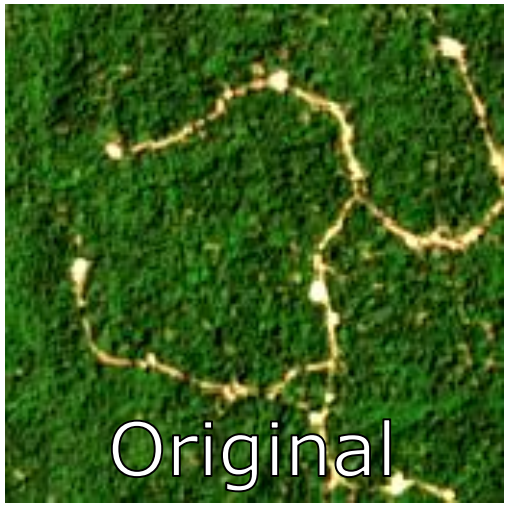


Kappa coefficient		
Compression vs.	**Raw**	**Control**
Lossless	1	0,9619
R50	0,8066	0,9713
R100	0,761	0,9753
R150	0,7282	0,9673

Conclusions

We've seen

- Thematic indicators are resilient even with high compression factors
- These thematic indicators are quantifiable
- Generic statistical indicators also provide with a quantifiable error
- Compression type and parameters depend on user needs



Other works

- Other compression algorithm tested like **HiFiC** (AI-based).
  That preserve textures in place of digital numbers.
- Test other indicator like **LPIPS** (Learned Perceptual Image Patch Similarity)
- Test on other data like Sentinel-1 radar data.
  Compression smooths uniform areas, decreasing the speckle.

Conclusions

Next steps

- Continue research on other lossy compression algorithms (AI-based),
- Optimize **DCT / iDCT** implementation (expected speed gain in range [x5; x40]),
- Improve quantization matrix,
- Study **variable error threshold** depending on the data physical properties,
- Study other entropy encoder than Huffman coding (arithmetic coding, asymmetric numeral systems...),
- Study “data partitioning” and variable DCT sizes (4, 8, 16),
- Add multiscale access,
- Study new use cases to determine a trade-off compression parameters.

Lossy compression may offer satisfactory trade-offs!