The impact of JPEG2000 lossy compression on the scientific quality of radio astronomy imagery

Sean M. Peters, Vyacheslav V. Kitaeff
International Centre for Radio Astronomy Research, The University of Western Australia, M468, 35 Stirling Hwy, Crawley 6009, WA, Australia

Abstract
The sheer volume of data anticipated to be captured by future radio telescopes, such as, The Square Kilometer Array (SKA) and its precursors present new data challenges, including the cost and technical feasibility of data transport and storage. Image and data compression are going to be important techniques to reduce the data size. We provide a quantitative analysis of the effects of JPEG2000’s lossy wavelet image compression algorithm on the quality of the radio astronomy imagery data. This analysis is completed by evaluating the completeness, soundness and source parameterisation of the Duchamp source finder using compressed data. Here we found the JPEG2000 image compression has the potential to denoise image cubes, however this effect is only significant at high compression rates where the accuracy of source parameterisation is decreased.

Keywords: data format, File formats, JPEG2000, Standards, 2000 MSC: 85-04

1. Introduction

The upcoming Australian Square Kilometer Array Pathfinder (ASKAP) telescope [DeBoer et al. 2009] is anticipated to capture spectral-imaging data-cubes (SIDCs) several orders of magnitude larger than any telescope ever before. The Square Kilometre Array (SKA) [Huynh and Lazio, 2013] SIDCs will be even larger, in the order of tens of terabytes per hour of observations. The networking, storing, processing and interrogating of such large datasets poses new technical, financial, and management challenges.

Such large SIDCs cannot be processed or stored on local user computers, even taking into account the projections for advances in HDD/SSD and network technologies, the cost of data storage will be significant. Any means to reduce the data volumes through compression will likely have significant benefits, especially in the reduction of the cost of the system.

Kitaeff et al. [2012] proposed the use of the JPEG2000 image standard as a method of storing SIDCs. JPEG2000 was first proposed in 1996 [Boliek, 1996] with remote sensing and medical imaging being the most demanding imaging fields in mind. The JPEG2000 group, established in 1999, has developed a standard, contain 12 parts, defining the mechanisms not only for compression, but also for servicing and interrogating a broad range of imagery data.

The most distinctive difference between JPEG2000 and the original JPEG image standard is the use of the Discrete Wavelet Transform (DWT). The DWT is an algorithm to convert a signal to the time-frequency domain. Data in this domain lends itself particularly well to quantisation and compression [Taubman and Marcellin, 2002]. As a result the JPEG2000 image standard can provide superior compression rates, with much less loss in visual quality, than any other standardised image format available. Additionally, JPEG2000 includes a variety of features especially useful for extremely large images such as: progressive transmission, the ability to decode any part of the image without having to decode the entire image, adaptive encoding with different fidelities through a precinct mechanism, multiple resolutions from a single master file. These features can significantly improve and enrich the interrogation of radio astronomy imagery data, as well as, reduce the overall volume of the data.

While JPEG2000 may retain visual content effectively at high compression rates, the effect of the lossy compression algorithm on the scientific quality of spectral image cubes captured in radio astronomy needs to be understood. One important way to study the impact of lossy compression on the quality of data is to see how well source finding algorithms identify the sources in compressed SIDCs.

While the volumes of data are large, the information density of the data is rather low. For example, for studies using HI emission of extragalactic objects, the majority of sources of such an emission will appear in data occupying only a few pixels. Such sources will be rather sparsely populating the volume of data-cube. The rest of imaging data is considered as noise. To distil the information from the volume of data it is therefore necessary to identify and parametrise the sources.
The development of fully automated source identification algorithms has recently become an intensive field of research (Whiting, 2012; Whiting and Humphreys, 2012; Floer and Winkel, 2012; Jurek, 2012; Boyce, 2003; Serra et al., 2012). These algorithms would normally construct a catalogue of sources from an astronomical image. Each entry in the constructed catalogue represents a source identified in the image, and its determined parameters. The unknown systematic errors introduced into the catalogue at any stage are of course very problematic. Thus, the source finding algorithms themselves need to be investigated, as well as, any additional data processing such as compression. The completeness and soundness are the main measures of how successful the algorithm is, in finding the sources (Popping et al., 2012). Such a set of metrics could be usefully utilised to study the impact of lossy compression on the quality of data, providing the guidance to the developers of the SKA as to whether JPEG2000’s lossy compression may be safely used.

The rest of the paper is laid out as follows. Section 2 details the dataset used and methodology taken in conducting our experiment. Section 3 provides the results and detailed discussion on the experiments performed in this paper. Finally, we draw some conclusions in Section 4.

2. Methodology

2.1. Synthetic dataset

In order to test the impact of compression on the scientific quality of data, we have chosen to use a simulated SIDC in preference to real radio astronomy imagery. By using a simulated set of galaxies to create our imagery the true catalogue of sources within the image cube is known with a higher certainty. This helps with measuring, not only the accuracy of source identification after image compression, but source parameterisation as well. The Deep Investigations of Neutral Gas Origins (DINGO) (Meyer, 2009) synthetic SIDC was used in all our tests.

DINGO is one of the planned surveys to be performed by the ASKAP telescope. The synthetic SIDC was generated using an analytical model of galaxies and their distribution. This helps with measuring, not only the accuracy of source identification after image compression, but source parameterisation as well. The Deep Investigations of Neutral Gas Origins (DINGO) (Meyer, 2009) synthetic SIDC was used in all our tests.

DINGO is one of the planned surveys to be performed by the ASKAP telescope. The synthetic SIDC was generated using an analytical galaxy simulation as described by Duffy et al. (2012), as part of survey planning. Significant effort was placed in creating a plausible DINGO survey simulation, as well as correctly inserting the sources while simulating instrumental noise and errors. A brief summary of the steps is outlined below:

1. An analytical model for each of the different types of galaxies was produced. This was followed by a model of the distribution of these galaxies throughout the DINGO survey space.

2. A cosmological simulation was produced using the analytical models of galaxies and their distribution. This provided position, flux, HI content etc. for almost 4 million galaxies, the vast majority of which would be unidentifiable. These sources formed a true catalogue in our experiments. Many of the simulated sources, however, are too faint to be observable with ASKAP. For the experiments performed in this paper a filter was applied on the true source catalogue (Duffy et al., 2012).

3. The galaxies were then injected into an empty image cube. At this point the dataset contains a perfectly clean set of galaxies distributed throughout the cube.

4. Mock visibility data (radio telescopes capture data in the visibility domain - the Inverse Fourier Transform of the image cube) was then generated from the image cube using the Miriad software (Sault et al., 1995).

5. The visibility data was then convolved with the dirty ASKAP beam, to introduce instrumental noise.

6. The resultant visibility data was passed through the Fourier Transform and then convolved with the clean image cube.

7. Finally Gaussian noise was distributed over the entire image cube following the profile of thermal noise expected in the ASKAP telescope.

The entire datacube is approximately 1 Terabyte in size with dimensions 3,600 × 3,600 × 23,060. The spatial region represented approximately 60°² in WCS (Greisen and Calabretta, 2002). Each frequency channel in the cube (third dimension) has a width of 18.518 KHz. Duchamp (Whiting, 2012) and several other used tools that were tested, required too much memory and computational power to feasibly process the entire cube at once, while meaningfully exploring the required parameter space of JPEG2000. Therefore extracts were taken from the original cube and each was tested independently.

The lower the frequency, the farther away the galaxies are in the cube due to the cosmological redshift of HI, and therefore the sources appear fainter towards the low frequency end of the cube. It was therefore important to test several subsets of data at different frequency ranges.

A subcube should also contain a sufficiently large number of sources in order to provide statistical significance in the tests. It was decided, arbitrarily, that the source finder should be able to identify at least 50 sources within the subcube.

At the low frequency end of DINGO datacube the signal-to-noise ratio for the sources becomes very small. It was established in tests that Duchamp was able to identify very few sources above the frequency plane ~15000, which was selected as a lower boundary for the frequency axes.

Table 1 shows the three subcubes extracted from DINGO cube to perform the tests. The Z references the frequency axes of the dataset. Larger Z corresponds to the lower frequency.

| Dataset | X  | Y  | Z   | Width  | Height | Depth |
|---------|----|----|-----|--------|--------|-------|
| A       | 0  | 0  | 4000| 3600   | 3600   | 100   |
| B       | 0  | 0  | 7000| 3600   | 3600   | 100   |
| C       | 0  | 0  | 10100| 3600  | 3600   | 100   |

Table 1: Subcubes selection within the DINGO datacube
2.2. Software

As JPEG2000 has not yet seen use in the radio astronomy domain many of the tools required, such as, source finders or image viewers did not provide direct support for the image standard. Several tools, therefore, were developed to support the experiments.

As the original dataset was stored using the HDF5 and FITS formats, we developed the skuareview-encoder and skuareview-decoder software[^2] to encode/decode the data to and from JPEG2000 using the JPEG2000 KDU library from Kakadu Software[^3].

The Duchamp source finder was used as the source identification software. Duchamp is the predecessor of the Selavy source finder (the source finder anticipated to be used in ASKAP and the SKA). Unfortunately Selavy was not made publicly available at the time of these works.

There are two particularly interesting aspects of Duchamp that are directly related to the experiments in this paper.

Firstly, Duchamp has the option to perform a wavelet reconstruction on the image cube before sources are searched for. This wavelet reconstruction aims to denoise the image. The “algorithme à trous” ([Starck et al., 1994]) is the wavelet transform used for this purpose as it maintains shift invariance where other wavelet transforms do not. The JPEG2000 image standard uses the Discrete Wavelet Transform (DWT) as a step in its compression algorithm. The DWT does not maintain shift invariance, however the algorithm is significantly less redundant as it includes subsampling ([Bradley, 2003]).

Secondly, Duchamp is known to be moderately inaccurate when parameterising sources found in the image ([Westmeier et al., 2012]). This must therefore be considered when measuring the accuracy of source parameterisation after JPEG2000 image compression.

2.3. Experiment design

2.3.1. Completeness and Soundness

The experiment aimed to provide a measure of the effect of the JPEG2000 image compression algorithm on the scientific quality of radio astronomy data. Here we define the scientific quality as the usefulness of the data to astrophysicists and their experiments. More specifically, we note that the only features important for DINGO science, found within a raw radio astronomy image cube are the HI sources within the cube; the vast majority of which represent the HI emission of distant galaxies. Everything else in the image cube is regarded as noise. As such, it is just the HI sources in the image that needed to be identified. This identification process needed to be as complete and as sound as possible, and each identified object needed to be parameterised as accurately as possible.

Source identification across an image cube is said to be complete if every source in the trueset catalogue is included in the identified set. The equation for completeness $C$ is given by

$$ C = \frac{P_{\text{true}}}{N} $$

where $P_{\text{true}}$ represents all true positive identifications, and $N$ is the number of sources in the trueset catalogue.

The source identification process is sound if every element in the identified set is included in the trueset catalogue. The equation for soundness $S$ is given by

$$ S = \frac{P_{\text{true}}}{P_{\text{total}}} $$

where $P_{\text{total}}$ is the total number of detections that includes both, positive and negative detections.

To obtain the effect the JPEG2000 compression has on the completeness and soundness of the source finders we use measure “completeness difference” $\psi$, and “soundness difference” $\omega$. These measurements are defined by

$$ \psi = C_{\text{compressed}} - C_{\text{original}} $$

$$ \omega = S_{\text{compressed}} - S_{\text{original}} $$

where $C_{\text{compressed}}$ and $S_{\text{compressed}}$ are the completeness and soundness respectively calculated on the compressed dataset, and $C_{\text{original}}$ and $S_{\text{original}}$ are the completeness and soundness respectively calculated the original dataset.

2.3.2. Source Parameterisation

After identification, a source also needs to be parameterised. Each source is attributed to a location in galactic coordinates Right Ascension (RA), Declination (Dec) and Frequency. Beyond this a variety of parameters are used to specify the attributes of each source. In this test we investigated the parameters listed below:

- Right Ascension
- Declination
- Frequency
- Right Ascension Width
- Declination Width
- Frequency Width
- Integrated Flux

We measured the difference in the performance of source parameterisation between the compressed dataset and the control dataset $\gamma_k$ for some JPEG2000 parameter $k$ and some source parameter type $p$ using the following equation

$$ \gamma_k = \text{RMSE}_\infty(p) - \text{RMSE}_k(p) $$

where $\text{RMSE}_\infty$ is the Root Mean Square Error (RMSE) between the source parameter value identified with the original dataset, and the true source parameter, while $\text{RMSE}_k$ is the RMSE between the source parameter value identified with the

[^2]: https://github.com/JCRAR/SkuareView
[^3]: http://www.kakadusoftware.com/
dataset compressed with JPEG2000 parameter $k$, and the true source parameter value.

The value $y_k$ will thus be positive if the compressed dataset allows for more accurate parameterisation and negative otherwise.

For example, to measure the effect of JPEG2000’s image compression algorithm on the parameterisation of sources with respect to the frequency width parameter of all sources identified, we calculate the RMSE of the frequency width of each source identified within the dataset by *Duchamp* against the true catalogue’s frequency width for the respective parameter. We then perform the same RMSE calculation using the sources identified by *Duchamp* after compression. If the difference between the resultant RMSE from the compressed dataset, and the RMSE from the original dataset is positive, then the compressed dataset has allowed *Duchamp* to provide more accurate parameterisation. However, if this difference is negative this would imply the compression has damaged the scientific quality of the data.

### 2.3.3. Cross matching

In order to correctly measure source parameterisation and the accuracy of the source finder, we need to ensure that the sources retrieved by the *Duchamp* are attributed to the correct source in the true catalogue.

This was done by iterating over all pairs $(u_i, v_j)$ where $u_i$ is the $i^{th}$ source obtained from the *Duchamp* source finder and $v_j$ is the $j^{th}$ source from the true catalogue. A pair was considered a match if the following conditions were true:

1. **Condition 1**
   - (a) The center of $u_i$ was contained within the bounds of $v_j$.
   - (b) OR the center of $u_i$ was within 3 voxels of the center of $v_j$.

2. **Condition 2**
   - (a) if multiple sources from the *Duchamp* catalogue were potential matches with a single $v_j$, the $u_i$ with the closest center to $v_j$ was chosen.

### 2.3.4. Comparison of JPEG2000 against the Wavelet Reconstruction in Duchamp

This experiment was intended to evaluate the JPEG2000 lossy compression as a denoising tool, by drawing a comparison between JPEG2000 lossy compression and the wavelet reconstruction algorithm used in *Duchamp*.

The differences in completeness, soundness and source parameterisation were calculated between *Duchamp* with a wavelet reconstruction and *Duchamp* without the wavelet reconstruction. These results were then graphed alongside the differences in completeness, soundness and source parameterisation of *Duchamp* between compressed imagery and uncompressed imagery.

Only one set of parameters was used for the wavelet reconstruction as described in Table 2 due to the fact that as the *Duchamp* wavelet reconstruction was computationally exhaustive. These parameters were chosen with reference to experiments performed on similar datasets (Popping et al., 2012; Westmeier et al., 2012).

| Parameter Name | Value | Comment |
|----------------|-------|---------|
| minVoxels      | 7     | Minimum voxels required to identify a source |
| flagAdjacent   | true  | Identified objects are merged using adjacency |
| snrCut         | 5     | Threshold in multiples of standard deviation for the IUWT (Starck et al., 2007) |
| scaleMin       | 2     | Minimum wavelet scale in the IUWT (Starck et al., 2007) |

Table 2: Duchamp source finder parameters

### 2.3.5. JPEG2000 parameter space

In our experiment we selected four of the most important parameters to investigate from the JPEG2000 parameter space:

- The **quantization step size** is extremely influential on the compression ratio and lossiness of the JPEG2000 compression algorithm. Smaller step sizes will result in less quantization and therefore less lossy compression. By exploring this parameter we can observe the influence of the JPEG2000 compression algorithm at different levels of lossiness.

- The number of **levels in the tree of the DWT** influences the structure of the wavelet domain before quantization and compression.

- **Precincts** partition the image cube into rectangles that are each encoded independently. This will effect how the wavelet domain will be supplied to quantization and compression resulting in differing compression ratios.

- The **Code block size** effects the size of the most granular partition in the JPEG2000 compression algorithm. Large block sizes will provide more opportunity for compression.

These parameters were explored as described in Table 3 due to the fact that as the

| Parameter Name               | Default | Start | End | Step |
|------------------------------|---------|-------|-----|------|
| Quantization step size       | 1/256   | 10^-5 | 0.01| ×10  |
| DWT levels                   | 5       | 1     | 32  | +4   |
| Precincts                    | 2^15    | 64    | 1024| ×2   |
| Code block size              | 64      | 4     | 64  | ×2   |

Table 3: JPEG2000 compression parameters iterated over in our experiment.
2.3.6. Procedure

A script was developed in order to perform this experiment over multiple JPEG2000 parameters and multiple datasets. The process is described below using the following function definitions;

- Duchamp($D$) takes dataset $D$ and returns catalogue $C$.
- Process($C$) takes catalogue $C$ and retrieves the completeness, soundness and source parameterisation results $R$.
- Encode takes dataset $D$, parameter type $j$ and parameter value $i$ and returns the JPEG2000 lossily compressed image.
- Decode takes the JPEG2000 lossily compressed image and decodes it into a FITS (required by Duchamp) dataset.

\[ C_t \leftarrow \text{true catalogue} \]
\[ D_o \leftarrow \text{original dataset} \]
\[ C_o \leftarrow \text{Duchamp}(D_o) \]
\[ R_o \leftarrow \text{Process}(C_o) \]

for all JPEG2000 parameter types $j$ in Table 3 do
  for all values $i$ for parameter type $j$ in Table 3 do
    item $C^i_j \leftarrow \text{Duchamp}(D^i_j)$
    item $R^i_j \leftarrow \text{Process}(C^i_j)$
  end for
end for

3. Results and Discussion

3.1. Compression Ratio and RMSE

Fundamentally, our most important choice when using any lossy compression algorithm, is to what degree the data can be compressed without the introduction of a significant error. We explored a variety of parameters in our experiments and measured the effect of a change on each parameter on the scientific quality of our dataset. This effect on scientific quality can be more intuitively understood with reference to each parameter’s effect on compression ratio and RMSE.

Figure 1 demonstrates a direct exponential correlation between the quantization step size and the compression ratio of the JPEG2000 compression algorithm. This correlation was consistent across each dataset used. The effect on scientific quality can be more intuitively understood with reference to each parameter’s effect on compression ratio and RMSE.

From Figure 2 we can observe that at a quantization step size of $1 \times 10^{-2}$, the RMSE has reached 1 standard deviation from the mean (as found in Table 4) for each respective dataset. This implies that the strongest sources may still be apparent in the image even at the extremely high compression ratio of over 15,000.

| Dataset | Mean       | Standard Deviation |
|---------|------------|--------------------|
| A       | -8.534e-10 | 3.650e-05          |
| B       | 9.588e-11  | 3.579e-05          |
| C       | 6.252e-09  | 3.532e-05          |

Table 4: Mean and Variance of each dataset.

Figures 3 and 4 show that more levels in the DWT in the lossy JPEG2000 compression algorithm, results in a higher compression ratio while decreasing the error in the samples. The degree of this change is, however, not as large as found in exploring the quantization step size parameter space. The compression increases from a minimum of approximately 8.5 to a peak of
higher than 11, while the RMSE differs only by as much as 1/100\(^{th}\) of the standard deviation of dataset A as seen in Table II.

### 3.2. Completeness and Soundness

#### 3.2.1. Completeness

In Figure 5 it can be seen that in dataset A the source identification algorithm achieves equivalent completeness when compressed for almost all values of quantization step size. There is a small dip just before an increase by as much as 3% in the completeness of the sources identified. This peak of completeness at a quantization step size of \(3 \times 10^{-3}\) corresponds with an extremely high compression ratio of over \(5 \times 10^{2}\). The final data point captured shows a significant drop in completeness which simply corresponds to the image eventually losing all scientific quality at extremely high compression.

Datasets B and C show similar results. In Figure 6 we observe that Duchamp achieves higher or equal completeness on a compressed dataset B for almost all data points. The completeness achieved in dataset B, was almost as high as 2%, occurring again at a high compression ratio of over \(4 \times 10^{2}\).

Finally Figure 7 shows a steady increase in the completeness of Duchamp with respect to the quantization step size in dataset C. Dataset C is particularly interesting as this dataset included, by far, the most sources. The vast majority of these sources are fainter and thus more difficult to observe. This experiment clearly shows the potential for JPEG2000 lossy compression to act as a denoising tool on spectral datasets to achieve higher completeness in source identification.

The number of DWT levels also positively influenced the completeness of Duchamp after JPEG2000 compression. Figure 8 shows that the completeness increases quickly when more than 1 level in the DWT is used, however increasing beyond 7 levels results in no further improvement. This result is expected as the DWT will divide each frame into a quad tree with the number of levels specified in this parameter. This binary (in 2 dimensions) structure results in a logarithmic diminishing
improvement from the DWT. A similar example of this kind of diminishing effect can be found in a simple binary search. If we aim to identify an interval that an element occurs in an ordered infinite element array, a binary search would logarithmically increase in accuracy after each iteration.

Our final two parameters, precinct size and block size, had no effect on the completeness or the soundness. This is somewhat expected as neither parameter has any direct effect on any of the lossy components within the JPEG2000 compression algorithm, rather they directly effect the lossless components e.g. run length encoding.

Figures 7, 8 and 9 show the completeness with respect to the integrated flux of a source for each dataset. The clear upward trend in all graphs is due to the fact that sources with higher total flux are easier identified. The dark ‘x’ data points correspond to the highest quantization step size where the image is extremely compressed, resulting in significant loss in completeness. Finally across all datasets it is apparent that more low integrated flux sources of less than 800 mJy per km/s are identified at higher quantization step size compressions, which again lends to the conclusion that JPEG2000 is denoising the spectral image cube allowing sources previously obscured by noise, to be identified.

In dataset A we found an improvement from the original completeness of 0.203 by ~3% to 0.23, in dataset B the improvement increased the completeness from 0.089 to 0.11 and in dataset C the improvement had a peak increase of completeness from 0.028 to 0.043. This result conclusively shows JPEG2000 to be having a denoising effect on the simulated DINGO cube dataset.

3.2.2. Soundness
Completeness cannot be considered independently from the soundness of the source identification algorithms.
In Figure 12 the soundness of Duchamp on dataset A under lossy JPEG2000 compression remains fairly consistent. A
small dip does exist in soundness at a quantization step size of approximately $1 \times 10^{-3}$. If we reference the respective completeness graph, Figure 5, we notice that the completeness appears to increase after this dip. This result is not unexpected as when the compression algorithm acts as a “filter” to the noise, the large structure instrumental noise will not be lost through the lossy compression. These stronger pieces of noise along with previously unidentified true sources are now more likely to be identified as a source. In fact with any source identification method that has a less than perfect soundness, if the completeness increases and the soundness remains stable both the true positives and false positives will increase. It is therefore quite probable that if just short of enough denoising is performed to identify a new true positive, there may be new false positives in our catalogue.

The soundness of Duchamp on dataset B and C stays the same or increases under JPEG2000 lossy compression for all data points excluding the final two as observed in Figures 13 and 14. In particular the soundness was only ever increased or exactly the same, at the average quantisation step size where completeness peaked. At high compression ratios JPEG2000 is known to occasionally cause ringing artefacts (Fang and Sun, 2007). If these ringing artefacts were to occur widely across the image the source finder may identify them as sources. This could potentially explain why we observe a loss in soundness at the highest compression ratios. We can note however, that this loss in soundness occurs after the average quantisation step size found to give peak completeness improvement to source identification. We therefore still find lossy JPEG2000 compression to improve completeness without loss in soundness at appropriate quantisation step sizes.

For all datasets however, the soundness fell dramatically for the final two quantisation step size. We can thus conclude that a quantisation step size of higher than $6 \times 10^{-3}$ will negatively effect accurate source identification.

The number of DWT levels on the other hand, had a surprisingly negative correlation with the soundness of the Duchamp source finder as seen in Figure 15. It should, however, be noted that all data points in this Figure indicated the soundness to decrease after compression. As we have seen this is not the case, when exploring over the quantisation step size space, we can conclude that this is the result of high default quantisation step size of $1/256$ used.

Overall we identify the quantisation step size to be the dominating parameter where a value of between $1 \times 10^{-3}$ and $3 \times 10^{-3}$ was found to be optimal to improve the completeness and soundness of source identification. Across all datasets the completeness and soundness trended upwards with respect to quan-
to see an increase in source identification completeness after reconstruction, where that increase was only 0.25%. The soundness was positively effected in datasets B in Figure 13 but notably less so than the peak soundnesses found by simply compressing the image cube. Dataset A saw absolutely no effect on either soundness or completeness after a Duchamp wavelet reconstruction.

This lack of significant effect can be attributed to the parameter space of the wavelet reconstruction not being fully explored. It is also important to note that the Duchamp source finder’s wavelet reconstruction has been improved in Duchamp’s successor Selavy, to the reconstruction used by the 2D-1D wavelet reconstruction source identification algorithm (Floer and Winkel, 2012). The “algorithme à trous” found in Duchamp was also far more computationally expensive than the JPEG2000 compression algorithm. Table 6 shows for each dataset how much longer Duchamp’s wavelet reconstruction took in comparison to how long encoding to JPEG2000, decoding back to FITS and then executing Duchamp without the wavelet reconstruction took.

It is, in fact, because of the computationally expensive nature of the wavelet reconstruction algorithm that a larger set of parameters could not be explored. Which lends to the hypothesis that the DWT in JPEG2000 as a denoising tool may be preferred over the “algorithme à trous” because while the DWT may include the undesirable trait of shift variance, the “algorithme à trous” is simply too slow for extremely large datasets. Furthermore, JPEG2000 can be used to generate compressed preview datasets for quality control and visual exploration of data. Once generated with optimal parameters the previews can be used for source finding purposes removing a need for prior denoising, and substantially improving I/O performance due to the smaller size of compressed datasets.

3.3. Source Parameterisation

3.3.1. Right Ascension and Declination

Overall the denoising effect of wavelet reconstruction in Duchamp was outperformed by the denoising effect of JPEG2000 image compression. In fact Duchamp’s wavelet reconstruction had little effect at all on completeness and soundness. As observed in Figure 14 dataset C was the only dataset
The spatial coordinates of sources identified by Duchamp appear to be more negatively affected by JPEG2000’s lossy compression. We can observe in Figure 16 that when compressed with a quantisation step size of higher than approximately $3 \times 10^{-4}$, the accuracy of RA source parameterisation drops off. This is notably well before the quantisation step size resulting in a peak in completeness.

While the Dec source parameterisation in Figure 17 follows a similar trend, we identified what appeared to be an error in either the trueset catalogue’s Dec parameter or Duchamp’s Dec parameterisation. Sources that were identified with the exact same centroid voxel as found in the true catalogue were often found to have a Dec difference by as much as a degree. We can therefore not make any conclusive judgement from our results with regard to Dec source parameterisation.

Figures 18 and 19 both show a loss in the accuracy of spatial width parameterisation at high compression ratios. As the voxels become more correlated to each other through compression the soft edges of the sources appear to be either stretched above or below the threshold of a source. This directly shows that at high compression ratios of the JPEG2000 lossy compression algorithm, the scientific quality of radio astronomy data is negatively effected. At much more reasonable compression ratios, however, the effect is zero.

Overall source parameterisation of spatial width remains unaffected until compressed with a quantisation step size greater than $1 \times 10^{-3}$ that corresponds to compression ratio approximately 1:100 (see Figure 1), and can concluded to be more stable than spatial location under lossy JPEG2000 compression.

### 3.3.2. Frequency

In Figures 20 and 21 one can observe that frequency and frequency width source parameterisation is almost entirely unaffected in every dataset. This result occurs for three reasons:

1. The sources in the image are point sources. They occupy
only a few voxels in the spatial domain, however they can span 100s of voxels in the spectral domain. This results in more stable source parameterisation in the spectral domain.

2. JPEG2000 performs the DWT in the spatial domain and as such will be most influential in the spatial domain.

3. The high frequency width of sources (100s of voxels) allows for the centroid frequency to differ more significantly when using our cross matching method described in Section 2.3.3. A source may be consistently identified across all quantization step sizes, while having an unusually high frequency difference. This type of outlier can have a dominating effect over the other sources’ frequency differences. Thus if the error in other sources changes dramatically and the outlier remains consistent the RMSE will not change as dramatically.

3.3.3. Integrated Flux

Figure 22 shows that across all datasets the parameterisation of integrated flux performs poorly at high compression rates (low quantisation step sizes). At high quantisation step sizes the domain of wavelet coefficients becomes more discretised. As these coefficients become more discretised the reconstruction will result in differing wavelet coefficient amplitudes. While the wavelet transform will still be capable of maintaining the structure of sources relative to the rest of the image, the original pixel amplitudes will change after reconstruction. Thus while the sources may still have amplitudes higher than neighbouring voxels, the actual value of the amplitude will have changed and while completeness can be increased, integrated flux source parameterisation is damaged.

In Section 3.2 it was identified that the peak completeness and soundness occurred between quantisation step sizes of $1 \times 10^{-3}$ and $3 \times 10^{-3}$. We can clearly see in Figure 22 that by this point the source parameterisation of the source identification algorithm had been damaged.

We thus conclude, that in order to maintain as accurate as possible source parameterisation while using JPEG2000’s lossy compression algorithm, one should not exceed the conservative limit of a compression ratio of 1:100 (or a quantisation step of approximately $1 \times 10^{-3}$). Any kind of measurable negative impact on source parameterisation had not been observed at all on the compression ratios below 12, thus we can conclude that the compression did not affect the data in any negative way.

4. Conclusion

4.1. Summary

JPEG2000 has been found to have a negligible effect on the scientific quality of radio astronomy imagery up to a compression ratio of approximately 12. Thereafter, source parameterisation would progressively become less accurate. However at compression ratios higher than 100 the completeness and soundness of the Duchamp source finder was increased. In particular the completeness increased by as much as 3%, while the soundness peaked with an increase of over 7%. This is the result of lossy wavelet JPEG2000 compression algorithm denoising the image such that previously obscured sources could be identified.

While the increase in completeness and soundness only occurred when source parameterisation had become less accurate, the result may still be useful. Further study is necessary to compare introduced errors due to the compression with other errors already present in the data. Such a study needs to be done with real rather than synthetic data.

Encoding using DWT compression found in JPEG2000 has been found an order of magnitude faster than the “algorithme `a trous” found in Duchamp, and yet our results indicated the denoising effects of both to be comparable with JPEG2000 being slightly better. Future work should be placed into the investigation and implementation of a DWT based wavelet reconstruction, with quantisation and thresholding, to be used in source finders.
5. Acknowledgements

A special thanks Prof. Amitava Datta for his guidance while preparing this work. Kakadu Software Ltd. and personally Prof. David Taubman for allowing us to freely use the software. Dr. Martin Meyer and Dr. Tobias Westmeier for their support in the use of the simulated DINGO cube dataset, and Dr. Attila Popping for other guidance in the Radio Astronomy domain.

References

Boliek, M., 1996. New work item proposal: JPEG2000 image coding system. Technical Report N390. ISO/IEC JTC1/SC29/WG1.

Boyce, P.J., 2003. Gamma Finder: a Java application to find galaxies in astronomical spectral line datacubes. M.sc. dissertation. School of Computer Science. Cardiff University.

Bradley, A., 2003. Shift-invariance in the discrete wavelet transform. in: DICTA, pp. 29–38. URL: http://www.cms.csiro.au/Hugues.Talbot/dicta2003/cdrom/pdf/0029.pdf

DeBoer D. et al., 2009. Australian ska pathfinder: A high-dynamic range wide-field 1-2 GHz Research on Galaxy Evolution. URL: http://www.publish.csiro.au/paper/AS11042

Fang, J., Sun, J., 2007. Ringing artifact reduction for jpeg2000 images., in: ICIC (3), Springer. volume 2 of Communications in Computer and Information Science. pp. 1026–1034. URL: http://dltp.uni-trier.de/db/conf/icic/icic2007-3.html#Fang07

Floer, L., Winkel, B., 2012. 2d-1d wavelet reconstruction as a tool for source finding in spectroscopic imaging surveys. Publications of the Astronomical Society of Australia 29, 244–250. URL: http://dx.doi.org/10.1071/AS11042

Greisen, E., Calabretta, M., 2002. Representations of world coordinates in fits. Astron. Astrophys. 395, 1061–1075. URL: http://dx.doi.org/10.1051/0004-6361:20021326

Huynh, M., Lazso, J., 2013. An Overview of the Square Kilometre Array. ArXiv e-prints:arXiv:1311.4288

Jurek, R., 2012. The characterised noise hi source finder: Detecting hi galaxies using a novel implementation of matched filtering. Publications of the Astronomical Society of Australia29, 251–261. doi:10.1071/AS11044 arXiv:1112.1561

Kitaeff, V.V., Wu, C., Wicenec, A., Cannon, A.D., Vinsen, K., 2012. Skaeview: Client-server framework for accessing extremely large radio astronomy image data., in: Proceedings of the 2012 Workshop on High-Performance Computing for Astronomy Date, ACM, New York, NY, USA. Astro-HPC ’12. pp. 25–32. URL: http://doi.acm.org/10.1145/2286976.2286984

Meyer, M., 2009. Exploring the HI Universe with ASKAP, in: Panoramic Radio Astronomy: Wide-field 1-2 GHz Research on Galaxy Evolution. arXiv:0912.2167

Popping, A., Jurek, R., Westmeier, T., Serra, P., Floer, L., Meyer, M., Koribalski, B., 2012. Comparison of potential askap hi survey source finders. Publications of the Astronomical Society of Australia 29, 318–339. URL: http://journal.cambridge.org/article_S132335800000134X
doi:10.1107/AS11067

Sauldt, J., Teuben, J., Wright, H., 1995. A retrospective view of MIRIAD, in: Astronomical Data Analysis Software and Systems IV, volume 77 of Astronomical Society of the Pacific Conference Series. pp. 433+. URL: http://adsa.org/adass/proceedings/adass94/saultr.html

Serra, P., Jurek, R., Floer, L., 2012. Using negative detections to estimate source-finder reliability. Publications of the Astronomical Society of Australia 29, 296–300. URL: http://journals.cambridge.org/article_S1323358000001314
doi:10.1071/AS11066

Starck, J., Murtagh, F., Bijaoui, A., 1994. Image restoration with noise suppression using a wavelet transform and a multiresolution support constraint, in: Image Reconstruction and Restoration, SPIE. volume 2302. pp. 132–143. doi:10.1117/12.188035

Starck, J.L., Fadili, J., Murtagh, F. 2007. The undecimated wavelet decomposi-
tion and its reconstruction. IEEE Transactions on Image Processing 16, 297–309. doi:10.1109/TIP.2006.887733

Taubman, D.S., Marcellin, M.W., 2002. JPEG2000 : Image Compression Fundamentals, Standards and Practice. Kluwer Academic Publishers, Boston.

Westmeier, T., Popping, A., Serra, P., 2012. Basic testing of the duchamp source finder. Publications of the Astronomical Society of Australia 29, 276–295. URL: http://journals.cambridge.org/article_S1323358000001302
doi:10.1071/AS11041

Whiting, M., 2012. Duchamp: a 3d source finder for spectral-line data. Monthly Notices of the Royal Astronomical Society , 3242–3256URL: http://arxiv.org/abs/1201.2710

Whiting, M., Humphreys, B., 2012. Source-finding for the australian square kilometre array pathfinder. Publications of the Astronomical Society of Australia 29, 371–381. URL: http://journals.cambridge.org/article_S1323358000001387
doi:10.1071/AS12028