# SKA SCIENCE DATA CHALLENGE 1: ANALYSIS DESCRIPTION

Document Number ........................................ SKA-TEL-SKO-0001081
Document Type ........................................ REP
Revision ........................................ 01
Author ........................................ A.Bonaldi
Date ........................................ 2019-07-12
Document Classification ........................................ UNRESTRICTED
Status ........................................ Released

| Name | Designation | Affiliation | Signature | |
|---|---|---|---|---|
| Authored by: | | | | |
| SKAO Science Team p.p. Anna Bonaldi | Project Scientist | SKAO | Anna Valentina Bonaldi | |
| | | | Date: | 2019-07-14 |
| Owned by: | | | | |
| Robert Braun | Science Director | SKAO | Robert Braun | |
| | | | Date: | 2019-07-15 |
| Approved by: | | | | |
| Robert Braun | Science Director | SKAO | Robert Braun | |
| | | | Date: | 2019-07-15 |
| Released by: | | | | |
| Joe McMullin | Programme Director | SKAO | J P McMullin | |
| | | | Date: | 2019-07-26 |

# DOCUMENT HISTORY

| Revision | Date Of Issue | Engineering Change Number | Comments |
|---|---|---|---|
| A | 2019-01-09 | - | First draft release for internal review |
| 01 | 2019-07-12 | | First Release |
| | | | |

# DOCUMENT SOFTWARE

| | Package | Version | Filename |
|---|---|---|---|
| **Word processor** | MS Word | Word 2007 | SKA-TEL-SKO-0001081-01-datachallenge analysis |
| **Block diagrams** | | | |
| **Other** | | | |

# ORGANISATION DETAILS

| | |
|---|---|
| Name | SKA Organisation |
| Registered Address | Jodrell Bank Observatory<br>Lower Withington<br>Macclesfield<br>Cheshire<br>SK11 9DL<br>United Kingdom<br><br>Registered in England & Wales<br>Company Number: 07881918 |
| Fax. | +44 (0)161 306 9600 |
| Website | www.skatelescope.org |

# TABLE OF CONTENTS



# LIST OF FIGURES



# LIST OF TABLES

# LIST OF ABBREVIATIONS

AGN……………Active Galactic Nucleus

FoV………….... Field of View

FS……………...Flat Spectrum

SDC1…………..Science Data Challenge 1

SFG……………Star-Forming Galaxy

SS……………...Steep Spectrum

# 1 Introduction

## 1.1 Purpose of the Document

The purpose of this document is to provide information on how the results submitted for the SKA data challenge #1 have been analysed and scored.

## 1.2 Scope of the Document

In this document, we describe how we scored the results of the SKA data challenge #1.

# 2 References

## 2.1 Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, **the applicable documents** shall take precedence.

[AD1] Applicable Document 1
[AD2] Applicable Document 2

## 2.2 Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, **this document** shall take precedence.

[RD1] SKA-TEL-SKO-00001001 (arXiv:1811.10454, Bonaldi et al. 2019)
[RD2] Riccio, et al. 2016 arXiv:1611.04431, http://dame.dsf.unina.it/c3.html

# 3 Introduction

Participants of the Science Data Challenge 1 (SDC1, RD1) were asked to submit results in the form of catalogues of detected sources (hereafter Submitted catalogues), ideally for the 3 frequencies $\nu = 560, 1400, 9200\ MHz$ and the 3 depths $T = 8, 100, 1000\ h$ for which images had been released. Submitted catalogues should contain the following entries:

COLUMN1: ID [none] Source ID
COLUMN2: RA (core) [degs] Right ascension of the source core
COLUMN3: DEC (core) [degs] DEcination of the source core
COLUMN4: RA (centroid) [degs] Right ascension of the source centroid
COLUMN5: DEC (centroid) [degs] Declination of the source centroid
COLUMN6: FLUX [Jy] integrated flux density
COLUMN7: Core frac [none] integrated flux density of core/total
COLUMN8: BMAJ [arcsec] major axis dimension
COLUMN9: BMIN [arcsec] minor axis dimension
COLUMN10: PA [degs] position angle
COLUMN11: SIZE [none] 1,2,3 for LAS, Gaussian, Exponential
COLUMN12: CLASS [none] 1,2,3 for SS-AGNs, FS-AGNs, SFGs

These catalogue entries match those of the input catalogues (hereafter True catalogues) that have been used to generate the SDC1 images. The analysis and the scoring of the submissions consist in the following steps:

1. Crossmatch of each Submitted catalogue with the corresponding True catalogue;
2. Categorise the detections as either true matches or false positives and produce Completeness and Reliability plots.
3. Compute an overall score for each team and identify a challenge winner.

We describe these steps in more details in the following sections.

# 4 True and Submitted catalogue crossmatch

The crossmatch of the True and the Submitted catalogues posed some challenges, which ultimately resulted in a custom procedure. The specific issues are:

1. Computational complexity. The Catalogues are very big (at 560 MHz, Ntrue=$10^7$ and typically Ndet=$10^5$) which would have made a procedure based on a simple Ntrue X Ndet cycle very time-consuming.
2. Source crowding. The True catalogue contains a very high density of sources (50 sources per square arcmin at 560 MHz on average). This means that a position-only crossmatch, such as implemented in several widely used tools, would likely fail to match sources correctly; given the likely error in the source position, there could be another true source that is closer to the estimated position but below the detection limit.

In order to overcome the two issues noted above, we introduced some additional features in the custom crossmatch procedure, described in the following two sub-sections.

## 4.1 Binning of the True and Submitted Catalogues

We binned the True and Submitted catalogues in 5 declination bins and 5 (logarithmic) flux density bins, therefore splitting them into 25 sub catalogues. The crossmatch described in 4.2 has been performed between catalogues belonging to the same bin, thus reducing computational complexity significantly.
To allow for errors in declination and in flux density, the True catalogue bins have the same bin centre but twice the size of the Submitted catalogue bins. This allows half-bin overlap between adjoint True catalogue bins.

Given that for each bin i=1,25 typically Ndet,i=Ndet/25 and Ntrue,i=2×Ntrue/25 (where the factor 2 is for the bin overlap) the computational complexity reduces from Ntrue×Ndet to 25×2×Ntrue/25 × Ndet/25 which is a factor 12.5.

## 4.2 Using flux density and size as well as position in the crossmatch

The crossmatch procedure minimizes a multi-dimensional distance between the true and estimated sources, which considers position, size and integrated flux density.

$$D = \sqrt{Dpos^2 + Dsize^2 + Dflux^2}$$

The three positional, size and flux mis-matches are defined as:

$$Dpos \propto \frac{\sqrt{(x - x')^2 + (y - y')^2}}{S'}$$

$$Dsize \propto \frac{[S - S']}{S'}$$

$$Dflux \propto \frac{[f - f']}{f'}$$

where

- The ∝ means each term has been further normalised, to ensures they contribute equally to D. The normalizations are 1/3σ of the distribution of the position/size/flux error across all the teams; once all three contributions added together, D is normalised as 1σ.
- Prime denotes the attributes of the True Catalogue;
- x,y are the pixel coordinates corresponding to ra,dec;
- S is the average source size =(Bmaj+Bmin)/2;
- f is the source integrated flux density.

In practice, to take advantage of existing crossmatch code and reduce the computational complexity of the purpose-written code, the custom crossmatch procedure has been implemented in two steps:

1. A positional-only crossmatch using the publicly available software package C3 [RD2]. This crossmatch was configured to select all true sources with Dpos <3 as candidate matches. As detailed in the list in Section 3, the source position could be entered as a position of the source

core and/or the source centroid. The positional crossmatch has been performed on both positions and all candidate matches have been retained.

2. The previous step results in a redundant cross-matched catalogue containing typically more than one true source for every detection. This is further analysed to reduce it to either one or no match per source, based on the multi-dimensional distance D as previously defined. The true source having minimum D has been kept as a match if D<5 (within 5σ of the global error distribution).

### 4.3 Running a null-test crossmatch for each Submitted catalogue

To quantify the residual level of chance matches between Submitted and True catalogue entries due to the source crowding, we created a copy of each Submitted catalogue (which we call Null-test catalogue) and randomised the position of sources within the field of view. In this way, the Null-test catalogue has the same flux and size distribution as the Submitted catalogue but should have no matches with the True catalogue, apart from chance matches. We applied to the Null-test catalogues the same cross-match procedure described in 4.1 and 4.2. This has been used initially to calibrate the thresholds for accepting matches as to minimise chance matches; with the set thresholds, this contamination is below 10% for most submissions. The number of chance matches quantified by these null tests have also been used to correct the Completeness and Reliability metrics, as detailed on the next section.

## 5 Completeness and reliability

We defined completeness C and reliability R as:

$$C(logF') = \frac{Nmatch(logF') - Nnull(logF')}{Ntrue(logF')}$$

$$R(logF) = \frac{Nmatch(logF) - Nnull(logF)}{Ndet(logF)}$$

Where:

- F is the integrated apparent flux density (before primary beam correction) and the prime denotes the True catalogue value;
- Nmatch is the histogram of the cross-match between the True and the Submitted catalogue;
- Nnull is the histogram of the cross-match between the True and the Null-test catalogue;
- Ntrue is the histogram of the True catalogue;
- Ndet is the histogram of the Submitted catalogue.

We note that C is measured as a function of the True catalogue entries (logF’) and R as a function of the Submitted catalogue entries (logF). This is because the True catalogue and the Submitted catalogue contain only logF’ and logF respectively. Since the cross-matched catalogues in the numerators contain both logF and logF’, we computed both histograms in order to achieve consistency with the denominators. This always guarantees the correct normalization of C and R even in the presence of errors in the estimation of the flux F. Whenever such errors are significant, this may however cause some discrepancy between C and R. We also note the Nnull term in the numerator, which is meant to correct for the contamination due to chance matches.

# 6 Overall score and challenge winner

The overall score has been defined to reflect not only completeness and reliability but also the accuracy of the estimation of the source properties.
Given a Submitted catalogue for observations at frequency $\nu$ and integration time T, the score per catalogue has been computed as:

$$\delta^T(\nu) = \frac{\sum_{i=1}^{Nmatch} w_i - Nfalse}{FoV(\nu)}$$

Where:

- Nmatch is the number of true detections;
- $0 \le w_i \le 1$ is a weight per source to take into account the accuracy of characterization of the true detections (see 6.1 for more details);
- Nfalse=Ndet-Nmatch is the number of false detections;
- FoV is the field of view for that frequency

The total score for integration time T has been obtained by adding the scores for the 3 frequencies considered:

$$\Delta^T = \sum_{\nu} \delta(\nu)$$

As initially stated in RD1, the challenge winner is the one with the **highest score for the deepest integration time (T=1000h)**.

## 6.1 Definition of the weight per source $w_i$

The source properties considered for computing $w_i$ are seven: position (the best between core and centroid position if both are present), flux density, core fraction, major axis, minor axis, position angle and class. For all of them except the source class, the score per source $i$ and per attribute $j$, $w_i^j$ , is defined as

$$w_i^j = \frac{1}{7} min\left\{1, \frac{err_i^j}{thr^j}\right\}$$

Where $err_i^j$ is the error on the attribute $j$ for the source $i$ and $thr^j$ is a threshold set on that attribute for all sources. The behaviour of $w_i^j$ is shown in Figure 1. The definition of errors and thresolds for all source properties are in Table 1. The thresholds have been calibrated on the error statistics for the submissions, so that the full range of scores from 0 to 1/7 are awarded on the full sample of results.
In the case of the source class, $w_i^j$ is either 1/7 or 0 depending on whether the source has been classified correctly or not. The final score per source is finally

$$w_i = \sum_{j=1}^{7} w_i^j$$

The 1/7 normalization of $w_i^j$ guarantees that the maximum value of $w_i$ is 1; this is awarded whenever all attributes of that source have been estimated with an error lower than the set thresholds.

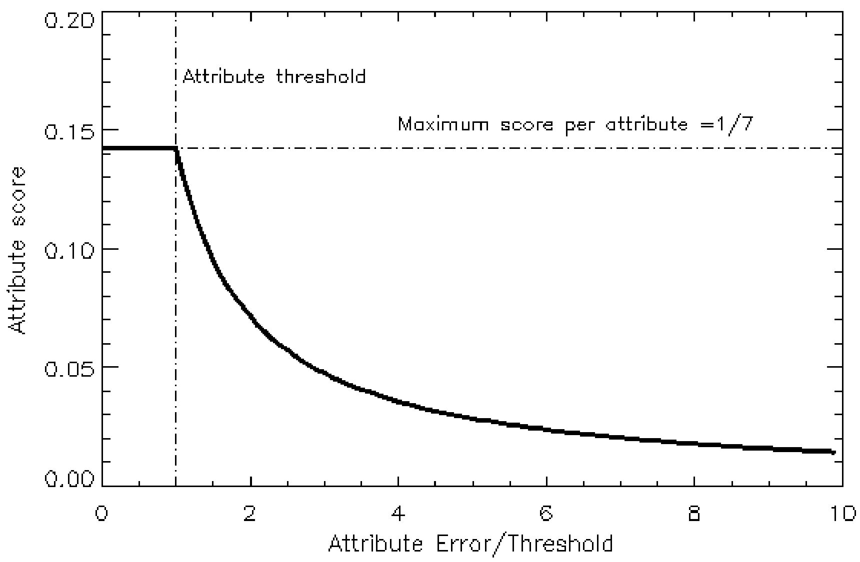


Figure 1: Score awarded per attribute as a function of attribute error/attribute threshold

Table 1: Definitions of Errors and Thresholds for the properties of sources

| | Error | Threshold |
|---|---|---|
| Position | $\frac{\sqrt{(x-x')^2+(y-y')^2}}{S'}$ | 0.3 |
| Flux density | $\frac{\lceil f-f' \rceil}{f'}$ | 0.1 |
| Major axis | $\frac{\lceil Bmaj-Bmaj' \rceil}{Bmaj'}$ | 0.3 |
| Minor axis | $\frac{\lceil Bmin-Bmin' \rceil}{Bmin'}$ | 0.3 |
| Position angle | $\lvert pa - pa' \rvert$ | 10 |
| Core fraction | $\frac{\lceil cf-cf' \rceil}{0.75}$ | 0.05 |