Optimization of galaxy identification algorithms in large HI surveys

T Gqaza[1], R C Kraan-Korteweg[1], B Frank[1], M Ramatsoku[1,3], T H Jarrett[1], E Elson[1] and A C Schroeder[2]

[1] Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch, 7700, Cape Town, South Africa

[2] South African Astronomical Observatory, PO Box 9, Observatory, 7935, Cape Town, South Africa

[3] Kapteyn Astronomical Institute, University of Groningen, Landleven 12, 9747 AV Groningen, The Netherlands

E-mail: gqztem001@myuct.ac.za

Abstract. The upcoming neutral hydrogen (HI) blind SKA-precursor surveys like Fornax and LADUMA, will produce extremely large volumes of spectral data cubes. Fully automated source-finding and parametrization algorithms will prove more efficient than visual examination methods. Such algorithms have been developed and rigorously tested on simulated HI data cubes. Their performance is not fully known when it comes to spectral cubes with true HI line emission. In this paper, we present preliminary results on the comparison of three galaxy identification methods (i.e. visual, semi-automated and fully automated). For these tests, the Westerbork Synthesis Radio Telescope (WSRT) Perseus-Pisces (PP) HI data cube is used. Visually, we detected 194 galaxy candidates, of which 90.2% have semi-automated cross-matches. We also present preliminary results from the initial run of SoFiA applied on 44.4% of the data cube. The final outcome, after the full comprehensive analysis is finalised, will be fed back to pipeline developers for possible optimization.

1. Introduction

An in-depth study of the neutral hydrogen content of the universe is of paramount importance for understanding star-formation, gas and galaxy evolution. As a result various HI surveys are planned for the Square Kilometre Array (SKA) precursors MeerKAT, APERTIF and ASKAP (e.g. LADUMA [1], Fornax HI survey [2], Northern sky HI shallow survey [3] and WALLABY [4]). These surveys will produce large quantities of HI data cubes containing up to hundreds of thousands galaxies. Hence the traditional method of galaxy identification through visual inspection will prove cumbersome. An HI data cube is a three-dimensional representation of HI-line emission where two of the axes are spatial positions (e.g. RA and Dec) and the third one is frequency (i.e. the velocity of the HI emission).

In preparation for this big data epoch, a handful of fully-automated source-finding and parametrization software have been developed. These software are based on advanced source-finding algorithms like the Characterised Noise HI (CNHI) source-finder [5]; 2D-1D Wavelet Reconstruction source-finder [6]; the Smooth Plus Clip (S + C) source-finder [7] and the standard source-finder for the Australian SKA pathfinder (ASKAP), DUCHAMP [8].

In 2012, Popping et al. [9] tested three of the aforementioned source-finding algorithms. Their focus was on testing the reliability and completeness of each algorithm. The tests were conducted on two different 3D spectral data cubes, which contained simulated HI sources and continuum sources, respectively. In

parallel the same group tested DUCHAMP on a data cube containing real sources [10]. These studies partially led to the development of a more advanced and flexible fully-automated software known as the Source-Finding Application (SoFiA) [11]. SoFiA's primary objective is to search a spectral cube and identify sources, then extract HI parameters. It is the first software that combines three different source-finding algorithms (i.e. CNHI, S + C and the basic threshold source-finder). SoFiA uses negative detections to quantify source reliability [7]. It also uses busy functions [**?**] to describe HI global profiles of the detected sources [12].

In this paper, we present preliminary results of a comparison of three different source-finding methods, namely visual inspection, semi-automated and fully automated. Fully automated software like SoFiA and its predecessors are ideal for bigger surveys. They can process large amounts of data in relatively short time frames. They also utilize recent statistical methods to quantify reliability. Knowing how these fully automated software tools compare to visual inspection on large spectral data cubes is critical. For smaller surveys (hundreds of sources) a combination of automated and Visual identification methods might produce complete and reliable results. Hence we also explore semi-automated source identification methods.

2. HI data cube

The data cube used in this paper is the Perseus-Pisces Supercluster (PP) hexagonal mosaic covering 9.6 sq.deg. It was observed in 2012 with the Westerbork Synthesis Radio Telescope (WSRT) in the Netherlands. The data cube is composed of 35 pointings and each is separated by 0.5 deg. Each pointing has a total integration time of 2×6 hours. The total effective bandwidth of the volume surveyed is 67 MHz. It covers a Doppler-shifted velocity range of $cz = 2400 - 16600$ km s^{-1}. The data cube has rms noise of 0.4 mJy.

3. Methodology

3.1. Visual inspection: Galaxy identification and parametrization

We intended to produce a reliable WSRT ZoA[1] PP source catalogue by visual search. We achieved this by having three authors search 4 out of 9 subcubes spanning the entire velocity range of $2400-16600$ km s^{-1}, using a visualisation tool (KVIS) from KARMA [13] . Each searcher compiled a candidate list. All three lists were handed to one author who acts as an adjudicator, to produce a final candidate list. In addition the adjudicator searched the rest of the subcubes. The HI parametrization was carried out using a specialized python script and MBSPECT module from MIRIAD [14]. For each candidate a corresponding sub-volume was extracted, from which the weighted emission sum along the spectral line was calculated. Each resulting one-dimensional spectrum was visualised and a lower-order polynomial was fit to the channels without HI emission and subtracted. The integrated flux density ($S_{\rm int}$) and the peak flux ($S_{\rm peak}$) were calculated across the channels with line emission. The systemic velocity ($V_{\rm sys} = cz$) was taken as the average of velocities corresponding to 50% of the peak flux from the line profile. Linewidths at 20% and 50% level of the peak flux density (ω_{20} and ω_{50}) were calculated using a width-maximiser method from MBSPECT. For each detection a zeroth moment (M_0) map was produced by collapsing the subcube along the spectral axis. Another miriad module (i.e. IMSAD) was used to fit a Gaussian to the histogram of the M_0 in order to get a flux-weighted centroid of the detected candidate.

3.2. Semi-automated source identification

In 2016, Ramatsoko et al. [15] published a source catalogue of the WSRT ZoA PP data cube. Here, we present the summary of the galaxy identification procedure they used (for details, see [15]). They first corrected for spatial noise variation by multiplying the cube by an inverse square weighted noise (σ^{-2}) in each of the 35 pointings. The original cube of spatial resolution (2300×1600) was smoothed up to (3000×3000). The resulting cube was then smoothed in velocity to four different resolutions, namely: Hanning smoothing (16.5 km s^{-1}) and a Gaussian smoothing kernel corresponding to four, six and eight channels (i.e. 33, 49.5 and 66 km s^{-1}). They ran the Groningen Image Processing System (GIPS) software on all eight different angular and spectral resolution combinations. A detection was accepted if it met

[1] Zone of Avoidance (ZoA) is the region of the sky which appears devoid of extragalactic objects when viewed on optical wavelengths.

the galaxy criteria explained in Ramatsoko et al. [5]. This method led to the detection of 683 galaxy candidates. After post visual inspection of all candidates, 235 out of 683 were identified as imaging artefacts or RFIs[2] and were rejected. Further analysis led to a rejection of another 237 candidates as they had features consistent with noise peaks. This resulted in a semi-automated catalogue with 211 galaxies.

4. Early results

4.1. Visual and semi-automated

A total number of 194 detections is achieved through visual inspection of the entire spectral cube. Figure 1 shows the distribution of total HI mass as a function of radial velocity (also known as a sensitivity curve). The black and the red curves show predicted HI mass limits of this survey assuming a 3σ flux detection for 100 and 250 $\mathrm{km\,s^{-1}}$ linewidth galaxies, respectively. Green dots indicate detections with ω_{50} less than 100 $\mathrm{km\,s^{-1}}$, red dots are detections with ω_{50} in-between 100 and 250 $\mathrm{km\,s^{-1}}$ and blue dots have ω_{50} greater than 250 $\mathrm{km\,s^{-1}}$. The detected candidates have a total HI mass ($\log(\mathrm{M_{HI}/M_{\odot}})$ ranging from 7.81 to 10.24 (see the right panel of Figure 2). The visual method finds low HI mass detections across the entire velocity range as well as narrow-linewidth galaxy candidates.

4.2. Semi-automated counterpart

We use a position-velocity-based algorithm to search for cross-matches. Each galaxy has a unique flux weighted centroid, but it can slightly differ to that of its counterpart due to the manner in which it was derived. To counterpoise this bias, we allow a spatial (Δs) and spectral deviation (Δv) of 3000 and 100 $\mathrm{km\,s^{-1}}$, respectively from the centroid. Let us suppose galaxy X with coordinates (ℓ, b, v) has a counterpart X', then X' coordinates (ℓ', b', v') must conform to Eq. 1, where s is either the Galactic longitude (ℓ) or latitude (b).

$$s - \Delta s \leq s' \leq s + \Delta s \quad v - \Delta v \leq v' \leq v + \Delta v \tag{1}$$

Table 1. Summary of cross match galaxies between visual and semi-automated output catalogue.

Measurements	**Visual**	**Semi-automated**
`No.`of galaxies	194	211
`No.`of galaxies with counterparts	175 (90.2%)	175 (82.9%)
`No.`of galaxies without counterparts	19 (9.8%)	36 (17.1%)
`Narrow-linewidths:`$\omega_{20} \leq 100$ $\mathrm{km\,s^{-1}}$	47.4%	19.4%
`Intermediate-linewidths:`$100 \leq \omega_{20}$ $\left[\mathrm{km\,s^{-1}}\right] \leq 250$	42.1%	77.8%
`Massive-linewidths:`$\omega_{20} > 250$ $\mathrm{km\,s^{-1}}$	10.5%	2.8%

[2] RFIs are man made Radio Frequency Inference signals (e.g. Global Positioning Satellites) that could be orders of magnitudes stronger than the observed celestial signal.

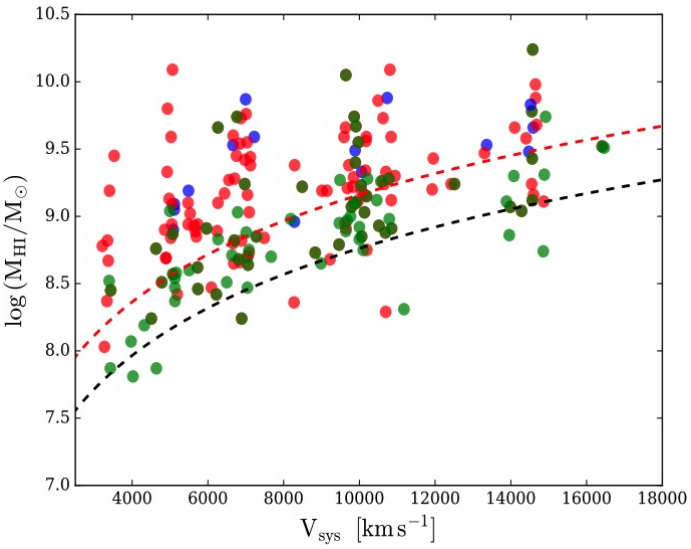

Figure 1. The logarithm of total HI mass of all detected galaxies (through Visual inspection) as a function of radial velocity. The red and the black dashed lines show the HI mass limit of the WSRT dataset assuming a 3σ detection with 100 and 250 $\mathrm{km\,s^{-1}}$ linewidths, respectively. The green dots indicate detected galaxies with measured linewidths less than 100 $\mathrm{km\,s^{-1}}$. The red dots indicate galaxies of linewidths between 100 & 250 $\mathrm{km\,s^{-1}}$ whereas the blue dots show galaxies with linewidths greater than 250 $\mathrm{km\,s^{-1}}$.

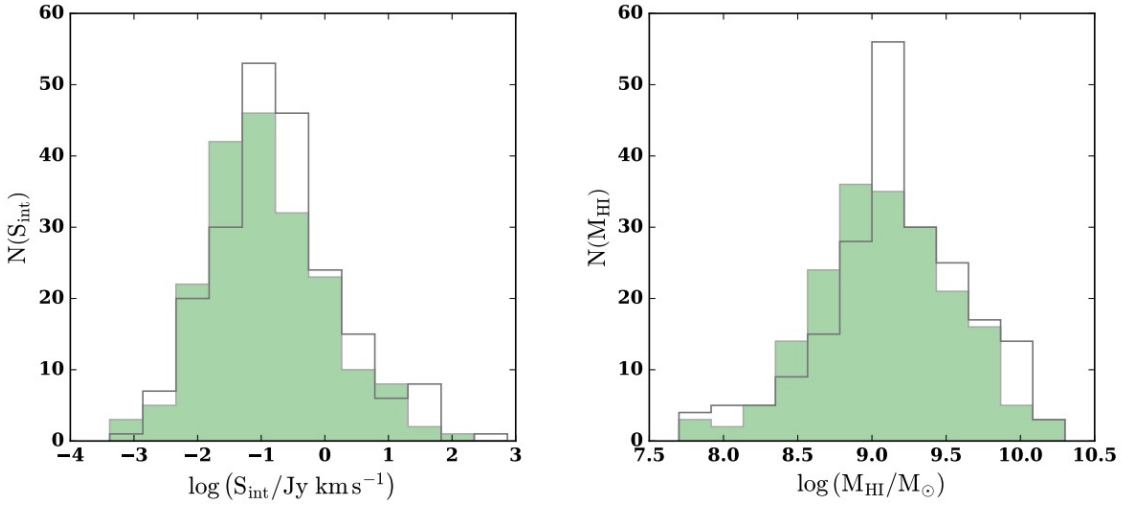

Figure 2. HI parameter comparison between visual and semi-automated galaxy detection in the WSRT PP HI data cube. Green indicates the HI distribution based on the visual catalogue, whereas the open grey histogram is based on semi-automated results. Left panel: logarithm distribution of integrated flux $\log(\mathrm{S_{int}/Jy\ km\,s^{-1}})$, right panel: the logarithmic distribution of HI mass $\log(\mathrm{M_{HI}/M_\odot})$.

Of the 194 visual detections, 175 (90.2%) semi-automated cross-matches were found. Table 1 presents a cross match summary between the two methods. In Figure 2 we compare the distribution of the HI parameters of identified galaxies from visual versus semi-automated catalogues (see Sect. 3.2). The right panel shows the logarithmic distribution of total HI mass. The green histogram represents visual detections, and the non-filled grey histogram represents the semi-automated distribution. The semi-automated HI mass distribution ranges from $\log(\mathrm{M_{HI}/M_\odot}) = 7.70$ to 10.30 with a mean HI mass of $\log(\mathrm{M_{HI}/M_\odot}) = 9.15$. On the other hand, the visual HI distribution ranges from $\log(\mathrm{M_{HI}/M_\odot}) = 7.81$ to 10.24, with a mean of $\log(\mathrm{M_{HI}/M_\odot}) = 9.08$. The left panel shows the logarithm of the integrated line flux, with the visual integrated mean line flux being 0.67 $\mathrm{Jy\,km\,s^{-1}}$ and the minimum detected flux

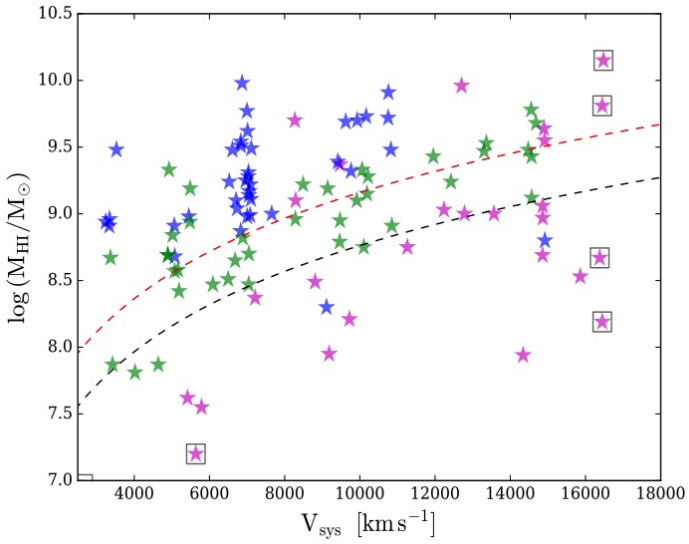

Figure 3. The logarithm of total HI mass of all detected galaxies (through visual inspection) as a function of radial velocity. The red and the black dashed lines show the HI mass limit of our survey assuming a 3σ detection with 100 and 250 km s^{-1} linewidths. Blue stars indicate galaxies identified by both visual inspection and SoFiA, green stars: galaxies identified by the visual method only. Magenta stars: galaxy candidates identified by SoFiA only. Stars enclosed by black open squares: SoFiA's false detections.

is 34 Jy km s^{-1}, while semi-automated method returns a minimum integrated flux of 40 Jy km s^{-1} and $S_{\rm int} = 0.84$ Jy km s^{-1}.

The two galaxy identification methods have a good overlap between. A large fraction of the sources that don't overlap are fainter detections below the sensitivity limit. A broader discussion is given in section 5.

4.3. SoFiA *preliminary result*

SoFiA has over 50 parameters that have to be set before it can be run successfully. For a quick look at the data, the default settings can yield reasonable results, but aiming for a more reliable and complete search fine tuning is required. There are at least 15 unique parameters for which their combinations lead to immediate differences in the total number of identified galaxy candidates. To get the most optimal results in terms of reliability and completeness, we extracted two subcubes containing a bright and a faint galaxy from the cube. Where reliability is defined as the ratio of True Positives (TP) over total number of detections (false positives + true positives), and completeness as the total number of TP over total number of all sources in the cube (both detected and undetected). One parameter file was tuned to identify the faint sources, a second one tuned SoFiA to detect the bright sources. The two parameter files were then merged into one file which in principle should then detect from faint narrow-linewidths sources to bright and wider-linewidths ones. We used the S+C source-finder with a flux threshold of 3σ (i.e. 1.20 mJy). All detections with reliability greater than 95% were accepted as positive candidates. The merged parameter file was run in four of the subcubes making up the WSRT PP HI data cube. Figure 3 presents the galaxy candidates obtained by running SoFiA. To get an idea of the performance of SoFiA, we plot the preliminary results alongside the visual results (see Figure 1). With SoFiA we identified 67 galaxy candidates, 56.7% have visual cross-matches (blue stars). But there are 43.3% without counterparts (magenta stars). The magenta stars enclosed by open black squares indicate galaxies that are consistent with noise or are found on the edges of the cube (i.e. $V_{\rm sys} \geq 16\,000$ km s^{-1}) where the noise is relatively high and a detection's reliability are compromised. Green stars show candidates within the searched fields that are not yet identified with SoFiA.

5. Discussion and conclusion

We have shown that both the visual and semi-automated galaxy identification methods extract a similar number of galaxies (194 and 211, respectively). Unlike the visual method, the semi-automated method was applied on smoothed cubes. Out of 194 visually identified galaxies, only 9.3% have no semi-automated cross-matches, compared to 17.6% of semi-automated. This means that the semi-automated has found

more sources than the Visual method. All the sources without cross-matches will be further assessed for their likelihood of being genuine and if so, why the respective methods were unsuccessful in uncovering them.

We compiled a parameter file (for running SoFiA) that in principle should return more than 80% of the galaxies identified through the visual method but so far we have managed 56.7%. In order to achieve higher completeness, further fine-tuning of the parameter file are currently underway. To quantify reliability of each detection, careful visual examination of all sources without counterparts is necessary. This will allow us to do a comprehensive analysis of all three methods; give feedback to the SoFiA developers on where possible optimization can be made to result to a more complete and reliable source catalogue and advise SoFiA uses on which combination of parameters to fine tune and under which conditions.

Acknowledgements

This work is based upon a research supported by the South African National Research Foundation (NRF), the Department of Science and Technology (DST) and National Astrophysics and Space Science Program (NASSP). We thank Prof. Patricia A. Henning for her contribution to this work. TG is immensely grateful to Dr. Paolo Serra, Dr. Gyula Jozsa and Dr. Khaled Said for useful discussions. We also thank the developers of SoFiA for having an open ear and showing interest on the outcome of this project.

References

[1] Holwerda B W, Blyth S L, and Baker A J 2011 *Proceedings of the International Astronomical Union* **7** 496499
[2] Serra P 2011, *The MeerKAT Fornax Survey Fornax, Virgo, Coma et al., Stellar Systems in High Density Environments* page 49
[3] Verheijen M, Oosterloo T, Heald G and Van Cappellen W 2009 *Panoramic radio astronomy: wide-field 1-2*
[4] Duffy, A R Meyer, M J Staveley-Smith, L Bernyk, M Croton, D J Koribalski, B S Gerstmann, D and Westerlund, S , 2012 , *Monthly Notices of the Royal Astronomical Society* , **426** , 3385-02
[5] Jurek R 2012 *Publications of the Astronomical Society of Australia* **29** 251-61
[6] Flöer L and Winkel B 2012 *Publications of the Astronomical Society of Australia* **29** 244-50
[7] Serra P, Jurek R and Flöer L 2012 *Publications of the Astronomical Society of Australia* **29** 296-00
[8] Whiting M T 2012 Monthly Notices of the Royal Astronomical Society **421** 3242-56
[9] Popping A, Jurek R, Westmeier T, Serra P, Flöer L, Meyer M and Koribalski B 2012 *Publications of the Astronomical Society of Australia* **29** 318-339
[10] Westmeier T, Popping A and Serra P 2012 *Publications of the Astronomical Society of Australia* **29** 276-95
[11] Serra P et. al 2015 *Monthly Notices of the Royal Astronomical Society* **448** 1922-29
Westmeier T Jurek R Obreschkow D Koribalski B S and Staveley-Smith L 2013 *Monthly Notices of the Royal Astronomical Society* **438** 1176-90
[12] Gooch R 2006 *KARMA Users Manual, ATNF* reference link
[13] Sault R and Killeen N 1996 *Multichannel image reconstruction image analysis and display (miriad) users guide, ATNF* reference-link
[14] Ramatsoku M Verheijen M Kraan-Korteweg R JÂťozsa G Schröder A Jarrett T Elson E van Driel W de Blok W and Henning P 2016 *Monthly Notices of the Royal Astronomical Society* **460** 923-41
[15] Westmeier T and Jurek R and Obreschkow D and Koribalski B S and Staveley-Smith L 2014 *The busy function: a new analytic function for describing the integrated 21-cm spectral profile of galaxies, mnras* **438** 1176-90