Automated Source Extraction for the Next Generation of Neutral Hydrogen Surveys

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Lars Flöer
aus
Bonn-Beuel

Bonn, März 2015

Dieser Forschungsbericht wurde als Dissertation von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem Hochschulschriftenserver der ULB Bonn `http://hss.ulb.uni-bonn.de/diss_online` elektronisch publiziert.

1. Gutachter:	PD Dr. Jürgen Kerp
2. Gutachter:	Prof. Dr. Pavel Kroupa
Tag der Promotion:	28. September 2015
Erscheinungsjahr:	2015

Contents

Summary

This thesis is a first step to develop the necessary tools to automatically extract and parameterize sources from future H I surveys with the Australia SKA Pathfinder (ASKAP, Johnston et al. 2008), the upgraded Westerbork Synthesis Radio Telescope (WSRT/Apertif, Oosterloo et al. 2010), and Square Kilometre Array (SKA). The current approach to large-scale H I surveys, that is, automated source finding followed by manual classification and parametrization, is no longer feasible in light of the data volumes expected for future surveys. We use data from the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011) to develop and test a completely automated source extraction pipeline for extragalactic H I surveys.

We apply a 2D-1D wavelet de-noising technique to H I data and show that it is well adapted to the typical shapes of sources encountered in H I surveys. This technique allows to reliably extract sources even from data containing defects commonly encountered in single-dish H I surveys.

Automating the task of false-positive rejection requires reliable parameters for all source candidates generated by the source-finding step. For this purpose, we develop a reliable, automated parametrization pipeline that combines time-tested algorithms with new approaches to baseline estimation, spectral filtering, and mask optimization. The accuracy of the algorithms is tested by performing extensive simulations. By comparison with the uncertainty estimates from the H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001) we show that our automated pipeline gives equal or better accuracy than manual parametrization.

We implement the task of source classification using artificial neural networks using the automatically determined parameters of the source candidates as inputs. The viability of this approach is verified on a training data set comprised of parameters measured from simulated sources and false positives extracted from real EBHIS data. Since the number of true positives from real data is small compared to the number of false positives, we explore various methods of training artificial neural networks from imbalanced data sets. We show that the artificial neural networks trained in this way do not achieve sufficient completeness and reliability when applied to the source candidates detected from the extragalactic EBHIS survey.

We use the trained artificial neural networks in a semi-supervised manner to compile the first extragalactic EBHIS source catalog. The use of artificial neural networks reduces the number of source candidates that require manual inspection by more than an order of magnitude. We compare the results from EBHIS to HIPASS and show that the number of sources in the compiled catalog is approximately half of the sources expected. The main reason for this detection inefficiency is identified to be misclassification by the artificial neural networks. This is traced back to the limited training data set, which does not cover the parameter space of real detections sufficiently, and the similarity of true and false positives in the parameter space spanned by the measured parameters.

We conclude that, while our automated source finding and parametrization algorithms perform satisfactorily, the classification of sources is the most challenging task for future H I surveys. Classification based on the measured source parameters does not provide sufficient discriminatory power and we propose to explore methods based on machine vision which learns features of real sources from the data directly.

CHAPTER 1

Introduction

The next generation of radio-astronomical facilities, the Square Kilometre Array (SKA) and its pathfinder experiments the South African SKA Pathfinder (MeerKAT, Booth et al. 2009), the Australia SKA Pathfinder (ASKAP, Johnston et al. 2008), and the upgraded Westerbork Synthesis Radio Telescope (WSRT/Apertif, Oosterloo et al. 2010), vastly increase the observational capabilities across a wide variety of scientific fields (Taylor 2013). Among the many fields the SKA will revolutionize are galactic and extragalactic surveys of neutral atomic hydrogen (H I) through the 21 cm line. The completed SKA will be able to conduct H I surveys detecting about a billion sources out to a redshift of about two (Yahya et al. 2014). But already the SKA pathfinders ASKAP and WSRT/Apertif will carry out all-sky surveys with unprecedented angular resolution and sensitivity through the use of focal plane arrays.

Focal-plane arrays increase the survey speed of radio-interferometric observatories by more than an order magnitude, allowing them to perform deep, all-sky surveys (Verheijen et al. 2008). This is a major revolution over the past approach to H I surveys, which were typically either deep observations of a sample of galaxies at high angular resolution with interferometric arrays or large-area, shallow surveys at low angular resolution with single-dish telescopes. Because of these limitations, the data volume from both kinds of surveys allow a manual analysis of the reduced data. This is no longer possible for the H I surveys with the next generation of radio observatories. These surveys combine a large survey area with high angular resolution. The data cubes from these surveys will be about 800 GB in size for each 30 square degrees of sky coverage (Whiting & Humphreys 2012). For comparison, a data cube covering 25 square degrees from the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011), the H I survey this thesis is based on, is about 300 MB in size. This requires to re-think the way in which data from H I surveys is analyzed.

In contrast to the optical community, where large-scale projects like the Sloan Digital Sky Survey (SDSS, York et al. 2000) already require automatic data analysis, the H I community only recently started concerted efforts on fully automated data processing (for a recent overview, see Koribalski 2012). While all large-area H I surveys employ some form of automated source-finding, the verification and parametrization of detected source candidates is performed manually. Considering the data volumes from the next generation of H I surveys, all tasks concerning source extraction need to be automated.

While many techniques from optical astronomy can be applied to H I data, the nature of the H I line and radio-astronomical observations give rise to additional complications for automated survey analysis. The 21 cm line is a particularly faint emission line and the signal from an individual galaxy can be spread out over many spectral channels, making the detection more difficult. Additionally, the frequencies at which red-shifted H I emission from galaxies is observed, are contaminated by radio-frequency interference (RFI). This interference gives rise to a large number of false positives during automated source finding and can mimic the appearance of astrophysical sources, requiring careful

inspection of every source candidate.

This thesis is a first effort to develop a completely automated pipeline for the analysis of data products from large-scale H I surveys. In Chapters 2, 3, and 4, we develop the methods required to perform fully automated source-finding, parametrization, and classification for H I surveys. The primary goal is complete independence from human supervision during data processing. While current source-finding algorithms already fulfill this criterion, there are no standard ways of reliable, automatic parametrization and candidate classification, necessary to reject false positives. We use EBHIS as a testbed to gain experience on which methods work and where the effort of future survey processing pipelines should be focused.

In this Chapter, we review the case for extragalactic H I science and review past, current, and future H I surveys. We also introduce EBHIS, a novel H I survey in the northern hemisphere conducted with the Effelsberg 100 m radio telescope, as it serves as the data base for this thesis.

1.1 Extragalactic H I Science

As the most abundant element in the universe, neutral hydrogen is the prime tracer for neutral gas. In extragalactic astrophysics, the two main aspects investigated with neutral hydrogen are the gas reservoir and the dynamics of galaxies. Neutral hydrogen can be observed out to many optical radii for most galaxies, and therefore traces the dynamics in the outskirts of galaxies (Bosma 1981). Since it is also the largest reservoir of material able to form stars, its accretion and recycling in galaxies is an important ingredient to understand the sustained formation of stars observed in galaxies (Putman et al. 2012).

These two science cases are usually addressed by two different kinds of H I surveys: One approach are high-resolution observations of individual or a small sample nearby galaxies, obtained with radio interferometers like the Jansky Very Large Array (JVLA), Westerbork Synthesis Radio Telescope (WSRT), and Australia Telescope Compact Array (ATCA). These observations allow detailed modeling of the gas dynamics and comparison with multi-wavelength data sets. This joint investigation allows to detect traces of past interaction and faint star-forming regions (e.g., de Blok et al. 2014).

The other approach to these scientific questions involves low-resolution, single-dish surveys having sample sizes between 10^3 and 10^4 galaxies. These surveys are typically conducted in a blind fashion although there are examples of large studies that preselect their targets based on optical catalogs (e.g., Catinella et al. 2010). Due to the larger number of sources but less detailed parametrization, these surveys are used to perform statistical analyses. An important statistical tool is the H I mass function (HIMF), which measures the number density of H I bearing sources as a function of H I mass. It allows to estimate the cosmological H I mass density, $\Omega_{\mathrm{H\,I}}$, which is important to understand the evolution of galaxies from redshifts $\gg 0$ until today (Lanzetta et al. 1995). As the gas fraction in galaxies increases with decreasing stellar mass, H I surveys are especially suited to study the distribution of low-mass systems (Geha et al. 2006). The faint-end slope of the HIMF is an important benchmark for simulations and semi-analytical modeling of galaxy evolution, as the gas content of galaxies depends on baryonic feedback processes (e.g., Baugh et al. 2005; Bower et al. 2006; De Lucia & Blaizot 2007; Obreschkow et al. 2009).

Large samples of the H I content of galaxies also allow to deduce the space density of galaxies with a given projected circular velocity (Zwaan et al. 2010; Papastergis et al. 2011). The observed distribution of projected circular velocities allows a direct comparison to predictions from dark matter (DM) simulations, as the circular velocities of disk galaxies are believed to be dominated by the DM halo of the galaxies. The studies conducted by Zwaan et al. (2010) and Papastergis et al. (2011) both find that the galaxies modeled by DM simulations over-predict the abundance of slow-rotating galaxies.

Parameter	EBHIS	HIPASS	ALFALFA
Observatory	Effelsberg	Parkes	Arecibo
Dish Diameter	100 m	64 m	300 m
Coverage	$\delta > -5°$	$\delta < 25°$	$0° < \delta < 36°$
Survey Area	22 424 deg^2	29 343 deg^2	7074 deg^2
Angular Resolution	10.8′	15.5′	3.5′
Spectral Resolution	10.24 km s^{-1}	18.0 km s^{-1}	5.4 km s^{-1}
Spectral Coverage	$cz < 18\,000$ km s^{-1}	$cz < 12\,700$ km s^{-1}	$cz < 18\,000$ km s^{-1}
Noise Level	23 mJy beam^{-1}	13 mJy beam^{-1}	2.4 mJy beam^{-1}

Table 1.1: Parameters of blind, extragalactic, single-dish H I surveys. ALFALFA is restricted to two separate areas between Right Ascension 7^h30^m to 16^h30^m and 22^h to 3^h hours. The spectral resolution for EBHIS and HIPASS is the effective resolution after binning or Hanning-smoothing. The channel separations for both surveys are 1.28 km s^{-1} and 13.3 km s^{-1}, respectively.

In combination with optical or infrared observations, the H I profile width can be used to determine the rotational velocity of the galaxy and allows a distance determination from the Tully-Fisher relation (Tully & Fisher 1977). The optical or infrared observations are required to correct the observed profile width for the inclination of the galaxy. This has been applied, for example, by Tully et al. (2013) to measure the large-scale flows of galaxies, groups, and clusters in the local universe.

Due to the intrinsic faintness of the 21 cm line, the fraction of galaxies detected in blind, large-area H I surveys is low. Recent studies exploit the fact, that all galaxies inside the survey volume are observed and use optically determined redshifts to shift all non-detections to a common radial velocity and average their spectra. This technique is called "stacking" and makes previously unused data from blind H I surveys scientifically valuable. Although no individual galaxies can be analyzed, stacking is used to determine $\Omega_{\rm H\,I}$ at redshifts where galaxies can not be detected individually (Delhaize et al. 2013). Another approach is to derive scaling relations by splitting up the non-detections in bins of, for example, optical color and investigate their average H I mass (Fabello et al. 2011a,b). Once calibrated, these scaling relations can be used to predict the H I content of galaxies based on their color in the optical or ultra-violet regime.

1.2 Extragalactic H I Surveys

1.2.1 Past and Current Surveys

To date there are three extragalactic, single-dish H I surveys that cover a significant fraction of the sky: The H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001), the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005) and the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011). Because of the different observatories they are conducted at, they vary in angular resolution, sensitivity, and coverage. We summarize the most important survey parameters in Table 1.1. The H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001) is the first large-area H I survey that covered a large enough volume to accurately determine the HIMF and $\Omega_{\rm H\,I}$. Originally, there were plans to conduct a complimentary survey on the northern hemisphere, the H I Jodrell All Sky Survey (HIJASS, Lang et al. 2003), but it was never completed. While covering a significantly smaller portion of the sky, the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005) is a much more sensitive survey than HIPASS because of the large collecting area of the Arecibo observatory. In only 40 % of the final survey volume,

Parameter	WALLABY	WNSHS
Observatory	ASKAP	WSRT/Apertif
Coverage	$\delta < 30°$	$\delta > 25°$
Survey Area	$30\,940\,\mathrm{deg}^2$	$11\,262\,\mathrm{deg}^2$
Angular Resolution	$30''$	
Spectral Resolution	$3.9\,\mathrm{km\,s^{-1}}$	
Spectral Coverage	$-2000\,\mathrm{km\,s^{-1}} < cz < 77\,000\,\mathrm{km\,s^{-1}}$	
Noise Level	$1.6\,\mathrm{mJy\,beam^{-1}}$	$1.5\,\mathrm{mJy\,beam^{-1}}$

Table 1.2: Key parameters of planned, all-sky, H I surveys with SKA precursors (Koribalski 2012). Note that WNSHS is a *proposed* survey project for WSRT/Apertif. The survey parameters are subject to change, depending on the final configuration of the observatories.

they detect nearly three times as many sources as HIPASS (Haynes et al. 2011). ALFALFA constrains the faint end of the HIMF to higher accuracy than HIPASS but the results from both surveys differ significantly on the high-mass end of the HIMF. Having started in 2008, EBHIS is the most recent among the three surveys. As it forms the basis of this thesis it is introduced in more detail in Sec. 1.3.

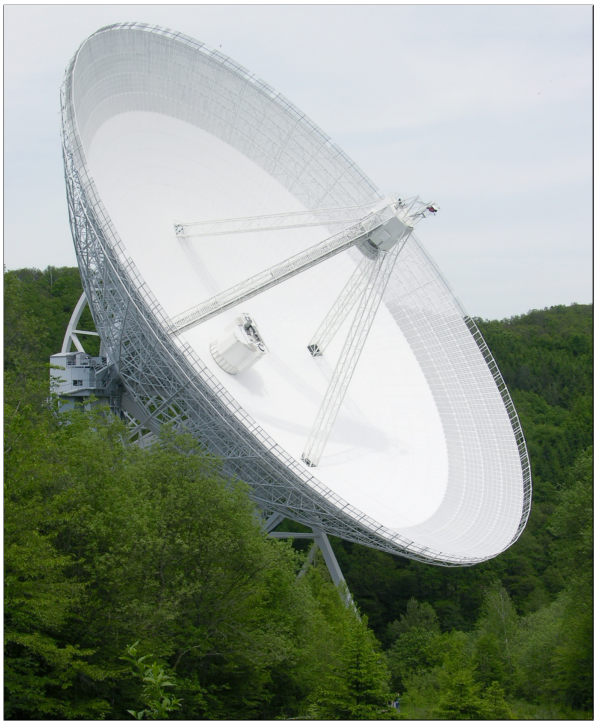
1.2.2 Future Surveys

With the construction of the first SKA pathfinder experiments, a number of new extragalactic H I surveys are planned. These surveys have the spatial coverage previously only accessible to single-dish surveys at the angular resolution of an interferometer. We summarize the most important parameters of the surveys in Table 1.2. Together, the WALLABY H I All-Sky Survey (WALLABY, Koribalski 2012) and the Westerbork Northern Sky H I Survey (WNSHS) will observe the complete sky. Both surveys cover 300 MHz with 16 384 spectral channels, allowing them to survey both galactic and extragalactic H I, although the low spectral resolution is not sufficient to perform detailed investigation of the H I content of the Milky Way; the planned GASKAP survey is more suited for this task (Dickey et al. 2013). It is expected to detect around 800 000 sources in both surveys combined of which 700 000 are expected to be resolved by less than three resolution elements (Duffy et al. 2012). While there will be around 8000 well resolved sources that allow detailed kinematic analysis as described above, the parametrization of marginally resolved sources still plays a major role.

While the all-sky surveys with ASKAP and WSRT/Apertif investigate the H I content of the present-day universe, there are multiple survey projects focusing on the evolution of the H I content of galaxies from redshift of about unity until today. The DINGO survey (Meyer 2009) uses ASKAP to observe a very deep $60\,\mathrm{deg}^2$ field and a shallower $150\,\mathrm{deg}^2$ sized field. The survey is designed to investigate the evolution of the H I content of galaxies for redshifts < 0.4. The LADUMA survey (Holwerda et al. 2012) uses MeerKAT to go even deeper than DINGO and will observe a single pointing for 5000 hours. Using stacking methods this survey is expected to make the first detections of H I in emission for redshifts of about unity.

1.3 The Effelsberg-Bonn H I Survey

The Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011) is an all-sky H I survey conducted with the Effelsberg 100 m radio observatory. Using modern spectrometers based on field-programmable gate arrays (FPGAs) the survey has both sufficient bandwidth and spectral resolution to conduct an galactic



Source: Wikimedia Commons, Dr. Schorsch

Figure 1.1: The Effelsberg 100 m dish.

Photo courtesy of K. Grypstra

Figure 1.2: The seven-feed receiver in the primary focus of the Effelsberg 100 m dish.

and extragalactic H I survey at the same time. This makes EBHIS the northern counterpart to both the Parkes Galactic All-Sky Survey (GASS, McClure-Griffiths et al. 2009; Kalberla et al. 2010) and HIPASS in one observing campaign. The main parameters of the survey are summarized in Table 1.1. The values stated in the table are given for the first coverage. A second coverage of the sky north of $\delta = 30°$ is currently observed to increase the sensitivity by 40 % and reduce the artifacts introduced by the scanning pattern (see below). In this section, we briefly describe the data acquisition, data reduction, and typical data products used in this thesis.

1.3.1 Observations and Data Reduction

First observations for EBHIS started in 2008 and the first coverage was completed in 2013. The observations for EBHIS are conducted using a seven-feed receiver in the primary focus of the Effelsberg 100 m dish, shown in Figures 1.1 and 1.2. Each feed delivers two orthogonal polarizations. The central feed is sensitive to circularly polarized radiation whereas the six outer feeds are sensitive to linearly polarized radiation. The front-end filter of the receiver has a bandwidth of 300 MHz from which 100 MHz are selected at the intermediate frequency (IF) stage and approximately centered on a sky frequency of 1380 MHz. The exact frequency changes depending on the local standard-of-rest (LSR) correction applied during observations [1]. Further technical details of the receiver are described in Keller et al. (2006).

The back-end for EBHIS consists of 14 identical spectrometers based on FPGAs (Stanko et al. 2005;

[1] Early observations of EBHIS applied the LSR correction on the fly. In later observations, the LSR correction is applied to the data after the fact.

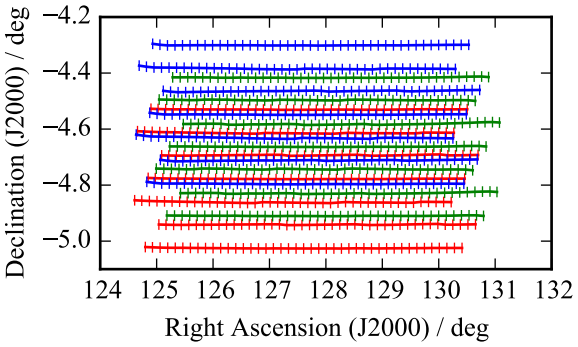


Figure 1.3: On-sky pattern of the seven-feed receiver for three subscans of an EBHIS observation. Each color indicates a different subscan. The vertical lines on each of the tracks indicates the location at which data is written to disk.

Klein et al. 2008), each providing 16 384 channels. The output from a single spectrometer is referred to as *baseband*. FPGA spectrometers do not suffer from Gibbs-ringing (Gibbs 1899) and allow more spectral channels than digital auto-correlators used by, for example, GASS and HIPASS. This setup allows EBHIS to conduct both an galactic H I survey with a spectral resolution of 1.45 km s^{-1} and cover extragalactic emission due to the 100 MHz bandwidth.

The observations for EBHIS are carried out in on-the-fly (OTF) mode by scanning individual 5 deg by 5 deg fields. Each field is referred to as a *scan* and individual scan-lines during an observation are referred to as *subscan*. In Fig. 1.3 we show the on-sky pattern for each of the seven feeds during the first three subscans of an observation. The seven-feed receiver is rotated such that the maximum distance between scan-lines still guarantees Nyquist-sampling of the telescope beam and that the sampling is as homogeneous as possible. To keep this pattern in Right Ascension and Declination homogeneous during observations, the receiver rotation is constantly adjusted to compensate for the changing parallactic angle. Data are written to disk every 500 ms for each of the 14 spectrometers. For calibration purposes, EBHIS uses both frequency switching and a noise diode. The combination of these methods leads to four separate switching phases.

Data calibration, flagging and reduction for EBHIS are described in Winkel et al. (2010). We briefly summarize the process from raw data to our intermediate storage format here. Each observation consists of approximately 50 000 individual spectra. We determine the gain curve, that is, the amplification of the receiving system as a function of frequency, of the receiver system for each of the basebands using the method described in Section 5 of Winkel et al. (2012b). To convert spectrometer counts into brightness temperatures, we use the S7 IAU standard calibrator (Kalberla et al. 1982) to perform absolute calibration for $v_{\mathrm{LSR}} = 0\,\mathrm{km\,s^{-1}}$ and use the determined gain curves to perform relative calibration for the whole baseband. The S7 position is observed regularly to track changes of the characteristics of the receiver system.

Once the data are calibrated, we use the algorithm described in Flöer et al. (2010) to recognize RFI in the data. The two main types of RFI encountered in EBHIS data are short, broadband events likely caused by aircraft RADAR and near-constant, narrow-band lines, affecting single channels. These narrow lines are likely caused by electronic equipment, located either on-site or remote. Data affected by broadband and variable, narrow-band RFI is excluded from further processing. Where the amplitude of the narrow-band RFI is stable enough, we try to subtract the RFI amplitude from the data. This process often leads to residual RFI present in the final data, but allows to re-use a large fraction of the affected data.

To separate the line emission from the continuum, we estimate a baseline in the time-frequency domain of each subscan, baseband, and switching phase. We first subtract the median from each dump to remove the elevation dependent baseline level and strong continuum sources. For the extragalactic

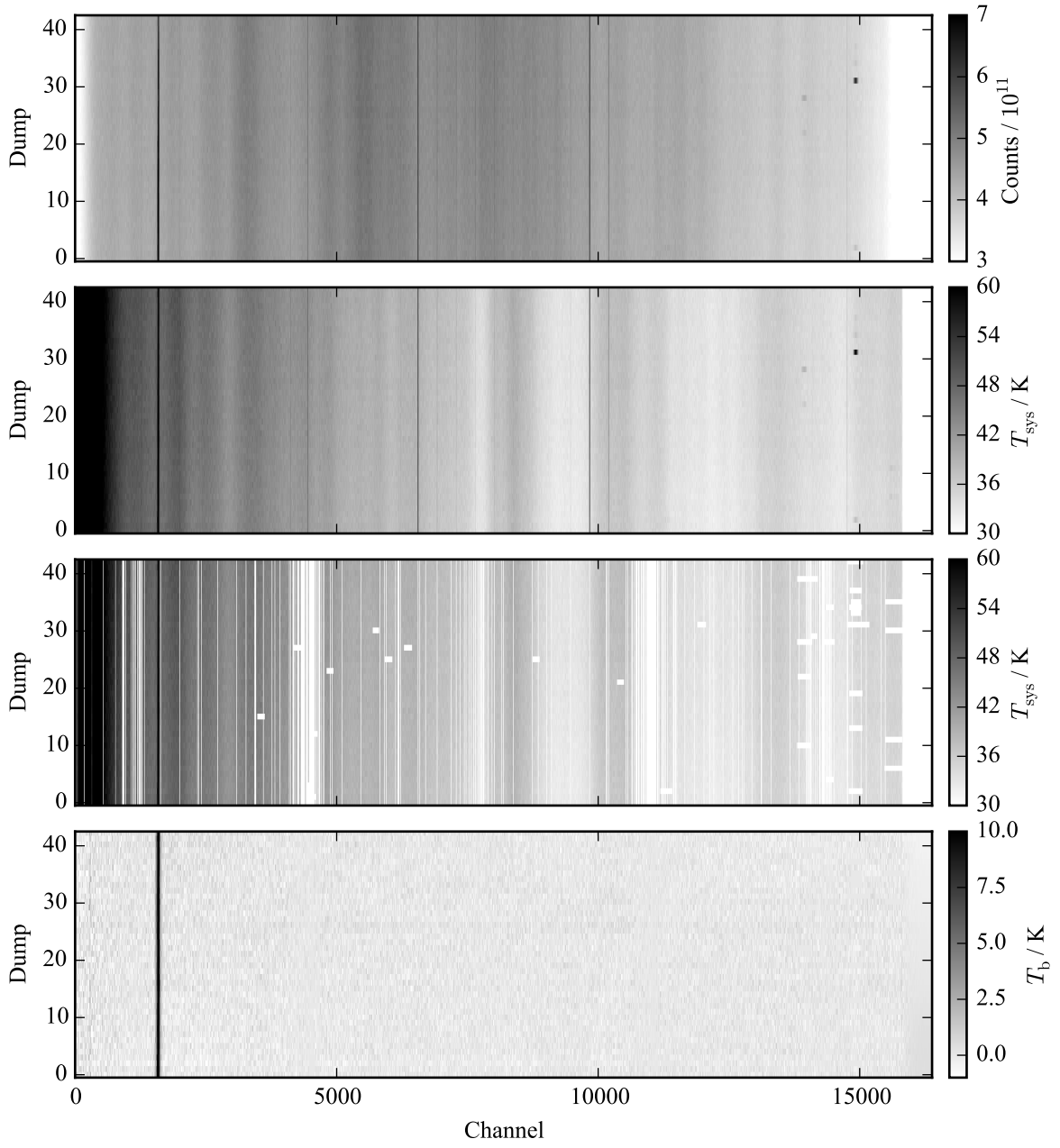


Figure 1.4: The various processing stages of EBHIS raw data. Each panel shows the same data from a single subscan, baseband, and switching phase. The topmost panel shows the raw data as written to disk from the spectrometer. In the next panel, the data are corrected for the system gain and calibrated against the S7 IAU standard calibrator. The panel below shows the flags found by the RFI flagging software. The last panel shows the baseline-subtracted data. The H I emission from the Milky Way is clearly visible around channel 1500. Note that most extragalactic sources are too faint to be visible in a single subscan.

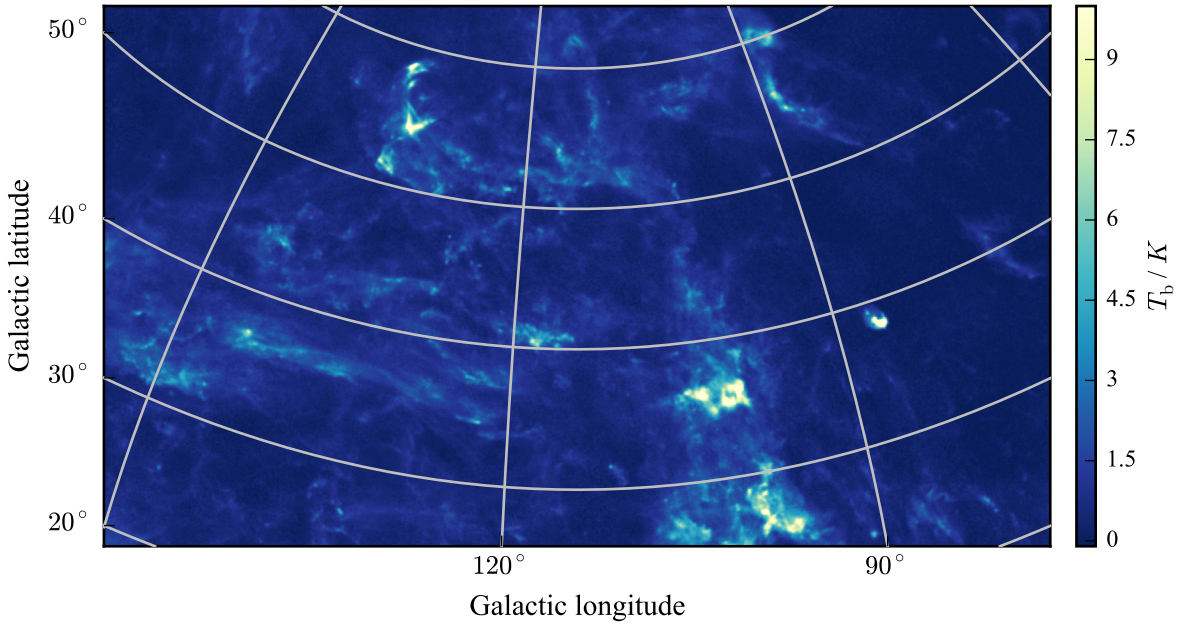


Figure 1.5: H I emission towards the high-velocity cloud complex C. The image shows a single channel of EBHIS data at a radial velocity of $-43\,\mathrm{km\,s^{-1}}$. The data are an example of intermediate velocity gas at high galactic latitudes.

survey, we estimate the spectral baseline by calculating the median spectrum over a full subscan and subtract it from each dump. Since this process makes the extended emission of the H I in the Milky Way unusable, the galactic survey uses a two-dimensional polynomial fit to estimate the baseline instead. We also use the spectral baseline as an estimate of the system temperature to determine the relative weights when combining separate EBHIS scans. The weights also reflect the decreased sensitivity due to data removed by the RFI flagging software.

In Fig. 1.4 we show the individual steps a single subscan, baseband, and phase go through to arrive at the final calibrated data. With the exception of the RFI flagging, all calibration steps are applied to the data by the program `bgrid` (Winkel 2009; Winkel et al. 2010). `bgrid` also applies the LSR correction and shifts the spectra to a common rest frame. After calibration, the data are interpolated onto a regular grid by convolving the individual spectra with a Gaussian kernel (see Winkel et al. 2012a, for a detailed description).

Data from H I surveys is usually visualized as a three dimensional data cube. The celestial position is represented by the first two axes and the spectral information is represented on the third axis. Each value in a data cube is the H I flux density at a given position and Doppler shift. To facilitate the quick creation of custom data cubes, the individual EBHIS scans are first interpolated onto a common HEALPix grid (Górski et al. 2005). This alleviates the need to re-perform the full data reduction process for each data cube and allows to construct data cubes that span multiple scans. We interpolate the calibrated spectra onto the HEALPix grid using a kernel full width at half maximum (FWHM) of 3.8′. When creating the final data cubes, the data are again convolved with the same kernel, yielding an effective kernel size of $\sqrt{2} \times 3.8' \approx 5.4'$ and an angular resolution of 10.5′. The process of interpolating irregularly spaced data from OTF measurements onto a regular grid is reviewed in Mangum et al. (2007).

In Fig. 1.5 we show an example of galactic H I emission as observed by EBHIS. The figure shows a large field towards the high-velocity cloud complex C (Wakker & van Woerden 1991) and is an example

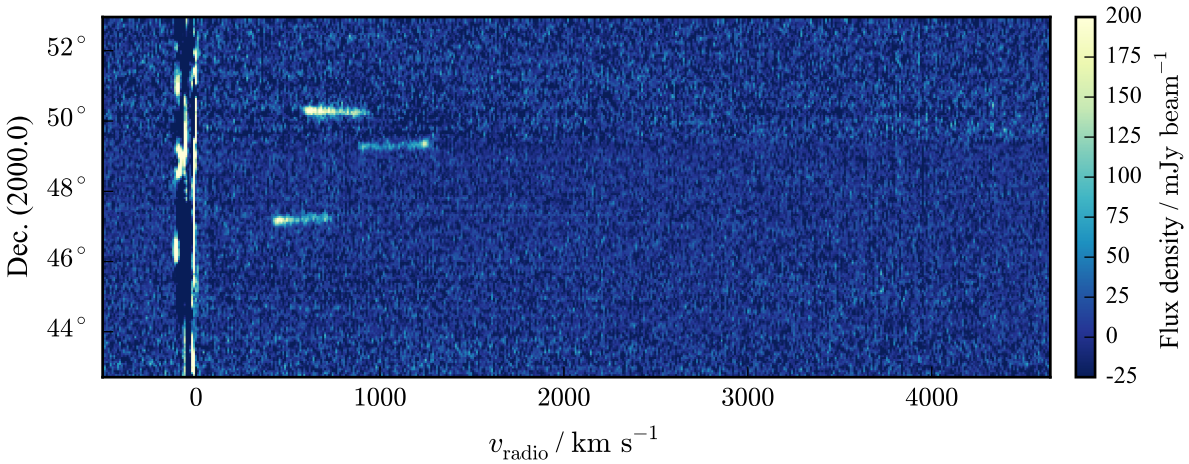


Figure 1.6: Position-velocity diagram of a 10°×10° degree field showing extragalactic EBHIS data. The data cube is cut at Right Ascension $12^{\mathrm{h}}2^{\mathrm{m}}52^{\mathrm{s}}$. The fluctuations at $v_{\mathrm{radio}} = 0\,\mathrm{km\,s^{-1}}$ are the left-over galactic H I emission, which is filtered out for extragalactic applications. There are three galaxies visible as bright, horizontal bars.

of a data cube created from our intermediate data storage. It contains data from 207 individual observations. In Fig. 1.6 we show a position-velocity diagram of extragalactic EBHIS data. The position-velocity diagram shows a 10° long slice at Right Ascension $12^{\mathrm{h}}2^{\mathrm{m}}52^{\mathrm{s}}$ containing the H I profiles of three galaxies. The shape of these profiles is caused by the spatially unresolved rotation of a gas disk in the galaxies. With a few exceptions, most galaxies appear in this form in EBHIS.

CHAPTER 2

Source Finding

The results from this Chapter are published in part in

- Flöer, L. & Winkel, B. 2012, PASA, 29, 244
- Flöer, L., Winkel, B., & Kerp, J. 2014, A&A, 569, A101

The literature has no strict definition of the term *source finding*. In the context of H I surveys, source finding usually describes one or more of the following tasks:

1. Extract regions of interest from the data based on some criterion.
2. Measure parameters of the detections using interactive or automated methods.
3. Determine whether the detection is of astrophysical origin or caused by some defect in the data.

For the purpose of this chapter, we define the term source finding to only refer to the first task: the extraction of regions from the data that are not noise-like and could therefore be of astrophysical origin. Automated approaches to parametrization and classification are explored in Chapters 3 and 4, respectively. In Sec. 2.1 we discuss the state of the art in source finding for spectroscopic surveys of H I. To motivate our choice of source finding algorithm, we discuss the challenges for automated source finding algorithms for spectroscopic data in Sec. 2.2. In Sec. 2.3 we introduce source finding by 2D-1D wavelet de-noising and highlight key advantages over other source finding methods. Before we conclude the chapter, we describe how we apply the developed algorithm to perform source finding for the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011) in Sec. 2.4.

2.1 State of the Art

With the H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001), data volumes of H I surveys became sufficiently large to make automated source finding a necessity. To generate a list of candidates, they use a combination of two source finders, MULTIFIND and TOPHAT (Meyer et al. 2004), both based on matched filtering (North 1963) and implemented using MIRIAD tasks (Sault et al. 1995). For MULTIFIND, each plane in the data cube is searched separately and values in excess of four times the noise level are accepted as candidates. The data are subsequently smoothed along the spectral domain to increase the sensitivity towards wide line profiles. TOPHAT searches the data by performing a matched filtering approach by using box-filters to be sensitive to wide spectral profiles. To reduce the impact of residual baselines and solar ripple, the data are pre-filtered with a median filter adapted to the width of the current box filter.

For the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005), Saintonge (2007) use templates derived from Hermite polynomials as templates for matched filtering. They argue that due to the two-peaked shape of the templates, they are particularly well suited to match the double-horned profile caused by the rotating H I disk in galaxies. Although these templates likely achieve a better sensitivity than box templates, they still do not resemble the true shape of double-horned profiles very well. This is mostly due to the smooth rise and fall of the template flanks, as can be seen from the plots shown in Saintonge (2007). Flöer et al. (2014) furthermore show that every matched filtering approach will still pick up residual baselines.

Wolfinger et al. (2013) use the `Duchamp` source finding application (Whiting 2012) for data from the H I Jodrell All Sky Survey (HIJASS, Lang et al. 2003). `Duchamp` uses isotropic 3D wavelet de-noising to detect possible H I signals in data cubes. Wavelet de-noising, which is discussed in more detail in Sec. 2.3, can be thought of as adaptive filtering (Starck & Bijaoui 1994), i.e., the amount of smoothing applied to the data is determined by the data itself.

2.2 Real Data Challenges

Apart from the astronomical signal, real H I data cubes are contaminated with unwanted signal of various origins. The approaches to source finding introduced above are mostly developed with clean data in mind, i.e., a combination of Gaussian noise and source signal only. This leads to a large number of false positives if no further filtering is applied to the candidates generated by these algorithms: For HIPASS, Meyer et al. (2004) generate 142 276 candidates of which only 4315 are classified as real sources. The remainder are either caused by inhomogeneous noise, residual baselines, radio-recombination lines (RRLs) or radio-frequency interference (RFI).

RFI introduces significant signal that can manifest itself in a variety of ways. In HIPASS, RFI is mostly present as intermittent, single-channel events or the broad signal caused by the GPS L3 beacon (Meyer et al. 2004). EBHIS data are mostly contaminated by narrow-band signals with a frequency width of less than 10 kHz but largely constant in amplitude over the course of an observation. These signals often exceed the brightness temperature of the observed H I emission by several orders of magnitude. This leads to the Gibbs phenomenon (Gibbs 1899) when performing spectroscopy with digital auto-correlator systems, as the RFI introduces a sharp spike in the data. For modern spectrometers based on field-programmable gate arrays (FPGAs) this phenomenon strongly suppressed, due to the much higher dynamic range. Nonetheless, in both cases RFI introduces signal in the data that is detected by the approaches to source-finding described above.

In addition to RFI, continuum emission from the Sun and other strong sources causes standing waves in the telescope support structure. These enter the data as a spectrally varying baseline. Since they are mainly caused by the Sun, the phase of these standing waves depends on the relative position of the Sun with respect to the observing direction. When observing in on-the-fly (OTF) mode, this leads to a constantly varying baseline level, as the observing direction is changed continuously. There are various approaches to remove these baseline variations through modeling (Briggs et al. 1997; Barnes et al. 2005). Despite the modeling efforts, data cubes from single-dish surveys still exhibit residual baselines. When using matched filtering or 3D wavelet de-noising, these are detected as significant and will generate false positives.

Another area of concern when working with real data is the homogeneity of the noise in the data. As mentioned above, algorithms like matched filtering and wavelet de-noising implicitly assume the noise level is constant across the data set. This is rarely the case in practice: Inhomogeneous sampling during OTF observations or flagging of bad data lead to a locally varying noise level. In addition to

these variations caused by observational and data reduction processes, ground and celestial continuum emission increase the system temperature and therefore the noise level. Since the elevation of the telescope changes during the course on an OTF observation, the contribution of ground emission varies. In addition to these external effects, the sensitivity of the receiver itself varies as a function of frequency. These effects cause a spatially and spectrally varying noise level and need to be taken care of when performing source finding in real data.

To reduce the impact of the aforementioned effects, we introduce source finding by 2D-1D wavelet de-noising in the following section. This wavelet de-noising scheme addresses all of the mentioned effects, with the exception of locally varying noise. We will address the topic of inhomogeneous noise in Sec. 2.4.

2.3 2D-1D Wavelet De-noising

The use of wavelet transforms in astrophysics has become very popular in recent years. Typical applications for wavelet-transform-based methods are morphological separation of sources in images and noise removal. The success of wavelet based methods in astrophysics is in part due to the fact that astrophysical data often contains information on different angular or spectral scales. For example, an optical image of a galaxy contains compact, bright stars as well as extended and faint emission from the bulge and spiral arms. Multi-scale methods, such as the wavelet transform, allow to investigate the different scales of an image separately. Starck & Bobin (2010) give a good overview of different applications of wavelet-based methods in various astronomical contexts.

2.3.1 The Isotropic Un-decimated Wavelet Transform

Starck et al. (2010) introduce the 2D isotropic un-decimated wavelet transform, which is a special case of discrete wavelet transforms that is particularly well adapted to astronomical images (Starck & Murtagh 1994, 2006). Because of this, and to avoid confusion with the more general un-decimated wavelet transform, they call this transform the *starlet transform*.

The *starlet transform* with J scales decomposes an image $c_0[k,l]$ into a set of wavelet coefficients $w_j[k,l]$ and a smooth approximation $c_J[k,l]$

$$c_0[k,l] = c_J[k,l] + \sum_{j=1}^{J} w_j[k,l] \quad . \tag{2.1}$$

Here the indices k, l indicate that the image is a discrete, two dimensional array of values. The coefficients $w_j[k,l]$ encode information about the variation in the signal at a given scale j, whereas the smooth approximation $c_J[k,l]$ contains the mean signal. We show a one-dimensional example of the wavelet transform in Fig. 2.1.

Calculating the wavelet coefficients at scale j of the *starlet transform* requires convolution of the data $c_j[k,l]$ to obtain $c_{j+1}[k,l]$ and taking their difference.

$$\begin{aligned} w_{j+1}[k,l] &= c_j[k,l] - c_{j+1}[k,l] \\ c_{j+1}[k,l] &= (h_j^k h_j^l \star c_j)[k,l] \end{aligned} \tag{2.2}$$

Here, the term $(h_j^k h_j^l \star c_j)$ indicates the separable convolution of c_j with the kernel h_j along each axis of the image individually. This allows to calculate the convolution in $O(2kN)$ instead of $O(k^2N)$, where

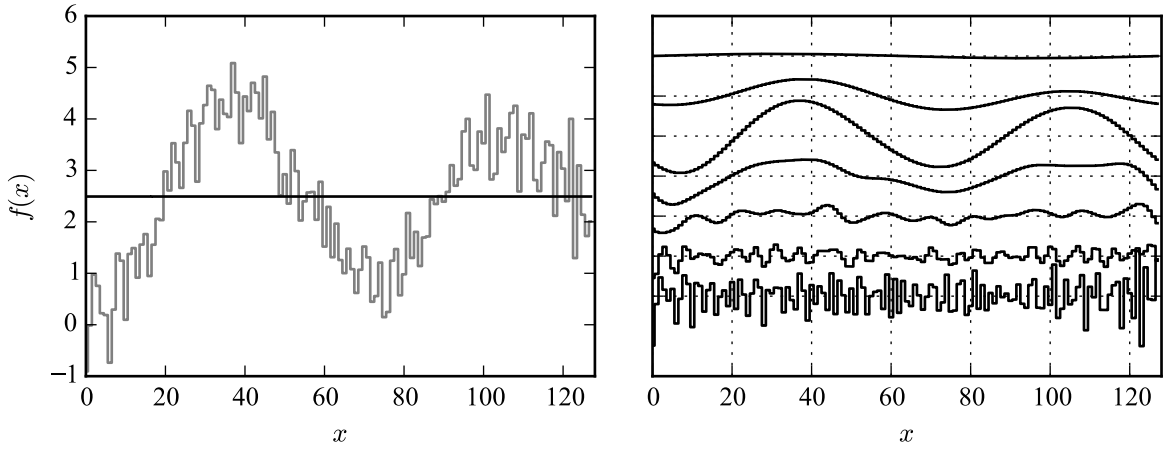


Figure 2.1: One-dimensional example of the un-decimated wavelet transform. The left panel shows the original signal in light gray. The solid black line corresponds to c_J, the smooth approximation. The right panel shows the wavelet coefficients w_j in increasing order of scale from bottom to top. The coefficients are offset vertically for visualization. All coefficients vary around 0.

k is the number of elements in the convolution kernel h_j.

The kernels h_j required to bring the data from resolution j to resolution $j+1$, are calculated using the *algorithme á trous* (Holschneider et al. 1989). The kernel h_0 is given by the so-called B3-spline

$$h_0 = [1, 4, 6, 4, 1]\,/16\,. \tag{2.3}$$

The kernels h_j are generated by inserting $2^{(j-1)}$ zeros between the kernel values of h_0. For example, the kernel for scale $j = 2$ would have the values

$$h_2 = [1, 0, 0, 4, 0, 0, 6, 0, 0, 4, 0, 0, 1]/16\,. \tag{2.4}$$

Since only non-zero kernel elements contribute to the convolution, the passage from resolution j to $j+1$ is calculated in the same time, regardless of the scale j.

2.3.2 Combination of Transformations

The *starlet transform* is well adapted to the typical shapes found in optical images. In H I astronomy, we are dealing mostly with data cubes. Although both optical galaxies and H I clouds can be composed of isotropic features, we have additional information in the data cube regarding the spectral extent of each H I source.

One way to include the additional information in the spectral domain to extent the filters in Eq. 2.2 to include the third dimension and calculate the coefficients according to

$$\begin{aligned} w_{j+1}[k,l,m] &= c_j[k,l,m] - c_{j+1}[k,l,m] \\ c_{j+1}[k,l,m] &= \left(h_j[k]\,h_j[l]\,h_j[m] \star c_j\right)[k,l,m] \end{aligned} \tag{2.5}$$

This is the 3D isotropic equivalent to the *starlet transform*. However, this does not account for the fact that, especially for single-dish surveys, H I sources can be spatially unresolved but have a well-resolved line profile.

To have a transform better adapted to the case of H I data cubes, we apply the 2D-1D transform proposed by Starck et al. (2009) for Fermi-LAT data. Here, instead of performing a full 3D wavelet transform, the spatial, two-dimensional transform and the spectral, one-dimensional transform are interleaved.

Omitting indices k, l, m, we first apply a 2D *starlet transform* to each spectral channel of a 3D data cube with J scales

$$c_0 = c_J + \sum_{j=1}^{J} w_j \tag{2.6}$$

We then apply a 1D wavelet transform along each line of sight with I scales to each cube of 2D coefficients at scale j

$$\begin{aligned} c_J &= c_{IJ} + \sum_{i=1}^{I} w_{iJ} \\ w_j &= w_{Ij} + \sum_{i=1}^{I} w_{ij} \end{aligned} \tag{2.7}$$

Reordering terms, we arrive at the 2D-1D wavelet transformation of the original data c_0

$$c_0 = c_{IJ} + \sum_{i=1}^{I} w_{iJ} + \sum_{j=1}^{J} w_{Ij} + \sum_{j=1}^{J} \sum_{i=1}^{I} w_{ij} \tag{2.8}$$

where c_{IJ}, w_{iJ}, w_{Ij} and w_{ij} are all three dimensional arrays of the same size as c_0.

To illustrate two key advantages of the 2D-1D transform over the 3D transform, we simulate a data cube as described in Appendix A. The data cube has dimensions $(x, y, z) = (128, 128, 256)$ and contains one unresolved galaxy. We calculate the isotropic 3D and 2D-1D wavelet coefficients and select the wavelet sub-band in which the galaxy has the highest SNR for a given transform. The simulated data as well as the two sub-bands are shown in Fig. 2.2. The first advantage of the 2D-1D transform over the isotropic 3D transform can be seen from a comparison of the SNR: the highest SNR achieved in the 2D-1D transform is 20.19, while the isotropic 3D transform has a maximum SNR of 11.48. This significant increase in SNR is highly advantageous when performing wavelet de-noising (see Sec. 2.3.3).

As described in Sec. 2.2, single-dish observations are commonly degraded by RFI and varying baseline levels. To illustrate how these data-defects propagate into the wavelet coefficients of the 2D-1D and 3D transform, we add simulated RFI and a varying baseline to the data cube described the above section. Figure 2.3 shows the same wavelet sub-bands as shown in Fig. 2.2. Both RFI and residual baselines data propagate through all sub-bands of the 3D transform, but are not present in every sub-band of the 2D-1D transform. This enables the accurate detection and localization of sources even in data containing the artifacts typically encountered in single-dish H I surveys.

2.3.3 Wavelet De-noising

Wavelet de-noising is the process of reconstructing data only from its significant wavelet coefficients. Following the notation introduced in Starck et al. (2010), suppose we have some measurements Y that

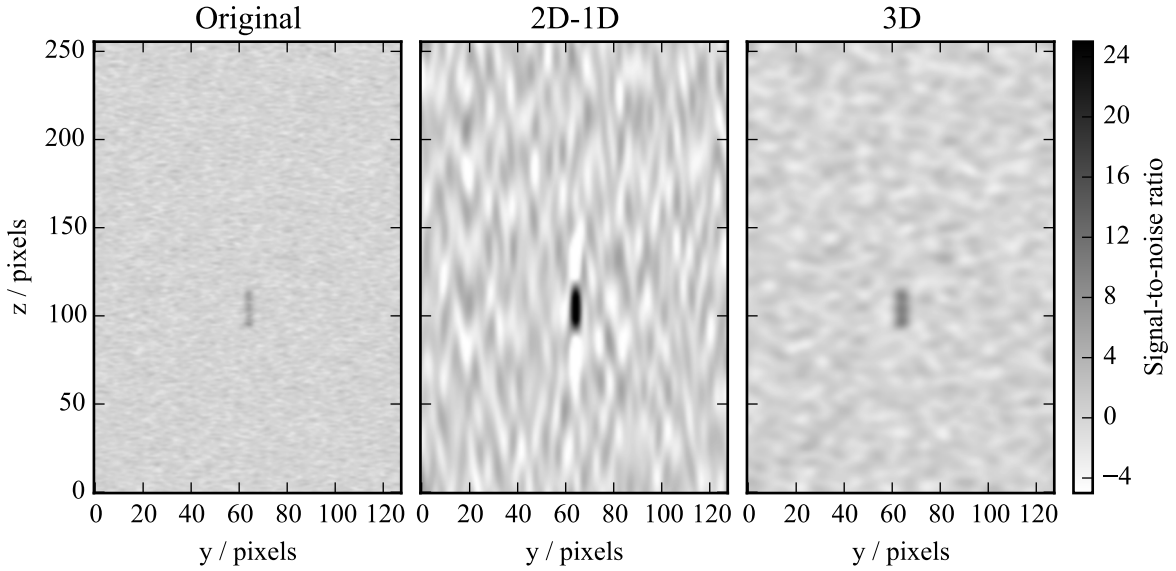


Figure 2.2: **Left**: Slice through a simulated data cube, containing one simulated, unresolved H I galaxy. **Middle**: Sub-band $w_{2,5}$ of the 2D-1D wavelet coefficients having the highest signal-to-noise ratio (SNR) for the simulated source. **Right**: Sub-band w_3 of the 3D wavelet coefficients having the highest signal to noise ratio for the simulated source.

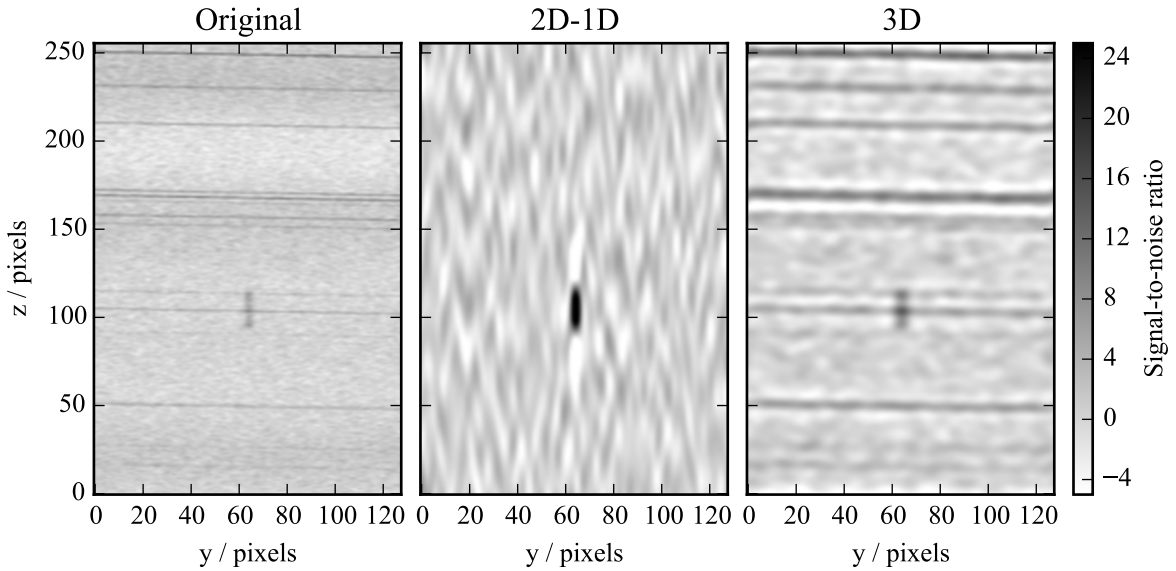


Figure 2.3: Same as Fig. 2.2, but with added noise, simulated baseline and RFI. Note how the RFI signal is propagated into the 3D wavelet sub-band but excluded from the 2D-1D sub-band shown.

are a combination of the true signal X corrupted by additive noise ϵ:

$$Y = X + \epsilon \tag{2.9}$$

To obtain an estimate of X, $\tilde{X}$, we perform the following operation:

$$\tilde{X} = \mathbf{R}\mathcal{D}(\mathbf{T}Y) \tag{2.10}$$

$\mathbf{T}$ is the wavelet transformation operator, e.g., the 2D-1D transform introduced in the previous section, to obtain the wavelet coefficients of Y for each sub-band. $\mathcal{D}$ is a non-linear operator that modifies the coefficients. For de-noising, $\mathcal{D}$ is commonly chosen to be either the hard thresholding rule (Starck & Bijaoui 1994)

$$\tilde{w}_j = \begin{cases} w_j & \text{if } \left|w_j\right| \geq t_j \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

or the soft thresholding rule (Donoho 1995; Donoho & Johnstone 1995)

$$\tilde{w}_j = \begin{cases} \text{sign}(w_j)(\left|w_j\right| - t_j) & \text{if } \left|w_j\right| \geq t_j \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

where t_j denotes the threshold in sub-band j. For un-decimated wavelet transforms, such as used in this thesis, hard thresholding gives better results (Starck et al. 2010). To obtain $\tilde{X}$, the data is reconstructed by applying the inverse wavelet transformation $\mathbf{R}$ to the modified wavelet coefficients $\tilde{w}$. In the case of the 2D-1D transform introduced in the previous section, this reduces to adding up all modified wavelet coefficients, i.e.,

$$\tilde{c}_0 = c_{IJ} + \sum_{i=1}^{I} \mathcal{D}(w_{iJ}) + \sum_{j=1}^{J} \mathcal{D}\left(w_{Ij}\right) + \sum_{j=1}^{J}\sum_{i=1}^{I} \mathcal{D}\left(w_{ij}\right) \tag{2.13}$$

Noise Estimation

To determine which wavelet coefficients are significant, it is important to have an accurate estimate of the noise level in each sub-band, σ_j. The value of σ_j depends on the noise in the data, σ, and the type of transform used. There are multiple methods described in the literature to estimate the noise level for each sub-band. A simple approach is proposed by Johnstone & Silverman (1997), who calculate the noise level in each sub-band separately using the median absolute deviation (MAD):

$$\sigma_j = \text{MAD}(w_j)/0.6745 \tag{2.14}$$

This avoids special treatment of spatially correlated noise present in H I data cubes. Estimating the noise level by this simple method assumes, that σ_j does not vary as a function of location in the data cube. We address this issue in Sec. 2.4.

Iterative Reconstruction

When modifying individual values of the wavelet coefficients in a given sub-band w_j, we introduce information that does not belong at this scale. This leads to a phenomenon not unlike the Gibbs phenomenon

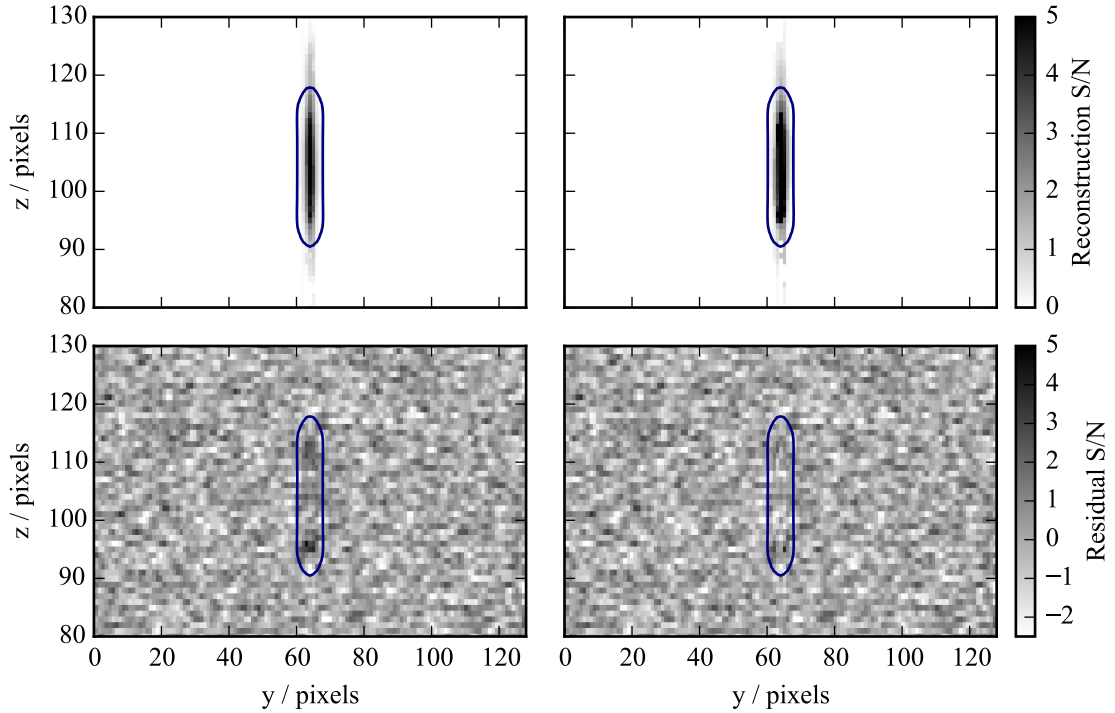


Figure 2.4: Wavelet reconstructions and residuals of the simulated data cube containing a galaxy and simulated noise. In all panels, the blue contour indicates the 0.1σ contour of the simulated source. **Left Column**: Reconstruction (top) and residual (bottom) after a single iteration of hard thresholding with a 5σ threshold. The residual clearly contains leftover structure of the line profile. **Right Column**: Reconstruction and residual after ten iterations of hard thresholding with a 5σ threshold and use of a MRS. No significant structure is left in the residual.

when introducing sharp edges in the Fourier transform of a signal. To obtain high reconstruction quality it is necessary to perform iterative reconstruction (Starck et al. 2007). For this purpose, Starck et al. (1995) introduce the multi-resolution support (MRS), **M**, which is unity if a wavelet coefficient is significant and zero if not. Using the MRS, we can use the Landweber iterative scheme (Landweber 1951; Combettes & Pesquet 2011) to obtain a solution

$$\tilde{X}^{n+1} = \tilde{X}^n + \mathbf{RMT}(Y - \tilde{X}^n) \quad . \tag{2.15}$$

During the first iteration $\tilde{X}^0 = 0$ everywhere and the operation **RMT** applies the wavelet transform to the data, performs hard thresholding of the wavelet coefficients, and applies the inverse wavelet transform. In subsequent iterations, the residual, $Y - \tilde{X}^n$, is searched for further significant coefficients which are added to the reconstruction. If the true signal is known to be positive, as is the case for H I data, we can enforce this constraint on the solution: in each iteration, negative values in $\tilde{X}$ are set to zero. This greatly improves the reconstruction quality (Starck et al. 2007). Starck et al. (2010) outline multiple ways of determining the MRS of the data. We choose to iteratively build up the MRS. During each iteration, we determine which coefficients are significant and set the MRS to one for these coefficients. Coefficients are never removed from the MRS.

Figure 2.4 shows the reconstruction and residual of the same data cube for three different approaches. The first result is from a single iteration of hard thresholding at a threshold of 5σ. This already recon-

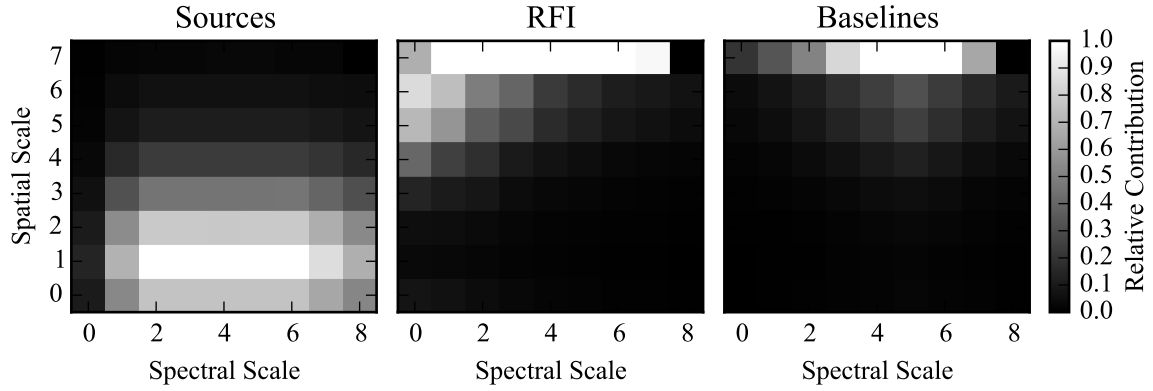


Figure 2.5: Relative contribution to the coefficients of the 2D-1D wavelet transform for various features present in H I data cubes as described in the text.

structs the bulk of the source signal but misses the sharp rise and fall at the edges of the line profile of the source. The second result is from the wavelet reconstruction as we use it for EBHIS source finding. Here, we perform ten iterations at a threshold of 5σ and add the significant coefficients to the MRS in each iteration. This approach is able to extract more signal from the data and the residual shows no significant structure. A larger number of iterations does not improve the reconstruction quality any further.

Scale Selection

The number of scales J with which a particular image or data set is decomposed is usually chosen by $J = \lfloor \log_2(d) \rfloor$ if d is the size of the largest dimension of the data set. But if the signal is only present on certain scales, e.g., H I galaxies are typically much smaller than the typical size covered by a data cube, we can limit our analysis to these scales.

This is especially advantageous when using the 2D-1D wavelet transformation for de-noising. Since it is an interleaved transform there is a much larger number of wavelet coefficients than, for instance, with the 3D wavelet transform. By excluding large spatial and spectral scales, we reduce the number of coefficients and therefore speed up calculations. Furthermore, as shown in Fig. 2.3, there might be scales in the wavelet decomposition of the data which contain undesirable signal. These scales can be treated with a higher threshold or excluded completely when de-noising the data.

To determine which scales of the 2D-1D wavelet transform are most susceptible to RFI and residual baseline, we simulated 1000 noise-free data cubes containing either only a single H I galaxy from our spectra library, simulated RFI or baseline ripple, respectively as described in App. A. For each data cube, we perform the 2D-1D wavelet decomposition and measure the highest SNR achieved in each sub-band w_{ij}. As the simulated data cubes are noise-free, the SNR is estimated by applying the 2D-1D wavelet decomposition to a data cube containing only simulated noise and measuring the noise level in each sub-band. To get an intensity-independent measure of the magnitude with which each particular feature contributes to each sub-band, we normalize the SNR for each sub-band to the highest SNR achieved in any sub-band of the cube. For each of the contributions, Fig. 2.5 shows the maximum normalized SNR across all simulated data cubes for each wavelet scale. This gives an intensity-independent measure of which scales are important for each feature.

Due to the different profile widths of H I sources, they span a range of spectral scales but are mostly

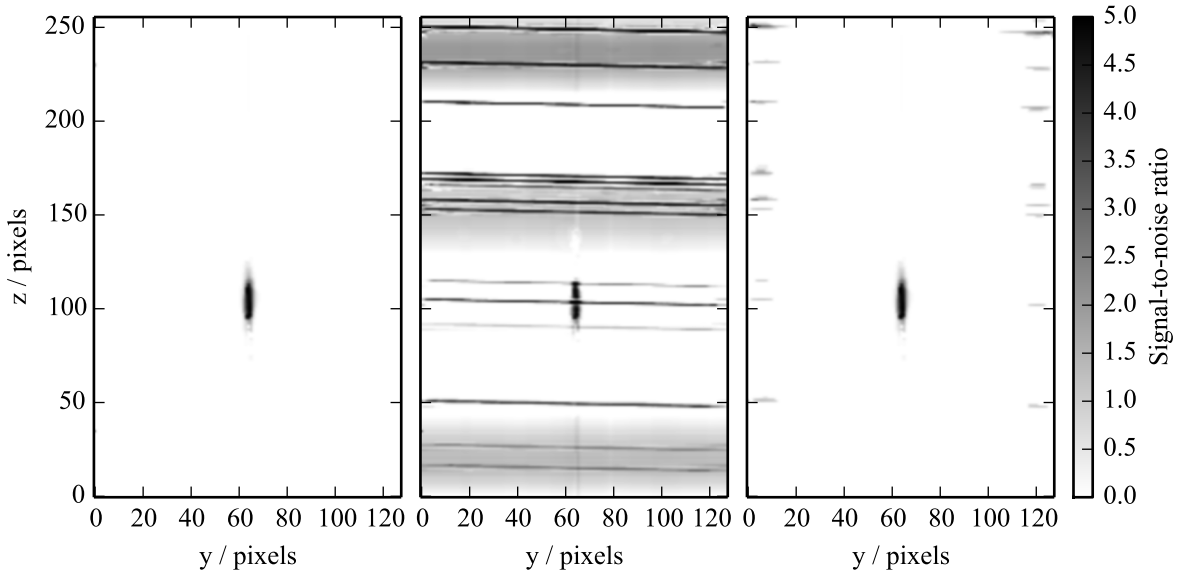


Figure 2.6: **Left**: 2D-1D wavelet reconstruction of the simulated clean data shown in Fig. 2.4 with limited decomposition scales, as described in the text. **Middle**: 2D-1D wavelet reconstruction of the simulated corrupted data shown in Fig. 2.3 without setting a limit the scales considered. **Right**: 2D-1D wavelet reconstruction of the same data shown in the middle but with limited decomposition scales as described in the text.

confined to the first three spatial scales. Here we can see the unresolved nature of the simulated sources. More extended sources will also contribute to larger spatial scales. But since the step in resolution is a factor of two between the spatial scales, even the most extended extragalactic sources will not contribute much to spatial scales beyond three.

The picture is very different for both RFI and baseline effects: They mostly contribute to very large spatial scales as they do not exhibit small-scale spatial variations. The range of spectral scales occupied by RFI is explained by the nature of the 2D-1D wavelet transform: Since the transform is linear, high-intensity spikes at small spectral scales propagate to larger scales. The baselines only contribute to the spectral decomposition of the largest spatial scale, i.e., the coefficients w_{iJ} in Eqn. 2.8. It is evident that sources and defects in the data occupy vastly different scales. We therefore limit the number of spatial scales considered for the reconstruction during source finding to $I = 4$ and exclude all coefficients of the smallest spectral scale, i.e., w_{0j}. Furthermore, we do not include the terms w_{iJ}, w_{Ij} and c_{IJ} in our reconstruction as they only contain information at the scale of the data cube.

To highlight the advantage gained by scale selection, we show three different wavelet reconstructions in Fig. 2.6. Using our limited scales, we first reconstruct the simulated clean data set used above. There is no significant difference in the reconstruction quality. We then perform a 2D-1D wavelet reconstruction of the corrupted data set used in Fig. 2.3, with and without limiting the reconstruction scales to the range described above. Carefully selecting the number of scales used in the reconstruction clearly improves the detection capabilities of the 2D-1D wavelet de-noising approach to source finding. Using the 3D transform, this scale selection would not be possible. Any kind of matched filtering would smear out the defects in the data cube, making a detection impossible.

2.3.4 Implementation

For the source finding pipeline, the 2D-1D wavelet transform and the components needed for de-noising are implemented in `Python`[1] with use of `Cython`[2] extensions. `Python` allows fast prototyping, experimentation and visualization. The time-critical operations, such as the convolutions involved in the *algorithme á trous*, are implemented in `Cython`, which generates C code that is subsequently compiled.

Since the 2D-1D transform of a data set yields $J \times I$ coefficients, it is not practical to calculate full transformation and keep all coefficients in memory concurrently. Instead, the de-noising process is implemented serially: Whenever a set of coefficients is calculated, the wavelet transform class calls a function `handle_coefficients`, which is handed the location of the current wavelet coefficients as an index to an internal work-array as well as the indices of the current scale, i.e., j and i. The full source code with comments is published on-line[3] and also part of the SoFiA source finding package (Serra et al. 2015), also published on-line[4].

This approach allows to calculate the complete transform of a data cube of size N with a work array of only $3N$ size. If a reconstruction is to be built up, a further array of size N is required to keep the reconstruction. For the MRS an array of $I \times J \times 3N$ boolean values is required. This is by far the largest part of the memory footprint for the transform and limits the practical size of the input data cube. This restriction is of no concern for single-dish data cubes as a typical 10° by 10° data cube for the extragalactic EBHIS survey is between 200 MB to 300 MB in size, depending on declination. This yields a total memory footprint of the 2D-1D transform using an MRS of 2.4 GB to 3.6 GB for four spatial scales and eight spectral scales.

2.4 Source Finding for the Effelsberg-Bonn HI Survey

We use the de-noising scheme developed in the previous section as the source finder for extragalactic EBHIS data. We typically apply the 2D-1D wavelet de-noising algorithm to data cubes covering 10° by 10° and the full spectral range of EBHIS. To use the de-noising scheme as a source finder, we have to take into account the variation of EBHIS sensitivity as a function of position and frequency and create initial source candidates from the de-noised data cubes.

The sensitivity of EBHIS is not constant but rather varies as a function of position and frequency. There are multiple factors influencing the sensitivity. First of all, EBHIS observations are split up in individual fields of 5° by 5° with an overlap region between the fields. When creating the larger data cubes from these individual observations, the sensitivity in the overlap region is higher and therefore has a lower noise level. Furthermore, the process of RFI flagging lowers the sensitivity for certain spectral channels. Also, continuum emission raises the system temperature of the receiving system, which also increases the noise level as a function of position. Lastly, the gain of both the radio frequency (RF) and intermediate frequency (IF) band passes for all 14 basebands varies slightly across the 100 MHz bandwidth.

With the exception of increased system temperature due do bright continuum point sources, all variations in the sensitivity of the survey are reflected in the weights associated with the data. By multiplying the data with the square root of the weights prior the de-noising process, the data has uniform variance. To also take into account the increased noise level due to bright continuum point sources, we measure

[1] `http://www.python.org`
[2] `http://www.cython.org`
[3] `https://github.com/lfloeer/sofia2d1d`
[4] `https://github.com/SoFiA-Admin/SoFiA`

the noise level for each line-of-sight individually with the robust MAD statistic. By dividing each line-of-sight by the measured noise level, the data cube has a uniform noise level and we can use the simple noise estimate described in Sec. 2.3.3 in each of the wavelet sub-bands.

Once the data is reconstructed we create a binary mask from the reconstruction. All non-zero values are set to 1 and everything else is set to 0. We then use an object generation code developed by Jurek (2012). The algorithm assigns a unique label to each neighboring group of non-zero voxels and thereby creates a three dimensional mask tracing the shape of each source candidate. The code also performs size thresholding on the generated source candidates. Candidates that occupy less than three spectral channels or whose spectrally collapsed, spatial mask has less than six pixels are removed from the binary mask.

For the purpose of source finding, we divide the sky into equally sized fields of 10 by 10 degree size with an overlap of one degree on each edge. The first field is centered on equatorial coordinates $(\alpha, \delta) = (0°, 0°)$ and subsequent fields are offset by 8° in Right Ascension, creating the desired overlap between the individual fields. Once the whole range in Right Ascension has been covered, we start a new set of fields from coordinates $(\alpha, \delta) = (0°, 8°)$, and so on. To keep the field size constant, the size of the fields in Right Ascension is scaled by a factor of $\cos^{-1}(\delta)$. The final location of all fields is shown in Fig. 2.7. The fields with field centers of declinations 80° and below use the Sanson-Flamsteed projection. The data at the equatorial north pole is projected into a data cube using the zenithal equal area projection. Both projections are equal area projections and are suited for pixel-based integration (Calabretta & Greisen 2002). The spectral axis of the data cubes is limited to the velocity interval $-500\,\mathrm{km\,s^{-1}} < v_{\mathrm{LSR}} < 15\,000\,\mathrm{km\,s^{-1}}$. The sensitivity of EBHIS is not sufficient as to expect any reliable detections beyond this upper velocity limit and the impact of broad-band RFI is more pronounced in this range.

2.5 Conclusions

In the preceding chapter, we describe the application of the 2D-1D wavelet de-noising technique in the context of source finding for H I surveys. We show that treating the spectral and spatial axes of data cubes separately has multiple advantages over state-of-the-art source finding methods based on matched filtering and isotropic 3D wavelet de-nosing. Since the 2D-1D transform is adapted to the shape of H I line profiles of galaxies, it achieves much higher SNR as compared to the isotropic 3D wavelet transform. This directly translates into increased sensitivity when performing source finding for spectroscopic surveys. The independent treatment of spatial and spectral axes furthermore allows to only select wavelet scales that contain the signal of interest. Our simulations show that common single-dish data defects like RFI and residual baselines occupy different wavelet scales than sources. By ignoring these scales, we can limit the influence data defects have on the source finding process.

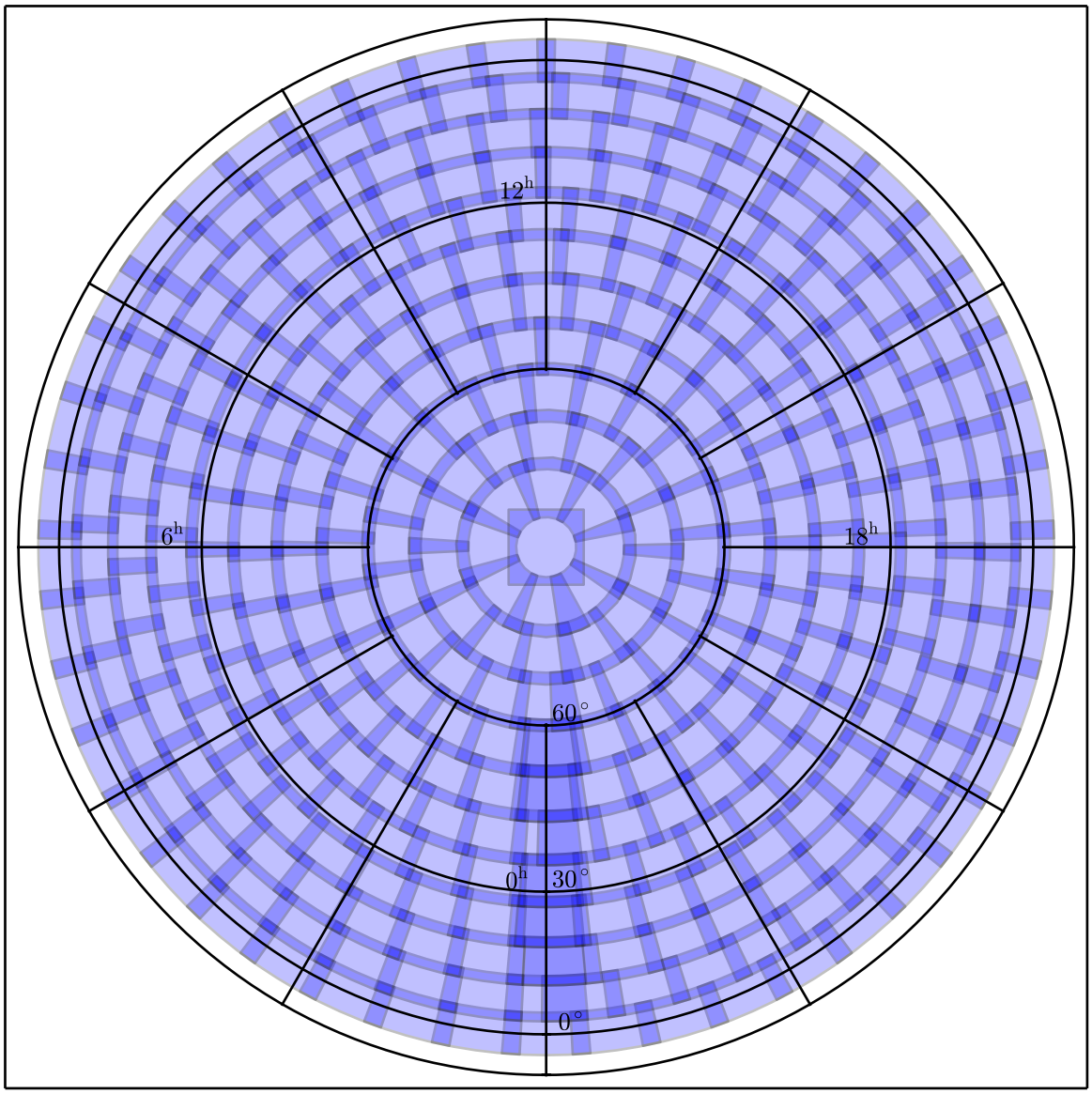


Figure 2.7: Zenithal equal area projection showing the approximate locations of the individual $10^\circ \times 10^\circ$ fields created for pipeline processing. Darker tones indicate the overlap between the individual fields. The increased overlap at Right Ascension 0^h is caused by the simple, periodic field layout.

CHAPTER 3

Automatic Parametrization

The results from this Chapter are published in part in

- Flöer, L., Winkel, B., & Kerp, J. 2014, A&A, 569, A101

The majority of manual labor involved in large-scale surveys of H I is spent on the parametrization of the signal detected from galaxies or other sources. For both the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005) and the H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001), the candidates generated by the source-finding software (see Sec. 2.1) are inspected by eye and, if classified as a real source, interactively parametrized. Even for the relatively modest data volumes of HIPASS and ALFALFA, this amounts to the inspection of over 100 000 candidates by multiple astronomers. Since the absolute number of false positives scales with the data volume as opposed to the physical survey volume, future H I surveys with the Australia SKA Pathfinder (ASKAP, Johnston et al. 2008), the upgraded Westerbork Synthesis Radio Telescope (WSRT/Apertif, Oosterloo et al. 2010), and the South African SKA Pathfinder (MeerKAT, Booth et al. 2009) will produce vastly more candidates that need to be inspected and parametrized.

There are efforts to develop automated parametrization algorithms for H I surveys, (e.g. Whiting & Humphreys 2012), but no complete survey has ever been parametrized without human supervision. In this chapter, we describe our approach to automated parametrization for the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011). For this, we combine an automation of time-tested algorithms with new approaches. In Sec. 3.1 we provide a brief overview of the current approach to parametrization for large-scale, single-dish H I surveys. Section 3.2 discusses in detail the individual steps performed by our automated pipeline. In Sec. 3.3 we perform large simulations involving 24 000 synthetic sources, to quantify the performance of our parametrization pipeline and derive highly significant completeness levels and 95% confidence regions for each measured parameter. Furthermore, we compare the performance of the automated pipeline to the uncertainties achieved with the manual parametrization for HIPASS.

3.1 Parametrization for Neutral Hydrogen Surveys

With the exception of a few blind, interferometric surveys, most large-scale H I surveys detect predominantly unresolved sources. Even though future surveys with the upcoming Square Kilometre Array (SKA) pathfinders will resolve most sources in the very local universe, most detections will still be only marginally resolved or completely unresolved (Duffy et al. 2012). It is therefore still worthwhile to

investigate the parametrization of unresolved H I sources. Typically, the parameters of interest for an unresolved H I source are its position on the sky, redshift, profile width and integrated H I flux density.

The position of H I sources are typically determined from their center-of-mass coordinates, which is measured on a velocity-integrated map of the source. As long as the source pixels are sufficiently bright, the center-of-mass coordinates give an accurate measure of the position of the source. If there are multiple sources in the field or the source pixels are not sufficiently bright, the center-of-mass coordinates become inaccurate. To circumvent issues in such low signal-to-noise ratio (SNR) cases, it is common to fit a two dimensional Gaussian function to the velocity-integrated map of the source.

Redshift and profile width are measured from the line profile of the source. The redshift of a particular H I detection is usually determined from the midpoint of its profile. There are numerous ways of measuring the width of a profile described in the literature. The simplest methods measure the width of the profile at a certain flux density. Most authors choose the flux density to be a fraction of the peak flux of the profile. Customary values are 50% and 20% of the peak flux density. Another approach is used by Courtois et al. (2011), who measure the width of the profile at 50% of the mean flux density across the profile. For high SNR cases, all of these methods perform equally well. Once the SNR ratio of the line profile drops below ten, these measurements become increasingly inaccurate (Lavezzi & Dickey 1997). There are recent efforts to combat this loss in accuracy through modeling of the line profile (Westmeier et al. 2014; Stewart et al. 2014).

Due to the optically thin nature of the 21 cm line of H I, the integrated line flux of a galaxy is proportional to its H I mass. This makes it one of most readily available physical parameters of a source and is of importance when determining the H I mass function (HIMF), that is, the space density of galaxies as a function of their H I mass. Most authors measure the integrated H I flux density by integrating over the spatial extent of the source in a velocity-integrated map (for example, Meyer et al. 2004; Haynes et al. 2011; Wolfinger et al. 2013). Another approach, especially viable for unresolved or marginally resolved sources, is fitting a small number of 2D Gaussian components to the velocity-integrated map of the source. The `Selavy` source finding pipeline developed for ASKAP is a recent example of this approach (Whiting & Humphreys 2012).

Up to now, the parametrization steps described above are used interactively. This allows for quick intervention if one of the parametrization algorithms fails, e.g., due to low SNR. Even though source-finding software like `Duchamp` offers automatic parametrization, the measured parameters are not considered reliable by many authors (for example, Wolfinger et al. 2013; Haynes et al. 2011). Manual parametrization of sources is still feasible for surveys like HIPASS or ALFALFA: False positives are identified by eye and only true detections are interactively parametrized. As argued in Sec. 2.2, the number of candidates for future surveys with ASKAP and WSRT/Apertif, will make an automated decision about the true nature of a candidate necessary. To perform accurate, automated classification, each candidate requires to be parametrized. Since the number of candidates for past surveys is on the order of 10^5 and future surveys are expected to generate more candidates, automated parametrization is clearly a necessity.

3.2 Automated Parametrization for the Effelsberg-Bonn H I Survey

Future large-area surveys need a full automation of the source-finding and parametrization process. We use EBHIS as a testbed to better understand the challenges posed by this requirement. A general overview of the EBHIS parametrization pipeline is shown as a flow diagram in Fig. 3.1. In general, the pipeline is split in two parts.

The first part is concerned with finding an optimal mask for the source. This is necessary for two

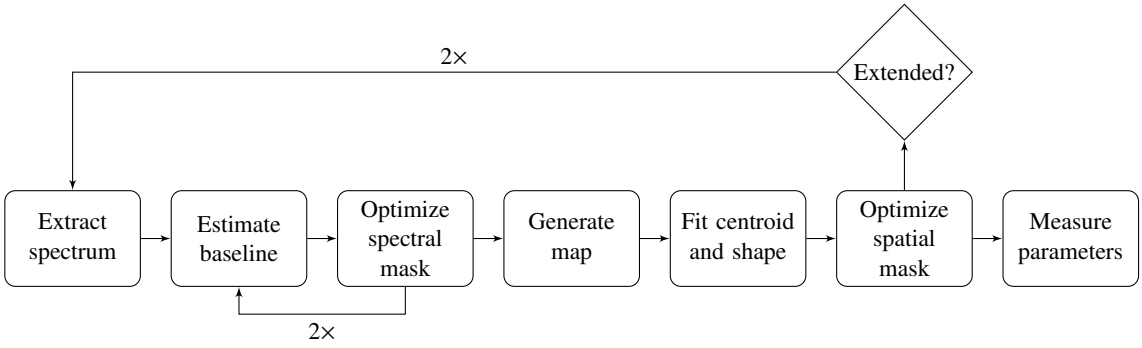


Figure 3.1: Flow diagram of the parametrization process for each source candidate.

reasons. Every source finder has some threshold below which emission is not considered significant. Since H I sources are either resolved or smeared out by the telescope beam, no single threshold will capture the whole emission. This leads to a bias when measuring the integrated H I flux density, as noted by Westmeier et al. (2012) for the `Duchamp` source-finding package. For this reason, both the spatial and spectral mask of the source candidate are optimized during the first stage of the pipeline. Since a change in the spatial mask can change the spectrum of the source and therefore the spectral mask, the mask optimization steps are iterated. Once the masks are optimized, the final parameters are measured in the second part of the pipeline.

3.2.1 Spectra Extraction

The first task of the pipeline is the extraction of the source spectrum, i.e., the flux density in units of Jy as a function of radial velocity v or frequency ν. Here we distinguish between the peak spectrum and the integrated spectrum.

The peak spectrum is the spectrum extracted at the angular coordinates where flux density is maximal in the velocity-integrated map of the source. During the first iteration of the mask optimization process, the maximum coordinate is approximated by masking the integrated-velocity map with the mask returned by the source finder and calculating the center-of-mass coordinate. In subsequent iterations, we use the fitted centroid of the source and the optimized spatial mask of the source.

We reconstruct the peak spectrum $f_p(v)$ according to the formula

$$f_p(v) = \frac{f(p_x, p_y, v)}{B(c_x - p_x, c_y - p_y)} \tag{3.1}$$

Here, p_x and p_y are the coordinates of the pixel closest to the determined location of the maximum at c_x, c_y. Since the source is sampled on a pixel grid, we need to interpolate the data to extract the true peak spectrum $f_p(v)$. We interpolate its values from the data f by dividing by the normalized[1] beam function B, where B is approximated by a 2D Gaussian with the appropriate full width at half maximum (FWHM). Here the data have units of Jy beam^{-1}.

The peak spectrum $f_p(v)$ fully describes the spectral shape for an unresolved source. If the source is resolved, we instead have to integrate over the angular extent of the source to obtain the integrated

[1] $B(0, 0) = 1.0$

spectrum $f_i(v)$. The integrated spectrum is calculated by

$$f_i(v) = \sum_{(x,y)\in M} f(x, y, v) \tag{3.2}$$

where the sum runs over all points (x, y) that are part of the spatial mask M of the source. For this simple pixel-based integration the data are converted to have units of Jy pixel^{-1}. The conversion between Jy pixel^{-1} and Jy beam^{-1} is given by

$$f\left[\text{Jy pixel}^{-1}\right] = \frac{\Omega_{\text{pixel}}}{\Omega_{\text{beam}}} f\left[\text{Jy beam}^{-1}\right] \tag{3.3}$$

where Ω_{pixel} and Ω_{beam} are the solid angle of a pixel and the beam, respectively. For this conversion to hold, all pixels are required to have the same solid angle, which is the reason why we are using the Sanson-Flamsteed and azimuthal equal area projection for EBHIS data cubes, as discussed in Sec. 1.3.

The first iteration of the pipeline assumes an unresolved source and performs the mask optimization steps using the peak spectrum $f_p(v)$. During the first iteration, the shape of the source is measured. If the source is determined to be resolved, the second iteration is performed using the integrated spectrum $f_i(v)$. Sources are treated as resolved if their spatial extent exceeds 1.5 times the FWHM of the EBHIS beam.

3.2.2 Baseline Estimation

As explained in Sec. 2.2, single-dish H I data cubes commonly exhibit non-flat baselines. To obtain accurate integrated flux densities and profile widths, we remove this residual baseline level during parametrization. The most common way of removing the baseline level of H I spectra involves masking the region which contains the emission of interest and interpolating the spectrum with a polynomial of sufficient order. When performing interactive parametrization, the order is chosen by the user, which is not possible in an automated pipeline.

Instead, we use smoothing splines for baseline estimation. Smoothing splines have long been used to smooth noisy data (Reinsch 1967, 1971). Their only free parameter, the smoothing factor, can be determined from the data using the method of generalized cross-validation. Garcia (2010) introduce a version of splines that are robust to outliers and gaps in the data. They are therefore very well adapted to the realities of automated baseline estimation. Meyer et al. (2004) already explore the use of smoothing methods for baseline estimation for HIPASS. As the method depends on interactive user input, it is not well suited for automation.

We simulate 10 000 spectra to verify the reliability of our approach. Each spectrum consists of 512 channels with uncorrelated Gaussian noise of unit variance. We generate a baseline by the superposition of two sine functions with wavelengths motivated by the typical scales of observed, large-scale variations:

$$B(x) = \sin\left[\frac{2\pi}{512}(x - \phi_1)\right] + \sin\left[\frac{2\pi}{1024}(x - \phi_2)\right] \tag{3.4}$$

The phases $\phi_{1,2}$ are chosen randomly for each spectrum. In each of the simulated spectra, we mask a central region between 3 and 50 channels wide and estimate the baseline from the remaining data using our smoothing spline approach. We then calculate the sum of the difference between the true and estimated baseline in the masked region. This quantity is an estimate of the error made when measuring the integrated flux density from the baseline-subtracted spectrum. We also perform a second simulation

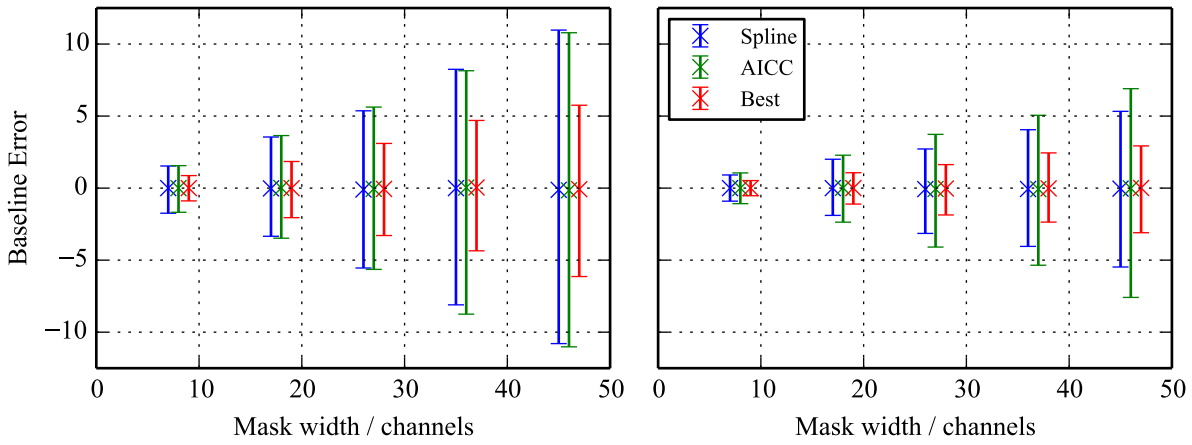


Figure 3.2: Results of the baseline simulations. Both panels show the absolute error between the true and estimated baseline as described in the text. The error bars indicate the range in which 95% of the values are contained. The crosses indicate the median of the values. The three measurements are slightly offset in mask width for clarity. "Spline" and "AICC" refer to the performance if the baseline is estimated with the respective methods. "Best" represents a lower bound for the uncertainties incurred during baseline fitting. **Left panel**: Results for spectra with a simulated baseline. **Right panel**: Results for spectra with no simulated baseline.

where the baseline amplitude is set to zero. The goal of this simulation is to test the stability of the baseline solution in the masked region.

As a comparison, we estimate the baseline with polynomials of degree between zero and ten and compare the results from our spline-based approach to the best-performing polynomial. This gives a best-case estimate, which is not achievable in practice. To have a more realistic comparison, we also compare our results to a polynomial fit selected using the corrected Akaike infromation criterion (AICC, Akaike 1974; Hurvich & Tsai 1989). The AICC is a relative measure of goodness-of-fit. It is calculated by

$$\mathrm{AICC} = \underbrace{n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2k}_{\text{classical AIC}} + \underbrace{\frac{2k(k+1)}{n-k-1}}_{\text{bias correction}} \tag{3.5}$$

where n is the number of data points in the fit, RSS is the residual sum of squares and k is the number of parameters in the model. Among a set of models, the best fitting model has the lowest AICC. The absolute magnitude of the AICC has no meaning.

The results from our simulation are shown in Fig. 3.2. The error for both the AICC and spline-based methods is between 1.5 and 2 times larger than for the best-case scenario. In most cases the spline-based baseline estimation performs slightly better than the polynomial baselines. For the simulation without baselines the AICC sometimes favors overly complex models, which lead to an increased scatter in the polynomial baselines. This effect is especially noticeable for wider profiles as the AICC can not penalize models that are overly complex in the masked part of the data as they do not contribute to the RSS. We note that there are model selections procedures related to the AICC that can handle missing data (e.g. Ibrahim et al. 2008). Generally, these methods are much more complicated and often involve computationally expensive operations. Since our spline-based approach yields satisfying results and has further advantages over polynomial fitting, like the robustness to outliers, we do not investigate this

further.

To apply the robust smoothing splines to EBHIS data, we modify the way the smoothing parameter is estimated. In the original implementation by Garcia (2010), the smoothing parameter is estimated from the full data set. This is only sensible if the smoothness of the data does not vary over its range. This is not always the case for EBHIS data. Sometimes data defects caused by radio-frequency interference (RFI) or strong continuum sources generate strong but localized variations in the data. If the smoothing factor is determined from the whole spectrum, the smoothness is underestimated which makes the baseline solution unstable. We therefore limit the range from which the smoothing factor is determined to ±100 channels centered on the location of a source candidate.

3.2.3 Profile Width Measurement

To optimize the spectral mask for each source, we subtract the baseline from its spectrum and measure w_{50}, the width of the profile at 50% of its peak flux density. As shown by Lavezzi & Dickey (1997) and others, the accuracy of this method decreases rapidly once the peak SNR is lower than about seven. We can increase the peak SNR by smoothing the spectrum prior to measuring the profile width. This lowering of the spectral resolution leads to an overestimation of the profile width. Additionally, the optimal amount of smoothing required depends on the SNR and shape of the profile.

To automate the strength of smoothing applied to the spectra, we use bilateral filtering developed by Tomasi & Manduchi (1998). Instead of smoothing the data only based on proximity a bilateral filter also considers the similarity of neighboring data points. This characteristic makes a bilateral filter edge-preserving. In the context of H I spectra, a bilateral filter preserves the steep flanks of high SNR profiles and smoothes low SNR profiles.

Following the notation of Tomasi & Manduchi (1998), a bilateral filter performs the following operation on a one-dimensional signal $f(x)$

$$\begin{aligned} h(x) &= k(x)^{-1} \int_{-\infty}^{+\infty} f(\xi)\, c(\xi, x)\, s(f(\xi), f(x))\, d\xi \\ k(x) &= \int_{-\infty}^{+\infty} c(\xi, x)\, s(f(\xi), f(x))\, d\xi \end{aligned} \tag{3.6}$$

For signals sampled at discrete locations ξ_i, the integrals in both equations are replaced by sums over all values of ξ. The functions $c(\xi, x)$ and $s(f(\xi), f(x))$ are the spatial and range or radiometric kernels, respectively. Together, they enforce spatial and radiometric similarity. This can be thought of as a spatial filter that changes its size depending on the similarity of the values beneath it. For our pipeline we use the kernel functions proposed by Tomasi & Manduchi (1998)

$$\begin{aligned} c(\xi, x) &= \exp\left[-\frac{1}{2}\left(\frac{\|\xi - x\|}{\sigma_d}\right)^2\right] \\ s(f(\xi), f(x)) &= \exp\left[-\frac{1}{2}\left(\frac{\|f(\xi) - f(x)\|}{\sigma_s}\right)^2\right] \end{aligned} \tag{3.7}$$

which are one dimensional Gaussian kernels for both spatial and radiometric similarity. The parameters of the kernels σ_d and σ_s are chosen to 5 and $\sqrt{2}\sigma_{\mathrm{rms}}$, respectively, where σ_{rms} is the noise level of the data. Since the function $s(\xi, x)$ measures the radiometric similarity, it makes sense to scale the width of the filter with the noise level of the data. The factor $\sqrt{2}$ comes from the fact that the difference of two independent Gaussian random variates with the same standard deviation is a Gaussian random variate

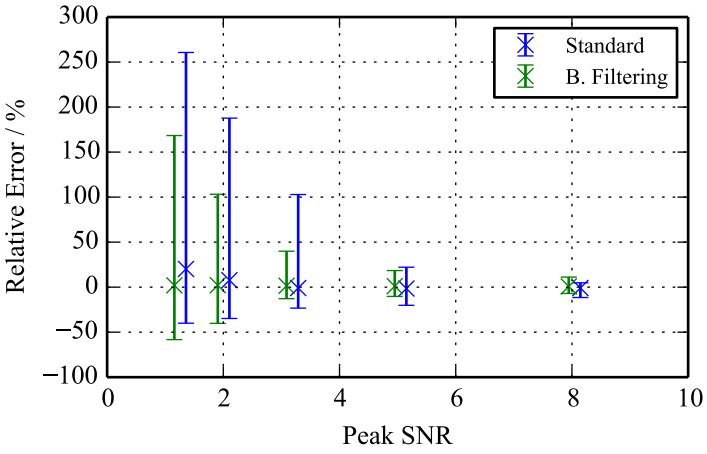


Figure 3.3: Results from the profile width measurement simulation. Shown are the relative errors in the w_{50} profile width over our whole spectra library. The error bars indicate the 95% range of errors. The cross indicates the median in each bin. The data points are offset horizontally for clarity. Bilateral filtering reduces both the scatter and bias of measuring the width of the profile at 50% of its peak flux density.

with a factor $\sqrt{2}$ larger standard deviation.

We test our approach on 10 000 simulated spectra (see App. A.1.4). The spectra are scaled to a random peak SNR between one and ten. To simulate a realistic uncertainty about the true location of the source, we create a mask for the emission in the spectra and extend this mask by five channels on either side. We measure the w_{50} profile width on the noise-free, noisy and filtered spectrum. Figure 3.3 shows the result of our simulations. The bilateral filter significantly outperforms the standard approach and yields usable results down to a peak SNR of three. Compared to the standard approach, the scatter is reduced up to 60%.

During mask optimization, we use the measured profile width of the source to define a new spectral mask. To capture all emission, we extend the mask by 50% of the measured width but at least four channels. After mask optimization, we use the same algorithm to measure the final profile width. The redshift of the source is measured as the midpoint between the two locations where the source profile rises above 50% of its peak flux density for the first time.

3.2.4 Extraction of Velocity-Integrated Maps

We create two different kinds of velocity-integrated maps for each candidate to measure an angular position and optimize its spatial mask. The first is a simple summation of all channels belonging to a source. This map gives a general overview of the data but can not be used for mask optimization as other sources present in the same spectral range are also present in this map. This can lead to the unintended merging of individual detections.

For these reasons, we create a second version of the map which we call *clean map*. For this, we first remove the current source candidate from the mask cube provided by the source finder. We then perform two iterations of binary dilation on the mask cube. This replaces each voxel in the data cube with the maximum of its surrounding values and thereby enlarges the masks for the source candidates currently not under consideration. We exclude emission from other candidates by applying this mask when calculating the velocity-integrated maps.

3.2.5 Centroid Fitting and Mask Optimization

As mentioned in Sec. 3.2.1, during the first iteration of mask optimization, the angular position of a source candidate is estimated from its center-of-mass coordinate. We improve the positional accuracy of the pipeline using iterative Gaussian fitting. Using the center-of-mass coordinate of the source as

starting parameters, we fit an elliptical Gaussian function to the brightness distribution of the source. The resulting model of the source is then used as the weights for subsequent iterations. After a small number of iterations (< 5), the fit converges and the process is stopped.

Once a good centroid is found, we optimize the spatial mask of the source. For this we use the clean map to calculate the integrated flux density in concentric annuli around the source candidate. We take the new mask to be the radius at which the integrated flux density for an annulus is negative for the first time. This is equivalent to increasing the size of the mask until the integrated flux density does not increase further. For unresolved sources the annuli are concentric circles. For resolved sources we use ellipses to better trace the shape of the source.

3.3 Performance of the Parametrization Pipeline

We evaluate the performance of our automated parametrization pipeline by simulating data. We simulate two sets of 120 data cubes, each containing 100 simulated sources using the method described in App. A.1.4. The first set of 120 data cubes contains sources covering the range of integrated flux densities from $1\,\mathrm{Jy\,km\,s^{-1}}$ to $30\,\mathrm{Jy\,km\,s^{-1}}$. This range covers the transition from 0 % to 100 % completeness for EBHIS. The second set of data cubes covers the integrated flux densities from $30\,\mathrm{Jy\,km\,s^{-1}}$ to $300\,\mathrm{Jy\,km\,s^{-1}}$. This set is useful to ensure that the pipeline behaves as expected from high SNR sources. In both data sets, the projected rotational velocity ranges between $30\,\mathrm{km\,s^{-1}}$ to $600\,\mathrm{km\,s^{-1}}$. Due to the turbulent motion included in the profile generation code, the actual w_{50} profile widths are up to $630\,\mathrm{km\,s^{-1}}$. The large number of sources is necessary to determine the transition from 0 % to 100 % with approximately 5σ significance. In creating this data set, we include three simplifying assumptions.

First, we only simulate unresolved sources. For single-dish H I surveys like EBHIS, the majority of sources are unresolved or can be treated as such. We only expect significantly resolved sources in the very local universe. Such sources are bright and easy to parametrize. We therefore limit our simulations to the more common and difficult case of unresolved sources.

Second, the sources are placed on a grid to avoid source confusion, i.e., multiple galaxies detected as one source because of angular and spectral proximity. For shallow spectroscopic surveys, Zwaan et al. (2003) estimate that the impact on derived cosmological parameters is negligible. Duffy et al. (2012) show that for the WALLABY H I All-Sky Survey (WALLABY, Koribalski 2012), confusion is even lower because of the increased angular resolution. We therefore do not consider this assumption a likely source of errors.

Third, we do not simulate the impact of non-flat baselines, RFI or other artifacts on the data. It is difficult to simulate a "typical" case data artifacts that would not bias the simulations or require vastly more data. Instead of simulating artifacts at this stage of the simulations, we include false-positives from EBHIS data in the performance evaluation of the classification stage discussed in Chapter 4. By not including artifacts in the simulations directly, we do not simulate the case in which a source is not detected because of corrupted data, which is a potential bias in the analysis.

We perform source-finding on the simulated data cubes as described in Sec. 2.4. After parametrization of all candidates, we match all candidates to sources in the input catalog. To count a given simulated source as detected, we require a spatial and spectral match. For spatial matching we require agreement of the angular position within one FWHM of the beam. For spectral matching we require the measured radial velocity to lie within the interval $v \pm w_{50}/2$ of the simulated source.

For bright sources, the 2D-1D wavelet reconstruction is prone to artifacts in the vicinity of the source. These are also treated as source candidates. If this occurs, we select the source with the highest measured integrated flux density to be the best match. We compile all measured parameters and the input values

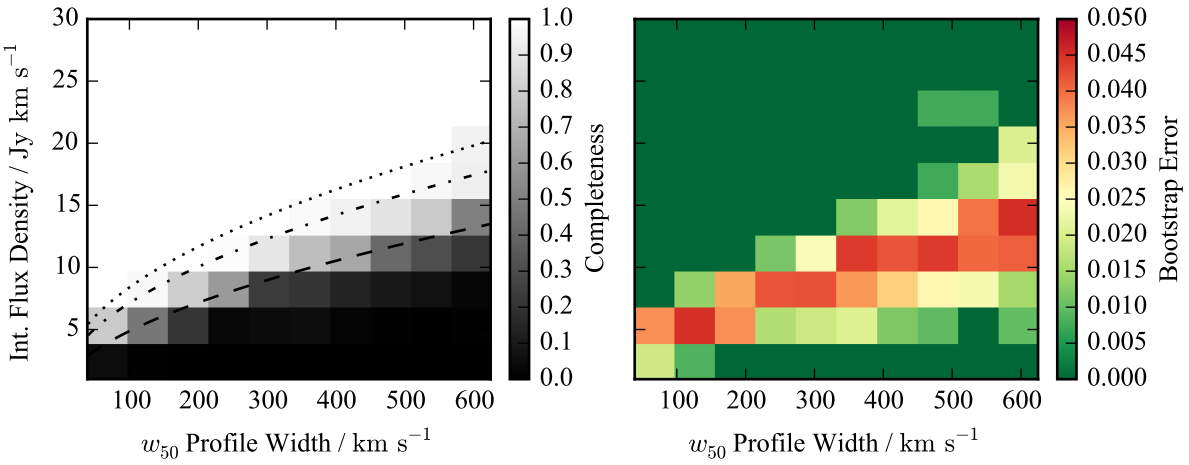


Figure 3.4: Bivariate completeness level for EBHIS derived from simulations. **Left**: Measured completeness level as a function of integrated flux density and w_{50} profile width. The dashed, dash-dotted and dotted lines indicate the 50%, 95% and 90% completeness level as derived from our model. **Right**: Bootstrap errors for the measured completeness level from 1000 bootstrap samples.

of the simulation into a single database which we use to investigate the expected completeness and parameter accuracy of EBHIS.

3.3.1 Completeness

The completeness of a survey describes the probability to detect a source as a function of its parameters. For photometric surveys, this can be expressed as a flux or magnitude limit. In spectroscopic surveys, the source signal is spread over a finite number of channels. The integrated SNR of a source with a given integrated flux density therefore decreases proportional to $N^{-0.5}$, where N is the number of channels it occupies. We therefore derive the completeness as a function of both integrated flux density and w_{50} profile width.

For statistical analysis it is advantageous to have an analytic description of the completeness level. We model the completeness of EBHIS using the function

$$C\left(F_{\mathrm{int}}\left[\mathrm{Jy\,km\,s^{-1}}\right], w_{50}\left[\mathrm{km\,s^{-1}}\right]\right) = \left[\exp\left(-\frac{F_{\mathrm{int}} - a_1\, w_{50}^{a_2}}{a_3\, w_{50}^{a_4}}\right) + 1\right]^{-1} \quad . \tag{3.8}$$

Here, the parameters a_1 and a_2 determine the shift of the completeness with increasing w_{50} profile width. The parameters a_3 and a_4 account for a broadening of the transition from 0% to 100% completeness. Wider profiles occupy more spectral channels and are therefore harder to detect as compared to narrow profiles given the same integrated H I flux density. Using standard least-squares fitting we find $a_1 = 0.37 \pm 0.02$, $a_2 = 0.56 \pm 0.01$, $a_3 = 0.15 \pm 0.04$ and $a_4 = 0.35 \pm 0.05$.

In Fig. 3.4 we show the measured completeness with corresponding bootstrap errors. We also show the 50%, 95% and 99% completeness level as derived from our model.

3.3.2 Parameter Accuracy

We quantify the parametrization uncertainties of our pipeline by comparing the absolute errors of the measured with the input parameters of the simulated sources. We define the absolute error in a given

parameter X to be $\Delta X = X_{\mathrm{measured}} - X_{\mathrm{true}}$. Therefore, positive errors point to an overestimation of the parameter. We use all detections from both data sets for which the completeness is above 90%. The cut in completeness is performed to avoid biasing our results due to increased uncertainties expected for sources that are barely detected. Since most sources detected in EBHIS will have a higher individual completeness, this cut does not diminish the validity of our analysis. For our simulations the cut at 90% completeness removes exactly 1000 of 20 710 detections (4.8 %).

The parameters under investigation are the peak flux density of the line profile P, the angular and spectral position v, the profile width w_{50}, and the integrated flux density F. Since P and F can be measured from both the peak and integrated spectrum, we investigate both cases. To differentiate between the two, we add a subscript "peak" or "int", respectively. Figure 3.5 shows the uncertainties in the measured parameters as a function of three input parameters, which we denote with a superscript I: the integrated flux density F^I, the profile width w_{50}^I, and the peak flux density P^I. Each panel shows the individual detections as grey points. All detections are split into ten bins of equal detections. We show the median and the region containing 95% of the errors for each bin as black crosses and error bars, respectively. We investigate the correlation between the uncertainties in the measured parameters and the input parameters in the following paragraphs.

Peak Flux Density For a given source, the observed peak flux density P is determined by the angular and spectral resolution of the observation. It has therefore no direct physical meaning. It is nonetheless an interesting parameter, as the uncertainty in the measurement of other parameters is correlated with the peak flux density.

From Fig. 3.5 we can see that the uncertainty in both P_{peak} and P_{int} correlates best with the true peak flux of the source P^I. This is the expected behavior, as P is measured from the largest value in a source spectrum. Additionally, P shows a pronounced positive bias for lower P^I. As we are taking the largest value in the source spectrum to be P, this value is boosted by the noise in the spectrum. For very large P^I, the 95% interval for ΔP_{peak} approaches ±46 mJy, which is twice the noise level of the simulations. Assuming Gaussian noise, this is the expected 2σ region for a noisy measurement. For ΔP_{int}, the situation is more complicated, as the noise level of the integrated spectrum depends on the aperture of the source. Since the simulated profiles are generated to uniformly cover w_{50} and F, the uncertainty in the measured peak flux density is uncorrelated with w_{50}. The uncertainty is slightly correlated with F, since fainter sources have lower peak flux densities.

Angular Position The uncertainty in the angular position is correlated with the integrated flux density F^I and the peak flux density P^I. The latter correlation is due to the correlation of F^I and P^I. As we measure the angular position by fitting a Gaussian to the velocity-integrated map of the source, the correlation with F^I is expected.

Redshift The uncertainty in the measured redshift v correlates with all three input parameters, but the correlation with P^I is most pronounced. Since we measure v as the midpoint between the two points in the source profile where its flux density first rises above 50% of its peak flux density, the uncertainties in w_{50} and v are correlated.

Profile Width As explained for the redshift measurement, the uncertainty in the w_{50} profile width correlates strongest with P^I. The trend observed in Fig. 3.5 is already evident from the simulations conducted in Sec. 3.2.3: Because of the width maximizing algorithm, the profile widths are more likely to be overestimated with decreasing P^I. For very low peak SNR the profile widths are sometimes

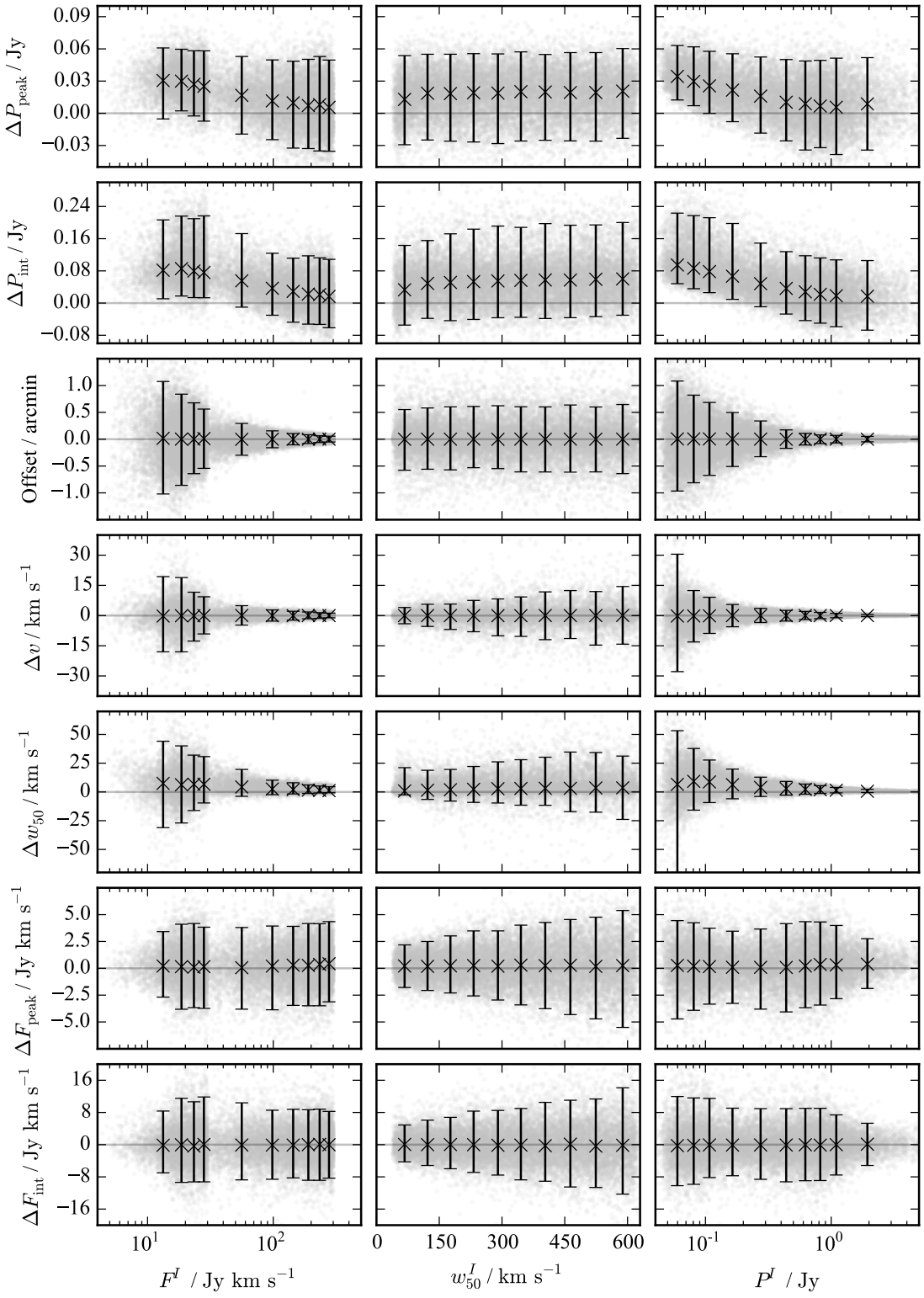


Figure 3.5: Dependence of pipeline parametrization uncertainties on the input parameters of the sources. From top to bottom, the rows show the uncertainties for the peak flux density measured from both peak and integrated spectrum $\Delta P_{\mathrm{peak,int}}$, the angular offset, the systemic velocity Δv, the profile width Δw_{50}, and the integrated flux density measured from both peak and integrated spectrum $\Delta F_{\mathrm{peak,int}}$. The columns represent the input parameters integrated flux density F^I, the profile width w_{50}^I, and the peak flux density P^I. For visualization, the measurements are divided in ten bins of equal counts. The error bars indicate the range of 95% of the values and the crosses are the median in each bin. The individual sources are shown as light grey points.

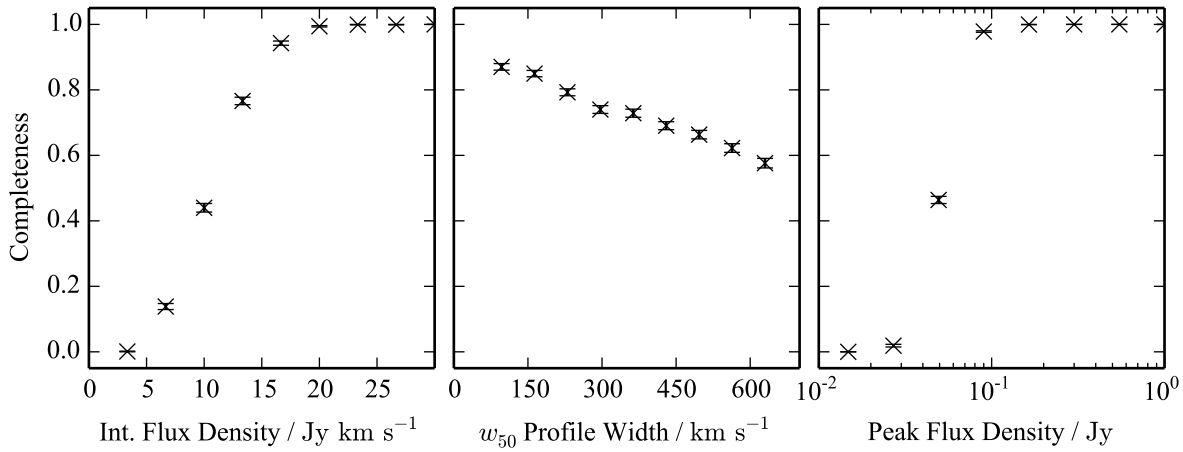


Figure 3.6: Univariate completeness levels for EBHIS derived from simulations. For each of the parameters, the crosses and error bars indicate the bootstrap mean and error in each bin from 1 000 bootstrap samples.

drastically underestimated, as the flanks of the profile become indistinguishable from the noise in the spectrum.

Integrated Flux Density Since the integrated flux density F is measured by summing up the flux in either spectral channels for F_{peak} or voxels contained in the source mask for F_{int}, the uncertainty is expected to scale with the square root of independent values summed up. For F_{peak} the channels are assumed to be independent, since the data are binned by a factor of eight for the extragalactic survey (see Sec. 1.3). For F_{int} the flux is summed over the volume created by the spatial and spectral mask. Since spatial pixels are correlated, the number of independent values $N_{\mathrm{ang}}\,N_{\mathrm{spec}}\,N_{\mathrm{kernel}}^{-1}$, where N_{ang} and N_{spec} are the number of pixels or channels in the spatial and spectral mask, respectively. N_{kernel} is the number of pixels contained in one gridding kernel.

Since all simulated sources are point sources, the spatial masks have similar sizes. Therefore the integration volume for both F_{peak} and F_{int} is determined by the spectral range which correlates with the w_{50} profile width. This is evident from Fig. 3.5, where the uncertainties in both F_{peak} and F_{int} only correlate significantly with w_{50}^{I}.

3.3.3 Comparison with HIPASS

To compare our fully automated pipeline to manual parametrization, we compare the results from the preceding section to the parameter uncertainties derived for HIPASS in Zwaan et al. (2004, hereafter Z04). Similar to our approach, Z04 use synthetic sources to determine their completeness and parametrization accuracy.

A major difference to our simulations is that their simulated sources uniformly sample peak flux density and profile width. By sampling peak flux density and profile width, they systematically underrepresent galaxies with high integrated flux densities and narrow profiles and therefore overestimate their completeness for sources with large profile widths. This biasing is immediately evident from Fig. 2 in Z04 where their completeness level for profile widths below 200 km s^{-1} drops until it reaches less than 0.5 for the narrowest profile widths. For comparison, we derive univariate completeness levels for integrated flux density, w_{50} profile width and peak flux density from our simulations and plot them in Fig. 3.6. While the overall shape of the completeness level for integrated and peak flux density are

comparable to the ones derived for HIPASS, the completeness in w_{50} profile width accurately reflects that wider profiles are more difficult to detect.

Another difference is the way, the uncertainties for HIPASS are stated. Z04 assumes a Gaussian distribution for all uncertainties with zero mean and continuously changing variance σ. For example, the 1σ uncertainty for integrated flux density is given as $\sigma_F = 0.5\sqrt{F}$. Our simulations show that the uncertainties in integrated flux density correlate most closely with the w_{50} profile width or the number of statistically independent data points integrated over. To compare our uncertainties to HIPASS, we compare their 2σ uncertainties to our 95% uncertainty intervals. Where we measure parameters from both the peak and integrated spectrum, we limit our analysis to the parameters derived from the peak spectrum, as the uncertainties derived by Z04 assume unresolved sources.

Peak Flux Density For the measured peak flux density, Z04 adopt a constant 95% uncertainty interval of ±22 mJy, which is slightly less than twice the typical noise level in their data. They observe a slight bias of 5 mJy which they attribute to the selection bias. In Sec. 3.3.2 we show that our automated parametrization pipeline has a strong positive bias which is expected from our measurement method. Meyer et al. (2004), who describe the parametrization for HIPASS, claim to use the same method as implemented in our pipeline. We can therefore not explain the difference in the behavior for lower peak flux densities. In agreement with the results found by Z04, our 95% uncertainty interval does converge to twice the noise level in our simulations for increasing integrated flux densities.

Angular Position Z04 model the uncertainty in the angular position of a source as a function of the integrated flux density, which agrees with our finding from the previous section. The equivalent 95% uncertainty interval ranges between ±5′ for sources at the HIPASS detection limit and ±2′ for the brightest sources in their simulations. Our automated pipeline achieves higher accuracy for the whole range of sources detected by EBHIS. Assuming that the positional accuracy depends linearly on the angular resolution, this difference can still not only be accounted for by the different angular resolutions of the respective surveys, 15.5′ and 10.5′, respectively. The measurement of the angular position of a source using iteratively re-weighted Gaussian fitting performs very well.

Redshift In agreement with our finding from the previous section, Z04 model the uncertainty in the measured redshift as a function of the peak flux density. Their 95% uncertainty interval ranges between $\pm 30\,\mathrm{km\,s^{-1}}$ for sources at HIPASS detection limit and $\pm 10\,\mathrm{km\,s^{-1}}$ for the brightest sources in their sample. Except for the faintest sources detected by EBHIS, our automated pipeline outperforms the manual parametrization for HIPASS.

Profile Width For the measured w_{50} profile width, Z04 adopt a constant 95% uncertainty interval of $\pm 15\,\mathrm{km\,s^{-1}}$. They do not observe a clear dependence of the uncertainty on the source parameters, which disagrees with our finding from the previous section. Z04 observe that for roughly one third of the sources, the adopted uncertainty is too small and can be better modeled by an 95% uncertainty interval of $\pm 50\,\mathrm{km\,s^{-1}}$. We suspect that the unclear dependence on the input parameters stems from the small size of their simulation. For comparison, Z04 use 1000 synthetic sources, whereas our simulations include 24 000 synthetic sources. The accuracy of the two surveys is approximately equal at the respective completeness limits of HIPASS and EBHIS. For brighter sources, our pipeline outperforms the manual parametrization. For fainter sources, our estimated uncertainty exceeds the HIPASS uncertainty interval of $\pm 15\,\mathrm{km\,s^{-1}}$ but agrees with the wider uncertainty interval of $\pm 50\,\mathrm{km\,s^{-1}}$.

Integrated Flux Density Z04 model the uncertainty in the integrated flux density as a function of the integrated flux density itself. In the previous section, we have shown that the uncertainty in the integrated flux density is roughly independent of itself but instead correlates with the w_{50} profile width. At the 99% completeness level of HIPASS they derive a 95% uncertainty interval of $\pm 3\,\mathrm{Jy\,km\,s^{-1}}$. For sources with a w_{50} profile width less than $200\,\mathrm{km\,s^{-1}}$, our automated pipeline exceeds the accuracy of the manual parametrization. Since most galaxies have w_{50} profile widths less than $200\,\mathrm{km\,s^{-1}}$ (Zwaan et al. 2010; Papastergis et al. 2011), the average uncertainty for unresolved galaxies in EBHIS is expected to be smaller than in HIPASS.

3.4 Conclusions

In the preceding chapter we introduce the algorithms employed in the automated parametrization pipeline for EBHIS. We argue that the current approach — manual parametrization — is not feasible for the upcoming surveys with SKA and, more imminent, the pathfinder surveys with ASKAP and WSRT/Apertif.

We use a novel approach to spectroscopic baseline estimation by using robust smoothing splines. The algorithm is shown to outperform baseline fitting using polynomials and has no free parameters, which makes it an ideal algorithm for a fully automated pipeline. We furthermore augment the classical method of measuring the profile width at 50% of its peak flux by using bilateral filtering. Since the bilateral filter preserves sharp edges in the spectrum, it avoids lowering the spectral resolution of high SNR profiles.

The automated nature of our parametrization pipeline allows us to perform a complete end-to-end simulation of 24 000 sources in 240 data cubes. This is a improvement of more than an order magnitude over previous approaches, e.g. Zwaan et al. (2004) for HIPASS. Using these simulations, we estimate both significant completeness levels and 95% uncertainty intervals for the measured parameters. To show the viability of our pipeline, we compare our results to the parametrization accuracy achieved for HIPASS. Our pipeline yields comparable results and even outperforms the manual parametrization for some parameters, despite the fact that EBHIS has a lower sensitivity than HIPASS.

CHAPTER 4

Classification

The results from this Chapter are published in part in

- Flöer, L., Winkel, B., & Kerp, J. 2014, A&A, 569, A101

During manual parametrization, an astronomer implicitly performs the task of determining whether a given source candidate is a genuine detection or a false positive caused by the various effects described in earlier chapters. An expert astronomer can arrive at an decision by using a large amount of information: numerical parameters, such as integrated flux density or profile width, but also the shape of the source candidate in an velocity-integrated map or a position-velocity diagram. Having an astronomer look at every source candidate gives a highly reliable classification, but requires a lot of man-hours. Additionally, the parametrization accuracy of an individual astronomer will vary over the course of a day or week. One way to combat this is the approach of Meyer et al. (2004) who use three different astronomers to inspect over 140 000 source candidates from the H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001). This approach further increases the required man-hours to process all source candidates and also prohibits reprocessing the data. To automate this task reliably is one of the greatest challenges in implementing a fully automated pipeline.

In terms of absolute numbers, data sets from future H I surveys can be expected to produce a larger amount of false positives. The main reasons for this are the vastly increased data volume, the automation of the data reduction process and the increased sensitivity. Since the majority of false positives are caused by defects in the data such as radio-frequency interference (RFI), they scale with data volume instead of survey volume. Since future H I surveys are expected to cover redshifts up to $z \approx 0.2$, they will observe at frequencies that are known to be more contaminated with RFI (see, for example, Fernández et al. 2013). Due to the necessary automation of the data reduction process for the Square Kilometre Array (SKA) and its pathfinder instruments and surveys, the probability of having some defect in the data will increase, as manual data reduction with careful flagging is no longer an option (Hotan et al. 2014). Lastly, the increased sensitivity will detect more faint RFI and amplify systematic errors in the data reduction, as is noted by Verheijen et al. (2010) in an ultra-deep H I survey with Westerbork Synthesis Radio Telescope (WSRT).

The task of deciding whether a given source candidate is a real, astronomical object given its parameters, can be formulated as a supervised classification task. For the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011) we solve this problem using an artificial neural network (ANN). In this chapter, we first introduce the techniques we use for training and benchmarking our ANN. This overview is limited to the techniques we use and does not aim to be a complete discussion of ANNs. We then discuss how we use the output from the parametrization pipeline to classify EBHIS source candidates. Here, we

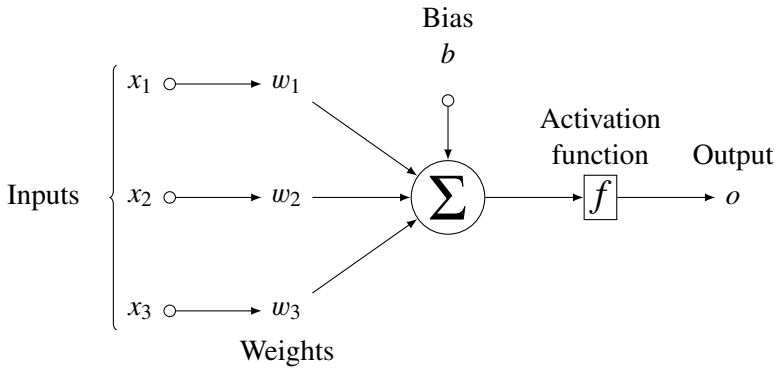


Figure 4.1: Schematic drawing of the inputs to a single neuron in an ANN. The inputs are the values in the input vector x, the weights $w_{1\ldots3}$ are one row of the weight matrix $\mathbf{W}$. Adapted from the answer of Gonzalo Medina to `http://tex.stackexchange.com/questions/132444`.

classify two different data sets. We use the simulated sources from Chapter 3 together with a sample of false positives from real data to show the general viability of our approach. Thereafter, we train ANN on real data only. Here we investigate different methods of training ANN from a highly imbalanced data set. We investigate the impact on completeness and reliability for both cases and close the chapter with a discussion of the results.

4.1 Artificial Neural Networks

The name artificial neural network stems from their structural similarity to the way the human brain processes information. The most commonly used type of ANN is the so-called feed-forward neural network. The network consists of multiple layers of neurons. Each neuron combines the outputs from all neurons of the preceding layer by applying two operations. The first operation is a weighted sum of all outputs of the preceding layer. The second operation is the addition of a bias term and the application of the activation function of the neuron. The resulting value is the output of the neuron. ANN like these are called fully connected, since each neuron takes all outputs of the preceding layer as input values. Mathematically such networks are described by a matrix-vector multiplication and element-wise application of the activation function:

$$o(x) = f\left(x\mathbf{W} + b\right) \ . \tag{4.1}$$

Here, x is the input vector, $\mathbf{W}$ is the weight matrix, b is the bias vector, and f is the activation function. Multi-layer networks are built by replacing x with the output vector o of the preceding layer. The activation function is also called the activation non-linearity, as its presence makes $o(x)$ a non-linear function of its argument. The presence of this non-linearity further allows a neural network to perform more complex regression or classification tasks. A common choice for f is the hyperbolic tangent function (LeCun et al. 1998b). The flow of signal for a single neuron in an ANN is shown in Fig. 4.1. When performing classification using ANNs the activation function last layer, also known as output layer, is often replaced by the softmax function. The value for the ith output neuron in a output layer of N neurons is calculated by

$$f_i(x) = \mathrm{softmax}_i(x) = \frac{e^{x_i}}{\sum_{j=0}^{N} e^{x_j}} \ . \tag{4.2}$$

This activation function ensures that the sum of the individual outputs in the output layer equals one. When using ANNs as classifiers, each class is represented by its own output neuron. Using the softmax function allows a probabilistic interpretation of the ANN output, that is, the value of the ith output gives the probability that a given example belongs to class i.

In the following sections, we describe the techniques used to train ANNs for the purpose of candidate classification for EBHIS.

4.1.1 Training

The ultimate goal for our ANN is to predict probabilities for a given source candidate to belong to a certain class. A source candidate can either be a true positive, that is, an object of astrophysical origin, or a false positive, meaning the candidate is caused by some kind of data defect. To achieve this, the ANN has to be trained on a training data set where we know the true class for each sample. During training, the weights $\mathbf{W}$ and biases b for each layer are updated to minimize an objective function which provides a measure of the classification accuracy. A simple measure of classification accuracy is the number of classification mistakes made, which is called the zero-one loss. As we show below, the objective function needs to be continuously differentiable which the zero-one loss does not fulfill. Instead, we maximize the probability that the ANN predicts the true labels for each training example from the data set $\mathcal{D}$. This defines the log-likelihood $\mathcal{L}$

$$\mathcal{L}\left(\theta = \{\mathbf{W}_0, \ldots, \mathbf{W}_n, b_0, \ldots, b_n\}, \mathcal{D}\right) = \sum_i \log P\left(Y = y_i \mid x_i, \theta\right) \ . \tag{4.3}$$

Here, $P\left(Y = y_i \mid x_i, \theta\right)$ denotes the probability that the predicted label Y for training example x_i equals the true label y_i, given the current parameters θ of the ANN. Note that when using the softmax function as the output layer, this term is given by the value of the output neuron representing the true class of the training example. During the training process, we minimize the negative log-likelihood $\ell(\theta, \mathcal{D}) = -\mathcal{L}(\theta, \mathcal{D})$ using back-propagation of errors with stochastic gradient descent, which is introduced in the next section.

Stochastic Gradient Descent

The most common way of training an ANN is gradient descent: After each iteration $k = 1, \ldots, N$ over the training set $\mathcal{D}$, also known as epoch, the gradient of the negative log-likelihood with respect to each of the parameters θ_i is calculated and the value of each parameter θ_i^k is adjusted by an amount ϵ_k in the direction of the gradient

$$\theta_i^{k+1} = \theta_i^k - \epsilon_k \frac{\partial \ell(\theta^k, \mathcal{D})}{\partial \theta_i} \ . \tag{4.4}$$

ϵ_k is also known as the learning rate and is either constant for all epochs or can be reduced during training. For multi-layered ANN, the gradients $\frac{\partial \ell(\theta, \mathcal{D})}{\partial \theta_i}$ are calculated recursively using the chain rule of differentiation. Since this differentiation starts from the output layer of the ANN, this process is called back-propagation of errors. The necessity of calculating these gradients also justifies the choice of objective function as it has to be continuously differentiable.

There are two variations of this training algorithm: stochastic gradient descent (SGD) and mini-batch stochastic gradient descent (MSGD). In SGD, the parameter updates in Eq. 4.4 are performed after every individual training example. This leads to a nosier estimate of the gradient, but speeds up training significantly. Especially for very large training data sets, it obviates the need to present all training

examples to the network for a weight update. Furthermore, the noisy nature of the procedure can help to avoid local minima in the gradient of the objective function, which leads to better solutions (LeCun et al. 1998b).

MSGD slightly modifies ordinary SGD in that the weight updates are not performed for individual examples from the training data set but for a small batch containing between 10 and 100 examples. This modification reduces the noise in the gradient estimation and lowers the computational burden as compared to SGD.

Before starting the training, the weights $\mathbf{W}$ and biases b have to be initialized. Glorot & Bengio (2010) show empirically that for the hyperbolic tangent activation function the weights should be initialized by uniformly sampling the interval

$$\left[-\sqrt{\frac{6}{n_{\text{in}}+n_{\text{out}}}}, +\sqrt{\frac{6}{n_{\text{in}}+n_{\text{out}}}}\right] \tag{4.5}$$

where n_{in} and n_{out} are the number of incoming and outgoing connections in the current layer, respectively. The biases are initialized to zero.

Momentum, Early Stopping and Regularization

A common practice to accelerate the training process is to use momentum. When momentum is used, the parameter update from Eq. 4.4 is modified to

$$\begin{aligned} \bar{g}_i^0 &= 0 \\ \bar{g}_i^k &= (1-\eta)\frac{\partial \ell(\theta^k, \mathcal{D})}{\partial \theta_i} + \eta \bar{g}_i^{k-1} \\ \theta_i^{k+1} &= \theta_i^k - \epsilon_k \bar{g}_i^k. \end{aligned} \tag{4.6}$$

Using this update, the effective parameter update is proportional to a moving average of the gradients, with η controlling the length of this average. Momentum has two main effects. On the one hand, it accelerates gradient based training by smoothing out the error surface the algorithm is moving on, avoiding small, local minima. On the other hand, in areas where the error surface is highly convex, it dampens oscillations around the minimum, enabling faster convergence (Bengio 2012).

A common problem with ANNs is their tendency to overfit, that is, they memorize the examples in the training data instead of learning general properties of the data. This decreases their capability of classifying new data. To avoid this overfitting, we employ early stopping. We monitor the performance of the ANN using a validation data set which is not used for training. At the beginning of the training process both the training and validation error decrease. Once the ANN starts overfitting the training data, the validation error increases again. Following Bergstra & Bengio (2012), we stop the training if at epoch k, the best validation error is measured before epoch $k/2$ but at least 100 epochs are completed. Once the training is stopped, we revert the parameters of the ANN to the parameters having the smallest validation error.

When the weights $\mathbf{W}$ become large, the layers of the ANN lose the ability to efficiently communicate information to the next layer, as the activation functions become saturated. Furthermore, the gradient of the activation function $f(x)$ is small for $x \to \pm\infty$ which slows down the training process. There are multiple regularization techniques that prevent this saturation of the weights. These techniques effectively limit the ability of the ANN to learn from the training data, which also reduces their tendency to overfit.

The simplest forms of regularization are L1 or L2 regularization, which add a term $\lambda \sum_i |\theta_i|$ or $\lambda \sum_i \theta_i^2$ to the objective function, respectively. They can also be used together. λ is the regularization parameter which determines the influence of the regularization term on the objective function. Both L1 and L2 regularization penalize large weights, but have different effects. L2 regularization only keeps the weights small. Since L1 regularization penalizes based on the absolute value of the weights, it has the effect of driving to zero unimportant parameters, for example, connections between layers that do not contribute significantly to the classification accuracy (Bengio 2012).

Hinton et al. (2012) introduce another form of regularization called *dropout*. For each training example presented to the ANN, each connection between the neurons has a probability $p = 0.5$ of being deactivated, that is, to drop out of the network. Since individual neurons can not count on the other neurons being active, they learn to perform a more general function rather than a specific one in concert with all other neurons in a given layer. Another way of interpreting *dropout* is in the context of model averaging: Since each training sample is presented to a different network configuration that share weights with each other, we are effectively training $O(2^n)$ sub-networks and average their output. In Hinton et al. (2012), *dropout* is shown to be effective in training very large ANN for image classification tasks and ANN training using *dropout* perform better than many previously trained networks using other forms of regularization. For small networks, Slatton (2014) shows that *dropout* can be detrimental to the performance of the network.

4.1.2 Hyperparameter Optimization

Apart from the optimization of the parameters θ, there are a number of so-called hyperparameters involved when training an ANN. Examples of hyperparameters are the number of layers and neurons in the ANN and parameters involved in the training of the ANN like the learning rate ϵ_k, regularization parameter λ, or momentum η, introduced in the previous section. Although there are empirical values for most parameters, the optimum parameter set is problem-specific.

A way of optimizing the hyperparameters is to perform a grid search. For each hyperparameter, a set of proposal values is chosen and the ANN is trained with all possible combinations of parameter values. The performance of a given parameter set is often determined using n-fold cross-validation. In n-fold cross-validation, the training data set is split into n parts, or folds. During training, the ANN is trained on $n - 1$ folds and the left out fold is used to determine the validation error. This process is repeated n times, each time leaving out another fold. The final cross-validation score of a given parameter set is often chosen to be the mean of the score of the individual folds. If the data set is strongly imbalanced, that is, there are many more examples of one class than the others, it is important to have the same class balance in all folds.

Depending on the number of hyperparameters and complexity of the ANN, a grid search is computationally expensive and there is no guarantee of finding the optimal parameter configuration. To get good results, usually a sequence of iteratively refined grid searches is necessary. Bergstra & Bengio (2012) show that it is more efficient to sample the parameter space spanned by all hyperparameters by drawing the proposal parameters from a random distribution for each parameter. This method is especially efficient if not all hyperparameters are equally important. The sampling process can be optimized by using distributions that include a priori knowledge of the typical range of the parameter.

4.2 Automated Classification for the Effelsberg-Bonn HI Survey

We implement our ANN with the `theano` package (Bergstra et al. 2010; Bastien et al. 2012). `theano` uses the `Python` language to define expression graphs that represent the ANN and operations required

for training which are then optimized and compiled to machine code for fast execution. To classify the source candidates from the automated source-finding and parametrization, we train an ANN using the techniques described in the previous section.

4.2.1 Feature Selection

We design our ANN to estimate class membership probabilities based on a number of features, derived from the parameters and data products measured by our automated parametrization pipeline. Instead of using derived parameters, it is conceivable to also use the spectrum and velocity-integrated map of each source candidate directly. Another approach might be to directly use a section of the data cube, containing only the source candidate. Indeed, a variant of the ANN, the so-called convolutional network, is commonly used in machine vision to classify images or recognize objects in them (for a recent example see Farabet et al. 2013). Such networks have also been used to classify images of galaxies from the Sloan Digital Sky Survey (SDSS, York et al. 2000) by using the data set created by GalaxyZoo[1] (Lintott et al. 2008). Nonetheless, we choose to use the derived parameters for classification. Using images or even data cubes requires vastly more complicated network architectures and increases the time required for training and hyperparameter optimization by many orders of magnitude. Convolutional networks are typically trained on small images with tens of pixels on each side, i.e., the first layer requires hundreds of inputs. In comparison, our approach requires only 53 inputs. Furthermore, it is not clear how one would fuse the information contained in the spectrum and velocity-integrated map or whether such an approach would yield more accurate results. The research into this topic is beyond the scope of this thesis.

For each source candidate we compile a feature vector containing the following parameters:

- the peak and integrated flux densities from both the peak and integrated spectrum (parameters $P_{\mathrm{int,peak}}, F_{\mathrm{int,peak}}$ discussed in Sec. 3.3.2),
- the peak and integrated signal-to-noise ratio (SNR) from both the peak and integrated spectrum,
- the measured w_{50} profile width from both the peak and integrated spectrum,
- the amplitude, major axis length and ellipticity of both Gaussian fits
- the cumulative and differential brightness profile,
- the noise in the velocity-integrated map,
- and the mean and standard deviation of the spectral baseline.

Since the two major source parameters obtained from single-dish H I surveys are the integrated flux density and profile width, they are a natural choice to base the classification on.

To have a measure of detection significance, we calculate the integrated and peak SNR. For both types of spectra, we measure the noise level in a range of 100 channels around the spectral mask of the source. For the peak SNR, we divide the peak of the line profile by this noise level. For the integrated SNR, we multiply the noise level by the square-root of the channels in the source mask and divide the measured total flux by this value.

The Gaussian fits are included to have a measure of the shape and concentration of the detections. Since the vast majority of sources in single-dish H I surveys are expected to be unresolved or marginally

[1] `http://www.galaxyzoo.org`

resolved, the major axis of the Gaussian fits is a good indicator whether the source candidate is real. Especially the non-iterative Gaussian fit gives a good measure for the extent of the source.

The cumulative and differential brightness profile provide a further, non-parametric measure of the extent of the source. Each profile measures the integrated flux density at a fixed set of radii between 1 pixels to 6 pixels. The points of the differential profile only represent the integrated flux density in concentric annuli around the source, whereas the nth point of cumulative profile is the integral of the n inner annuli.

The noise level in the velocity-integrated map is a good measure of how clean the data are overall. There are some cases where solar interference degrades the quality of the data which often generates false positives. In these cases, the noise level in the velocity-integrated map is exceptionally high.

The mean and standard deviation of the spectral baseline are a good measure to detect the residual effects of continuum sources. Even though the total power level of these sources is subtracted during data reduction, strong continuum sources still cause artifacts in the data. Especially glitches in the spectral baseline can mimic source candidates that have a reasonable profile width, integrated flux density and shape in the velocity integrated map. Since these sources cause strong variations of the baseline level over the whole band, they can be detected by characterizing the fitted baseline. We have found the mean and standard deviation of the baseline level to be a good description of the quality of the baseline.

Since the parameters of the ANN are initialized to be uniformly distributed around zero (see Sec. 4.1.1) it is advantageous to standardize the features presented to the ANN during training (LeCun et al. 1998b). This is achieved by subtracting the mean from each feature and divide each feature by its standard deviation. For our ANN, we use the median and median absolute deviation (MAD) as a replacement for mean and standard deviation, respectively, to be more robust against outliers. As a side effect of this scaling, outliers have extremely large values in the features. This poses another problem during training, since these few outliers dominate the estimate of the gradient in Eq. 4.4. We therefore transform the features f using

$$f' = \arctan\left(\frac{1}{5}\frac{f-\hat{f}}{\hat{\sigma}_f}\right) \tag{4.7}$$

where $\hat{f}$ and $\hat{\sigma}_f$ are the median and MAD of the respective feature. Using the arctan function compresses parameter distributions with large extreme values to the interval $[-\pi/2, \pi/2]$. Since $\arctan(x) \approx x$ on the interval $[-0.5, 0.5]$ and by using the scaling factor $1/5$, we ensure that features falling the range $\left[-2.5\sigma_f, 2.5\sigma_f\right]$ are not modified and only outliers are compressed.

4.2.2 Simulated Data

To test our approach, we first create a data set from the parametrized simulated sources used in Sec. 3.3. Since our simulations do not include real-data artifacts, we run our source-finding and parametrization pipeline on the real EBHIS data as described in Sections 2.4 and 3.2. We then inspect a random sample of detection candidates, classify them by eye, and verify potential true positives with the astronomical on-line data bases Simbad[2] and NED[3]. To exclude sources confused with the Milky Way H I emission or false positives caused by residual Milky Way H I emission, we limit this search to radial velocities in excess of 200 km s^{-1}.

With this process we compile a training data set consisting of 38 102 examples, 20 710 of which are true positives. For each example we compile the features as described above and shuffle them. This

[2] http://simbad.u-strasbg.fr/simbad/
[3] http://ned.ipac.caltech.edu

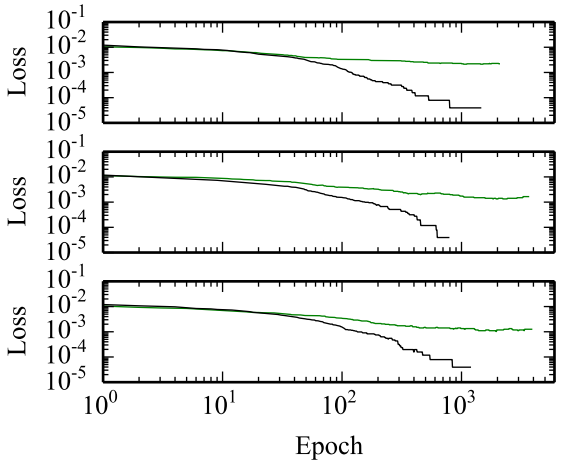


Figure 4.2: Training and validation loss curves of the best-performing ANN for the simulated data set. Each panel shows the curves for one of the folds used in cross-validation. The black and green curve show the fraction of mis-classified sources for the training and test data, respectively.

Parameter	Value
Layers	1
Hidden neurons	39
ϵ_0	0.083
τ	9 861
η	0.065
λ L1	3.2×10^{-5}
λ L2	0.0
Accuracy	0.9985

Table 4.1: Hyperparameters of the best-performing ANN for the task of simulated source classification as determined by randomized optimization.

is done to ensure equal class balance between individual mini-batches and cross-validation folds. To optimize the hyperparameters of this network, we perform a random search of the parameter space as introduced in Sec. 4.1.2. We use a three-fold cross-validation scheme and compare the different networks on the average of the minimum validation error across the three folds. The validation error is calculated after each epoch. For each hyperparameter, we draw proposal values from appropriate random distributions. The following list summarizes the hyperparameters we optimize and how proposal values are generated.

- Each ANN has either one, two, or three layers with equal probability.
- The number of neurons in each layer is drawn from a discrete uniform distribution in the interval $[30, 100]$. We use a rather narrow range of values to speed up training of individual networks. From other experiments we know that we do not need a large number of hidden neurons to achieve high classification accuracy.
- The initial learning rate ϵ_0 is drawn logarithmic[4] on the interval $\left[10^{-3}, 10^{-1}\right]$.
- For each epoch k, we use the learning rate $\epsilon_k = \epsilon_0 \tau (k + \tau)^{-1}$. During optimization, we draw τ from a uniform distribution on the interval $[100, 10\,000]$.
- The momentum parameter η is drawn for a uniform distribution on the interval $[0.01, 1]$.
- With 50% probability each[5], we apply L2 or L1 regularization to the objective function with a regularization parameter λ drawn logarithmic on the interval $\left[10^{-6}, 10^{-2}\right]$.

[4] With the term "logarithmic on the interval $[A, B]$" we denote the process of drawing a random number from a uniform distribution on the interval $[\log_{10} A, \log_{10} B]$ on and exponentiating the value.

[5] Meaning we have equal number of cases having only L1, only L2, both, or no regularization.

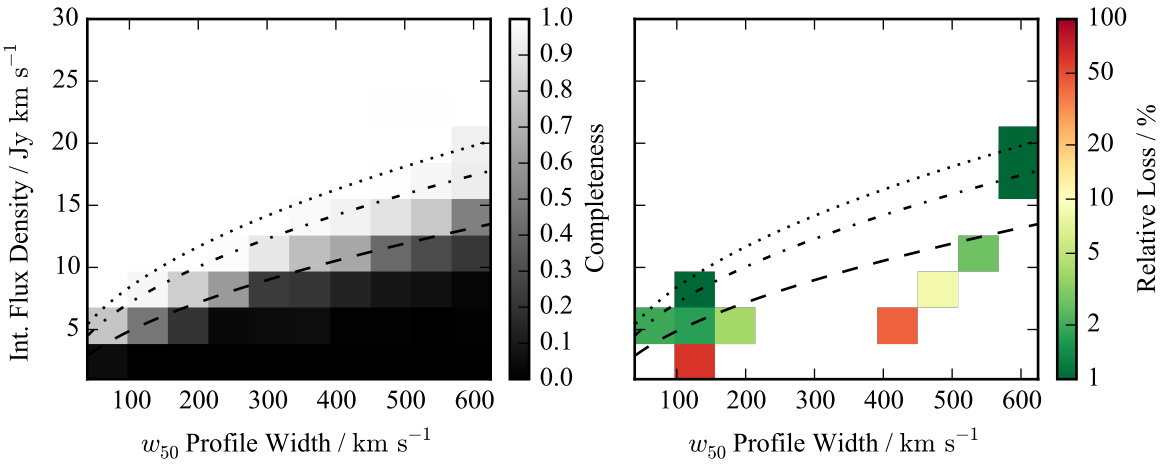


Figure 4.3: Completeness levels derived from correctly classified simulated sources and relative loss as compared to all sources (see Fig. 3.4). The dashed, dash-dotted and dotted lines show the 50%, 95% and 99% completeness level for all sources as a comparison. The derivation of the relative loss in each bin is described in the text.

In Fig. 4.2 we show the training and validation error for all three cross-validation folds. The optimal hyperparameters are summarized in Table 4.1. For the case of simulated sources, the classification performance is exceptionally good: our best performing ANN only mis-classifies 0.15% of all sources. This can be explained by the similarity of the simulated sources, as they are all perfect point sources. They can therefore be very well classified by the estimators for shape like the major axis of the Gaussian fits. Although most sources in EBHIS are expected to be unresolved, the measurement of the shape parameters from real data is less reliable.

The mis-classifications by the ANN determine two quantities that are interesting for blind H I surveys: completeness and reliability. The completeness for the best-case scenario, that is, all detected sources are classified as true positives, is investigated in Sec. 3.3.1. Mis-classifications reduce the completeness. The reliability on the other hand is defined as the probability that a candidate that is classified as true is actually a true source.

We quantify the average reduction in completeness by creating 10 000 bootstrap samples from the full simulated source catalog. We classify each bootstrap sample using the ANN and compare the resulting completeness to the completeness derived in Sec. 3.3.1 for all sources. We then average over all completeness estimates from the bootstrap samples to determine the relative loss of sources due to mis-classification. In Fig. 4.3 we show both the average completeness after classification as well as the relative loss when compared to the completeness from all sources. Since our ANN only mis-classifies eight sources, the completeness does not change significantly. Additionally, only sources at or below the completeness limit are lost, which have unreliable parameters and are often excluded from analysis anyway.

In Fig. 4.4 we show the reliability for our best-performing ANN on the simulated data. As expected from the high accuracy, the reliability is very high and never drops below 95% for any bin. As comparison, the overall catalog reliability for HIPASS is 95%.

4.2.3 Real Data

The challenge in training an ANN to classify detection candidates from real data is in creating a representative training data set. We compile our training data set by cross-matching the detection candidates

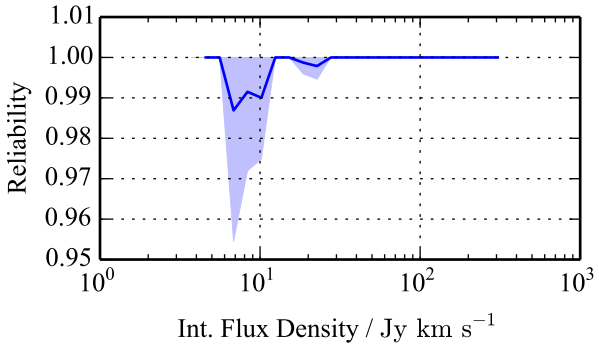


Figure 4.4: Reliability of the best-performing ANN for simulated data as a function of integrated flux density. The blue shaded area indicates the 95% uncertainty interval determined by 1 000 bootstrap samples.

from our pipeline with the source catalogs from HIPASS and the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005). We apply the same cross-matching criterion as in Sec. 3.3 and manually check the matches. The manual check is necessary to exclude matches of HIPASS or ALFALFA detections with false positives in the EBHIS pipeline output. We identify 767 true positives among the source candidates generated by the pipeline.

Since we have 17 647 examples for false positives, training an ANN with only 767 examples for true positives implies a strong weighting in favor of the false positives. Training from such an imbalanced data bears the risk of missing true positives when applying the ANN for classification. This is a common problem in machine learning and not only applies to ANN. He & Garcia (2009) provide an overview of the various methods developed to deal with imbalanced data. We examine four different approaches:

1. As a benchmark, we train the ANN on the unbalanced data set.

2. We balance the training data set by randomly replicating true positives until they match the false positives in number. This approach is described as "random oversampling" by He & Garcia (2009) and is one of the simplest balancing techniques that operates on the data. Since we are using cross-validation, the replication is performed only on the folds used for training. Otherwise, the ANN would be presented with all examples of true positives which biases the cross-validation score.

3. We create artificial examples of true positives using the Borderline Synthetic Minority Oversampling Technique (BLSMOTE) algorithm introduced by Han et al. (2005). BLSMOTE generates synthetic examples for the minority class which are located on the border separating the examples of the minority and majority class in the space spanned by the features of each example. The algorithm uses the nearest neighbors of each minority example and generates the synthetic examples based on their features. As before, we only augment the training data set with synthetic sources and leave the validation data for each cross-validation fold untouched.

4. We modify the cost function used in training. For this we add a weighting factor w to Eq. 4.3, which is unity for examples from the majority class, that is, the false positives, and the ratio $N_{\text{false}}/N_{\text{true}}$ for the true positives, where N_{false} and N_{true} are the number of examples of false and true positives in the training data set, respectively:

$$-\hat{\ell}(\theta, \mathcal{D}) = \hat{\mathcal{L}}(\theta, D) = \sum_i \log\left(P\left(Y = y_i \mid x_i, \theta\right) w(y_i)\right) \tag{4.8}$$

This weighting effectively increases the learning speed for true positives during training. Kukar & Kononenko (1998) show that approaches like this can be effective in learning from imbalanced

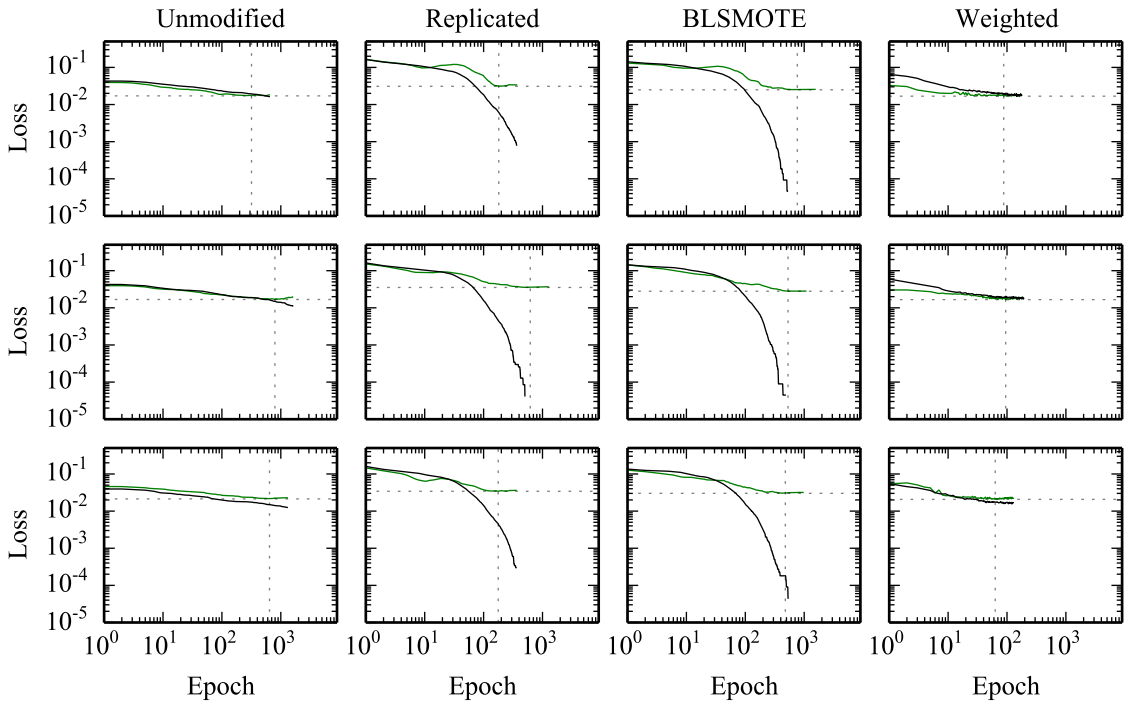


Figure 4.5: Training and validation loss curves of the best-performing ANN for the real data set. Each panel shows the curves for one of the folds used in cross-validation. The black and green curve show the average fraction of mis-classified examples over all mini-batches for the training and test data, respectively. The dotted grey lines indicate the loss at the best epoch.

data sets.

For all four approaches we perform separate hyperparameter optimization as described in the previous section with the following changes:

- Since we anticipate the classification of real sources to be a more difficult task, we search a wider range of neurons per layer. We draw the proposal number of neurons logarithmic on the interval $\left[10^{1.5}, 10^{3}\right]$ and round to the nearest integer.
- With increased number of neurons per layer, the possible size of our ANN becomes large enough for dropout-regularization to be a practical approach. With 50% probability each, we add dropout-regularization to the input or hidden layers. The dropout probability is $p = 0.5$ in all cases.

For all four approaches, Fig. 4.5 shows the training curves of the best performing ANNs for each approach. The corresponding parameters and performance characteristics are summarized in Table 4.2. The accuracy is the average fraction of correctly classified examples over all folds. To better quantify the differences between the four approaches, we also show the average reliability and completeness over all training folds for each ANN. In the machine learning community, these are also known as precision and true-positive rate, respectively.

An interesting observation is that all ANNs except for the ANN trained on the data set augmented by BLSMOTE employ some form of regularization (L1, L2, dropout or a combination thereof). This is a direct consequence of the larger, more diverse training data set created by using BLSMOTE. Having a

Parameter	Unmodified	Replicated	BLSMOTE	Weighted
Layers	2	3	3	2
Neurons	269	37	368	388
ϵ_0	0.037	0.033	0.036	0.024
τ	776	6927	6112	3499
η	0.06	0.73	0.85	0.72
λ L1	—	1.85×10^{-5}	—	5.28×10^{-3}
λ L2	2.46×10^{-6}	—	—	—
Input Dropout	—	0.5	—	0.5
Hidden Dropout	—	—	—	0.5
Accuracy	0.98	0.97	0.97	0.98
Reliability	0.87	0.58	0.66	0.90
Completeness	0.66	0.70	0.68	0.64

Table 4.2: Parameters of the best-performing ANNs for the task of real data classification.

large, diverse training data set prevents overfitting very effectively and yields the lowest generalization error, that is, the ability of the ANN to learn general properties of the examples instead of memorizing the training data set (Bengio 2012).

Another interesting observation is that the ANN trained on the unmodified data set performs relatively well and even outperforms the ANN trained using the weighted error function in terms of completeness. From the training curves in Fig. 4.5, it is clear that the training is much slower than for the other approaches, which is also explained by the small, initial learning rate ϵ_0 and fast learning rate decay τ as compared to the other ANN.

In all four cases, the accuracy is higher than 95%. This is explained by the large number of false positives in the training data set, which the ANNs are able to correctly categorize.

Since the average completeness of the ANNs is at most 70%, we investigate the impact on the survey completeness as function of integrated flux density and w_{50} profile width. We use our ANNs to predict the classes of the true positives for each validation fold individually. In Fig. 4.6 we show the correctly and incorrectly labeled true positives as a function of measured integrated flux density and w_{50} line width. For comparison, we show the completeness levels derived from the simulated data. Since they are derived from artifact-free data, they can be interpreted as the best-case scenario. Note also that the completeness for simulated data was calculated from the input values of integrated flux density and w_{50} profile width, as compared to the measured values used for sources from real data. From the plots, it is evident that the bulk of missed sources is located in the transition area from 100% to 0% completeness. Additionally, there is a number of very high SNR sources that are not correctly classified.

The most significant difference between the four approaches is their reliability. Whereas the ANNs trained with the weighted objective function or the unmodified data set achieve close to 90% reliability, the ANNs trained on the replicated or BLSMOTE-augmented training data sets achieve 58% or 66% reliability, respectively. This significant drop in reliability gives a slight advantage in completeness. In Fig. 4.7 we show the reliability as a function of measured integrated flux density. In all cases, the reliability decreases as sources become fainter. In case of the ANNs trained on the augmented training data sets, the transition to low reliability sets in at higher integrated flux densities and the transition is smoother as compared to the other two ANNs.

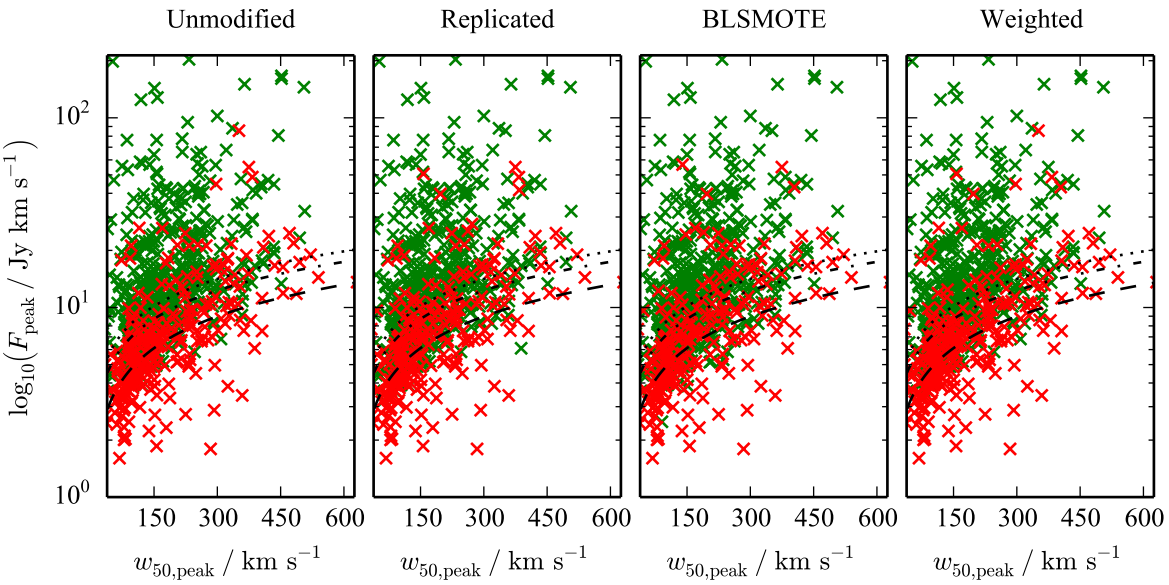


Figure 4.6: Classification accuracy for true positives for the four best performing ANN as a function of integrated flux density and w_{50} profile width. The dashed, dash-dotted and dotted lines indicate the 50%, 95% and 99% completeness level as derived from the simulations discussed in Chapter 3. Each cross indicates a true positive from the training data set. The red and green colors indicate wether the training example was missed or recognized by the ANN, respectively.

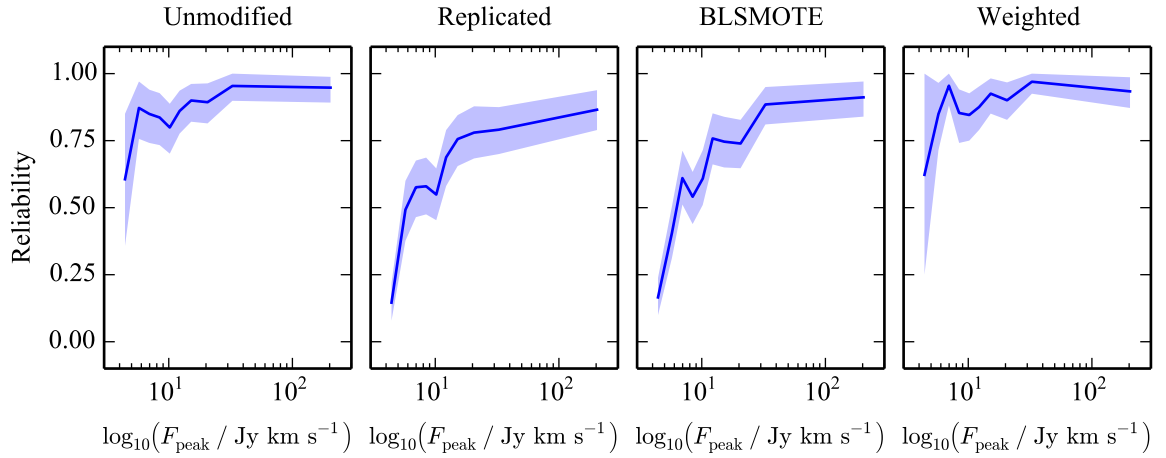


Figure 4.7: Reliability as a function of measured integrated flux density for the four best performing ANN. The blue shaded area indicates the 95% uncertainty interval derived from 1 000 bootstrap samples.

4.3 Conclusions

The results from this chapter show that the automated classification of source candidates generated by our automated source finding and parametrization pipeline is clearly a viable approach. For unresolved sources from simulated, clean data we achieve near-perfect classification. This is mainly explained by two effects. First, the large number of simulated sources and examples for false positives from the real data yield a large training data set. This allows us to train a highly accurate ANN with good generalization properties. Second, since all sources are perfectly unresolved, their properties are very similar and they can be efficiently described and recognized using the features provided to the ANN.

In applying the same method to classify source candidates from real data only, we have to deal with a highly imbalanced training data set, as the number of examples for true positives is relatively low. Using different approaches designed to train ANN from imbalanced data sets show mixed results. By augmenting the training data set using random replication or synthetic data, we can increase the completeness of the classification at the cost of reliability. Weighting the objective function in favor of the true positives increases training speed but does not increase accuracy, reliability, or completeness as compared to an unmodified ANN trained from the imbalanced data directly.

Neither of our approaches achieves an average completeness and reliability which is competitive with manual classification. We show that the bulk of classification-loss occurs for sources for which EBHIS has low completeness. Similarly, the reliability is a clear function of integrated flux density. Both observations suggest that the impact of automated classification on survey completeness could be modeled with smooth functions, which greatly reduces mathematical complexity when analyzing survey data.

To further improve automated classification, we identify two main approaches to be explored in future work. The first approach consists of creating a larger and more diverse training data set. It is a well known fact that ANN trained on larger training data sets show better performance, and the results from training ANNs on data sets augmented by synthetic sources or random replication show an increase in completeness.

Another approach to improve classification accuracy of the ANN would be to have better features to distinguish true and false positives. As mentioned before, it might be possible to use the moment maps, spectra or even data cubes as input for an ANN. Hinton & Salakhutdinov (2006) show that networks with many, large layers that use unsupervised pre-training methods are capable of learning features from images. Using the data itself rather than classifying sources based on their measured parameters is also closer to the process an astronomer performs when manually classifying source candidates.

CHAPTER 5

The Effelsberg-Bonn H I Survey Extragalactic Catalog

In this chapter, we apply the developed source-finding, parametrization, and classification software to the survey data from the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011). This is the first time that an H I survey is processed in a completely automated manner and is an important milestone on the way to process large-scale H I surveys with the next generation of radio observatories.

Based on the results of the prior Chapter it is obvious that a catalog that is based on simply the positively classified source candidates will be neither complete nor reliable. We therefore use the artificial neural networks (ANNs) for semi-automated classification, where every source candidate classified as a true source is confirmed by manual inspection. This does not alleviate the problem of completeness but ensures the reliability of the catalog.

The process of catalog creation and its properties are detailed in the following section. In Sec. 5.2 we compare our catalog to the published catalog from the H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001) and we discuss two commonly encountered issues with the creation of the catalog in Sec. 5.3.

5.1 Catalog Creation

To create the EBHIS extragalactic catalog, we apply source-finding and automated parametrization as described in Sections 2.4 and 3.2. As shown in Sec. 4.2.3, the reliability and completeness of the automatically classified catalogs is not satisfactory when compared to the completeness and reliability of manually classified catalogs. For this reason we adopt a semi-automated scheme for source candidate classification explained in the following section.

5.1.1 Semi-automated Classification

To increase the completeness and reliability of the extragalactic catalog, we use the ANNs in an interactive manner. We train each of four ANNs described in Sec. 4.2.3 on the full training data set, that is, without performing three-fold cross-validation. This is done to increase the number of training examples and therefore increase classification accuracy and generalization performance of the ANNs. Since the number of training epochs τ is now unable to be optimized by early-stopping, we fix it to the values determined by random optimization, that is, the values for τ shown in Table 4.2.

The processing of EBHIS data by our pipeline yields 218 490 source candidates. Using the process described in the previous Chapter, we use the ANNs to predict the nature of each source candidate. Whenever any network predicts that a given source candidate is of astrophysical origin, we manually

confirm or reject the candidate. The decision whether a source candidate is genuine is based on various criteria. Apart from its spectral shape or appearance in the velocity-integrated map, we search for known sources with matching position and redshift in the Simbad[1] and NED[2] data bases.

If the catalog search is inconclusive, we inspect 10′ images centered on the position of the source candidate from the Digitized Sky Survey (DSS)[3], the NRAO VLA Sky Survey (NVSS, Condon et al. 1998) and Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010). We use DSS blue and red filter images to identify optical counterparts outside of the zone of avoidance, that is, the region around galactic latitudes $|b| < 10°$. Where the extinction from the Milky Way renders DSS unusable for visual identification, we use data from all four wavelengths covered by WISE. The images from NVSS are used to identify strong continuum sources, as they are known to produce artifacts in the data that can mimic the shape of galaxies.

After a first pass through all candidates proposed by the ANNs, we create a new training data set, including the newly classified true and false positives and re-train the ANNs on this set. We then repeat the process of visual verification as described above. After two iterations of the process, 2756 true positives are identified among 11 026 inspected source candidates, including multiple detections due to the overlap of the individual data cubes shown in Sec. 2.4. These numbers show the advantage of using ANNs even for semi-automated classification. The number of candidates that has to be inspected is more than an order of magnitude lower than for HIPASS. Furthermore, due to the automated parametrization, the time spent to verify each candidate is typically less than 10 s.

5.1.2 Source Compilation

To compile the final source catalog, multiple detections caused by the field overlap are merged by applying the same cross-matching criterion used in Sec. 3.3. Once all multiple detections of a source are found, the detection with the largest distance to the edge of the data cube is chosen as the best. Since the angular extent of the sources is typically much less than 30′ and the data cubes have an overlap of 2°, this criterion is sufficient.

Since we derive the integrated flux density and profile width from both the peak and integrated spectrum (see Sec. 3.2), we need to decide which measurement is appropriate for any given source. We make use the simple criterion derived by Meyer et al. (2004). Broeils & Rhee (1997) show that there is a correlation between H I mass and H I diameter in a sample of optically selected galaxies. Together with the standard formula for H I mass given the distance D and integrated flux density S, this relation can be used to estimate the angular extent of a given source $\theta_{\mathrm{H\,I}} \approx \sqrt{0.083\,S}$. We estimate the angular extent of a given source using the integrated flux density as determined from the peak spectrum, F_{peak}. Once the estimated size exceeds half the angular resolution of EBHIS, that is 5.25′, we use the parameters measured from the sum spectrum.

5.1.3 Catalog Properties

The final EBHIS catalog contains 1847 sources. Figure 5.1 summarizes the parameter distribution of the detected sources. A comparison of the detected sources with the completeness limits from the simulations in Chapter 3 shows good agreement in the shape of the transition from 100% to 0% completeness. From the performance of the ANNs determined in Sec. 4.2.3 it is expected that at most 70% of all

[1] http://simbad.u-strasbg.fr/simbad/
[2] http://ned.ipac.caltech.edu
[3] http://www3.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/dss/index.html

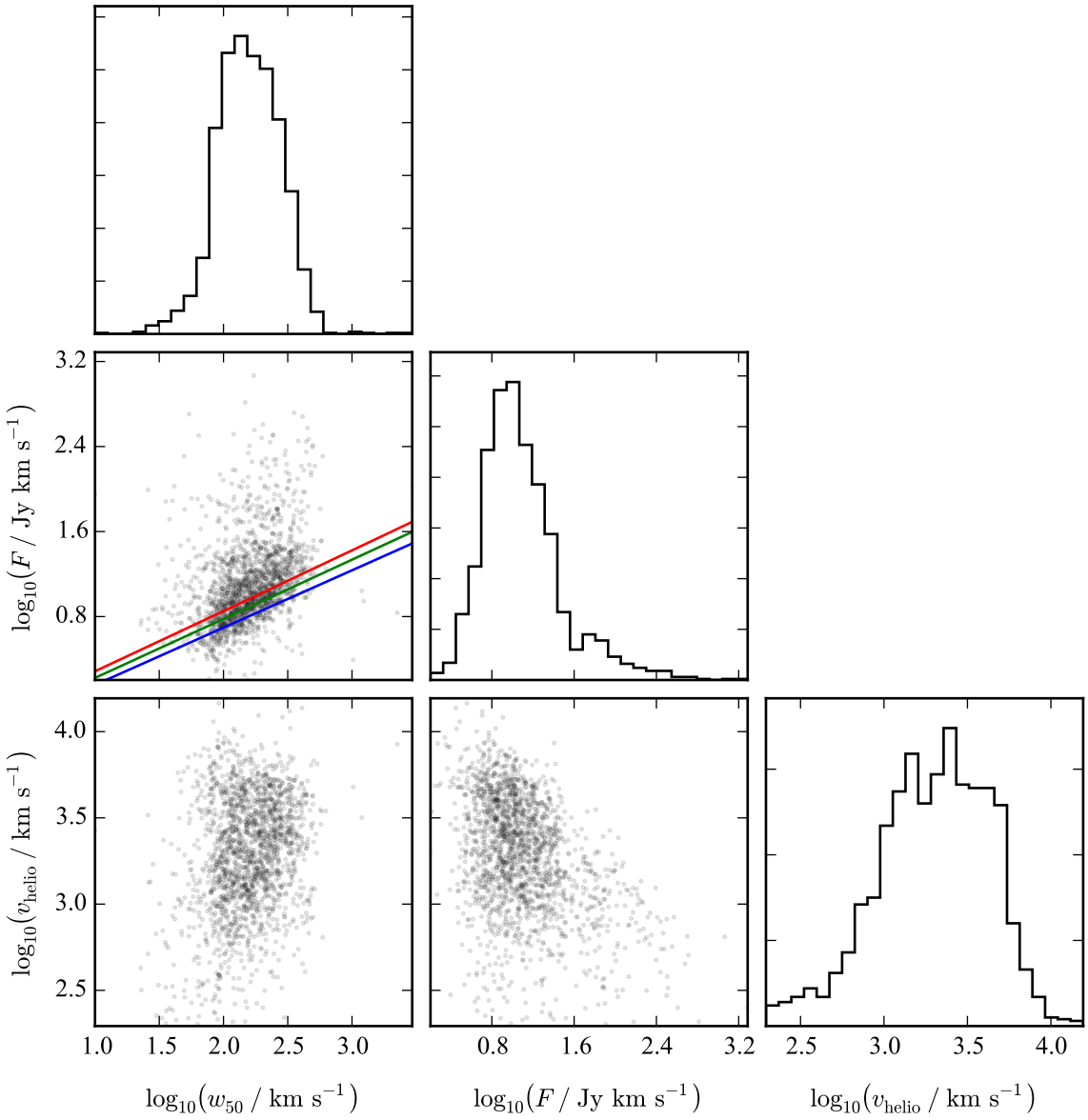


Figure 5.1: Logarithmic parameter distributions of the compiled EBHIS catalog. The scatterplots show the bivariate distribution of detected sources as grey dots. The blue, green, and red lines in the w_{50}-F_{int} plane indicate the 50%, 95% and 99% completeness levels for EBHIS as derived from the simulations in Chapter 3. The histograms show the relative distribution of sources for the parameters.

sources are found using our semi-automated classification, assuming the properties of the training data adequately reflect the general source population.

We show the sky distribution of sources in Fig. 5.2. The catalog clearly reflects the prominent large-scale structure on the northern hemisphere. Figure 5.2 also shows the median noise level of the survey as a function of sky position. We determine the noise level by calculating the median absolute deviation (MAD) on 15 non-overlapping chunks of 100 channels and taking the median of the individual estimates. Prominent features in the noise map are the increased noise level towards the galactic center and for declinations larger than 60°.

To get an estimate for the real completeness of the survey, we follow the approach of Zwaan et al. (2004) who use two different tests to verify the completeness of HIPASS.

We first implement the *test for completeness* statistic T_C as described byRauzy (2001). The test consists of repeatedly calculating the statistic T_C on subsamples of the survey, truncated at decreasing integrated flux densities F^{lim}. As long as the survey is complete down to a given integrated flux density, the statistic T_C is expected to have zero mean and unit variance. If the survey starts to become incomplete, the T_C statistic will give negative values. The integrated flux density limit where the T_C statistic is significantly below -2 determines the completeness level of the survey at the 97% confidence level.

To calculate the T_C statistic, we first define a distance modulus $Z = \log_{10}(F) - \log_{10}(M_{\mathrm{H\,I}})$. For each limiting integrated flux density F^{lim}, we can now calculate the minimum H I mass $M_{\mathrm{H\,I}}^{\mathrm{lim}}$ each galaxy could have to still be included in the truncated sample. We calculate the masses by converting the radial velocities of the galaxies to the CMB rest frame and use Hubbles law to obtain their distances. For each galaxy in a given truncated sample, we now determine the numbers r_i, the number of galaxies with $M_{\mathrm{H\,I}} \geq M_{\mathrm{H\,I},i}$ and $Z \geq Z_i$, and n_i, the number of galaxies with $M_{\mathrm{H\,I}} \geq M_{\mathrm{H\,I},i}^{\mathrm{lim}}$ and $Z \geq Z_i$. Given these numbers, the T_C statistic for a given F^{lim} is calculated by

$$\xi_i = \frac{r_i}{n_i + 1} \tag{5.1}$$

$$V_i = \frac{1}{12}\frac{n_i - 1}{n_i + 1} \tag{5.2}$$

$$T_C = \frac{\sum_i (\xi_i - 0.5)}{\sqrt{\sum_i V_i}} \quad . \tag{5.3}$$

We estimate uncertainties for T_C by calculating the statistic on 100 bootstrap samples from the EBHIS catalog. We furthermore investigate the dependence of the statistic on measurement uncertainties in F and v_{helio} by artificially corrupting the measured parameters in the EBHIS catalog. We corrupt the integrated flux densities and radial velocities by adding Gaussian noise with zero mean and dispersion $5\,\mathrm{Jy\,km\,s^{-1}}$ and $250\,\mathrm{km\,s^{-1}}$, respectively. Because the EBHIS catalog is dominated by the super-galactic plane, we also divide the catalog in six equal slices in Right Ascension and calculate the T_C statistic on each slice individually with the first slice centered on Right Ascension 0^{h}.

The results for the T_C statistic are shown in Fig. 5.3. The T_C statistic determined from the full catalog shows the expected behavior down to approximately $30\,\mathrm{Jy\,km\,s^{-1}}$. At this point T_C becomes strongly positive before steeply declining at $10\,\mathrm{Jy\,km\,s^{-1}}$. The failure of the T_C statistic can be explained by its implicit assumption, that the completeness only depends on the integrated flux density of each source. Since the classification by the ANNs adds another, difficult to quantify, selection function to the survey this assumption is clearly violated.

The results from the corrupted catalog also show that this failure can not be explained by measurement errors. The only significant effect of the corrupted measurements is a change in the location where T_C becomes negative, that is, a reduction of the measured completeness level. This is explained by the

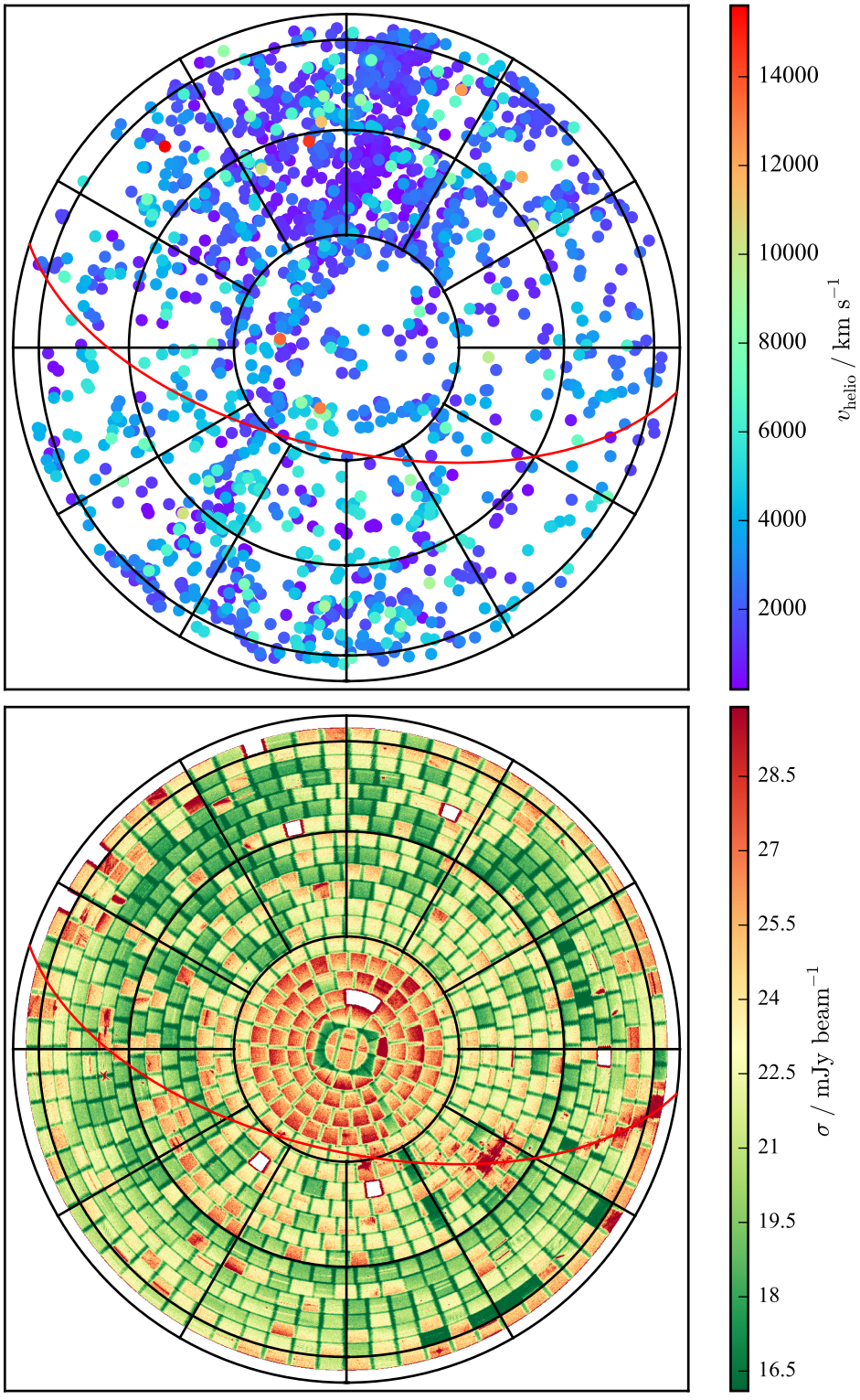


Figure 5.2: Sky and redshift distribution of the sources in the compiled EBHIS catalog (**top**) and median noise level of the survey as a function of sky position (**bottom**). The coordinates are equatorial J2000 coordinates. Right ascension 0^{h} is down. The circles are drawn at declinations $-10°$, $0°$, $30°$, and $60°$. The red line indicates the galactic plane.

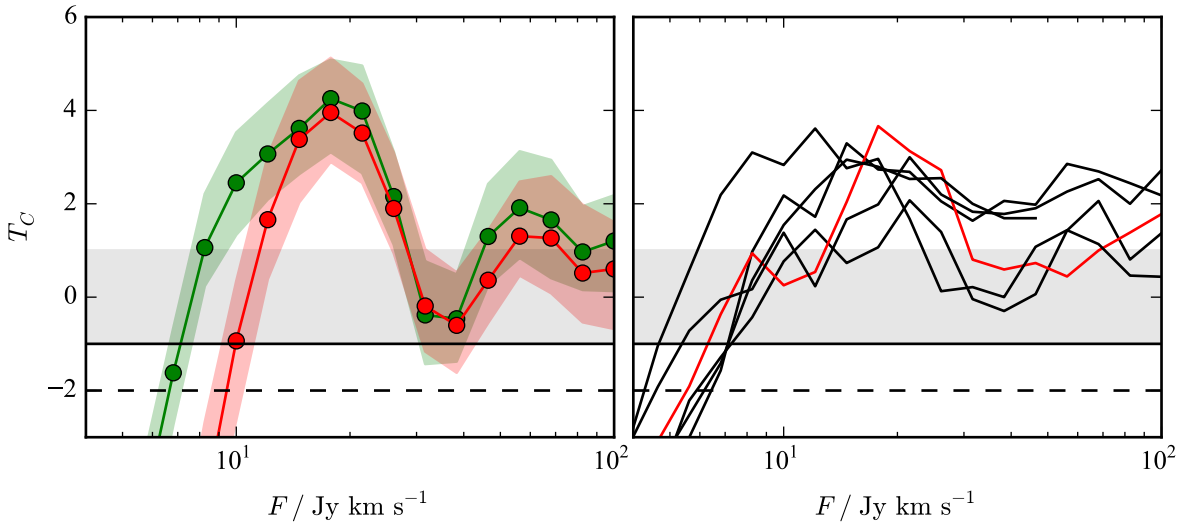


Figure 5.3: Results for the T_C statistic after Rauzy (2001) for the EBHIS catalog. The gray area indicates the expected range of T_C for a complete sample at a given integrated flux density F. **Left**: The T_C statistic for the full EBHIS catalog. The green shaded area and line indicate the 68% confidence interval and median for the statistic. The red shaded area and line show the same for the statistic determined from the artificially corrupted EBHIS catalog. **Right**: The T_C statistic for the six Right Ascension slices of the EBHIS catalog. Each line indices the median of the T_C statistic from 100 bootstrap samples. The red line indicates the bin centered on Right Ascension 12^{h} which contains 636 galaxies.

galaxies which have their integrated flux density reduced below $0\,\mathrm{Jy\,km\,s^{-1}}$ by the added noise and therefore effectively drop out of the catalog.

The estimation of T_C from the six redshift slices does not show any significant difference between any two slices. This is expected as the T_C statistic was designed to be unaffected by large-scale structure in the survey volume. It also shows that the selection function of EBHIS does not vary across the sky.

The other test Zwaan et al. (2004) use, relies on the fact that the number of sources N as a function of integrated flux density F scales as $dN \propto F^{-5/2}dF$ for a complete, flux-limited survey. We can compare the slope of this prediction to a histogram of the sources detected in the survey. Once the the histogram significantly deviates from the predicted slope, the survey completeness limit is reached. For comparison to a histogram of sources, the relation is integrated over the width of a bin spanning the interval $[a, b]$:

$$N_{[a,b]} \propto \frac{2}{3}\left(a^{-\frac{3}{2}} - b^{-\frac{3}{2}}\right) \quad . \tag{5.4}$$

The results for the dN-dF-test are presented in Fig. 5.4. We plot the expected shape of the source distribution scaled to the faint-end turnoff and the bright end of the source distribution, respectively. The bright end of the observed source distribution is not well described by either relations. As in case of the T_C statistic, this shows that the underlying selection function of the catalog does not depend on integrated flux density alone.

As the dN-dF-test is not robust against large-scale structure in the survey volume, we test the robustness of our result by extracting 10 000 random Right Ascension slices of 90° size from the catalog and perform the dN-dF-test on each individually. Since the test uses the absolute number of sources, we randomly oversample the number sources each slice to have the same number of sources as in the full

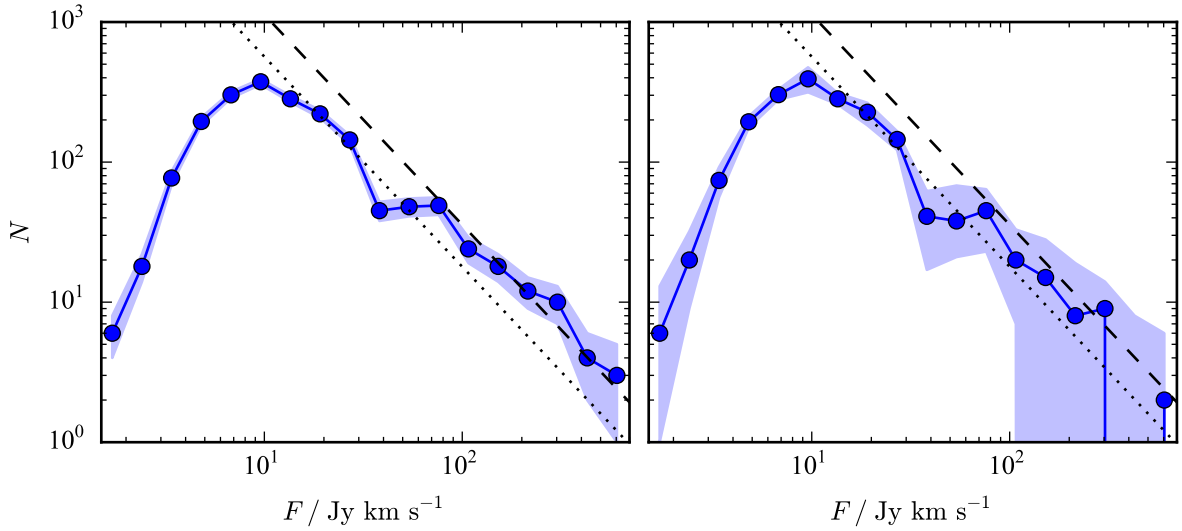


Figure 5.4: **Left**: The *dN-dF*-test for the full EBHIS catalog. The blue line and shaded area indicate the median and 68% confidence interval as determined from 10 000 bootstrap samples. The dashed and dotted line are the expected shape of the source distribution for a complete survey, scaled to the bright end and peak of the distribution, respectively. **Right**: Same as the left panel but for the catalog randomly sampled in Right Ascension.

catalog. As can be seen in the right panel of Fig. 5.4, there is evidence that the bright end of the source distribution is strongly affected by the large scale structure in the survey. This is expected as the catalog is dominated by the Virgo cluster region.

5.2 Comparison with HIPASS

Using the HIPASS catalog and the selection functions of both surveys, we can estimate how many sources EBHIS should be able observe. Since the selection functions for HIPASS and EBHIS depend on the integrated flux density F, the peak flux density P, and the profile width w_{50}, we assume that the distribution of these parameters is statistically identical between the northern and southern hemisphere (Zwaan et al. 2004, and Sec. 3.3.1). We further have to assume that HIPASS has detected all sources that EBHIS can detect. Since the noise level in HIPASS is only 30% lower than in EBHIS, this assumption might lead to a underestimation of the expected number of faint sources in EBHIS.

A more sophisticated way of estimating the expected number of sources in EBHIS would be to extract the survey volume out of a cosmological simulation and measure the parameters relevant for detection for each galaxy. Since this requires a realistic simulation of baryon physics, this is not straight forward to do. Recent simulations show significant disagreement with previous results on the distribution of baryons (Vogelsberger et al. 2012). In addition the observed distribution of w_{50} is in disagreement with the expected velocity width distribution from dark matter simulations (Zwaan et al. 2010; Papastergis et al. 2011). For these reasons, we use the simple catalog resampling technique described in the following paragraphs to obtain an approximate estimate of the expected source distribution in EBHIS.

For each source in HIPASS, we can calculate a probability that it would be observed by EBHIS from the completeness function for EBHIS derived from the simulations in Chapter 3. By using these probabilities as the parameter for a Bernoulli distribution, $\mathcal{B}(p)$, we can obtain random draws for the number of sources expected in EBHIS. Since most sources in HIPASS have a completeness smaller than unity, we need to weight each source by its inverse completeness in HIPASS, C_{HIPASS}. A single random

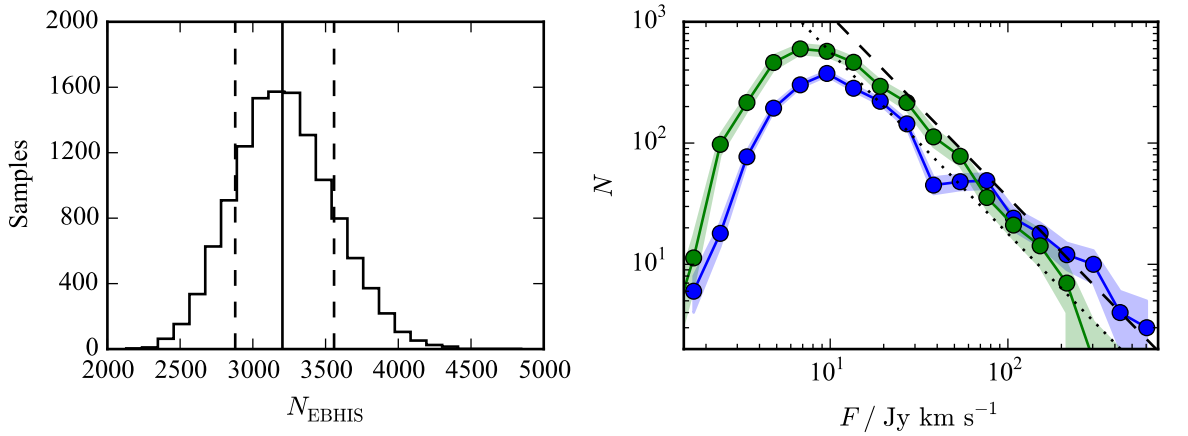


Figure 5.5: **Left**: Histogram of 12 500 random draws of the expected source counts for EBHIS determined from HIPASS. The vertical lines indicate the median and 68% confidence interval. **Right**: Expected distribution of source integrated flux densities as derived from 12 500 random draws from HIPASS. The blue curves as well as the dashed and dotted line have the same meaning as in Fig. 5.4. The green line and shaded area indicate the median and 68% confidence level of the expected distribution of source fluxes in EBHIS.

draw of the expected number of sources is obtained by

$$N_{\mathrm{EBHIS}} = \sum_i \frac{\mathcal{B}(C_{\mathrm{EBHIS}}(F_i, w_{50,i}))}{C_{\mathrm{HIPASS}}(P_i, F_i)} \tag{5.5}$$

where the index i runs over all sources in the HIPASS catalog and F_i, P_i, and $w_{50,i}$ are the parameters of the ith source. Since the northern and southern hemisphere exhibit a different large-scale structure, we do not obtain the random draws from the full HIPASS catalog. We instead perform the sampling on eight random, 45°-sized slices in Right Ascension separately and add the number of sources to obtain a single estimate of the number of sources expected in EBHIS. This sampling method averages-out the large-scale structure contained in HIPASS and gives a less biased estimate. In the left panel of Fig. 5.5 we show the histogram of 12 500 random draws[4]. The number of sources that should be detected by EBHIS is 3207^{+347}_{-322}, which is approximately twice as many sources as included in the EBHIS catalog.

We use the same method to obtain an estimate on the expected number of sources with a given integrated flux density. This can be compared to the results from Sec. 5.1.3 to obtain information about where the current EBHIS catalog is incomplete. For each random draw, we create a weighted histogram using the same binning as used to create Fig. 5.4 and plot the median and 68% confidence interval in the right panel of Fig. 5.5. The expected distribution of source fluxes follows $dN \propto F^{-5/2}dF$ very well up to 20 Jy km s^{-1}. The observed number of sources is less than expected for the complete range of integrated flux densities below 40 Jy km s^{-1}. The slight excess of detected sources brighter than 100 Jy km s^{-1} is explained by the influence of the Virgo cluster regions, as shown in the previous section and Fig. 5.4.

The northern extension of HIPASS overlaps with EBHIS and allows us to compare the integrated flux densities and measured w_{50} profile widths (Wong et al. 2006). We crossmatch the northern HIPASS catalog with the EBHIS catalog and obtain 518 matching detections. To correctly treat the measurement uncertainties in both surveys and account for the expected intrinsic scatter in the parameter comparison, we implement the bayesian fitting routine outlined in Sections 7 and 8 of Hogg et al. (2010). We use the posterior distribution for each fit parameter to report the median and 68% uncertainty interval for each

[4] 100 000 draws of 45°-sized slices in Right Ascension

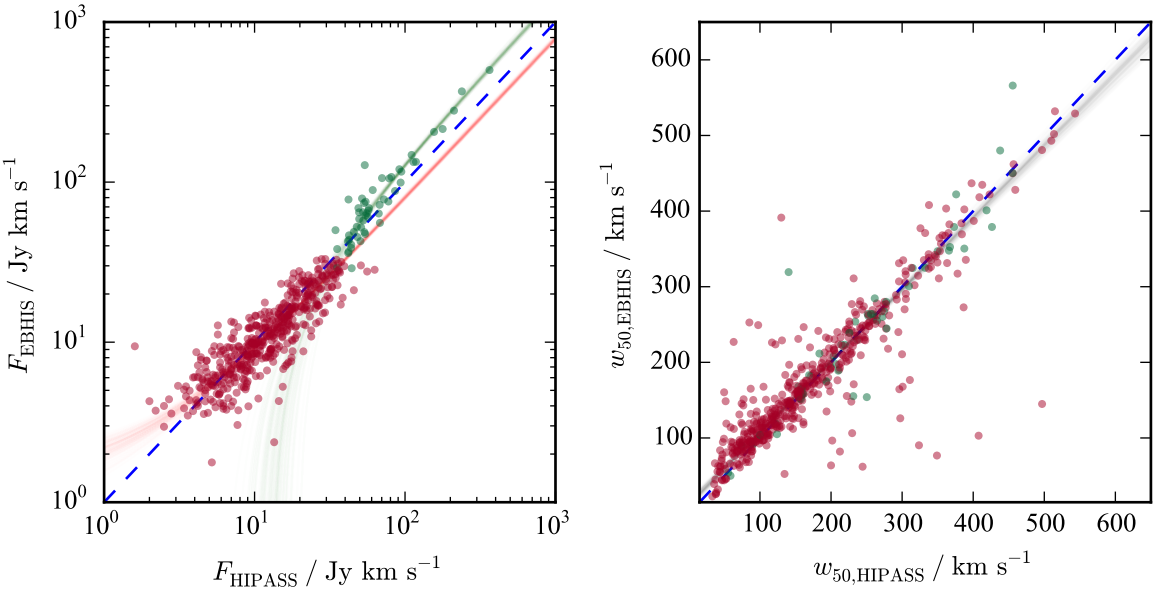


Figure 5.6: Comparison of measured galaxy parameters between EBHIS and HIPASS. The dashed blue line indicates the one-to-one relationship. Each group of lines represents 100 random samples from the posterior distribution of the linear model. The differently colored dots indicate whether a source is treated as point-like (green) or extended (red) by the EBHIS pipeline. Although used in the analysis, we do not show error bars for clarity. **Left**: Comparison of the integrated flux density. The red and green lines are samples from the posterior for peak and integrated photometry, respectively. **Right**: Comparison of the w_{50} profile width.

parameter. Figure 5.6 shows the fit results for the integrated flux density and w_{50} profile width.

Since the integrated flux densities in our catalog are measured either from the peak or integrated spectrum, we fit the correlation between our catalog and NHICAT separately for both methods and once for the full catalog. The best fit for slope and intercept for flux densities determined from the peak spectrum are 0.78 ± 0.02 and 1.3 ± 0.3 Jy km s^{-1}. The intrinsic scatter of the relation is 2.1 ± 0.1 Jy km s^{-1}. For these sources, EBHIS seems to detect about 20 % less flux than HIPASS does. This picture reverses when we only consider sources whose integrated flux density is determined from the integrated spectrum. Here the best-fit parameters for slope and intercept are 1.46 ± 0.05 and −18 ± 4 Jy km s^{-1}. The intrinsic scatter of this relation is 7 ± 1 Jy km s^{-1}. For these sources EBHIS seems to measure nearly 50 % higher integrated flux density. Haynes et al. (2011) perform a similar comparison between NHICAT and the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005) and find that, on average, ALFALFA measures about 10 % higher integrated flux densities. They attribute this difference to the poorer sensitivity and subsequently less accurate parametrization for HIPASS.

A possible explanation for the systematic difference for bright sources between EBHIS and HIPASS lies in the way the integrated flux density is measured in both surveys. In HIPASS, the integrated flux density is measured in a box centered on the source. The side length of this box is chosen by the observer parametrizing the source and is either 28′ or 44′ for most sources in the HIPASS catalog, depending on whether the source is treated as point-like or extended (Meyer et al. 2004). In the northern extension of HIPASS, with which we compare our parametrization, there are only two sources treated as extended (Wong et al. 2006). In EBHIS, the size of the aperture in which the integrated flux density is measured is determined from the source itself (see Sec. 3.2.1). This enables larger integration areas and captures the brightness distribution of an extended source more accurately. It does however not explain the significant difference for sources treated as unresolved. In Sec. 3.3.2 we show that for

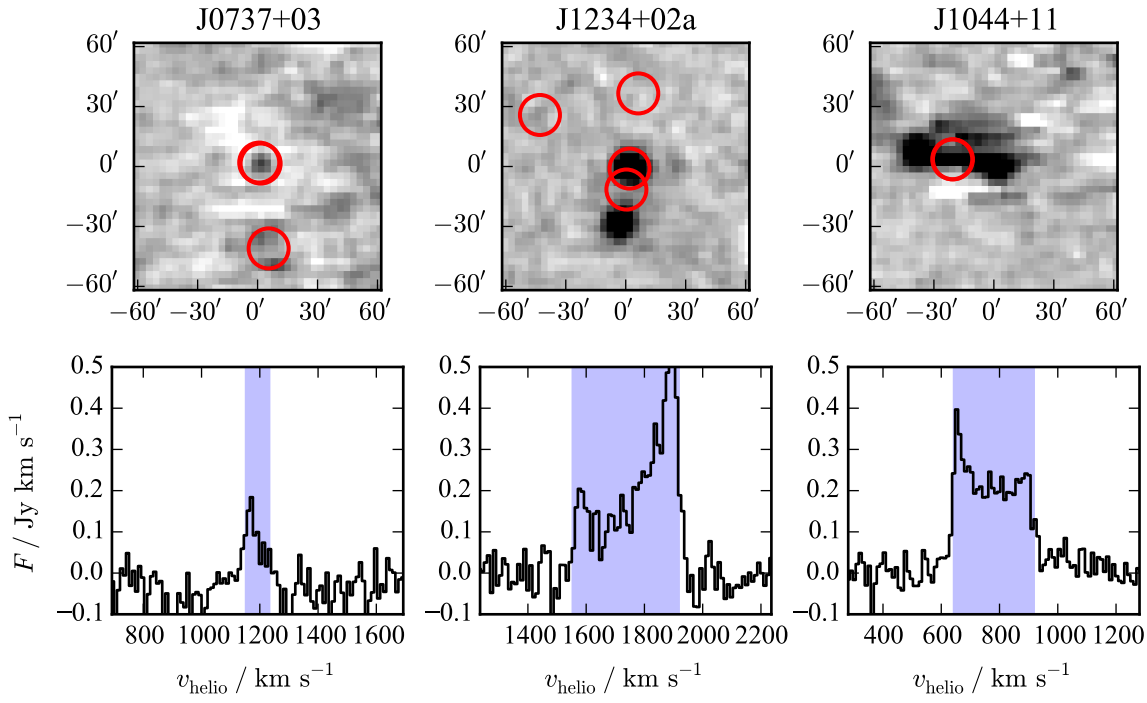


Figure 5.7: Examples for sources detected in the northern extension of HIPASS which are not classified as true positives by the ANNs. The names above each column refer to the designation in the HIPASS catalog. **Top Row**: Velocity-integrated maps from EBHIS data, integrated over $1.5 \times w_{50}$ and centered on the radial velocity given in the HIPASS catalog. The red circles indicate the positions of source candidates from EBHIS. **Bottom Row**: Spectra from EBHIS data centered on the angular position of the HIPASS detection. The blue shaded area indicates the radial velocity and w_{50} as stated in the HIPASS catalog.

unresolved sources, the method of measuring the integrated flux density from the peak spectrum is accurate and unbiased. It is possible that the sources in EBHIS are more resolved than assumed by the simple criterion used in Sec. 5.1.2.

The best fit parameters and 68% confidence interval for the slope and intercept for the w_{50} parameter are 0.95 ± 0.02 and $12 \pm 4\,\mathrm{km\,s^{-1}}$, respectively. The intrinsic scatter of the relation is $31 \pm 1\,\mathrm{km\,s^{-1}}$. In comparison to the rather tight relations between the integrated flux densities, there are a few outliers in the measurement of the w_{50} profile width. This is also reflected in the large intrinsic scatter in the relation. Apart from the outliers, the measured w_{50} profile widths agree very well between the two surveys.

5.3 Common Issues

In light of the results from the preceding sections, we will discuss two commonly encountered issues in our automated pipeline. We will discuss these issues by inspecting sources detected in the northern extension of HIPASS for which EBHIS has a theoretical completeness of larger than 90%, based on the parameters determined by HIPASS and the completeness function from Sec. 3.3.1.

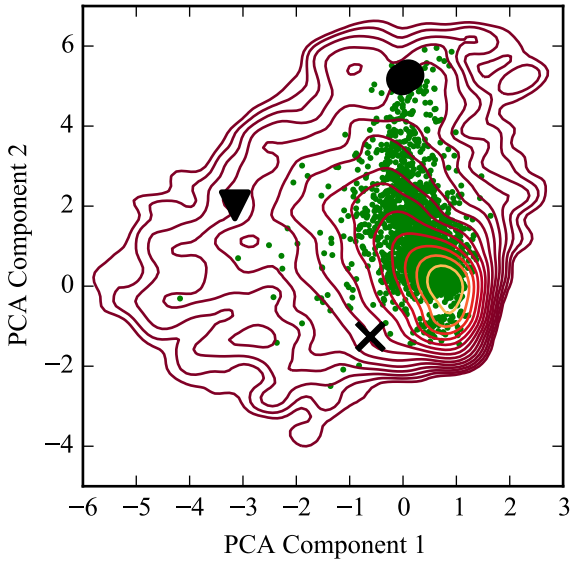


Figure 5.8: Distribution of the two principal components explaining the most variance for the parameters of manually classified true and false positives. The red contours indicate the empirical distribution function of the false positives and are spaced logarithmically. The green dots are the true positives. The black cross, triangle, and filled circle indicate the most likely matches for J0737+02, J1234+02a, and J1044+11, respectively.

5.3.1 Mis-classification

As anticipated by the results from Chapter 4, the most common reason why a source is not included in the EBHIS catalog is mis-classification by the ANNs. In Fig. 5.7, we show examples of three galaxies from NHICAT clearly detected in EBHIS data but not classified as true positives by any of the ANNs.

In all cases there is at least one plausible source candidate in EBHIS. To understand why these galaxies are mis-classified by the ANNs we investigate the location of the most-likely matches in EBHIS in the 53-dimensional parameter space used for source classification. For the purpose of visualization, we perform a principal component analysis (PCA) on all confirmed true and false positives from EBHIS data. We use `RandomizedPCA` from the `scikit-learn` package (Pedregosa et al. 2011, `http://scikit-learn.org`) to extract the two principal components which explain the most variance in the original parameter space. We show the distribution of true and false positives together with the most likely matches for each of the sources in Fig. 5.7. From this visualization it is evident that the source candidates matching the HIPASS detection are located in a region with few true positives. This also means that there are very few training examples in this region of the parameter space which allows the ANNs to optimize their objective function by classifying all candidates in this region as false positives. This is especially true for the source J1044+11, which corresponds to NGC 3628 and is known to possess a giant H I stream, believed to be the result of an interaction with NGC 3627 (Rots 1978). This feature gives the galaxy a highly unusual shape and parameters, which make it difficult to train a machine learning algorithm recognize such objects.

5.3.2 Data Quality

After mis-classification, the most common issue is the data quality of the extragalactic EBHIS data. Until now, the data is optimized for the radial velocity regime in which galactic H I emission can be found, since the galactic EBHIS survey is expected to have higher scientific impact. Consequently, there are still many artifacts present in the extragalactic data. The two most commonly encountered defects are residual radio-frequency interference (RFI) not flagged by the automated flagging algorithm and solar ripple. While the former mostly creates a large number of false positives, the latter can severely reduce

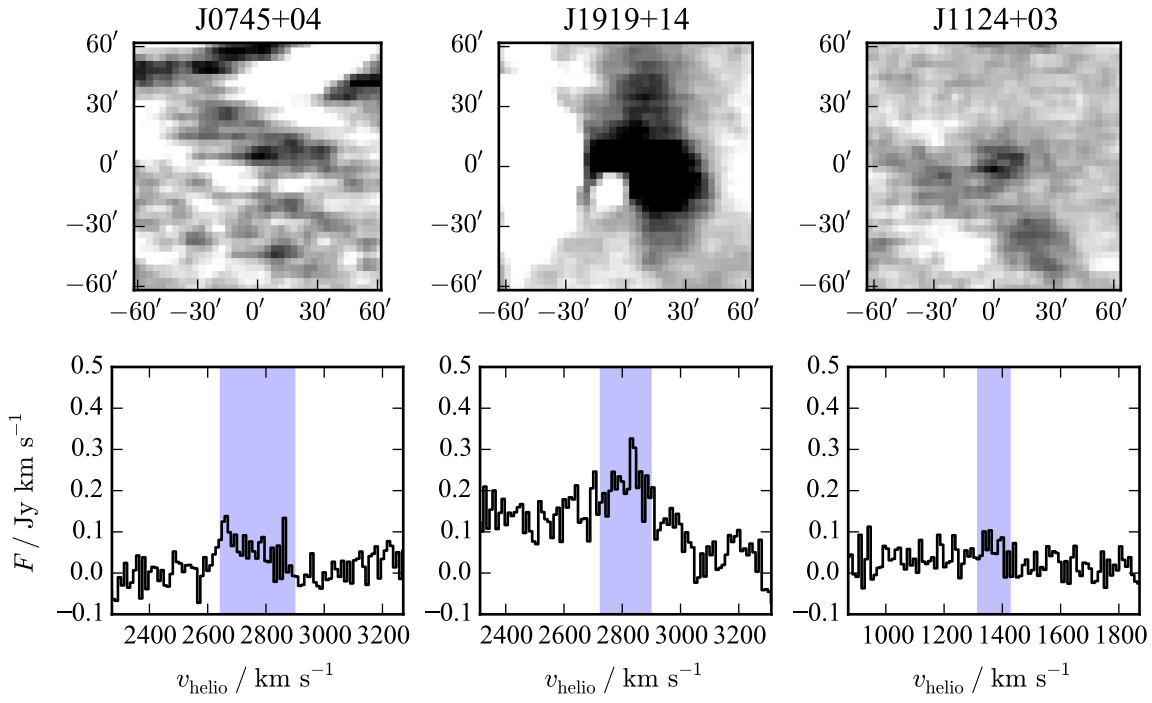


Figure 5.9: Examples for sources detected in the northern extension of HIPASS which are not detected in EBHIS data due to data artifacts. The description of the panels is the same as in Fig. 5.7.

the pipeline's ability to detect sources in the data. Another common artifact are corrupted baselines due to strong continuum sources, which have the same effect on the data as the solar ripple.

In Fig. 5.9 we show three examples of sources which are not detected in EBHIS data due to quality of the data. While the sources J0745+04 and J1124+03 are examples of data affected by solar ripple, J1919+14 shows the effect of a strong continuum source on the data. There are methods to mitigate the impact of both solar ripple and strong continuum sources (Meyer et al. 2004; Barnes et al. 2005) on the data and we fully expect to increase the data fidelity once an effort is made to incorporate these methods in the EBHIS data reduction pipeline.

5.4 Conclusions

In the preceding Chapter we use the methods and algorithms developed in Chapters 2 through 4 to create the first extragalactic catalog from EBHIS data. As the performance of the ANNs is not sufficient to perform completely unsupervised classification, we use them in an interactive way to reduce the number of source candidates that are inspected. Instead of having to inspect all 218 490 candidates generated by the pipeline, we only inspect 11 026 candidates to obtain 2756 true positives. After merging multiple detections caused by the overlap between the individual data cubes, we obtain 1847 unique detections. Comparing the resulting catalog to the number of galaxies expected from resampling the HIPASS catalog, we note that the number of sources contained in our catalog only contains about 58% of the expected number of sources. Although the absolute estimates of expected sources from HIPASS are only rough estimates, the incompleteness of the catalog is further corroborated by two independent statistical tests. The tests also show that the incompleteness is not due to insufficient sensitivity but the

classification performance by the ANNs.

The statistical properties of the catalog in its current stage make it unsuitable to derive statistical properties like the H I matter content of the local universe, $\Omega_{\mathrm{H\,I}}$, or the shape of the H I mass function (HIMF). This is mainly due to the difficult to quantify selection function imposed by the ANNs used for semi-automated classification. We identify two approaches which would improve the performance of the ANN on EBHIS data.

The extragalactic EBHIS data still contains a large amount of artifacts. This is evident from large number of source candidates generated by the automated pipeline, despite using a source-finding algorithm that is proven to be robust against common defects. If the number of artifacts that overlap with the true astrophysical sources in the parameter space used for classification is reduced, the classification performance is expected to improve drastically.

Another approach to the problem is the use of more information for the task of classification. This could be achieved by using ANNs to perform image recognition rather than classification based on derived parameters. As hinted at above, this requires vastly more complex networks and machine-learning techniques but would open new approaches to classifying detections from spectroscopic data.

The comparison of the measured integrated flux densities and w_{50} profile widths between HIPASS and EBHIS shows a mixed picture. The integrated flux densities measured in both surveys differ significantly depending on the method used. While integrated flux densities measured from the peak spectrum are about 20 % smaller than in HIPASS, extended sources have a nearly 50 % higher integrated flux density in EBHIS. We identify the different parametrization schemes used by both surveys as a possible source for the discrepancy. Apart from a small fraction of outliers, the measured w_{50} profile widths agree very well between the two surveys.

Although the results from this chapter leave a lot of room for improvement they still represent a milestone for the future of H I surveys. This is the first time a survey is processed in a completely automated manner. The complete process of source-finding, parametrization and classification for the full EBHIS data set takes less than a week of time using 16 CPU cores. Re-running only the parametrization pipeline requires approximately 24 h. This is a vast improvement over the manual labor and time involved in processing past surveys like HIPASS and ALFALFA. Future H I surveys with the Australia SKA Pathfinder (ASKAP, Johnston et al. 2008), the South African SKA Pathfinder (MeerKAT, Booth et al. 2009) and Apertif on Westerbork Synthesis Radio Telescope (WSRT) have orders of magnitude more processing power available to them, making more sophisticated techniques computationally feasible. The reduction of manual labor involved in processing a survey from the automation of the parametrization process alone might make it possible to use semi-automated or manual classification for future H I surveys.

CHAPTER 6

Conclusions and Outlook

In the last Chapter of this thesis we will summarize the main findings and give an outlook on possible future projects improving on them.

6.1 Conclusions

In this thesis, we describe the development of a fully automated source-finding, parametrization, and classification pipeline for large-scale H I surveys, with specific application to the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011). The pipeline focuses on the detection and parametrization of unresolved sources, as they represent the bulk of sources detected in current and future H I surveys (Duffy et al. 2012). It is a first step to develop the necessary algorithms to adequately process data from future H I surveys.

With the application of 2D-1D wavelet de-noising to H I data cubes, we use a detection algorithm well adapted to the signal caused by H I in galaxies. We show that the separate treatment of spatial and spectral axes of H I data cubes yields both higher sensitivity and robustness to data artifacts, as sources and common data artifacts contribute differently to each wavelet sub-band. While source finding is already automated by past and current H I surveys, the presented algorithm is an improvement over state-of-the-art techniques and uses more information to extract relevant signal from H I data cubes.

The way source candidates are parametrized and classified is the major departure from current H I source extraction techniques. Past surveys like the H I Parkes All-Sky Survey (HIPASS, Barnes et al. 2001) and the Arecibo Legacy Fast ALFA Survey (ALFALFA, Giovanelli et al. 2005) use the expertise of individual astronomers to first decide whether a given source candidate is a real galaxy. If a given source candidate is determined to be of astrophysical origin the astronomer then proceeds to derive the source parameters of interest. It is impractical to scale this process to the data volumes of future H I surveys with the Australia SKA Pathfinder (ASKAP, Johnston et al. 2008) and the upgraded Westerbork Synthesis Radio Telescope (WSRT/Apertif, Oosterloo et al. 2010) and, because of its human component, is non-deterministic.

We apply machine learning to solve the task of candidate classification. This requires to revert the usual process of classification and parametrization, meaning that all source candidates need to be parametrized to perform classification using artificial neural networks (ANNs). For this reason, we develop a reliable parametrization pipeline that does not require human supervision. Apart from enabling machine learning, it also greatly reduces the processing time required as compared to manual parametrization. We use this increased speed to perform large end-to-end simulations and derive precise parametrization uncertainties.

Using simulations, we verify that source classification based on their measured parameters is indeed a viable approach. In a data set containing simulated sources and real data artifacts from EBHIS, we achieve near-perfect completeness and reliability. This is in part owed to the fact that the simulated sources are completely unresolved and therefore occupy a narrow region in the parameter space describing their spatial structure. When applying the same methodology to true and false positives both taken from real data, the performance drops significantly. We identify two main reasons for this:

1. The number of training cases is for true positives is dramatically lower than in case of the simulations. The simulations were performed with a data set containing 38 102 total examples, 20 710 of which are true positives. The ANNs used for real-data classification consisted of 17 647 examples, only 767 of which are examples for true positives.

2. As mentioned above, real sources exhibit a larger diversity in the parameters describing their spatial structure, either caused by real sources being slightly resolved or more inaccurate measurements due to inhomogeneous data quality. This increases the overlap with false positives in the feature space used for classification and also makes generalization by the ANNs less efficient.

We address the first issue by using several methods for learning from imbalanced data sets by either data augmentation — random replication or Borderline Synthetic Minority Oversampling Technique (BLSMOTE) — or modifying the objective function. While the data augmentation methods lead to a 4 % increase in completeness, they dramatically reduce the reliability of the classification. There is clearly a need to have a larger and more diverse training data set.

The second issue becomes evident when we use the ANNs to compile an extragalactic source catalog from the pipeline output. By re-sampling the HIPASS catalog we estimate the expected number of sources detected by EBHIS to be 3207^{+347}_{-322}. The actual number of galaxies found using semi-automated classification is 1847 or about 58 % of the expected count. Using two independent statistical tests, we show that this is not due to insufficient sensitivity, but can be attributed to imperfect classification by the ANNs. Although intuitively bright sources should be easy to recognize we show that they populate regions in the parameter space that have very few examples of true positives. Additionally, there are enough false positives in these regions of the parameter space that the ANNs perform better when classifying all sources in these regions as false positives. In general, there is a strong parameter overlap between true and false positives. This can be attributed in part to the current data quality of EBHIS which still contains many artifacts. This also means that the features we choose for classification, the majority of which are standard astronomical parameters, do not possess sufficient discrimination power.

Lastly, we test the parametrization accuracy of our pipeline by comparing the measured values of F and w_{50} for 518 galaxies also contained in both EBHIS and the northern HIPASS catalog. While the measured w_{50} agrees reasonably well between the two surveys, the measured integrated flux densities disagree significantly. For sources treated as unresolved, our pipeline finds about 20 % lower integrated flux densities, whereas sources treated as resolved have close to 50 % higher integrated flux densities. While not as significant as the discrepancy described here, Haynes et al. (2011) note a 10 % difference in the integrated flux densities measured by HIPASS and ALFALFA. We conclude that the most likely reason for this discrepancy is a difference in the algorithms used for parametrization. While our simulations show that measuring the integrated flux density from the peak spectrum is accurate and unbiased, it is possible that the sources in EBHIS are resolved enough for our approach to be invalid. However, the higher integrated flux densities for resolved sources can not be explained in this way. Here we argue that our method of aperture optimization is more accurate than the method used for HIPASS.

6.2 Outlook

The methods and pipeline developed in this thesis represent a first step towards fully automated processing of large-scale H I surveys. The results from this thesis bring to light where there is significant need of improvement over the current state of the art. In this section we point out areas that are in need of improvement and hint at possible solutions. We discuss the outlook for source finding, parametrization, and classification separately.

6.2.1 Source Finding

In addition to the algorithm developed in this thesis, a wide variety of source-finding methods for spectroscopic surveys is available (e.g., Whiting 2012; Jurek 2012; Serra et al. 2012). While we argue that our approach is superior to many established methods, especially in terms of reliability, there clearly is a fair amount of time spent developing source-finding algorithms and associated software. An area of concern is the increase in data volume, which makes the application of computationally expensive algorithms like wavelet de-noising difficult.

The fact that the 2D-1D wavelet transform can be split into a 2D transform of each channel of a data cube, followed by a 1D transform along each line of sight, might make it feasible to perform all-sky source-finding for a hypothetical combined survey from WSRT/Apertif and ASKAP. Starck et al. (2010) show how the HEALPix scheme (Górski et al. 2005) can be used to calculate the isotropic un-decimated wavelet transform on the sphere. For a survey like the WALLABY H I All-Sky Survey (WALLABY, Koribalski 2012), we would need a N_{side} parameter of 16 384 to achieve full sampling for an angular resolution of 30″. This means that each full-sky channel map would require 24 GB of storage using 32 bit floating point and there would be 16 384 of these maps, one for each spectral channel. Performing source finding in this way would remove all field overlap and give many opportunities for distributed and shared memory parallelization. All 2D transforms are calculated independently for each channel map and can be distributed to a multi-node cluster. The 2D transform on the sphere requires a spherical harmonics transform of the data, which is efficiently calculated using `anafast`[1]. It is part of the HEALPix software suite and itself parallelized to take advantage of multiple cores in a shared memory system. After the 2D transform, each line of sight is independently transformed, which allows for a massively parallel processing, as there are 3 221 225 472 spectra. The most difficult part of implementing source-finding in this way is the calculation of the spherical harmonic transform for maps as large as required to achieve the desired all-sky resolution. All-sky source finding as proposed here is clearly a major challenge in terms of software and hardware requirements. But future missions to survey the cosmic microwave background, like the proposed COrE mission[2], have similar requirements which will guarantee a rapid development of the necessary software. Another approach might be to image the survey data at a lower resolution. At a resolution of 1′, the storage requirements would decrease by a factor of 12, since $N_{\mathrm{side}} = 8192$ would be sufficient.

6.2.2 Parametrization

The approach to parametrization in this thesis is mainly focused on automating time-tested methods used by H I astronomers for decades. To improve the performance of the parametrization, we introduce robust spline fitting and bilateral filtering, but do not alter the general approach to parametrization. There

[1] `http://healpix.jpl.nasa.gov/html/facilitiesnode7.htm`
[2] `http://www.core-mission.org`

are some recent developments in both the optical and H I community that are worth considering for an automated parametrization pipeline.

Recently, two independent papers propose the use of a model for the spectral shape of global H I profiles to derive parameters like the profile width, center and integrated flux density. Westmeier et al. (2014) use a piecewise analytic function to model the flanks and the central dip of a global H I profile. Although most parameters of their model have no physical meaning, they can be used to infer the typical profile parameters. Stewart et al. (2014) develop a model that is expressed in the Fourier domain and then transformed to give the profile. The parameters of this model all have physical meaning and the formulation in the Fourier domain allows to include the thermal broadening of the line profile by the gas motion in a straight-forward manner. Both of these models would allow to perform Bayesian inference of the model parameters which would result in precise uncertainties for each parameter. In Sec. 3.3.2 we show the complicated relation between the uncertainty of w_{50} and the peak flux density of a line profile. Being able to perform Bayesian inference would give a direct estimate of the uncertainties.

In this thesis, we use an iterative Gaussian fitting procedure to determine the centroid of our sources. This fitting process could be extended to derive more informative parameters describing the spatial appearance. This is especially useful for sources which are resolved by less than five beams, as these can not be reliably modeled by current tilted-ring fitting codes (Józsa et al. 2007; van der Hulst et al. 1992). Refregier (2003) introduce a set of orthogonal basis functions — the so called "shapelets" — that are used to represent images of galaxies by a small set coefficients. These coefficients can be used to succinctly describe the centroid, asymmetry, and concentration of galaxies. There is also a version of these basis functions that is expressed in polar coordinates which exploits the rotational symmetry present in astronomical sources (Massey & Refregier 2005). Typically, tens of coefficients are sufficient to represent even complex optical images of galaxies. These coefficients would also lend themselves to perform machine learning and also would allow to perform Bayesian inference similar to the probabilistic approach of Lang et al. (2014).

6.2.3 Classification

This thesis shows that the reliable, automated classification of source candidates is one of the most difficult tasks for automated survey processing. Classifying source candidates by their measured parameters clearly does not provide sufficient discriminative power. We give a short overview of alternative methods that are applicable to the problem of source candidate classification.

Instead of using pre-computed features for classification, it is common to use convolutional neural networks to learn features from the data itself, which are usually small images (LeCun et al. 1998a). The latest versions of such networks are able to recognize and locate thousands of different objects in real-world images (Krizhevsky et al. 2012; Erhan et al. 2014; Vinyals et al. 2014). Since objects like trees or cars in a real-world context are much more complicated scenes than astronomical sources in images or data cubes, it is conceivable that such neural networks could be trained to recognize H I sources from their velocity-integrated maps or even the full three-dimensional data cube. As hinted at in Sec. 4.2.1, such ANNs require vastly more inputs than the approach pursued in this thesis, but are far from being impractical as shown by the references given above. Basing the classification of sources on the data itself would furthermore allow to re-establish the order of operations humans use to analyze survey data: Find interesting objects in the data using source-finding methods, investigate whether the object is of astrophysical origin and, if so, perform parametrization. This would alleviate the need to fully parametrize all source candidates, speeding up pipeline processing and freeing up the resources for more sophisticated parametrization methods as hinted at in the previous section.

APPENDIX A

Simulated Data

For the evaluation of various algorithms and methods, simulated Effelsberg-Bonn H I Survey data is required.

A.1 Simulated Datacubes

In the following, we describe how we simulate different types of data cubes containing unresolved galaxies and simulated baseline and radio-frequency interference (RFI) artifacts.

A.1.1 Noise

Due to the data-reduction process, the noise in the Effelsberg-Bonn H I Survey (EBHIS, Kerp et al. 2011) data cubes is spatially correlated (see Chapter 1.3). The correlation length is determined by the gridding kernel used to create the data cubes. Since EBHIS data are gridded in a two step process, the final correlation length is the square-root of the quadratic sum of the individual gridding kernels. For standard EBHIS data cubes, this is 5.4′. At full spectral resolution, there is also a slight correlation between adjacent channels. Since the data are binned by a factor of eight for extragalactic applications, we assume the noise to be spectrally uncorrelated. To reproduce realistic simulated noise, we generate data cubes of uncorrelated white noise and spatially convolve it with a Gaussian kernel of 5.4′. The resulting data cube can then be scaled to have the desired amplitude. An example of this is shown in Fig. A.1.

A.1.2 Radio-Frequency Interference

The bulk of RFI encountered in EBHIS data is in the form of constant amplitude, narrow-band signals. They can therefore be very well simulated by generating a data cube having zero value everywhere, except for a few, random channels. We model the amplitude distribution of RFI by a Rayleigh distribution with a mode parameter of $0.1\,\mathrm{Jy\,beam^{-1}}$. This simple model does not account for the local standard-of-rest (LSR) correction that is applied to real data. Since the LSR correction changes over the course of a 1.5 h observation and the RFI is usually emitted at constant frequency, there is a time-depended shift of the RFI relative to the H I signal. Because of the specific mapping mode of EBHIS data, this leads to a slight spectral skewing of narrow-band RFI. We simulate this effect by first generating a data cube as described above, but applying an affine transform that skews the data along the spectral axis. The resulting data cube is a reasonable approximation of the real narrow-band RFI signal observed in EBHIS data. An example for the simulated RFI contribution is shown in Fig. A.1.

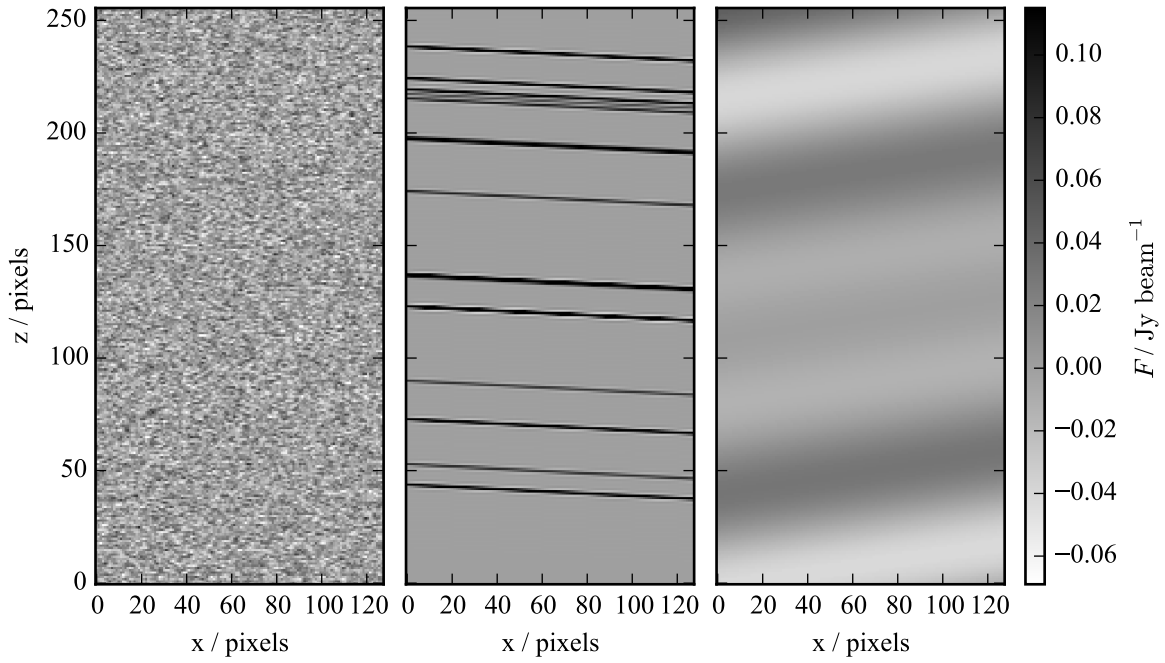


Figure A.1: Position-velocity diagrams of the contributions to the simulated data cubes. **Left**: Simulated correlated noise. **Center**: Simulated RFI. **Right**: Simulated residual baselines.

A.1.3 Baselines

The residual baseline found in EBHIS data is usually limited to the ripple caused by strong and unresolved continuum sources. Real EBHIS data is therefore mostly free of large-scale baseline ripples. For investigation of the 2D-1D wavelet decomposition it is nonetheless interesting to see, at what wavelet-scales typical baseline ripple contributes to the data. We model the large-scale baseline ripple by a superposition of squared sine functions with varying wavelength, phase and amplitude. This provides a reasonable model of the ripple found in single-dish data, which is mostly caused by multiple reflections in the support structure of the telescope. An example of the simulated baseline ripple is shown in Fig. A.1.

A.1.4 Sources

At the angular resolution of EBHIS, most extragalactic H I sources are unresolved. We can therefore simulate a large part of the sources by only generating their spectral profile. Once we have the profile, we model the beam of with a two dimensional Gaussian function and insert the source into a data cube. We generate our profiles using the model by Stewart et al. (2014). They create galaxy profiles by modeling the global profile as a superposition of a disk rotating as a solid body and a ring representing the differential rotation. Apart from the integrated flux density, which determines the overall amplitude of the profile, the model has five parameters determining its appearance: systemic velocity v_{sys}, rotational velocity v_{rot}, turbulent motion σ_{turb}, asymmetry, and the fraction of solid body rotation f_{solid}.

The systemic velocity determines the location of the profile in the synthetic spectrum. The effects of the other parameters are illustrated by the examples in Fig. A.2. We show a profile with an integrated flux density of 30 Jy km s^{-1} and shape parameters $v_{\mathrm{rot}} = 200\,\mathrm{km\,s^{-1}}$, $\sigma_{\mathrm{turb}} = 10\,\mathrm{km\,s^{-1}}$, $f_{\mathrm{solid}} = 0.2$, and

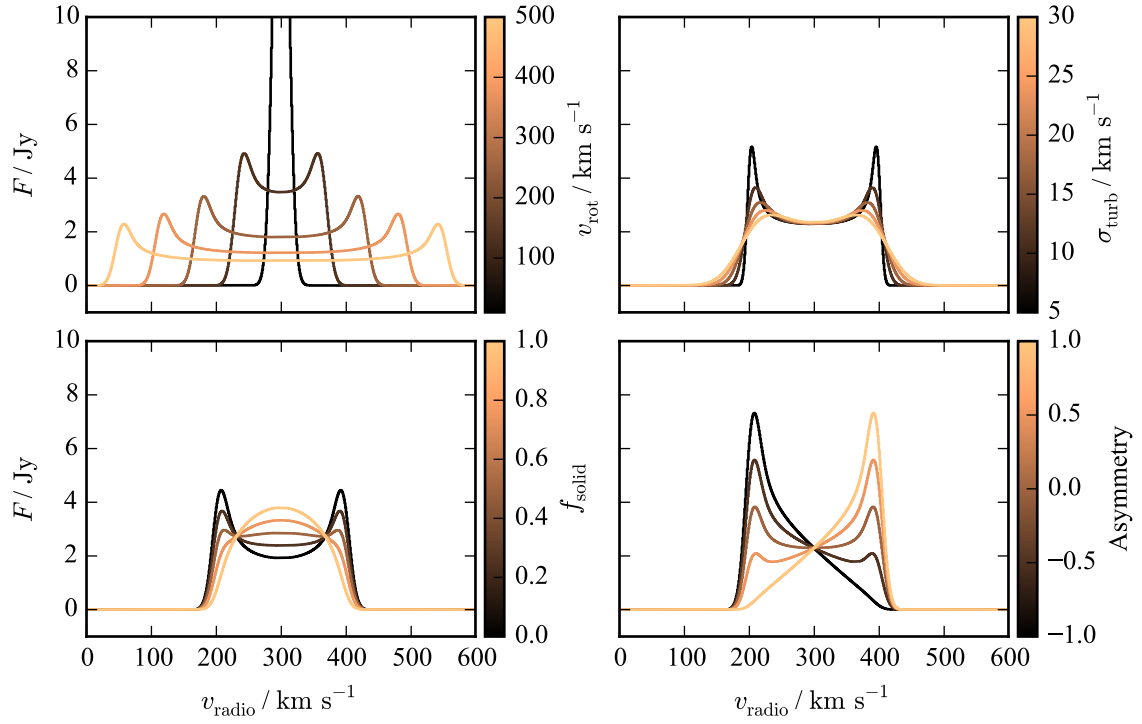


Figure A.2: Showcase of different global H I profiles generated by the model by Stewart et al. (2014). In each panel, one parameter is varied according to the values given by the color scale. The other parameters are fixed at their default values (see text). The model is evaluated at a velocity resolution of 1 km s^{-1} for visualization purposes.

zero asymmetry. The effect of the four shape determining parameters is shown by keeping the other parameters fixed and varying each one in turn. Even in its simplicity, the model is clearly capable of representing a wide variety of integrated H I profiles. The model is also capable of transitioning from a double-horned to a Gaussian appearance.

List of Abbreviations

AICC corrected Akaike information criterion.

ALFALFA Arecibo Legacy Fast ALFA Survey.

ANN artificial neural network.

ASKAP Australia SKA Pathfinder.

ATCA Australia Telescope Compact Array.

BLSMOTE Borderline Synthetic Minority Oversampling Technique.

DM dark matter.

EBHIS Effelsberg-Bonn H I Survey.

FPGA field-programmable gate array.

FWHM full width at half maximum.

GASS Parkes Galactic All-Sky Survey.

HIJASS H I Jodrell All Sky Survey.

HIMF H I mass function.

HIPASS H I Parkes All-Sky Survey.

IF intermediate frequency.

JVLA Jansky Very Large Array.

LSR local standard-of-rest.

MAD median absolute deviation.

MeerKAT South African SKA Pathfinder.

MRS multi-resolution support.

MSGD mini-batch stochastic gradient descent.

NHICAT northern HIPASS catalog.

NVSS NRAO VLA Sky Survey.

OTF on-the-fly.

PCA principal component analysis.

RF radio frequency.

RFI radio-frequency interference.

RRL radio-recombination line.

SDSS Sloan Digital Sky Survey.

SGD stochastic gradient descent.

SKA Square Kilometre Array.

SMOTE Synthetic Minority Oversampling Technique.

SNR signal-to-noise ratio.

WALLABY WALLABY H I All-Sky Survey.

WISE Wide-field Infrared Survey Explorer.

WNSHS Westerbork Northern Sky H I Survey.

WSRT Westerbork Synthesis Radio Telescope.

WSRT/Apertif upgraded WSRT.

Bibliography

Akaike, H. 1974, Automatic Control, IEEE Transactions on, 19, 716

Barnes, D. G., Briggs, F. H., & Calabretta, M. R. 2005, Radio Science, 40, 5

Barnes, D. G., Staveley-Smith, L., de Blok, W. J. G., et al. 2001, MNRAS, 322, 486

Bastien, F., Lamblin, P., Pascanu, R., et al. 2012, Theano: new features and speed improvements, Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop

Baugh, C. M., Lacey, C. G., Frenk, C. S., et al. 2005, MNRAS, 356, 1191

Bengio, Y. 2012, in Neural Networks: Tricks of the Trade (Springer), 437–478

Bergstra, J. & Bengio, Y. 2012, The Journal of Machine Learning Research, 13, 281

Bergstra, J., Breuleux, O., Bastien, F., et al. 2010, in Proceedings of the Python for Scientific Computing Conference (SciPy), oral Presentation

Booth, R. S., de Blok, W. J. G., Jonas, J. L., & Fanaroff, B. 2009, ArXiv e-prints

Bosma, A. 1981, AJ, 86, 1825

Bower, R. G., Benson, A. J., Malbon, R., et al. 2006, MNRAS, 370, 645

Briggs, F. H., Sorar, E., Kraan-Korteweg, R. C., & van Driel, W. 1997, PASA, 14, 37

Broeils, A. H. & Rhee, M.-H. 1997, A&A, 324, 877

Calabretta, M. R. & Greisen, E. W. 2002, Astronomy and Astrophysics, 395, 1077

Catinella, B., Schiminovich, D., Kauffmann, G., et al. 2010, MNRAS, 403, 683

Combettes, P. L. & Pesquet, J.-C. 2011, in Fixed-point algorithms for inverse problems in science and engineering (Springer), 185–212

Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693

Courtois, H. M., Tully, R. B., Makarov, D. I., et al. 2011, MNRAS, 414, 2005

de Blok, W. J. G., Józsa, G. I. G., Patterson, M., et al. 2014, A&A, 566, A80

De Lucia, G. & Blaizot, J. 2007, MNRAS, 375, 2

Delhaize, J., Meyer, M. J., Staveley-Smith, L., & Boyle, B. J. 2013, MNRAS, 433, 1398

Dickey, J. M., McClure-Griffiths, N., Gibson, S. J., et al. 2013, PASA, 30, 3

Donoho, D. L. 1995, Information Theory, IEEE Transactions on, 41, 613

Donoho, D. L. & Johnstone, I. M. 1995, Journal of the american statistical association, 90, 1200

Duffy, A. R., Meyer, M. J., Staveley-Smith, L., et al. 2012, MNRAS, 426, 3385

Erhan, D., Szegedy, C., Toshev, A., & Anguelov, D. 2014, in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2155–2162

Fabello, S., Catinella, B., Giovanelli, R., et al. 2011a, MNRAS, 411, 993

Fabello, S., Kauffmann, G., Catinella, B., et al. 2011b, MNRAS, 416, 1739

Farabet, C., Couprie, C., Najman, L., & LeCun, Y. 2013, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35, 1915

Fernández, X., van Gorkom, J. H., Hess, K. M., et al. 2013, ApJ, 770, L29

Flöer, L. & Winkel, B. 2012, PASA, 29, 244

Flöer, L., Winkel, B., & Kerp, J. 2010, in Proceedings of the RFI Mitigation Workshop. 29-31 March 2010. Groningen, the Netherlands

Flöer, L., Winkel, B., & Kerp, J. 2014, A&A, 569, A101

Garcia, D. 2010, Computational Statistics & Data Analysis, 54, 1167

Geha, M., Blanton, M. R., Masjedi, M., & West, A. A. 2006, ApJ, 653, 240

Gibbs, J. W. 1899, Nature, 59, 606

Giovanelli, R., Haynes, M. P., Kent, B. R., et al. 2005, AJ, 130, 2598

Glorot, X. & Bengio, Y. 2010, in International Conference on Artificial Intelligence and Statistics, 249–256

Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759

Han, H., Wang, W.-Y., & Mao, B.-H. 2005, in Advances in intelligent computing (Springer), 878–887

Haynes, M. P., Giovanelli, R., Martin, A. M., et al. 2011, AJ, 142, 170

He, H. & Garcia, E. A. 2009, IEEE Transactions on Knowledge and Data Engineering, 21, 1263

Hinton, G. E. & Salakhutdinov, R. R. 2006, Science, 313, 504

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv preprint arXiv:1207.0580

Hogg, D. W., Bovy, J., & Lang, D. 2010, ArXiv e-prints

Holschneider, M., Kroland-Martinet, R., Morlet, J., & Tchamitchian, P. 1989, Wavelets, Time-Frequency Methods and Phase Space, Springer, Berlin

Holwerda, B. W., Blyth, S.-L., & Baker, A. J. 2012, in IAU Symposium, Vol. 284, IAU Symposium, ed. R. J. Tuffs & C. C. Popescu, 496–499

Hotan, A. W., Bunton, J. D., Harvey-Smith, L., et al. 2014, PASA - Publications of the Astronomical Society of Australia, 31

Hurvich, C. M. & Tsai, C.-L. 1989, Biometrika, 76, 297

Ibrahim, J. G., Zhu, H., & Tang, N. 2008, Journal of the American Statistical Association, 103, pp. 1648

Johnston, S., Taylor, R., Bailes, M., et al. 2008, Experimental Astronomy, 22, 151

Johnstone, I. M. & Silverman, B. W. 1997, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59, 319

Józsa, G. I. G., Kenn, F., Klein, U., & Oosterloo, T. A. 2007, A&A, 468, 731

Jurek, R. 2012, PASA, 29, 251

Kalberla, P. M. W., McClure-Griffiths, N. M., Pisano, D. J., et al. 2010, A&A, 521, A17

Kalberla, P. M. W., Mebold, U., & Reif, K. 1982, A&A, 106, 190

Keller, R., Nalbach, M., Müller, K., et al. 2006, Multi-Beam Receiver for Beam-Park Experiments and Data Collection Unit for Beam Park Experiments with Multi-Beam Receivers, Tech. rep., Max-Planck-Institut für Radioastronomie

Kerp, J., Winkel, B., Ben Bekhti, N., Flöer, L., & Kalberla, P. M. W. 2011, Astronomische Nachrichten, 332, 637

Klein, B., Krämer, I., Hochgürtel, S., et al. 2008, in Ninteenth International Symposium on Space Terahertz Technology, ed. W. Wild, 192–+

Koribalski, B. S. 2012, PASA, 29, 359

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in Advances in neural information processing systems, 1097–1105

Kukar, M. & Kononenko, I. 1998, ECAI

Landweber, L. 1951, American journal of mathematics, 615

Lang, D., Hogg, D. W., & Schlegel, D. J. 2014, ArXiv e-prints

Lang, R. H., Boyce, P. J., Kilborn, V. A., et al. 2003, MNRAS, 342, 738

Lanzetta, K. M., Wolfe, A. M., & Turnshek, D. A. 1995, ApJ, 440, 435

Lavezzi, T. E. & Dickey, J. M. 1997, AJ, 114, 2437

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998a, Proceedings of the IEEE, 86, 2278

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. 1998b, in Neural networks: Tricks of the trade (Springer), 9–50

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Mangum, J. G., Emerson, D. T., & Greisen, E. W. 2007, Astronomy and Astrophysics, 474, 679

Massey, R. & Refregier, A. 2005, MNRAS, 363, 197

McClure-Griffiths, N. M., Pisano, D. J., Calabretta, M. R., et al. 2009, The Astrophysical Journal Supplement, 181, 398

Meyer, M. 2009, in Panoramic Radio Astronomy: Wide-field 1-2 GHz Research on Galaxy Evolution, 15

Meyer, M. J., Zwaan, M. A., Webster, R. L., et al. 2004, MNRAS, 350, 1195

North, D. O. 1963, Proceedings of the IEEE, 51, 1016

Obreschkow, D., Croton, D., De Lucia, G., Khochfar, S., & Rawlings, S. 2009, ApJ, 698, 1467

Oosterloo, T., Verheijen, M., & van Cappellen, W. 2010, in ISKAF2010 Science Meeting

Papastergis, E., Martin, A. M., Giovanelli, R., & Haynes, M. P. 2011, ApJ, 739, 38

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Putman, M. E., Peek, J. E. G., & Joung, M. R. 2012, ARA&A, 50, 491

Rauzy, S. 2001, MNRAS, 324, 51

Refregier, A. 2003, MNRAS, 338, 35

Reinsch, C. H. 1967, Numerische mathematik, 10, 177

Reinsch, C. H. 1971, Numerische Mathematik, 16, 451

Rots, A. H. 1978, AJ, 83, 219

Saintonge, A. 2007, AJ, 133, 2087

Sault, R. J., Teuben, P. J., & Wright, M. C. H. 1995, in Astronomical Society of the Pacific Conference Series, Vol. 77, Astronomical Data Analysis Software and Systems IV, ed. R. A. Shaw, H. E. Payne, & J. J. E. Hayes, 433

Serra, P., Jurek, R., & Flöer, L. 2012, PASA, 29, 296

Serra, P., Westmeier, T., Giese, N., et al. 2015, ArXiv e-prints

Slatton, T. G. 2014, PhD thesis

Stanko, S., Klein, B., & Kerp, J. 2005, A&A, 436, 391

Starck, J.-L. & Bijaoui, A. 1994, Signal Process., 35, 195

Starck, J. L. & Bobin, J. 2010, in Proceedings of the IEEE

Starck, J.-L., Fadili, J., & Murtagh, F. 2007, IEEE Transactions on Image Processing, 16, 297

Starck, J.-L., Fadili, J. M., Digel, S., Zhang, B., & Chiang, J. 2009, A&A, 504, 641

Starck, J.-L. & Murtagh, F. 1994, A&A, 288, 342

Starck, J.-L. & Murtagh, F. 2006, Astronomical Image and Data Analysis

Starck, J.-L., Murtagh, F., & Bijaoui, A. 1995, Graphical models and image processing, 57, 420

Starck, J.-L., Murtagh, F., & Fadili, J. M. 2010, Sparse Image and Signal Processing, Wavelets, Curvelets, Morphological Diversity (Cambridge University Press)

Stewart, I. M., Blyth, S.-L., & de Blok, W. J. G. 2014, A&A, 567, A61

Taylor, A. R. 2013, in IAU Symposium, Vol. 291, IAU Symposium, ed. J. van Leeuwen, 337–341

Tomasi, C. & Manduchi, R. 1998, in Computer Vision, 1998. Sixth International Conference on, IEEE, 839–846

Tully, R. B., Courtois, H. M., Dolphin, A. E., et al. 2013, AJ, 146, 86

Tully, R. B. & Fisher, J. R. 1977, A&A, 54, 661

van der Hulst, J. M., Terlouw, J. P., Begeman, K. G., Zwitser, W., & Roelfsema, P. R. 1992, in Astronomical Society of the Pacific Conference Series, Vol. 25, Astronomical Data Analysis Software and Systems I, ed. D. M. Worrall, C. Biemesderfer, & J. Barnes, 131

Verheijen, M., Deshev, B., van Gorkom, J., et al. 2010, ArXiv e-prints

Verheijen, M. A. W., Oosterloo, T. A., van Cappellen, W. A., et al. 2008, in American Institute of Physics Conference Series, Vol. 1035, The Evolution of Galaxies Through the Neutral Hydrogen Window, ed. R. Minchin & E. Momjian, 265–271

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. 2014, ArXiv e-prints

Vogelsberger, M., Sijacki, D., Kereš, D., Springel, V., & Hernquist, L. 2012, MNRAS, 425, 3024

Wakker, B. P. & van Woerden, H. 1991, A&A, 250, 509

Westmeier, T., Jurek, R., Obreschkow, D., Koribalski, B. S., & Staveley-Smith, L. 2014, MNRAS, 438, 1176

Westmeier, T., Popping, A., & Serra, P. 2012, PASA, 29, 276

Whiting, M. & Humphreys, B. 2012, PASA, 29, 371

Whiting, M. T. 2012, MNRAS, 421, 3242

Winkel, B. 2009, PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn

Winkel, B., Flöer, L., & Kraus, A. 2012a, A&A, 547, A119

Winkel, B., Kalberla, P. M. W., Kerp, J., & Flöer, L. 2010, The Astrophysical Journal Supplement, 188, 488

Winkel, B., Kraus, A., & Bach, U. 2012b, A&A, 540, A140

Wolfinger, K., Kilborn, V. A., Koribalski, B. S., et al. 2013, MNRAS, 428, 1790

Wong, O. I., Ryan-Weber, E. V., Garcia-Appadoo, D. A., et al. 2006, MNRAS, 371, 1855

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

Yahya, S., Bull, P., Santos, M. G., et al. 2014, ArXiv e-prints

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Zwaan, M. A., Meyer, M. J., & Staveley-Smith, L. 2010, MNRAS, 403, 1969

Zwaan, M. A., Meyer, M. J., Webster, R. L., et al. 2004, MNRAS, 350, 1210

Zwaan, M. A., Staveley-Smith, L., Koribalski, B. S., et al. 2003, The Astronomical Journal, 125, 2842

Danksagung

Kein Mensch ist eine Insel[1] und so habe auch ich vielen Leuten für diese Arbeit zu danken.

Ich möchte mich bei meinem Doktorvater PD Dr. Jürgen Kerp bedanken. Er hat mir enorme Freiheiten bei der Entwicklung meiner Arbeit geben und mich stets unterstützt. Darüber hinaus möchte ich ihm danken meine Teilnahme an mehreren internationalen Konferenzen und Workshops zu ermöglichen.

Ebenso möchte ich Prof. Dr. Pavel Kroupa für sein kurzfristiges Einspringen als Zweitgutachter und Prof. Dr. Ulrich Klein für die anfängliche Übernahme des Zweitgutachtens danken.

Mein Dank gilt auch Prof. Dr. Jochen Dingfelder und Prof. Dr. Armin Cremers für die Teilnahme an meiner Prüfungskommission als fachangrenzender und fachfremder Gutachter.

Ich möchte auch allen meinen Kollegen für die schöne Zeit am Argelander-Institut danken, insbesondere meinen Bürokollegen Shahram Faridani, Milan den Heijer und Daniel Lenz. Mein besonderer Dank gilt Benjamin Winkel der mir immer mit Rat und Tat zur Seite stand.

Auch möchte ich mich ganz herzlich bei Christina Stein-Schmitz bedanken die stets mit größter Sorgfalt und Kompetenz in alle Verwaltungsangelegenheiten erledigt hat.

Ich danke auch Anna Kumpf die mich mit viel Liebe durch meine Dissertation begleitet hat.

Zuletzt möchte ich auch meinen Eltern Ulrike und Thomas Flöer danken die mich stets in all meinen Zielen unterstützt und mir mein Studium ermöglicht haben.

[1] Frei nach John Donne, Meditation XVI