The Effelsberg–Bonn H I Survey

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Benjamin Winkel

aus

Forst/Lausitz

14. November, 2008

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn unter
`http://hss.ulb.uni-bonn.de/diss_online`
elektronisch publiziert.

Erstgutachter und Betreuer:	PD Dr. Jürgen Kerp
Zweitgutachter:	Prof. Dr. Uli Klein
Fachnaher Gutachter:	Prof. Dr. Herbert Hübel
Fachangrenzender Gutachter:	Prof. Dr. Joachim K. Anlauf

Tag der Promotion: 4. Feb 2009
Erscheinungsjahr: 2009

"If both the past and the external world exist only in the mind, and if the mind itself is controllable — what then?"

(George Orwell, "1984")

Contents

List of Figures

List of Tables

Abstract

Since Summer 2008 a new L-band 7-Feed-Array is operated for astronomical science at the 100-m radio telescope at Effelsberg. This receiver will be used to perform an unbiased, fully sampled H I survey of the whole northern hemisphere observing both the galactic and extragalactic sky in parallel — the Effelsberg–Bonn H I survey (EBHIS). The integration time per position will be 10 min towards the area of the Sloan Digital Sky Survey and 2 min for the remaining sky. The sensitivity is chosen to be competitive with the Arecibo ALFALFA and GALFA surveys which are restricted to a much smaller portion of the sky. The use of state-of-the-art digital Fast Fourier Transform spectrometers based on field programmable gate arrays — superior in dynamic range and allowing fast dumping of spectra — makes it possible to apply sophisticated radio frequency interferences (RFI) mitigation.

The EBHIS survey will be extremely valuable for a broad range of scientific disciplines ranging from the study of the low-mass end of the H I mass function in the local volume, as well as, environmental and evolutionary effects on the HIMF, the search for galaxies near low-redshift Lyman-alpha absorbers, to the analysis of multiphase and extraplanar gas, H I shells, and ultra-compact high-velocity clouds.

The thesis focuses on developing the data reduction pipeline and software. The program to perform the stray-radiation correction was already available, other tasks like RFI detection, gain-curve correction, intensity calibration, gridding, and source detection were completely redesigned. The implementation of the algorithms was done in the programming language C++, using multi-threading techniques to significantly improve the computation speeds on multi-core or -processor platforms. This aspect is crucial, as the total amount of recorded data expected during five years of observing will exceed several Terabytes.

The software was tested extensively on simulated data, not only revealing the impact of RFI signals on the results, but showing several potential bias and selection effects which need to be considered during the scientific analysis of the data.

Finally, several test observations using the new instrument were carried out at different stages of the development of the receiver. These measurements were used to evaluate the status of the 7-Feed-Array and the FPGA backends, and to test whether the data reduction pipeline provides viable results. As a cross check two of the astronomical sources, Leo T and NGC 2403, were re-observed with the old 21-cm single-beam receiver (but using the new backends). The data were analyzed and basic physical properties derived.

CHAPTER 1

Introduction

Ground-based astronomy is only feasible in few wavelength regimes, the so-called observational windows; see Fig. 1.1. Due to the absorption of the radiation at many wavelengths mainly optical and radio astronomical measurements are usually performed, as well as to some extent near-infrared (IR) observations. Many aspects of modern astronomy were only discovered using space-based telescopes (X-rays, Gamma rays, UV, Far-IR, and Sub-mm), or in some cases observatories at highest altitudes.

This is one of the reasons, why radio telescopes gained high importance. The first radio measurement of an astronomical source was done by Jansky (1932) who studied (interference) radiation from the atmosphere caused by thunderstorms and found additionally "a steady hiss type static of unknown origin". Later, it was realized that the source of this radiation was in the center of the Milky Way (MW) — the strong radio emitter Sagittarius A. To honor his achievements the flux unit commonly used in radio astronomy is called

$$1\,\text{Jansky} = 1\,\text{Jy} = 1 \cdot 10^{-26}\,\text{W}\,\text{cm}^{-2}\,\text{Hz}^{-1}.$$

G. Reber built the first radio telescope having a paraboloidal dish and performed the first survey of the (accessable) sky at various wavelengths (Reber 1940).

Even today, among the largest science projects in astronomy is the construction of giant radio telescopes, like the square-kilometre array (SKA, `www.skatelescope.org`; see Carilli & Rawlings 2004; Hall 2005) or its pathfinder experiments, such as the low-frequency array (LOFAR, `www.lofar.org`; see Falcke et al. 2007, and references therein) combining huge collecting areas (high sensitivity) and superior resolution by using thousands of relatively small antennas and making use of sophisticated receivers and antenna designs (e.g., phased aperture and plane feed arrays).The SKA will be complemented by the Atacama Large Millimeter/Submillimeter Array (ALMA; Bachiller & Cernicharo 2008) observing at higher frequencies. It will allow to measure molecular transitions and continuum radiation. The Atacama Pathfinder Experiment (APEX; Güsten et al. 2006) is already constructed and operating.

Many interesting phenomena can be studied in the radio regime. Here, the most important spectral transitions are the 21-cm line of neutral atomic hydrogen (H I) being

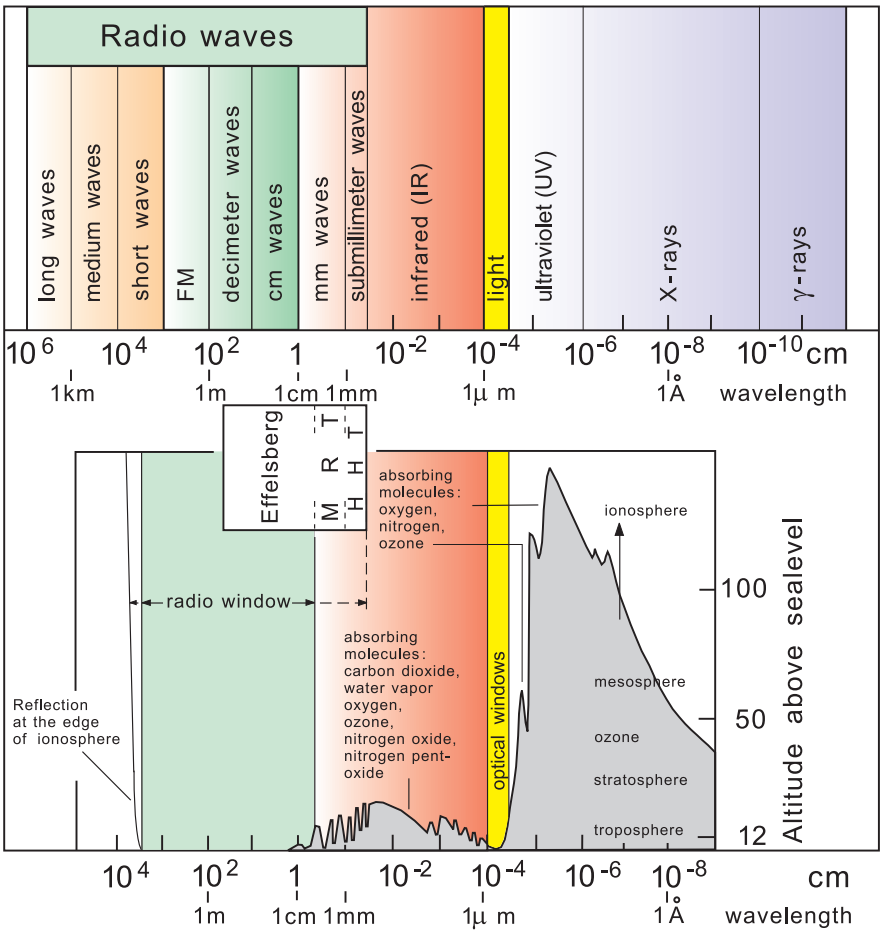


Fig. 1.1: Transmissivity of the Earths atmosphere and classification of the wavelength regimes. The atmosphere is transparent only in some so-called observational windows allowing only optical and radio astronomical observations, and to some extent IR measurements. (Courtesy of Fusshöllers Public Graphics Archive, MPIfR)

the most abundant element, and lines from simple molecules, e.g., CO. But also continuum emission plays a major role, allowing the analysis of supernova remnants (synchrotron radiation from relativistic electrons in magnetic fields) or H II regions (bremsstrahlung from thermal electrons) and many more objects.

1.1 Gas in galaxies and their environment

The space between stars and even between galaxies is not empty but contains gas, although having extremely low densities of at most $\sim 10^6$ particles per cubic centimeter. While compared to stars the density is negligible, the total mass of this gas can not be neglected, since the total volume filled is huge. Usually one refers to the gas and dust between stars in a single galaxy as the interstellar medium (ISM; typical mean density of $1\,\mathrm{cm}^{-3}$), while the baryonic matter between galaxies, having much lower densities ($10^{-6} \ldots 10^{-5}\,\mathrm{cm}^{-3}$),

Table 1.1: Gas phases in the interstellar medium of the MW. (Source: Wikipedia (en), based on Ferrière 2001, `http://en.wikipedia.org`; April, 15, 2008).

Component	Fractional Volume	Scale Height (pc)	Temperature (K)	Density (cm^{-3})
Molecular clouds	$< 1\%$	70	10–20	10^2–10^6
Cold Neutral (CNM)	1–5%	100–300	50–100	20–50
Warm Neutral (WNM)	10–20%	300–400	6–$10 \cdot 10^3$	0.2–0.5
Warm Ionized (WIM)	20–50%	1000	8000	0.2–0.5
Hot Ionized (HIM)	30–70%	1000–3000	10^6–10^7	10^{-4}–10^{-2}

is called intergalactic medium (IGM). An exception is the so-called intra-cluster medium (ICM) — gas filling the volume in large galaxy clusters — having densities of up to $\sim 10^{-3}\,\mathrm{cm}^{-3}$.

In galaxies the fraction of baryonic mass provided by the gas ranges from almost zero (i.e., in elliptical galaxies) to a few percent (in gas-rich galaxies, i.e., spirals). The MW, for example, has a total gas mass of $\sim 10^8 \ldots 10^9\,\mathrm{M}_\odot$, the mass in stars is about $2 \cdot 10^{11}\,\mathrm{M}_\odot$.

Different gaseous phases are observed in the ISM. Table 1.1 contains a compilation (Ferrière 2001) of their physical parameters for the Milky Way. The coldest and densest phase consists of molecular gas within dark clouds, being the places of star formation. In the widely accepted picture, this molecular phase can be condensed out of the cold neutral medium (CNM) mainly consisting of atomic hydrogen. Molecules are effectively formed in the presence of a catalyzer (dust particles), and under certain physical conditions. Another neutral phase is the warm neutral medium at slightly higher temperatures being more diffuse than the CNM which is mainly observed in clouds. Two ionized phases exist, the warm and hot ionized medium (WIM, HIM). The ionization is driven by hard radiation from hot stars (or the general background field) and shocks (e.g., caused by supernovae). However, the largest (mass) fraction of the material in the ISM is neutral.

The IGM and ICM consist of various gaseous phases, as well. Both, IGM and ICM are mainly ionized and can hardly be observed using the 21-cm line radiation (though the neutral H I is measured in absorption, i.e., producing the so-called Lyman-α-forest; see Bahcall & Salpeter 1965; Bahcall 1966). In the IGM, Braun & Thilker (2005) observed for the first time structure, filaments, in the 21-cm line emission of the neutral hydrogen part of the IGM connecting the galaxies M 31 and M 33. This cosmic web was known before from absorption spectroscopy. Especially, the neutral gas traces the gravitational potential of the underlying dark matter (DM) distribution.

An important aspect is that matter, once having formed stars, is eventually reinjected into the surrounding ISM by supernova explosions or stellar winds/outflows, enriching the ISM with metals[1], generated by fusion processes in stars. Not only such relatively localized feedback processes need to be considered, but the interaction of the outflows and jets of active galactic nuclei (AGN) with the host galaxy and/or the surrounding medium. Morganti et al. (2007) detected fast ($\sim 1000\,\mathrm{km\,s}^{-1}$) outflows of neutral gas (21-cm ab-

[1] In astronomy all elements having higher atomic weight than hydrogen and helium are referred to as metals.

sorption) in strong radio sources and estimated the mass outflow rates to be comparable to those of moderate starburst-driven superwinds, which must have a significant impact on the evolution of the host galaxy. The outflows and supernovae can produce density waves in their surroundings which can trigger new fragmentation processes leading to new star formation (SF). Hence, measuring the metallicity of sources can give hints about the history in terms of SF and origin of the gas.

On a much larger scale observations of the gas provides extremely valuable information, e.g., how structures may have formed during the ages of the universe. In the standard cosmological scheme, which is widely accepted today, structures have formed hierarchically in a bottom-up scenario. Dark matter clustered in the density inhomogeneities caused by microscopical quantum fluctuations, which in the early phases of the universe were inflated to macroscopical scales. Baryonic matter mainly follows the potential wells produced by DM. Gravitational interaction caused this initial density contrast to enlarge leading to the formation of DM halos. The halos grew through merging up to the super cluster sized halos observed today.

Observing other galaxies and their environment can help to better constrain the cosmological models and their parameters by comparing the predictions made by cosmological models and simulations (either purely DM based, or including baryonic matter). For example, using H I observations of Messier 31 (M31) with the Effelsberg and Green Bank 100-m telescopes, Carignan et al. (2006) measured the rotation curve out to a radial distance of ~ 35 kpc and inferred a minimum dark-to-luminous mass ratio of ~ 0.5. Weijmans et al. (2008) mapped the elliptical galaxy NGC 2974 with the Very Large Array (VLA) and claim that the previously detected H I disc in this galaxy is in fact a ring. Analyzing the velocity field provides constraints to the shape of the underlying gravitational potential. In the case of NGC 2974 data are consistent with an axisymmetric shape and to reproduce the observed flat rotation curve of the H I gas, a dark halo needs to be included in the mass model. Unfortunately, a pseudo-isothermal sphere, as well as a theoretically predicted Navarro-Frenk-White (NFW; Navarro et al. 1996) halo or modified Newtonian dynamics (MOND; Milgrom 1983) provide good fits to the observations, not allowing to rule out one of the concurrent models.

Using H I emission one is unfortunately restricted to the local universe (up to at most $z \sim 0.3$) due to the lack of sensitivity of current instruments. Only the huge collecting area of the SKA (and possibly its pathfinder experiments, e.g., LOFAR) will provide the sensitivity to detect gas at much larger redshifts. Nevertheless, 21-cm surveys (see Section 1.2.2) proved to be a valuable tool to explore the current values of the cosmological parameters. Furthermore, it is today still not very clear how galaxies and clusters form. Although it is widely accepted that larger DM halos form through merging, the detailed picture still needs to be fixed. For a long time it was thought that merging spiral galaxies form ellipticals, but recent simulations suggest, that the results of major mergers of spirals can again lead to spirals (Mayer et al. 2008).

An important role plays the accretion of matter from the surroundings of galaxies (either primordial continuously accreted matter, or residuals from tidal or ram-pressure interactions from earlier mergers, e.g., Oosterloo et al. 2007b; Putman et al. 2008; Sancisi et al. 2008). Furthermore the galaxies themselves inject material into their own (gaseous) halo by (super-)galactic winds (Chevalier & Clegg 1985; Cooper et al. 2008) or in the framework of the so-called Galactic fountain model (Shapiro & Field 1976; Bregman 1980)

were strong stellar winds and supernovae produce large shells and cavities driving the ambient medium out of the Galactic disk. Today it is widely accepted, that also magnetic fields and cosmic ray electrons play an important role in this context. However, it is still an open question which of these processes are dominant on specific types of galaxies, i.e., whether there is outflow or infall.

The hot ionized gas in the halo would then eventually cool down (forming neutral gas clouds via unstable cooling) and fall towards the plane due to gravity. Such neutral halo clouds are observed in our own galaxy and nearby sources, e.g., in the M81/M82 group observed with the Green Bank Telescope (GBT; Chynoweth et al. 2008). Several H I clouds and filamentary structures close (spatially and in velocity) to the group members were detected. In the vicinity of M31 (Thilker et al. 2004; Westmeier et al. 2005, 2007) and M33 (Grossi et al. 2008) similar clouds and complexes were found as well.

1.2 Surveys of the ISM

To extend the knowledge about the ISM in particular and all different aspects of structure formation (stars, galaxies and clusters) surveys have been carried out. In contrast to observations of individual objects they provide statistical information. Surveys cover different aspects, being blind or pointed, mapping a large fraction of the sky or only small areas (but deeper), and whether they are limited to the local universe or include highly redshifted objects. While these points are mainly determined by the amount of available telescope time, also instrumental constraints exist. Depending on the type of the backend used, surveys can be spectroscopic or photometric. While in the latter case usually much better sensitivity is provided, the additional information contained in the spectra of astronomical sources is missing. A second fundamental technical category is the type of the telescope, being a single-dish or an interferometer. While interferometers are in principle not really limited in angular resolution and collecting area — although they are limited by their construction expenses, not only for the telescope structures but also for the quadratically increasing computing power per baseline which is needed for the correlation — they suffer from the short-spacings problem: in contrast to single-dish telescopes, interferometers can not observe emission having spatial frequencies below a limit determined by the shortest possible distance between two antennas. Therefore, for more diffuse sources one still needs single-dish observations in order to measure total fluxes.

It is important to note, that most astronomical objects can, due to their physical properties, only be observed in specific wavelength regimes. Therefore, a survey carried out with a certain instrument can always deliver information only about some of the phases of the ISM, or more generally, astrophysical phenomena. This is one of the reasons why surveys are done in many wavelengths supplementing information to build up a general picture.

1.2.1 The Effelsberg–Bonn H I Survey

A new L-band 7-Feed-Array is available for astronomical measurements at the 100-m radio telescope at Effelsberg since August 2008. Using this instrument an unbiased, fully sampled H I survey of the full northern hemisphere will be performed — the Effelsberg–Bonn H I Survey (EBHIS). EBHIS will map both, the galactic and extragalactic sky, in

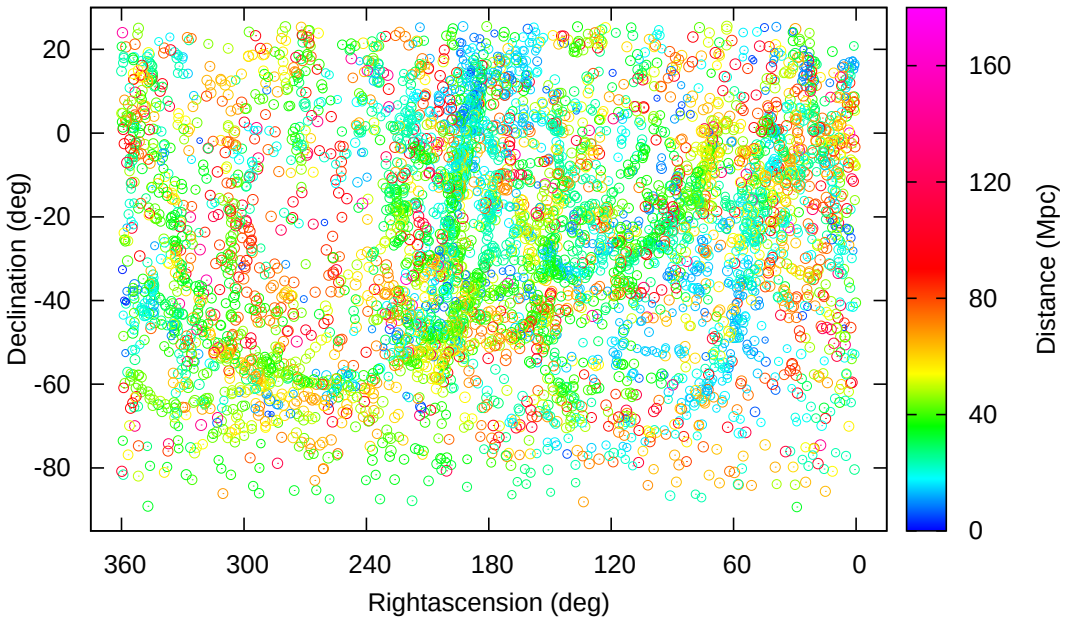


Fig. 1.2: Positions and distances of all sources in the HIPASS catalog (HICAT). The survey covers the complete southern hemisphere up to a declination of +20°. The sizes of the circles were defined such that their area is proportional to the logarithmic H I mass of each detection. Due to the large number of sources, the plot is quite crowded. Fig. 1.3 and Fig. 1.4 contain subsamples showing certain mass intervals and distance cuts, respectively, for better visualization. HICAT data were obtained through the Aus-VO Skycat data service, `http://hipass.aus-vo.org/`; see Meyer et al. (2004); Zwaan et al. (2004); Wong et al. (2006). HIPASS data was acquired with the Parkes 64-m radio telescope, operated by the CSIRO's Australia Telescope National Facility.

parallel. The integration time per position will be 10 min towards the Sloan Digital Sky Survey (SDSS; see Adelman-McCarthy et al. 2008, and references therein) area and 2 min for the remaining sky, resulting in a sensitivity much higher than that of any previously performed large H I survey. Newly developed state-of-the-art digital Fast Fourier Transform spectrometers based on field programmable gate arrays will be used as backends. They are superior in dynamic range and allow fast dumping of spectra, which is the basis for a sophisticated data reduction pipeline. Their high bandwidth of 100 MHz will cover the required redshift range (out to $z \sim 0.07$) and, at the same time, their large number of spectral channels (16k) provides a good spectral resolution of 1.25 km s^{-1}.

A comprehensive list of scientific aims and drivers for the Effelsberg–Bonn H I Survey (EBHIS) is provided in Chapter 2. Of course, one has to answer the question, whether there is really need for a new survey, as well. Before this issue is discussed (see Section 1.2.4), some of the existing surveys of the ISM will be presented.

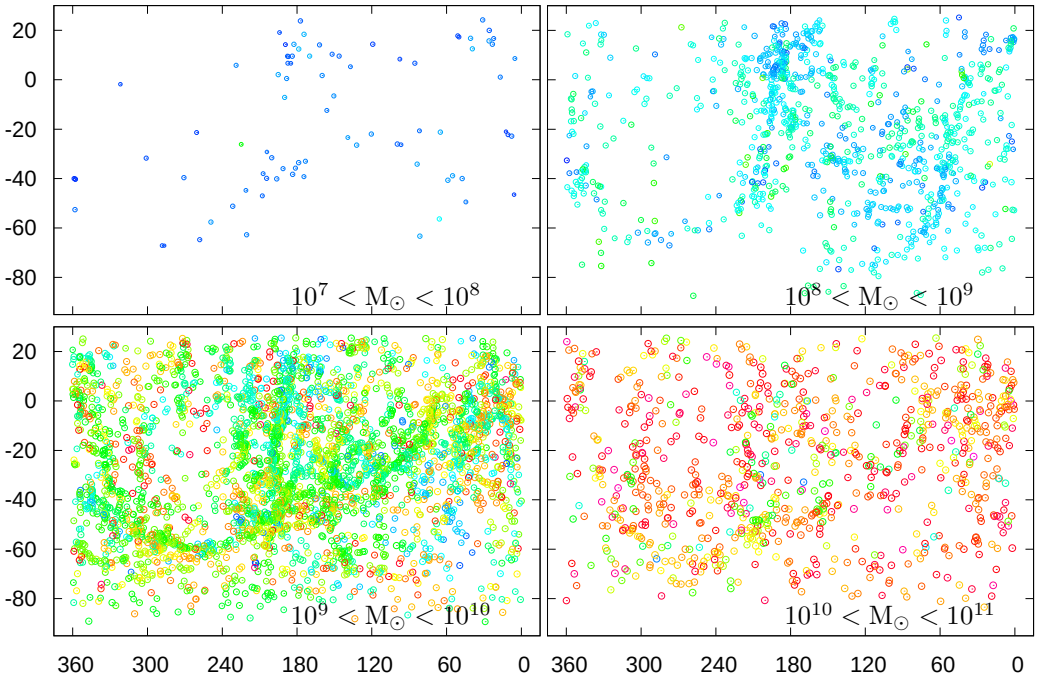


Fig. 1.3: Positions and distances of all sources from the HIPASS catalog (HICAT) for four mass intervals. Plot labels are omitted, but are, as well as the colors, identical to those in Fig. 1.2. From the top left panel it follows that only a tiny fraction of low-mass (H I) galaxies is present in the sample, showing the selection effect of a blind survey — fainter galaxies can be observed to much smaller distances than higher mass sources. The top right panel reveals the filamentary structure of the local volume, which is as well visible in Fig. 1.4.

1.2.2 Existing single-dish and interferometric H I surveys

During the last two decades several big projects to map the H I content of the Milky Way and the local volume were initiated. The H I Parkes All Sky Survey (HIPASS; Barnes et al. 2001; Meyer et al. 2004) used a 13-beam receiver to measure the complete southern hemisphere (up to a declination of 25°, covered in the northern extension, Wong et al. 2006) in a reasonable amount of time. In total 5317 galaxies were detected, providing a valuable database to infer the H I properties of galaxies in the local universe. Using the HIPASS catalog (HICAT), Zwaan et al. (2005) could calculate the local gas fraction in the universe, $\Omega_{\rm HI}$, determine the H I mass function (HIMF), as well as investigate environmental effects on the HIMF. All these quantities help to improve the basic cosmological parameters and, therefore, their determination is suited to test cosmological models and predictions based on simulations.

Fig. 1.2 shows positions and distances of all sources from the HICAT. Due to the large number of sources, the figure is not very clear. Fig. 1.3 and Fig. 1.4 contain subsamples showing mass intervals and distance cuts, respectively, for better visualization. From the top left panel in Fig. 1.3 it follows that only a tiny fraction of low-mass (H I) galaxies is present in the sample, showing the selection effect of a blind, shallow survey. Due to

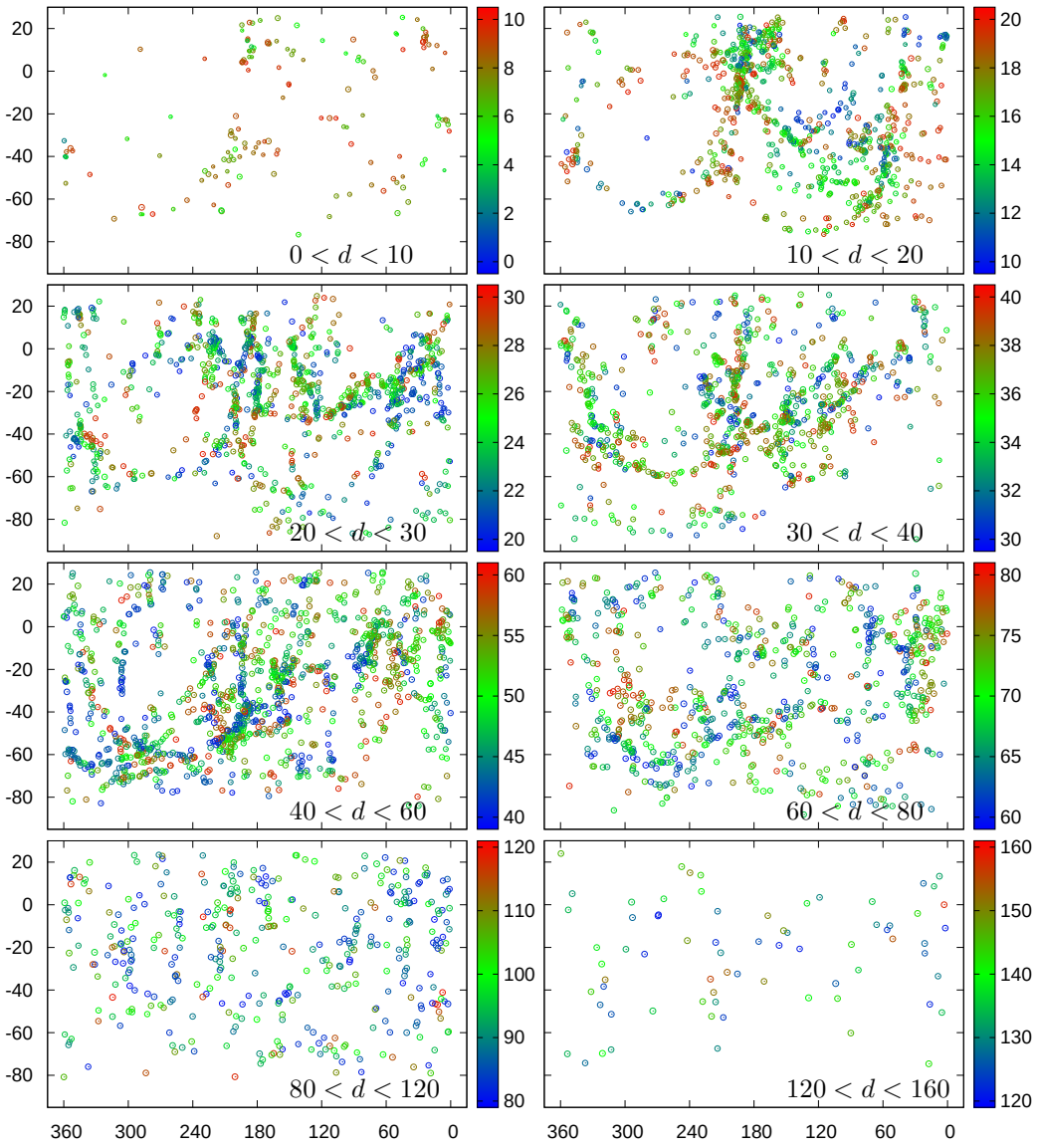


Fig. 1.4: HICAT positions and distances for different distance slices (plot labels and axes as in Fig. 1.2). In several distance ranges the filamentary structure as well as the prominent local void (at about $\alpha \approx 300°$, $\delta \approx -20°$) are clearly visible. Neglecting peculiar motions the radial velocities directly convert to distances providing the three-dimensional structure of the "cosmic web".

sensitivity limits faint galaxies can be observed only to much smaller distances than higher H I-mass sources. The top left right reveals the filamentary structure of the local volume, which is as well visible in Fig. 1.4 at several distance ranges. Furthermore, the local void (at about $\alpha \approx 300°$, $\delta \approx -20°$) shows up clearly. Neglecting peculiar motions, the radial velocities were directly converted to distances providing the three-dimensional structure of the "cosmic web".

An interesting question in extragalactic astronomy is whether Dark Galaxies exist, i.e., dark matter halos containing gas but no stellar component. Minchin et al. (2007) claim the detection of a dark galaxy, VirgoHI 21, showing evidence for rotation[2]. However, according to Haynes et al. (2007), VirgoHI 21 is embedded into a tidal feature (not surprising in the vicinity of a cluster) originating from NGC 4254 as their deeper and higher-resolution data from the Arecibo telescope reveal. It is still under debate if VirgoHI 21 is part of this tidal feature or truly a dark galaxy. Up to now, no stellar counterpart to this stream was observed. If produced by tidal forces stars should be affected as well, but their detection would not be an easy task due to the low total brightness. Recently, Bonamente et al. (2008) performed X-ray observations of VirgoHI 21 and report a non-detection (in the 0.3–2.0 keV band) of extended emission corresponding to a limit of $2.1 \times 10^{-14}\,\mathrm{ergs\,cm^{-2}\,s^{-1}}$ (99% confidence). The authors conclude, that it is unlikely (though not impossible) that VirgoHI 21 is a gravitationally-bound rotating disk within a massive DM halo.

In the northern hemisphere, the HIPASS will be supplemented by the currently ongoing Arecibo legacy fast ALFA[3] survey (ALFALFA; Giovanelli 2005), which already mapped the Virgo cluster region (di Serego Alighieri et al. 2007), and the survey our group will perform, the Effelsberg–Bonn H I Survey (EBHIS; Winkel et al. 2008). While ALFALFA is limited to a relatively small fraction of the sky (the telescope is not steerable), EBHIS will survey the complete northern hemisphere. Both, being an order of magnitude more sensitive than HIPASS, will provide a much larger database for galaxy detections.

The large-area surveys are blind surveys, which has the advantage that no bias due to pre-selecting sources enters the galaxy catalog. A drawback is the smaller effective integration time per source compared to pointed observations, like the Nançay Interstellar Baryons Legacy Extragalactic Survey (NIBLES; van Driel et al. 2008), which will measure the H I content and dynamics of about 4000 galaxies selected from the Sloan Digital Sky Survey (SDSS; see Adelman-McCarthy et al. 2008, and references therein) in the local volume. Estimations are that about 40–45% of their sources would be detected by ALFALFA and EBHIS.

In contrast to the EBHIS, which will observe both the galactic and extragalactic H I emission simultaneously, two separate galactic surveys are performed at the Parkes telescope and Arecibo. The data acquisition of the Galactic All Sky Survey (GASS; McClure-Griffiths et al. 2006) is already completed and the data are currently reduced. The Galactic ALFA (GALFA; Goldsmith 2004; Heiles et al. 2004) is still collecting data. Due to the high angular resolution and sensitivity, both surveys have revealed new interesting insights to the galactic science. McClure-Griffiths et al. (2006) find evidence for a chimney breakout in the Galactic Supershell GSH 242–03+37 supporting the idea of the galactic fountain model. The shell has a diameter of about 500 pc and an expansion energy of 10^{53} ergs. Several hundreds of hot O/B stars were necessary to produce a cavity of this size, by stellar outflows and supernovae. H I filaments associated to the source are visible up to 1.6 kpc above and below the galactic plane. This height is also found in recent simulations of Melioli et al. (2008). They performed a number of 3D hydrodynamical radiative cooling simulations of the gas in the Milky Way where the whole Galaxy structure, the Galactic differential rotation and the supernova explosions generated by a single OB association

[2] An H I mass of $3 \cdot 10^7\,\mathrm{M_\odot}$ was measured. The physical size is about 8 kpc and the inferred circular velocity is of the order of $100\,\mathrm{km\,s^{-1}}$.

[3] ALFA is the Arecibo L-band feed array, a newly installed 7-beam receiver.

(100 type II supernovae) are considered. The gas cools down and falls back to the disk (within relatively low radial distance of less than one kpc). The condensed clumps possibly match the small low-velocity clouds observed by Stanimirović et al. (2006). At least, the latter indicate that gas is still accreted onto the Milky Way disk.

Today, the largest database for galactic H I science is the Leiden/Argentine/Bonn survey (LAB; Kalberla et al. 2005) the first full-sky survey corrected for stray-radiation. At galactic radial velocities the measured brightness distribution of the source of interest can be severely distorted by signals from nearby objects entering the system via the sidelobes of the telescope beam. Stray-radiation correction is crucial for analyses requiring high sensitivity. Using the LAB, Haud & Kalberla (2007) could apply a Gaussian decomposition of Milky Way gas showing three or four groups of preferred line-widths. While two of these line-widths of about 3.9 and 24.1 $\mathrm{km\,s^{-1}}$ are well understood in the framework of traditional models of the two-phase interstellar medium a component of around 11.8 $\mathrm{km\,s^{-1}}$ most likely traces a phase transition between warm and cold neutral gas.

Levine et al. (2006a) use an unsharp masking procedure on the LAB data to reveal the spiral structure of the MW. Logarithmic spirals can be fitted to the arms with pitch angles of 20° to 25°. Furthermore, a warp is found (e.g., Levine et al. 2006b) which can be parametrized by a vertical offset ($m = 0$) and two additional Fourier modes of frequency $m = 1$ and $m = 2$ growing with radius. The $m = 2$ mode accounts for the large asymmetry between the northern and southern warps.

Another interesting result from the LAB is a revised map of the HVC sky (Westmeier 2007), as shown in Fig. 1.5. To define gas not belonging to the disk material a model of the kinematics of the Milky Way H I gas (Kalberla et al. 2007) was applied. Today, it is still an open question whether the HVCs originate from the Galactic disk itself (fountain model), from tidal and/or ram-pressure interactions with satellite galaxies, or even from primordial gas. While for low- and intermediate velocity gas the connection to the disk is very likely (although the origin of accreting material is still an issue) it is not even clear how the high velocity gas is connected to the disk material. Only for some of the large HVC complexes the origin is more or less known. Exemplary, the Magellanic System (mapped in the Parkes H I Survey of the Magellanic System; Brüns et al. 2005) consisting of the Magellanic Stream and its northern extension, the Leading Arm, are physically connected to the Small and Large Magellanic Clouds (SMC, LMC). However, there is discussion if the gas was tidally disrupted, or produced by shells (galactic fountain) and only thereafter formed the Stream due to the gravitational potential of the interacting galaxies: MW, SMC, and LMC (see Nidever et al. 2008; McClure-Griffiths et al. 2008; Stanimirović et al. 2008).

A big issue in galactic science using H I observations is the impossibility to determine distances to the gas clouds. Only indirect methods exist, using stellar absorption features in spectra of foreground and background stars (e.g., Thom et al. 2008), applying models of the galactic rotation (Kalberla et al. 2007), or correlating the H I sky with X-ray data (Pradas et al. 2004).

The VLA Galactic Plane Survey (VGPS, $l = 18° \ldots 67°$, $|b| < 1.^{\circ}3 \ldots 2.^{\circ}3$; Stil et al. 2006b), the Canadian Galactic Plane Survey (CGPS, $l = 74° \ldots 147°$, $b = -3.^{\circ}6 \ldots 5.^{\circ}6$; Taylor et al. 2003), and the Southern Galactic Plane Survey (SGPS, $l = 253° \ldots 358°$ and $l = 5° \ldots 20°$, $|b| \leq 1.^{\circ}5$; McClure-Griffiths et al. 2005) mapped the Milky Way, mainly tracing the galactic disk using interferometric measurements. The missing short

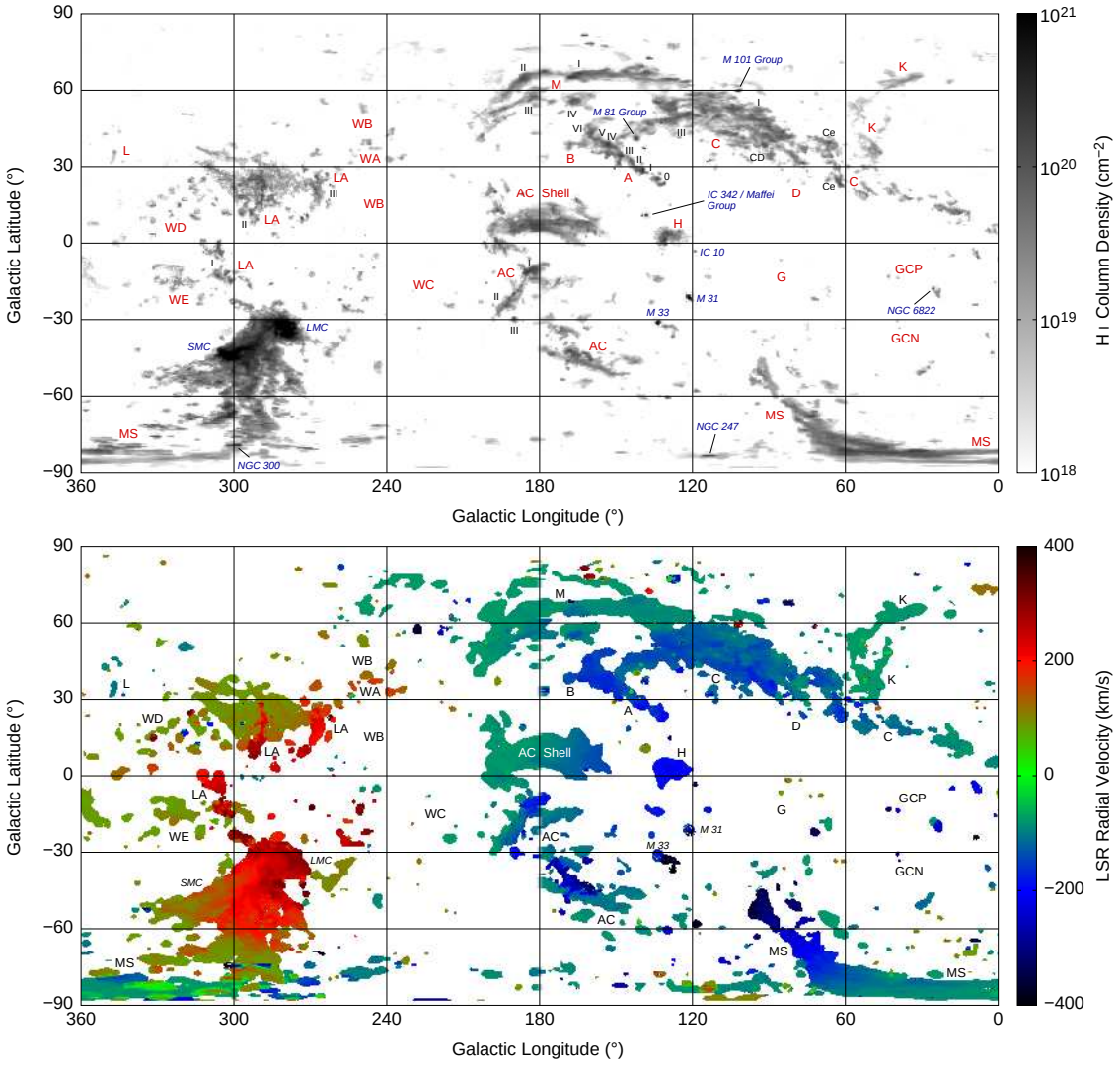


Fig. 1.5: HVC sky based on the LAB survey and a Milky Way model. The upper panel shows the total column density map, the lower panel the velocity distribution of the high velocity gas. (Courtesy of T. Westmeier, CSIRO Australia Telescope National Facility.)

spacings were provided from single-dish observations in order to reconstruct the complete fluxes. The superior angular resolution of these surveys allows to supplement the single-dish surveys, which mostly observe the WNM, with information on the phase transition between the warm and cold neutral state, as well as to study the dynamical effects of stars on the ISM. Small H I shells were discovered (e.g., GSH 23.0–0.7+117; Stil et al. 2004), most likely produced by stellar winds, which drag away the surrounding ISM leaving under-dense cavities. Stil et al. (2006a) report the detection of 17 parsec-sized clouds at large forbidden velocities, which would have been overlooked by single-dish telescopes due to beam smearing effects. The data show a wealth of substructure within the galactic disk.

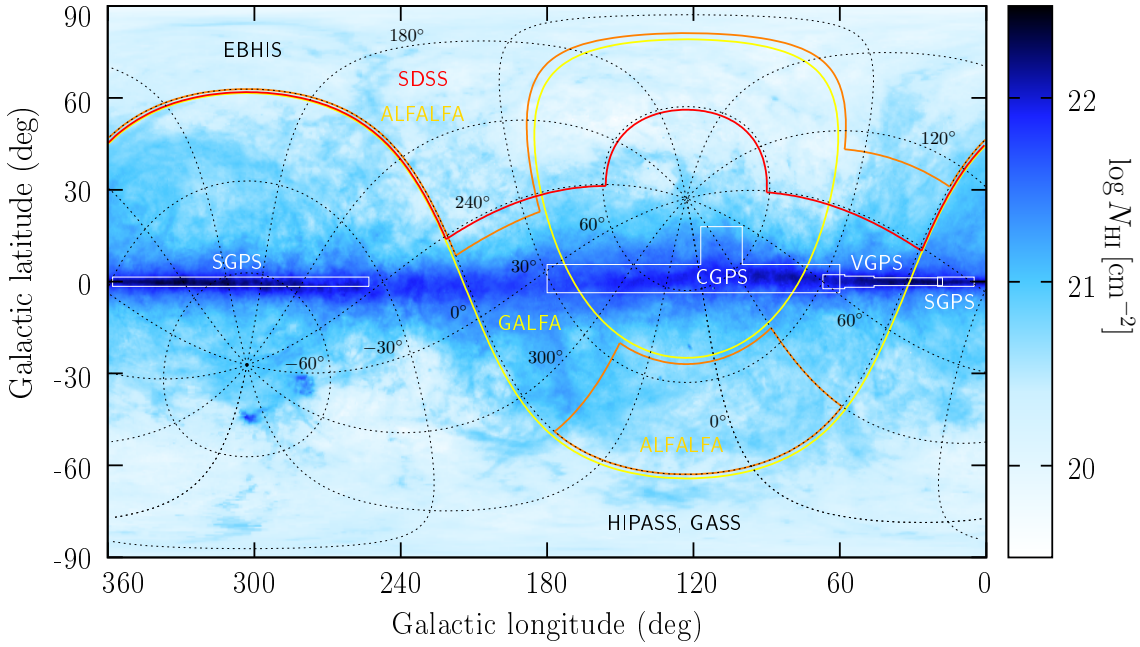


Fig. 1.6: Sky coverage of some of surveys of interest for this work overlaid on an H I column density map (integrated from -400 to $+400\,\mathrm{km\,s^{-1}}$) obtained from the LAB survey (Kalberla et al. 2005). ALFALFA and GALFA are rather limited to a small fraction of the sky (between $\delta \approx 0 \ldots 36°$). HIPASS and GASS cover the southern hemisphere ($\delta \leq 0°$, northern extension of HIPASS $\delta \leq 20°$, respectively), complemented by EBHIS ($\delta \geq 0°$). The interferometric galactic plane surveys provide high-resolution data of the Milky Way disk but only to very small galactic latitudes. SDSS data will be used to determine optical counterparts for overlapping areas. Note, that SDSS contains also a small number of stripes on the southern hemisphere, which have not been included.

Levine et al. (2008) find a vertical drop-off in the Milky Way rotation curve of about $-20\,\mathrm{km\,s^{-1}\,kpc^{-1}}$ within 100 pc of the Galactic midplane. Its magnitude is larger than expected (though consistent with measurements from halo gas of other galaxies), giving a hint on other physical processes being involved.

Also carried out with the VLA was The H I Nearby Galaxy Survey (THINGS, de Blok et al. 2005) observing 34 galaxies in the local volume ($3 \ldots 10\,\mathrm{Mpc}$) with highest resolution (FWHM $\sim 7''$, $\Delta v \leq 5\,\mathrm{km\,s^{-1}}$). The main goal was to investigate the key characteristics of galaxies along the Hubble sequence, covering a range of different evolutionary stages and physical properties. The THINGS data can be used to reveal the three-dimensional structure of the ISM down to small scales (100–300 pc) tracing the interaction of star formation with the ambient medium or measure the matter distribution in the observed galaxies (and thus also providing constraints to the dark matter potentials).

Fig. 1.6 shows for some surveys of interest the sky coverage in galactic coordinates on top of an H I column density map obtained from the LAB survey.

1.2.3 Surveys at other wavelengths

A wealth of information is contained in the (optical) Sloan Digital Sky Survey (SDSS; see Adelman-McCarthy et al. 2008, and references therein) containing after the sixth data release about 287 million (photometric) objects over 9583 $\deg^2$ also covering a large fraction of galactic longitudes and latitudes. These data are supplemented by about 1.27 million spectra of stars, galaxies, quasars, and blank sky (for calibration purposes). To describe all the scientific projects which will make use of these data is outside the scope of this work. In the context of an H I survey SDSS data can used to find the stellar counterparts of the H I detections, study correlations between the gas content and stars (e.g., Tully-Fisher relation, Tully & Fisher 1977), compute mass-to-light ratios, and determine metallicity of the stars, giving as well hints on the history of the gas.

The strongest optical emission line, Hα, plays an important role in exploring the ionized gas in the Milky Way and thus helps to estimate the general UV radiation field produced by the stellar distribution. The Wisconsin Hα Mapper (WHAM; Haffner et al. 2003) Northern Sky Survey provides spectra between local-standard-of-rest (LSR) velocities $-100 \ldots 100\,\mathrm{km\,s^{-1}}$ for Declinations $\delta \geq -30°$. While the HVC sky is unfortunately not included, several IVC detections were made and new H II regions discovered. In the meantime, several regions of the sky were reobserved to include the HVC regime, as well (Tufte et al. 2002). These measurements suggest, that the detected HVCs are located in the galactic halo and not distributed throughout the Local Group, as previously proposed by Blitz et al. (1999) and Braun & Burton (1999). In the extra-galactic regime the Wyoming Survey for Hα (WySH; Dale et al. 2008) recorded photometric data in several redshift bins between $z = 0.16 \ldots 0.81$. Hα is a good tracer of the star formation rate (SFR). Hence, the redshift binning can reveal the star formation history of galaxies. Three low column-density regions (e.g., the famous Lockman hole, Lockman et al. 1986) were chosen to avoid foreground effects as much as possible.

Dame et al. (2001) presented large-scale CO surveys of the first and second Galactic quadrants and the nearby molecular cloud complexes in Orion and Taurus. CO is an important tracer of the most abundant molecule, H_2, which can hardly be detected directly. There also exist some high-resolution surveys of individual objects, e.g., of galaxies in the Virgo cluster (Sofue et al. 2003). From the CO emission it was possible to distinct the spiral morphology (barred, armed, or amorphous). Using the data, Nakanishi et al. (2005) found an elliptical ring-like structure (semimajor axis about 720 pc) in the circumnuclear region of NGC 4569. The bar potential of this galaxy is thought to induce non-circular motions (seen in position–velocity diagrams).

Generally, observations in any wavelength regime would supplement the discussed surveys with new information about the physical conditions, kinematics and dynamics, or structure of the sources. To discuss all surveys would be outside the scope of this work, but exemplarily the SIRTF[4] Nearby Galaxies Survey (SINGS; Kennicutt et al. 2003) shall be mentioned. Here, the infrared (IR) data are complemented with deep surveys from the H I, radio continuum, CO, submillimeter, $BVRIJHK$, Hα, Paschen-α, ultraviolet, and X-ray regimes to maximize the scientific impact of the SINGS.

Finally, the pathfinder experiment for ALMA, APEX, is already operated and will be used for several surveys. E.g., a Sunyaev-Zel'dovich survey (APEX-SZ; Dobbs et al. 2006)

[4] The Space Infrared Telescope Facility, also known as the Spitzer space telescope.

of high-redshift clusters which can in combination with X-ray observations provide the density, temperature, and pressure profiles of the intracluster medium, just relying on the spherical symmetry of the cluster and the hydrostatic equilibrium hypothesis (Morandi et al. 2007). Furthermore, it is possible to calculate the angular diameter distance, and therefore, applying a joint analysis involving the baryon acoustic oscillations (BAOs) as given by the SDSS catalogue, it is possible to infer the cosmological parameter H_0 with high precision (Cunha et al. 2007).

1.2.4 Do we need a new deep H I survey?

An important question is obviously, why there is need for a new large-area H I survey in the northern hemisphere. Computing statistical properties of extra-galactic objects, as the HIMF, was already possible using the HIPASS. In galactic science with the LAB data a lot of analyses are possible. Furthermore, with ALFALFA, GASS, and GALFA going on, new deeper observations will be available shortly. However, an issue for the GALFA survey is that it is restricted to a relatively small overlap region with the galactic plane, increasing the need for EBHIS. A comprehensive list of science drivers behind the EBHIS project will be discussed in Section 2.1. Here, only some key arguments shall be noted.

A first important point is the fact, that the HIPASS data are quite restricted at the lower mass end of the HIMF due to its limited sensitivity. While ALFALFA will be an order of magnitude deeper, it is restricted to a relatively small fraction of the sky (of about $7000\,\mathrm{deg}^2$). In a statistical sense this would not be a major drawback, if the properties of the galaxies and their mass distribution would be more or less independent of the specific part of the sky which is observed. However, one must take care of the filamentary structure of the universe being the source of local over- and underdensities. Hence, the larger the survey area the better. Furthermore, many of the most interesting nearby groups and clusters of galaxies lie outside of the area accessible by the Arecibo telescope. It should also be mentioned that at the moment there is no extragalactic database for the northern hemisphere comparable to HIPASS.

For galactic observations, the EBHIS will prove valuable not only in the search for small-scale structure, e.g., ultra-compact HVCs (not detected by LAB due to spatial undersampling and poor angular resolution), but also providing deep measurements needed by future experiments which need to correct the galactic foreground contamination. This is not only true for X-ray observations but will be crucial also for the Planck mission (measuring the cosmic microwave background, CMB) being sensitive to dust, which recently was claimed to be detected in the HVC Complex C. Miville-Deschênes et al. (2006) conclude, that "in order to separate the HVC emission from the Galactic cirrus emission, the use of 21-cm observations will be mandatory".

1.3 Structure of this work

As the scientific drivers behind the EBHIS were only briefly discussed, in Chapter 2 a comprehensive list of possible projects is presented. The survey will be carried out in several stages, beginning with test observations, mapping most interesting objects. Then the SDSS area is observed and finally the complete northern hemisphere. A roadmap

and the current status of the instrument are described. Telescope, receiver and backend properties are included as well.

The most work while preparing the survey went into the development of data reduction software. A sophisticated RFI mitigation scheme, stray-radiation correction, intensity and bandpass calibration, and gridding of the data are discussed in Chapter 3. Performance-critical routines have widely been programmed to make use of multi-processor or multi-core computers to improve runtimes. For the extra-galactic data a graphical user interface for the (automatic and manual) search and parametrization of sources is provided.

To test the data reduction software and analyze the impact of RFI on the results simulations were performed emulating realistic spectra containing extra-galactic sources (Chapter 4). The resulting parameters (such as peak and total intensities, velocity widths, etc.) were compared to the generated properties. Using several statistical tests, bias effects were found, being partly due to (expected) selection effects, but also due to the specific parametrization scheme, which was adopted from the HIPASS galaxy catalog. The latter is an important result and should be considered when evaluating the survey. Such simulations also play an important role in the determination of the completeness function for EBHIS which should be relatively easy based on this work.

Finally, the results from first test observations are presented in Chapter 5, followed by the summary (Chapter 6). During these tests several technical issues with the receiver were observed. The overall quality of the new system is described in terms of bandpass and noise stability and system temperature. The long-term properties were investigated during the mapping of two interesting nearby galaxies — the dwarf galaxy Leo T and the spiral galaxy NGC 2403. The spectra have been processed and gridded to datacubes. Basic physical properties were calculated being compatible with previous findings.

CHAPTER 2

The Effelsberg–Bonn H I Survey

A new unbiased fully sampled H I survey — abbreviated as Effelsberg–Bonn H I Survey (EBHIS) — of the full northern hemisphere will be performed using a new L-band 7-feed array installed at the 100-m telescope in Effelsberg (Fig. 2.1). Both, the galactic and extragalactic sky are observed simultaneously utilizing state-of-the-art FPGA-based digital Fast Fourier Transform spectrometers having superior dynamic range and making fast dumping of spectra possible. The integration time per position will be 10 min towards the SDSS area (8 500 sq.deg) and 2 min for the remaining sky (20 600 sq.deg) pushing sensitivity to be comparable to that of the Arecibo ALFALFA (Giovanelli 2005) and GALFA (Goldsmith 2004; Heiles et al. 2004) surveys (see Fig. 1.6 for sky coverages).

The survey is a joint project of the Argelander-Institut für Astronomie (AIfA) der Universität Bonn and the Max-Planck-Institut für Radioastronomie (MPIfR) Bonn and is funded by the Deutsche Forschungsgemeinschaft (DFG).

The regular measurements will start in the beginning of 2009. In the first four years of observing the SDSS area will be mapped, the remaining part of the northern sky will take about one year.

In this chapter the scientific aims of the EBHIS are presented in detail, as well as the technical properties of the receiving system, containing the telescope, front-end (receiver), intermediate frequency chain, back-end (spectrometer), and telescope software issues. Furthermore, important survey parameters are compared to other prominent H I surveys. The data reduction pipeline itself is covered in Chapter 3 and the current status of the instrument will be discussed in Chapter 5.

2.1 Scientific aims

The following main scientific questions have been identified to be covered by the EBHIS. For the purpose of clarity, the list is separated into extragalactic tasks and analyses of (the vicinity of) the Milky Way.

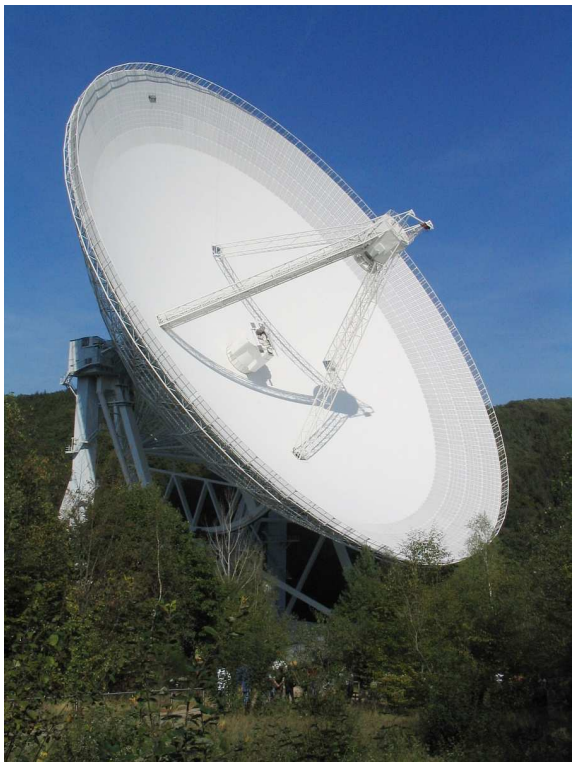


Fig. 2.1: The 100-m telescope in Effelsberg. The new 7-feed array receiver for the EBHIS is placed in the primary focus of the telescope. (Photograph: Courtesy of T. Westmeier.)

2.1.1 EBHIS — Extragalactic Survey

The survey is especially designed to provide a complete census of H I galaxies with a mass range down to $10^7\,M_\odot$ at the distance of the Virgo cluster (16 Mpc). This allows very detailed studies of the H I mass content in the local universe. The backends in use have a bandwidth of 100 MHz corresponding to a maximum observable redshift of $z = 0.07$.

The H I mass function of galaxies in the local universe

One of the fundamental problems of modern astronomy is the quantitative discrepancy between the predicted number of theoretically predicted low mass dark matter halos (with less than $10^8\,M_\odot$) and the number of observed dwarf satellites in the Local Group (Klypin et al. 1999; Moore et al. 1999). This fact is well known as the "missing satellite problem". One possible explanation would be, that the mini-halos exist, but did not form stars. However, baryonic matter should have settled down in their gravitational potential and might be observable in H I.

Nevertheless, today no "dark" galaxy (containing a substantial amount of gas embedded in a DM halo, without any star formation) was unambiguously detected so far. Minchin et al. (2007) claim the detection of a dark galaxy, VirgoHI 21. However, according to Haynes et al. (2007) VirgoHI 21 is embedded into a tidal feature (and most likely simply part of it) originating from NGC 4254 as their deeper and higher-resolution data from the Arecibo telescope reveal. Whether dark galaxies exist or not is an issue which might be solved by deeper surveys like EBHIS and ALFALFA.

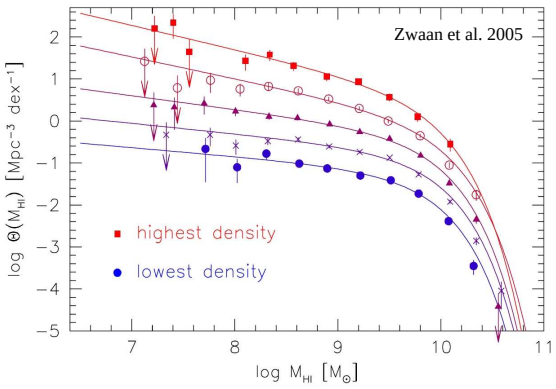


Fig. 2.2: The dependence of the H I mass function on the local environment (number density) as observed by Zwaan et al. (2005) using HIPASS data. The slope of the low-mass end of the HIMF is steeper in higher (number) density regions, meaning there are more low-mass galaxies relative to high-mass sources. This is expected, as high-density regions trace groups and clusters which contain more gas-poor (early-type) galaxies, than found in the field.

An important statistical measure for the structure of the local universe is the H I mass function (HIMF) which is the number of galaxies per mass interval. It is usually fitted with a Schechter function (Schechter 1976); see Eq. (4.2).

Using sophisticated methods previous extra-galactic surveys (e.g., HIPASS) allowed to derive the H I mass function of the local universe independent of local over- and under-densities associated with the cosmic web. Using the mass function, the total H I matter content $\Omega_{\rm HI}$ of the local universe can be determined. However, due to sensitivity limits the low-mass part of the HIMF is still not well-constrained, as the catalogs contain few sources in the regime of $M_{\rm HI} \lesssim 10^7\,{\rm M}_\odot$. While deep surveys exist as well, they are not useful to determine the global HIMF, covering only a small fraction of the sky. The EBHIS, and the ALFALFA will be suited to measure the H I mass function with unprecedent quality. Although, the collecting area of Arecibo is ten times larger than that of Effelsberg EBHIS is going to have comparable sensitivity limits by spending 10 minutes of integration time on each position. In the overlap regime with the SDSS area, being 8 500 square degrees, which is four times larger than the ALFALFA area (2 000 sq.deg).

The high sensitivity together with the large mapped area leads to an expected high number of low-mass galaxies. It will be possible to calculate the HIMF for subsets of the data, either separated into several redshift bins to trace evolutionary effects[1], or into samples of different number densities, showing how galaxies rely on their environment. The latter analysis was performed by Zwaan et al. (2005) using HIPASS data. Their results imply that in high-density regions (groups, clusters) the relative fraction of low-(H I-)mass galaxies compared to high-mass galaxies is higher than in the field; see Fig. 2.2. However, the total number of sources is not sufficiently high to provide good constraints on the results (especially in the low-mass regime).

The Cosmic Web

While the HIMF is a valuable (statistical) tool to analyze the properties of galaxies it neglects[2] the tendency of matter to form structures, filaments, groups, and clusters —

[1] Clearly, the low-mass regime ($\sim 10^7\,{\rm M}_\odot$) can only be investigated for small redshifts up to about $z = 0.004$ (20 Mpc). As the mass detection limit goes roughly with the quadratic distance, the HIMF should be determinable for $\gtrsim 10^8\,{\rm M}_\odot$ up to $z = 0.02$ (80 Mpc).

[2] An exception is of course the analysis of the HIMF in different environments.

the "Cosmic Web". In fact, not only the number of satellites is an important measure for cosmological models, but the distribution of (different types of) galaxies, as well. By performing a spectroscopic blind survey the three-dimensional structure of the local universe can be explored (compare to Fig. 1.2), showing the filamentary structure and voids in-between. Basilakos et al. (2007) compute the two-point correlation function based on the HICAT sources. Their preliminary result is that the correlation function seems to be anti-biased with respect to the underlying matter fluctuation field.

Galaxies and their environment

Out to a distance of few Mpc H I clouds down to masses of about $10^5\,M_\odot$ can be detected. Accordingly, it is feasible to determine the local neutral baryon fraction towards the barycenter of the Local Group of Galaxies (comprising the large spirals Andromeda (M 31), the Milky Way, and their satellite galaxies).

Not only the local group is of interest, but gas around other nearby galaxies, tracing ongoing merging, tidal interactions, and accretion of gas. The large number and variety of galaxy clusters and groups in the northern sky allows to study galaxy evolution and mass accretion as a function of different environmental conditions or mass concentrations. The overlap with SDSS makes it possible to correlate optical properties of the clusters and groups as a function of the H I mass distribution. Gas in the vicinity of galaxies plays an important role in refueling them and, thus, has a big impact on the star formation history.

In this context several detections of high-velocity gas around other galaxies (in H I and metal absorption lines, e.g., Mg), very likely being the counterpart to the HVCs around the Milky Way, is not surprising. Nevertheless, the main origin of HVCs is still under debate, either produced in gravitational interactions between galaxies, by galactic fountains, or being primordial. At least it is clear today (e.g., Westmeier 2007), that HVCs are not the "missing satellites" as proposed by Blitz et al. (1999).

The observations of the neutral gas content can be supplemented by X-ray data to provide information on the ionized state of the intergroup and intracluster medium, tracing the hot phases of the gas produced by shock-heating and photoionization (UV, X-ray) in the radiation field of galaxies.

High-mass H I galaxies

Of particular interest are the highest mass H I galaxies, as only they can be traced out to largest distances — the redshift limit of EBHIS is about $z = 0.07$ (with a mass detection limit of about $10^{10}\,M_\odot$). Evolutionary effects as a function of redshift should be measurable, since the $(1 + z)^3$-dependence means an enhancement of any evolutionary effect of up to 23% at the limit of the frequency coverage (e.g., Schneider 2006). Unfortunately, one is biased towards gas-rich galaxies, while, for example, early-type ellipticals are not traced being relatively gas-poor.

Search for galaxies near low-redshift Lyman-alpha absorbers

This might be a short term project within the Effelsberg survey, because one can start from an already existing source catalogue and search for H I emission towards particular lines of sight. If H I emission is detected, follow-up measurements with sensitive radio

interferometers are necessary to clarify the origin of the absorption by the galaxy halo or by accreted neutral clouds within the gravitational potential of the mass concentration.

Furthermore, one can search for H I absorbers against very bright background continuum sources, which has the advantage, that the optical depth τ can be determined. A difficulty in the measurements is, however, the degradation of the baselines due to the strong continuum sources.

2.1.2 EBHIS — Milky Way Survey

While an extragalactic H I survey is mainly about the detection of unresolved sources (compared to the single-dish beam size of about $9'$), in the galactic regime the diffuse extended emission makes a very precise calibration of the data necessary. This is not only technically challenging, regarding for example the receiver stability and advanced baseline algorithms, but must incorporate a correction for stray radiation (SR). This is emission from (strong) Milky Way H I sources entering via the sidelobe pattern of the telescope. EBHIS will benefit from the experience at the AIfA, where the SR corrections for the LAB survey and the GASS were successfully applied.

The Milky Way halo

Recent studies suggest a physical connection of low- and intermediate-velocity CNM clouds with the Milky Way disk (e.g., Stanimirović et al. 2006), being an important source of refueling the Milky Way by means of accretion. Still, it remains unclear which processes are mainly involved (candidates are galactic fountains, produced by H I shells and the fragmentation of the outflows, IGM accretion, tidal remnants, etc.), but EBHIS, as well as GASS and GALFA, will play an important role in studying this gas.

The discovery of parsec or AU-sized clouds within the Galactic halo by Lockman (2002) focused our interest on the tiniest structures which can exist within the extreme environment of the Milky Way halo. The Effelsberg survey will be the major resource to determine the mass and size spectrum of structures within the Milky Way halo, because of its unique signal-to-noise ratio, dense angular sampling and sensitivity of $10^{18}\,\mathrm{cm}^{-2}$ (24 mK RMS at a line width of $\Delta v = 20\,\mathrm{km\,s}^{-1}$; 11 mK RMS towards the SDSS area).

Supplemented by metal (e.g., Ca II and Na I) absorption line spectroscopy against background quasars (see Ben Bekhti et al. 2008), H I observations can be used to infer the chemical composition of the clouds in the galactic halo, giving hints about the physical origin and evolution of the gas. While this kind of analysis is only possible for a limited number of sightlines, a complete statistical census of all H I clouds can be used to infer large-scale kinematics, measure the gravitational potential of the MW, or determine the net mass infall/outflow rate due to the clouds. Furthermore, the emission line profiles will be analyzed, investigating the multiphase structure. From the LAB survey it is already clear that not only two phases exist, but at least one interface regime connecting the CNM and WNM (Haud & Kalberla 2007).

High-Velocity Clouds

While being part of the Milky Way halo, HVCs are an interesting topic on its own. High-velocity cloud research is a traditional theme of radio astronomy at Bonn University. Here,

the interaction of HVCs with the Galactic halo were discovered (Kerp et al. 1994; Pietz et al. 1996; Brüns et al. 2001). Kerp et al. (1999) performed a large-scale correlation with ROSAT All-Sky-Survey data. Of special interest is the transition region between dwarf galaxies and compact high-velocity clouds (CHVC), as well as the frequency and mass spectrum of the ultra-compact high velocity clouds (UCHVC), discovered almost simultaneously by Brüns & Westmeier (2004) and Hoffman et al. (2004). They might trace a population of Galactic clouds that have been overlooked so far, due to the relatively low angular resolution of existing H I surveys.

Many of the known CHVCs show unique structures, e.g. head-tail shapes (Brüns et al. 2000), which possibly form through ram-pressure interaction stripping away the warm neutral envelopes around the cold cores. Such clouds give valuable constraints on the ambient medium in the halo, when distance limits can be provided (e.g., by using absorption measurements; see Thom et al. 2008). The resolution and sensitivity of EBHIS will enable to identify a much larger number of sources than are known today.

H I shells

The VLA Galactic Plane Survey (VGPS; Stil et al. 2006b), the Canadian Galactic Plane Survey (CGPS; Taylor et al. 2003, 2002), and the Southern Galactic Plane Survey (SGPS; McClure-Griffiths et al. 2005) provide a wealth of information on shells and cavities produced by the stellar evolution within the Milky Way. The shallower Effelsberg survey of the low Galactic latitude sky will supplement these radio interferometer surveys with information on the warm neutral gas with unique signal-to-noise quality data. This will allow to obtain a coherent view of the evolution of shells within the solar neighborhood and the inner galaxy, and of the imprint of the density waves on these structures, only traceable by the warm gas.

H I shells probably play an important role in galactic fountain processes. McClure-Griffiths et al. (2006) find evidence for a chimney breakout in the Galactic Supershell GSH 242–03+37 supporting the idea of this model. It is remarkable that, to produce a cavity of that size, at least several hundred hot (O/B) stars are necessary, which provide the energy to pull away the surrounding medium by strong stellar winds and supernova. A single association of stars would not be sufficient, but several generations of stars are needed.

X-ray absorption

Future X-ray missions like SIMBOL-X (Ferrando 2002; Slane et al. 2008) need to observe the early universe through the X-ray attenuating gas distribution of the Milky Way[3]. The warm neutral medium is, by an order of magnitude, the most efficient absorber for soft X-ray photons with energies less than 0.3 keV (Kerp & Pietz 1998; Kerp 2003). To overcome the "cosmic conspiracy" it is necessary to observe the sky through the Lockman window or the Chandra Deep Field South window, but towards multiple lines of interest with extremely well-studied H I column density distributions. EBHIS in combination with

[3] XEUS and Con-X will be transformed into a joint project, under the name International X-ray Observatory (IXO).

the Parkes narrow band survey will be the standard resource to identify these regions of "simple" soft X-ray absorbing regions.

2.2 Telescope properties

2.2.1 Antenna

The Effelsberg radio telescope is located at longitude $6^\circ 53' 00\overset{''}{.}3$ east and latitude $50^\circ 31' 30''$ north at an altitude of 319 m. Its primary mirror is 100 m in diameter. The original secondary mirror (subreflector) was replaced by a new subreflector providing a much higher surface accuracy of about 60 μm (the old surface had 800 μm); see Bach et al. (2007). It has a radius of 3.25 m and an elliptical shape with a major axis of 14.3 m and a minor axis of 7.4 m. The main advantages are the possibility to adjust the position using six parameters (three before) and due to changes in the construction of the primary focus cabin so-called multi-frequency receiver boxes can be fitted allowing to position up to four different receivers simultaneously into the prime focus. Therefore, the desired receiver (in primary focus) can be brought online in a few minutes. The telescope is operated at frequencies between 400 MHz and 96 GHz as a multipurpose instrument used for continuum, spectroscopy, pulsar and VLBI observations. Table 2.1 contains a list of basic technical properties of the antenna.

To shield the antenna from unwanted radio frequency interference (RFI) produced by the technical equipment in the observatory a Faraday-room was built, where backends, etc., are placed in. A LOFAR station is operated at the site, providing long baselines to the central LOFAR array mainly placed in the Netherlands. Appropriate measures have been applied to shield the telescope against RFI from this station (Reich 2006).

For the description of the telescope properties some quantities must be introduced, the notation of which is adopted from Rohlfs & Wilson (1996). Most antennas have direction-dependent sensitivity to incoming radiation. This so-called power pattern is defined as

$$P(\vartheta, \varphi) = |\langle \vec{S} \rangle| \tag{2.1}$$

which is the mean absolute value of the Poynting flux $|\langle \vec{S} \rangle|$ the antenna would radiate if used as emitter. Usually, the normalized power pattern

$$P_\mathrm{n}(\vartheta, \varphi) = \frac{1}{P_\mathrm{max}} P(\vartheta, \varphi) \tag{2.2}$$

is used. To measure $P_\mathrm{n}(\vartheta, \varphi)$ one can use a small point-like radio source. The directive gain

$$G(\vartheta, \varphi) = \frac{4\pi P(\vartheta, \varphi)}{\int \mathrm{d}\Omega \; P(\vartheta, \varphi)} \tag{2.3}$$

is the fractional power from a certain direction compared to the mean power per unit solid angle which is received/emitted.

The beam solid angle of an antenna is given by

$$\Omega_\mathrm{A} = \int_{4\pi} \mathrm{d}\Omega \; P_\mathrm{n}(\vartheta, \varphi) \,. \tag{2.4}$$

Table 2.1: Technical data of the 100-m radio telescope at Effelsberg. (Source: `http://www.mpifr-bonn.mpg.de/div/effelsberg/antenna/antenna_spec.html`; May, 4th, 2008)

Reflector Diameter	100 m
Aperture	7 854 m^2
Number of Surface Elements (Panels)	2 352
Shape Accuracy of Surface	< 0.5 mm
Focal Length in Prime Focus	30 m
Secondary Mirror Diameter (Gregory-Reflector)	6.5 m
Aperture Stop	
– in Prime Focus	f/0.3
– in Secondary Focus	f/3.85
Angular Resolution (Beam Width)	
– at 21cm wavelength (1.4 GHz)	$9'.4$
– at 3cm wavelength (10 GHz)	$1'.15$
– at 3.5mm wavelength (86 GHz)	$10''$
Azimuth Track Diameter	64 m
Setting Accuracy of Track	± 0.25 mm
Azimuth Range	480°
Maximum Rotation Speed	30°/min
Pointing Accuracy	
– Blind Pointing	$10''$
– Repeatability	$2''$
Power Output of the 16 Azimuth-Drives	10.2 kW each
Radius of Elevation Gear Track	28 m
Elevation Movement	from 7° to 94°
Maximum Tilt Speed	16°/min.
Power Output of the 4 Elevation-Drives	17.5 kW each
Total Weight	3 200 t
Construction Period	1968–1971
Height of Track above Sea Level	319 m
Commencement of Operation	August 1st, 1972
Constructed by	AG KRUPP/MAN

Analogous the main beam solid angle is defined as

$$\Omega_{\mathrm{mb}} = \int_{\substack{\mathrm{main}\\ \mathrm{beam}}} \mathrm{d}\Omega \; P_{\mathrm{n}}(\vartheta, \varphi) \,. \tag{2.5}$$

Using these quantities the beam efficiency can be inferred

$$\eta_{\mathrm{B}} = \frac{\Omega_{\mathrm{mb}}}{\Omega_{\mathrm{A}}} \,. \tag{2.6}$$

The beam efficiency indicates how much of the power pattern is concentrated in the main beam. The directivity D is defined as the maximum gain

$$D = G_{\max} = \frac{4\pi}{\Omega_{\rm A}} \,. \tag{2.7}$$

The antenna receiving a electro-magnetic (plane) wave extracts a certain amount of power $P_{\rm e}$ from the wave. The fraction

$$A_{\rm e} = \frac{P_{\rm e}}{|\langle \vec{S} \rangle|} \tag{2.8}$$

is called effective aperture of the antenna and is connected to the geometric aperture via

$$A_{\rm e} = \eta_{\rm A} A_{\rm g} \,. \tag{2.9}$$

It can be shown that the relation

$$D = G_{\max} = \frac{4\pi A_{\rm e}}{\lambda^2} \tag{2.10}$$

is valid for all antennas.

In the Rayleigh-Jeans limit the total power per unit bandwidth received by an antenna is given by

$$W = \frac{1}{2} A_{\rm e} \int {\rm d}\Omega \; T_{\rm b}(\vartheta, \varphi) P_{\rm n}(\vartheta, \varphi) \,. \tag{2.11}$$

In radio astronomy often the equivalent antenna temperature $T_{\rm A} = W/k_{\rm B}$ is used which leads to

$$T_{\rm A}(\vartheta_0, \varphi_0) = \frac{\int {\rm d}\Omega \; T_{\rm b}(\vartheta, \varphi) P_{\rm n}(\vartheta - \vartheta_0, \varphi - \varphi_0)}{\int {\rm d}\Omega \; P_{\rm n}(\vartheta, \varphi)} \,. \tag{2.12}$$

Hence, the measured antenna temperature, $T_{\rm A}$, is the convolution of the true brightness distribution, $T_{\rm b}(\vartheta, \varphi)$, of the sky with the antenna diagram (or pattern) $P_{\rm n}(\vartheta, \varphi)$. The possible spatial resolution of observations is determined by the directivity of the antenna pattern. Following Eq. (2.10), to increase the directivity (making the beamwidth smaller) for a fixed wavelength the effective aperture must be enlarged. This is one of the reasons why single-dish radio telescopes are as large as possible.

The quantity finally measured is $T'_{\rm A}$ which also accounts for atmospheric and electronic losses. However, as of physical interest is the true brightness temperature distribution on the sky, Eq. (2.12) must be inverted using numerical approaches; see also Section 3.4 which discusses this so-called stray-radiation correction. Furthermore, $P_{\rm n}(\vartheta, \varphi)$ is known only for a limited range of (ϑ, φ)-values and suffers from measurement errors. Consequently, the power pattern needs to be modelled to some extent (Kalberla 1978).

Of practical importance is the so-called aperture function $g(x, y)$ (for plane apertures). The power pattern is directly related to the aperture via Fourier transformation

$$P_{\rm n}(\vec{n}) = \frac{|f(\vec{n})|^2}{|f_{\max}|^2} \tag{2.13}$$

where

$$f(\vec{n}) = \frac{1}{2\pi} \int \frac{{\rm d}x'}{\lambda} \frac{{\rm d}y'}{\lambda} \; g(\vec{x}') e^{-\imath k \vec{n} \vec{x}'} \,. \tag{2.14}$$

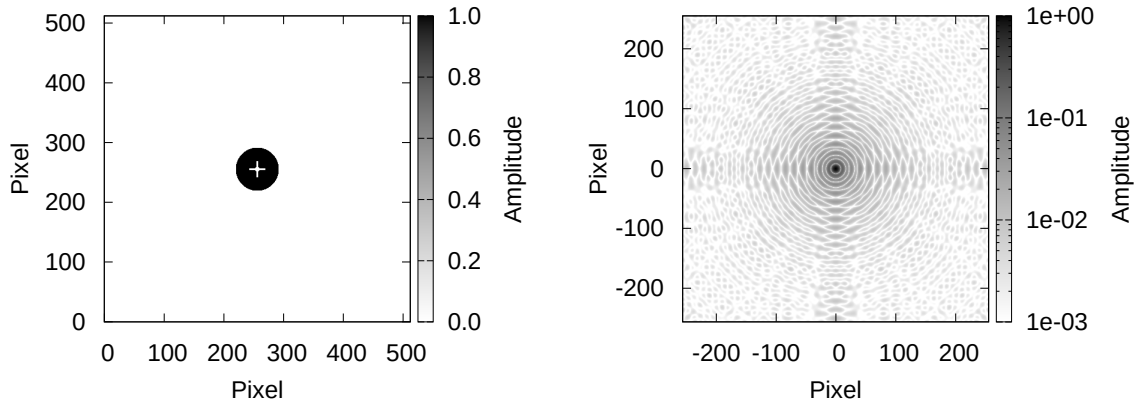


Fig. 2.3: If the aperture function $g(x, y)$ (left panel) is known the antenna pattern $P_{\rm n}(\vec{n})$ (right panel) can be numerically computed via Fourier transform (see Eq. (2.13) and (2.14). For this example, the aperture is a spherical box function (uniform illumination) from which a smaller circle and a cross were subtracted to mimic the blocking of the focus cabin and support legs of a typical radio telescope. Note, that in reality also phase errors due to reflections and diffraction, as well as, misalignments of the focal point from the optical axis (introducing phase errors) play an important role.

Even, if the aperture is more complex (shadowing by the primary focus cabin and support legs) this gives us a handle on how to calculate the beam pattern. Fig. 2.3 shows this schematically. However, in reality the computation is more complex as there are also reflections and diffraction of incoming radiation, as well as, misalignments of the focus horn from the optical axis introducing phase shifts (e.g., so-called coma lobes).

2.2.2 Receiver

The new 21-cm multi-beam receiver was primarily built for beampark experiments to measure space debris in the Earths environment using radar measurements[4]. It works as a (cooled) single conversion heterodyne system having 14 separate channels (seven beams á two polarization channels). The electromagnetic waves are focussed by the antenna and couple via the feed horns into circular waveguides placed in a cryogenic Dewar which is situated in the primary focus of the telescope. An initial amplification of 40 dB is applied to the spectral band between 1200 and 1700 MHz (5 K excess noise contribution due to the HEMT preamplifiers). A lot of effort went into arranging the beams providing best possible beam efficiency while minimizing inter-beam coupling. Before down-conversion to the intermediate frequency (IF) using a local oscillator (LO) further filtering is applied, limiting the bandwidth to the range 1290 MHz to 1430 MHz. The IF band lies between 80 MHz and 220 MHz (the center frequency is 150 MHz). A complete technical report is

[4] The Project "Multi-Beam Receiver for Beam-Park Experiments" of the European Space Operation Centre (ESOC), Contract No. 16173/02/D/HK, pays material cost of the receiver which is developed at the Max-Planck-Institut für Radioastronomie (MPIfR). The TIRA antenna of FGAN in Wachtberg–Werthofen near Bonn is used as transmitter, the 100-m telescope as receiver, working as bistatic radar arrangement. Particles down to 9 mm diameter can be detected at heights between 800 to 1000 km.

Fig. 2.4: The new multi-beam receiver. In the lower part of the photograph the seven feed horns are visible which transform the electromagnetic wave collected by the antenna to a circular waveguide. After conversion to an electric signal (pre-)amplification of 40 dB is applied to the spectral band of interest. The main part of the receiver is placed inside a Dewar which is cryogenically cooled to provide low-noise properties of the electronics.

found in Keller et al. (2006). The receiver utilizes a noise diode at constant temperature to monitor relative changes of the system temperature.

The central feed measures circular polarization (to allow measurements of magnetic fields), while the outer beams provide linear polarization. The beam separation is 0.25°, or 1.6 beam widths, placed on a hexagonal grid. During test observations preparing the survey, the system temperature of the inner beam was measured to be about 60 K, while the outer horns had between 20 and 30 K (elevation dependent). While the latter values are acceptable, the central signals are not — most likely strong RFI causes the problems, coupling in via circular polarization much more effectively. For a complete description of the test observations, see Chapter 5.

For completeness the measured beam pattern for the multi-feed array is shown in Fig. 2.5. It is the convolution of the telescope beam with the response function(s) of the feed horns.

2.2.3 Intermediate frequency chain

The observed frequencies in the range of 400 MHz to 86 GHz are technically challenging in terms of amplification and signal transmission from the receiver in one of the two foci to the backends placed in the observatory. For that reason the heterodyne principle is used to downconvert the radio frequencies (RF) to a much lower intermediate frequency by mixing the RF signal with the output of an ultra-stable LO. After (possibly) preamlifying the signals, the IF signal can be much easier amplified to the level which is needed by the

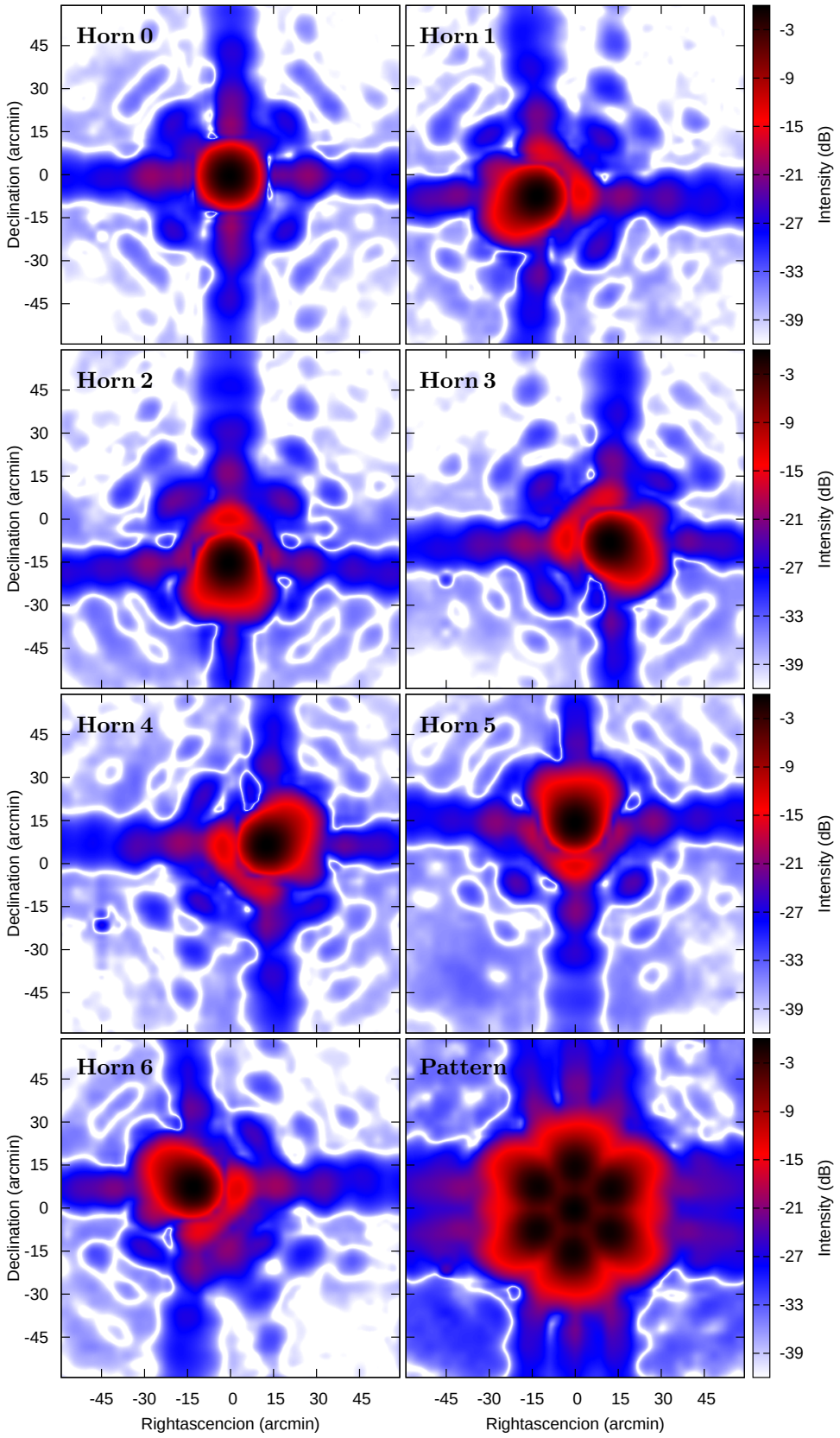


Fig. 2.5: Antenna pattern (bottom right) of the new multi-beam receiver as measured by E. Fürst (MPIfR) using a continuum backend. The other panels show the single feed responses of each of the seven horns. Horn 0 is the central horn. The outer (off-axis) feed horns, 1–6, produce a more complex shape than the inner feed, being not exactly in the focus of the telescope.

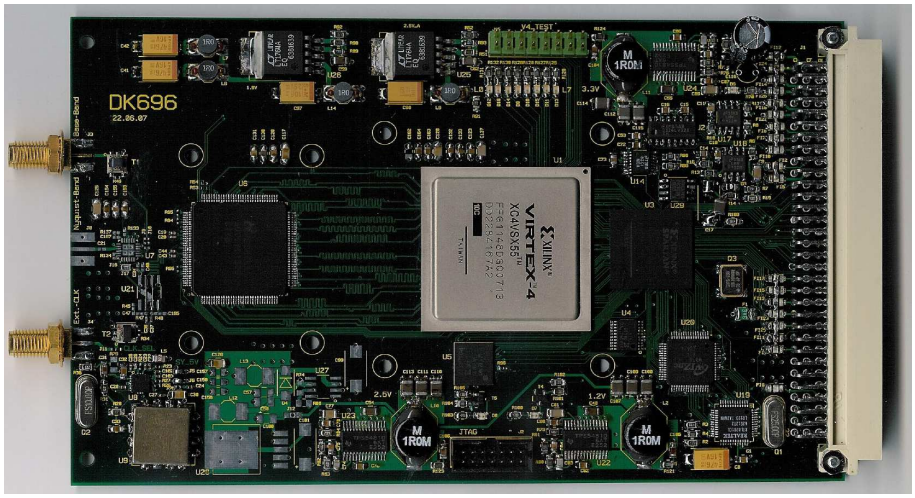


Fig. 2.6: Custom-built spectrometer board based on a Virtex-4 FPGA. (Courtesy of S. Hochgürtel, MPIfR.)

backend. Furthermore, the IF frequency can be transmitted using coaxial cables having acceptable attenuation. Another advantage is, that each receiver can be used with the same IF chain, making many of the amplification and filter elements reusable.

At the moment, two IF bands are used at the Effelsberg telescope, a narrow band in the range of 50 to 250 MHz, and a broad band (500 to 1000 MHz). The latter is currently not used by most of the receivers, but they will be upgraded successively. An optical fiber system is going to be installed which makes it easier to transport the larger bandwidth to the backends.

Two "Universal Local Oscillators" (ULO) are available. To allow frequency switching (compare to Section 3.5) ULO 1 utilizes two synthesizers connected to a frequency standard. It provides frequencies in the range of 1.0 to 2.0 GHz, while ULO 2 provides 0.01 to 1.0 GHz. Both oscillators are tunable in steps of 2 Hz, or 1 Hz, respectively.

2.2.4 Backend

The the amplitude of the electric signal can not be detected directly, but only the power of it. Therefore, after amplification and filtering a detector must be used, measuring the power. There exist total power detectors, continuum backends, spectrometer, and pulsar dedispergers, but here we focus on spectroscopic backends.

For many astronomical observations the spectrum of a source is of interest. It can be used to determine temperature limits, radial velocities, or relative abundances of elements. In radio astronomy there are several types of spectrometers known, filter banks, acousto-optic spectrometers (AOS), autocorrelators (AC), and fast Fourier transform spectrometers (FFTS; e.g., Benz et al. 2005; Stanko et al. 2005; Klein et al. 2005, 2006) based on field programmable gate arrays (FPGA). While, due to lack of computing power, for long time the real-time FFT could not be calculated with desired bandwidths and numbers

of spectral channels, with the advent of modern reconfigurable FPGAs it became possible. The sensitivity of a receiving system is described by the radiometer equation

$$\frac{\Delta T}{T_{\rm sys}} = \frac{k}{\Delta f \cdot \tau} \tag{2.15}$$

with the bandwidth (per channel), Δf, per given integration time τ. The constant k depends on the type of spectrometer. As most autocorrelators use a low quantization (up to at most 3 or 4 bits) of the analog input signal, k is larger than one (for 1 Bit, $k = \pi/2$). FFT spectrometers have $k = 1$, and due to their better input quantization of 8 Bits or more, they are less affected by strong interference signals, which can cause leveling problems on a AC system. Apart from that, FPGA based spectrometers are much cheaper, are reusable, need less space and power, provide superior stability, and allow fast dumping of spectra. With modern ethernet systems they can be easily embedded into a telescope control system. For a more detailed comparison of the various spectrometers, we refer to Winkel (2005).

For the EBHIS a FFT spectrometer developed at the Max-Planck-Institut für Radioastronomie (MPIfR) will be used. It is based on a Xilinx Virtex-4 FPGA, fed by an analog-digital converter (ADC) that has a sampling rate up to 3 GHz at 8 bit resolution; see Fig. 2.6. The original design was introduced for molecule spectroscopy at the APEX telescope (needing large bandwidths in the GHz-regime), but can be easily adapted to match the desired specifications for an H I survey. Since August 2008 seven backends were installed at the telescope and are up and running. They utilize a total bandwidth of 100 MHz and 16k spectral channels, providing a frequency resolution of 6.1 kHz, or a velocity resolution of $1.3\,\mathrm{km\,s^{-1}}$, respectively (Hochgürtel, priv. comm.). Every 500 ms the (integrated) spectrum ("dump") is transmitted via Ethernet to a personal computer, which writes the raw data (in binary file format) to disk.

2.2.5 Control system and software

All observations at the 100-m telescope are controlled per software using the program $\texttt{OBS}_{\rm e}$.[5] It allows to use scripts/macros for better handling of complex tasks. Sources can be saved in catalogs, which then can be accessed by an (unique) identifier. Several monitoring systems provide information about the telescope-state, environmental conditions, and receiver status. For many observations preliminary online-data-reduction is possible.

Each single observation is assigned to a so-called *Scan* number. When mapping an area, the Scan is divided into several *Subscans* based on the different positions in the map. As the maximum counter for the Scan number is 9999, previous observations will be over-written (beginning with 0001) after this limit is reached.

Depending on the type of observations, the (raw) data are recorded in various file formats. The FFT spectrometer mean spectra per (Sub)scan are written primarily to `fits`-files; see the subsequent Section 2.2.6. However, for the EBHIS the best possible temporal resolution of the spectral data (500 ms) is desired, to allow RFI mitigation (see Section 3.2). Hence, the raw data provided by the spectrometer are used as well (and later merged to the fits-files, containing all relevant header information; see the following Section).

[5] A manual can be found under `http://www.mpifr-bonn.mpg.de/div/effelsberg/DOC/OBSE/index.html`.

2.2.6 The `MBfits` file format

In astronomy the `fits` file standard is widely used (Flexible Image Transport System; Wells et al. 1981; Hanisch et al. 2001). Being only a container format it allows to adapt it to almost every need. The advantage is, that programming libraries and tools are publically available, making it easy to read and write data (almost) independently of the type of observation and telescope specialties. In general each fits file consists of so-called header data units (HDU), being either image (cube) or table. Each of those HDUs has a header containing important information (e.g., source, time, type of observation, coordinate specifications, calibration information, or dimensions of the image/table). How the data are logically organized is not part of the standard, making fits very flexible.

Beginning with the tests of the new Multi-beam receiver, the so-called Multi-beam fits (MBfits; Muders 2007) format was migrated to the Effelsberg telescope. This is a standard based on the more general fits specifications which more precisely describes how Single-dish (Multi-beam) observations should be written. It was initially developed for the APEX telescope, and is now also in use at the IRAM 30-m telescope.

2.3 Survey parameters and comparison with other surveys

In this Section a brief comparison of the most important parameters and properties of the different single-dish H I surveys is presented. Table 2.2 contains the spatial and spectral resolution, the survey areas, radial velocity ranges, and integration times, leading to the given noise limits, which are converted to column density limits (galactic surveys) or mass limits (extragalactic), respectively.

Due to the larger Arecibo antenna, much higher integration times are needed in the EBHIS to reach the same sensitivity limits. However, this does not affect the (area) mapping speed, as the Effelsberg beam is accordingly larger. Of course the spatial resolution of ALFALFA and GALFA is higher, but EBHIS can access a three times larger fraction of the sky in return. Together with HIPASS and GASS a complete census of the full sky will be provided, having superior spatial resolution and sensitivity than any full-sky H I survey before.

In the galactic regime the detection limit is $9 \cdot 10^{17}\,\mathrm{cm}^{-2}$ ($4 \cdot 10^{17}\,\mathrm{cm}^{-2}$ towards the SDSS area). The integration time per position was chosen to match a H I mass detection limit of $10^7\,\mathrm{M}_\odot$ at a distance of the Virgo cluster ($\sim 16\,\mathrm{Mpc}$) which should be sufficient to detect a large number of (even small) dwarf galaxies.

Table 2.2: Parameters of several existing or ongoing H I surveys in compared to EBHIS. Towards the SDSS area EBHIS will have a higher sensitivity than for the remaining sky.

Survey	Telescope	Type	Status	Area (sq.deg.)	Beam size (arcmin)	$z_{\max}$	Δv (km/s)
LAB	Dwingeloo/Arg.	g	Completed	41 300	36		1.3
HIPASS	Parkes	e	Completed	29 300	14.1	0.05	18
GASS		g	Ongoing	20 600			0.8
ALFALFA	Arecibo	e	Ongoing	7 100	3.4	0.06	11
GALFA		g	Ongoing				0.7
EBHISe	Effelsberg	e	Planned	20 600	9	0.07	7
EBHISg		g	Planned	(8 500*)			1

Survey	RMS noise (mJy/Beam)	$N_{\rm HI}$ limit (10^{18} cm^{-2})	Mass limit (10^7 M$_\odot$)	Velocities (km/s)	Integr. time per beam (s)
LAB	620	3.3		− 450 . . . 400	
HIPASS	13		8.7	−1 280 . . . 12 700	450
GASS	95	2.6		− 400 . . . 450	90
ALFALFA	1.6		0.8	−2 000 . . . 18 000	28
GALFA	13.2	3.2		− 700 . . . 700	18
EBHISe	4.5 (2*)		1.9 (0.8*)	−2 000 . . . 18 000	120(600*)
EBHISg	12 (5.5*)	0.9 (0.4*)		− 500 . . . 500	

* towards SDSS

CHAPTER 3

Data reduction

Performing blind H I surveys is undoubtly a major challenge with respect to a variety of different aspects. Modern receiving systems provide spectra containing thousands of spectral channels with total bandwidths of several hundreds MHz or even GHz. Digital high-dynamic-range backends on the basis of field programmable gate arrays (FPGA) allow to dump spectra on time-scales of seconds. While all these properties will provide to substantial improved data quality they lead to overwhelmingly increased data volume. Reducing these data is a task which is impossible to accomplish manually. But automatic reduction pipelines need very sophisticated methods in order to fulfil the demand to detect faintest sources within the data.

In terms of radio frequency interference (RFI) the situation at the telescope sites is getting worse, due to the increasing amount of radio transmission all over Earth. Even far from any civilization RFI signals contaminate a substantial fraction of the observed bandwidth especially at low frequencies of several hundred MHz (where SKA and LOFAR will observe highly redshifted H I emission). It is obvious that RFI mitigation/detection mechanisms must be treated at the beginning of any data reduction chain as those interferences would affect each reduction step significantly. There are very different approaches from real-time applications — adaptive filters (Bradley & Barnbaum 1996a), post-correlation with the signal of a reference antenna (Briggs et al. 2000), or higher-order statistical analysis (Fridman 2001) — to various software-based solutions, e.g., simple (manual) flagging of (known) spectral channels, or more sophisticated methods as described in Winkel et al. (2007), which is also described in detail in Section 3.2. On the software side many astronomers tend to use very basic filtering techniques often based on robust statistics (median) which strongly suffer from their nonlinear characteristics and in some cases provide only poor results.

A second major issue is the gain curve calibration. Using large bandwidths often causes strong ripples (e.g., solar ripples) due to strong continuum emitters which produce standing waves between the dish and the receiver. Together with more or less complex IF gain characteristics the proper bandpass calibration is very complex. One solution may be the implementation of the least-squares frequency switching (LSFS) method recently

developed by Heiles (2007), but would need minor changes of the hardware in the IF processing chain. However, Winkel & Kerp (2007) could show the disastrous influence RFI signals have on the reconstruction of the signals of interest. Fortunately, a workaround to the RFI problem was identified, if a 'flag' database is available, containing accurate time and frequency information of RFI events. Indeed, it turns out that flagging data points affected by RFI signals is a cheap but well-working possibility to deal with RFI during all steps of the data reduction pipeline. Both, the LSFS method and the improvements will be described in Section 3.5.

Very important for galactic astronomy is the stray-radiation (SR) correction. Stray-radiation — emission entering via the sidelobes of the antenna — can seriously affect the flux measured from the source of interest. For the EBHIS the method described by Kalberla et al. (2005) will be used; see Section 3.4. If the antenna diagram is well-known (by measurements and modelling), the measured data can be corrected for the influence of the sidelobe pattern to a good precision.

After RFI detection and stray-radiation correction the data must be calibrated in terms of intensity or brightness temperature, respectively. A two-step method will be used, first applying an absolute calibration via measuring a calibration source of well-defined flux density, and then relative changes are determined using a noise diode which output is fed into the receiver at certain intervals (Section 3.3).

The overall data processing is organized as flexible pipeline (see Section 3.1). Because the raw- and processed data products have all the same data format, it is possible to build up a data processing chain consisting of individual data reduction modules. At the end of all data processing steps, via the merger module the gridder (Section 3.6) will produce a final data cube for scientific analysis. Accordingly, it is feasible to produce data cubes i.e. corrected for RFI events but without stray-radiation correction or vice versa. The advantage of this approach is the opportunity to modify individual modules without the need of having to recompute all the corrections. Many modules are optimized to be performed on multi-processor computers.

Finally, a graphical user interface (GUI; see Section 3.7) was developed especially suited for the search and parametrization of sources in the data cubes of the extragalactic H I survey. A very promising finder algorithm was implemented, based on the Gamma test (Boyce 2003). The GUI is designed to allow a fast working flow and is able to compute statistical errors of the fitted parameters using Markov-Chain Monte-Carlo methods.

3.1 The reduction scheme

As mentioned in Section 2.2.6, the spectrometric data will be saved in the `MBfits` file format. The temporal resolution of the raw data, i.e., the integration time per dump, is of the order of one second, meaning a huge amount of data has to be handled. To avoid (unnecessary) redundancy, each module (RFI detection, SR correction, flux and bandpass calibration) works independently on the data and all correction terms needed to describe the outcome are saved to a database, instead of producing complete sets of corrected spectra after each intermediate step. In example, the RFI detection algorithm returns only the list of spectral channels per dump containing an RFI event. Therefore, only this list needs to be stored and can be directly used by the other tasks to flag bad

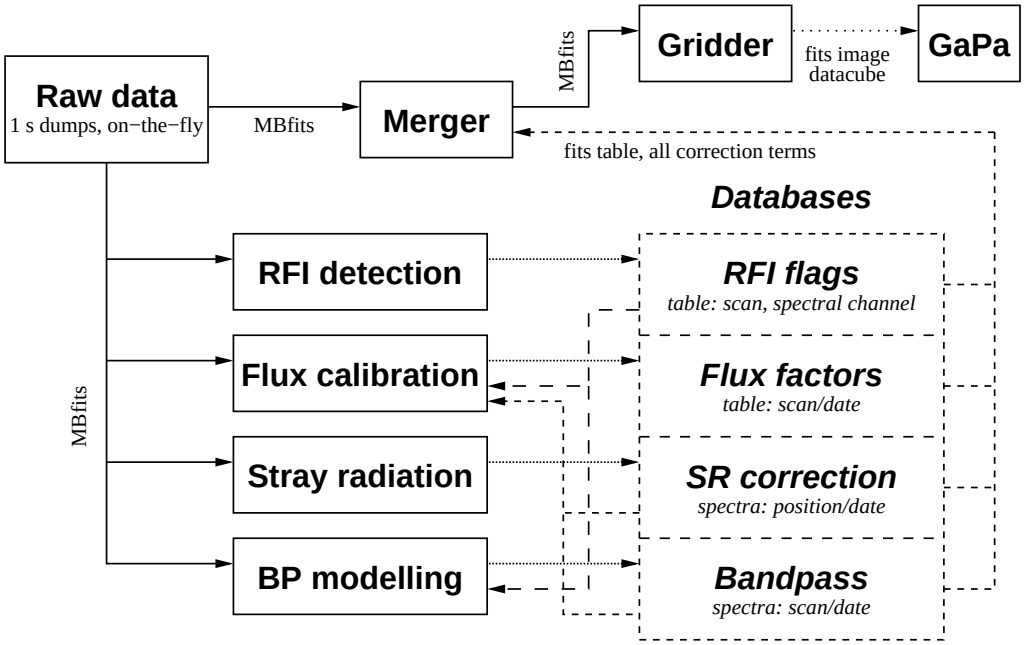


Fig. 3.1: Data reduction scheme used for EBHIS as explained in the text.

data. The same principle holds for the flux calibration, where only one correction factor per spectrum must be present. A schematic view is shown in Fig. 3.1.

In contrast to the usual pipeline approaches, a new task is necessary, which is called "merger", to combine all the correction terms to form the final dataset. It will be possible to apply all different combinations of corrections to the data, without the need arising to re-execute a single calibration task.

After merging the data, to produce a data cube a gridder is used. Extragalactic data can be (automatically) searched for sources, by the "Galaxy Parametrizer". Many of the modules make use of multi-processor or -core platforms boosting the computation speed.

3.2 RFI mitigation

In Radio astronomy the signals of interest are often polluted by man-made artificial radio emission, so-called *radio frequency interference* (RFI), caused by Radio/TV broadcasts, mobile communications, or radar applications, etc. Due to the high sensitivity of modern radio telescopes already weak sources can lead to corrupted data. Furthermore, even in the protected bands one has to deal with "legal" transmitters in adjacent spectral regimes, leaking part of their emission into the band of interest. Especially, the use of digital signals in radio communications produces (sub-)harmonics at various frequencies. The dynamic range of RFI includes events hardly noticeable in the spectral noise up to strong bursts, which can even lead to permanent damage of the receiver.

Two possibilities exist to deal with interferences — keeping artificial radiation away from the telescope (passive mitigation) or trying to detect or even mitigate RFI once it has entered the receiving system (active mitigation). For the latter, several methods have

been developed today. In general, three different approaches can be identified. Firstly, a manual search for RFI signals in already recorded spectra can be carried out (the "classical" method). Secondly, using sophisticated algorithms it is possible to search for RFI signals in recorded data automatically (Bhat et al. 2005). Thirdly, one can use real-time applications which have to be implemented into the signal chain of the telescope. Various approaches are under consideration, e.g. adaptive filters (Bradley & Barnbaum 1996b), post-correlators (Briggs et al. 2000), or real-time higher order statistics (HOS; see Fridman 2001).

In this work, we follow the second approach using spectral data of high temporal resolution of the order of few hundreds of milliseconds up to one second. This accounts for the quickly changing signal signature of most RFI events. It became obvious during test observations (Winkel et al. 2007), that the RFI signal variations at Effelsberg occur on the order of less than 100 ms. The necessary short integration times yield very high data rates. In practice, one must find a compromise between reasonable time-resolution for RFI detection and the amount of data. Moreover, the read-out time of a spectrometer is a technical constraint. Practically, only modern FPGA-based FFT spectrometers provide the dynamic range and temporal resolution needed for off-line RFI detection applications.

3.2.1 Passive mitigation

One of the most useful RFI mitigation "techniques" is to simply avoid interferences as much as possible. This starts with choosing a telescope location far from the civilisatory influences (communications, etc.) but also includes proper shielding to reduce as much as possible leaking emission emerging from all kind of electronic devices in the vicinity of the antenna. Obviously, the former is only practicable for the construction of new observatories, while the 100-m telescope, which will used for EBHIS, can hardly be moved into a radio-quiet zone.

According to national laws and international agreements, the spectral bands in the radio regime are regulated for economic or public users. Each application must follow the legal power limits and must — to a certain degree — ensure, that no emission leaks into adjacent wavelength bands. Unfortunately, as the radio spectrum is finite, there is a lot of competition between the potential users. Commercial interests are directly opposed to the interests of radio astronomy science. The former usually is to transmit as many data as possible at lowest expences, which means digital transmission with relatively high power levels — analog transmission would need more spectral bands, and lower power levels are only feasible if more basis stations (repeaters) are used. On the other hand, digital transmission produces (sub)harmonics leaking into adjacent bands. As a consequence, radio observations today easily detect lots of RFI signals of various terrestrial origins, as well as from satellites.

In Germany the *Regulierungsbehörde für Post und Telekommunikation* defined several protected bands (e.g., 1400–1427 MHz) for (passive) use in radio astronomy; see also Table 3.1 for a list of several bands. Unfortunately, on the one hand, this band is considerably narrow for many modern aspects of astronomy (higher redshifts), while on the other hand, even the protected regimes are affected by interferences leaking into the receivers, making a large portion of the data scientifically unusable.

Table 3.1: Some examples of allocated spectral bands in the cm-regime for commercial and public applications in Germany. (Source: Regulierungsbehörde für Post und Telekommunikation 2004)

Frequency (MHz)	Application
87.5 – 108.0	Radio broadcasting
108.0 – 117.8	Aircraft navigation
144.0 – 146.0	Amateur radio
174.0 – 223.0	TV broadcasting (DVB-T)
890.0 – 915.0	Digital cellular phone network (GSM)
1370.0 – 1400.0	Military applications (e.g., radio comm., geo-exploration)
1400.0 – 1427.0	Reserved for radio astronomy
1427.0 – 1452.0	Cellular phone network (civilian, military)
1452.0 – 1479.5	Digital radio broadcasting (DAB)
1479.5 – 1492.0	Digital satellite radio broadcasting (DAB)
1646.5 – 1656.5	Service links for satellite communication
1660.0 – 1660.5	Reserved for radio astronomy
1820.1 – 1875.5	Digital cellular phone network (GSM)
1880.0 – 1900.0	Wireless telecommunication (DECT)
1900.0 – 1980.0	Digital cellular phone network (UMTS)

It is clear, that transmissions of any kind do not respect national sovereignty. Therefore, international agreements have gained high importance. The International Telecommunication Union (ITU) is the leading United Nations agency for information and communication technologies. One of its tasks is to manage the radio-frequency spectrum and define certain limits to the radiated power, as well as propose standards for radio communications (ITU-R recommendations); see Table 3.2. The recommendation having the largest impact on radio science is ITU-R RA.769, which defines the "Protection criteria used for radio astronomical measurements". Power levels in the protected bands are given for continuum and spectral line measurements. It is very important for radio astronomers to stay involved in the associated regulation boards. Power and cross-band limits are continuously under debate, as they mean direct costs for (active) users, e.g., by applying better filters.

A nice visualization of the matter is given in van Driel (2007), called the complex "problem space" of spectrum management; see Fig. 3.2. The conflicting interests of active and passive users, are mediated by the administration on national, regional, or global level (depending also on the type of application). Accordingly, different organizations/boards are involved. The description of all those committees is outside the scope of this work. Only the Expert Committee on Radio Astronomy Frequencies (CRAF, `www.craf.eu`) of the European Science Foundation (ESF) shall be mentioned, as for the 100-m telescope it has (apart from the German regulations) the largest impact. It represents radio observatories of 19 (mainly European) countries and several multi-national organizations (e.g., the European VLBI network, EVN, the European space agency, ESA, and, the Institut de Radio Astronomie Millimétrique, IRAM).

Table 3.2: List of ITU-R recommendations for radio astronomy. (Source: `http://www.itu.int/rec/R-REC-RA/e`; May, 14th, 2008)

RA.314	Preferred frequency bands for radio astronomical measurements
RA.479	Protection of frequencies for radioastronomical measurements in the shielded zone of the Moon
RA.517	Protection of the radio astronomy service from transmitters operating in adjacent bands
RA.611	Protection of the radio astronomy service from spurious emissions
RA.769	Protection criteria used for radio astronomical measurements
RA.1031	Protection of the radio astronomy service in frequency bands shared with other services
RA.1237	Protection of the radio astronomy service from unwanted emissions resulting from applications of wideband digital modulation
RA.1272	Protection of radio astronomy measurements above 60 GHz from ground based interference
RA.1417	A radio-quiet zone in the vicinity of the L2 Sun-Earth Lagrange point
RA.1513	Levels of data loss to radio astronomy observations and percentage-of-time criteria resulting from degradation by interference for frequency bands allocated to the radio astronomy on a primary basis
RA.1630	Technical and operational characteristics of ground-based astronomy systems for use in sharing studies with active services between 10 THz and 1 000 THz
RA.1631	Reference radio astronomy antenna pattern to be used for compatibility analyses between non-GSO systems and radio astronomy service stations based on the epfd concept
RA.1750	Mutual planning between the Earth exploration-satellite service (active) and the radio astronomy service in the 94 GHz and 130 GHz bands

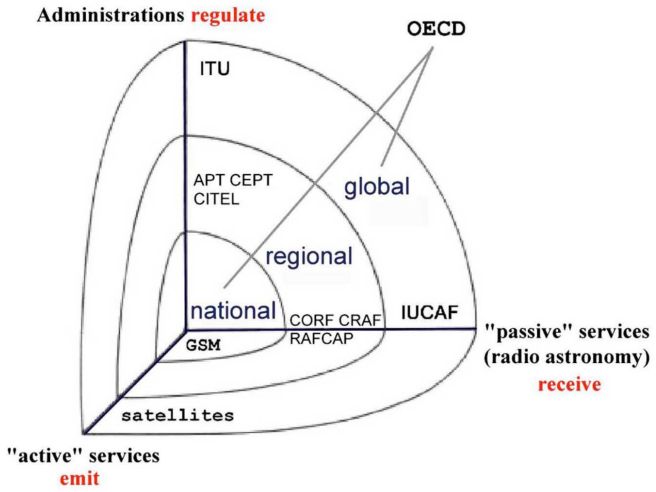


Fig. 3.2: The complex "problem space" of spectrum management, with three orthogonal axes (passive services, active services, Administrations) and three spheres of different radii (national, regional, global). (Source: van Driel 2007)

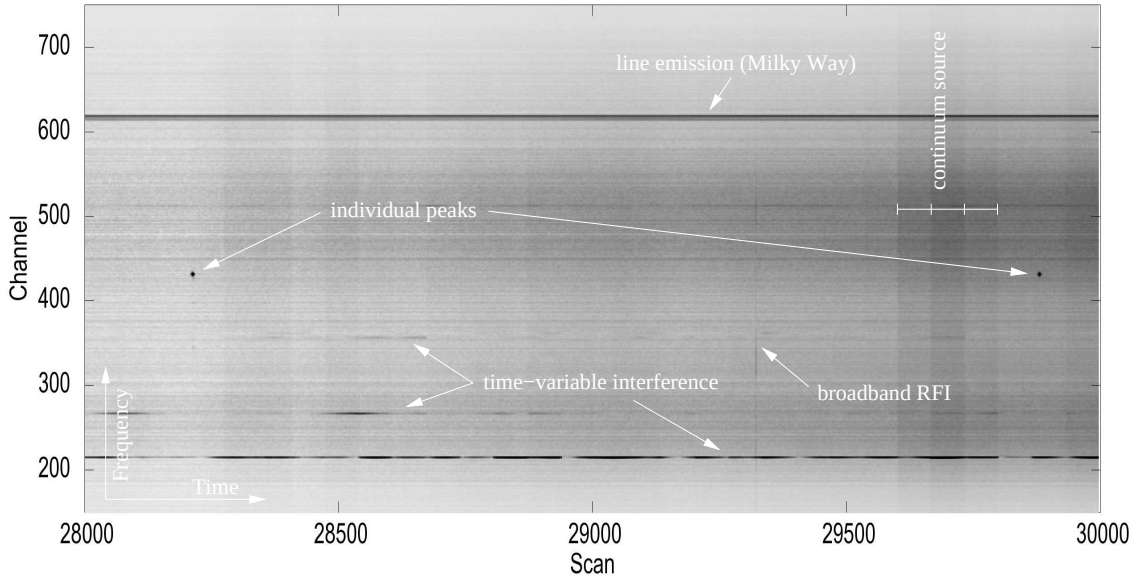


Fig. 3.3: Grey-plot (from Winkel et al. 2007) visualizing one of the datasets recorded with an FPGA-based FFTS (Stanko et al. 2005). The data are composed of several subscans, each of about 30 single spectra containing data from a specific position on the sky. Three different types of RFI signals were detected. Note, that due to its sensitivity the spectrometer reveals different continuum fluxes towards spatially separated positions.

3.2.2 The detection algorithm

In this paragraph the algorithm which will be use to detect RFI signals in a set of spectral data is described. It implements an automated spectral feature detection supplemented with follow-up statistical analyses, performing the separation between RFI signals and astronomical line emission.

The prime signature of RFI signals, in comparison to astronomical line emission, is the temporal variability. To make best use of this behavior Bhat et al. (2005) suggested to use grey-plots which are a representation of the one dimensional astronomical spectra plotted as a function of time t. In Fig. 3.3 a typical grey-plot of an astronomical H I 21-cm line observation at the 100-m telescope is shown. The continuous and constant intensity line traces the H I emission of the Milky Way. Throughout the whole grey-plot statistical distributed RFI peaks, as well as long duration RFI signals can be identified. Based on our experience most types of RFI signals can easily be identified by eye in the time-frequency plane. This is because of the excellent pattern recognition ability of the human brain. However, handling several thousands of spectra is a task which demands pattern-recognition algorithms, which are also of great interest in many computer science problems such as classification of images, face recognition and many more applications. In principle, one could try to use artificial neural networks (Duda et al. 2001) for such an approach, but this would need a huge amount of preprocessing and fine-tuning of the network which is outside the scope of this work.

To optimize the RFI signal detection rate, the underlying envelope (henceforward denoted as baseline) should ideally be constant in both frequency and time domain. This requirement is not fulfilled in real H I 21-cm line data. The shape of the baseline is determined by the overall gain curve of the receiving system. In the best case the shape of the bandpass is independent on the incident radiation power. As the flux of most cosmic sources can be considered as constant on timescales of minutes all changes of the baseline are due to drifts of the system, atmospheric effects, radiation from the ground, etc. For subsets with a duration of the order of one minute these changes can be assumed to be small, such that the functional dependence of the values in one spectral channel can be described by a low order polynomial. Usually, a baseline-fit is performed separately for each (longer) integrated difference spectrum, which refers to a specific position on the sky. At this step, in the reduction pipeline the only interest is in finding RFI signals, therefore (un-calibrated) spectra are used. A baseline-fit demands setting of so-called baseline windows, which define the spectral channels containing emission of cosmic sources. All spectral features of interest have to be covered within such a window, otherwise they would degrade the baseline-fit. Accordingly, RFI signals affect the baseline-fit strongly because in contrast to the emission of cosmic sources RFI signals are randomly distributed over time-frequency plane, so that in general no unique window can be defined. To overcome this limitation a fitting procedure was developed which automatically sets appropriate windows around spectral features, either of astronomical sources or of RFI signals. The data are separated into "tiles" — smaller parts of the frequency-time plane. Each tile consists of data of a certain number of spectral channels (denoted as columns) of one Subscan, or sky position, respectively (denoted as rows). For simplicity and to grant robustness versus fit-related computational uncertainties, it is assumed that each spectral channel in a tile has the same dependency of time. This means the baseline can be described by

$$f(\nu, t) = \sum_{i=0}^{\nu_{\text{order}}} c_i \nu^i + \sum_{i=1}^{t_{\text{order}}} d_i t^i \tag{3.1}$$

which represents a hyper-surface of $(\nu_{\text{order}} + t_{\text{order}} + 1)$ degrees of freedom. This extends the widely used 1-dimensional baseline fit procedure to a 2-dimensional fitting.

The algorithm automatically sets windows with a certain width, e.g., 5 pixels (5 spectral channels times 5 successive scans) around all values above a *trigger* level of $x_{\text{trig}}\sigma_{\text{rms}}$. All data points within the windows are excluded from the baseline-fit. The fitting procedure and calculation of the residual are repeated until the fit has converged in terms of the reduced χ^2 test. In practice it is found that in general less than five iterations are sufficient. If very broad RFI signals are present the number of iterations is slightly higher because in each iteration the window may change its size only by a few pixels. A major issue on the robustness of the procedure is the first iteration. If one tile contains too many features, their initial impact on the fit could cause unexpected results. To overcome this, first a

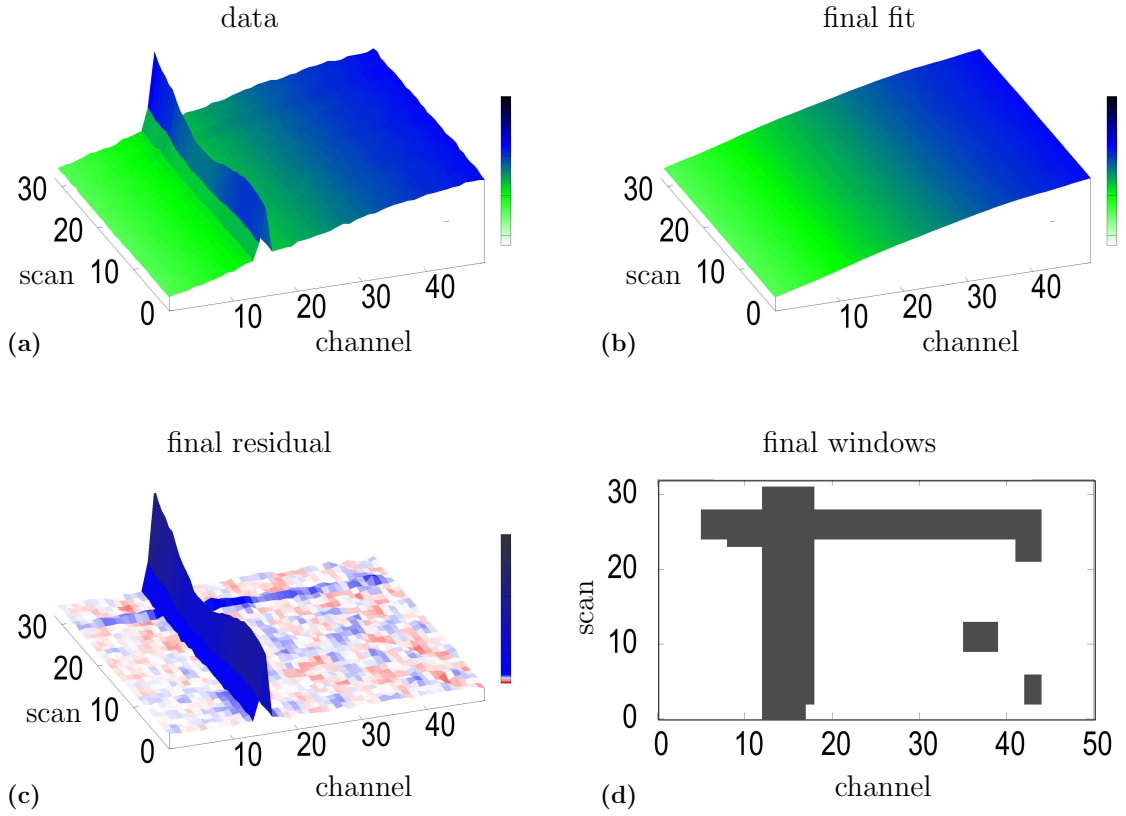


Fig. 3.4: **(a)** H I 21-cm line data of the Effelsberg telescope containing a broadband interference superposing a narrow-width time-variable RFI signal. Shown is a part of a subset (tile) in time-frequency domain. **(b)** Fit of the 2-dimensional baseline after five iteration steps to find the location of line emission and RFI signals in the data tile. **(c)** The difference between the data and the baseline in the time-frequency domain. The line emission and the noise are easy to identify. **(d)** Windows defined by the automatic window-fit algorithm (from Winkel et al. 2007).

guess of window positions and sizes is made by applying horizontal/vertical matched filters $A_{\parallel,\perp}$ (also known as edge-enhancement)

$$A_{\parallel} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 & -1 \end{pmatrix}, \; A_{\perp} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}. \tag{3.2}$$

The noise level is calculated robustly (by cutting off lower and upper 10% percentiles) and is used to search for peaks ($\geq 5\sigma_{\mathrm{rms}}$). This procedure sets initial constraints on the windows.

Figure 3.4 shows exemplarily one tile and the results of the automated feature detection procedure. In the final step all data points above a threshold level $x_{\mathrm{thresh}}\sigma_{\mathrm{rms}}$ which are enclosed within a window are taken into account. Note, that the discrimination of trigger

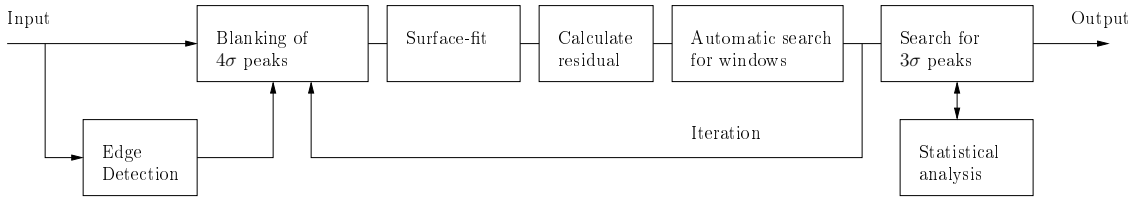


Fig. 3.5: Flow-chart of the detection algorithm. First, an edge-enhancement algorithm (utilizing horizontal/vertical matched filters) is applied to evaluate a guess for appropriate windows in the time-frequency plane. In the successive iteration steps the fitted baseline is subtracted from the data. The result is inspected for signals in excess of a threshold $x_{\mathrm{trig}}\sigma_{\mathrm{rms}}$ which in return define new windows. Ideally, the final residual contains only astronomical line emission and RFI signals. To distinguish between both a follow-up statistical analysis has to be performed; see Fig. 3.6.

and threshold levels is in some cases helpful to avoid an unnecessarily high number of erroneous detections of noise peaks. The trigger level x_{trig} defines the intensity threshold to determine the extent of a window enclosing all pixels exceeding this threshold in the frequency–time domain. The general scheme is depicted in Fig. 3.5.

The remaining task is to distinguish between RFI peaks and astronomical line emission within the windows. This is done by a statistical analysis; see Fig. 3.6 for a flowchart. The intensity of the astronomical line emission should approximately be constant within a certain amount of time and for one sky position. Hence, in the absence of RFI the scatter (standard deviation) in each row of a tile should be similar. The same holds true for the columns in the tile. As nearly all RFI signals are varying or modulated either in time or in frequency, an RFI event will change the statistical properties within the tile. Each data point within a window belongs to a specific row and column of the tile. If the standard deviation of this row/column is significantly (two times) higher than the median of the standard deviations of all rows/columns this is most likely due to RFI.

Such a statistical consideration does not work very well for relatively constant RFI signals, e.g. broadband interferences or impulsive but faint signals, as isolated weak peaks. Broadband interferences seem to be impulsive in time but enhance the continuum level across dozens or even hundreds of adjacent spectral channels. This easily triggers a window but is not sufficiently strong to enhance the standard deviation in a single tile significantly. To overcome this limitation, a workaround mechanism was implemented (referred to as *workaround (a)*), which is optimized to identify RFI signals polluting more than a half of the total spectral channels of a single row.

Isolated RFI peaks in the frequency-time domain are identified via *workaround (b)*: the algorithm searches for features in the tile which show up with an intensity in excess of $(x_{\mathrm{thresh}} + 1.5)\sigma_{\mathrm{rms}}$ (to minimize the detection of noise peaks) which are narrower than the minimum expected line widths of astronomical sources and are short term events. At this stage the properties of the used backend have to be considered. Modern FPGA-based spectrometers will allow a high number of spectral channels and large bandwidth (Benz et al. 2005; Stanko et al. 2005). The combination of both determines the spectral resolution. The narrowest spectral line of interest should be sampled by at least three spectral channels. If a spectral feature shows up being narrower as this lower limit, it is

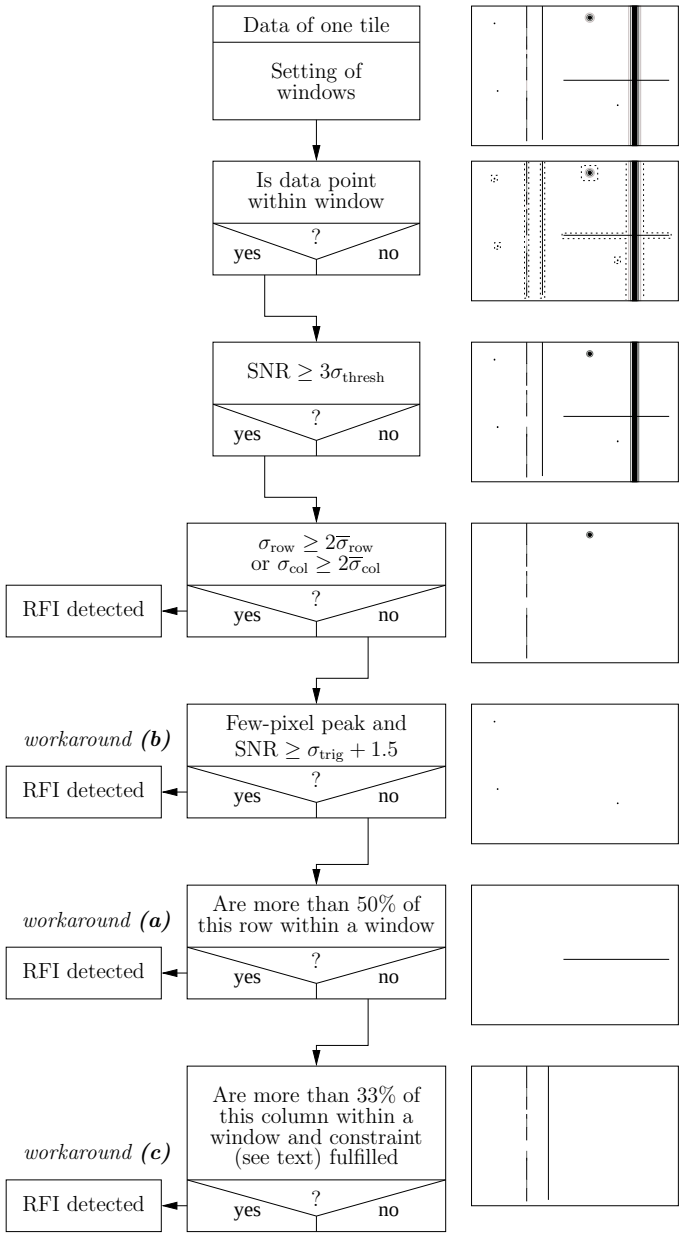


Fig. 3.6: The viewgraph shows the processing chain of the statistical analysis. After feature extraction all signals above a threshold level $x_{\mathrm{thresh}}\sigma_{\mathrm{rms}}$, which are localized in a window, are considered for follow-up statistical analysis. The main classifier uses variability of a signal, but to improve the detection quality additional morphology-based classifiers were implemented.

considered an RFI-event. In the extreme case of 50 kHz frequency resolution, the major fraction of the astronomical lines belonging to the cold neutral medium are not adequately sampled. Accordingly, one has to account for these "unresolved" astronomical lines to differentiate them from RFI-events.

Finally, a third workaround mechanism is implemented which allows to detect narrow-band RFI signals which have nearly constant intensities during the observation. Here, *workaround (c)* searches for channels which are associated with enhanced emission during at least 33% of all scans. Because this also accounts for astronomical emission lines histograms are used counting the pixels within windows and above $x_{\mathrm{thresh}}\sigma_{\mathrm{rms}}$ vs. spectral channels. There-in equivalent widths of those candidates are calculated. If the signal has

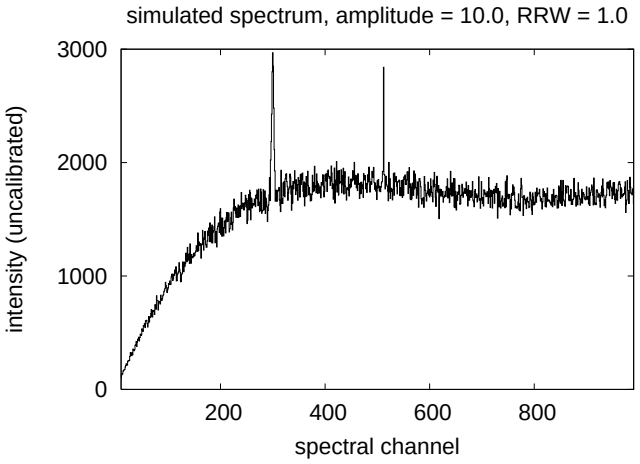


Fig. 3.7: Simulated spectrum containing noise, an "astronomical" line, and an interference peak. The result was multiplied with a polynomial of fifth degree to emulate a bandpass shape.

been identified as being smaller than the expected minimal astronomical line width, this signal is attributed as an interference. Of course, this width parameter should be chosen carefully and makes only sense, if the spectral resolution is high enough such that all astronomical signals are well resolved.

3.2.3 Computer simulations

To determine the detection probabilities of the different RFI signal identification algorithms, simulated H I spectra were used containing statistical noise, an irregularly shaped bandpass, and an "astronomical" line superposed by some randomly generated interferences (see Fig. 3.7). The RFI signals were parametrized by (2-dimensional) sinc^2-functions (cutoff at first root) of adjustable width (interval between first roots, RRW) and scaled intensity (in units of SNR). This allowed to test for a wide range of possible morphologies of the RFI signals. Note, that due to sampling issues the constraints on minimum astronomical line widths were set to 1.5 pixel although an astronomical emission line (of cold gas) at frequency resolution of 50 kHz might also be as narrow as one pixel.

Four basic types of RFI signals were generated: (1) A time-variable but frequency-stable RFI signal (within each individual spectrum). This signal has a certain intensity modified by a sine modulation (with a period of 30 scans, which is slightly less than number of scans per subset) of 20% of the mean interference intensity. (2) A signal with variable intensities drawn randomly from a power-law distribution. In several test observations most of time-variable interferences followed a probability distribution represented by a power-law with an exponent $\nu \sim -1.6$. If having narrow widths, both time-variable types should be detected by workaround (c). (3) Single peaks distributed randomly in the time–frequency plane. These kinds of RFI signals are detected by workaround (b) if they cover only a few pixels in the time-frequency domain. Otherwise their intensity or width is sufficiently high to enhance the variance statistically significant above the statistical noise level. (4) Broad signals which extend over a large number of channels with intensity modulations determined by a sine-wave. This type is likely detected by workaround (a).

Figure 3.8 shows the detection rate of different RFI signal signatures as a function of both their amplitude (SNR) and width (RRW, in units of spectral channels or pixels,

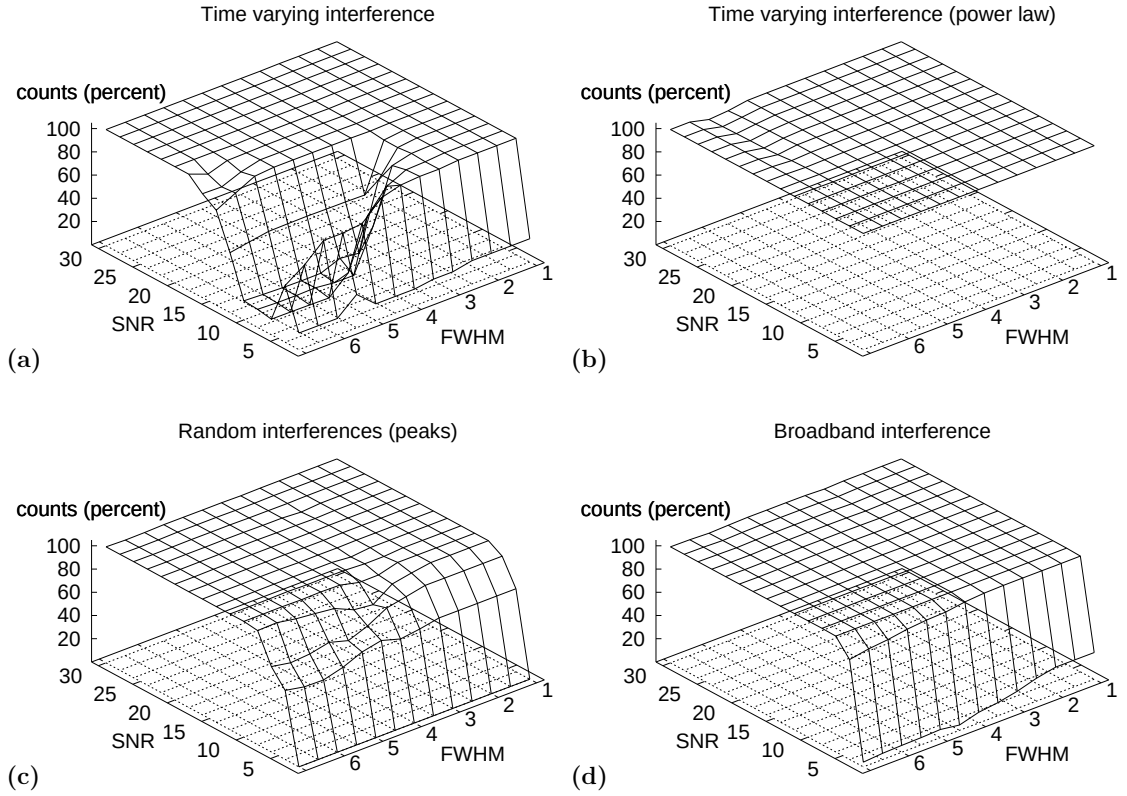


Fig. 3.8: Histograms showing the total detection rate of the proposed algorithm. For each of the four simulated types of RFI — **(a)** time-varying, **(b)** time-varying with underlying power law, **(c)** random peaks, and **(d)** broadband events — the number of correctly found interference-containing data points was calculated as a function of signal-to-noise ratio (SNR) and root-to-root width (RRW, in pixels) of the sinc-shaped specific events (from Winkel et al. 2007).

respectively). For the power law type the value of SNR denotes the base intensity of a constant signal onto which the power law values where added. Detection rate denotes the ratio of pixels suspected by the algorithm to contain RFI signals to the generated number of interference-polluted pixels. The corresponding trigger and threshold level (see Sect. 3.2.2) were set to $x_{\mathrm{trig}} = x_{\mathrm{thresh}} = 3$. In Fig. 3.8(a) one can clearly distinguish between the main (variance-driven) mechanism and workaround (c). Workaround (c) is responsible for the detection of interferences of lower SNR with narrow widths. These are already well-detected at a SNR of 3.5. In general (for larger widths), non-power-law time variable interferences can only be successfully detected if the SNR is sufficiently high, because then the variance is also high enough. In our simulations a SNR of about 10 to $15\,\sigma_{\mathrm{rms}}$ is needed. In Fig. 3.8(b) 100% detection rate is reached because of the high variance of these kinds of signals. Moreover, the width of the RFI signal does not affect the detection rate as in the case of weaker sine-modulated signals. The bulk of the randomly distributed events (Fig. 3.8(c)) are well detected above a SNR-threshold of $6\sigma_{\mathrm{rms}}$. Again, there is no proportionality between the width of the RFI signal and

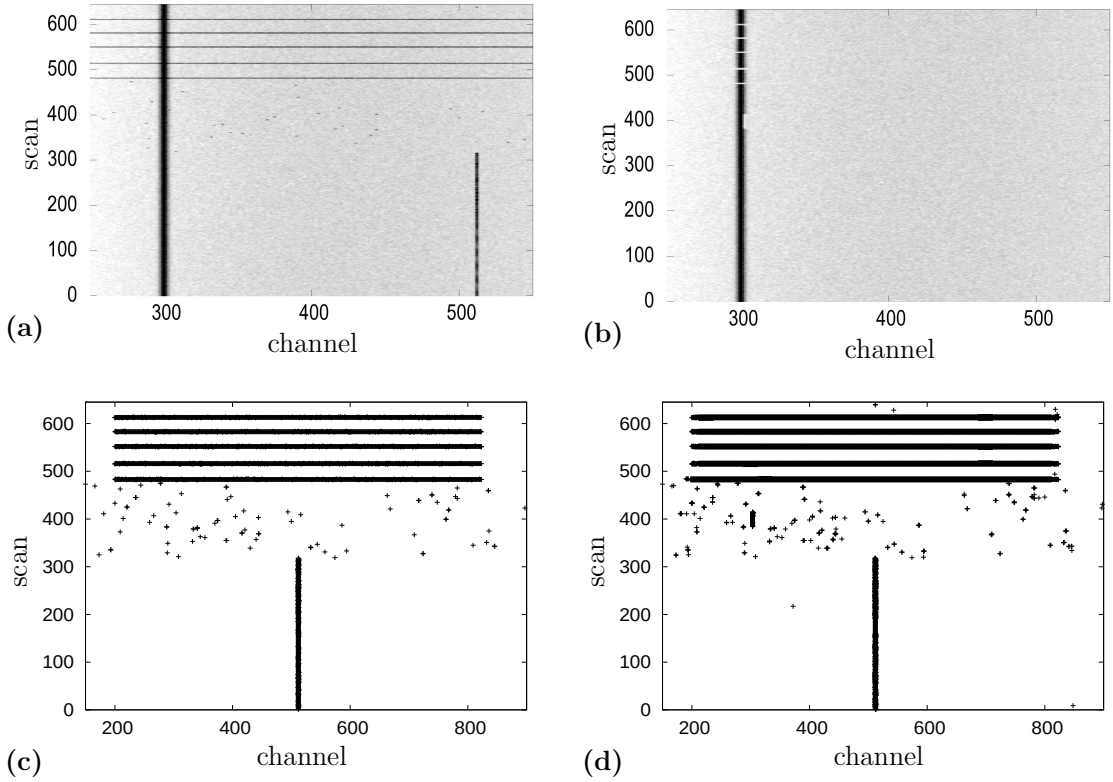


Fig. 3.9: Example, showing the results obtained from application of the proposed algorithm to simulated data (SNR = 15, RRW = 3.0). Four different types of interferences were added to the spectra. In panel **(a)** a grey-plot of the polluted data is shown with time-varying, random and broadband interferences. In the lower panel the detected peaks **(d)** vs. generated peaks **(c)** in time-frequency plane are shown. Panel **(b)** again shows a grey-plot but with cleaned data. Cleaning refers simply to substitute values in pixels, which are affected by an interference, with the value from the baseline fitting function (from Winkel et al. 2007).

its detection probability. Figure 3.8(d) shows the results for the broadband interference signals. Here the 100% rate is reached at $4\sigma_{\mathrm{rms}}$. It is remarkable that even in the case of large widths (RRW) the algorithm is robust. At a width of 7 scans already 25% of all values are affected. This robustness is only possible because of pre-filtering (edge-enhancement) the data as described above.

Figure 3.9 shows the example of a generated and cleaned dataset for an amplitude of 15 and a RRW of 3.0 channels. Cleaning denotes the substitution of the RFI enhanced spectral channels with values consistent with the neighboring baseline function. We did not invest too much effort into a sophisticated cleaning procedure, because our primary intention was to detect interferences for flagging "bad" data. This has important consequences for automatic data reduction pipelines. Flagging of data leads to a decrease of the signal-to-noise ratio (as SNR $\sim \sqrt{\tau}$), but is by far better than the analysis of contaminated data (see also Fig. 3.10). The fast temporal sampling is necessary to identify the

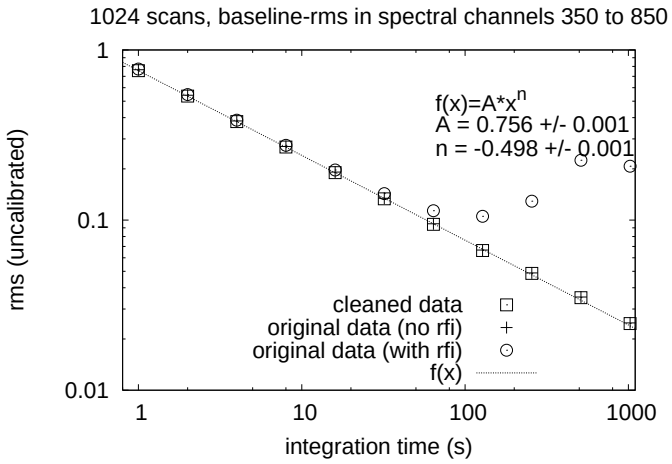


Fig. 3.10: Baseline RMS vs. integration time for SNR = 10 and RRW = 1.0. The theoretic noise level decreases with slope $\nu = -0.5$. The processed data yield this result, although the used cleaning algorithm is most simple. If no RFI-mitigation is performed after $\sim$ 1 min integration time there is no further improvement on the sensitivity limit. Compared to typical integration times, of the order of several minutes, this becomes unacceptable (from Winkel et al. 2007).

RFI affected pixels in the time–frequency domain and to differentiate those events from unresolved astronomical emission lines.

The influence of integration time on the detection process shall be discussed briefly . By calculating mean spectra[1] the noise (RMS) is decreasing by a factor of $\sqrt{n}$ where n is the number of spectra (assuming that all spectra have the same RMS). If a signal has constant strength in all of the spectra its SNR is increasing by $\sqrt{n}$, as well. Thus, it is easier to detect. On the other hand, if a signal is short-lived that is it arises only in a single spectrum, its SNR will decrease by an overall factor of $\sqrt{n}$ when averaging. Most real events are mixtures of these extreme cases. But one can say that short-lived events are harder to detect with increasing integration times while time-varying (long-lived) events are easier to detect.

3.2.4 Impact of RFI signals on sensitivity

A further indicator for interference detection quality is given by the impact of the RFI signals and mitigation method on sensitivity. The noise level decreases with increasing integration time according to the radiometer equation

$$P(t) \sim \frac{1}{\sqrt{t \cdot \Delta f}} \sim t^{-0.5} \tag{3.3}$$

with integration time t and receiver bandwidth Δf.

To evaluate the noise level as a function of integration time spectra (1024) were simulated as before. Figure 3.10 shows RMS vs. integration time of the simulated data with RFI, without RFI, and cleaned. The term integration time has in principle no meaning for artificial spectra. It is assumed, that one scan refers to 1 s as it will be the case in the EBHIS observations. The RMS is calculated by the baseline fit of a 4^{th} order polynomial

[1] The FFT spectrometer internally compute spectra having integration times of the order of μs. The I/O rate, however, is not able to write the data to a disk at this speed. Consequently, the spectrometer internally integrates several thousands of dumps.

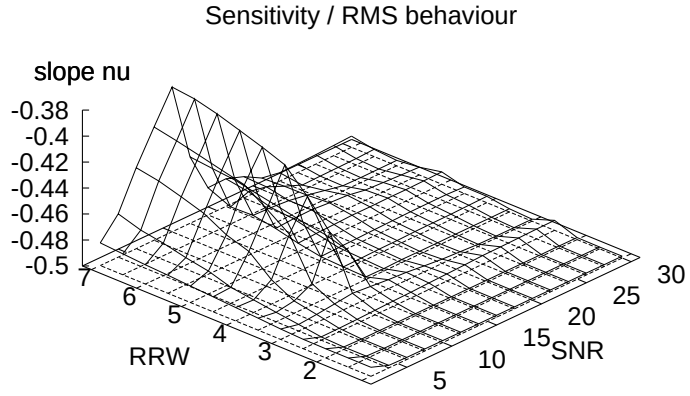


Fig. 3.11: Dependence of the slope ν (radiometer equation) on amplitude and width of simulated interferences. There is a small region in SNR-RRW plane where the theoretical sensitivity could not be reached. This is due to non-detection of broad time-varying (non-power-law) interferences; see Fig. 3.8(a). In all other cases the theoretical value of -0.5 could be reconstructed (from Winkel et al. 2007).

to the spectral channels 350 to 850. For simplicity, this range contains no simulated line emission, but many RFI signals. Next, the mean of all RMS values was computed as an estimator for the noise level in all scans. This is repeated for different integration times by calculation of the mean of each of two successive scans. After 10 steps one single spectrum remains which is the mean of all generated 1024 spectra.

Probably one of the most important results is that without RFI mitigation the highest reachable sensitivity is far below that of clean data and the proposed mitigation method is able to reconstruct the original RMS level extremely well; see Fig. 3.11. Here, only a small region in the SNR-RRW plane is below the theoretical expected value.

3.2.5 RFIDE – a graphical front-end

The detection algorithm described in the previous paragraph was initially implemented using the C programming language. In the mean-time the code was ported to C++, as well, allowing easier integration into a graphical user interface using the Qt-library (`www.trolltech.com`) and enabling to use Qt's threading classes. The latter provides a significant boost of performance when the software is used on multi-core or -processor platforms by distributing computational expensive procedures over several processes.

3.3 Flux calibration

In order to convert the measured spectral information, $S(v)$, recorded in arbitrary units (counts) to a physical quantity as the brightness temperature, T_B, the system temperature T_sys must be derived. Since

$$T_\mathrm{B}(v) = T_\mathrm{sys} \frac{S_\mathrm{sig}(v) - S_\mathrm{ref}(v)}{S_\mathrm{ref}(v)}, \tag{3.4}$$

also a reference spectrum is needed optimally containing no emission features. There exist a couple of different methods to obtain such a reference signal, e.g. by position or frequency switching (compare Section 3.5.4). If no other method is applicable also a polynomial fit

of the baseline might work, but problems can arise if broad emission features are present which are in some cases indistinguishable from the underlying baseline.

One method to derive the absolute system temperature is to use a calibration source, e.g. S 7, having a well-known flux $T_{\mathrm{B}}(v)$. Using Eq. (3.4) T_{sys} can be directly computed. Note, that in practice only the integral value $\int_{v_{\mathrm{min}}}^{v_{\mathrm{max}}} \mathrm{d}\, T_{\mathrm{B}}(v)$ is used for this calculation. Usually such a measurement is done before and/or after the observation. However, T_{sys} varies with time due to changes of the elevation angle (varying fraction of radiation from ground and atmosphere) or intrinsic temperature fluctuations of the receiver. Hence, it is necessary to do a secondary relative flux calibration.

The 7-Beam L band receiver utilizes a stabilized noise diode whose power (flat spectrum) is coupled into the receiver during well-defined intervals (leading to two different phases, denoted as *with cal* and *without cal*). The noise temperature of the diode is known in principle. However, it should be determined by one of the absolute flux measurements using a calibration source to obtain higher accuracy, as this approach leads to the correct antenna temperature, T_{A}, by considering atmospheric attenuation. The system temperature is linked to the total power contained in the spectrum. The mean spectral intensity per frequency channel containing no astronomical emission should be equal to T_{sys} once calibrated.

As RFI signals and emission lines are likely present in the spectra, a more robust method to derive the mean intensity should be used. One possible solution is discussed in Section 5.5 for the data recorded during test observations.

Making use of a calibration source (providing the initial value of T_{sys}) and the estimated mean intensity $\langle S_{\mathrm{diode}}(v)\rangle$ of the noise diode (in arbitrary units, the "difference" between the two spectra with/without noise diode coupled in) a calibration factor T_{cal} (the "effective" temperature of the noise diode) can be calculated independently for the signal and reference phase

$$T_{\mathrm{cal}}^{\mathrm{sig,ref}} = T_{\mathrm{sys}}^{\mathrm{sig,ref}} \left[\frac{\langle S_{\mathrm{diode}}(v)\rangle}{\langle S_{\mathrm{sig,ref}}(v)\rangle} \right] = T_{\mathrm{sys}}^{\mathrm{sig,ref}} \left[\frac{\langle S_{\mathrm{sig,ref}}^{\mathrm{cal}}(v)\rangle - \langle S_{\mathrm{sig,ref}}(v)\rangle}{\langle S_{\mathrm{sig,ref}}(v)\rangle} \right] . \tag{3.5}$$

Assuming T_{cal} to be stable during the complete measurement, Eq. (3.5) can be used to calculate T_{sys} for each single spectrum, regardless on changes in $\langle S_{\mathrm{sig,ref}}(v)\rangle$ and $\langle S_{\mathrm{sig,ref}}^{\mathrm{cal}}(v)\rangle$.

3.4 Stray-radiation correction

In the previous Section the procedure to obtain the antenna temperature T_{A} was described. In order to compute the measured brightness temperature — which is the true brightness temperature convolved with the main beam — the signal must be corrected for the sidelobes of the antenna diagram (stray-radiation correction) and antenna gain[2]

$$T_{\mathrm{B}} = \frac{T_{\mathrm{A}}}{\eta_{\mathrm{eff}}} - T_{\mathrm{SR}} \,. \tag{3.6}$$

Stray-radiation (SR) can provide a significant fraction to the measured spectrum especially in galactic science where Milky Way emission is observable towards all lines of sight.

[2] The 100-m telescope has an efficiency of 70% at 21-cm.

A comprehensive review and application to the Leiden/Dwingeloo survey is given in Hartmann et al. (1996, see also references therein). Here, only some fundamental aspects should be summarized.

First, a few quantities shall be defined. Using the definitions for the beam solid angle, Ω_A, the main beam solid angle, Ω_B in Eq. (2.4) and (2.5), the stray-pattern solid angle Ω_SP can be written as

$$\Omega_\mathrm{SP} = \Omega_\mathrm{A} - \Omega_\mathrm{B}\,. \tag{3.7}$$

Analogous to Eq. (2.6) the stray factor

$$\eta_\mathrm{SP} = \frac{\Omega_\mathrm{SP}}{\Omega_\mathrm{A}} \tag{3.8}$$

is the fraction of radiation power entering through the side lobes.

Slightly changing Eq. (2.12) by introducing a normalized beam pattern, such that

$$\int \mathrm{d}x'^2\; P(\vec{x}-\vec{x}') = 1\,, \tag{3.9}$$

leads to

$$T_\mathrm{A}(\vec{x}) = \int \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}') \tag{3.10}$$

which is written using (2-dim.) Cartesian coordinates for the sake of simplicity. In reality, the equations must be transformed to spherical coordinates. The integral can be separated into a main-beam and a stray-pattern component

$$T_\mathrm{A}(\vec{x}) = \int_\mathrm{MB} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}') + \int_\mathrm{SP} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}')\,. \tag{3.11}$$

Any variations of T_B inside the main beam are not of interest, hence, it can be set to a constant value in the first part

$$\begin{aligned} T_\mathrm{A}(\vec{x}) &= T_\mathrm{B}(\vec{x}) \int_\mathrm{MB} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}') + \int_\mathrm{SP} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}') \\ &= T_\mathrm{B}(\vec{x})\,\eta_\mathrm{B} + \int_\mathrm{SP} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}')\,. \end{aligned} \tag{3.12}$$

Consequently,

$$T_\mathrm{B}(\vec{x}) = \frac{1}{\eta_\mathrm{B}} T_\mathrm{A}(\vec{x}) - \frac{1}{\eta_\mathrm{B}} \int_\mathrm{SP} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}')\,. \tag{3.13}$$

This equation can be iteratively solved by inserting a sky model, T_B, which in each step leads to a better model for the true brightness distribution (convolved with the main beam). However, when it was realized (e.g., van Woerden 1962) that SR correction is essential, there was no good approximation for T_B available. As a consequence the computational effort (for a significant fraction of the sky) would have been too large compared to the computing power at that time. Kalberla (1978) could show, that for $\eta_\mathrm{B} > 0.5$ the solution can be obtained in a single step using the *resolving kernel method*. It is possible to transform the integral

$$\int_\mathrm{SP} \mathrm{d}x'^2\; P(\vec{x}-\vec{x}')T_\mathrm{B}(\vec{x}') = \int_\mathrm{SP} \mathrm{d}x'^2\; Q(\vec{x}-\vec{x}')T_\mathrm{A}(\vec{x}')\,, \tag{3.14}$$

where

$$Q(\vec{x}) = \sum_{i=0}^{N} (-1)^i K_i(\vec{x}) \tag{3.15}$$

is the resolving kernel function and

$$\begin{aligned} K_0(\vec{x}) \equiv K(\vec{x}) &= \begin{cases} 0 & \text{inside main beam} \\ \dfrac{1}{\eta_{\mathrm{B}}} P(\vec{x}) & \text{outside main beam} \end{cases} \\ K_{i+1}(\vec{x}) &= \int \mathrm{d}x'^2 \; K_i(\vec{x}') K(\vec{x} - \vec{x}') \end{aligned} \tag{3.16}$$

can be solved iteratively. Once $Q(\vec{x})$ is known, Eq. (3.13) can be used to directly calculate T_{B}. Nowadays, this is not necessary any longer, as good models of the sky are known from existing surveys.

A few caveats still exist. The stray-pattern is not easy to measure with highest precision. This is on the one hand due to the fact, that no true bright isolated point sources on the sky exist, which can be used for such measurements. On the other hand, the accurate determination of the antenna pattern would need a lot of observing time, competing with other scientific observing programs. Furthermore, the pattern is in principle a function of frequency and time, as different portions of the sky and ground are accessed during observation (elevation angle changes). Due to the movement of the telescope relative to the sky, the stray-patterns rotates with respect to the sky brightness distribution. Also the different LSR velocity corrections during a year need to be considered. Therefore, the measured pattern must be supplemented by models of the antenna response function, providing the far-sidelobe characteristics (spill-over ring, stray cones form radiation scattered of the feed support lags, features caused by blocking, or "shadowing", of the pattern by the support legs, and small components caused by reflections of the roof of the apex cabin; see Kalberla 1978). These features in the far sidelobes must be known down to power levels of $-60\,\mathrm{dB}$ (Hartmann et al. 1996), which simply can not be measured in reasonable amounts of time for various elevations and observing epochs/hour angles.

3.5 Bandpass calibration

Radio receivers following the heterodyne principle (using analog devices) give rise to problems as unknown gain functions ("the bandpass") and instabilities, both in frequency and time. Up to now mainly two different schemes are used to overcome these problems: position switching and frequency switching. Both methods reduce the receiver gain instabilities by measuring a reference spectrum, either off-source (position switching) or with a detuned local oscillator (LO) frequency to move the line of interest outside the observed spectral band (frequency switching). There are a lot of drawbacks applying both methods. Using position switching it is necessary to avoid different continuum levels (or very broad line emission) towards the ON and OFF position which of course limits the usage of position switching in Milky Way observations, having emission from all directions. This is especially valid for the important spectral lines of neutral atomic Hydrogen or CO, due to their high area filling factor. Furthermore, there is loss of observing time either when redirecting the telescope between on and off positions or during the retuning of the

LO frequency, respectively. Frequency switching suffers also from the gain curve changes while shifting the spectral range. The most problematic aspect of both methods is the loss of half of the observing time. In-band frequency switching (both LO phases provide the line of interest lying in the observed bandwidth) avoids this loss. Unfortunately, one loses half of the available bandwidth (which is equivalent to the loss of velocity coverage or number of spectral channels), because a proper separation of both signals is needed. When observing single objects this is usually not a major problem, but when performing, for example, a blind survey one would strongly suffer from such a restriction.

Heiles (2007) presented a new method called *Least-squares frequency switching* (LSFS). LSFS is able to deal with all problems discussed above, making it the best choice for future spectral line observations with radio telescopes especially well suited for H I. However, it requires minor hardware changes at the telescope in order to provide not only two, but a set of three or more LO frequencies within one switching cycle. This is not a substantial problem, and there already is a working system using LSFS at the Arecibo telescope (Heiles 2007; Stanimirović et al. 2006).

Of major interest in radio spectral observations — especially in high-redshift H I astronomy — is the ability of a "bandpass removal" tool to provide high-quality results in a statistical sense. Most observations today need to integrate at least a couple of minutes to reach the desired sensitivity limit. While in theory the noise level scales as $1/\sqrt{t\Delta\nu}$ according to the radiometer equation, with $\Delta\nu$ being the bandwidth and t the integration time, this is not necessarily true for a real receiving system. Modern backends have Allen times, t_0, (for $t \leq t_0$ the radiometer equation holds) of hundreds of seconds (Stanko et al. 2005). Winkel et al. (2007) show that in presence of radio frequency interferences (RFI) this can be limited to less than few tens of seconds. Here, we analyze, if the LSFS method has an impact on the RMS level (the sensitivity) achievable. It turns out, that in statistical sense the LSFS performs very well, yielding only a slightly decrease in sensitivity. Apart from that, the robustness of the LSFS versus several typical problems at radio telescope sites is tested. These are for example RFI events, possibly bandpass instabilities in time and frequency, as well as (strong) continuum sources. LSFS works well under most tested circumstances except for RFI and strong emission lines. However, small changes to the original LSFS method already provide meaningful workaround mechanisms. These are discussed in subsequent paragraphs.

Most spectroscopic observations in radio astronomy use the heterodyne principle where the radio frequency (RF) signal is multiplied with a monochromatic signal of a LO. An appropriate low-pass filter applied after this operation provides the desired IF signal at much lower carrier frequencies. The whole system can be described by

$$P_{\mathrm{IF}}(f_{\mathrm{IF}}) = G_{\mathrm{IF}}(f_{\mathrm{IF}})G_{\mathrm{RF}}(f_{\mathrm{RF}}) \times [T_A(f_{\mathrm{RF}}) + T_A + T_R(f_{\mathrm{RF}}) + T_R] \tag{3.17}$$

with P_{IF} being the power in the IF chain. G_{IF} and G_{RF} are the gain functions at IF and RF stage, respectively. The gain acts on the signals which enter the feed — the astronomical signal of interest plus the contribution from the sky which we denote as T_A — as well as on the noise of the receiver, T_R. Heiles (2007) separates T_A and T_R into a frequency dependent and independent (continuum) part. Note, that the gain is not simply a scalar but has a spectrum due to the filter curves.

To recover from the measured signal, P_{IF}, the signal of interest, T_A, one needs to know the gain spectrum G_{IF} (G_{RF} can be treated as constant with frequency on modern

receivers). Traditionally, this is achieved by measuring a reference spectrum without any spectral features either by position or frequency switching.

3.5.1 Position switching

In position switching the reference spectrum is obtained by pointing the telescope to an OFF position which does not contain any emission from the source of interest (the ON). By calculating (ON-OFF)/OFF one can get rid of the gain spectrum. From Heiles (2007) the equation

$$\frac{P^{\mathrm{on}}(f_{\mathrm{IF}}) - P^{\mathrm{off}}(f_{\mathrm{IF}})}{P^{\mathrm{off}}(f_{\mathrm{IF}})} = \left[T_A^{\mathrm{on}}(f_{\mathrm{RF}}) + (T_A^{\mathrm{on}} - T_A^{\mathrm{off}})\right] \left[\frac{1 - \frac{T_R(f_{\mathrm{RF}})}{T_A^{\mathrm{off}}+T_R}}{T_A^{\mathrm{off}} + T_R}\right] \tag{3.18}$$

is adopted with explicit separation of frequency-dependent and independent terms and the assumption $T_R = T_R^{\mathrm{on}} = T_R^{\mathrm{off}}$. The correction factor on the right is only mildly frequency-dependent. However, if the difference $T_A^{\mathrm{on}} - T_A^{\mathrm{off}}$ of the continuum temperatures is large this can produce a noticeable effect. Position switching can therefore fail for weak lines with strong continuum sources.

3.5.2 Frequency switching

In the case of spatially extended sources position switching becomes unwieldy. For these observations one shifts the LO frequency to obtain a reference spectrum. This yields

$$\begin{aligned}\frac{P^{\mathrm{on}}(f_{\mathrm{IF}}) - P^{\mathrm{off}}(f_{\mathrm{IF}})}{P^{\mathrm{off}}(f_{\mathrm{IF}})} = &\Big[T_A^{\mathrm{on}}(f_{\mathrm{RF}}) + \Big(T_R^{\mathrm{on}}(f_{\mathrm{RF}}) - T_R^{\mathrm{off}}(f_{\mathrm{RF}})\Big) \\ &+ \frac{\Delta G}{G}\left(T_A^{\mathrm{on}} + T_R^{\mathrm{on}} + T_A^{\mathrm{on}}(f_{\mathrm{RF}})\right)\Big] \times \left[\frac{1 - \frac{T_R^{\mathrm{on}}(f_{\mathrm{RF}})}{T_A^{\mathrm{on}}+T_R^{\mathrm{on}}}}{T_A^{\mathrm{on}} + T_R^{\mathrm{on}}}\right]\end{aligned} \tag{3.19}$$

where $\Delta G/G$ is the relative difference of the gain curves of both LO phases which is hopefully $\lll 1$. It is in general not sufficient that $\Delta G/G \ll 1$, because T_R is relatively large. If this condition is not fulfilled, again, the right-hand factor can become a severe limitation for the usefulness of this method. $T_R^{\mathrm{on}}(f_{\mathrm{RF}}) - T_R^{\mathrm{off}}(f_{\mathrm{RF}})$ is usually negligible. Fortunately, in most cases $\Delta G/G$ varies smoothly and slowly with f_{RF} leaving behind a baseline which can be described by a low-order polynomial. A more complete review of position and frequency switching is given by Heiles (2007).

3.5.3 Least-squares frequency switching

Eq. (3.17) reads in a simplified form as

$$P_{\mathrm{IF}}(f_{\mathrm{IF}}) = G_{\mathrm{IF}}(f_{\mathrm{IF}}) S_{\mathrm{RF}}(f_{\mathrm{RF}}) \tag{3.20}$$

where all signals entering the mixer were combined to $S_{\mathrm{RF}}(f_{\mathrm{RF}})$. In contrast to Heiles (2007) the assumption that the input signals are a superposition of frequency dependent and -independent parts is dropped. Using modern broadband spectrometer backends, the

continuum signal likely cannot be treated as constant over the entire observed bandwidth. Moreover, the continuum emission itself might be of interest. Broadband spectrometers could also be used to generate continuum maps. Furthermore, from the mathematical point of view, it is not necessary to perform the separation — all subsequent equations do not rely on it.

As before, one is interested in obtaining $G_{\mathrm{IF}}(f_{\mathrm{IF}})$. Introducing not only two, but a set of N different LO frequencies (for a detailed analysis of how to choose appropriate LO frequencies, see Heiles 2007), results in $N \cdot I$ equations

$$P_{i,\Delta i_n} = G_i S_{i+\Delta i_n} \,. \tag{3.21}$$

In this representation integer indices are used representing the spectral channels of the backend. Hence, i is the i^{th} out of I channels, Δi_n is the frequency shift of LO n versus LO 0 (the unshifted LO) given in channels. By using different LO frequencies of course somewhat different spectral portions of the input spectrum S are observed. Without loss of generality the input signal can be normalized to have a mean value of unity, $S_{i+\Delta i_n} = 1 + s_{i+\Delta i_n}$, leading to

$$P_{i,\Delta i_n} = G_i + G_i s_{i+\Delta i_n} \tag{3.22}$$

This equation can be solved using nonlinear least-squares techniques. However, Heiles (2007) converted the equation to an iterative linear least-squares problem by solving for the difference of guessed values of G_i^{g} and $s_{i+\Delta i_n}^{\mathrm{g}}$ from their true values. From these guessed values one can of course compute the associated power output $P_{i,\Delta i_n}^{\mathrm{g}}$ for each spectral channel and LO setting. After some simplifications (dropping higher order terms; see Heiles 2007) Eq. (3.22) transforms into

$$\frac{\delta P_{i,\Delta i_n}}{G_i^{\mathrm{g}}} = \frac{\delta G_i}{G_i^{\mathrm{g}}} + \delta s_{i+\Delta i_n} \,. \tag{3.23}$$

The δ-terms denote the difference between the true and the guessed value of the corresponding quantity. A further constraint is needed in order to keep the mean RF power approximately constant, namely $\sum_{i,n} \delta s_{i+\Delta i_n} = 0$.

For convenience we use matrix notation for Eq. (3.23)

$$\boldsymbol{p} = \mathbf{X}\boldsymbol{a} \tag{3.24}$$

$$\boldsymbol{p}^{\mathrm{T}} \equiv \left(\boldsymbol{p}_{i,0}^{\mathrm{T}}, \ldots, \boldsymbol{p}_{i,N-1}^{\mathrm{T}}\right) \tag{3.25}$$

$$\boldsymbol{a}^{\mathrm{T}} \equiv \left(g_0, \ldots g_{I-1}, \delta s_0, \ldots, \delta s_{I-1+\Delta i_{N-1}}\right) \tag{3.26}$$

$$p_{i,n} \equiv \frac{\delta P_{i,\Delta i_n}}{G_i^{\mathrm{g}}}, \qquad g_i \equiv \frac{\delta G_i}{G_i^{\mathrm{g}}} \,. \tag{3.27}$$

Least-squares fitting is achieved by computing

$$\boldsymbol{a} = \left(\boldsymbol{\alpha}\, \mathbf{X}^{\mathrm{T}}\right) \boldsymbol{p}, \qquad \boldsymbol{\alpha} \equiv \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \tag{3.28}$$

with the covariance matrix $\boldsymbol{\alpha}$. Computing $\boldsymbol{\alpha}$ requires matrix inversion which in general does not exist necessarily. To deal with degeneracies, Heiles (2007) proposes the Singular-Value Decomposition (SVD) of matrix $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}[\mathbf{W}]\mathbf{V}^{\mathrm{T}} \tag{3.29}$$

where the diagonal matrix $\mathbf{W}$ contains the so-called singular values w_i. In case of degeneracies one or more of the w_i are close to zero, leading to infinite (or huge) numbers when inverting. It turns out that

$$\left(\boldsymbol{\alpha}\mathbf{X}^{\mathrm{T}}\right) = \mathbf{V}\left[\frac{1}{\mathbf{W}}\right]\mathbf{U}^{\mathrm{T}}\,. \tag{3.30}$$

The critical singular values can be treated separately (e.g. setting the inverse values to zero). By computing the SVD of the matrix $\mathbf{X}$ Eq. (3.25) can be solved directly without encountering any problems caused by degeneracies. The computation of the SVD of a matrix, e.g. for $N = 8,\ I = 1024$, is possible on a modern PC but is not finished within fractions of a second; see Section 3.5.4 for details. Nevertheless, the SVD calculation fortunately needs to be done only once per LO setup, as the matrix itself is independent from the measurements.

The assumed normalization of the signal seems to be somewhat arbitrary. But, be the bandpass (gain) curve normalized, in practice associated gain factors can be easily attributed to the signal (which nevertheless has to be calibrated in terms of intensity). This way, the overall power of the input signal (in arbitrary units) can be uniquely reconstructed by computing the mean, m, of the measured signal, normalizing (dividing by m), calculating the LSFS which gives a signal of mean value of unity and finally multiplying the reconstructed signal with m. It is clear that this scheme will only work if the gain curve remains constant. This can be expected at least for the duration of the observing session so that the computed gain factor remains constant for a single observation.

The LSFS method allows reconstructing the continuum part of the input signal, as well. While for position and frequency switching the separation of RF-dependent and -independent parts was necessary, the LSFS algorithm enables the reconstruction of the complete mixture of signals which are fed into the mixer. Of course, this implies that the need might arise for further disentangling these signals into emission line and continuum components (from astronomical sources, ground, and receiver noise) which may even have different spectral indices.

Note, that the equations hold only for small values of $s_{i+\Delta i_n}$ as they were computed only to first order approximation. In most cases this is easily fulfilled in radio astronomy, because the observed lines are much weaker than the typical intensity of the unavoidable continuum level produced by the atmosphere, ground, and receiver noise, which sum is the system temperature T_{sys}. However, there is one case known where one gets indeed a signal much brighter than the continuum level: the H I emission of the Milky Way which can reach intensities of a few 100 K while the system temperature for a typical telescope is $20 - 40$ K. This issue will be addressed in the following analysis.

3.5.4 The robustness of LSFS

Setup

The LSFS algorithm was implemented within the programming language C. Using this code various tests could be performed to investigate statistical stability, response to possible variations of bandpass shapes, impact of RFI signals, and performance in the presence of continuum sources and strong emission lines (MW).

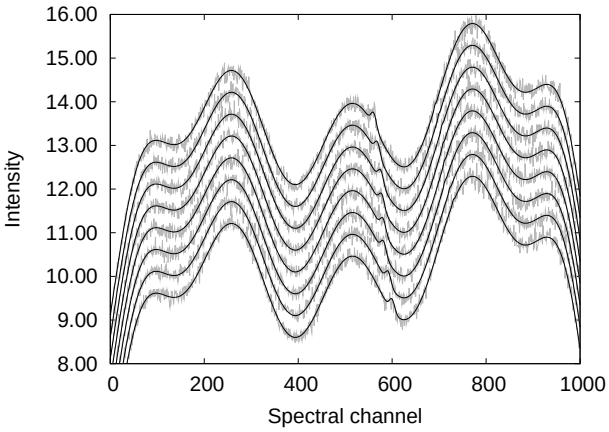


Fig. 3.12: Raw input spectra as would be measured by the receiving system using the MR8 scheme. For better visualization the spectra (grey solid lines) were stacked and a noise-free analogon (black solid lines) was overplotted. Each spectrum is the multiplication of the "true" input signal and the IF gain function. Due to the different LO frequencies within a LO cycle the signals of interest are folded to different spectral channels.

For testing, spectra (1024 spectral channels) were simulated containing several Gaussian-shaped (faint) "emission lines" of different intensities and widths on top of a constant signal (which shall resemble those continuum signals with spectral index of zero). After adding Gaussian noise, these emission lines are partly well below the noise level; compare for example the signal spectrum of Fig. 3.13 and Fig. 3.14. This "true" input signal is then multiplied with a gain function

$$G_{\mathrm{IF}}(f_{\mathrm{IF}}) = G_{\mathrm{IF}}^{\mathrm{filter}} \cdot G_{\mathrm{IF}}^{\mathrm{wave}} \cdot G_{\mathrm{IF}}^{\mathrm{poly}} \tag{3.31}$$

$$G_{\mathrm{IF}}^{\mathrm{filter}} = 0.5\left[\tanh\left(5f+5\right) - \tanh\left(5f-5\right)\right] \tag{3.32}$$

$$G_{\mathrm{IF}}^{\mathrm{wave}} = 1 + 0.1\cos\left(F\pi f\right) \tag{3.33}$$

$$G_{\mathrm{IF}}^{\mathrm{poly}} = 1 + Af + 0.5f^2 \tag{3.34}$$

which was adopted from Heiles (2007) to allow for a better comparison of our results to that work. The quantity f is the frequency which was transformed to integers i, the spectral channels, using $f = 2.1(i - 512)/1024$. In contrast to Heiles (2007) a spectral portion was chosen where the gain curve does not gets too close to zero. This would break the assumption that G_i is of order unity and distorts the normalization scheme which was presented previously. Heiles (2007) did not encounter that problem because continuum emission was neglected. In practice this is no drawback, as one can easily choose those portions of the spectra which fulfill $G_i \simeq 1$. The two parameters $A = 0.1$ and $F = 4$ were varied to change the bandpass in amplitude and shape for some of the tests. A small variation of F already has a dramatic impact on overall shape of the gain curve.

Statistical stability

First, the statistical stability of the method was examined. For this purpose spectra were generated for a set of 8 LO frequencies using the MR8 scheme (Heiles 2007); see Fig. 3.12. As shown in Fig. 3.13, the solution for a single set of spectra does not necessarily provide the "true" signal and bandpass shape, but there are small systematic effects. These can partly be attributed to the influence of single (strong) noise peaks to the overall solution. One should also keep in mind, that the set of equations which are used to solve the decoupling of signal and gain is in linear-order approximation. It is important to note,

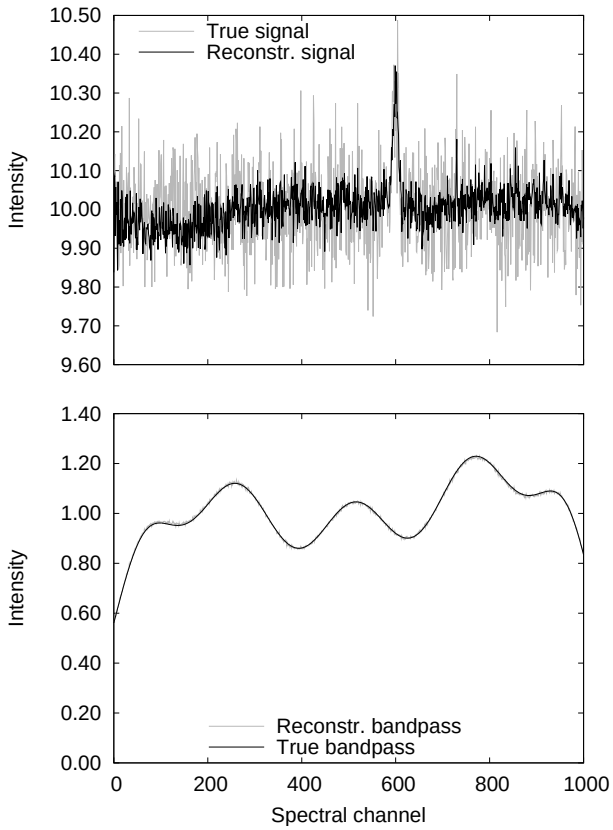


Fig. 3.13: Reconstructed and original signal (top) and bandpass shape (bottom) of a single spectrum from our simulations. The signal is a superposition of noise and three unrelated line signals — two of them are well below the noise level but are significantly detected after integration of several spectra; see Fig. 3.14 (top). Note, that the noise level of the reconstructed signal is about a factor $\sqrt{8}$ smaller because 8 spectra (the different switching phases) result in a single reconstructed signal spectrum.

that the iteration was not interrupted until the solution had converged. The question is whether these systematics cancel out after integration of several spectra or whether they remain. In order to have a measure of the "goodness" of the solution two quality indicators are adopted from Heiles (2007) — the *RMS level of the reconstructed signal* (denoted as RMS) and the *RMS of the residual gain curve* (denoted as σ). The latter quantity uses the residual which is the difference between the true gain curve (noise-free) and the reconstructed gain. The other quality indicator, the RMS, is calculated making use of all spectral channels except those containing the signals of interest. For the purpose of comparison with the noise level of the signals, σ was rescaled with the same gain factor, formerly used to normalize the signals. It is clear, that the reconstructed signals ideally should have a factor of $\sqrt{8}$ lower noise compared to the originally generated signals, because each reconstructed signal was calculated using eight "observed spectra" (one LO cycle).

Furthermore, the behavior of the quality indicators as a function of integration 'time' is analyzed. This is done by successively summing up adjacent spectra. In each step this reduces the number of spectra by a factor two. For convenience, we start with a total number of 1024 generated true spectra (or $8 \cdot 1024$ measured spectra, respectively) which is a power of two. The LSFS was performed on each of the 1024 spectral sets, then the summation of the reconstructed signal and gain curve was performed stepwise.

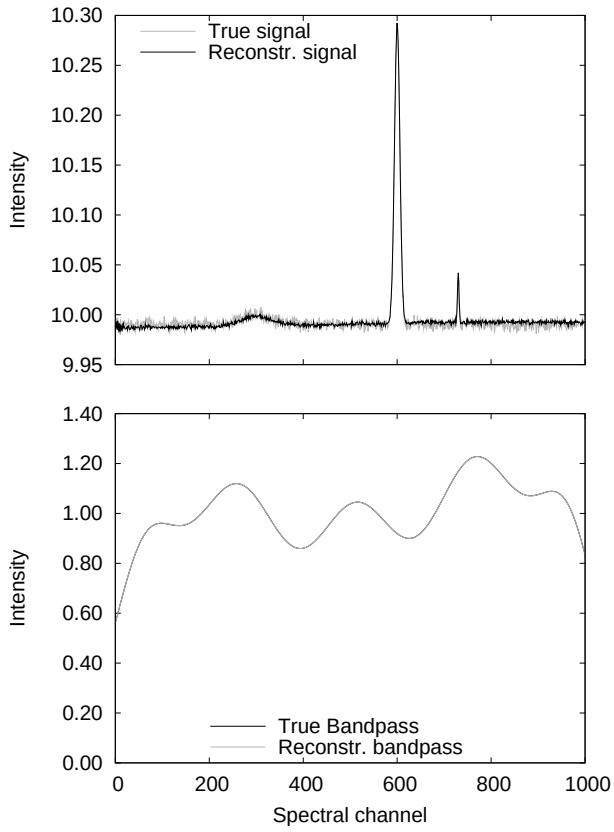


Fig. 3.14: After integration of 1024 spectra the signals previously hidden in the noise show up. Both signal (top) and bandpass shape (bottom) are well recovered, as the functional behaviors of the quality indicators in Fig. 3.10 reveal.

Theoretically, the functional dependence of RMS vs. integration time is given by the radiometer equation

$$P(t) \sim \frac{1}{\sqrt{\Delta\nu \cdot t}} \sim t^{-0.5} \,. \tag{3.35}$$

Fig. 3.14 clearly shows, that despite the fact that individual spectra were not perfectly handled the integrated signal, as well as the bandpass visually match well. However, Fig. 3.15 reveals increased RMS values of about 30% for the reconstructed signal and about 35% higher noise for σ. Calculating the RMS and σ values with respect to a 3rd-order polynomial (fitted after integration) results in significantly lower noise values which are only slightly increased compared to the theoretical expectation value by 8% (RMS) and 10% (σ), respectively. Obviously the residual systematics can be described by a low-order polynomial. We point out, that the RMS behavior of the signal is of much greater interest from the observers point of view. If the gain curve is sufficiently stable with time, one can also compute the systems gain dependence to high precision by using thousands of spectra. The results show that one is effectively not losing sensitivity using this method as this was the case in earlier attempts (e.g. Liszt 1997) with increased noise levels of about 100%.

LSFS and strong line emission

It was already mentioned that strong emission lines (as would be the case in galactic H I observations) can lead to problems, as they violate the assumption of small variations of

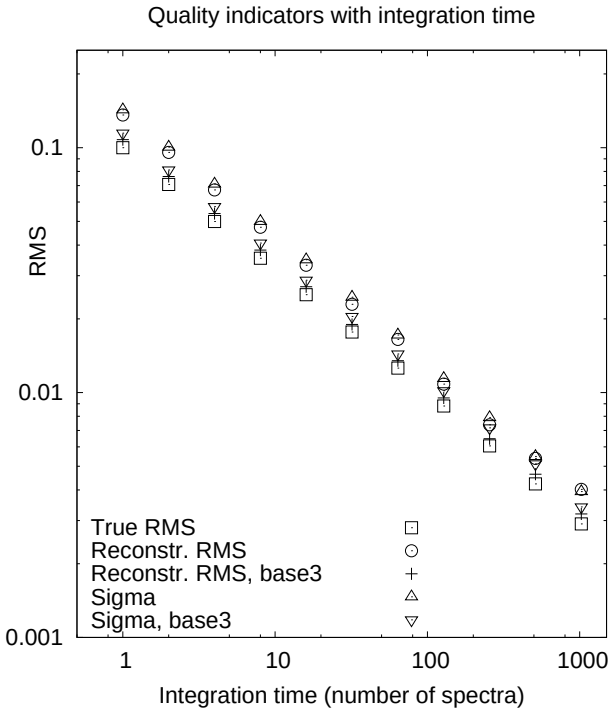


Fig. 3.15: Functional dependence of the different quality indicators vs. integration time (scans). The boxes mark the noise level (RMS) of the true signal, the circles represent the noise of the reconstructed signal. The triangles (up) mark the RMS values of the gain residual, σ, calculated using the difference of the true and reconstructed gain curves. After subtracting a third-order polynomial both the RMS (crosses) and σ (triangles, down) quantities are closer to the theoretical value.

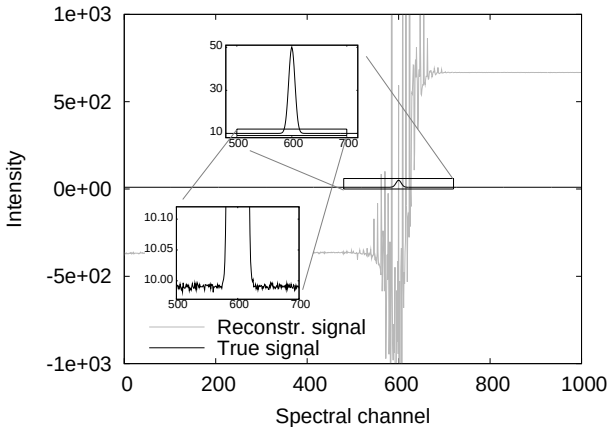


Fig. 3.16: In the presence of a strong emission line the LSFS method fails. After normalization, all spectral features must be close to unity, otherwise the linear approximation is no longer applicable. A solution to the problem is remapping of the input signal; see Fig. 3.17 and Fig. 3.18.

the (normalized) signal around unity. Fig. 3.16 shows that for a strong emission line LSFS fails completely; the intensity of the signal at spectral channel 600 is about 5 times higher than the baseline level (system temperature). This causes heavy distortions during the reconstruction process (comparing the relative amplitudes of the input and reconstructed signal).

A solution to the problem could be identified: by remapping the observed (normalized) signal, P, in terms of a nonlinear function one can treat strong signals into the realm of small variations around unity. In the example the transfer function $P \rightarrow \sqrt[x]{P}$, with $x = 4$, was applied. It is not clear, though, whether the reconstructed signal and bandpass can be transformed back by simply using the inverse $P \rightarrow P^x$. But actually it turns out, that this is possible; see Fig. 3.17.

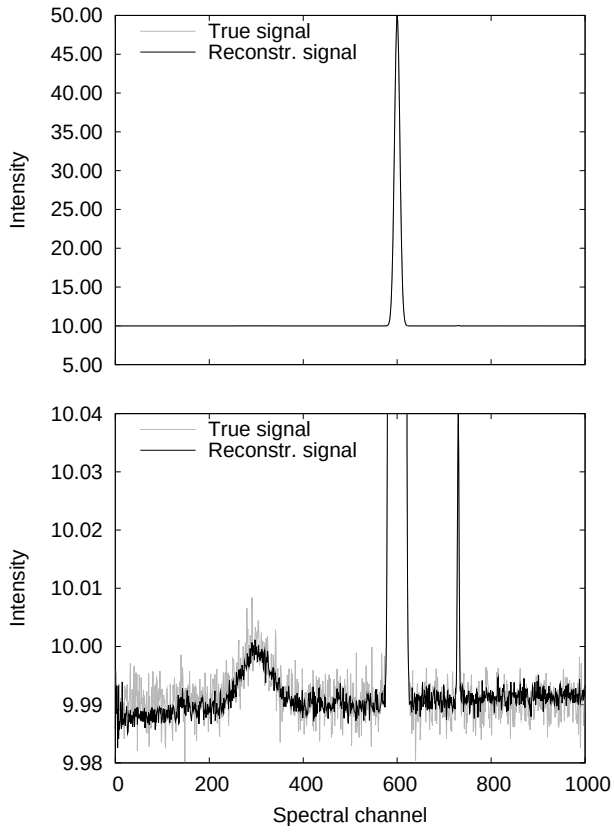


Fig. 3.17: Strong emission lines (top) can be handled by remapping the measured signal with a nonlinear transfer function, e.g. $P \rightarrow \sqrt[4]{P}$. This ensures the LSFS method to be in the linear regime. The bottom panel shows a zoom-in for better visualization. The quality indicators are shown in Fig. 3.18.

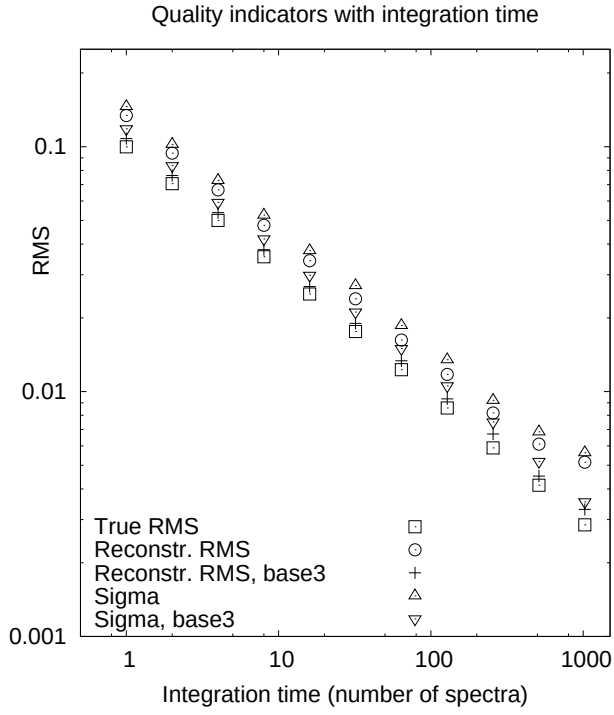


Fig. 3.18: Strong emission lines (top) can be handled by remapping the measured signal. The quality indicators (see Fig. 3.10 for the explanation of the symbols) show that the remapping works correctly in a statistical sense. There is no significant increase of the RMS or σ values compared to the undisturbed case; see Fig. 3.10.

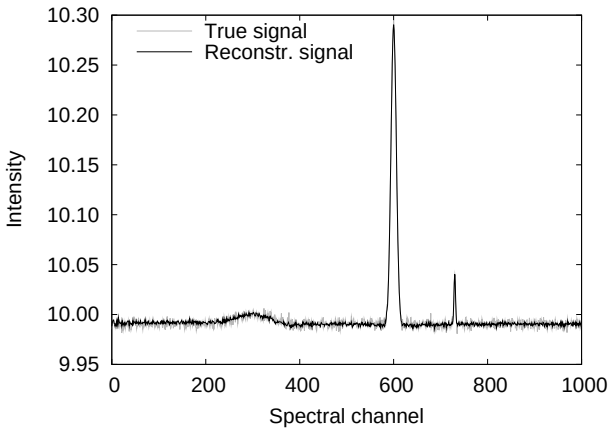


Fig. 3.19: A slowly changing bandpass shape (constant gain curve during one LO cycle) has no significant influence on the LSFS method. The signals can be treated as well recovered (see Fig. 3.20).

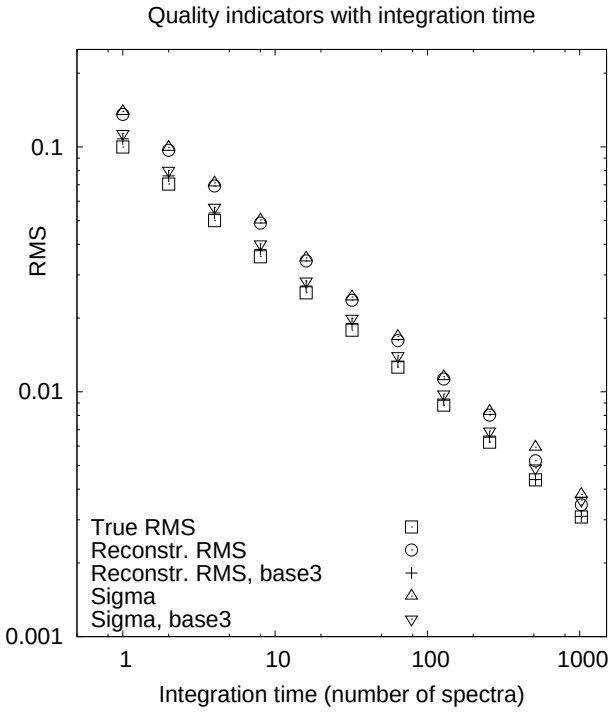


Fig. 3.20: A slowly changing bandpass shape (constant gain curve during one LO cycle) has no measurable influence on the LSFS method. The quality indicators (see Fig. 3.10 for the explanation of the symbols) show no significant increase of the RMS or σ values compared to the undisturbed case; see Fig. 3.10.

Bandpass instabilities

To further test the statistical stability, the bandpass shape and amplitude were changed with time. First, only the shape was slowly changed, though using the same shape for each bandpass within a single switching cycle (8 adjacent spectra have the same shape). This should resemble the situation at the telescope site as we can (hopefully) expect the bandpass shape to be independent of switching frequency, keeping in mind that the frequency shifts are very small compared to the total bandwidth. No significant difference to the undisturbed case was observed; see Fig. 3.19 and Fig. 3.20.

As the slowly changing bandpass shape was no challenge for the LSFS algorithm, the bandpass shape was also changed more rapidly, but in a manner that there are no

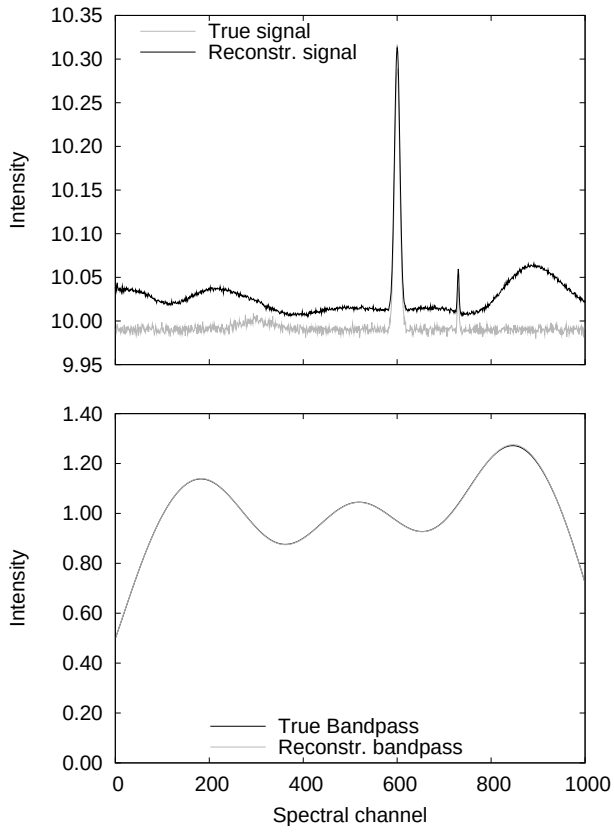


Fig. 3.21: A rapidly changing bandpass shape (see text) pushes the LSFS method to its limits. The bandpass and signal (top and bottom panel) were not well reconstructed. The difference between true and reconstructed gain curve as well as the baseline of the reconstructed signal can not be described by a low-order polynomial; see Fig. 3.22 for the quality indicators. Note, that due to the rescaling of the signal the uncertainties are much more visibly prominent in the signal domain than in the gain curves.

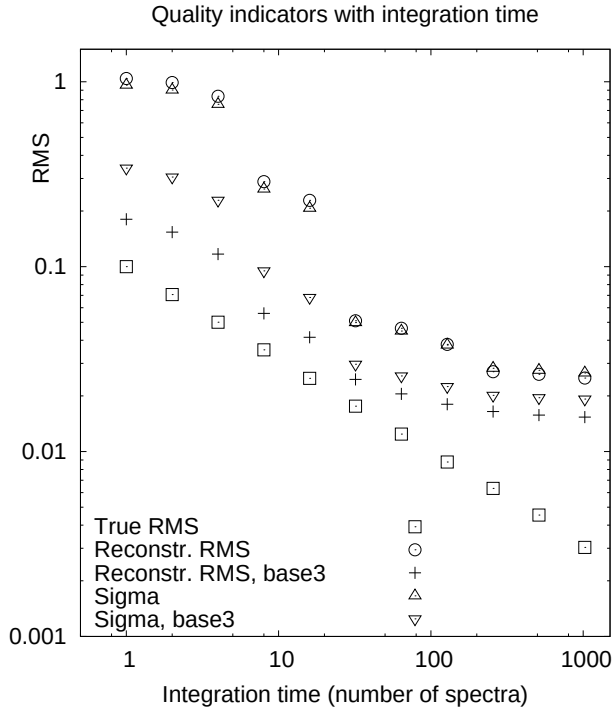


Fig. 3.22: Quality indicators (see Fig. 3.10 for the explanation of the symbols) for a rapidly changing bandpass shape (see text). The difference between true and reconstructed gain curve as well as the baseline of the reconstructed signal can not be described by a low-order polynomial. Both the RMS and σ values are much higher than in the undisturbed case and their functional behavior is far from linear.

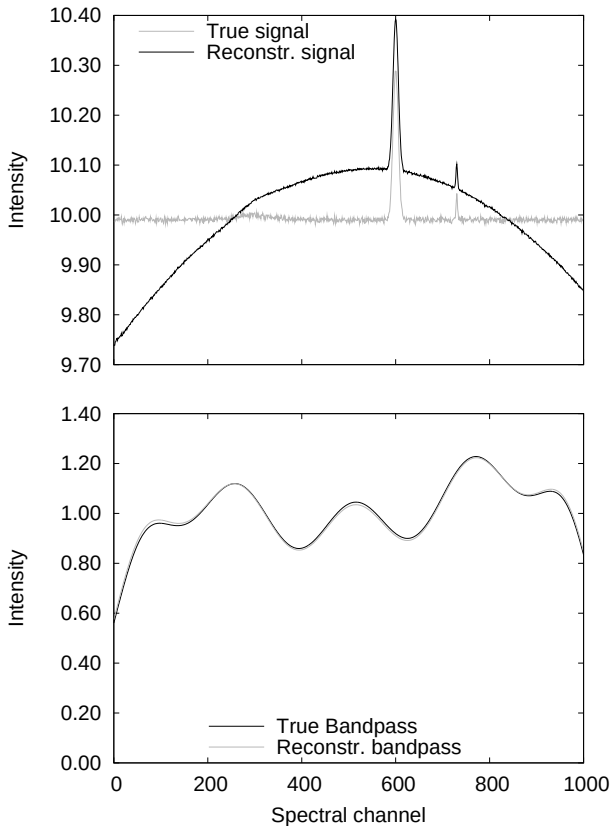


Fig. 3.23: In case of a systematic change of the bandpass shape dependent on the shift frequency (each bandpass was multiplied with a linear function, the slope of which was steeper for higher shifting frequencies) LSFS fails to reconstruct the signal (top) and bandpass (bottom); see Fig. 3.24 for the quality indicators.

systematic differences between the different LO phases. The outcome of this is shown in Fig. 3.21 and Fig. 3.22. The LSFS method could not reconstruct the signal and gain curve. The residual is smooth but can only be described by a high-order polynomial. In fact, computing the RMS with respect to a third-order polynomial results in significantly increased noise values and σ.

For completeness, the shape of the bandpass was also changed in a more systematic manner by multiplying each bandpass with a linear function which slightly drops off towards higher frequency (negative slope). The slope of this function was steeper with higher shifting frequencies. This mimics one of our early test observations with a digital Fast Fourier transform spectrometer prototype (Stanko et al. 2005; Winkel et al. 2007), where we were forced to use an LO frequency far off any specifications. This systematic error was combined with a slowly overall change of the bandpasses; see Fig. 3.23 and Fig. 3.24. This time, the outcome was slightly better — calculating RMS and σ with respect to a 3rd-order polynomial leads to acceptable results in case of the signals RMS. However, the value of σ does not decrease significantly after the summation of about 100 spectra. At the end the noise is about a factor of four higher than in the ideal case. The RMS level of the reconstructed signal is not significantly increased compared to the undisturbed case. On the other hand, subtracting a baseline the RMS and σ values are even nearly independent on integration time.

The latter two cases of very strong bandpass instabilities are far from any realistic scenario at modern radio telescopes. IF filter devices may have response to temperature

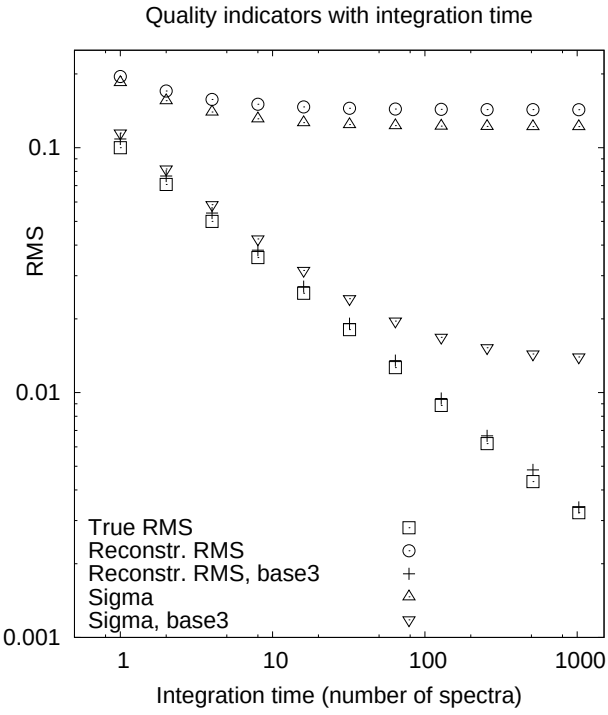


Fig. 3.24: Quality indicators for a systematic change of the bandpass shape. The residual gain curve can — to some extent — be described by a low-order polynomial, but after integration of 100 spectra the σ value (see Fig. 3.10 for the explanation of the symbols) no longer decreases.

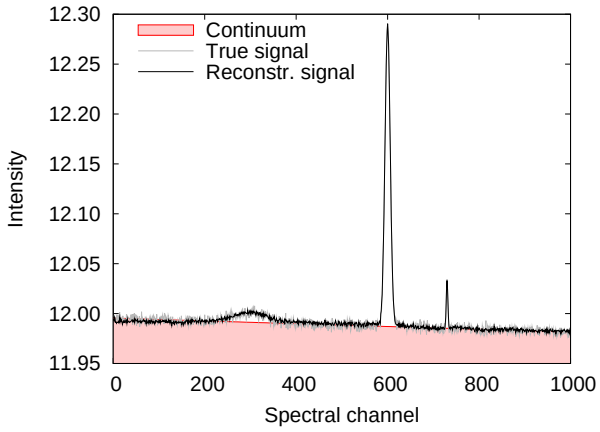


Fig. 3.25: The presence of a continuum source does not affect the result of the LSFS algorithm. The spectrum displayed, contains a continuum source of spectral index $\alpha = -2$ (shaded area). Both spectral and continuum emission are well recovered as the quality indicators (see Fig. 3.26) reveal.

and frequency variations but on a much smaller scale than were used to test the robustness of LSFS against those instabilities. In case of slowly varying gain curves the LSFS performs as good as without bandpass variations.

Continuum sources

When mapping a region of the sky one often encounters the situation that continuum sources contribute significantly to the observed signal. Using common in-band frequency-switching algorithms, it is assumed that the spectra of these continuum sources are sufficiently flat and, hence, do not lead to any significant difference between both switching phases. LSFS is much less sensitive to this bias requiring to switch only by a small frac-

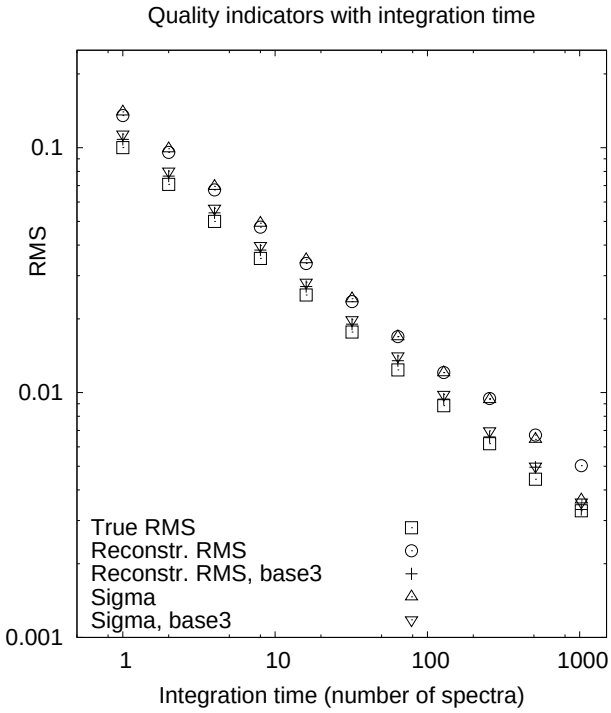


Fig. 3.26: The quality indicators (see Fig. 3.10 for the explanation of the symbols) show that the LSFS method is not affected by continuum sources. There is no significant increase of the RMS or σ compared to the undisturbed case.

tion of the total bandwidth. The advantage of LSFS in this context is the recovery of the continuum signal as part of the signal spectrum, when using the proposed normalization scheme. Since spectrometer bandwidths have grown up to hundreds of MHz or even GHz nowadays it, therefore, becomes possible to also map continuum sources 'for free' during a spectroscopic observation.

Fig. 3.25 shows the result for the case that a continuum source is superposed to the spectral lines. Its intensity is described by

$$I_\nu = A \left(\frac{\nu}{\nu_0} \right)^\alpha \tag{3.36}$$

with spectral index $\alpha = -2$ and amplitude $A = 2$ assuming $\nu_0 = 1420\,\mathrm{MHz}$ and a frequency resolution of $50\,\mathrm{kHz}$ ($\delta v \approx 10\,\mathrm{km\,s^{-1}}$) per spectral bin. Both the continuum signal as well as the spectral lines were nicely recovered, as the quality indicators (Fig. 3.26) show no increase in the RMS or σ values.

Radio frequency interference

One of the key properties of each data reduction pipeline used in radio astronomy today is the capability to handle radio frequency interferences (RFI). These artificial signals are in general variable on timescales down to μs (Fisher 2002). Therefore, one of the most interesting analyses in this Section is the impact of such interferences on the LSFS method. Starting simple, two narrow-band interferences were added whose amplitudes obey a power law. This is — at least at the 100-m telescope at Effelsberg — one of the most common types of interference (Winkel 2005).

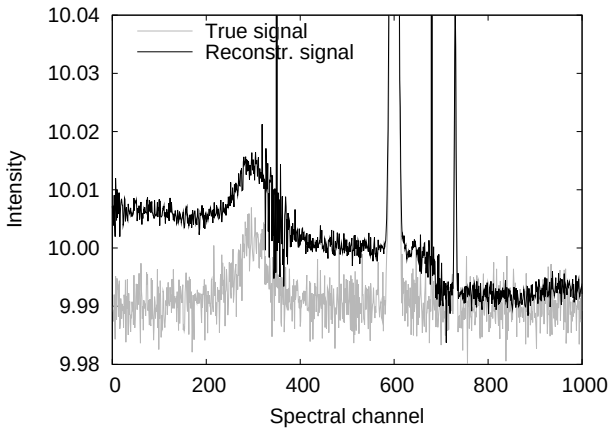


Fig. 3.27: RFI signals can have a severe effect on the solution of the LSFS. In spectral channels 350 and 680 a narrow-band interference signal was added. The fast-varying behavior of the simulated RFI signals produces strong distortions in the reconstructed signal. For better visualization the plot shows a zoom-in. Fig. 3.28 contains the quality indicators.

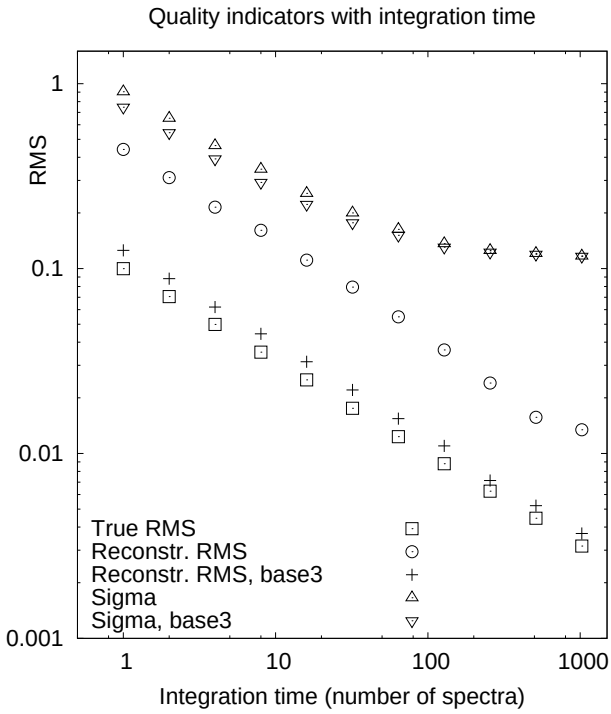


Fig. 3.28: The quality indicators (see Fig. 3.10 for the explanation of the symbols) are very sensitive for RFI signals. After subtracting a third-order baseline the signals noise level is enhanced by about 20%, while σ is even increased by a factor of $\gtrsim 4$.

As the LSFS algorithm assumes the signals to be stable in time, it was not surprising that the resulting spectra had to low quality for a scientific analysis. Due to the coupling of channels with different frequency shifts, one ends up with a number of contaminated spectral channels which is higher than the initial number of channels affected[3]; see Fig. 3.27. The only solution is actually to address the RFI problem before performing the LSFS.

Winkel et al. (2007) presented an algorithm which detects interferences down to the $\lesssim 4\sigma_{\mathrm{rms}}$ level; see Section 3.2. Having detected interference peaks, bad data can be flagged in order to exclude them from following computations. Flagging data points is equivalent

[3] Usually interferences enter the system via the front-end, meaning that for each LO shift the RFI peak is mixed into a different spectral channel. If the RFI is varying, the effect is similar to a "randomly" and rapidly changing gain curve.

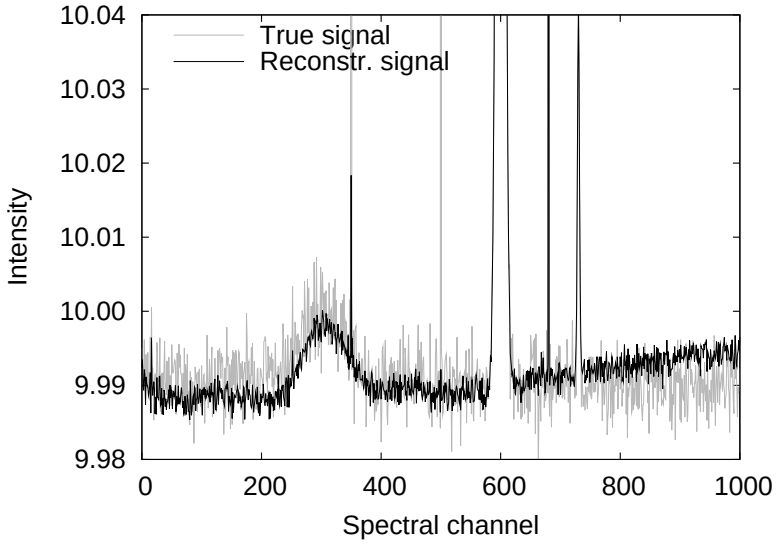


Fig. 3.29: Proper handling of RFI contaminated data points allows the reconstruction of the original spectra. In the signal domain residual RFI peaks remain, but have a low intensity and no significant influence on the neighboring channels as shown in Fig. 3.27. Three narrow-band interferences were simulated. The RFI signal in spectral channel 350 was added to all LO phases except for LO 1 and 2. The signal in spectral channel 500 affected every second LO phase while the signal in channel 680 was present in all phases. It turns out, that if an RFI signal is not persistent for a whole LO cycle the unaffected data points in the associated spectral bin can even be sufficient to reconstruct the signal without artifacts. The less LO phases are affected, the less impact the interferences have on the reconstructed signal.

to projecting the correlation matrix in Eq. (3.25) to a subspace which does not contain contaminated spectral channels. This, however, would require to recompute the SVD of the matrix when the RFI signals change their frequencies. This is far from practical as the computation of the SVD for 1024 spectral channels using eight LO frequencies lasts at least a few minutes on a modern PC.

By far easier is the following alternative: setting all spectral channels containing an RFI signal (those are of course different channel numbers for different shift frequencies) in $\boldsymbol{p}$ to zero. Of importance is here a robust calculation of the mean signal strength by dropping all disturbed spectral channels. Otherwise, the gain factors would depend on the actual strength of the RFI signals.

The current RFI detection software does not try to identify RFI below the noise level (though an iterative scheme may be possible, were the search for interferences is performed at different integration times). Hence, only those spectral channels were set to zero containing an interference signal larger than $4\sigma_{\mathrm{rms}}$. This time, three narrow-band RFI signals were added with amplitudes following a power law with spectral index $\nu = -1.5$. The leftmost signal at spectral channel 250 was persistent in all LO settings except 1 and 2. The signal at channel 600 was only added in every second LO and the rightmost interference at channel 680 was present for each LO.

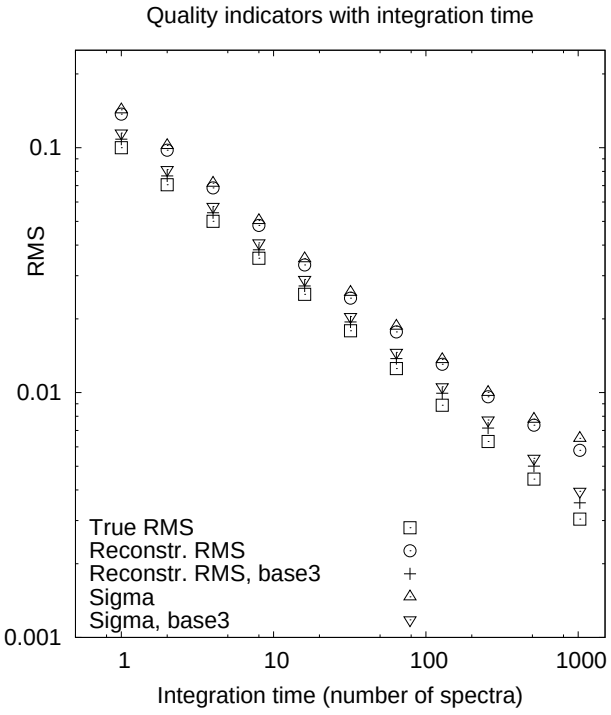


Fig. 3.30: Using the flagging scheme to suppress distortions by RFI signals provides noise level (RMS) values (bottom; see Fig. 3.10 for the explanation of the symbols), being only $\sim 14\%$ higher and a value of σ, which is $\sim 25\%$ higher than theoretically expected.

The outcome is shown in Fig. 3.29 and Fig. 3.30. The bandpass is well-recovered. However, each RFI leaves behind some 'fingerprint' in the reconstructed spectrum, the residual strength of which obviously depends on the number of affected LOs. These "left-overs" are, nevertheless, easy to handle as the spectral channels and LOs containing RFI are more or less known (otherwise also the flagging would not have been possible).

Implementing RFI flagging enables the analysis of the response of the LSFS to different types of RFI. During our measurements broad-band events (which last for only a second or less but affect several hundred spectral channels; Winkel et al. 2007) were rarely encountered. Fig. 3.31 and Fig. 3.32 show the result for affecting the 4th LO within spectral channels 200 to 400 — the reconstruction was successful when using the proposed flagging scheme — otherwise the method fails. The intensities of the broad-band signal are drawn from a power law but lie within the range $4 \ldots 20\sigma_{\mathrm{rms}}$. The interference was added to each spectrum of the 4th LO which would hardly be the case for a real observation (this type of RFI is rare).

Computational efficiency

When we started the analysis by implementing the LSFS within the C programming language we chose for the sake of simplicity the SVD algorithms delivered with the GNU Scientific Library (GSL)[4]. They make use of the modified Golub-Reinsch algorithm. However, while calculating the LSFS the computational speed can be increased, using the fact that the matrix is sparse. There exist a few libraries (mainly for FORTRAN) which use the

[4] `http://www.gnu.org/software/gsl/`

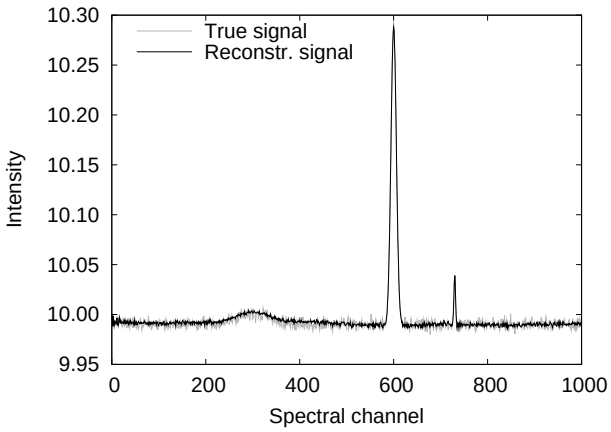


Fig. 3.31: LSFS for a broadband interference signal. As only one LO frequency is affected, the signal and bandpass could be very well recovered. Fig. 3.32 shows the quality indicators which are only slightly increased compared to the undisturbed case.

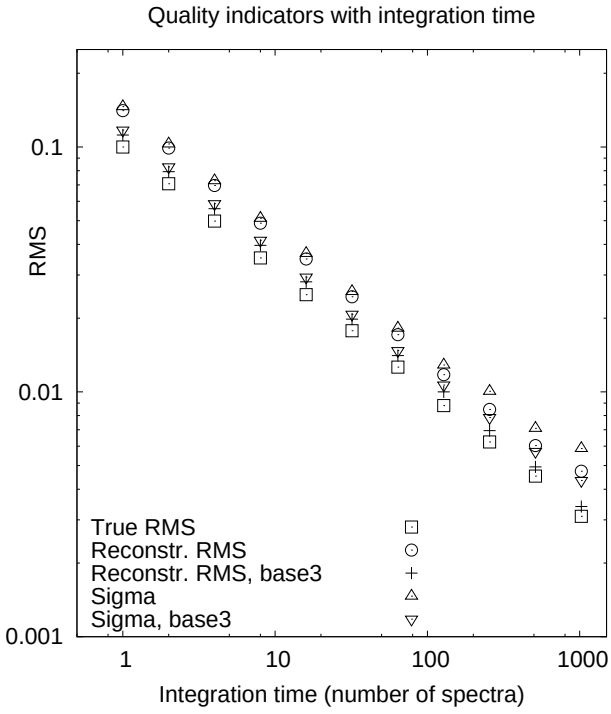


Fig. 3.32: Quality indicators (see Fig. 3.10 for the explanation of the symbols) for spectra containing a broadband interference signal.

Lanczos (SVD) algorithm for sparse matrices. Here, the `las2` routine from SVDPACKC[5] was utilized through the interface library SVDLIBC[6]. Table 3.3 lists run-times of the pure SVD computation for different I, comparing both methods. The `las2` algorithm is about an order of magnitude faster, which means the SVD of a two times larger matrix can be calculated within the same time (as the SVD computation scales as I^3). Therefore, it is the preferred method for large values of $N \cdot I$. Note also, that the main memory needed scales roughly as NI^2. For the largest of our problems ($I = 2048$, $N = 8$) a 1-GB-machine was barely sufficient using double precision arithmetic.

[5] `http://www.netlib.org/svdpack/`
[6] `http://tedlab.mit.edu/ ~dr/SVDLIBC/`

Table 3.3: Computing times needed to calculate the SVD using different algorithms.

I	N	rows	cols	matrix density	time (s)[a] sparse	gsl
128	8	1025	300	0.72	$\simeq 1$	2
256	8	2049	556	0.39	3	20
512	8	4097	1068	0.20	28	168
1024	8	8193	2092	0.10	220	2454
2048	8	16385	4140	0.05	2573	N/A

[a]Using a 2.0 GHz x86 CPU.

As the SVD needs only to computed once per LO setup the more important contribution to computing times needed is due to the LSFS calculation itself. Based on our experience in many cases the convergence is reached after few ($\lesssim 5$) steps. When confronted with instabilities (RFI, etc.) this increases up to 20 or more iterations. To account for this changes of the solution are monitored. This allows to stop the iteration once the solution has stabilized. The computation of the LSFS ($N = 8$, $I = 1024$) lasts $\sim 0.05\,\mathrm{s}$ per iteration step on a modern desktop PC (2.0 GHz, x86). An optimized BLAS library was used, which makes use of SSE or equivalent features of modern x86 CPU's meaning there is probably not much potential to further speed up the computation of the LSFS.

Some tests were done utilizing the compiler extension OpenMP[7] to parallelize the LSFS for use on multi-processor/core machines. This improved the run-time by about 25% on a Dual-Xeon machine and about 15% on a Dual-Core processor. The maximum speed-up one could expect would be a factor of two, which is not reached practically. In fact, the LSFS computation depends mainly on the multiplication of the (huge) correlation matrix with the input vector. Here, the memory bandwidth has large impact on the overall speed, as well.

3.5.5 LSFS at the 100-m telescope Effelsberg

It is under investigation, whether the LSFS method can be implemented at the 100-m telescope. Currently two LO frequencies are provided by two ultra-stable oscillators (which is needed for VLBI). A drawback is, that the programming (of a new frequency) needs several seconds and can only be done between recording data. For frequency switching both LO are used and a hardware switch is applied. This can be done also during a measurement.

However, the LSFS needs several different frequencies and obviously long duty cycles should be avoided. Eventually, it is possible to implement a new oscillator, which can be programmed faster.

[7] `http://www.openmp.org/`

3.6 Gridding

The visible sky can be described by spherical coordinates. Each position is defined by two coordinates, azimuth and elevation. Being for many cases quite impractical they are usually converted to more general coordinate systems, e.g., equatorial or galactic coordinates. The spherical systems, however, make it rather complicated to measure larger areas of the sky on regular (cartesian) grids. Locally, a tangential plane can be used to approximate positions. In that case the observer can relatively easy map a (small) region equidistantly. However, if larger fractions of the sky must be described, the cartesian approximation no longer holds.

For such cases, the data must be gridded (or regridded, if an alternative projection system is desired). This means that from the raw input spectra (or data points, if photometric) a cartesian data cube[8] (or pixel map) will be computed. Note, that only the pixel coordinates lie on a rectangular grid, while the associated true (sky) coordinates do not, as for a 2-dim map of the world. Another example is Fig. 1.6 where the pixel grid matches the galactic coordinates, while the equatorial system is "distorted" in this representation. Many projection systems are in use, mostly being area- or distance-conserving, depending on the kind of application and part of the sky which is to map (complete, or in the vicinity of poles or equator).

While it is (relatively) easy to convert the sky coordinates to the regular pixel representation, the gridding procedure itself is more complex. Here, it must be considered, that in the general case the measured data points (spectra) are not equally distributed with respect to the pixel grid. Then, the resulting value in each pixel is influenced by a different number of data points. Regardless of the gridding algorithm, this can be problematic itself, as the final noise level might vary in distinct areas in the map. One of the key problems for large-scale observations is, therefore, the proper distribution of measured data points on the sky.

In this Section it is assumed, that the data are properly distributed in this respect. Then, only the specific method of combining/merging the data around each pixel must be developed. Note, that many typical (re)gridding algorithms, as (bi)linear or spline interpolation cannot be used, as they do not work for scattered data (positions).

The most simple gridding method would be to calculate the mean of all data points accessible within a certain radius for each pixel. Because the data are fully sampled this guarantees the proper reconstruction of the underlying intensity distribution (features of sizes smaller than the sampling rate are smoothed).

A very basic approach is to compute for each pixel the average of all data points accessible within a certain radius, eventually weighted with inverse distance. A bit more computational expensive was the method used initially by Barnes (1998) for the HIPASS data who chose a median estimator instead of the mean, to suppress RFI signals and increase robustness against various instabilities. This works on scattered data and provides a proper normalization in Cartesian coordinates — it is desired to conserve the (total) flux of each source, but *not* the peak flux, which depends on the beam size. However, for different projection systems in use the normalization scheme needs to be adapted to

[8] A data cube contains spectral information for each spatial position (within a certain range) on a regular grid. Each frequency (or velocity) plane contains a 2-dimensional intensity map for that specific frequency. One can also easily compute projections in (longitude, frequency) or (latitude, frequency).

the different metrics. Hence, the weighted averaging method was modified slightly to provide conserved flux regardless of the projection without slowing down the computation too much. It should also be mentioned, that Barnes et al. (2001) finally used a more complicated method to grid the HIPASS data. Their method is still based on the median estimator and has some drawbacks, as an increased noise level, a (arbitrarily chosen) scale radius which affects the result, and it is computationally expensive. The median being a non-linear operator could also introduce unexpected effects.

Before the gridding procedure is described in detail a few notes should be made. First, it is very important to avoid aliasing, which can be introduced when the sampling interval does not match the physical resolution of the data, leading to a wrong "reconstruction" of the true sky. The aspect of sampling (and aliasing) is discussed in Section 3.6.2. Second, in radio astronomy two common measures of the flux exist, the brightness temperature, T_B in units of Kelvin and the flux density S in units of $\mathrm{Jy\,Beam^{-1}}$. Both are related via

$$T_\mathrm{B} = \frac{\lambda S}{2\mathrm{k_B}\Omega} \tag{3.37}$$

with wavelength λ, the Boltzmann constant $\mathrm{k_B}$ and beam solid angle $\Omega = \theta_a\theta_b$. The flux density, S, is dependent on the size of the telescope beam. This must be considered when comparing measurements from different telescopes. T_B is not influenced by the beam size due to the correction with Ω. Furthermore, the beam filling factor, i.e., the fraction of the main beam which is filled by the astronomical source, has a large impact on the results. While for an extended object, the filling factor equals 100% it can be significant smaller for point-like objects. In this case, the measured fluxes do not represent the true value. Also, different telescopes (having different main beam sizes) would not observe the same flux. Usually, it is not possible to correct for the filling factor, when the size of the object is not known.

The basic principle of the gridding is as follows: the datapoints are filtered (convolved with a Gaussian kernel), which is nothing else than averaging with respect to a well-defined scale, the kernel-size σ_sm. This only slightly increases the resulting angular resolution according to

$$\sigma_\mathrm{total} = \sqrt{\sigma_\mathrm{tel}^2 + \sigma_\mathrm{sm}^2}\,. \tag{3.38}$$

In example, the beam-width of the 100-m telescope of about $9'$ (FWHM) leads to a spatial resolution of $10\farcm5$ when convolved with a Gaussian of width $5\farcm4$. The filtering ensures, that for every pixel of a given coordinate grid a good estimate of the nearby datapoints is calculated, regardless of the number of accessed datapoints. It should be noted, that this not yet conserves the fluxes of the sources. Furthermore, this basic principle is rather computationally expensive, as for the filtering all relevant data must be considered. Hence, a slightly modified algorithm is used, which is equivalent. Using spectral dumps of about 1 s integration time produces a huge amount of data — up to a million of spectra for a single datacube of size $10^\circ \times 10^\circ$. To keep all input spectra within the computer memory is not feasable today. Therefore, a serially working algorithm is preferred.

Such a procedure can be setup, by first, convolve the flux value for single each datapoint with a Gaussian and add the result to the data grid. In addition the associated Gauss weighting value is added to the kernel grid. The kernel grid can then be used to account for unequally distributed data points, by simply divide the final data grid by the

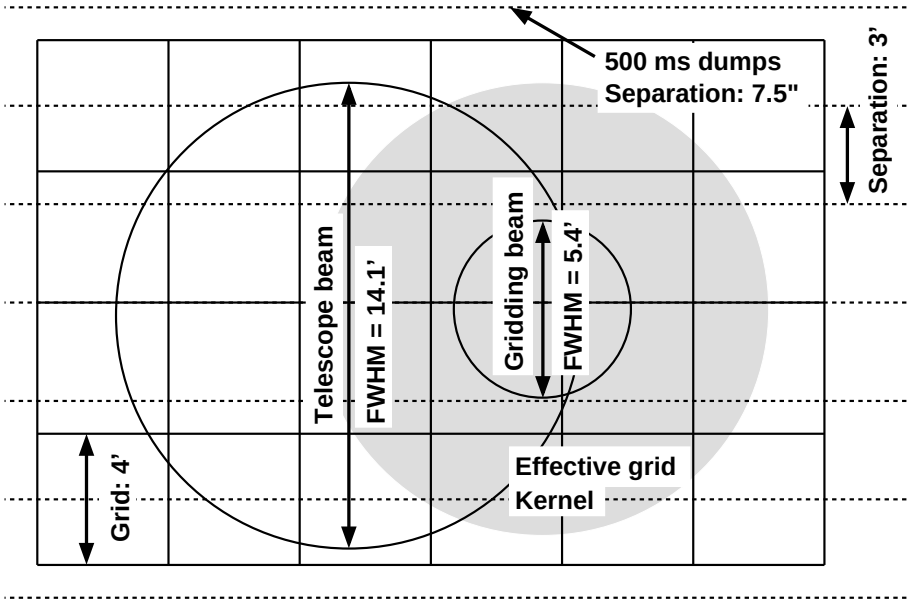


Fig. 3.33: The relevant gridding parameters — telescope beam size, size of the gridding kernel, and grid cell size — are schematically plotted. In order to avoid aliasing the data must be (spatially) fully sampled: the separation between two positions, i.e., the between two scanlines, must not be too large (according to the Shannon-Nyquist theorem; see Section 3.6.2). Note, that the choice of the angular resolution or size of the grid cells, respectively, is independent on the Nyquist condition. The Gaussian function representing the gridding beam which is used for the filter kernel has FWHM of $5.4'$. The gridding kernel is computed to a size of $3\sigma_{\mathrm{sm}}$ (shaded area) including more than 99% of its integral value.

"normalization weights" from the kernel grid. This scheme works also for general projection systems, because the normalization is intrinsically adapted to the same projection system.

Another advantage of this serial approach is, that flags (e.g., from RFI detection) can be easily accounted for. If a datapoint is (completely) flagged as bad, it can just be neglected. This would not affect the normalization. But even if only some spectral channels are flagged, the same scheme can easily be used — one only needs to store a kernel datacube, containing the normalization for each pixel and each velocity plane. The serialization makes it easy to parallelize the computation to make use of multi-processor or -core platforms. Here, the input data stream can be easily split-up into several parts, which then are processed individually.

In Fig. 3.33 the relevant gridding parameters are shown. Exemplary, the values chosen are those, which will be used for the simulations in Chapter 4. The size of the grid cells is $4'$, the telescope beam has $14\farcm1$. For a hypothetical dump rate of 500 ms and if the observation would have been made using drift scans, each single dump would be separated by $7\farcs5$, while the scan lines should have a separation of $3'$ to guarantee full angular sampling. The

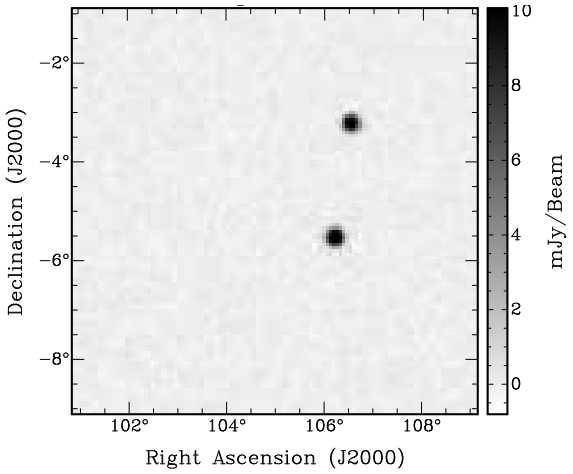


Fig. 3.34: Plane of a data cube gridded from simulated data using the proposed method. It contains two Gaussian-shaped galaxies.

gridding kernel has 5′.4, which means that the effective kernel area[9] is nearly as large as the telescope beam.

3.6.1 The gridding software

In order to provide a convenient interface to the gridding algorithm explained in Section 3.6 the implemention was done in C++/Qt[10]. Qt provides a sophisticated widget library, as well as container classes (arrays, lists) and multi-threading support. The latter makes it easy to use multi-processor or -core systems to accelerate the gridding. The Fits-I/O is done using the CFITSIO[11] library. For computations regarding projected coordinate systems the world coordinate system (WCS) enhancement of CFITSIO is applied (WCSlib; Greisen & Calabretta 2002; Calabretta & Greisen 2002; Greisen et al. 2006).

The graphical user interface (GUI) allows the user to load the input files containing the spectra to be gridded.The program will show the spatial positions of the data points (spectra) according to the currently chosen coordinate system.At the moment the WCSlib provides $\sim$ 20 projection systems. The user can also convert the Ra–Dec coordinates to Galactic coordinates, as well as define the relevant fits header entries (which define the number of the pixels in the grid, spatial resolution, the reference pixel, etc.). The viewport will update its content on-the-fly. For convenience, one can also change the displayed coordinate mesh.Finally, the beam-size (according to the input spectra), the desired kernel-size and the maximum radius out to which the filtering is carried out can be adjusted. For supervision purposes a weight map is created, as well.

[9] The effective kernel area is for computational reasons limited to $3\sigma_{\mathrm{sm}}$. Note also, that the FWHM size is not equal to σ_{sm}. The latter is defined via the normalized Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right], \qquad \int_{-\infty}^{\infty} \mathrm{d}x\, f(x) = 1\,. \tag{3.39}$$

The FWHM is linearly connected with σ, as $\mathrm{FWHM} = \sqrt{8\ln 2}\,\sigma \approx 2.35\sigma$. Within the interval $[-3\sigma, 3\sigma]$ about 99% of the integral value is enclosed.

[10] `http://trolltech.com`

[11] `http://heasarc.nasa.gov/docs/software/fitsio/fitsio.html`

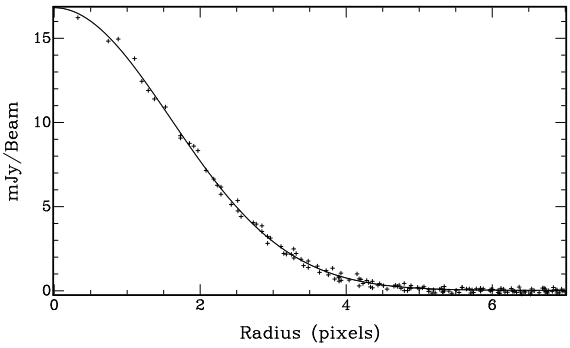


Fig. 3.35: Radial profile of one of the galaxies shown in Fig. 3.34. The fit represents a Gaussian profile with radius of 3.77 pixels corresponding to 15.07′ (FWHM) which matches the theoretical expected angular resolution of 15.1′ very well.

Fig. 3.34 shows a plane of a data cube gridded from simulated data using the proposed method. It contains two Gaussian-shaped "galaxies". In Fig. 3.35 the radial profile of one of the galaxies was computed. The best fit of a Gaussian profile has a FWHM of 15.07′ which matches our theoretical expectation. After applying the 2-dimensional Gaussian fit the spatial position generally is not aligned to the pixel grid. This causes the data points in the plot to be not equidistantly distributed. The number of grid cells increases quadratically with distance to the central position.

A small modification to this scheme is necessary if one wants to include handling of flagged data points as delivered by the RFI detection software. Data points found to be polluted by interferences are not added to the cube. This can result in pixels containing very few (or no) information (this is indeed likely for spectral channels influenced by long-lasting narrow-band RFI). To deal with this problem a second data cube is assembled consisting of the numbers of values contributing to each pixel. After gridding pixels having not enough "informational content" are filled with the median value of their surrounding. An iterative scheme is used, which increases the box size used to calculate the median until enough "good" data points lie within.

3.6.2 Aliasing

An important effect which may occur when sampling and gridding mapped observations is aliasing. From Shannon's theorem[12] (Shannon 1949) it is known that a measurement using discrete sampling intervals can only used to reconstruct features above the Nyquist frequency which is determined by the sampling interval. If signals of a higher bandwidth are involved, downfolding into the first Nyquist zone occurs producing so-called aliasing. One simple method to avoid aliasing (e.g., when downsizing digital images) is to first convolve the signal with a bandwidth-limited filter (smoothing) neglecting all higher-frequency features according to the desired sampling interval.

In radio astronomy, however, this is not possible, as the the true sky brightness distribution is convolved with the telescope beam itself. The beam size cannot be changed, therefore it is necessary to fulfill the Nyquist condition during observing. This means that

[12]Often referred to as Nyquist's theorem. However, the theorem and it's proof were stated by Shannon based on the work of Whittaker (1915) and Nyquist (1928).

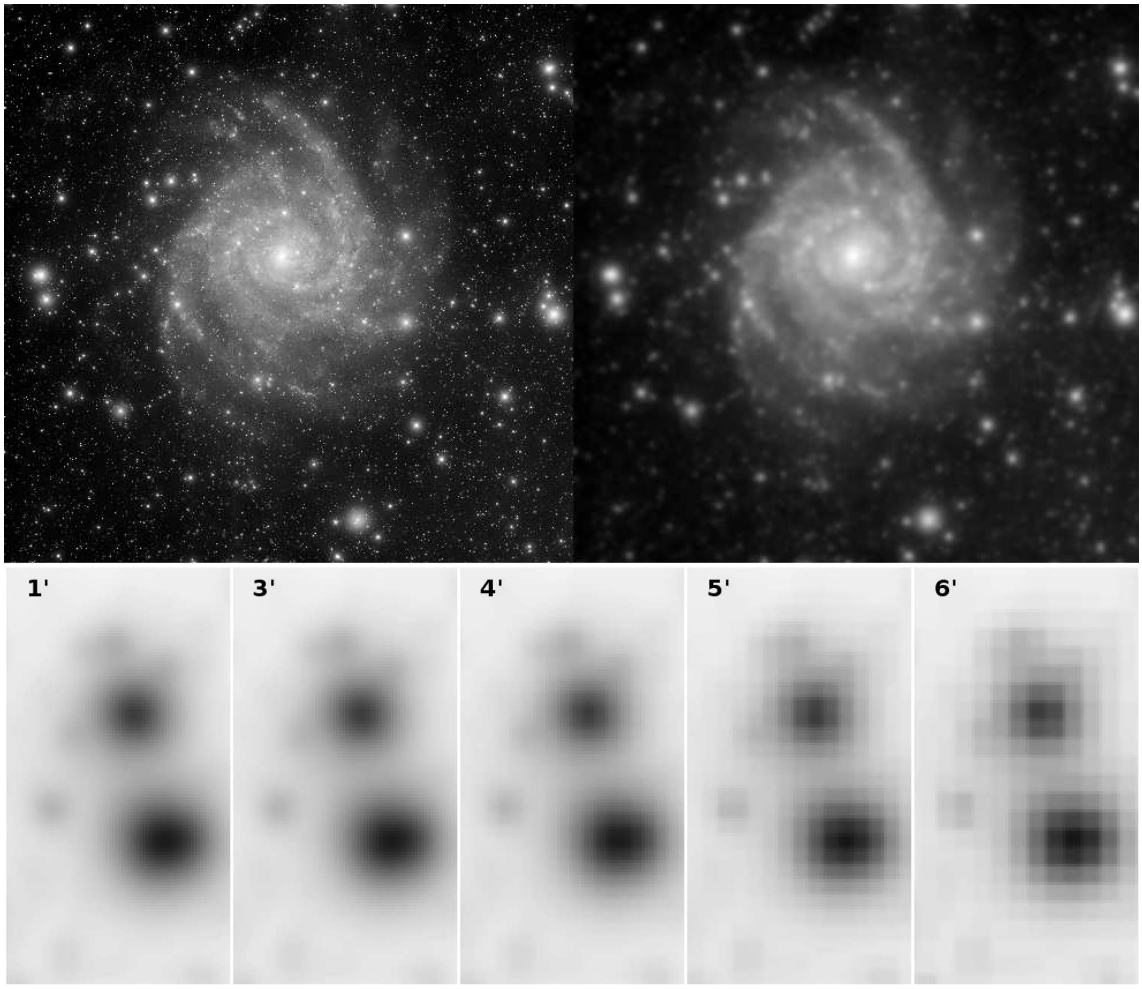


Fig. 3.36: Simple simulation to visualize the effect of aliasing. The image in the top left panel is convolved with a Gaussian of $9'$ (assuming $1'$ equals the size of a pixel) resulting in a "smoothed" image. Using the values in each n-th individual pixel is equivalent to a sampling interval of n arcmin. By (re-)gridding one can reconstruct the image. Using a sampling interval of $1'$ leads to the map in the upper right panel. The lower panels show a zoom-in region for different sampling intervals of $1'$, $3'$, $4'$, $5'$, and $6'$. Starting with a sampling interval of $4'$ a rectangular pattern appears, produced by not fully sampling the data, i.e. not fulfilling the Nyquist condition.

the sampling interval, i.e., the distance between each adjacent pointed observation (within the map) must be smaller than half of the beam size[13].

To visualize the effect of aliasing a simple simulation was performed, see Fig. 3.36. First, the true brightness distribution (upper left panel of Fig. 3.36) was convolved with a Gaussian of width $9'$ (FWHM, for convenience the size of each pixel was define to be $1'$)

[13]On rectangular grids this must also be fulfilled for the diagonal distances, i.e., the grid cell size needs to be FWHM/$(2\sqrt{2})$.

leading to the "observed" brightness distribution. Then, different sampling intervals were chosen to generate a "raw" data set of pointed observations. The data were gridded using a kernel width of $4'$ (slightly enlarging the effective beam size to $\sim 10'$; see Section 3.6 for details). The upper right panel shows the regridded map for a sampling interval of $1'$. The lower part of Fig. 3.36 shows the result for a small region of the full image. Starting with a sampling interval of $4'$ a rectangular pattern appears caused by not fully sampling the data, i.e. not fulfilling the Nyquist condition.

3.7 Galaxy finder algorithm and source parametrization

For the extra-galactic part of the EBHIS several thousands of sources are expected to be in the data cubes. Such a huge number can hardly be searched for and parametrized manually. One very challenging task is to develop algorithms which are able to automatically detect (point) sources down to the sensitivity limit of the survey. In example for the HIPASS data two different source finders were applied (see Meyer et al. 2004, and references therein), a simple peak flux threshold method (MULTIFIND), and an algorithm based on cross-correlation of the spectra with a top-hat profile of various widths (TopHat). While both tasks find about 90% of the source, the former method produces many more false positive detections. For that reason Wong et al. (2006) used only a revised version of TopHat for the Northern HIPASS Catalogue. However, the large overhead made it necessary for HIPASS to manually check each of the detections meaning a huge effort for the astronomers involved. In this Section two promising finder algorithms (using the Gamma test and "X-rays") are discussed, as well as a graphical user interface which was developed for the detection (manually and automatically) and parametrization of extra-galactic sources. One has to keep in mind, that automatic source detection routines will probably never reach the same completeness level as a careful manual search would do. Here, a best compromise has to be found.

3.7.1 The Gamma test finder algorithm

One of the promising source finders, based on the so-called Gamma-test, was proposed by Boyce (2003). It was mainly developed for the HIJASS survey (Kilborn 2002), which was going to be the northern complement of HIPASS, but unfortunately due to extreme pollution with RFI turned out to be not very useful. Although preliminary results were good, it was not considered for source detection in the HIPASS data. Here, the algorithm will be described, followed by a discussion of implementation details. Its detection performance is investigated in Chapter 4.

The Gamma test is a statistical method which allows to estimate the noise level in data with an arbitrary function underlaid. This function has not necessarily to be known, but one major condition for the Gamma test to work, is, that the function is smooth (which implies continuity). The test turns out to be useful in time series modelling, where one tries to find models describing data using neural networks or genetic algorithms. A problem in training a neural network is to know when to stop the training, i.e. determine the scatter which is incorporated in the data. In astronomy for example, it is useful to have a noise estimator which is robust against baseline fluctuations. The Gamma test was

originally developed at the University of Cardiff; for a proof see Evans (2002). Here, we follow basically the work of Boyce (2003).

In order to understand the Gamma test, it is useful to start with the following simple consideration. Assume a function $f(\vec{x})$ (the so-called model). Then,

$$y = f(\vec{x}) + r, \qquad y' = f(\vec{x}') + \tilde{r} \tag{3.40}$$

with r and $\tilde{r}$ being different noise values which are drawn from the same noise distribution. In the limit $|\vec{x} - \vec{x}'| \to 0$ one can safely assume that $|f(\vec{x}) - f(\vec{x}')| \to 0$ approaches zero because the model is continuous. It follows that

$$\frac{1}{2}(y - y')^2 \to \frac{1}{2}(r - \tilde{r})^2 \,. \tag{3.41}$$

Computing the expectation values of both sides leads to

$$E[\frac{1}{2}(y - y')^2] \to E[\frac{1}{2}(r - \tilde{r})^2] = \mathrm{Var}[r] \,. \tag{3.42}$$

The right side turns out to be the variance of the r, which was of interest. The problem in this calculation is, that for observational data the sampled points are usually not sufficiently dense to calculate this limit. The Gamma test is nothing else than the generalization of this simple procedure to discrete data sets.

First, some definitions need to be introduced. Suppose, M observations

$$(\vec{x}(i), \vec{y}(i)) \,, \qquad 1 \leq i \leq M, \qquad \vec{x}(i) \in \mathcal{C} \subset \mathbb{R}^m \tag{3.43}$$

where samples were generated according to an underlying unknown function f with noise added

$$y = f(x_1, \ldots, x_m) + r \qquad f : \mathcal{C} \subset \mathbb{R}^m \to \mathbb{R} \,. \tag{3.44}$$

Here, it is assumed that the noise distribution has zero mean, which is without loss of generality, because the model could be transformed to fulfill this.

From the continuous case it follows, that it might be a good idea to explore the continuity of f. Considering two points $\vec{x}(i)$ and $\vec{x}(j)$ which are close together, it is expected that $f(\vec{x}(i))$ and $f(\vec{x}(j))$ are also close together or if not this can only be due to noise, respectively. The Gamma method is based on the statistic

$$\gamma = \frac{1}{2M} \sum_{i=1}^{M} \left[y'(i) - y(i)\right]^2 \,. \tag{3.45}$$

Here, y' denotes the function value of the nearest neighbor of $\vec{x}(i)$, therefore Eq. (3.45) is called sometimes *Near-neighbor statistic*. Writing a similar form not only considering the first neighbor but the p^{th} nearest neighbors leads to

$$\gamma(p) = \frac{1}{2M} \sum_{i=1}^{M} \frac{1}{L(N[i,p])} \sum_{j \in N[i,p]} |y(j) - y(i)|^2 \,. \tag{3.46}$$

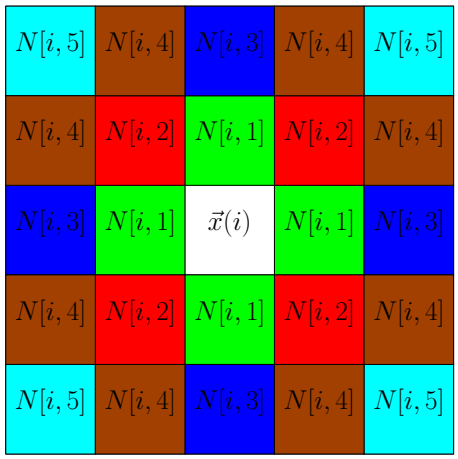


Fig. 3.37: On regular gridded data there often exists more than one near neighbor with a specific distance. This is not restricted to the 2-dim case shown here. The lengths of the lists are $L(N[i,1]) = L(N[i,2]) = L(N[i,3]) = 4$, while $L(N[i,4]) = 8$ and so on.

Furthermore, a distance measure

$$\delta(p) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{L(N[i,p])}\sum_{j\in N[i,p]}|\vec{x}(j)-\vec{x}(i)|^2 \tag{3.47}$$

$$\delta(p) = \frac{1}{M}\sum_{i=1}^{M}|\vec{x}(N[i,p])-\vec{x}(i)|^2 \tag{3.48}$$

is needed. The distance between a point and its p^{th} nearest neighbors is called $\delta(p)$. Because, even for a regular one-dimensional set of data, each sample has more than one p^{th} nearest neighbor, the distance is computed as the overall average (of *all* samples) of the distance (mean-squared difference) of all p^{th} nearest neighbors to $\vec{x}(i)$. The notation $N[i,p]$ means the list of all (equidistant) p^{th} neighbors of $\vec{x}(i)$, $L(N[i,p])$ is the size of the associated list. The second equation is only another notation for convenience. Fig. 3.37 shows exemplarily the case of a two-dimensional regular grid to clarify the notation.

Evans (2002) showed

$$\lim_{M\to\infty}\gamma = \mathrm{Var}[r] + A(p)\delta + o(\delta) \overset{\delta\to 0}{=} \mathrm{Var}[r]\,. \tag{3.49}$$

If $\delta(p) \to 0$ this becomes the pure variance of r, as long as the constant A is finite. It turns out that

$$0 < A(p) \leq \frac{1}{2}\max|\nabla f|^2\,. \tag{3.50}$$

The smoothness of the model f was required, because then the partial derivatives of f are bounded and A becomes finite. For a valid equation the first four moments of the noise distribution must to be bounded (for technical reasons in the proof of the Gamma test). Furthermore the noise on different outputs should be independent, as well as that the noise on the output needs to be homogeneous over the input space. However, if the latter is not fulfilled, this is not fatal for practical applications. The Gamma test will still provide a useful estimate for the average of the variance.

In the next step, the Gamma test computes all values of $\delta(p)$ and $\gamma(p)$ up to a certain maximum $p_{\max}$. By linear regression the vertical intercept is determined which is

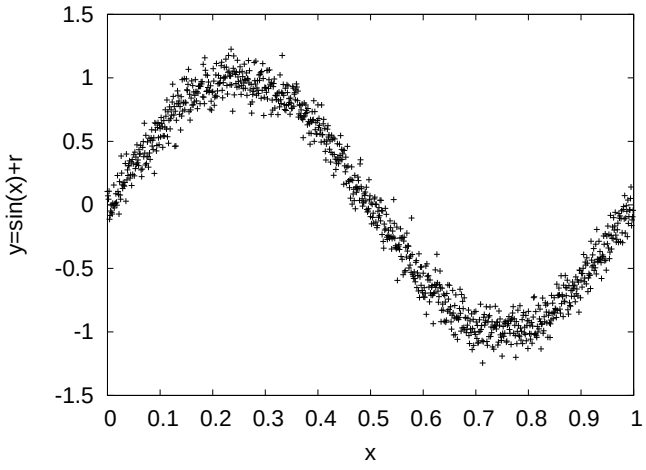


Fig. 3.38: Simple example showing a sine function with Gaussian noise added ($\sigma = 0.1$). Fig. 3.39 shows the regression plot.

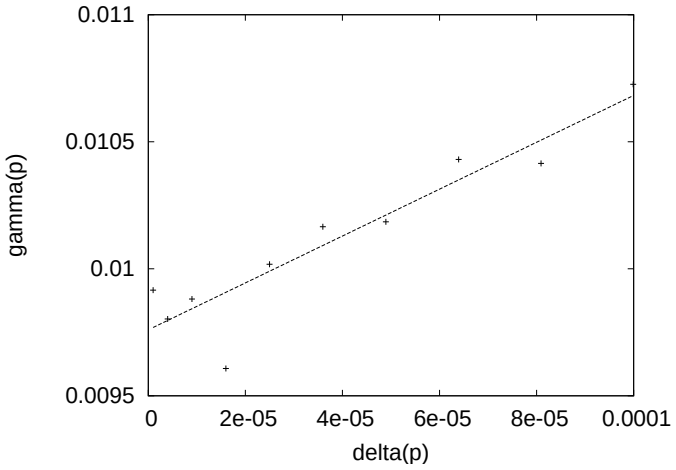


Fig. 3.39: Regression plot for the sine model example. The vertical intercept provides the Gamma test value which is an estimator for the noise variance of the data.

effectively the limit $\lim_{\delta\to 0}\gamma = \mathrm{Var}[r]$. The slope of the regression line provides further information about the complexity of the model, being proportional to the constant $A(p)$, which in turn is related to the partial derivatives of f. The computational complexity of the algorithm is of order $O(M \log M)$.

Fig. 3.38 shows an example, where a simple sine function was used as model. 1000 samples were generated and Gaussian noise with a standard deviation of $\sigma_{\mathrm{rms}} = 0.1$ was superposed. Performing the Gamma test leads first to the γ and δ values as a function of p. As shown in Fig. 3.39 these values lie more or less on a straight line. Note, that this is not necessarily true for all cases, as the method only makes a statement on the value of the intercept, not on the points themselves. After linear regression the variance was extracted, being about $9.8 \cdot 10^{-3}$. Taking the square root gives $\sigma^{\Gamma}_{\mathrm{rms}} = 0.099$ which matches very well the input noise. For comparison, the standard mean square error of the signal is about $\sigma_{\mathrm{rms}} = 0.7$, which is of course totally off, because the amplitude of the sine function is much higher than the amplitude of the noise.

An important question is, how many samples are needed to estimate the variance to a certain precision, or equivalent, how precise is the estimated noise value for a fixed number of samples. This problem can be solved using the so-called M-test. Computing the

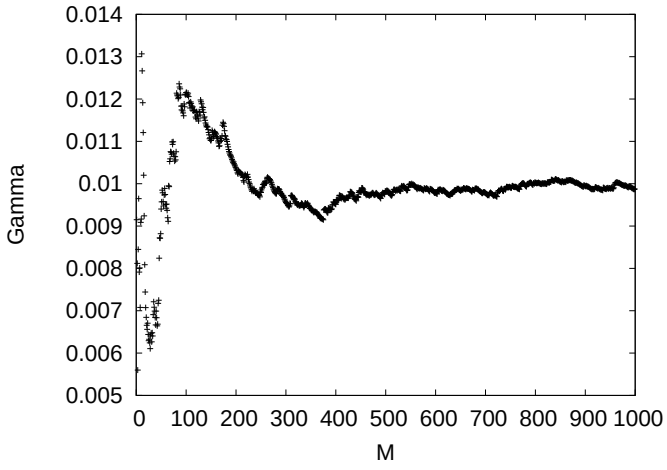


Fig. 3.40: M-test for the sine model example. As can be seen after using about 400 samples the Gamma value gets asymptotically close to the correct value of about 0.01.

Gamma values involved the complete set of samples taking mean values for all combinations of distance relations. For the M-test one chooses subsets of the data incorporating an increasing number of samples. In this way the Gamma value can be calculated as a function of M. If an asymptotic behavior of this function is observed a certain level of confidence can be estimated. Application to the sine wave example reveals that for $M \geq 400$ the Gamma values lie within a relatively sharp regime around the true variance; see Fig. 3.40. M-plots are able to answer both questions. First, if a specific precision is desired the minimal number of samples M needed can be determined. Second, one can estimate how precise the Gamma test works if only a subset of the data is used. Of course, the second analysis has to be done on the full data set (or at least a larger subset).

As already mentioned above, the application of the Gamma test as source finder does *not* use the property to determine the noise in astronomical spectra. Instead, the idea is, that the H I emission lines break the smoothness assumption for the underlying model (the baseline), leading to locally increased values of Γ. The latter is obviously a contradiction to what was said before, because the Gamma test is always applied globally. However, it turns out, that even if the computation is done on smaller subsamples, at least an enhancement of the spectral features is gained, increasing "sensitivity". Boyce (2003) proposed to use a "running" Gamma test, i.e., performing the algorithm in windows of size M centered around each of the pixels in the data cube (also denoted as voxels in the three-dimensional case). As the neighbors at certain (larger) distances might be located outside the current window, Boyce (2003) uses a kind of "wrapping" — folding in values from the opposite side of the window. In our implementation this scheme is slightly modified, by using data points outside the window, but with correct distances. Furthermore, in this work the software explicitly allows to enhance the test to three dimensions, while Boyce (2003) performed the algorithm only in the spectral dimension. After having computed an associated Gamma value for each voxel resulting in the "Gamma cube", a simple peak search provides a source candidates list. To account for different velocity profile widths of the emission lines it is possible to first smooth (e.g., with a Hanning filter as proposed by Boyce 2003) the data cube on different scales (spectral direction), and regrid it accordingly. Note, that while continuum sources and baseline ripples (e.g., caused by standing waves due to the sun) are supposably smooth, other artifacts or baseline defects as RFI peaks are not. The

latter indeed introduce a lot of problems for the source finding based on the Gamma test. This underlines the importance of a sophisticated RFI detection scheme.

3.7.2 The "X-Rays" finder algorithm

X-raying is a denotation for applying the following mathematical recursion formula, being equivalent to the radiative transfer equation,

$$I_n = I_{n-1}\left[1-(S_n)^\alpha\right] + \left[t(S_n)\right]^\beta (S_n)^\alpha, \qquad t(x) = \begin{cases} x : x \geq x_{\mathrm{thresh}} \\ 0 : x < x_{\mathrm{thresh}} \end{cases} \tag{3.51}$$

where S_n is the intensity ("emissivity") of the n^{th} pixel in the line-of-sight counted from the back plane. The emissivity is affected by a threshold function $t(x)$ so that only pixels above a certain brightness account. The opacity $\tau = (S_n)^\alpha$ determines how much light from the background a pixel will let pass. I_n is the summed total intensity of the sightline from the back plane up to the n^{th} pixel. Two free parameters $\alpha, \beta > 0$ are used to tune the operation in terms of sensitivity to faint/bright emission and how "deep" the X-rays penetrate into the cube. As the Gamma test, this method greatly enhances the contrast of sources in the datacube with respect to the noise level, being less affected by the peak flux of a source but the total (integrated) flux of an emission line. Nevertheless, it is stronger influenced by baseline fluctuations, as solar ripples, which provide a certain "flux", where, due to the thresholding, only the positive ripples are enhanced.

3.7.3 The Galaxy Parametrizer

The Galaxy Parametrizer (GaPa) is a graphical user interface (GUI) for semi-automatic finding galaxies in H I data cubes and their parametrization in terms of position (spatial, distance), flux (peak, integrated), velocity widths (w_{20}^{min}, w_{20}^{max}, w_{50}^{min}, w_{50}^{max}), ellipticies, position angle, and mass. It was especially designed to allow a fast work flow. This is reached by the combination of modern finder algorithms, based on the Gamma test (see Appendix 3.7.1; Evans 2002; Boyce 2003), and sophisticated methods for a manual search, e.g. the simultaneous view of all three projections of the data cube (Ra–Dec, Ra–Velo, and Dec–Velo).

Fig. 3.41 (top panels) shows exemplarily the Ra–Dec and Dec–Velo projected plane. By inspecting the whole data cubeone will find brighter galaxies as spots in the Ra/Dec projections or as "sausages" in the two Velo views. Of course also artifacts could mimic such a shape. It is very rarely the case, that interferences and other artificial sources look very much like a galaxy. It is indeed true that fainter galaxies appear often similar to those image artifacts. In that sense the experience of the astronomer is needed to classify the candidate (the GaPa provides several tools to assist this task).

The search for galaxies has not to be restricted to the data domain. The GaPa provides two further representations, the Gamma cube and the X-ray cube. The Gamma cube contains for each pixel the Gamma value as calculated by the Gamma test; see Section 3.7.1. To keep the computational effort as low as possible the Gamma values are determined locally using a subset of the data cube around the current pixel. The construction of the Gamma cube still takes several minutes. Therefore, the user can save it to a fits file for future usage. Fig. 3.41 (middle panels) shows two planes of the Gamma cube which

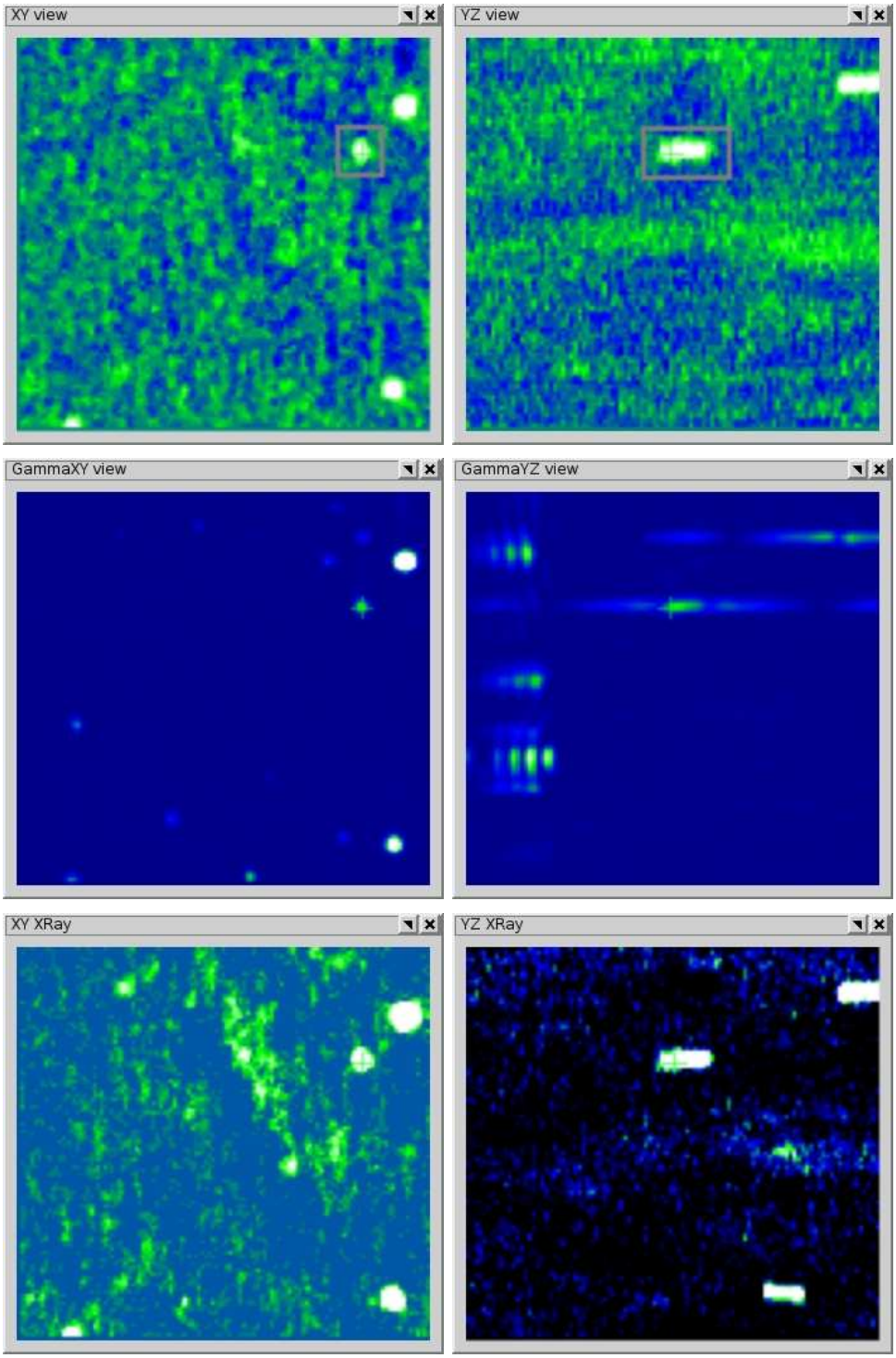


Fig. 3.41: Galaxy Parametrizer: Data (top), Gamma (middle), and X-ray (bottom) cube projections of the HIPASS data cube H364. Exemplarily, the x–y (left panels) and y–z (right panels) projections are shown. In the middle right panel image artifacts, ripples, are visible, caused by Milky Way emission.

corresponds to the planes of the data cube (top panels). Obviously the source emission is greatly enhanced making the Gamma values a supreme tool for galaxy finding. Note, that compared to Boyce (2003) GaPa also explicitly allows to calculate the Gamma values using three dimensions.

Another useful representation is the X-rayed data cube; see Fig. 3.41 (bottom panels). The underlying mathematical operation was introduced in Section 3.7.2. Compared to the Gamma cube the sources are less enhanced, but are still much more prominent than in the data view. The advantage is that the X-ray cube can be computed much faster than the Gamma cube. The X-ray view also works in real-time without the need to compute the complete cube beforehand, though the X-ray cube is still needed if an automatic finder (based on the X-ray cube) shall be used. Note, that the source enhancement is simply due to the fact, that the "X-rays" are sensitive to the integrated flux of the source while the data view only shows the intensities of the currently viewed plane.

After defining a subcube (manually or automatically) completely containing the source emission, the spatial profile fitting in the mean intensity (Moment 0) map of the source can be started. The function

$$f(x,y) = A\exp\left[-\frac{r^2}{2w^2}\right] + b, \qquad r = \sqrt{x^2+y^2} \tag{3.52}$$

is a two-dimensional Gaussian of width w, peak intensity A, and offset b. It does, however, not account for any ellipticity a galaxy might have. Therefore, a two-dimensional elliptical Gaussian profile is introduced based on the transformation

$$w \to w' = \frac{w_b}{\sqrt{1 - w_\varepsilon^2 \cos^2(\varphi - \alpha)}}\,, \tag{3.53}$$

which parametrizes any elliptical profile using the width of the minor axis, w_b, the eccentricity, w_ε, and position angle, α. The use of one axis and the eccentricity is equivalent to using both axes. They are related via $\varepsilon^2 a^2 = a^2 - b^2$. Finally, in addition to the radius r in Eq. (3.52) the polar angle $\varphi = \arctan(y/x)$ is needed for the general functional dependence. This leads to the function of a *two-dimensional elliptical Gaussian*

$$f(r,\varphi) = A\exp\left[-r^2\left(\frac{1 - w_\varepsilon^2\cos^2(\varphi-\alpha)}{2w_b^2}\right)\right] + b\,, \tag{3.54}$$

which is used for (spatial) Galaxy fitting. This function does not represent real galaxies in principle, but provides a useful approximative description — especially for non-resolved sources.

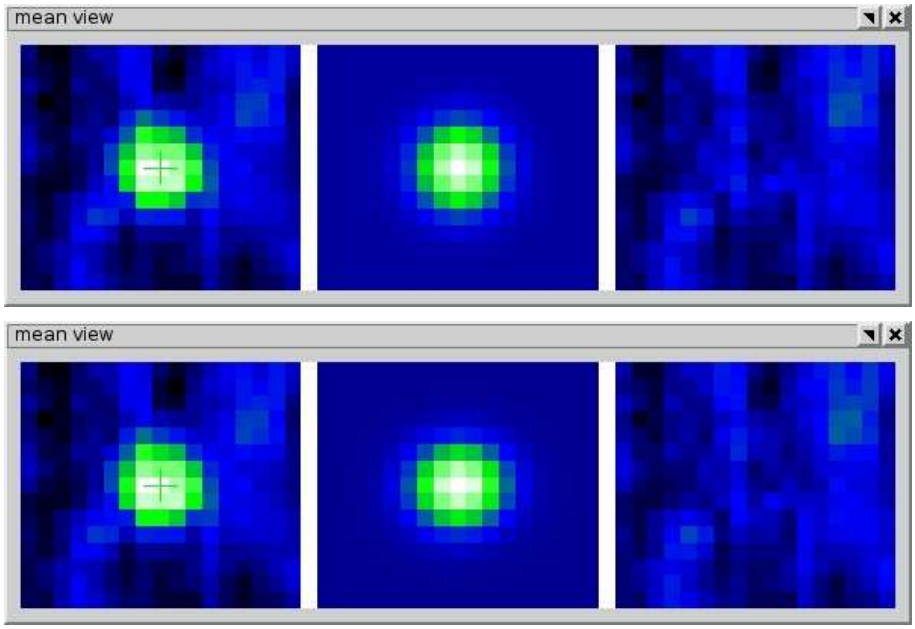


Fig. 3.42: Galaxy Parametrizer: Mean intensity map of a source. In the top panel the analytic Levenberg-Marquardt algorithm was applied, while the bottom panel was fitted based on a *simulated annealing* solver. The middle column contains the actual fit, the right column the residual.

Two different fit methods were implemented: (1) a "standard" algorithm using the derivative solver[14] `gsl_multifit_fdfsolver_lmsder` of the GNU scientific library (GSL)[15] which is based on the Levenberg-Marquardt (LM) algorithm (Fig. 3.42, top panel Levenberg 1944; Marquardt 1963) and (2) a *simulated annealing* (SM) solver (bottom panel Kirkpatrick et al. 1983; Cerny 1985) also from the GSL. Method (1) worked fine during extensive testing but the LM algorithm (as all analytic solver) is known to often underes-

[14]Derivative solvers are faster but the Jacobian of the function with respect to the fit parameters must be known. The partial derivatives of $f(r,\varphi) = A\,E + b$ are

$$\begin{aligned}
\frac{\partial f}{\partial A} &= E \\
\frac{\partial f}{\partial b} &= 1 \\
\frac{\partial f}{\partial w_\varepsilon} &= A\,E\,\frac{2r^2 w_\varepsilon \cos^2(\varphi-\alpha)}{2w_b^2} \\
\frac{\partial f}{\partial w_b} &= A\,E\,\frac{2r^2\left(1-w_\varepsilon^2\cos^2(\varphi-\alpha)\right)}{2w_b^3} \\
\frac{\partial f}{\partial \alpha} &= A\,E\frac{r^2 w_\varepsilon^2}{2w_b^2} 2\cos(\varphi+\alpha)\sin(\varphi-\alpha) \\
E &\equiv \exp\left[-r^2\left(\frac{1-w_\varepsilon^2\cos^2(\varphi-\alpha)}{2w_b^2}\right)\right]
\end{aligned}$$

Note, that some denominators can become zero, but using L'Hospital's rule leads to defined results. It is, therefore, necessary to check for too small values during the computation.

[15] `http://www.gnu.org/software/gsl/`

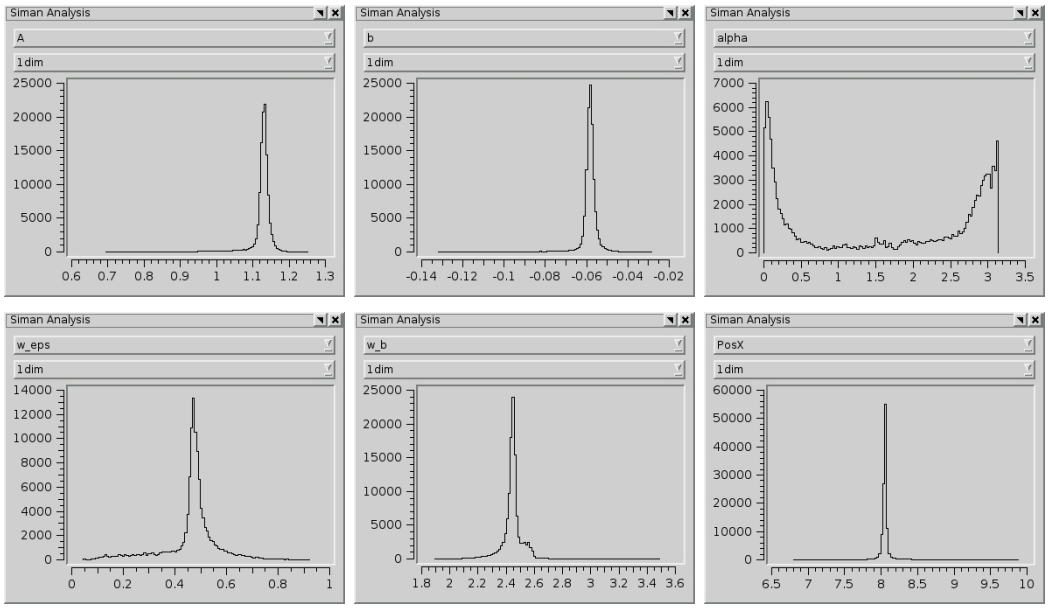


Fig. 3.43: Galaxy Parametrizer: Accurate determination of statistical errors is possible using a sophisticated Markov-Chain Monte-Carlo (MCMC) method. The plot for the y-position is omitted as it is qualitatively indifferent from the x-position plot.

timate the statistical errors of the fit (which are estimated from the correlation matrix). This is the reason that also a Monte-Carlo (MC) procedure, the SM method, is provided. One can use a Markov-Chain Monte-Carlo (MCMC) method to investigate the parameter space in detail. This ensures the estimated statistical fit errors to be more accurate. The main disadvantage is the higher computational demand — not for finding the best fit, but to run the fit several hundred times to collect enough MCMC paths for the error estimation. Hence, for simple determination of the best fit the analytic method should be preferred.

It is now described, how the fit errors are estimated using the MCMC data. Starting from the best fit solution, random starting points (in parameter space) are generated around the best parameter vector. For each of these starting points, the simulated annealing method is performed, while the current parameter vector during each iterations is saved, building-up "chains". All chains are used to generate histograms for each parameter; see Fig. 3.43. If the problem is well defined (and the noise is not to large) clustering around the best fit parameters is expected. The density and morphology of this parameter cloud defines the statistical errors of the fit. If the cluster is very sharply peaked the errors are small and vice versa. Starting from the highest peak in the histogram (which has not necessarily to be the best fit, though it is in most cases) one goes to the left and right until 68% (1σ) of all values are enclosed. The left and right marker positions determine the errors on the specific parameter, which can have different lower and upper values.[16]

[16]For many galaxies, which are only mildly elliptical the histogram for α and w_ε is often not very sharply peaked, due to the near-degeneracy in these two parameters. The user can activate the "bias ellipticities" check-box which results in a 75% chance to take a step towards a higher w_ε compared to a 25% chance

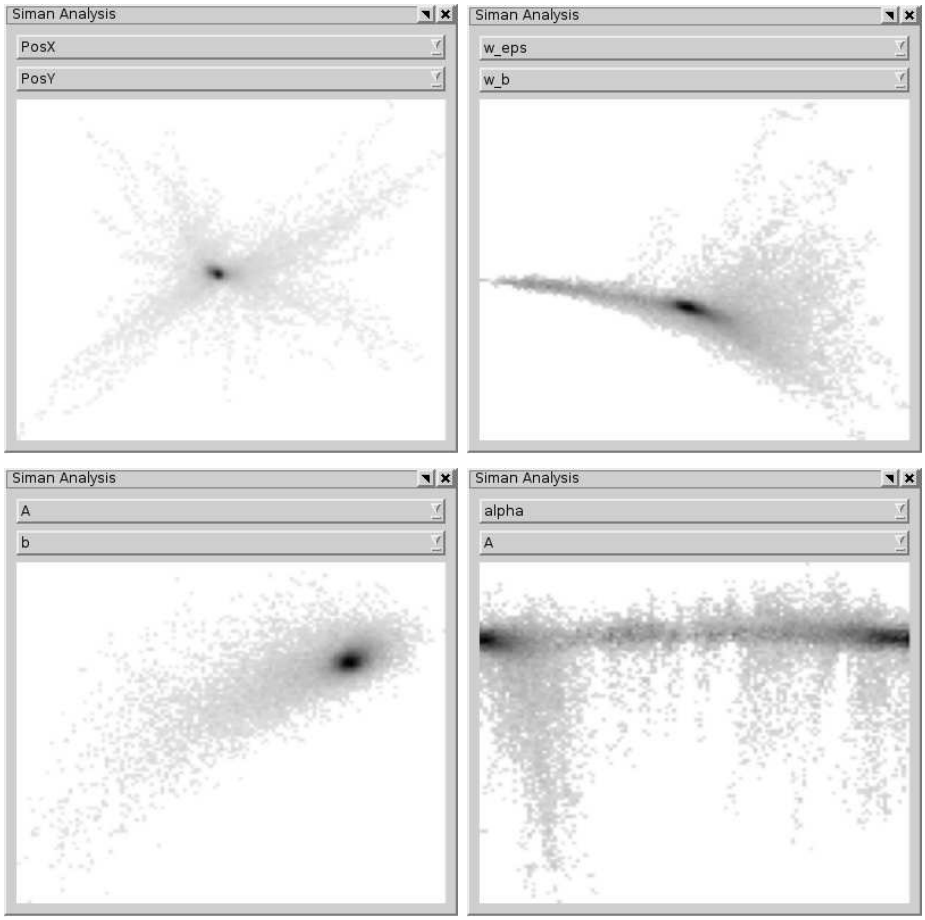


Fig. 3.44: Galaxy Parametrizer: Cross-correlations of the fit parameters as obtained by computing MCMC paths. Exemplary, the following correlations are shown, positional (x–y, top left), w_ε–w_b (top right), A–b (bottom left), and α–A (bottom right).

For completeness the user can also inspect cross-correlations of different parameters. They are plotted as density maps from two-dimensional histograms; see Fig. 3.44. Note, that the visual inspection of the histograms is only possible for the galaxy currently parametrized, because the MCMC data are not held in memory for performance reasons. The correlation plots are actually not used to infer any quantitative values, but are thought to aid the manual inspection of the data, to check the quality of a certain fit.

If the fit was successful GaPa computes the total flux from the mean intensity of the subcube. It does this by integrating the fit function for the best fit parameters. One could of course sum up the intensities of the mean map, but this would in some cases be more error-prone, e.g., when another nearby galaxy disturbs the mean map or in case of baseline residual effects. Nevertheless, calculating the flux from the fit function is not easy, due to the fact, that the rather complex fit function leads to an elliptical integral

to decrease w_ε during the simulated annealing random walk. If there is a slight ellipticity this increases the probability to find it during the solution. Still, both errors can be quite large.

	ra	dec	velo	sub	pf	if	ifs	wo	po	so	comments
1	189.83	-0.53	1141	102, ...	1120	214.98	253.49	1	1	1	finder,
2	190.08	-5.78	1101	99, ...	391	71.05	-57.66	1	1	1	finder,
3	190.18	-5.14	2645	98, ...	136	38.11	-112.15	1	1	1	finder,
4	190.65	-1.34	1062	90, ...	200	41.52	-137.25	1	1	1	finder,
5	190.64	-1.33	1101	91, ...	200	37.25	45.75	1	1	4	user,
6	190.65	-1.34	1101	91, ...	200	39.80	-118.24	1	1	1	user,
7	190.64	-0.06	1761	89, ...	344	64.84	-80.63	1	1	1	finder,
8	190.78	-1.23	3304	90, ...	119	28.85	127.16	1	1	1	finder,
9	190.93	-0.57	2684	86, ...	167	35.61	-166.36	1	1	1	finder,
10	191.03	-5.70	1471	86, ...	360	42.05	-98.06	1	1	1	finder,
11	191.15	-2.33	1629	83, ...	120	19.52	116.68	1	1	1	finder,
12	191.28	-0.49	1682	80, ...	344	74.35	-224.84	1	1	1	finder,
13	191.45	-6.06	1405	78, ...	360	20.85	38.18	1	1	1	finder,
14	192.44	-4.57	1418	64, ...	133	13.00	-38.43	1	1	1	finder,
15	192.69	-4.13	1510	60, ...	136	15.74	-50.35	1	1	1	finder,
16	192.76	-6.37	1418	62, ...	631	116.27	135.92	1	1	1	finder,
17	193.77	0.16	1246	43, ...	164	29.66	97.90	1	1	1	finder,
18	194.32	-4.14	1563	35, ...	166	21.55	-12.61	1	1	1	finder,
19	194.33	-5.33	1194	36, ...	265	42.17	52.98	1	1	1	finder,
20	196.12	-3.57	1392	7, ...	426	36.45	52.74	1	1	1	finder,

Fig. 3.45: Galaxy Parametrizer: List of detected galaxies and their parameters as found by the various procedures described in the text.

which can not be solved analytically. A numerical integration procedure (Monte Carlo MISER algorithm) from the GSL is used. Also, in cases were the EBHIS beam resolves the source and the fit function is inappropriate, the flux has to be calculated by directly summing up.

Usually the Monte-Carlo integrator returns an estimated error, though being valid for the best fit function only. As described in the previous paragraph a lot of effort was invested to precisely compute the parameter uncertainties of the fit. Obviously the integrated flux heavily depends on the actual fit parameters, so it would be meaningless to neglect these errors while calculating the errors of the flux. GaPa will estimate the errors of the flux more accurately by enabling the check-box "compute flux errors". Then the integration is repeated 1000 times, each time using a different parameter vector randomly drawn (Gaussian distributed) from the 1σ-interval. A good measure for the flux is the mean (or median) of all individual results, while the error is reflected by their standard deviation. As before, the drawback of the precise error calculation is the much higher time complexity. Note, that all of the above calculations are independent of the projection system in use (compare Section 3.6), as from the pixel coordinates the associated world coordinates are derived in which then the fit is applied.

After the fitting process the source appears in the list of galaxy candidates; see Fig. 3.45. The table shows positions, size of the subcubes, and fit parameters. Each galaxy/candidate appears also in the views of the data cubes (the sub-cube is marked using rectangles). The galaxy catalog and all parameters can be saved to an XML file (Extended Markup Language[17]). XML is today widely used to store a wide range of

[17] `http://www.w3.org/TR/2006/REC-xml-20060816/`

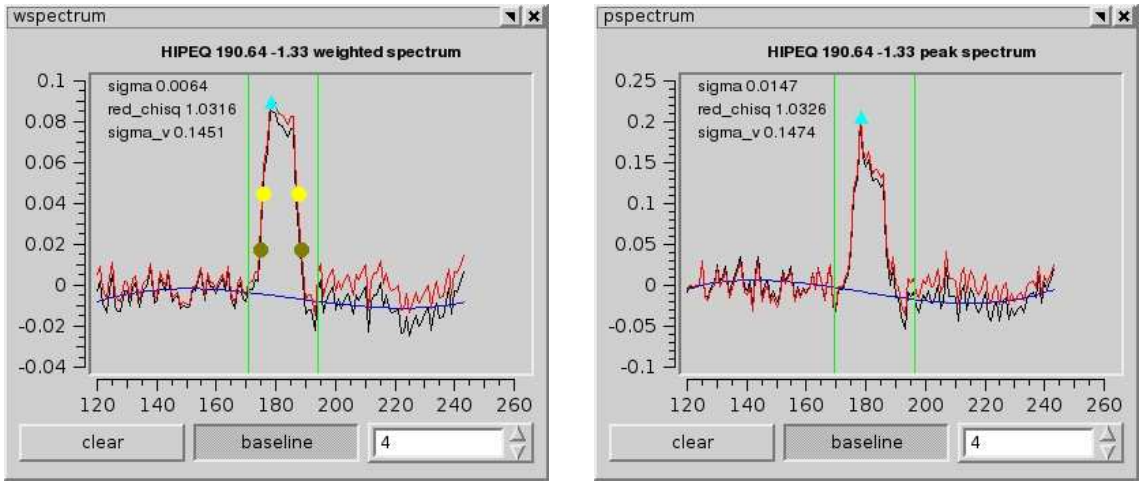


Fig. 3.46: Galaxy Parametrizer: Weighted (left panel) and peak (right panel) spectrum of a parametrized galaxy. Various profile parameters are automatically determined. If necessary the spectra can be corrected for baseline effects. Each of the spectral views shows also the RMS level of the data and measures regarding the fit quality of the baseline, e.g., the reduced χ^2-value[19].

(strongly) structured data. If the user provides an XSL style sheet the catalog can be directly viewed by a variety of web browsers.

Each detected galaxy can be further inspected using several spectral views. Most important in terms of parametrization is the weighted spectrum, which is used to determine the velocity widths of the profile; see Fig. 3.46 (left panel). It is calculated by computing the mean spectrum of the subcube using the function values of the elliptical Gaussian fit as weighting factors. This has the effect, that the asymmetries of the profile (due to a rotation of the galaxy opposite line of sights may have different mean velocities and profile shapes) are better taken into account. Following Meyer et al. (2004), 20% and 50% minimum/maximum widths $w_{20,50}^{\mathrm{min,max}}$, as well as mean velocities $v_{20,50}^{\mathrm{min,max}}$ are computed. These are the ranges at which the maximum intensity has dropped to 20%/50%. In many cases the minimum/maximum values are equal. For double horned profiles it may happen that there are two points at which the intensity reaches 20%/50%. In those case the larger width determines the maximum value. Another spectral view shows the peak spectrum. That is the single spectrum of the galaxy which contains the pixel with the highest intensity; see Fig. 3.46 (right panel). Note, that this parametrization scheme suffers from bias effects for faint sources. This issue is discussed in detail in Section 4.5. A third spectral view is the summed spectrum (not shown here). It is not used to find any parameters although it can be used for an alternative measurement of the integrated flux. This is — at least for the HIPASS data cubes — not the preferred method due to baseline deficiencies which often lead to unreliable results.

[19] The reduced χ^2 is the variance, $V = \chi^2/(n-q)$, with n being the number of data points and q the number of independent fit parameters. Hence, $n-q$ are the degrees of freedom of the fit. V can be used as indicator for the goodness-of-fit, as for the ideal model (and $n \gg 1$) its mean value should be within $1 \pm \sigma_v$, with $\sigma_v = \sqrt{2/(n-q)}$ (see Sormann 2008).

Finally, the built-in automatic galaxy finder algorithm shall be discussed. It is implemented in a very general way to be used with both the Gamma cube, as well as the X-ray cube. The finder is invoked by clicking on the associated entry in the "Finder" menu. First, it robustly calculates the mean and variance (noise) levels of the Gamma/X-ray cube. In the next step all pixels which have a value of $\geq x_{\mathrm{trigger}}\sigma_{\Gamma,X}$ above the mean value are appended to the so-called peak list, which contains coordinates and intensities (not the Gamma values but values from the data cube) of all pixels potentially containing a source. Of course in general a single source will have got more than one of the peaks attributed.

It is computationally efficient to start with those peaks having highest intensities, therefore the peak list is sorted. A very complex task is now a meaningful definition the subcube containing *all* flux which can be attributed to a galaxy candidate. For the HICAT a standard size was used for the spatial extend and the velocity range was interactively defined by the user. It is not desirable to apply a standard subcube size for the velocity extend, because the possible range of profile widths is quite high (from 50 $\mathrm{km\,s^{-1}}$ to $\gtrsim$ 500 $\mathrm{km\,s^{-1}}$). In a first approach the subcube was defined in a way that all pixels were enclosed having intensities above a certain threshold (e.g. $5\sigma_{\mathrm{rms}}$) and are physically connected to the actual pixel. It turned out, however, that especially fainter galaxies could not be found. The same is true for sources having moderate integrated flux but low peak flux (i.e., having very broad velocity profiles).

Consequently, a lot of effort went into the development of an adaptive scheme to work out for each galaxy an optimal size of the subcube. During extensive testing it was verified that in almost all cases a good solution was calculated — even for faint sources, which are obscured by noise and only detectable by their integrated flux (mean intensity of the source is significantly higher than the RMS of the mean intensity of the surrounding). The latter, obviously, is only possible if the candidate was incorporated into the peak list. In that sense the Gamma test performed good — all galaxies are visibly traced by the Gamma values. How many sources are not incorporated into the candidate list is only a matter of the threshold level, which also determines the number of "false positives" (i.e., noise peaks or image artifacts) in the list.

The adaptive procedure to define the subcube size shall be discussed in the following. Drawing consecutively values from the peak list, it can be assumed that the current peak is always the highest one which was not yet associated to a galaxy (as peaks which are nearby a galaxy already processed are dropped). Therefore, one can treat it as spatial center of a new subcube containing the candidate. Some quantities need to be defined. Let $w_{\mathrm{min/max}}$ be the minimum/maximum velocity width allowed for a galaxy. Within the subcube the size of the source is denoted as the core radius, r_{core}, while the surrounding of the emission is parametrized by a minimum/maximum radius, $r_{\mathrm{surr}}^{\mathrm{min}}$ and $r_{\mathrm{surr}}^{\mathrm{max}}$. These radii define two areas, which can be used to compute the significance of a source — the total flux within the core, I_{core}, related to the noise level of the surrounding.

The algorithm now iterates over: (1) all possible values of w, and (2) all possible lower boundaries (although both velocity boundaries, $v_{\mathrm{low,upp}}$, should at least have a distance of 25% of w_{max} from the starting peak). During this iteration the "weighted mean flux" of the core

$$I_{\mathrm{core}}^{\mathrm{weighted}} = I_{\mathrm{mean}}\sqrt{v_{\mathrm{upp}} - v_{\mathrm{low}}} \tag{3.55}$$

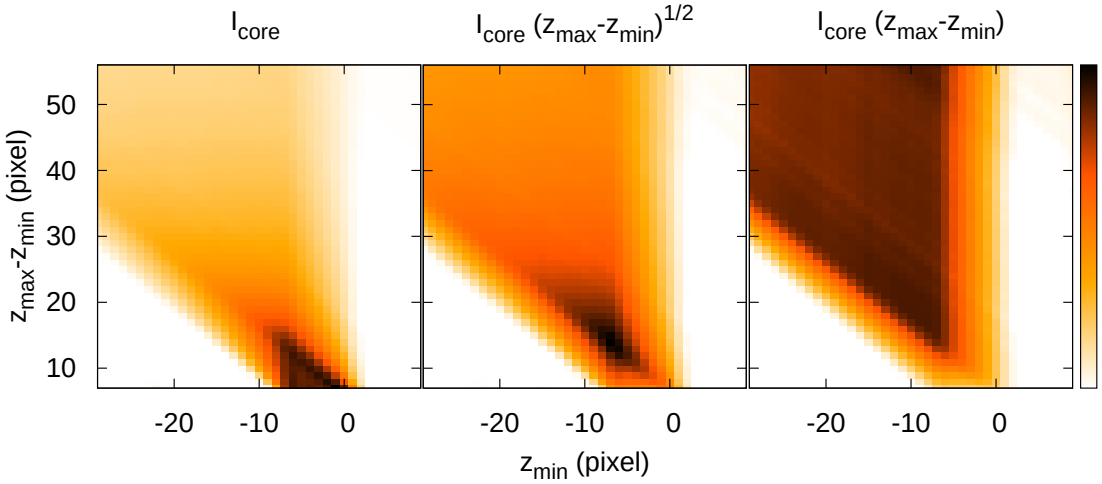


Fig. 3.47: For a proper determination of the subcube size containing a source, GaPa utilizes an effective "brightness" of the source (the "core"). The plot shows the normalized core intensities I_{core} — weighted with unity (left panel), $\sqrt{v_{\mathrm{upp}} - v_{\mathrm{low}}}$ (middle panel), and $v_{\mathrm{upp}} - v_{\mathrm{low}}$ (right panel) — as a function of the size (velocity range) of the subcube. The size is parametrized as lower boundary z-value relative to the starting pixel and the z-width of the subcube (in pixels). If weighted with the square root of the width the highest $I_{\mathrm{core}}^{\mathrm{weighted}}$ is neither concentrated to an extreme value (as with unity weighting) nor wide-spread over a large parameter space (as in the linear case). The distribution is well concentrated avoiding large scatter of the calculated subcube size around its optimal value. For visualization purposes the boundaries were changed within a range larger than usual.

is computed. Note, that because I_{core} is the mean intensity of core pixels of the current subcube changes with the velocity during the iteration. $I_{\mathrm{core}}^{\mathrm{weighted}}$ will have a maximum when the subcube has optimal size (and placement).

The use of the weighting width the square root of the subcube width is necessary. Assume one would use I_{core} as tracer. This quantity would easily be maximized, if the width $(v_{\mathrm{upp}} - v_{\mathrm{low}})$ is minimal, because the innermost values are highest. Therefore, the current subcube width has to be taken into account somehow. The simplest possibility would be weighting with the width, i.e., compute $I_{\mathrm{core}}(v_{\mathrm{upp}} - v_{\mathrm{low}})$. This is also not meaningful, as it would become larger with increasing width even if no additional flux is enclosed by the growing subcube. Using the square root of the width as weighting factor provides a good compromise for typical profile shapes and has proven to be well suited during tests; see also Fig. 3.47.

The performance of the finder algorithm in terms of detection probability is analyzed in Section 4 in detail. Here, one of the HIPASS data cubes (H364) was inspected manually for sources and then compared to the results of an automatic galaxy finder algorithm using the Gamma cube. The assiduous visual inspection of the data cube produced only three more detections than the automatic approach. One of them was previously masked by another galaxy, while the remaining two galaxies had extremely low intensities. More importantly, only minor overhead ("false positives") of about 25% was produced. It would

be very complicated to improve the rejection of false detections. Those which are still left in the catalog have parameters very similar to real galaxies and might only be distinguished by very sophisticated shape/pattern matching. As a manual verification is recommended anyway the manual rejection of the few false positives would not be too time-consuming. Note also, that the number of galaxies in this data cube was very high, which can introduce confusion.

CHAPTER 4

Preparing the survey — Simulations

The aim of this chapter is to analyze the performance of the data reduction software in terms of source detection probability and data quality. Furthermore, the influence of RFI signals on the results of an automatically compiled galaxy catalog is investigated. An empirical approach, i.e., running simulations, was chosen, providing the possibility of a quantitative analysis and knowledge about typical processing times. Also, the free parameters of the detection algorithm could be tested.

While the details are discussed in subsequent sections, the basic scheme shall be briefly introduced. The simulations comprise the generation of (artificial) spectra, containing emission lines (Gaussian shaped), noise, and eventually narrow-band interference signals with time-varying amplitude. An RFI detection scheme (see Section 3.2) was applied resulting in a *flagged database*. The data were gridded (see Section 3.6). Three different sets of data cubes were produced, one set without interferences, the remaining two including RFI, but one with and one without RFI mitigation. The source detection was completely automatted using a special (scripted) version of GaPa (see Section 3.7). The obtained results are directly compared to the simulated galaxy properties, leading to empirical measures of the data reduction quality and quantity.

4.1 Sampling of galaxy properties based on real data

In order to maximize the explanatory power of the results the shape and flux of the sources, i.e. the velocity profile widths (in spatial dimension galaxies can be considered point-like), were based on real data. I had access to the source catalog of the Northern HIPASS Extension which was analyzed by Garcia-Appadoo (2005). From this catalog it was possible to sample line profiles subject to the observed statistical properties. Some technical key parameters as beam-size of the telescope and spectral resolution were adjusted to match those of the Parkes telescope. This ensures the comparability of the simulated and reconstructed source parameters with the HIPASS catalog and derived properties. Some simplifications had to be made. First, only Gaussian velocity profiles were used, and, second, the spatial grid (i.e., the "observed" positions) was equally distributed in Carte-

sian coordinates (right ascension, declination). As the data cube were also gridded on a Cartesian grid, the noise level does not show any dependence on declination. On the other hand, the source positions were sampled such that they lie uniformly on a sphere. Hence, in the data cubes the distribution is declination-dependent, meaning lower source density at high elevations. This effect is not a drawback for the statistical analyses, as the correlation between sources is not investigated. Note, that not the full sky was simulated but only for Declination angles of $-60° \ldots 60°$, saving some computational effort (neglecting high-latitude data cubes being poor of sources).

The use of Gaussian profile shapes only has certainly an impact on the results. In reality box-shaped and double-horn profiles are observed, as well. The former has a relatively lower peak flux compared to its integrated flux than a Gauss curve. However, the source finder is mostly sensitive to the "unsmoothness" of a profile. In this respect, box profiles are not expected to reveal much lower detection rates. The same is true for double-horns being a mixture of both cases.

4.1.1 The H I mass function (HIMF)

One of the key properties of the local H I universe is the so-called H I Mass Function (HIMF) ϕ, the number density of galaxies per mass bin and unit volume. For a certain survey (with specific detection limits) it can be expressed as

$$\mathrm{d}N(M_{\mathrm{HI}}) = \phi(M_{\mathrm{HI}})V(M_{\mathrm{HI}})\mathrm{d}M_{\mathrm{HI}} \tag{4.1}$$

with the number of detected sources, $\mathrm{d}N(M_{\mathrm{HI}})$, per mass interval $[M_{\mathrm{HI}}, M_{\mathrm{HI}} + \mathrm{d}M_{\mathrm{HI}}]$. $V(M_{\mathrm{HI}})$ is that volume out to which a galaxy of mass M_{HI} would be observable. The HIMF was determined by several authors, with the best accuracy today is provided by the HIPASS survey (Zwaan et al. 2005), who measured the HIMF within the mass range $M_{\mathrm{HI}} = 10^7 \ldots 10^{11}\,\mathrm{M}_\odot$. However, due to the low number of sources below $M_{\mathrm{HI}} = 10^8\,\mathrm{M}_\odot$ the results for the lower mass-part of the HIMF are not very well constrained, one of the main goals for future surveys will be to investigate the low-mass end of the local HIMF down to $M_{\mathrm{HI}} \lesssim 10^7\,\mathrm{M}_\odot$ which would include dwarf galaxies. The HIMF can be fitted using a Schechter function (Schechter 1976)

$$\phi(M)\mathrm{d}M = \theta^* \left(\frac{M}{M^*}\right)^{\alpha} \exp\left(\frac{M}{M^*}\right) \mathrm{d}\left(\frac{M}{M^*}\right) \tag{4.2}$$

with α being the slope in the lower mass regime, M^* is the mass of the "knee", while θ^* is a normalization factor (results from the HIPASS are: $\alpha = -1.37 \pm 0.03$, $\log(M^*_{\mathrm{HI}}/\mathrm{M}_\odot) = 9.80 \pm 0.03\,h_{75}^{-2}$, and $\theta^* = (6.0 \pm 0.8) \cdot 10^{-3}\,h_{75}^3\,\mathrm{Mpc}^{-3}\,\mathrm{dex}^{-1}$; see Zwaan et al. 2005). Integrating the Schechter function returns the total H I mass content per unit volume in the local universe $\Omega_{\mathrm{HI}} = (3.5 \pm 0.4 \pm 0.4) \cdot 10^{-4}\,h_{75}^{-1}$ (with different errorbars accounting for distinct systematic effects; details are discussed in Zwaan et al. 2005).

To infer from the observed emission line profiles the mass M_{HI} is not directly possible, but there is a correlation between mass and integrated flux, S_{int}, (Roberts 1962)

$$\begin{aligned} M_{\mathrm{HI}}\,[\mathrm{M}_\odot] &= 2.356 \cdot 10^5\,(D\,[\mathrm{Mpc}])^2\ S_{\mathrm{int}}\left[\mathrm{Jy\ km\,s^{-1}}\right] \\ &= 2.356 \cdot 10^5\,(D\,[\mathrm{Mpc}])^2 \int_{\mathrm{source}} \mathrm{d}\Omega\,\mathrm{d}v\ \left(S\left[\mathrm{Jy\,Beam^{-1}}\right]\right) \end{aligned} \tag{4.3}$$

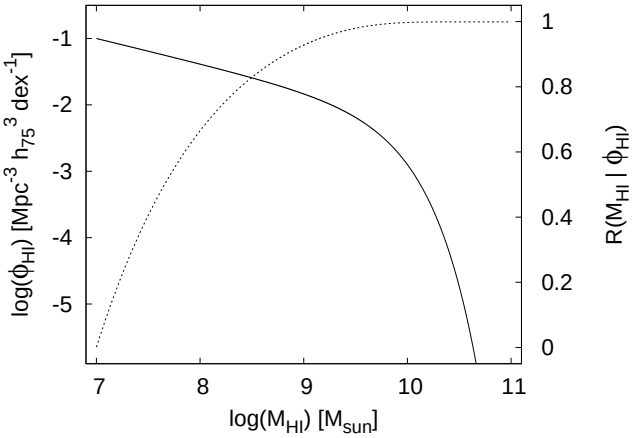


Fig. 4.1: The HIMF, ϕ_{HI} (solid line; with parameters as described in the text), and its cumulative density distribution, $R(M_{\mathrm{HI}} \,|\, \phi_{\mathrm{HI}})$ (dotted line), used for the simulations. Inserting uniformly distributed random numbers y from $[0, 1]$ into the inverse function $R^{-1}(y)$ provides samples which follow the desired density function ϕ_{HI}.

which depends on the distance D of a galaxy. S is the flux density of the source. The relation holds only for the optically thin case. The distance D can be calculated from redshift z or, equivalently, radial velocity of the source

$$D = \frac{v}{H_0} \tag{4.4}$$

with H_0 being the Hubble constant. Here, it is assumed that $H_0 = h_{75} = 75\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Eq. (4.3) disregards any peculiar motion of the observed galaxies, which is statistically correct for survey volumes large enough, but not for local regions of over- or underdensity (e.g., the Virgo cluster).

The HIMF ϕ can be used to sample masses according to the observed distribution. In order to do that a function is needed which maps the output of a basic uniform random number generator to the desired distribution. For arbitrary density distributions $r(x)$ this can be achieved via the cumulative density

$$R(x) = \int_0^x \mathrm{d}x'\, r(x') \tag{4.5}$$

which has the important property $R(x) \in [0, 1]$, because $r(x)$ — being a density distribution — is normalized by definition. Now, sampling uniformly distributed random numbers y from the interval $[0, 1]$ one just needs to insert them into the inverse function $R^{-1}(y)$ to get random numbers which follow the density distribution r. Both the HIMF and the cumulative density are plotted in Fig. 4.1 using parameters as found by Zwaan et al. (2005). The observed volume (a sphere with radius according to the maximum redshift possible according to the bandwidth of the receiving system) can now be filled with galaxies of different masses until the total mass limit (according to Ω_{HI}) is reached. The mass per "unit volume" ($1\,\mathrm{Mpc}^3$) can be analytically determined using

$$\rho_{\mathrm{HI}} = \theta^* \Gamma(2 + \alpha) M^*_{\mathrm{HI}} = 4.56 \cdot 10^7\,\mathrm{M}_\odot \tag{4.6}$$

(Zwaan et al. 2003). For the simulations only masses in the range between $\log(M_{\mathrm{HI}}) = 7 \ldots 11$ were sampled, the total number of sources generated might be slightly to large, which is, nevertheless, not of interest in this work. Note, that the generated number of sources will strongly differ from what can be observed. Due to the detection limits of the observation a source can only be seen up to a certain distance, according to its flux.

4.1.2 Detection limits

If all generated sources would be kept in the catalog, the computational effort would grow dramatically. First, for each position a huge list would have be searched for associated sources, second, even for faintest sources (which would never be detectable at a certain noise level) a Gaussian would need to be added, and third, after the application of the automatic galaxy finder, the detections must be assigned to generated sources. The first and third task have computational complexity of order $O(n^2)$. To avoid a huge number of (non-detectable) low-mass galaxies at larger distances, in this paragraph the expected detection limits are estimated. All generated sources below these limits were not added to the initial galaxy catalog.

After assigning to each galaxy a (3-dimensional) position in order to compute its spatial coordinates and distance, the linewidth, Δv_{50}, of the velocity profile must be determined. As mentioned in the introduction a catalog including the galaxy parameters of the Northern Strip/Extension of the HIPASS was available. Hence, it was decided to simulate galaxy properties similar to what was found for this dataset. Garcia-Appadoo (2005) found the weak correlation $\Delta v_{50}[\mathrm{km\,s^{-1}}] = 0.022 \cdot M_{\mathrm{HI}}^{0.39}[\mathrm{M_\odot}]$. However, the scatter around this correlation was quite high, so a different approach was followed. Using the source catalog a two-dimensional histogram (logarithmic binning) was calculated connecting both quantities, mass and velocity width. The resulting histogram can be treated similar to a probability density distribution and makes it possible to infer the (discrete) two-dimensional cumulative distribution function, which again can be used to sample random numbers in agreement to the desired probability distribution. After sampling a mass from the HIMF, pairs $(M_{\mathrm{HI}}, \Delta v_{50})$ are drawn according to the mass–width distribution until both masses are equal. This ensures that each sampled mass (from the HIMF) is associated to a realistic velocity width. Then, all free parameters are determined. From mass and distance the value for the integrated flux, S_{int}, can be inferred (Eq. (4.3)), while the peak flux, S_{peak}, is connected to S_{int} and Δv_{50} (for Gaussians) via

$$S_{\mathrm{peak}} = \left(\frac{8\ln 2}{2\pi}\right)^{\frac{3}{2}} \frac{S_{\mathrm{int}}}{\theta_{\mathrm{b}}^2\,\Delta v_{50}}\,. \tag{4.7}$$

The factor $((8\ln 2)/(2\pi))^{\frac{3}{2}} \approx 1$ results from integrating the three-dimensional Gaussian having spatial width θ_{b}^2 and spectral width Δv_{50}. For the simulations the beamwidth of the Parkes telescope was used, $\theta_{\mathrm{b}} = 14\farcm1$. However, usually S_{peak} is usually expressed in units of $\mathrm{Jy\,Beam^{-1}}$, meaning $\theta_{\mathrm{b}} \equiv 1$. Consequently, the simplification $S_{\mathrm{int}} \approx S_{\mathrm{peak}}\,\Delta v_{50}$ holds. Note, that S_{int} is the flux integrated over both, velocity and spatial coordinates, while S_{peak} is only the value of the single pixel (or better voxel) containing the highest flux.

For most H I surveys the detection limit for a specific type of galaxy will be a rather complex (unknown) function of peak and integrated flux, velocity width and profile shape, and of course the sensitivity (noise) limit of the survey.

A pure peakflux detection limit is reasonably given by the noise level (RMS) of a data cube, $S_{\mathrm{peak}}^{\mathrm{limit}} = 5\sigma_{\mathrm{rms}}$, which is about $65\,\mathrm{mJy\,Beam^{-1}}$ if the noise level of the HIPASS shall be reproduced ($\sigma_{\mathrm{rms}} = 13\,\mathrm{mJy\,Beam^{-1}}$; see also Section 4.2). To define a limit for the total flux is more complicated. It is clear, that still the noise level of the data cube

must play a role, as well as the line width of the velocity profile. Higher Δv_{50} require a larger integration interval, increasing the absolute noise value (though the relative noise level decreases). Arbitrarily, the flux limit was chosen to be equivalent to three velocity channels having a peakflux detection, i.e.,

$$S_{\rm int}^{\rm limit} = 3\, S_{\rm peak}^{\rm limit}\, \delta v\, \frac{2\pi\, \theta_{\rm b}^2}{8 \ln 2} \sqrt{\frac{\Delta v_{50}}{\delta v}}\,. \tag{4.8}$$

To match the properties of the HIPASS data a velocity width per spectral channel of $\delta v = 13.2\,{\rm km\,s^{-1}}$ was used.

To combine both detection limits, the most simple condition would be

$$S_{\rm int} + S_{\rm peak}\Delta v_{50} > \alpha \left(S_{\rm int}^{\rm limit} + S_{\rm peak}^{\rm limit}\Delta v_{50} \right)\,. \tag{4.9}$$

where both, the peak and integrated values contribute evenly. A value of $\alpha = 0.4$ was applied to include also sources below the detection limit, which allows to study the detection properties of the simulations. Note, that for the case of Gaussian-shaped emission lines the above criterion could be merged into a detection condition only dependent on two properties (e.g., width and peakflux) instead of three parameters.

4.2 Generation of spectra

After compiling a galaxy source catalog the next step was to simulate H I spectra. In the final data cube (pixel size $4'$) an RMS level of $\sigma_{\rm rms} = 13\,{\rm mJy\,Beam^{-1}}$ was desired. In on-the-fly scanning mode the telescope is hold at fixed azimuth (and elevation) angles utilizing the rotation of the Earth. For hypothetical dump intervals of 500 ms the "mapping speed" would be $7.5''$ per second.[1] To avoid unnecessary complexity we do not simulate a multi-beam observation. A fully sampled grid is desired, therefore the offset of adjacent "stripes" should be $\sim 3'$ (compare to Section 3.6.2). One could calculate the number of spectra contributing to each pixel in the final data cube, leading to the desired noise level of the input spectra. However, the beam response function would have to be taken into account, hence, it is easier to determine the noise value experimentally. It turned out that $\sigma_{\rm rms} = 159\,{\rm mJy\,Beam^{-1}}$ was adequate.

One key aspect was the influence of RFI signals on the data reduction quality and derived physical parameters. Generating spectra, therefore, must also incorporate the simulation of RFI events. From Winkel et al. (2007) it was known, that at the 100-m telescope narrowband interferences are most common. As those signals usually remain for minutes to hours they outnumber other types of RFI in relative occurrence. Hence, it was sufficient to implement only this type of interference. For highest comparability the same clean input spectra were used for each of three runs: *run (1)* – without RFI, *run (2)* – with RFI but without mitigation, *run (3)* – with RFI and with mitigation. While generating the spectra the narrowband interferences were simulated according to a power law with exponent -1.5.

[1] For a typical beamsize of $\sim 10'$ this would be sufficient to provide enough spectra per source to allow our RFI detection algorithm to work.

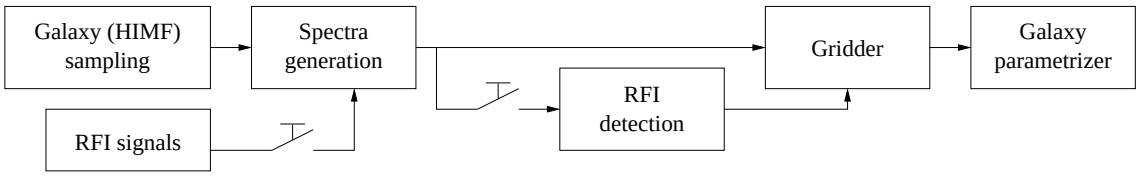


Fig. 4.2: The data reduction pipeline used for the simulations.

4.3 Data reduction pipeline

The data reduction pipeline used for the simulations is in principle not different from what was already presented in Chapter 3. For simplification, the generated spectra are assumed to be already calibrated in terms of bandpass and intensity. Furthermore, simulating the extra-galactic part of the survey, no stray-radiation must be considered. To allow a completely automatted processing, all reduction tasks can read in a "parameter file" containing desired values and states for all GUI elements (e.g., threshold levels, fitting methods, fits header entries, etc.). Fig. 4.2 shows the (reduced) pipeline schematically.

4.4 Results

In total about 800 data cubes ($\sim$ 270 per run) were simulated. The statistical analysis was performed using the programming language Python. In a first step, all "matches" had to be identified, those sources which were correctly detected by the galaxy finder in terms of position. To account for scatter, a the spatial difference had to be less than $0\overset{\circ}{.}4$, the distance deviation less than 2 Mpc. In case of confusion, i.e., more than one source was within the search radius, the closest candidate was chosen. Having the list of *matches*, the list of galaxies which were *not found* as well as those, which were wrongly identified as galaxies (*false-positives*) could be easily compiled.

Fig. 4.3 shows histograms of the peak (left panels) and total (right panels) fluxes for all runs. Each plot contains the numbers of sources per peak flux interval for the matched galaxy pairs, the false-positive events, and sources not found. The colored numbers state the size of each list. As expected, galaxies not detected lie at the fainter end of the flux scale. Except for *run (2)* this holds also true for the false-positives. The latter resemble an overhead of about 15% to 25% compared to the correctly found galaxies (again for *run (2)* the number far outreaches the matched galaxies). In *run (2)* the number of matched galaxies decreased by a factor of three due to the high number of false-positives. The plots showing the integrated fluxes are similar to those of the peak fluxes except that the false-positive detections inhibit a rather low total flux when compared to the values of the non-detections. The RFI signals during *run (2)* generate a second distribution of false-positives having very large total fluxes.

Furthermore, it is important to note, that the numbers of matched and not-found galaxies in *run (1)* and *run (3)* are comparable, meaning that the RFI mitigation worked sufficiently well. The number of undetected galaxies appears very high ($\sim$ 50%), but is due to sources below the detection limit. Obviously, in the low-flux regime the number of non-detections is larger than the number of matched galaxies, while for higher fluxes the "survey" gets more and more complete. Generating sources below the actual detection

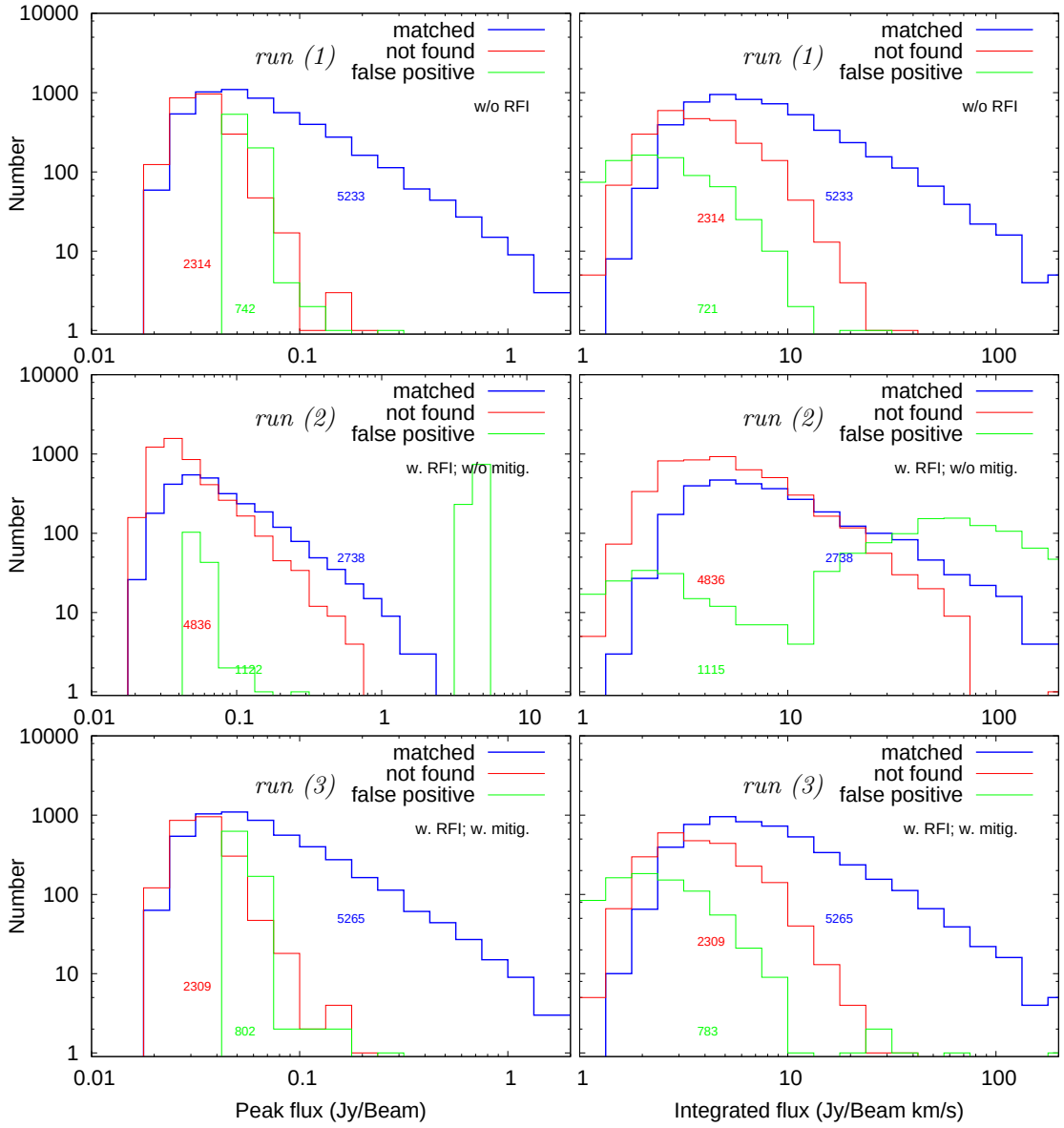


Fig. 4.3: Histograms of peak (left panels) and total (right panels) fluxes for all runs. Each plot shows the number of sources per peak flux interval for the *matched* galaxy pairs, the *false-positive* events, and sources which are *not detected*. The colored numbers mark the total size of each list. The plots showing the integrated fluxes are similar to those of the peak fluxes except that the false-positive detections inhibit a rather low total flux when compared to the values of the non-detections. The RFI signals during *run (2)* generate a second distribution of false-positives having very large total fluxes.

limit, as well, is useful to reveal the selection function. Such a procedure was used by Zwaan et al. (2003) and Zwaan et al. (2005) to compute the completeness function for the

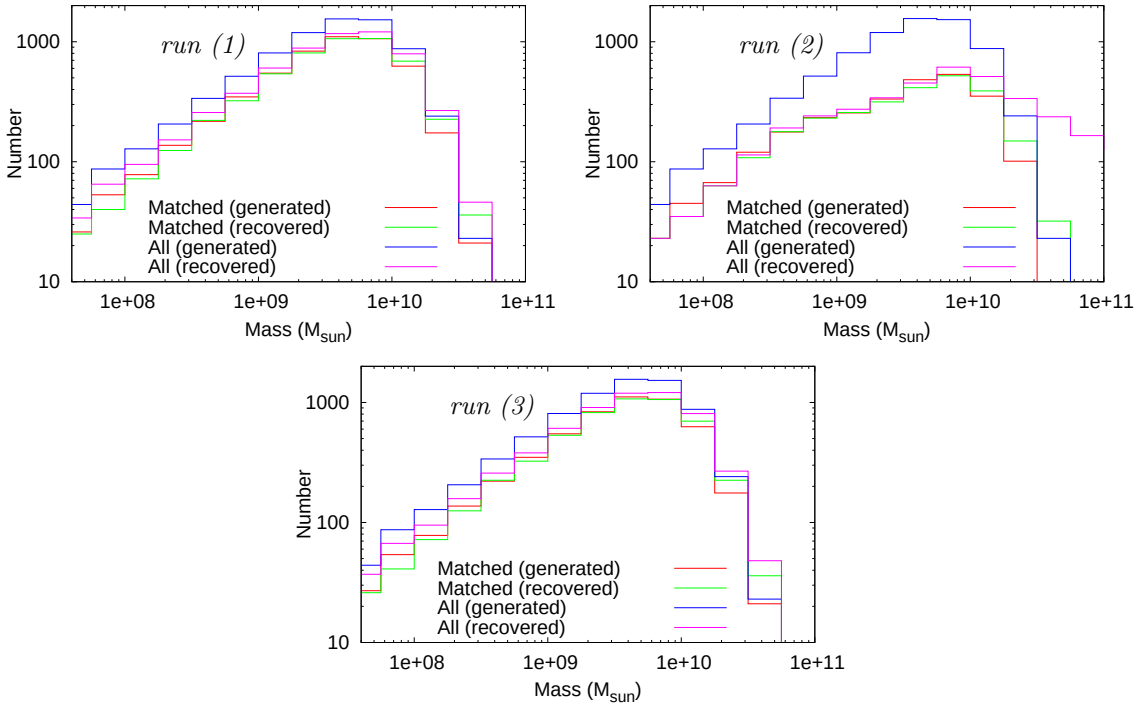


Fig. 4.4: Number of detections per mass interval for all runs. Each plot shows the numbers for all generated and recovered sources, as well as the numbers of all matched pairs. In the latter case the masses match quite well except for the faint end. Again *run (2)* makes the exception.

HIPASS survey. Note, that the simulations are not detailed enough to be used for the EBHIS in this respect, as only Gaussian profile shapes were used.

The number of detections per mass interval for all runs is shown in Fig. 4.4. Each plot shows the numbers for all generated and recovered sources, as well as the numbers of all matched pairs. In the latter case the number of generated galaxies equals the number of recovered galaxies (by definition) but their individual mass might be different. It turns out that the masses match quite well except for the faint end part, the reasons for which are discussed in Section 4.5. The number of all generated galaxies is larger than the number of all recovered galaxies (due to the non-detection of faint sources). Again *run (2)* makes an exception. The high fluxes of the false-positives cause an excess of the total mass to the high-mass end of the distribution and as the overall number of correct detections is low, the recovered total mass is significantly lower.

In Fig. 4.5 the difference of the generated and recovered spatial coordinates of each matched pair was computed. Ideally, these offsets should be scattered around zero. The median difference was computed and is marked with a red circle. It differs at most $0\farcm02$ from zero in both coordinate directions, the variance is about $1\farcm5$ which is much smaller than a beam size and corresponds approximately to one third of the size of a pixel. The fact that in the individual cases the recovered position has an offset from the generated coordinates is simply due to noise in the data cubes, the fainter the source the more uncertain the localization.

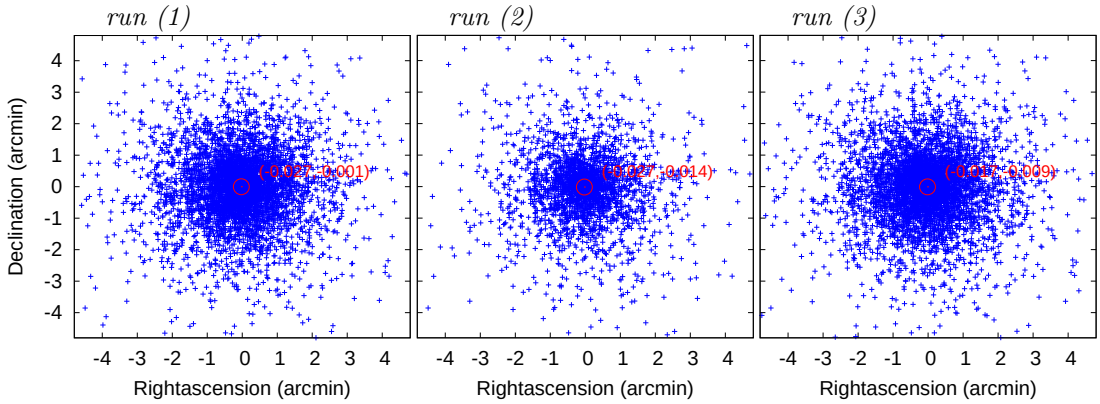


Fig. 4.5: Positional differences of matched galaxies for all runs. The red circle marks the median of all points. It differs at most 0.″02 from zero in both coordinates, the variance is about 1.″5 corresponding to one third of a pixel of the datacubes.

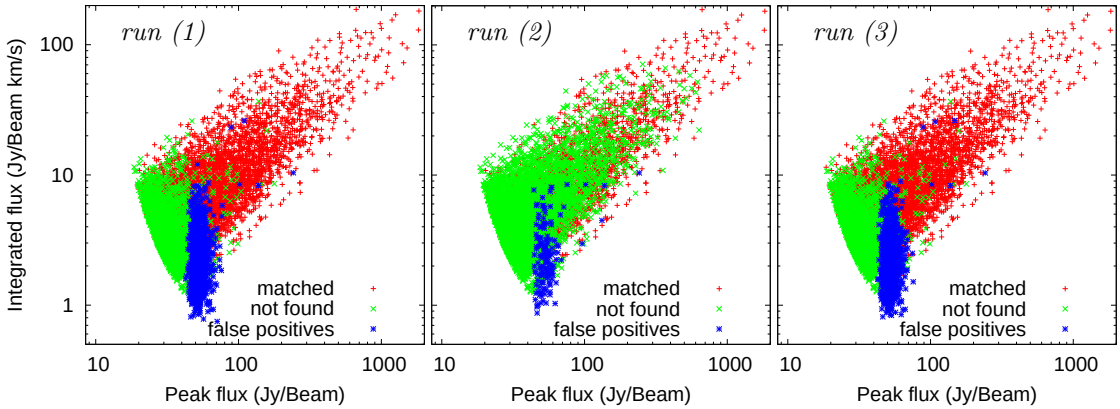


Fig. 4.6: The plots show the peak and integrated fluxes of matched pairs, non-detected sources, and false-positives in all three runs. While the matched and not detected sources cover in principle the same parameter region, the false-positives lie within a different (partly overlapping) regime. The plots reveal that the typical peak fluxes trace noise peaks, agglomerated around $S_{\mathrm{peak}} \approx 40 \ldots 70\,\mathrm{mJy\,Beam^{-1}}$; the lower value corresponds to $3\sigma_{\mathrm{rms}}$.

Also of interest are the properties of the false-positive detections. In Fig. 4.6 the distribution of the three populations in the peak and total flux plane are shown. Matched pairs and non-detections cover in principle the same region, the latter are more condensed to the lower flux regime. The false-positives have a different underlying distribution, most likely due to the fact that they are cause by single noise peaks. Hence they undergo a sharper cut with respect to the peak flux at about $40\,\mathrm{mJy\,Beam^{-1}}$ corresponding to $3\sigma_{\mathrm{rms}}$. Their total fluxes are less confined. Note, that for *run (2)* the second distribution of false-positives (compare to Fig. 4.3) is outside the plotted range. The properties of the false-positive detections make it likely to distinct them from true sources during a manual post-inspection of the datacubes. Additionally, the number of false-positives as a function of distance (or spectral channel) were analyzed; see Fig. 4.7. For *run (1)* and *run*

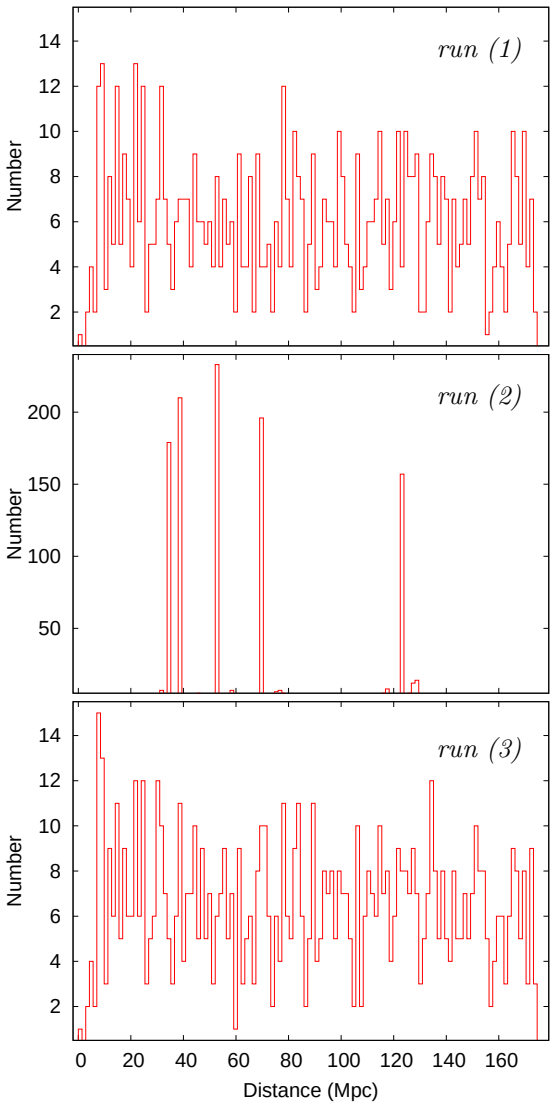


Fig. 4.7: The plots shows histograms for the distances of the false-positives resulting from all three runs. It is an important result that in *run (1)* and *run (3)* a homogeneous distribution is found, which is expected in case that noise peaks trigger the detections. Furthermore, the datacubes from *run (3)* show no significant increase in false-positive detection rate compared to *run (1)*, while without mitigation (*run (2)*) a large excess is visible at five distances, where RFI signals were added. For better visualization the data were binned with a width of 1.4 Mpc corresponding to eight spectral channels.

(3) the false-positive detections are equally distributed with respect to distance, which is expected for a homogeneous noise distribution, while for *run (2)* there are large excesses in the spectral channels containing interferences. The relative occurrence of events in these spectral channels is about two orders of magnitude higher than in the remaining channels. Furthermore, the datacubes from *run (3)* do not show any significant increase in false-positive detection rate compared to *run (1)* meaning the RFI detection worked sufficiently well.

In Fig. 4.6 a (partly) separation of false-positives and matched sources was observed. To further analyze the parameter coverage the distribution of the velocity profile widths of the false-positives and the matched pairs is plotted as well; see Fig. 4.8. Here, the separation of both populations becomes even more pronounced — only a small "interface" regime at $\Delta v_{50} \approx 50 \ldots 100\,\mathrm{km/s}$ is visible.

Finally, it is to say that in *run (2)* it is not that problematic that the false-positives are so numerous (in the results), because they could be (partly) filtered out using their

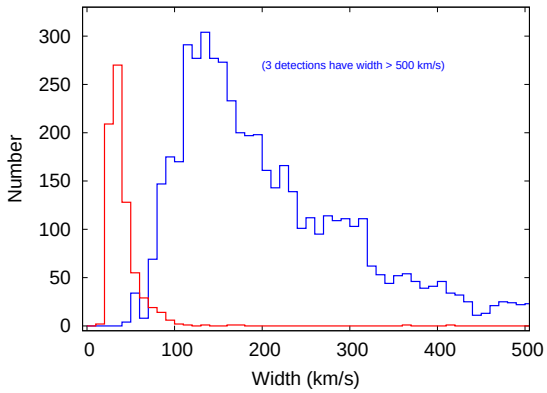


Fig. 4.8: Velocity profile widths of matched galaxies and false-positives for *run (1)*. The separation of both populations, matched pairs and false-positives, becomes more pronounced, overlapping only in a narrow "interface" regime at $\Delta v_{50} \approx 50 \ldots 100\,\mathrm{km/s}$.

small velocity profile width. What forbids the use of a dataset without RFI mitigation is that the number of candidates grows enormously causing the number of matched galaxies (especially the fainter ones) to decrease. In a certain sense this is due to some limitations of the Galaxy Parametrizer software; compare to Section 3.7.3. There, the number of voxels in the initial peak list (not the number of candidates) is restricted for computational reasons to be less than a 100 000. If too much RFI is present, the brightest 100 000 voxels in the peak list will not include all real sources, but only those which can compete with RFI fluxes. Another reason might be of course that some RFI signals mask true sources.

4.4.1 Completeness

An important issue when analyzing the final data is always the completeness of the survey. In most cases astronomical surveys suffer from selection effects — bright objects are visible to larger distances than faint sources. But there might be also systematics influencing the completeness, e.g., the used finder algorithm could fail to identify a certain kind of sources. As Zwaan et al. (2005) pointed out, for the HIPASS data the actual detection limit of that survey was not given by the peak flux or integrated flux limit but by a unknown combination of both. Lacking analytic solutions they propose to use an empirical approach by straying in artificial sources into the data during the reduction process. The comparison of generated and recovered objects leads to a completeness function

$$\mathcal{C}(\mathcal{P}^{\mathrm{gen}})\mathrm{d}\mathcal{P}^{\mathrm{gen}} \equiv \frac{n(\mathcal{P}^{\mathrm{rec}})}{n(\mathcal{P}^{\mathrm{gen}})}\mathrm{d}\mathcal{P}^{\mathrm{gen}}, \qquad n(\mathcal{P}^{\mathrm{gen}}) > 0 \tag{4.10}$$

with $n(\mathcal{P})$ being the number density of generated/recovered sources as a function of the parameter $\mathcal{P}$. For finite samples only the discrete completeness can be computed by binning the data. For this work the empirical completeness function with respect to the integrated and peak fluxes as well as to the velocity profile width were calculated; see Fig. 4.9 to Fig. 4.11. The plots additionally contain the numbers of sources generated per interval. For the integrated flux $S_{\mathrm{int}}^{\mathrm{gen}}$ (Fig. 4.9) the completeness reaches the 100% level at $S_{\mathrm{int}}^{\mathrm{gen}} \gtrsim 20\,\mathrm{Jy/Beam\,km/s}$. With respect to the peak flux $S_{\mathrm{p}}^{\mathrm{gen}}$ the source catalog is complete for $S_{\mathrm{int}}^{\mathrm{gen}} \gtrsim 60\,\mathrm{mJy/Beam}$ while for the velocity width w_{50}^{gen} it is found that the sample is incomplete for all values within the observed interval $0 \ldots 500\,\mathrm{km/s}$.

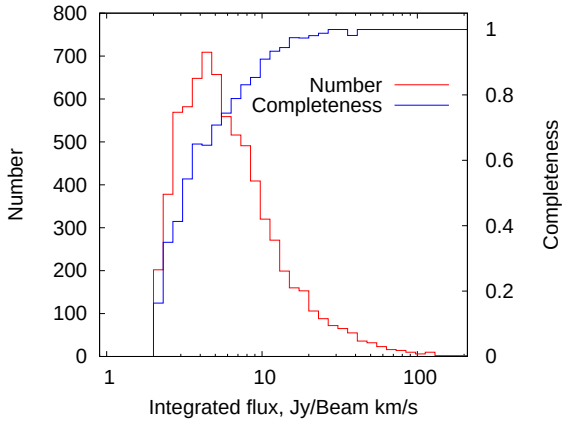


Fig. 4.9: Survey completeness as a function of integrated flux for *run (1)*. The plot shows the number of generated galaxies vs. $S_{\rm int}^{\rm gen}$ (red solid line) and the completeness $\mathcal{C}(S_{\rm int})$ (blue solid line). The completeness reaches the 100% level for $S_{\rm int}^{\rm gen} \gtrsim 20\,{\rm Jy/Beam\,km/s}$.

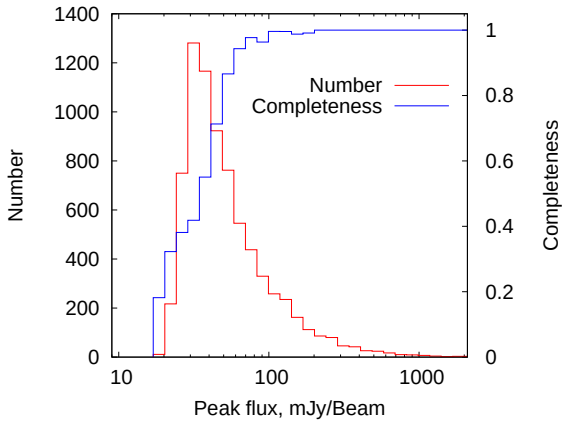


Fig. 4.10: Survey completeness as a function of peak flux for *run (1)*. The plot shows the number of generated galaxies vs. $S_{\rm p}^{\rm gen}$ (red solid line) and the completeness $\mathcal{C}(S_{\rm p}$ (blue solid line). The completeness reaches the 100% level for $S_{\rm int}^{\rm gen} \gtrsim 60\,{\rm mJy/Beam}$.

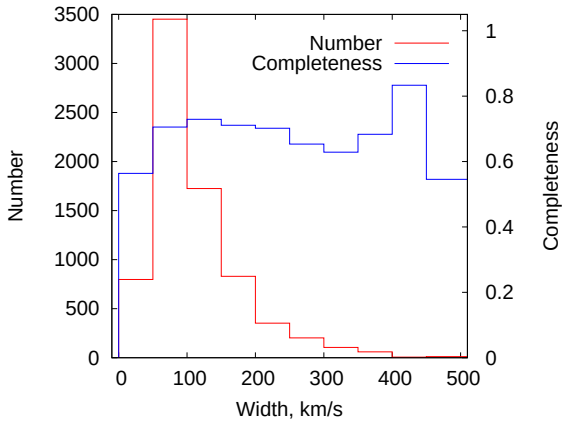


Fig. 4.11: Survey completeness as a function of velocity profile width for *run (1)*. The plot shows the number of generated galaxies vs. $w_{50}^{\rm gen}$ (red solid line) and the completeness $\mathcal{C}(w_{50}$ (blue solid line). For the full range of observed velocity widths the completeness is below 100%.

These empirical flux limits are rather large. However, this result can be explained by correlations between the different parameters. Equation 4.10 can easily be extended to the two-dimensional case,

$$\mathcal{C}(\mathcal{P}_1^{\rm gen},\mathcal{P}_2^{\rm gen}){\rm d}\mathcal{P}_{1,2}^{\rm gen} \equiv \frac{n(\mathcal{P}_1^{\rm rec},\mathcal{P}_2^{\rm rec})}{n(\mathcal{P}_1^{\rm gen},\mathcal{P}_2^{\rm gen})}{\rm d}\mathcal{P}_{1,2}^{\rm gen}\,, \qquad n(\mathcal{P}_1^{\rm gen},\mathcal{P}_2^{\rm gen}) > 0\,. \tag{4.11}$$

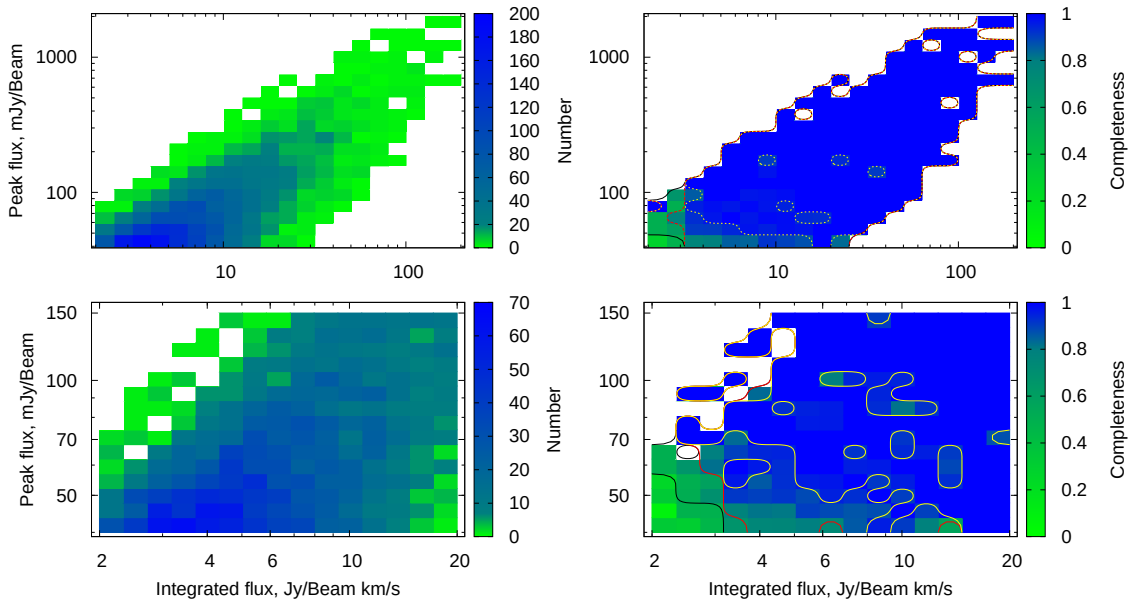


Fig. 4.12: Survey completeness as a function of integrated and peak flux for *run (1)*. The upper left panel shows the number of generated galaxies vs. $S_{\rm int}^{\rm gen}$ and $S_{\rm p}^{\rm gen}$. In the upper right panel we computed the completeness $\mathcal{C}(S_{\rm int}, S_{\rm p})$. The lower panels show a zoom-in of the low-flux regime. Plotted as contours are the 50% (black), 75 (red), and 95%-levels (yellow).

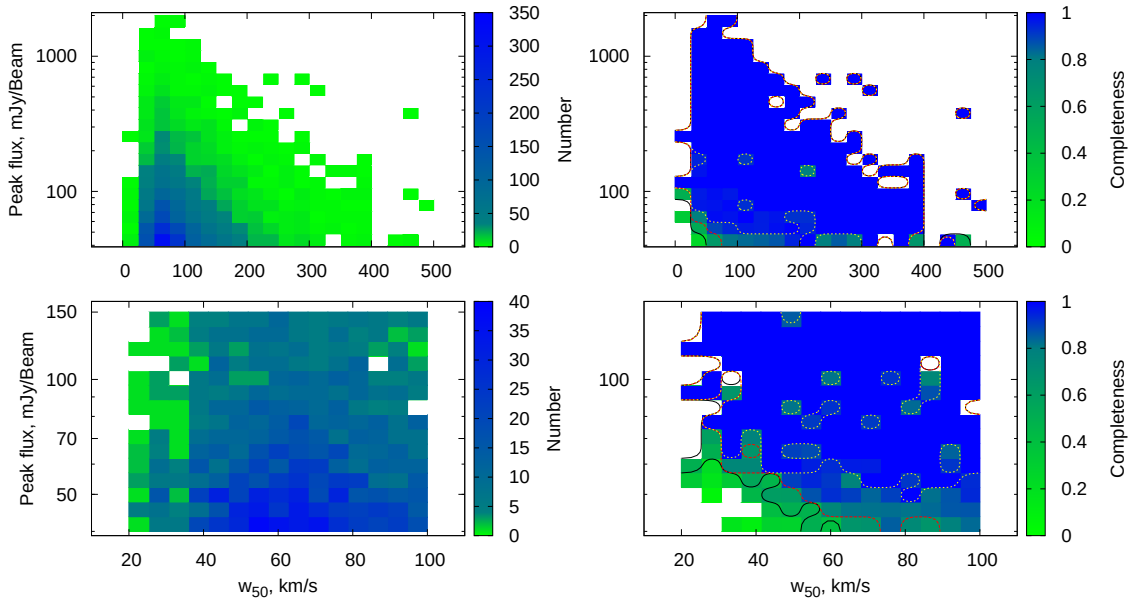


Fig. 4.13: As Fig. 4.12 but showing the completeness $\mathcal{C}(w_{50}, S_{\rm p})$ as a function of peak flux and velocity profile width for *run (1)*.

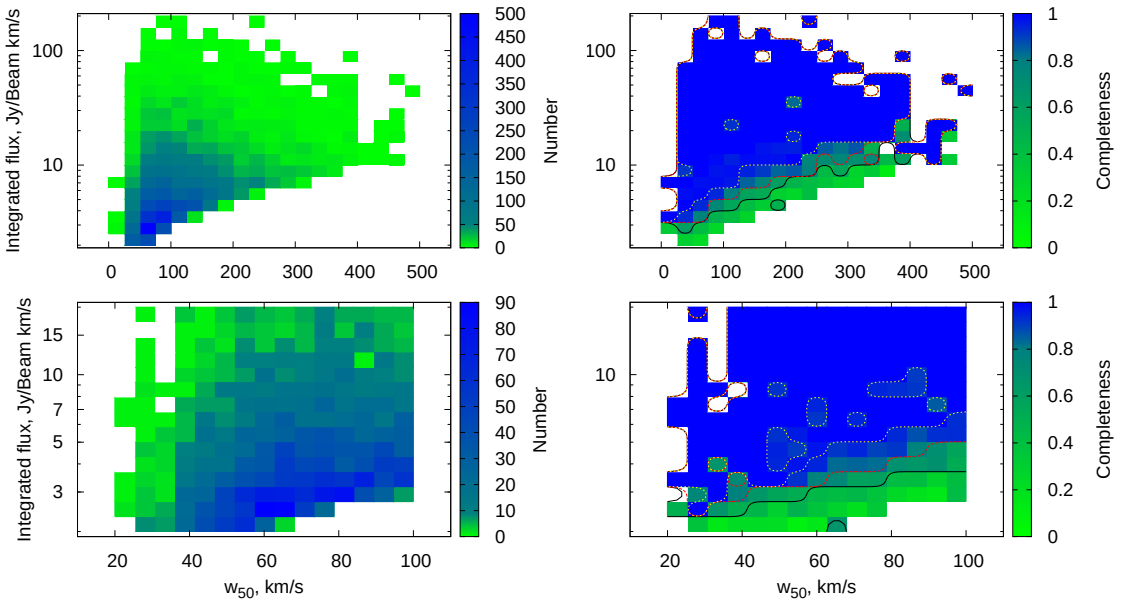


Fig. 4.14: The same display like in Fig. 4.12 but showing the completeness $\mathcal{C}(w_{50}, S_{\rm int})$ as a function of integrated flux and velocity profile width for *run (1)*.

In Fig. 4.12 to Fig. 4.14 the two-dimensional completeness functions $\mathcal{C}(S_{\rm int}, S_{\rm p})$, $\mathcal{C}(w_{50}, S_{\rm p})$, and $\mathcal{C}(w_{50}, S_{\rm int})$ are shown as derived from the simulations. The figures contain a zoom-in of the low-value regimes of the parameters as well. It turns out that the non-detections having rather large integrated flux are sources with very low peak flux and vice versa. Therefore, the completeness of the sample is much better than expected from Fig. 4.9 and Fig. 4.10, though one still can not give an analytic expression for the detection limit.

From Fig. 4.14 it follows that the incompleteness with respect to w_{50} is caused by the lowest integrated flux generated within each specific width interval. Sources, having small peak fluxes, can produce rather large integrated fluxes due to broad velocity widths. In that case they might be overlooked by the Gamma test source finder algorithm. A broad Gaussian with low amplitude is a rather "smooth" spectral feature, hence, the Gamma value could be to low for the applied threshold (compare to Section 3.7.1).

Finally, for better comparison with the HIPASS results, shown in Fig. 4.15 are the weighted completeness functions

$$\mathcal{C}_{\rm w}(\mathcal{P}^{\rm gen}|\tilde{\mathcal{P}}^{\rm gen})\mathrm{d}\mathcal{P}^{\rm gen} \equiv \frac{\sum_{\tilde{\mathcal{P}}^{\rm gen}} n(\mathcal{P}^{\rm gen}, \tilde{\mathcal{P}}^{\rm gen})}{\sum_{\tilde{\mathcal{P}}^{\rm gen}} n(\mathcal{P}^{\rm gen}, \tilde{\mathcal{P}}^{\rm gen})/\mathcal{C}(\mathcal{P}^{\rm gen}, \tilde{\mathcal{P}}^{\rm gen})}\mathrm{d}\mathcal{P}^{\rm gen} \qquad (4.12)$$

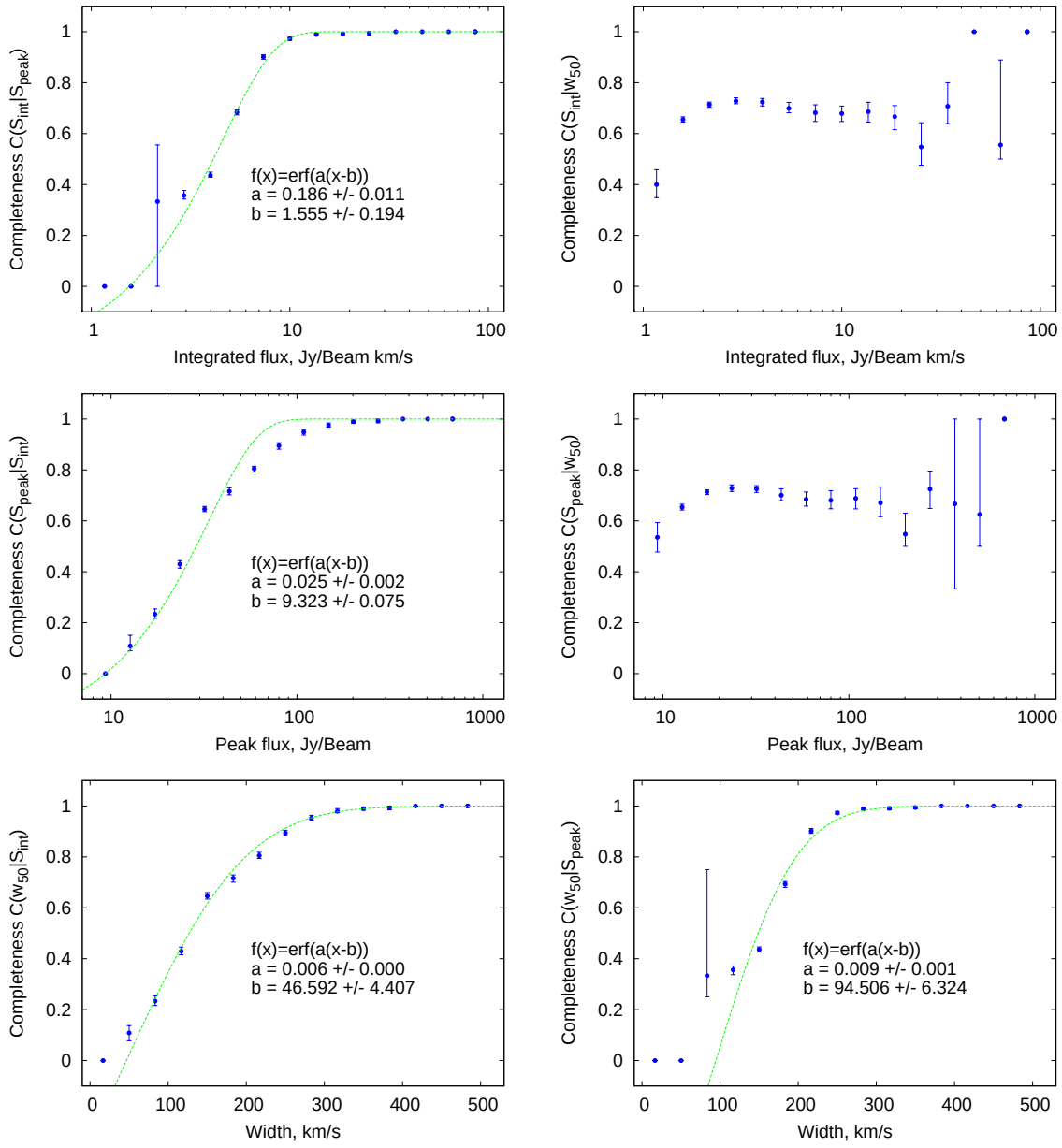


Fig. 4.15: One-dimensional weighted completeness functions $\mathcal{C}_{\mathrm{w}}(\mathcal{P}_{\mathrm{gen}}|\tilde{\mathcal{P}}_{\mathrm{gen}})\mathrm{d}\mathcal{P}_{\mathrm{gen}}$ computed by marginalization over the second parameter of the two-dimensional completeness functions. The error bars were computed using a bootstrap method. When reasonable, an error function $\mathrm{erf}\left(a\left(\mathcal{P}_{\mathrm{gen}}-b\right)\right)$ was fitted to the data, the associated fit parameters are given in the plots.

which can be computed by marginalization over the second parameter of the two-dimensional completeness functions. The error bars were computed applying the same bootstrap method used by Zwaan et al. (2004)[2]. Error functions $\mathrm{erf}\left(a\left(\mathcal{P}_{\mathrm{gen}}-b\right)\right)$ with

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x \mathrm{d}t\ e^{-t^2} \tag{4.13}$$

[2] "From the parent population of N synthetic sources, N sources are chosen randomly, with replacement. This is repeated 200 times, and for each of these 200 regenerated samples the completeness $\mathcal{C}'$ is calculated

were fitted through the data points when possible, i.e., for w_{50} as marginalization parameter this was not reasonable, because the completeness never reaches the 100% level. The fit results are similar to what was found by Zwaan et al. (2004).

For the final survey data it will be necessary to repeat the presented simulations incorporating additional velocity profile shapes, i.e., top-hat and double-horn profiles. This will not only solve some issues (completeness vs. profile width too low), but is essential for a realistic estimation of the survey completeness, as Gaussian line profiles only account for a part of the sources. Note, that the exact knowledge of the completeness, and hence detection properties, is crucial for the calculation of the HIMF.

4.5 Systematic effects

During the analysis a number of systematic effects were observed which are very likely produced by the kind of parameter extraction. An analysis of systematic effects was already performed by Zwaan et al. (2004) for the HIPASS. They compared generated and recovered parameters by computing the difference $\Delta\mathcal{P} \equiv \mathcal{P}^{\mathrm{rec}} - \mathcal{P}^{\mathrm{gen}}$. To have a quantitative description, the resulting values $\Delta\mathcal{P}$ were binned and Gaussians were fitted to the data. The mean of the Gaussian is an estimator for bias effects while the variance describes the scatter of the recovered quantities around the mean value. The problematic aspect using this approach is, that it is not clear whether the scatter of the parameters can be described by a Gaussian distribution function. This issue will be discussed in more detail in a subsequent paragraph.

In Fig. 4.16 both the differences between generated and recovered integrated, ΔS_{int}, and peak flux, ΔS_{p}, respectively, are plotted as a function of reconstructed flux[3]. Following (Zwaan et al. 2004) the mean and variance were not only computed for the overall histogram (upper panel) but for certain flux ranges (lower panel) as well. These ranges are indicated by the red points, the errorbars mark the 3σ interval of the underlying Gaussian. The left panel of Fig. 4.16 shows the peak flux ΔS_{p}. Especially, in the low-flux region a strong bias of $\mu \sim 25\,\mathrm{mJy/Beam}$ is observed, while in the high-flux regime the difference becomes negative ($\mu \sim -20\,\mathrm{mJy/Beam}$), suggesting two systematic effects counteracting each other. In the latter case, however, the scatter is large, making the value less significant.

The peak flux is not a physically relevant quantity, as it depends strongly on the beam pattern, i.e, the beam filling factor, and the gridding scheme. Of much more interest is, therefore, the integrated flux, ΔS_{int}; see Fig. 4.16 (right panel). A slight overall shift $\mu = 0.3\,\mathrm{Jy/Beam\,km/s}$ of the recovered integrated fluxes (right panel) to larger values is observed. The scatter is moderately larger than that measured for HIPASS ($1.6\,\mathrm{Jy/Beam\,km/s}$ compared to $1.1\,\mathrm{Jy/Beam\,km/s}$). Zwaan et al. (2004) do not report an excess but were observing an increasing spread towards larger integrated fluxes, while there the determined variances are approximately independent on the flux. All of the discussed systematic effects can be explained; see Section 4.5.1. For the lowest integrated

following [Eq. (4.12)]. The 1σ upper and lower errors on the completeness are determined by measuring from the distribution of $\mathcal{C}'$ the 83.5 and 16.5 percentiles." (Zwaan et al. 2004)

[3] Note, that it would eventually be more meaningful to plot the differences as a function of the generated values, as these are the "fixed" quantities. However, to allow for comparison with the results of (Zwaan et al. 2004) this is not done here.

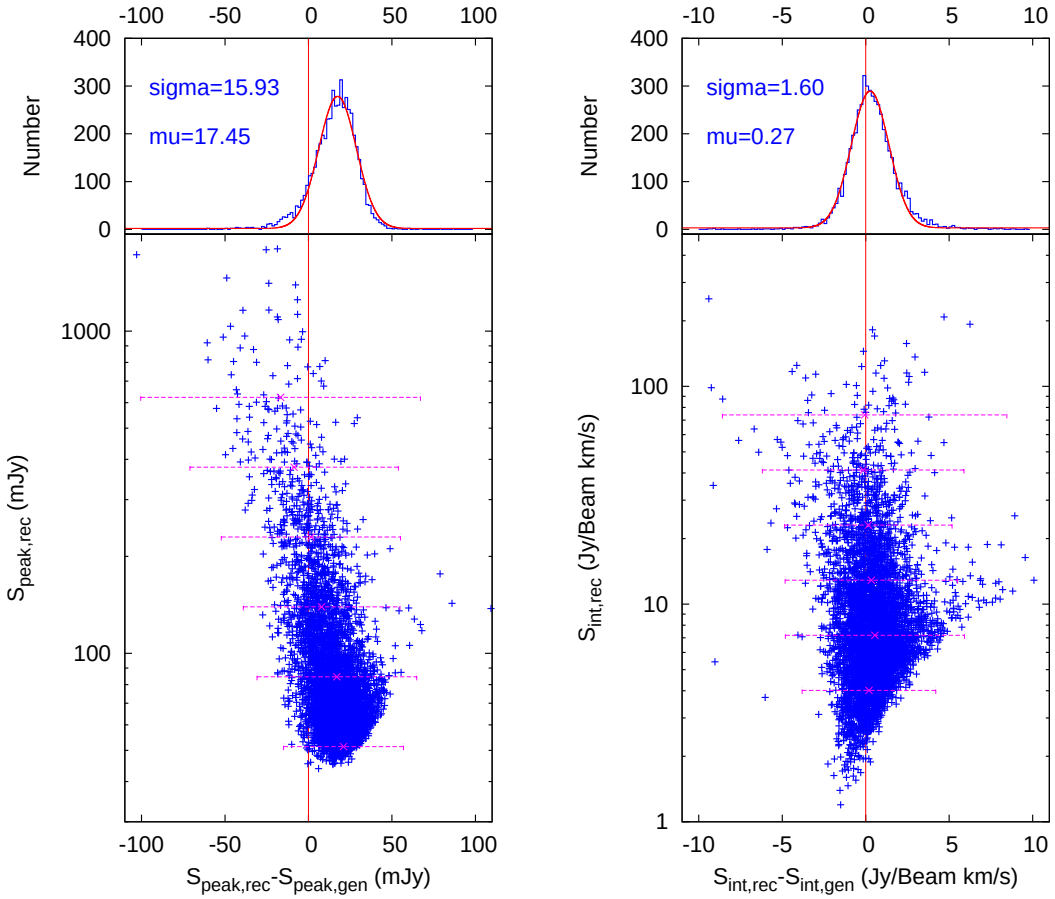


Fig. 4.16: Systematic differences in integrated (right panel) and peak (left panel) flux as a function of absolute peak and integrated flux, respectively. The lower panels show scatter plots, while in the upper part the data were binned. A Gaussian profile was fitted to the distribution to have a quantitative handle on any bias effects and to estimate the scatter of the data around the mean value. The peak fluxes show a significant bias in the low-flux regime as well as in the higher-flux regime, counteracting each other. Furthermore, a slight overall shift of the recovered integrated fluxes to larger values is observed. All three effects can likely be explained; see Section 4.5.1. For the lowest integrated fluxes, the recovered fluxes are slightly underestimated, which probably depends on the determination of the matching size of subcubes in case of broad profile widths. However, this issue is only be observed for a small fraction of sources. Additionally, there is a kind of "wing" for medium integrated flux values, produced by overestimation $S_{\rm int}^{\rm rec}$.

fluxes, the recovered fluxes are slightly underestimated, which is probably due the determination of the matching extent of subcubes covering broad H I profile widths. However, this is only observed for a small fraction of sources. Additionally, there is a kind of "wing" for medium integrated flux values, produced by overestimation $S_{\rm int}^{\rm rec}$; the further discussion of this "wing" is postponed here to Section 4.5.1.

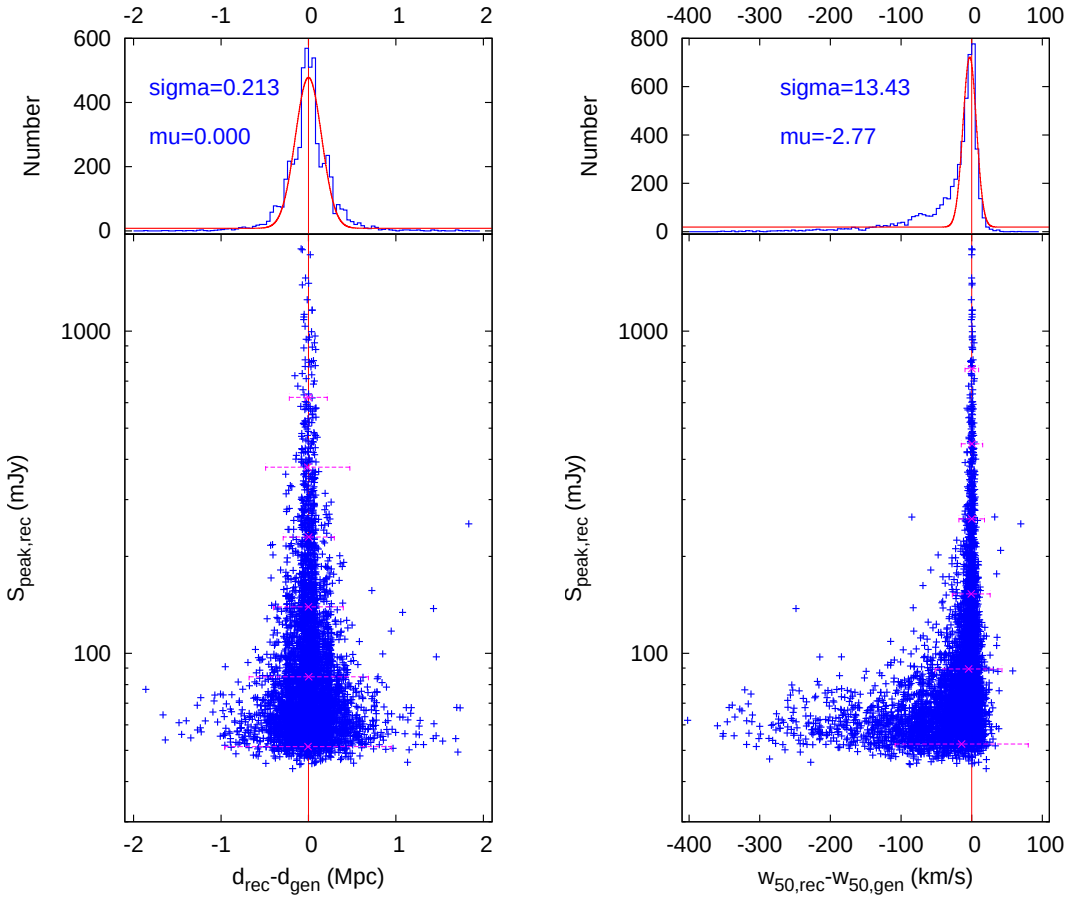


Fig. 4.17: The same plot as in Fig. 4.16, but showing the distance d and velocity profile width w_{50}. While the distances are distributed symmetrically around zero mean, the widths show a large negative excess for small peak fluxes.

In Fig. 4.17 the distance d (left panel) and velocity profile widths w_{50} errors are plotted. While the distances are symmetrically distributed around zero (mean value of 0.004 Mpc, scatter of 0.2 Mpc), the widths show a strong negative excess in the low flux regime. In the high-flux regime this bias is not seen. The effect, is likely directly related to the parametrization of w_{50}. Again, the discussion of this effect will be postponed to a subsequent paragraph (Section 4.5.1).

Finally, Fig. 4.18 shows the scatter in the positional parameters right ascension and declination. As for the distance, the distributions are symmetric and have small a scatter, pointing out the high positional accuracy of the fitting procedure. Note, that the variances are still larger than those measured by Zwaan et al. (2004), but for the fit they probably did not take into account the clearly visible wings of the distribution profile. Again, the question arises, whether the use of Gaussian profiles for the description of the data is a meaningful quantity. In Section 4.5.3 another approach is presented to quantify the quality of the reconstructed parameters.

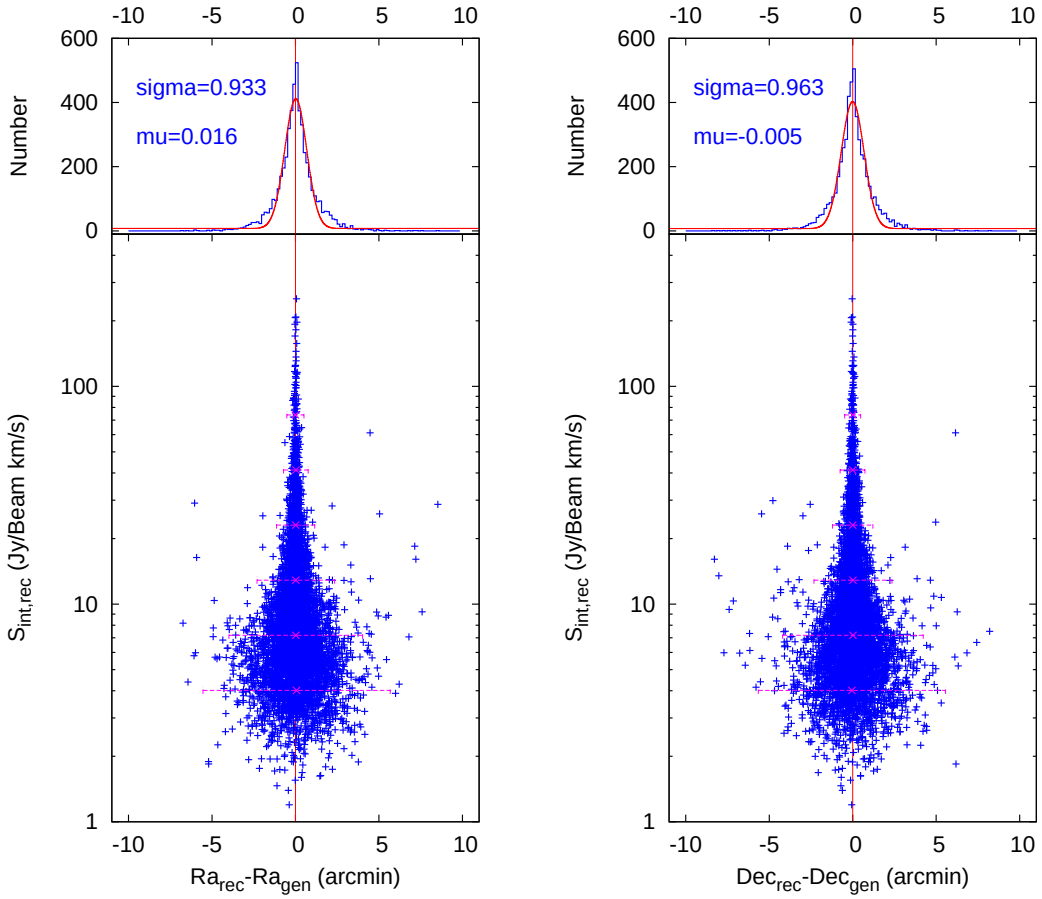


Fig. 4.18: The same plot as Fig. 4.16, but showing the positional offsets in Right-Ascension and Declination. The distributions are scattered symmetrically around zero mean and reveal small variances.

It is also interesting to reveal any correlations between different parameters, which was not possible by applying the approach of Zwaan et al. (2004). Fig. 4.19 to Fig. 4.21 show the difference of the peak and integrated flux, as well as of the velocity profile width of the generated and the recovered quantities (of all matched pairs). In each diagram the difference of the actual quantity is color-coded and drawn as a function of the generated fluxes and widths. In the special case of using only Gaussian velocity profiles, a degeneracy between the parameters peak flux, integrated flux, and velocity width is encountered. This means that in the parameter space spanned by those quantities all points lie on a plane. Remember, that the simulated values are the Gaussian parameters as sampled from the various input distributions (see Section 4.1), while the recovered values are not being determined by Gaussian fitting procedures (see Section 3.7.3 for details). This means, that observed systematic effects might have their origin within the data reduction itself or in the way the determination of the parameters is done. In fact both mechanisms are involved. Before this is discussed in detail, the observed effects will be described.

In Fig. 4.19 the systematic differences of the peak flux, $\Delta S_{\mathrm{peak}} = S^{\mathrm{gen}}_{\mathrm{peak}} - S^{\mathrm{rec}}_{\mathrm{peak}}$, are shown. The left panel shows the absolute deviations, the right panel the relative deviations

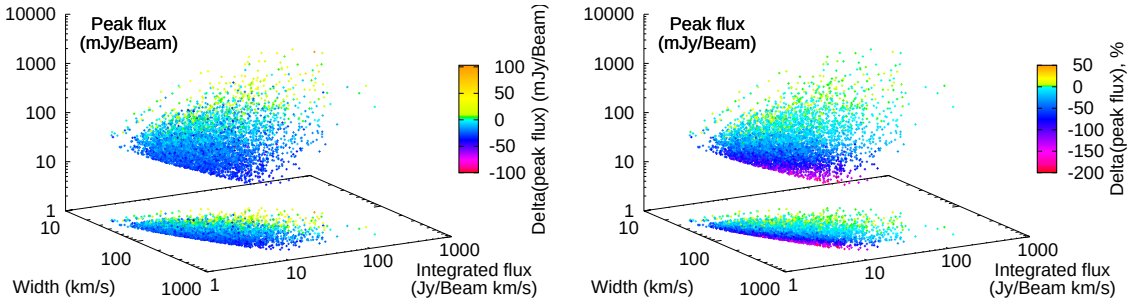


Fig. 4.19: Systematic effects and correlations for $\Delta S_{\mathrm{peak}} = S^{\mathrm{gen}}_{\mathrm{peak}} - S^{\mathrm{rec}}_{\mathrm{peak}}$. The left panel shows the absolute difference as a function of peak flux, width and integrated flux of each generated galaxy. Note, that the true distribution of the latter three quantities is two-dimensional, as the three parameters are not independent. The right panel, showing relative differences, reveals that for smaller peak fluxes, $S^{\mathrm{rec}}_{\mathrm{p}}$ is systematically higher than $S^{\mathrm{gen}}_{\mathrm{p}}$, while for higher peak fluxes, $S^{\mathrm{rec}}_{\mathrm{p}}$ is on average slightly lower than $S^{\mathrm{gen}}_{\mathrm{p}}$. Both effects were already observed in the one-dimensional distribution (Fig. 4.16). However, using the three-dimensional presentation also a slight dependence on w_{50} can be observed for the strength of the flux overestimation for low $S^{\mathrm{gen}}_{\mathrm{p}}$.

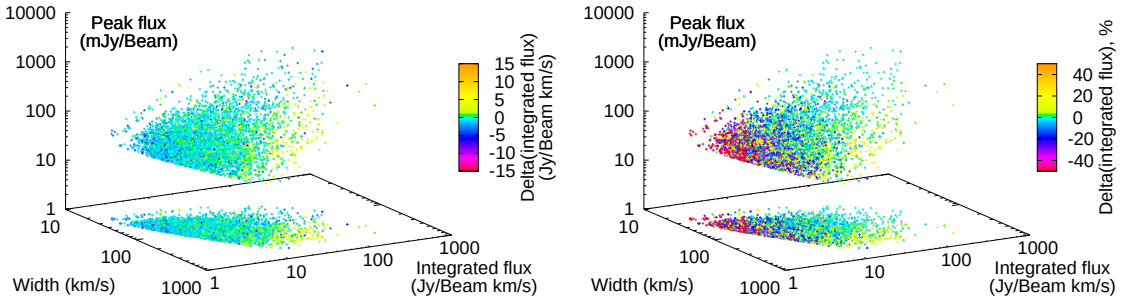


Fig. 4.20: Systematic effects and correlations for $\Delta S_{\mathrm{int}} = S^{\mathrm{gen}}_{\mathrm{int}} - S^{\mathrm{rec}}_{\mathrm{int}}$. The left (right) panel shows the absolute (relative) difference of the integrated fluxes as a function of peak flux, width and integrated flux. Two relatively small systematic deviations are visible. First, the overestimation of $S^{\mathrm{rec}}_{\mathrm{int}}$ for small $S^{\mathrm{gen}}_{\mathrm{int}}$. Second, a slight underestimation of $S^{\mathrm{rec}}_{\mathrm{int}}$ for large w^{gen}_{50}.

of the recovered peak fluxes compared to the generated values. Especially, the right panel reveals that for smaller peak fluxes, $S^{\mathrm{rec}}_{\mathrm{p}}$ is systematically higher than $S^{\mathrm{gen}}_{\mathrm{p}}$, while for higher peak fluxes, $S^{\mathrm{rec}}_{\mathrm{p}}$ is on average slightly lower than $S^{\mathrm{gen}}_{\mathrm{p}}$. Both effects are the result of two counteracting mechanisms discussed in Section 4.5.1 and were already detected in the one-dimensional distribution (Fig. 4.16). However, using the three-dimensional presentation also a slight dependence on w_{50} can be observed for the strength of the flux overestimation for low $S^{\mathrm{gen}}_{\mathrm{p}}$.

Next, in Fig. 4.20 the differences in the integrated flux $\Delta S_{\mathrm{int}} = S^{\mathrm{gen}}_{\mathrm{int}} - S^{\mathrm{rec}}_{\mathrm{int}}$ are shown. The integrated flux is not subject to the error-prone estimation of peak fluxes and widths (see Section 4.5.1, Fig. 4.19, and Fig. 4.21). Nevertheless, two relatively small systematic

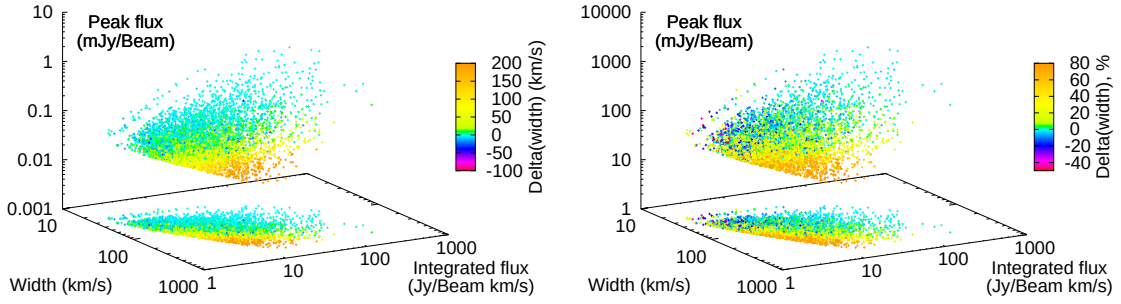


Fig. 4.21: Systematic effects and correlations for $\Delta w_{50} = w_{50}^{\mathrm{gen}} - w_{50}^{\mathrm{rec}}$. The left (right) panel shows the absolute (relative) difference of the widths as a function of peak flux, width and integrated flux. The recovered widths, w_{50}^{rec}, are systematically underestimated for small $S_{\mathrm{p}}^{\mathrm{gen}}$ and large w_{50}^{gen}.

deviations are detectable. First, the overestimation of $S_{\mathrm{int}}^{\mathrm{rec}}$ for small $S_{\mathrm{int}}^{\mathrm{gen}}$. This bias could be a selection effect; see Section 4.5.1. Second, a slight underestimation of $S_{\mathrm{int}}^{\mathrm{rec}}$ for large w_{50}^{gen}.

Finally, Fig. 4.21 visualizes the difference $\Delta w_{50} = w_{50}^{\mathrm{gen}} - S_{50}^{\mathrm{rec}}$ in the profile widths. The recovered widths are systematically underestimated for smaller peak fluxes and larger widths. This, again, has to do with the specific way of the HICAT parametrization; see the following section.

4.5.1 Explanation of the systematic effects

The following compilation discusses several mechanisms which could cause the observed systematic effects.

1. The gridding causes the recovered peak flux to be slightly lower than the simulated one, because the gridding algorithm, which was used, conserves the total flux and increases the spatial resolution by a few percent. This effect is practically not important, as the peak flux is by no means a useful quantity — it depends on the angular resolution as well as on the (unknown) beam filling factor. This effect is one of the reasons for deviations in the peak flux (Fig. 4.16 and Fig. 4.19).

2. The determination of peak flux and profile width using the HICAT parametrization scheme according to Meyer et al. (2004); see also Fig. 4.22.

   (a) The peak flux is determined by finding the pixel with the highest flux in each subcube. This is not a good estimator in the presence of noise, as $S_{\mathrm{p}}^{\mathrm{rec}} \equiv \max_i(S_i^{\mathrm{rec}}) = \max_i(S_{\mathrm{i}}^{\mathrm{gen}} + S_i^{\mathrm{noise}}) \neq \max_i(S_{\mathrm{i}}^{\mathrm{gen}}) \equiv S_{\mathrm{p}}^{\mathrm{gen}}$, with the noise contribution S_i^{noise} in the i-th spectral channel; see Fig. 4.23. This effect causes the peak fluxes to be overestimated, which has a relatively larger impact on smaller $S_{\mathrm{p}}^{\mathrm{gen}}$. Furthermore, the larger w_{50} the larger the likelyhood that a specific large noise value adds up to the observed emission line, introducing a slight dependence on velocity profile widths. Nevertheless, this method of parametrization was chosen because it was used for the HICAT galaxies. Both systematics can be seen in the deviations of the peak flux (Fig. 4.19), while only one of the

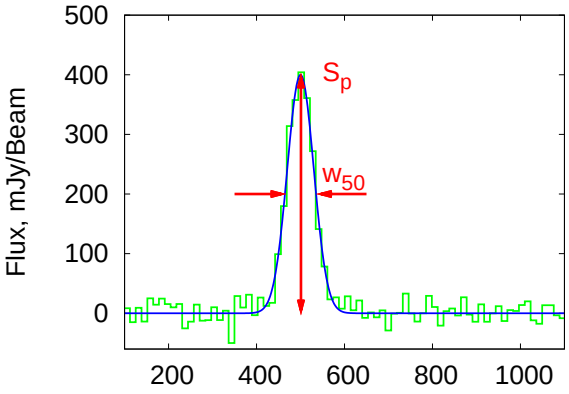


Fig. 4.22: The parametrization used to obtain peak fluxes and velocity profile widths as used for the HIPASS catalog (HICAT). The peak flux S_p is determined by searching for the highest flux value in the velocity interval given by the finder algorithm. The width w_{50} is calculated by searching those spectral channels at which the spectral value has fallen below 50% the peak value (of the weighted spectrum). The 20%-width is defined accordingly. This kind of parametrization can lead to problems; see Fig. 4.23 and Fig. 4.24.

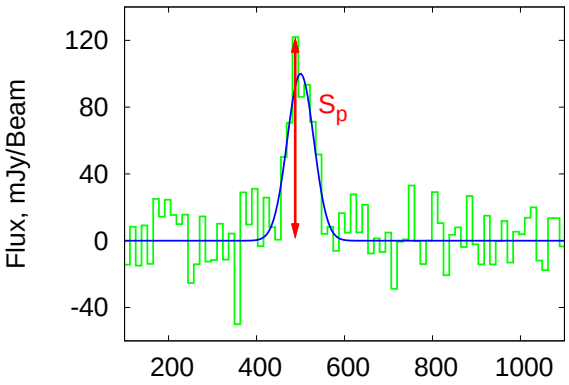


Fig. 4.23: The algorithm used to obtain peak fluxes (see Fig. 4.22) can lead to a bias. Using the HICAT parametrization scheme, the peak flux is overestimated due to the addition of noise to the true spectrum. As the figure sketches, a single noise peak can add significantly to the underlying profile. The measured highest peak has not even to match the underlying function's peak in velocity. The relative error introduced not only depends on the amplitude of the velocity profile but also on its width and shape.

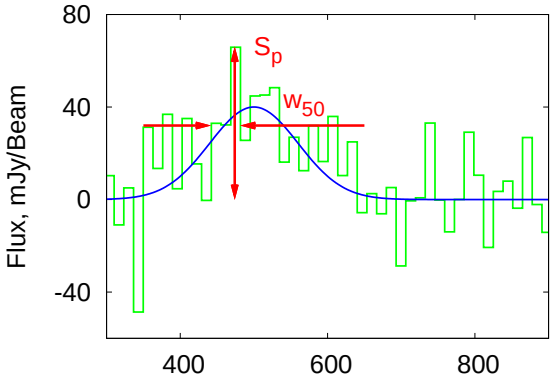


Fig. 4.24: The algorithm used to obtain velocity profile widths (see Fig. 4.22) can lead to a bias. For low peak values (in the weighted spectrum) negative noise values can force the algorithm to stop too early. Hence, the velocity width, w_{50}^rec, can be strongly underestimated.

discussed biases (the overestimation) is visible in the one-dimensional plot in Fig. 4.16.

(b) The width of the galaxy profile is determined using the peak value in the weighted spectrum (which has nothing to do with the peak flux and is less sensitive to noise). Going to higher and lower velocities those pixels in the ve-

locity profile are searched, where the profile falls below half of the peak value. If the peak intensity is already low (compared to the noise) it may happen that the halfwidth points lie within the noise. Then the width will likely be underestimated, because a single negative noise value can stop the algorithm; see Fig. 4.24. The broader the profile the higher the likelyhood that the effect occurs, being probably the reason for the deviations of the profile widths (Fig. 4.17 and Fig. 4.21).

Additionally, the profile width after computing the weighted spectrum is not necessarily equal to the width of the Gaussian velocity profile during the generation of the spectra (the sources were parametrized as the multiplication of a Gaussian in spatial coordinates with a Gaussian in velocity space). At least, one can assume that this only introduces a constant error in the relative differences of the profile widths.

Both effects are highly dependent on the type of the profile shape. In the simulations only Gaussian shaped emission lines were used. In order to be able to correct for the bias introduced, other velocity profile shapes would need to be incorporated into the analysis.

3. Another source of uncertainty is due to a selection effect. From the completeness function it follows that objects with low integrated fluxes are found only to a certain fraction. After adding noise the uncertainty of the recovered total flux of a faint source (of small width) will be relatively increased. Sources which flux was increased during this process will have a slightly higher detection probability, introducing a bias — galaxies with small integrated fluxes will more likely exhibit overestimated integrated fluxes, as long as the completeness is below 100%; see Fig. 4.25. Note, that this effect is independent of profile shape. The resulting bias is presumably the reason for deviations in the total flux (Fig. 4.16 and Fig. 4.20) in the low flux regime.

4. For the largest profile widths, the determination of the subcube sizes using the Galaxy Parametrizer (Section 3.7.3) probably does not include the faint outermost parts of the emission lines. This results in a slight underestimation of $S_{\mathrm{int}}^{\mathrm{rec}}$ for large w_{50}^{gen} (see Fig. 4.20), producing the "wing" in Fig. 4.16. However, the error is small and only relevant for a small number of sources. Note also, that Gaussian profiles of such large width are not very realistic, but usually different profile shapes are found for broad sources.

4.5.2 Quantification of the parametrization bias

To quantify the bias introduced by HICAT parametrization scheme (overestimation of the peak flux and underestimation of widths in the low flux regime), further simulations were performed. By generating single spectra with known underlying velocity profile the influence of the parametrization for a broad range of peak fluxes and profile widths can be empirically determined, while any effects caused by the data reduction pipeline itself will not distort the results.

For the plots in Fig. 4.26 in total 2000 individual spectra were simulated by adding noise ($\sigma_{\mathrm{rms}} = 13\,\mathrm{mJy/Beam}$) to a Gaussian of width w_{50} and amplitude S_{p}. Both parameters

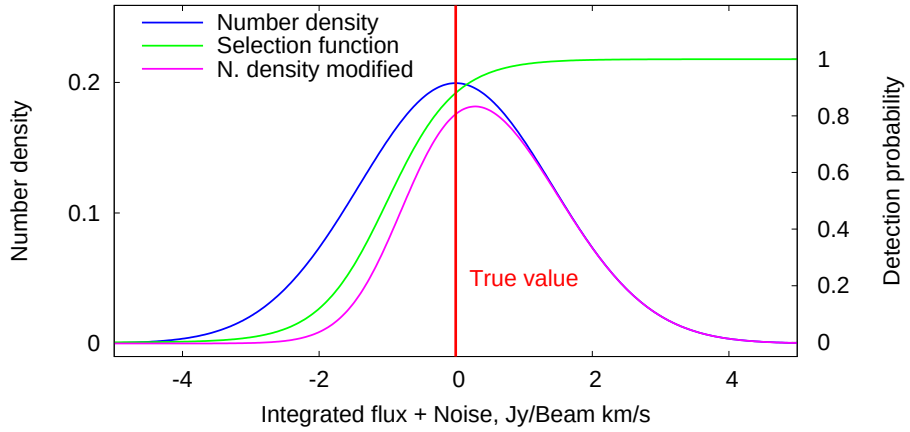


Fig. 4.25: For integrated fluxes, $S_{\rm int}^{\rm gen}$, in the incompleteness regime of the survey (i.e., the selection function is less than one) $S_{\rm int}^{\rm rec}$ can be biased, i.e., on average reveal overestimated fluxes. Due to noise, the actual measured value of the integrated flux can differ from the true value. In the sketch the magenta curve describes the probability of having a certain measurement error after applying the selection function (green curve) to the original probability density (blue curve). The expectation value is shifted to higher fluxes, causing the bias. Furthermore, the modified distribution function of the errors becomes non-Gaussian. As the selection function is unknown the modified distribution can not be analytically described.

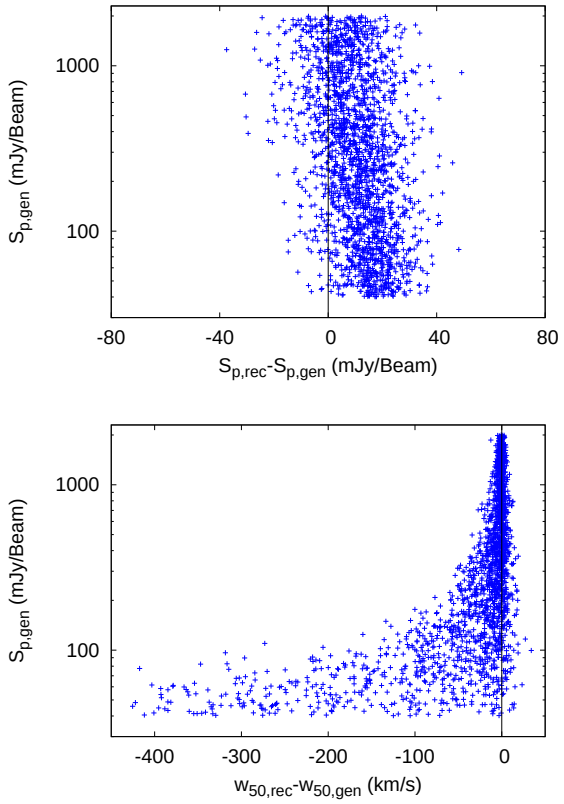


Fig. 4.26: To test if the systematics observed (see Fig. 4.19 and Fig. 4.21) can be caused by the HIPASS parametrization scheme we simulate single spectra containing a Gaussian profile plus noise. Peak fluxes $S_{\rm p}$ and widths w_{50} of the velocity profile were uniformly sampled in the region of interest. (For the peak flux $\log(S_{\rm p})$ was evenly distributed.) After that the Parametrization scheme was applied (see also Fig. 4.22). In the left panel the difference $S_{\rm p}^{\rm rec} - S_{\rm p}^{\rm gen}$ as a function of the generated flux $S_{\rm p}^{\rm gen}$ is shown. The right panel shows $w_{50}^{\rm rec} - w_{50}^{\rm gen}$ again as a function of the generated flux $S_{\rm p}^{\rm gen}$.

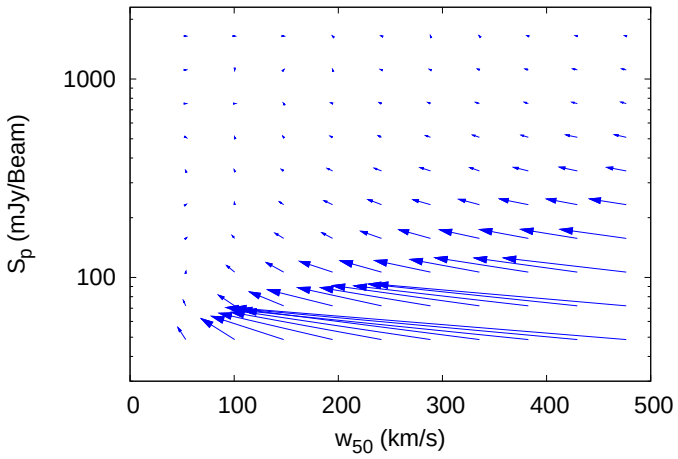


Fig. 4.27: Similar plot as in Fig. 4.26 but the sampled peak fluxes and widths were distributed on a regular grid. For each grid cell 100 spectra were generated. The vectors point from the generated parameters to the recovered ones. For large fluxes there is no significant bias whereas low fluxes reveal an overestimation of the peak flux and underestimation of the profile width.

$(\log S_\mathrm{p}^\mathrm{gen}, w_{50}^\mathrm{gen})$ were uniformly sampled from the intervals $w_{50}^\mathrm{gen} = 30 \ldots 500\,\mathrm{km/s}$ and $S_\mathrm{p}^\mathrm{gen} = 40 \ldots 2000\,\mathrm{mJy/Beam}$, respectively. The plots show the difference of the peak fluxes, $S_\mathrm{p}^\mathrm{rec} - S_\mathrm{p}^\mathrm{gen}$, (upper panel) and widths, $w_{50}^\mathrm{rec} - w_{50}^\mathrm{gen}$, (lower panel) as a function of $S_\mathrm{p}^\mathrm{gen}$. As expected, the peak fluxes are overestimated in the low-flux regime, while the widths are on average strongly underestimated. Note that both graphs look very similar to Fig. 4.16 and Fig. 4.17.

To reveal possible correlations between both quantities a vector plot is used in Fig. 4.27. To improve the visualization the flux–width pairs were produced on a regular grid to avoid overlapping arrows. To lower noise effects 100 spectra were generated per grid cell and the extracted parameters for each cell were averaged. The origin of each vector marks the simulated parameters pointing to the recovered values. In this representation systematic effects show-up easily. For high fluxes no significant bias is visible but towards lower fluxes and larger widths, S_p is increasingly overestimated, and w_{50} underestimated. The plot can be used to quantify the average bias for each flux and width interval though it neglects possible additional systematics (e.g., originating from the data reduction).

Consequently, it is of interest whether these systematics can be suppressed. One possibility to estimate the profile widths would be to calculate the second moment which, for Gaussian profiles, equals $w_\sigma = w_{50}/\sqrt{8\ln 2}$. The first three moments are defined as

$$\mathcal{M}_0 = \sum_i p_i \tag{4.14}$$

$$\mathcal{M}_1 = \frac{1}{\mathcal{M}_0} \sum_i p_i v_i \tag{4.15}$$

$$\mathcal{M}_2 = \frac{1}{\mathcal{M}_0} \sum_i p_i \left(v_i - \mathcal{M}_1\right)^2 \tag{4.16}$$

with the flux p_i in velocity channel v_i. Unfortunately, $\mathcal{M}_2$ is very sensitive to noise in the spectrum. Meaningful results were only obtained if the noise is very low compared to the Gaussian peak value or if only spectral channels within 3σ(Gaussian) were used for the computation. Both conditions are not fulfilled in practice.

Another possibility is to "smooth" the spectra with a low-pass filter before applying the parametrization procedure[4]. In fact, the HIPASS spectra were filtered, first by a Tukey filter (giving a new spectral resolution of 18 km/s), then Hanning smoothing was applied (final resolution: 26.4 km/s); see Barnes et al. (2001).

The Tukey filter was applied in the lag domain to the HIPASS data to reduce Gibbs ringing. The filter kernel is given by

$$T_f(x) = \begin{cases} 1 & \text{for } |x| < f x_{\max} \\ \frac{1}{2} + \frac{1}{2}\cos\left(\pi \frac{|x| - f x_{\max}}{x_{\max} - f x_{\max}}\right) & \text{for } f x_{\max} \leq |x| \leq x_{\max} \end{cases} \tag{4.17}$$

with maximum lag $x_{\max}$. The parameter f controls the fraction of the lag spectrum that is tapered. The discrete Hanning filter of width M is defined as

$$H_n = \begin{cases} 0 & \text{for } |n| > \frac{M}{2} \\ \frac{1}{2}\left[1 + \cos\left(\frac{2\pi n}{M}\right)\right] & \text{for } |n| \leq \frac{M}{2}\,. \end{cases} \tag{4.18}$$

For the simulations spectra were directly generated. Hence, it was not possible to apply the Tukey filter. Note, that for the EBHIS ringing will not be a problem, because FPGA-based spectrometers have a much better dynamic range than autocorrelators, i.e., the system used at the Parkes telescope.

In the following the influence of different smoothing schemes on the systematic effects will be analyzed. In Fig. 4.28 (top panel) a Gaussian filter kernel of width $w_\sigma = 20\,\mathrm{km\,s^{-1}}$ was applied. This results in a reduction of the bias but the influence of the filter operation is visible, as well — spectral features are "smeared", becoming broader and their amplitude decrease. For a Gaussian spectral line the convolution with another Gaussian results in a wider Gaussian of width

$$w_{\mathrm{out}} = \sqrt{w_{\mathrm{in}}^2 + w_{\mathrm{filt}}^2}\,. \tag{4.19}$$

If the filter kernel is normalized to have an integral value of unity, i.e.

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{4.20}$$

the integral of the filtered function is conserved. For a spectrum this means the total (integrated) flux remains constant after Gaussian filtering. Hence, a spectral Gaussian-shaped line of amplitude $S_{\mathrm{p}}^{\mathrm{gen}}$ has an integral value of

$$S_{\mathrm{p}}^{\mathrm{gen}} \sigma^{\mathrm{gen}} \sqrt{2\pi} = S_{\mathrm{p}}^{\mathrm{rec}} \sigma^{\mathrm{rec}} \sqrt{2\pi}\,, \tag{4.21}$$

which in the ideal case should match the recovered total flux. Assuming the spectral lines are Gaussian-shaped and neglecting any systematics one can correct via

$$w_{\mathrm{gen}} = \sqrt{w_{\mathrm{rec}}^2 - w_{\mathrm{filt}}^2} \tag{4.22}$$

$$S_{\mathrm{p}}^{\mathrm{gen}} = \frac{w_{\mathrm{rec}}}{w_{\mathrm{gen}}} S_{\mathrm{p}}^{\mathrm{rec}}\,. \tag{4.23}$$

[4] Filtering is often used to decrease the noise level, but leads to an increased spectral resolution. Mathematically, the process convolves the signal of interest with a certain function — the so-called filter kernel.

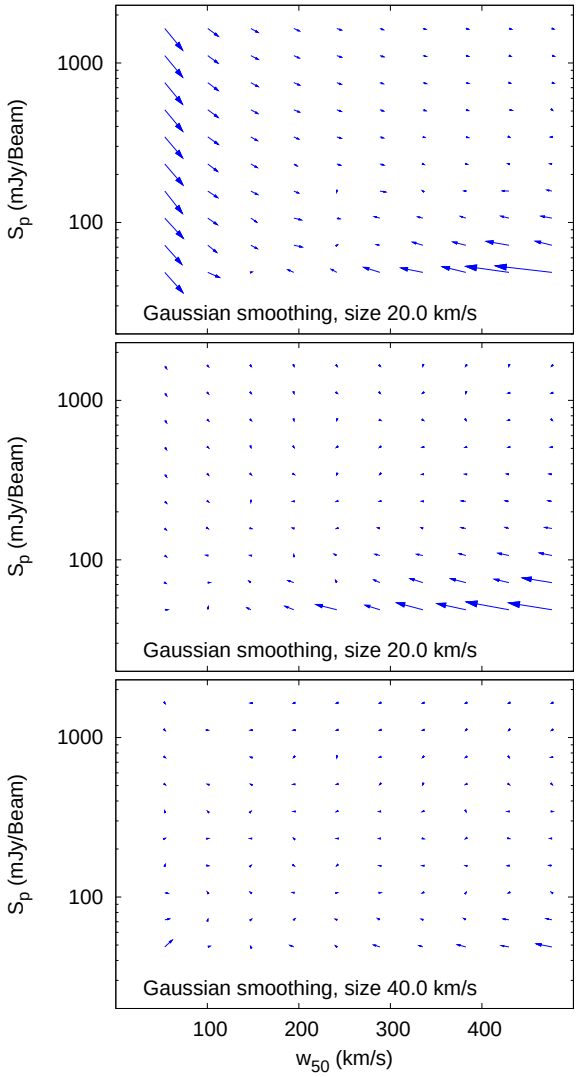


Fig. 4.28: After "smoothing" (filtering with a Gaussian of width $\sigma_{\rm filt} = 20\,{\rm km\,s^{-1}}$) the bias due to the HIPASS parametrization scheme changes significantly. The upper panel shows, that for low widths the peak flux decreases while the measured width increases. This is a direct (expected) consequence of the filtering. Additionally, for low fluxes and large widths there is an underestimation of w_{50}. The first effect can be corrected by applying a scaling correction (see text). The result is shown in the middle panel ($\sigma_{\rm filt} = 20\,{\rm km\,s^{-1}}$). The latter effect is influenced by the size of the filter kernel. In the lower panel $\sigma_{\rm filt} = 40\,{\rm km\,s^{-1}}$ is used leading to less underestimation of the width for large widths and low fluxes.

Fig. 4.28 (middle and bottom panel) shows this rescaling applied after filtering with two different kernel sizes of $\sigma_{\rm filt} = 20\,{\rm km\,s^{-1}}$ and $40\,{\rm km\,s^{-1}}$. In the first case, a bias for the smallest peak fluxes and largest widths is still visible, while for the broader filter kernel the systematic errors have significantly reduced.

For a better comparison of these results with the HIPASS data, additionally Hanning smoothed spectra were investigated. Fig. 4.29 shows the results for Hanning filters of kernel sizes 5, 7, 9, and 11 pixels (the pixels on the edges equal zero). While for small kernel sizes the outcome is similar to the unsmoothed case larger filter widths lead to a situation comparable to the (unscaled) Gaussian smoothing. Unfortunately, the rescaling relations Eq. (4.22) and (4.23) are are only valid for Gaussian filtering and if the spectral lines are Gaussian-shaped. Hence, Hanning filtering is not superior to Gaussian smoothing, at least for Gaussian spectral lines. Consequently, an important result is that the HIPASS spectra are likely affected by the parametrization bias, as well, though due to smoothing, the effect is less severe than for unfiltered data.

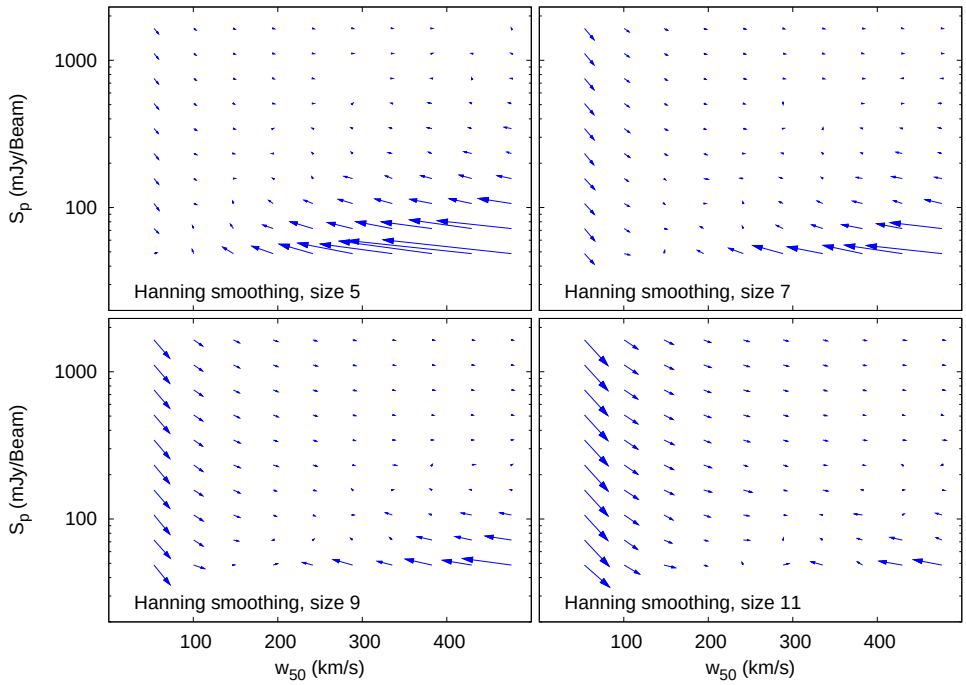


Fig. 4.29: As seen in Fig. 4.28 filtering the spectra can improve the quality of the peak flux and width estimator used for the HIPASS parametrization. As Barnes et al. (2001) pointed out for the HIPASS a Hanning filter was applied leading to a spectral resolution of 26.4 km/s. The plot shows the results for for different Hanning kernels of size 5, 7, 9, and 11 pixels (corresponding velocity root-to-root widths: 66, 92, 119, and 145 km/s). Although all plots reveal less bias than the unsmoothed version (Fig. 4.27) the Gaussian smoothing provides — at least for Gaussian profile shapes — better results; see Fig. 4.28.

Finally, it is important to note that Gaussian profiles account only a for fraction of observed shapes. Due to the different galaxy types, morphologies (possibly warped), sizes, and inclination angles the spectral lines measured have a variety of shapes. Zwaan et al. (2004) used three different types for their simulations, a Gaussian, a flat-top, and a double-horn profile. Saintonge (2007) followed a different approach by developing a general profile which is analytically known (based on Hermite functions) and evolves with a width parameter. Starting from a Gaussian the profile becomes double-horned for larger input widths. However, the actual shape of the latter is not very realistic, when compared, e.g., to the profiles of the 1000 brightest HIPASS galaxies (BGC; Koribalski et al. 2004). Another disadvantage is that two arbitrarily chosen scales enter the definition of the profile function, which makes it hard to use it for analytical fitting algorithms. Furthermore, the broadest profiles are only stretched versions of a certain template, meaning there is no true evolution of shape after a certain threshold width.

Lacking a general sophisticated functional description of profile shapes only the systematics from using a flat-top profile were further analyzed here. Fig. 4.30 presents the outcome for unsmoothed, Gaussian-, and Hanning-filtered data. In the second row the peak fluxes and widths were rescaled again. It turns out that for flat-top profiles this ap-

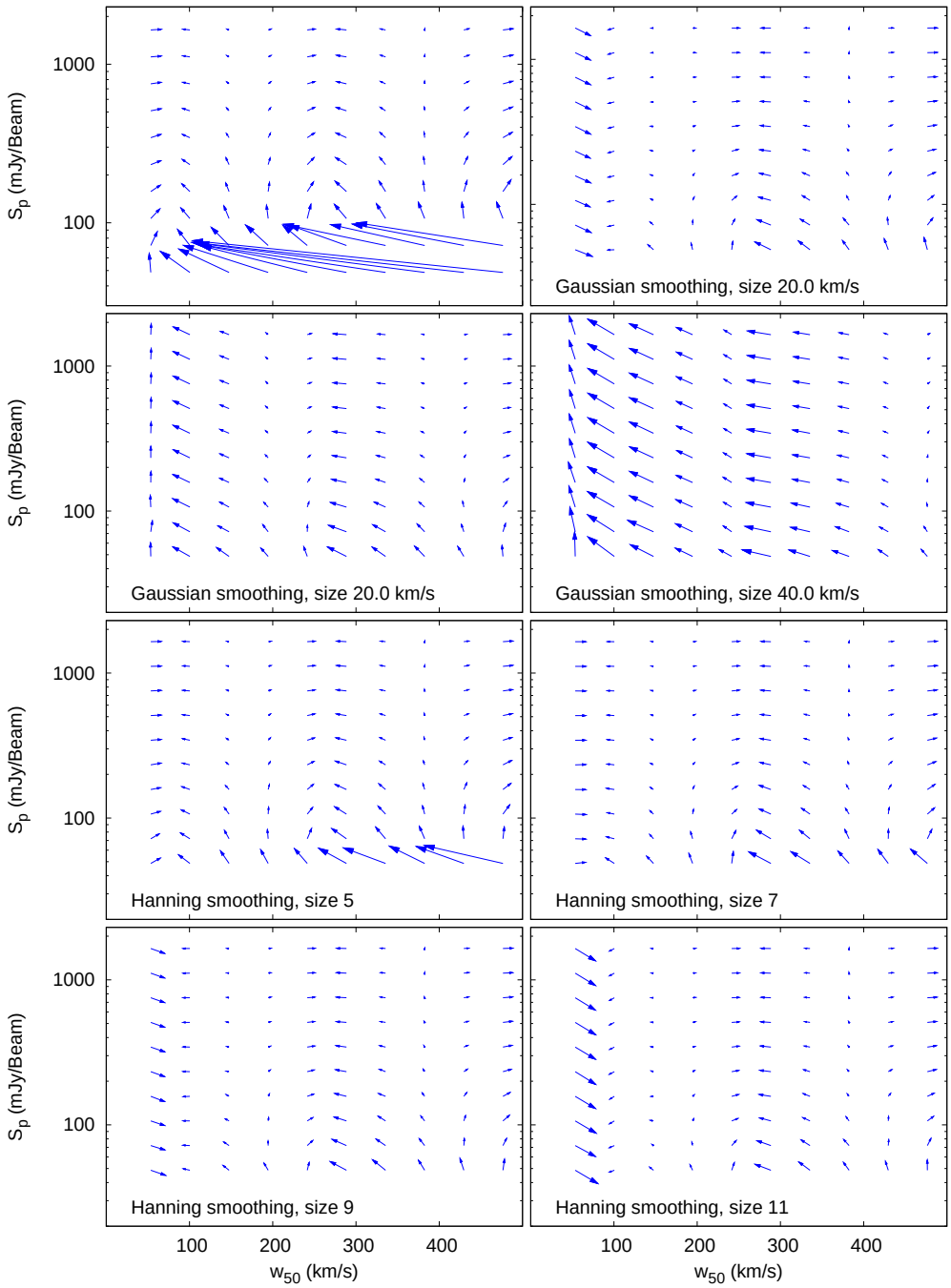


Fig. 4.30: Simulation of the effect of parametrization bias for flat-top profiles. As Fig. 4.27 to Fig. 4.29 the plots show the difference between recovered and generated parameters. The top left panel shows the unsmoothed version, for the top right panel the spectra were Gaussian smoothed. In the second row again Gaussian filtering was applied but with rescaling peak fluxes and widths. The third and fourth row show results for the Hanning-filtered data.

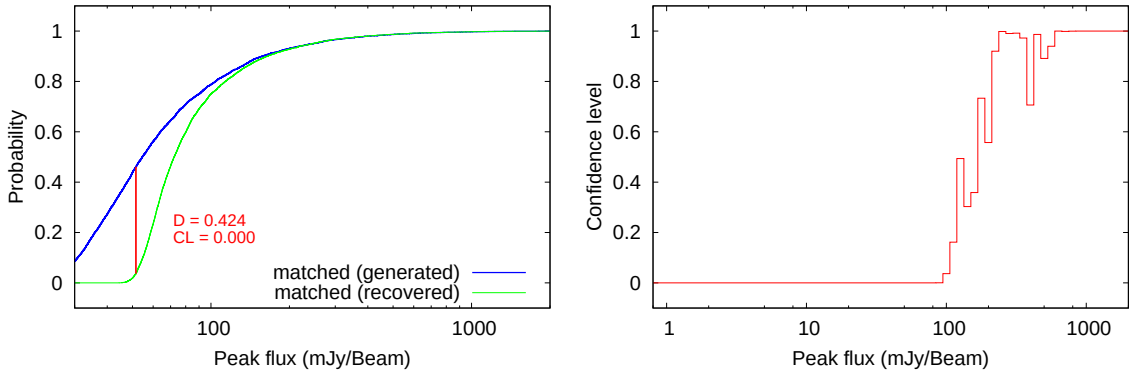


Fig. 4.31: Left panel: Cumulative distribution function of the generated and recovered peak fluxes of matched galaxies. The confidence that both samples are likely originating from the same distribution is zero, rejecting this hypothesis. The right panel shows the confidence CL after thresholding (with respect to peak fluxes) of both samples. Above $S_\mathrm{p} \approx 200\,\mathrm{mJy\,Beam^{-1}}$ the confidence reaches the 100% level.

proach does not improve the parameter estimation, but the bias is even increased compared to the case of simple Gaussian smoothing (top right panel) and to the Hanning-filtered spectra (third and fourth row). This is not surprising, because the rescaling scheme was developed for Gaussian profiles, and is indeed invalid for any other line shape.

In summary, there is not yet a generally valid approach to deal with the parametrization bias. Direct profile fitting would provide probably the best results once a general functional description could be found. Moderate filtering of the spectra improves the estimation but residual systematic effects are measured. The latter fact is important for the HIPASS data, though it was not analyzed in detail by Zwaan et al. (2004).

4.5.3 The Kolmogorov-Smirnov test

In the previous sections possible bias and selection effects were described and quantitatively analyzed. In order to study their impact on some of the recovered distributions, e.g, flux and mass properties, the Kolmogorov-Smirnov (KS) test was applied. It can be used to determine whether two samples likely do originate from the same distribution function and has the advantage that it is non-parametric and is working on unbinned data as well. The KS test calculates the cumulative probability distribution of both samples and searches for the largest difference d.[5] Depending on d and the sample sizes n_1 and n_2 the confidence CL can be computed whether the samples follow the same distribution. For further details see Press et al. (1992).

Fig. 4.31 (left panel) shows the cumulative distribution for the simulated and recovered peak fluxes of matched galaxies. The computed confidence is CL = 0 which means that the hypothesis of equal distributions can be rejected at the 100% level. This result is not surprising, as in the lower peak flux regime the survey is biased visible on the different curve shapes — the recovered cumulative distribution increases only after about $50\,\mathrm{mJy\,Beam^{-1}}$.

[5] One might also think of different more complex statistics using the area between the two probability curves or the mean squared distance.

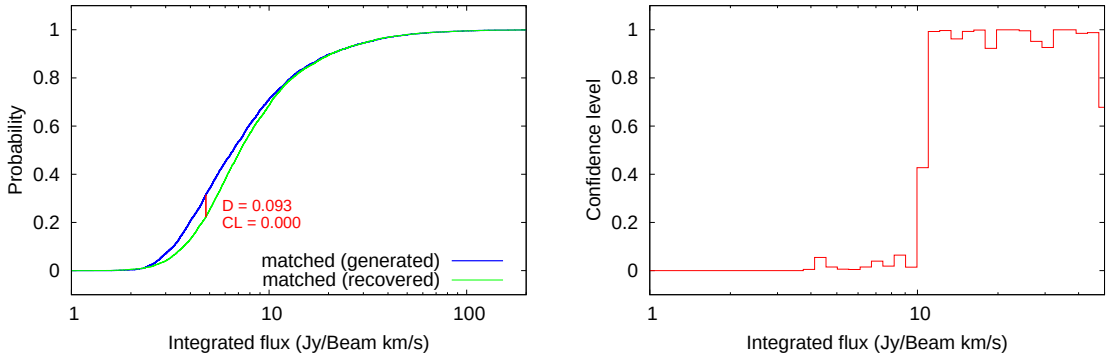


Fig. 4.32: Left panel: Cumulative distribution function of the generated and recovered integrated fluxes of matched galaxies. The confidence that both samples are likely originating from the same distribution is zero, rejecting this hypothesis. The right panel shows the confidence CL after thresholding (with respect to integrated fluxes) of both samples. Above $S_{\mathrm{int}} \approx 10\,\mathrm{mJy\,Beam^{-1}\,km\,s^{-1}}$ the confidence reaches 100%.

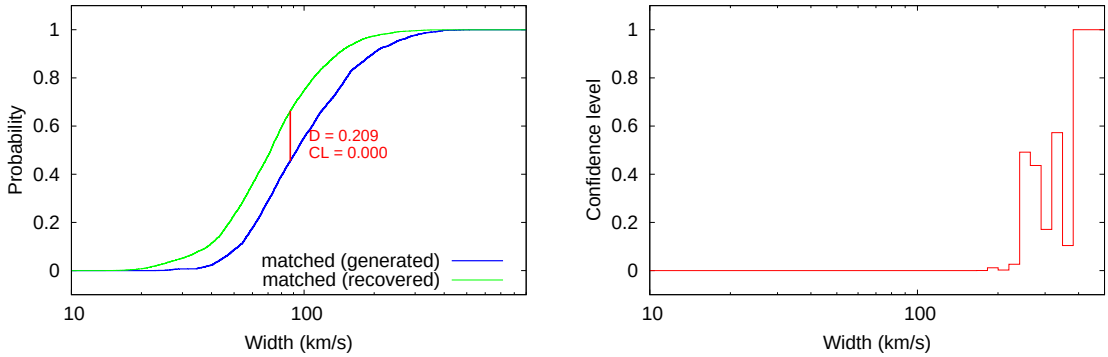


Fig. 4.33: Left panel: Cumulative distribution function of the generated and recovered velocity profile widths of matched galaxies. The confidence that both samples are likely originating from the same distribution is zero, rejecting this hypothesis. The right panel shows the confidence CL after thresholding (with respect to width) of both samples. Above $w_{50} \approx 400\,\mathrm{km/s}$ the confidence approaches 100%. However, as only a small fraction of all sources have such large widths (compare Fig. 4.19 to 4.21) this result is not very reliable.

It is important to note, that this offset is not due to the incompleteness of the survey, as only matched pairs are compared, but due to the overestimated S_{p}. The Figure reveals that for higher peak fluxes the distributions become equal. Introducing a flux threshold could therefore increase CL. The right panel of Fig. 4.31 shows CL as a function of peak flux threshold which was applied to the samples before performing the KS test. Obviously above a value of $S_{\mathrm{p}} \approx 200\,\mathrm{mJy\,Beam^{-1}}$ the influence of the systematic effects described in Section 4.5 becomes less pronounced. Here, CL reaches the 100% level.

The same behavior is observed for the integrated fluxes; see Fig. 4.32. Here, above $S_{\mathrm{p}} \approx 10\,\mathrm{mJy\,Beam^{-1}\,km\,s^{-1}}$ the confidence reaches 100%. For completeness also the results of the KS test for the velocity width w_{50} (Fig. 4.33) are shown. The latter never

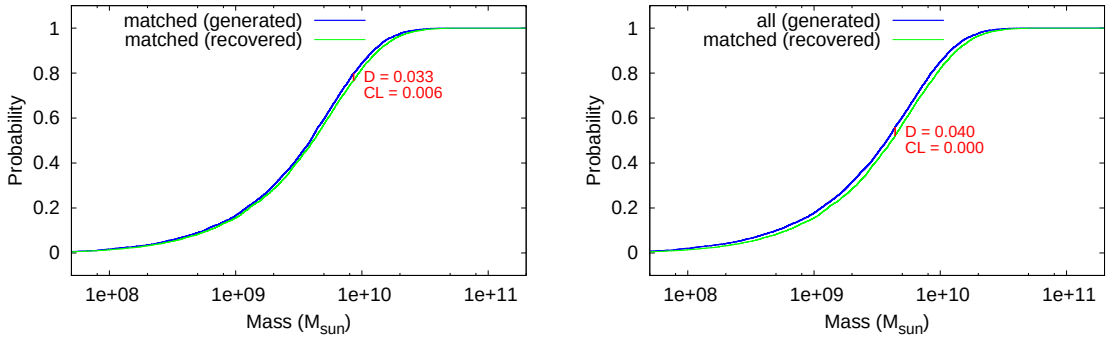


Fig. 4.34: Left panel: Cumulative distribution function of the generated and recovered H I masses of matched galaxies. The confidence that both samples are likely originating from the same distribution is less than 1%. The right panel shows the cumulative distributions for all generated and recovered galaxies, including non-detections and false-positives. The hypothesis that both distributions are equal is rejected.

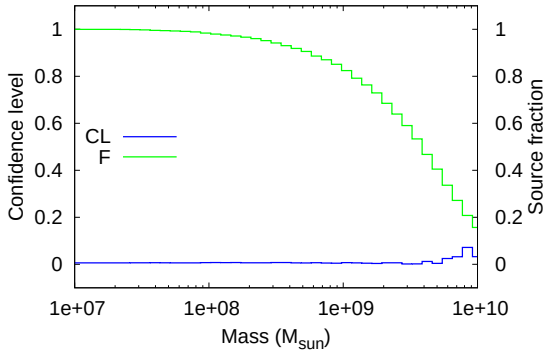


Fig. 4.35: The confidence CL after thresholding (with respect to masses) of both samples. As the mass is a function of distance there is no threshold value which significantly increases CL. The green curve marks the fraction of sources remaining after applying the thresholds.

reaches a satisfactory CL, which is due to the widths being underestimated for low peak fluxes over the complete profile width range. Again, the low CL values for small $S_{\rm int}$ and w_{50} is caused by systematic effects and not incompleteness of the survey; see Section 4.5.

The situation is slightly different for the mass distributions. Fig. 4.34 contains the cumulative distributions for the masses of matched galaxies (left panel) and all generated/recovered galaxies (right panel), respectively. In both cases the distributions are found to be different. However, the visual matching is much better than in the previous cases. Also the largest difference, d, is much smaller. That CL is nevertheless small is a consequence of the KS test being very sensitive when having a large sample. By using only a half (a tenth) of the sources the CL equals 18% (88%). A simple mass threshold is not appropriate to improve the confidence; see Fig. 4.35. This can be understood, as the mass is a function of distance, hence, the mass distribution is not the result of a scaling of the integrated flux distribution. However, when peak or integrated flux thresholds are applied CL increases significantly as before; see Fig. 4.36

One problem of applying (one-dimensional) flux thresholds is that the fraction of sources at the CL = 100% level has dropped to below 20%. This is a consequence of the completeness function being better described two-dimensionally. Therefore, also two-

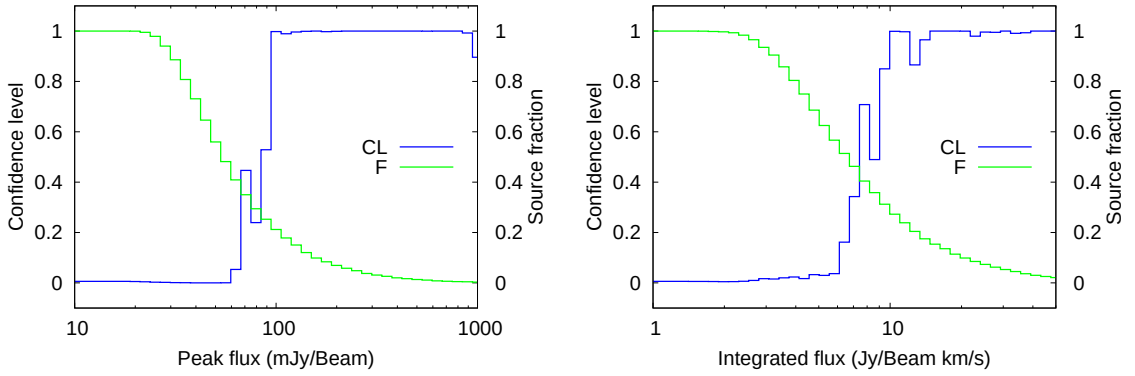
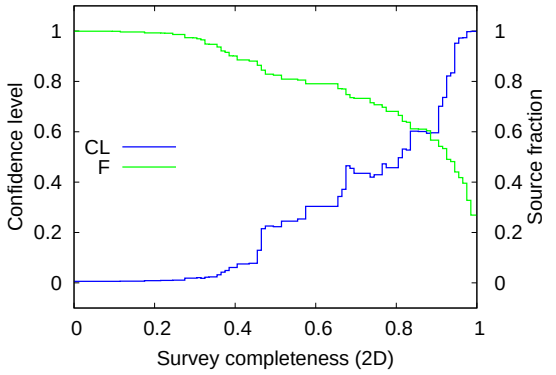


Fig. 4.36: Applying flux thresholds to the H I mass samples increases the confidence CL (blue line). The green line marks the fraction of sources remaining after applying the thresholds. However, using 1D thresholds neglects the correlation of peak and integrated flux completeness as discussed in Section 4.4.1. Therefore, at the threshold flux values where the confidence CL reaches the 100% level the catalog only contains about 20% of the sources.

Fig. 4.37: The plot shows the confidence CL (blue line) after applying two-dimensional completeness thresholds to the H I mass samples. Although the CL reaches the 100% level only after the completeness approaches $\sim$ 0.9 the fraction of sources remaining in the catalog at that point is still about 50%, which is much higher than after applying one-dimensional flux thresholds; see Fig. 4.36. The green line marks the fraction of sources remaining after applying the thresholds.



dimensional thresholds were applied by first, binning the data in the peak–integrated flux plane, and second, mapping the bins to their completeness value (see Fig. 4.12. This results in a list of sources as a function of the two-dimensional completeness, which in turn can be used to apply appropriate thresholds. Fig. 4.37 that the distributions can be considered equal at a completeness of about 0.9, at which about 50% of the total number of sources remain in the catalog (about 30% of sources lie above the completeness level).

4.6 Outlook

While the main aim of the simulations were to analyze the impact of RFI signals on the data reduction pipeline and on the detection rate, it is certainly clear that for the final analysis of the EBHIS data the actual completeness function needs more investigation. This would need to involve various profile shapes into the simulation. Using the Gamma

test as source finder algorithm probably gives similar detection rates for box- or double-horn profiles. Nevertheless, the bias- and selection effects which were observed likely show different characteristics for other emission line shapes. Furthermore, by directly inserting artificial sources into the measured spectra instead of completely generating artificial data would also account for the complete reduction pipeline. During the setup of the simulations it was also not clear what the survey parameters, i.e., velocity resolution, noise levels, would be. This should have minor impact though.

The simulations revealed how important a proper RFI mitigation scheme is for the source detection rate. As soon as interferences enter the data the number of galaxies found drops significantly. The RFI detection software which was developed in the framework of this thesis led to results, very similar to the ideal case. A second major issue is the quality of the source finding algorithm. The Gamma test proved to be more than satisfactory.

During the analysis of the data several bias and selection effects were found and could be explained. Especially, the problems introduced via the HICAT parametrization scheme should be kept in mind. It might be worthwile to develop better measures for the profile widths, having the largest impact on physical important quantities drawn from the line profiles (the peak flux is not very meaningful and the integrated flux is unaffected). Also, the selection effect which slightly increases the mean integrated flux of fainter sources should be considered. Eventually, one could correct for this effect.

CHAPTER 5

Preparing the survey — Test observations

Several test measurements have been conducted not only to test the new multi-beam receiver and the FFT spectrometers but also to prove the data reduction tasks to be fully functional. The latter were already described in detail in Chapter 3. However, some of the calibrational issues are more specific to the data obtained making it more meaningful to discuss them here. As some of the reduction steps and the receiver are still in a test phase the results must be considered as preliminary. Section 5.1 briefly describes both the flux and bandpass calibration, as well as the stray-radiation correction, which was derived from the LAB data (Kalberla et al. 2005), lacking the complete EBHIS survey data at this point. Up to now two test observations with the new multi-beam receiver took place. The first measurement (November, 20/21 and 23/24, 2007) was intended to check the overall system quality by performing *Allan* tests, measurements of the system temperature, and bandpass stability. These data are discussed in Section 5.2. To test the receiver stability on longer time scales and check the data reduction pipeline a second observation campaign took place on Dec, 27-30, 2007. Two sources of astronomical interest, NGC2403 and the recently detected (Ryan-Weber et al. 2008, in H I;) dwarf galaxy Leo T, were measured. Furthermore, both objects were re-observed with the (older) 21-cm single-beam receiver but using the same FFT spectrometer. This enables the direct comparison of the receiver performances. All aspects regarding the data quality for these map observations are discussed in Section 5.5. Finally, the compiled data cubes were used to analyze the sources for basic physical parameters (Section 5.6).

5.1 Calibration

5.1.1 Intensity calibration

The intensity calibration was applied according to Section 3.3: the calibration signal produced by the noise diode can be absolutely calibrated using the S7 measurement to compute the effective temperature $T_{\rm cal}$ of the diode. This is used for the subsequent intensity calibration of the map observations.

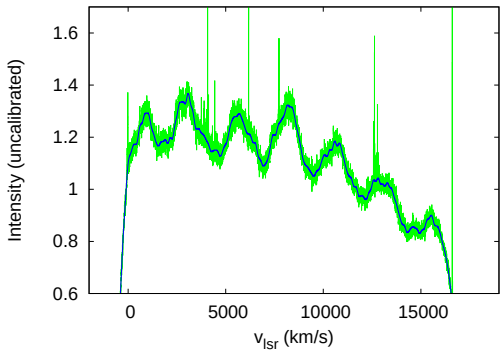


Fig. 5.1: In order to have a good estimate for the gain curve (necessary for robust computation of calibration factors) an iterative algorithm is used; see text for details. Outliers as RFI peaks are excluded. The green line shows a spectrum, the solid blue curve is the bandpass curve obtained.

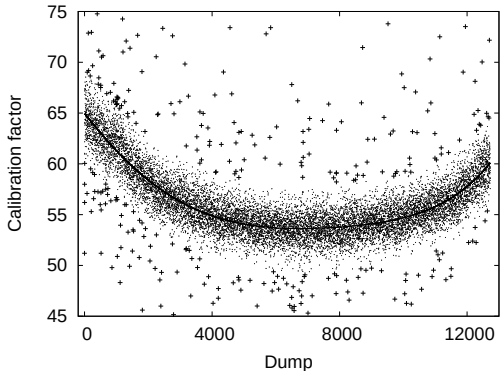


Fig. 5.2: Time dependend behavior of the calibration factor shown for one of the Leo T observations (Scan 2621, ch1, ph0/2). The solid line shows the best fit after applying the algorithm described in the text. Data points in excess of 5σ in the residual are flagged as "bad" (crosses, good data are plotted as dots). These dumps can be excluded in the subsequent data processing.

A major problem before applying Eq. (3.5) is the calculation of the mean $\langle S_{\mathrm{sig,ref}}(v)\rangle$ of the spectra being very sensitive to RFI signals and baseline instabilities. A more robust algorithm was developed and provides good results for the Effelsberg data. In a first step a good estimate of the bandpass curve is calculated by iteratively (1) low-pass filtering the data, (2) calculating the residual, (3) search for outliers, i.e., peaks above 5σ in the residual, (4) subtract these outliers from the initial data, and repeat the procedure starting from step (1). After five iterations the final filtered spectrum is a smooth gain curve insensitive to (narrow-band) RFI peaks and low- to medium-intense emission lines (strong sources as S7 have a small but neglectable impact). Fig. 5.1 shows the derived gain curve for one of the Leo T spectra.

The second step is to calculate for each dump the calibration factor, T_{cal}; see Fig. 5.2. The overall changing of the system temperature is clearly visible. Still, single dumps exhibit large differences to the mean behavior. Therefore, a best (polynomial) fit is computed (solid line). Again an iterative scheme allows to flag dumps having calibration factors above 3σ for subsequent fitting. After five iterations a stable solution is reached. All "bad" factors (crosses) are marked and can be neglected for the further data processing. In order to calculate the system temperature from the calibration factors the fit values are used for each dump.

5.1.2 Stray-radiation correction

In Section 3.4 the stray-radiation (SR) correction scheme was described. Optimally, the sky model is obtained from data observed with the same measurement configuration, i.e., the

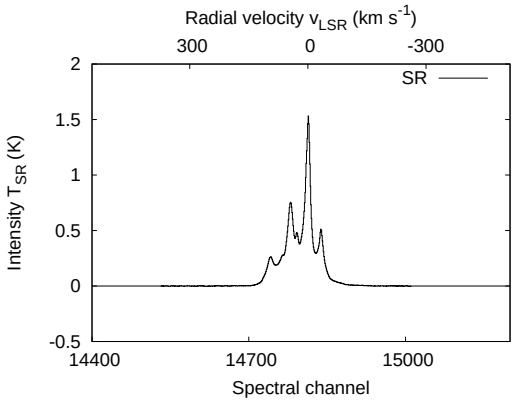


Fig. 5.3: Example of the stray-radiation in the direction of Leo T (Scan 2621). The spectrum was calculated using the LAB data for each of the map positions (subscans). As EBHIS data are not yet available the SR model must be considered preliminary. However, it should provide a satisfactory approximation.

same telescope, receiver, data reduction etc. However, no such data exist yet for the new multi-beam receiver. Preliminary, the sky model obtained from the LAB survey (Kalberla et al. 2005) is used but incorporating a theoretical aperture pattern for the Effelsberg beam, leading to a good approximation though having a worse spatial resolution.

5.1.3 Bandpass calibration

In Section 3.5 several methods were introduced to calibrate the bandpass (gaincurve) of the receiving system. Although the least-squares frequency switching method (LSFS) has advantages, it is not yet available at the 100-m telescope at Effelsberg, because changes to the hardware would be necessary in order to implement a LO switching scheme having more than two different frequencies. Therefore, the test observations were performed in in-band frequency switching. The exact frequency setup varied during the measurement due to testing purposes and/or to give consideration to the source velocity.

In frequency switching two LO setups ($\nu_1 = \nu$ and $\nu_2 = \nu + \Delta\nu$) lead to a "shifting" of the astronomical line of interest in the channel representation of the spectra. Converting the channel number to sky frequency or radial velocity must take this extra frequency shift $\Delta\nu$ into account. As only spectral features entering the system before downconversion (mixing with the LO signal) are shifted, the gaincurve, usually defined by the IF properties, is more or less the same in both switching phases. Hence, Eq. (3.4) can be directly applied.

All spectral features present in the reference spectrum will, however, appear inversely in the residual spectrum. This would mean a loss of information, as only half of the spectra (referred to as signal spectra) would be taken into account. Usually, one overcomes this shortfall by interchanging signal and reference spectrum. Obviously the "inverse" feature will then appear on the opposite side of the emission line.

5.2 First test observations and system quality

To evaluate the current status of the new multi-beam 21-cm receiving system test measurements were performed on November, 20/21 (hereafter referred to as *tm1*) and November, 23/24, 2007 (*tm2*). Three FPGA spectrometers were used in parallel, one narrow band unit (1024 spectral channels, 50 MHz) for the detection of the emission of the Milky Way (*B1*; Stanko et al. 2005), a broad band (8k/16k, 100 MHz) unit originally developed for the

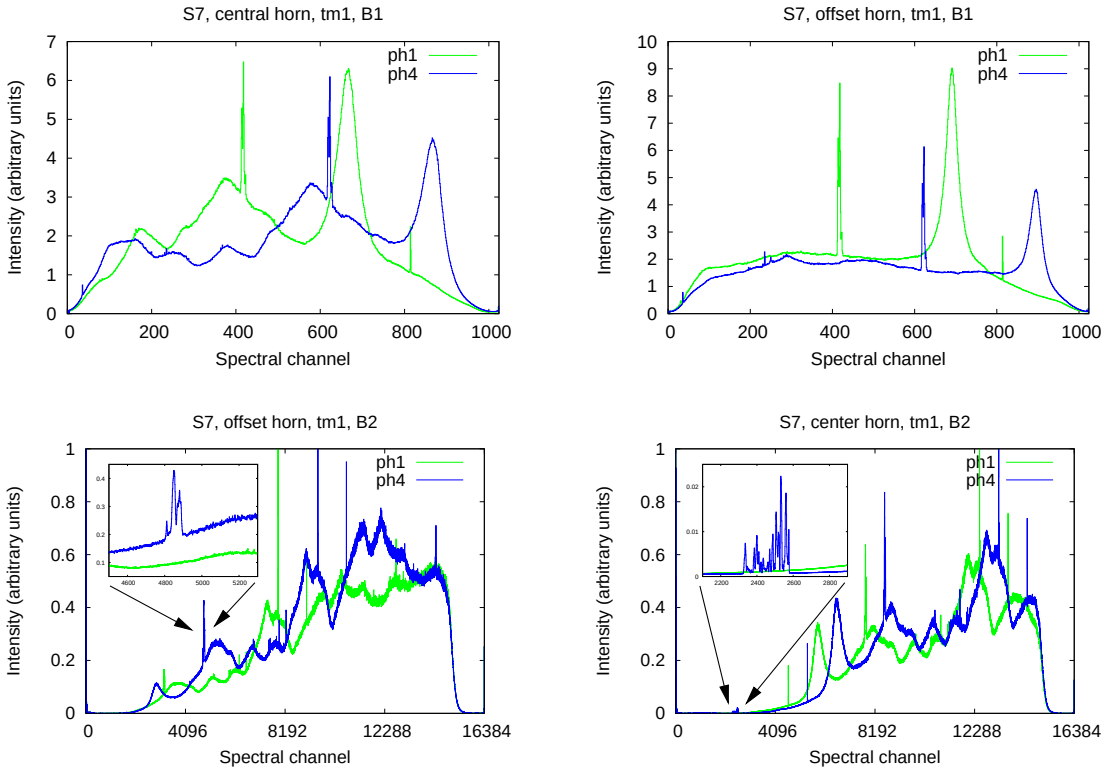


Fig. 5.4: Uncalibrated S7 spectra as observed during the test measurement *tm1*. **Upper left:** Scan 6698, *B1*, ch1, central feed. Most of the spectral features, i.e., the bandpass shape follow the frequency shifting. This makes a bandpass calibration using a standard frequency switching scheme impossible. **Upper right:** Scan 6699, *B1*, ch1, offset feed. **Lower left:** Scan 6699, *B2*, ch1, offset feed. The inset shows the signal of interest, the calibration source S7. **Lower right:** Scan 6690, *B2*, ch1, central feed. The inset shows a very prominent RFI signal located beyond the stopband of the filters which hence should not be detectable. It is caused by terrestrial digital radio (DAB) at a frequency of 1450 MHz.

APEX telescope covering the whole frequency band of the receiver (*B2*; Hochgürtel et al. 2008; Klein et al. 2006), and a 16k channel spectrometer (20, 100 MHz) which is currently the standard spectrometric backend at the 100-m telescope (*B3*; Klein et al. 2005). Setup *B2* is equivalent to the one which will be used during the proposed Effelsberg survey.

Due to the limited number of backends, feeds could be tested only consecutively. In total 14 channels are available. To have data from all horns, the backends were successively attached to all channels, the inner horn (circular polarization) and the six outer beams (linear polarization) each of them providing two polarization channels (referred to as ch1 and ch2). All measurements were carried out in frequency switching mode (inband) leading to four measurement phases ph1 to ph4 (ph1: sig+cal, ph2: ref, ph3: sig, ph4: ref+cal).

Figure 5.4 shows in example several snapshot spectra of the calibration source S7 for the central horn and one of the outer horns as observed during tm1 with *B1* (upper panels) and *B2* (lower panels), respectively. The data were not calibrated with respect to bandpass

profile and intensity. In fact, a bandpass calibration using a standard frequency switching scheme (so-called *folding*) is not possible because it can only correct for the intermediate frequency (IF) gain curve. Here, most spectral features (i.e., the bandpass curve) follow the frequency switch meaning they are already present before mixing the high frequency (HF) signal to IF. This effect is clearly visible in the upper left panel showing the central feed, but is also present in the outer horn channels (e.g., lower right panel). Unfortunately, these signals entering the frontend (origin at the moment unknown) introduce a very complex gain curve, making it almost impossible to calibrate the bandpass by other means, e.g., using polynomial fits (baselining). Note, that the lower panels show spectra recorded with different local oscillator (LO) settings.

Another prominent feature is a feed resonance line (Keller, priv. com.) producing a strong broad Gaussian-like signal (right edge of the bandpass for *B1* and left edge for *B2*). The insets in the lower row show the emission line of the source of interest, S7 (left panel), and a very prominent RFI signal caused by terrestrial digital radio (DAB) at a frequency of 1450 MHz. The latter is far outside the bandpass filter but still detectable, showing that the stopband suppression of the frontend filters is not sufficient. In fact, some of the temporal variations of the baseline of the central feed (see below) can be attributed to the very intense linear polarized interference signal of the DAB.

The subsequent paragraphs compile a more quantitative analysis of the receiver.

5.3 Receiver stability

The stability and sensitivity of the receiver was of highest interest. To evaluate this quantitatively, the so-called *Allan*-plots can be used. A qualitative analysis is also feasible, using greyplots to visualize the dependence of the bandpass shape and gain of time.

5.3.1 *Allan*-plots

For receiving systems in astronomy it is very important that the noise of the system decreases with the square root of the integration time. For an ideal radio receiver the radiometer equation

$$P(t) \sim \frac{1}{\sqrt{t \cdot \Delta f}} \sim t^{-0.5} \tag{5.1}$$

is applicable, stating that the noise power $P(t)$ decreases with the square root of the integration time t and bandwidth Δf. Practically, each system suffers from instabilities on a certain time scale t_A, yielding a divergence from the theoretical behavior. It is obvious, that a system should be designed in a way to maximize t_A. One possibility to measure the stability is to compute an Allan-plot. Several hundred or thousands of short (in time) spectra or "spectral dumps" are recorded and subsequently integrated (e.g., in steps of 2^n) to evaluate the noise on the time scale t_n. The noise level can be measured in various ways: simply compute the RMS for the time series in a given spectral channel, or calculate the baseline RMS using polynomial fits, etc. To deal with the complex bandpass shape, first, a mean spectrum was calculated and each individual spectral dump divided by this mean spectrum. Accordingly, the baseline RMS is then directly given by the standard deviation within a certain spectral range. Note, that this procedure does not

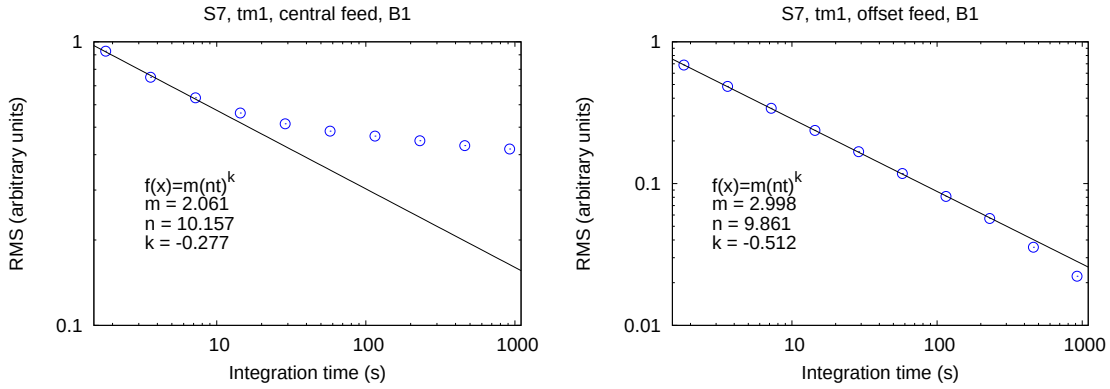


Fig. 5.5: Allan-plot calculated for spectra of the calibration source S7, as observed during the test measurement *tm2*. The left plot shows the noise behavior of the central feed (scan 6698, *tm1*, B1, spectral channels 450 to 550) for 512 dumps of 0.45 s integration time each (separated by 2 s). To handle the complex shape of the gain curve a mean spectrum was computed and each individual spectrum was divided by this mean spectrum prior to the calculation of the Allan plots. The right panel shows the Allan-plot of one of the offset feeds (scan 6699). While the central feed diverges from the theoretical behavior even after $t_A \sim 10$ s the offset feed reveals a very good stability/sensitivity up to the maximum measured integration time of about 1000 s.

have any affect on the noise properties of the data. The mean spectrum can be treated as an arbitrary constant[1].

In Fig. 5.5 Allan-plots for the the central feed (left panel) and one of the offset feeds (right panel) are shown. The total sample consists of about 900 spectra (0.45 s dump time, recorded every 2 s due to the measurement with four phases) which are used to calculate a mean spectrum. For the actual Allan-plot only 512 of them are used. The central feed has a noise property incompatible with the radiometer equation Eq. (5.1), with $t_A \sim 10$ s. On the other hand, the offset feed follows the expected behavior even after integration of about 1000 s.

5.3.2 Bandpass instabilities

The reason for the short Allan-time t_A of the central feed (circular polarization) is a very instable gain curve probably caused by the DAB interference signal (linearly polarized) pumping power into the system. The offset horns are much less sensitive as they measure linear polarization only. The instability can be visualized by using grey-plots, showing the development of the spectra with time; see Fig. 5.6. Again, a mean spectrum was computed and each spectral dump was divided by it to improve the dynamic range of the grey-plots.

[1] The only drawback is, that the RMS values referring to the longest integration times are slightly underestimated. This bias is produced if the overall dataset, for which the mean spectrum was calculated, is not much larger than the sample which was used to calculate the Allan plot.

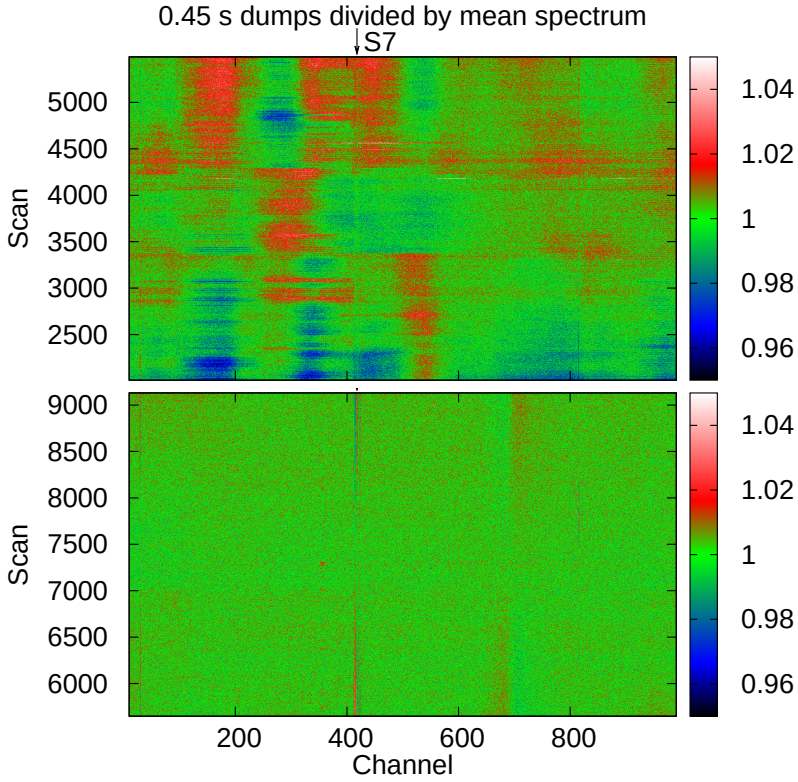


Fig. 5.6: To test the temporal bandpass stability of the central feed (top panel) and one of the offset feeds (bottom panel), the spectral energy distribution of the calibration source S7 (*tm1*, *B1*, Scan 6698/6899, ch1, ph1). Because the overall bandpass curves of both feeds are complex, each individual dump was divided by a mean spectrum. The top panel reveals time- and frequency-dependent residuals showing the instability of the central feed. The offset horn does not show this problem.

5.4 System temperature and noise

In order to calculate the system temperatures, T_{sys}, not the switched spectra were used to obtain a reference gain curve, but a baseline fit was applied (with appropriate windows set to exclude the S7 emission line). This was necessary to deal with the very complex and time-dependent baseline shape even after application of the frequency switching scheme.

Of further interest is the baseline noise (RMS). It is determined from the residual (calibrated) spectrum by computing the standard deviation of a spectral portion of the baseline. Practically, this is not easy to fulfill, as often RFI signals degrade the spectra. Furthermore, the bandpass shape might hardly be described by a low-order polynomial. In fact, both problems make the measurements of the baseline noise very errorprone.

Figure 5.7 shows calibrated S7 spectra (*tm2*, *B3*) for the central feed (left circular, hzl) and three of the offset feeds (2-4, vertical, h2v, h3v, h4v) each for two phases (ph2, ph4). The integration times, as well as the measured system and noise temperatures are given in Table 5.1. The panels in the figures contain the (uncalibrated) raw spectra (red line), the polynomial fit (green line), and the calibrated spectrum (in terms of antenna temperature T_{A}).

As expected, the central feed reveals a very high system temperature $T_{\mathrm{sys}} \sim 65\,\mathrm{K}$ which is a direct consequence of the bandpass instabilities discussed in Section 5.3. The offset feeds have much lower system temperatures between $\sim 20 \ldots 25\,\mathrm{K}$. The RMS noise values are good and in the expected range, but show large scatter due to the complicated bandpass shape and probably residual RFI contamination (it is not easy to distinct both). Both features might not be well approximated by the polynomial fit leading to increased

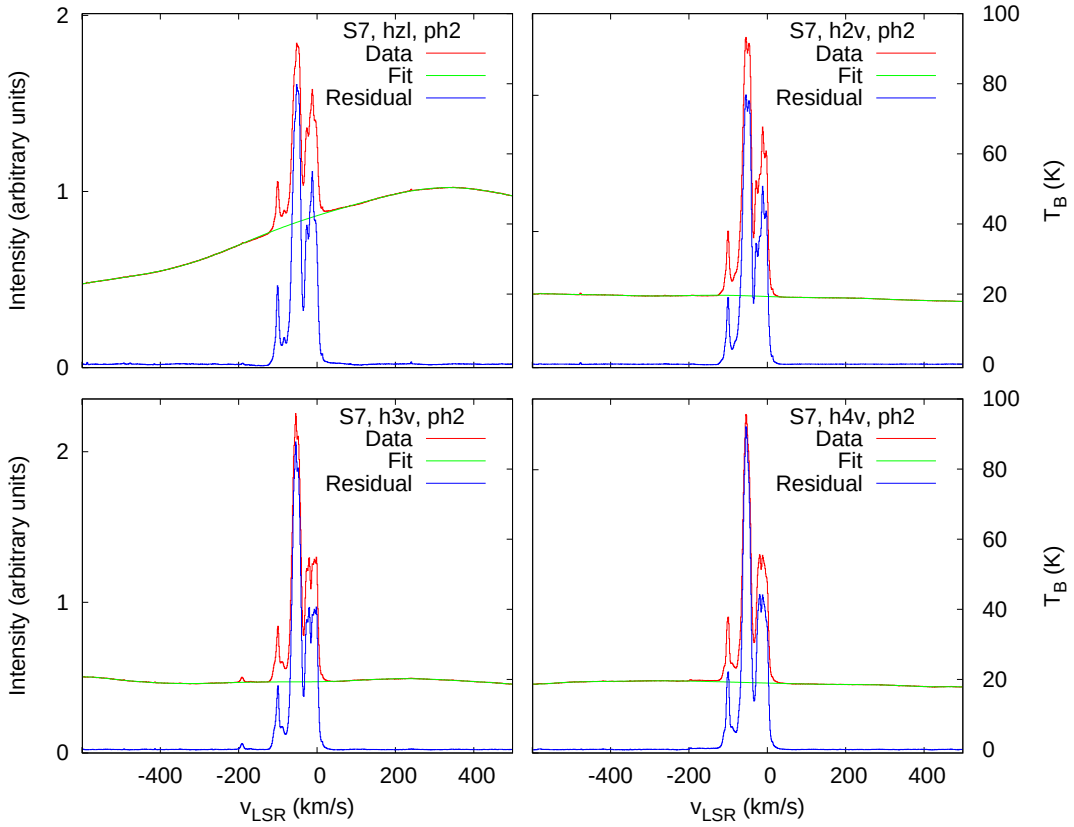


Fig. 5.7: Calibrated S7 spectra obtained with the central feed measured during the test *tm2*. The upper row figures shows the left-hand circular polarization channel of the central feed (left panel) and one of the linear polarization channels of the second feed (right panel) after 650 s of integration (*tm2*, *B3*, Scan 6813, ph2). The lower row displays one linear polarization channel of feed three (left panel) and feed four (right panel) after 150 s integration (*tm2*, *B3*, Scan 6814, ph2). A polynomial fit (green line) was applied to the raw (uncalibrated) data (red line). The calibrated spectrum was calculated (blue line) as described in the text.

RMS values. However, the properties of the receiver — especially for the offset feeds — are very promising.

5.5 Testing the long-term behavior — the observation of NGC 2403 and Leo T

A second measurement campaign was performed during Dec, 27-30, 2007, to test the long-term properties of the new receiving system. Two sources of astronomical interest NGC 2403 and the recently detected (in H I, Ryan-Weber et al. 2008) dwarf galaxy Leo T were mapped several times. To compare the 21-cm multi-feed with the older single-beam receiver both objects have been re-observed in June/July 2008, though with shorter integration times per position (60 s vs. 120 s).

Table 5.1: System temperatures and RMS-noise of the calibrated S7 spectra for the 7-feed receiver as measured during *tm2*. The noise temperatures are the RMS values converted to a spectral resolution of 1 km/s. RMS was measured two times, to the left and to the right of the spectral line. Depending on the specific bandpass shape and possibly RFI signals the values have a scatter of up to 100%. Nevertheless, the system temperature $T_{\rm sys}$ is estimated with higher accuracy, as the determination is less affected by the quality of the baseline fit. The central horn has a much higher system temperature — a direct consequence of the bandpass instabilities discussed in Section 5.3. For the outer horns $T_{\rm sys}$ has good values which are lower than for the old 18/21-cm receiver at Effelsberg.

Horn	Phase	$\tau_{\rm int}$ (s)	$T_{\rm rms1}$ (mK)	$T_{\rm rms2}$ (mK)	$T_{\rm sys}$ (K)
z	2	650	92	131	64.7
z	4	650	93	128	65.3
2	2	650	61	22	21.4
2	4	650	61	23	21.8
3	2	150	56	57	23.3
3	4	150	50	57	23.7
4	2	150	55	37	24.2
4	4	150	57	37	24.6

Table 5.2 contains observational parameters as map size, spatial resolution, integration time per position, etc. As the first test revealed bandpass instabilities and very high system temperatures within the central feed (see Section 5.2, one of the outer (linearly polarized) feeds was used during this observation period. Another goal of the measurements was to test the data reduction software and pipeline, which so far had only be used on simulated data. Furthermore, the flux and bandpass calibration, as well as SR correction — which were neglected in the simulations — needed to be checked.

From Table 5.2 follows that for NGC 2403 only one complete coverage was obtained with the multi-feed, while for Leo T exist four. The spatial resolution of the maps was $3'$. The calibration source S7 was measured several times before maps were observed.

Fig. 5.8 shows an uncalibrated spectrum of the source S7 as measured with the new multi-beam receiver for two of the four switching phases (ph1, ph2 — the signal and reference phase without calibration diode fed in). Note, that the gain curve stays more or less fixed in spectral channels, while the signals entering through the antenna are shifted (meaning the conversion to LSR velocities is different for both phases). The upper row contains all 16k spectral channels, revealing lots of strong RFI signals all over the band. The gain curve inhibits a very complex shape probably caused by ripples the origin of which is not yet very clear; see also Section 5.2. It might be possible that standing waves between the primary and secondary mirror produce them. During the tests the focus was shifted by $\pm\lambda/8$ and both spectra were added. It seems that by this procedure a single wave mode of the distortions vanished. However, further investigation need to be performed. The lower panels of Fig. 5.8 show a zoom-in for a better visualization of the

Table 5.2: Observational parameters of the NGC 2403 and Leo T measurement using both the multi-feed and (older) single-beam 21-cm receiver. The angular sampling interval of the maps was 3′. Each position within a map refers to a specific subscan number having the integration time, $t_{\rm int}$. The spatial coordinates stated for the map scans are first given as the range of coordinates of the first to the last subscan (for completeness). Furthermore, the rectangular limits of positions in the specific scan are presented.

Scan	Type	Source	Receiver	Date	Time	$t_{\rm int}$	Subscan		Spatial range (first subscan to last)				Spatial range (rectangular limits)			
							From	To	From		To		From		To	
						[s]			Long [°]	Lat [°]	Long [°]	Lat [°]	Long [°]	Lat [°]	Long [°]	Lat [°]
2532	calib	S7	multi-beam	2007-12-27	18:53:57	30			131.973[g]	-1.179						
2539	ref	NGC2403	multi-beam	2007-12-27	19:23:58	120			114.213[e]	65.598						
2540	map	NGC2403	multi-beam	2007-12-27	19:27:03	120	1	53	112.993[e]	65.085	113.921	65.183	112.993	65.085	115.434	65.183
2545	ref	NGC2403	multi-beam	2007-12-27	21:59:07	120			114.212[e]	65.599						
2546	map	NGC2403	multi-beam	2007-12-27	22:02:00	120	29	50	114.736[e]	65.134	113.571	65.183	112.988	65.134	114.736	65.183
2547	ref	Leo T	multi-beam	2007-12-27	23:10:27	120			143.721[e]	17.051						
2548	map	Leo T	multi-beam	2007-12-27	23:13:22	120	1	220	143.257[e]	16.607	143.853	17.495	143.257	16.607	144.101	17.495
2597	calib	S7	multi-beam	2007-12-28	12:59:03	30			131.965[g]	-1.011						
2600	ref	NGC2403	multi-beam	2007-12-28	13:10:03	120			114.213[e]	65.598						
2601	map	NGC2403	multi-beam	2007-12-28	13:13:00	120	1	210	112.993[e]	65.085	114.154	65.525	112.988	65.085	115.435	65.525
2602	ref	Leo T	multi-beam	2007-12-28	22:13:07	120			143.721[e]	17.051						
2603	map	Leo T	multi-beam	2007-12-28	22:16:03	120	1	225	143.257[e]	16.607	144.187	17.495	143.257	16.607	144.185	17.495
2604	ref	NGC2403	multi-beam	2007-12-29	07:45:20	120			114.212[e]	65.599						
2606	map	NGC2403	multi-beam	2007-12-29	08:12:58	120	209	377	114.272[e]	65.525	115.232	65.917	112.969	65.525	115.457	65.917
2607	map	NGC2403	multi-beam	2007-12-29	15:25:36	120	378	483	115.112[e]	65.917	113.064	66.112	112.951	65.917	115.477	66.112
2609	ref	NGC2403	multi-beam	2007-12-29	20:01:53	120			114.213[e]	65.599						
2610	map	NGC2403	multi-beam	2007-12-29	20:06:05	120	1	46	112.993[e]	65.085	113.105	65.183	112.988	65.085	115.435	65.183
2611	ref	Leo T	multi-beam	2007-12-29	22:12:51	120			143.721[e]	17.051						
2612	map	Leo T	multi-beam	2007-12-29	22:15:47	120	1	225	143.257[e]	16.607	144.186	17.496	143.257	16.607	144.185	17.496
2613	ref	NGC2403	multi-beam	2007-12-30	07:46:07	120			114.212[e]	65.599						
2614	map	NGC2403	multi-beam	2007-12-30	07:49:19	120	45	127	112.988[e]	65.182	113.567	65.329	112.986	65.182	115.443	65.329
2616	ref	NGC2403	multi-beam	2007-12-30	11:26:02	120			114.212[e]	65.599						
2617	map	NGC2403	multi-beam	2007-12-30	11:28:53	120	1	88	112.992[e]	65.084	112.986	65.231	112.988	65.084	115.439	65.231
2618	ref	NGC2403	multi-beam	2007-12-30	15:15:29	120			114.213[e]	65.598						
2619	map	NGC2403	multi-beam	2007-12-30	15:18:25	120	1	162	112.993[e]	65.085	114.625	65.427	112.984	65.085	115.444	65.427
2620	ref	Leo T	multi-beam	2007-12-30	22:16:48	120			143.721[e]	17.051						
2621	map	Leo T	multi-beam	2007-12-30	22:19:38	120	1	184	143.257[e]	16.607	143.455	17.369	143.256	16.607	144.186	17.369
4520	calib	S7	single-beam	2008-07-09	22:08:22	60			131.961[g]	-1.180						
4522	ref	NGC2403	single-beam	2008-07-09	22:16:54	60			114.212[e]	65.599						
4523	map	NGC2403	single-beam	2008-07-09	22:19:04	60	1	388	112.994[e]	65.086	113.913	65.917	112.994	65.086	115.435	65.9165
4655	calib	S7	single-beam	2008-07-11	14:30:34	60			131.961[g]	-1.180						
4656	ref	NGC2403	single-beam	2008-07-11	14:36:50	60			114.212[e]	65.599						
4657	map	NGC2403	single-beam	2008-07-11	14:38:19	60	380	484	114.876[e]	65.917	112.964	66.113	112.955	65.917	115.480	66.113
4659	calib	S7	single-beam	2008-07-11	16:58:53	60			131.961[g]	-1.180						
4660	ref	Leo T	single-beam	2008-07-11	17:06:15	60			143.721[e]	17.051						
4661	map	Leo T	single-beam	2008-07-11	17:09:17	60	1	111	143.389[e]	16.733	143.438	17.076	143.389	16.733	144.055	17.076
4663	calib	S7	single-beam	2008-07-11	19:40:09	60			131.961[g]	-1.180						
4664	ref	NGC2403	single-beam	2008-07-11	19:46:09	60			114.212[e]	65.599						
4665	map	NGC2403	single-beam	2008-07-11	19:48:24	60	1	199	113.027[e]	65.100	115.423	65.527	113.027	65.100	115.411	65.527
5178	calib	S7	single-beam	2008-07-19	05:07:07	60			131.961[g]	-1.180						
5179	ref	NGC2403	single-beam	2008-07-19	05:14:33	60			114.212[e]	65.599						
5180	map	NGC2403	single-beam	2008-07-19	05:16:53	60	190	484	114.499[e]	65.479	112.979	66.099	112.980	65.479	115.453	66.099
5182	calib	S7	single-beam	2008-07-19	11:46:23	60			131.961[g]	-1.180						
5183	ref	Leo T	single-beam	2008-07-19	11:56:51	60			143.721[e]	17.051						
5184	map	Leo T	single-beam	2008-07-19	11:58:38	60	100	196	144.003[e]	17.076	143.387	17.372	143.386	17.076	144.055	17.372

[g] galactic coordinate system, [e] equatorial coordinate system

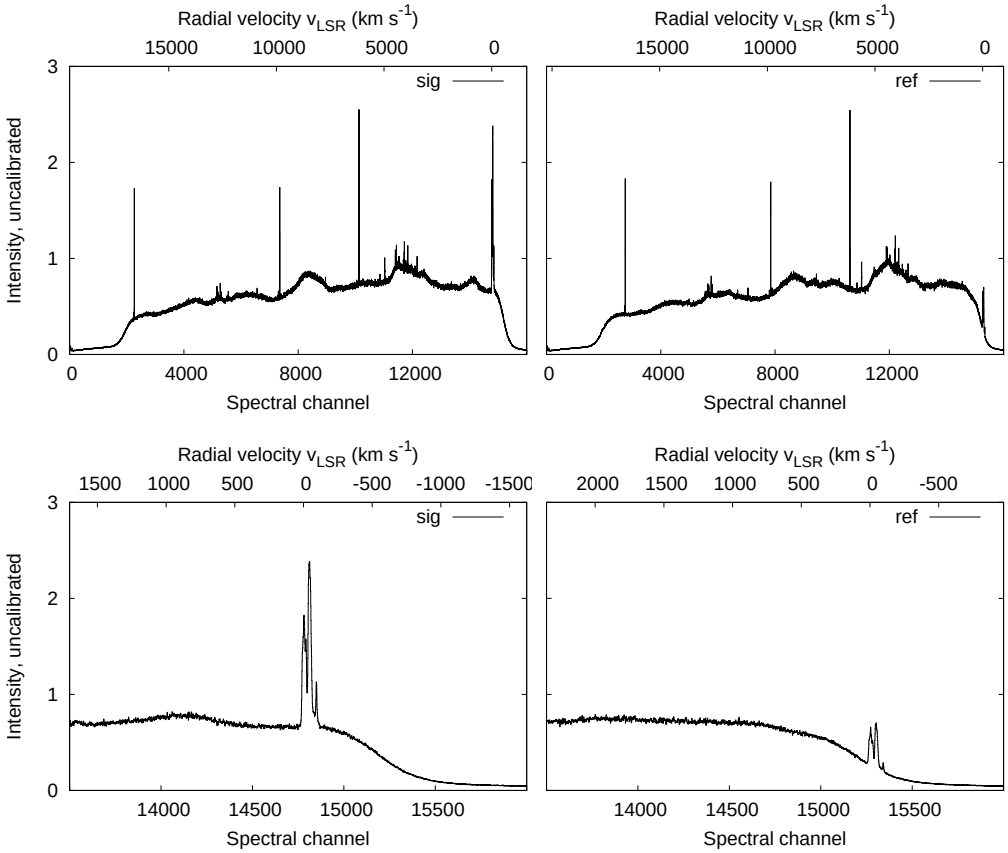


Fig. 5.8: Scan 2597 containing a measurement of the calibration source S7 showing strong gain curve ripples and several (strong) RFI signals. The lower panels show a zoom-in containing the spectral line of S7. The left panels display the spectral distribution for the signal phase, while the right panels show the reference phase. For the latter the emission line lies at the edge of the bandpass increasing the flux calibration error for the reference phase.

S7 emission line. In case of the multi-feed observations the LO setup had to set the mid frequencies to around 1380 MHz to avoid the feed resonance and the strong DAB RFI signal (Section 5.2). As a consequence the LSR velocities of $\sim 0\,\mathrm{km\,s^{-1}}$ and lower moved to the edge of the bandpass making the measurements in ph1 and ph3 less accurate. It is also clearly visible that the bandpass curve in both phases is similar but not matching. This leads to quite strong residual baselines after applying the frequency switching scheme; see Fig. 5.9. In the top panels the result is plotted.

To deal with the residual baseline wiggles a polynomial fit of (rather large) 10^{th} order was applied piecewise to spectral portions of 1024 channels width. This was done using the baseline fit procedure with automatic windowing provided by the RFI detection tool; see Section 3.2. The polynomial order in time-direction was two. The task was also used to find RFI signals which were flagged as bad for the gridding task. However, a lot of interference signals were constant in amplitude and/or well below the noise level. In those cases the detection algorithm failed. The fitting flattened the residual baseline

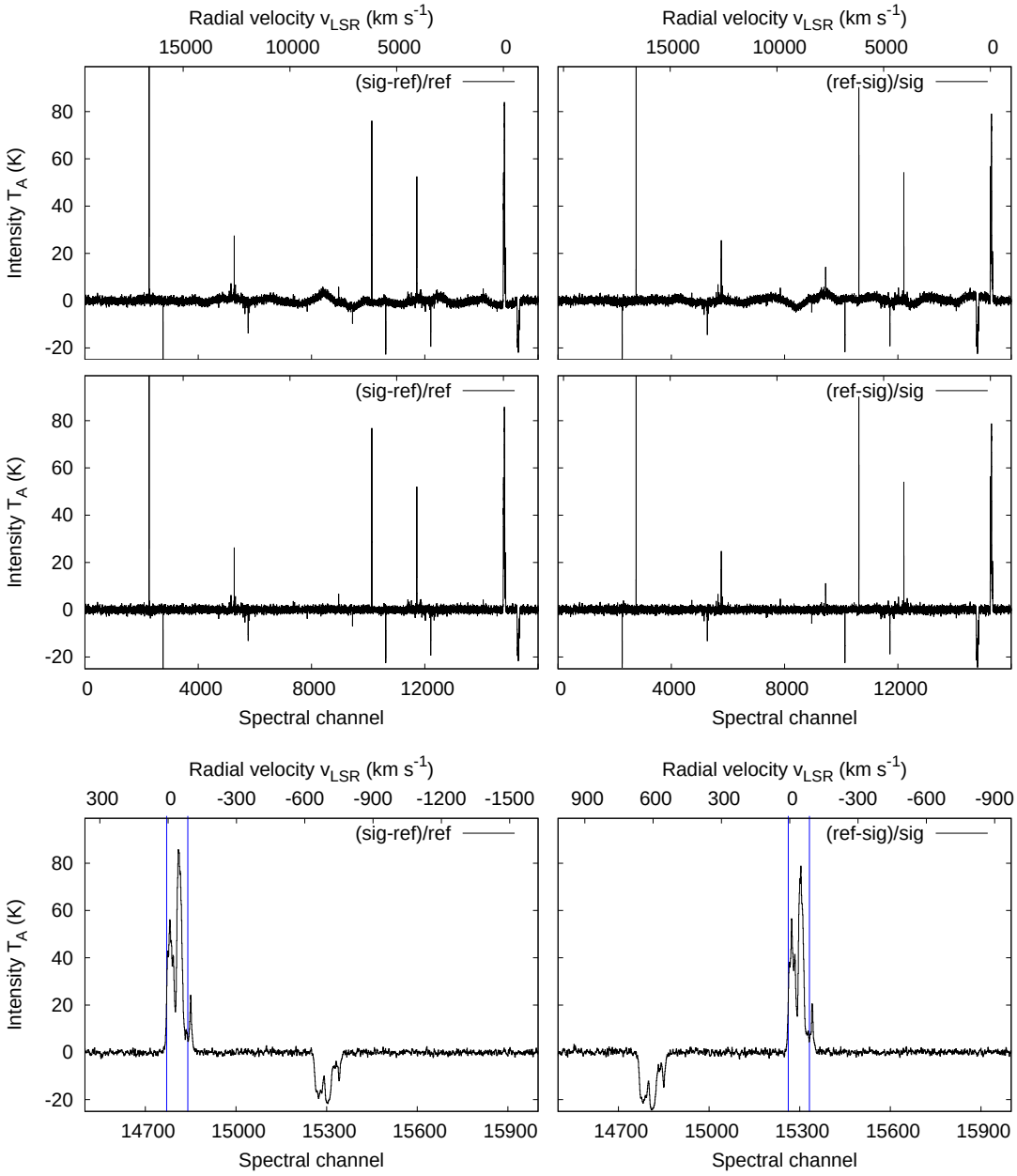


Fig. 5.9: One spectrum out of a measurement of the calibration source S7. The upper panel shows residual spectra (for a 500 ms dump) after bandpass calibration (frequency switching) was applied. Unfortunately, due to the problems with the receiver strong baseline features remain. To visualize the effect of frequency switching the left column contains the signal phase, the right column the reference phase. Note, that the radial velocities convert to different spectral channels due to the two LO frequency settings. Strong RFI peaks are visible all over the spectrum. To deal with the baseline ripples a 10^{th} order polynomial was fitted piecewise (using 1024 channels each) to the data. The residual spectra are shown in the middle panel. The lower panel shows a zoom-in containing the S7 emission line. The blue markers indicate the spectral region ($v_{\text{LSR}} = -86.5 \ldots 6.5\,\text{km/s}$) which is used for the absolute flux measurement of S7.

(Fig. 5.9 middle panels) but is not very sophisticated due to the large polynomial order. Broad emission features might have influence on the baseline fit. Another problem of the piecewise fitting is that the connection between adjacent spectral ranges is not necessarily smooth.

Using the method described in Section 5.1 the system temperature was determined providing the calibration for the S7 spectra. The lower panels show a zoom-in for better visualization. The velocity range used for measuring the S7 flux are marked by the vertical lines. Due to the method of frequency switching each spectral feature "generates" a negative and shifted signal. Unfortunately, this is especially problematic for RFI signals effectively "doubling" their impact.

Fig. 5.10 shows a single spectrum (500 ms dump) out of the Leo T observation (Scan 2621, w/o cal). The left (right) column contains the signal (reference) phase. After applying frequency switching, intensity calibration, and SR correction the residual gain curve is again not useful. As before, baseline fitting was performed (same parameters, middle panel) resulting in a sufficiently flat baseline, but eventually having the same drawbacks as discussed above. The lower panel shows a zoom-in showing the Milky Way emission line. Note, that the reference phase reveals lower intensity caused by the source being shifted to the edge of the bandpass.

The data were gridded (signal phase only, for a first test only Scan 2621, hence one coverage, and polarization channel one) to a datacube using a resolution of $1'$ per pixel. The result reveals a clear detection of Leo T. However, right on top of the emission line residuals of an RFI signal appear. Although a large fraction of it where correctly flagged (as a manual inspection of the flag database showed), the interference was hidden in the noise for some spectra.

The same reduction procedure was applied to the NGC 2403 dataset. As can be seen in Table 5.2 the map was observed during three sessions. Unfortunately, during the second measurement — covering the largest fraction of the source — strong bandpass instabilities occurred leading to extremely bad baseline quality. Fig. 5.11 exemplarily shows two spectra (500 ms dumps) with (right panel) and without (left panel) such baseline fluctuations. The issue appears to be similar to the problems observed during the first test measurements (Secion 5.3) and is probably caused by strong RFI events or resonances in the front-end, although it remains unclear what could cause such resonances. Note, that these fluctuations do not cancel out after applying the frequency switching scheme.

Compared to the above results, the measurements with the single-beam receiver (but the same backend) led to much better data quality. In contrast to the multi-feed observations the backend was used with a bandwidth of only 20 MHz leading to a much higher velocity resolution ($0.25\,\mathrm{km\,s^{-1}}$). The exactly same reduction procedure was applied with the only difference that here a low order polynomial was used to fit the baseline for the region of interest[2] providing much better constraints and avoiding discontinuities at the interface points. The only drawback is that in case of NGC 2403 the emission profile is extremely broad compared to the total number of channels which causes overfitting at the edges of the spectral line. As a consequence in some areas negative baseline residuals occur. (A piecewise fitting would be even more problematic.)

[2] For NGC 2403 a 6^{th} order baseline was fitted for the range of $-550 \ldots 1500\,\mathrm{km\,s^{-1}}$ (about 8 000 spectral channels). In case of Leo T a 5^{th} order baseline describes the range of $-550 \ldots 800\,\mathrm{km\,s^{-1}}$ (about 5 000 spectral channels).

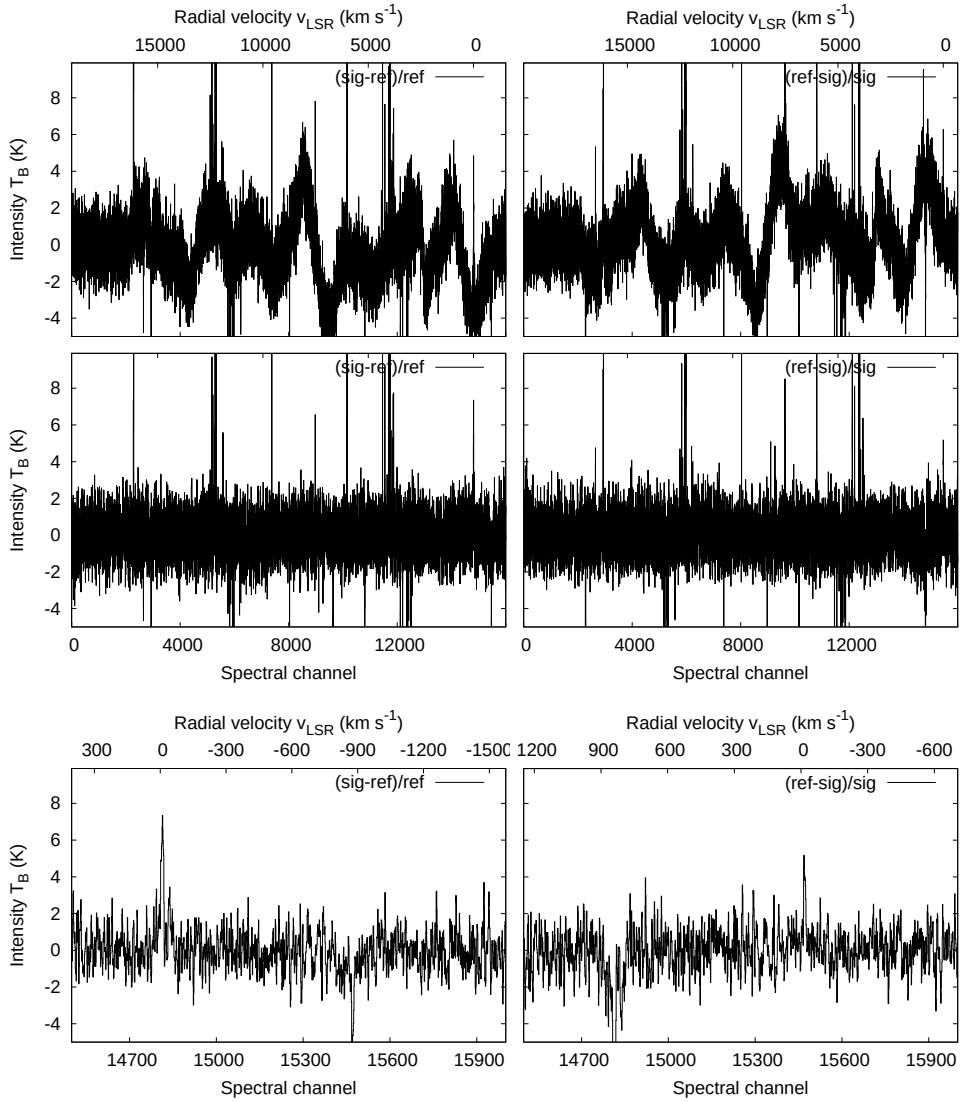


Fig. 5.10: Spectrum from one of the Leo T observations (Scan 2621). The upper row shows spectra (for a 500 ms dump) after applying the frequency and intensity calibration (using the absolute intensity provide by a measurement of S7, see Fig. 5.9, and the noise diode for relative a correction of the system temperature during the mapping), as well as stray-radiation correction; see Fig 5.3. Strong bandpass ripples remain which are due to the problems with the receiver (see Text). To deal with them a 10th order polynomial was fitted piecewise (using 1024 channels each) to the data. The residual spectra are shown in the middle panel. The lower panel shows a zoom-in containing the Milky Way emission line. Note, that in the reference spectrum the peak line intensity of the Milky Way emission is lower than in the signal phase. This is due to the fact, that the LO setup shifted the reference phase to much causing the emission line to appear at the edge of the bandpass curve. Hence, the signal quality is non-optimal in this regime (compare also to Fig. 5.8).

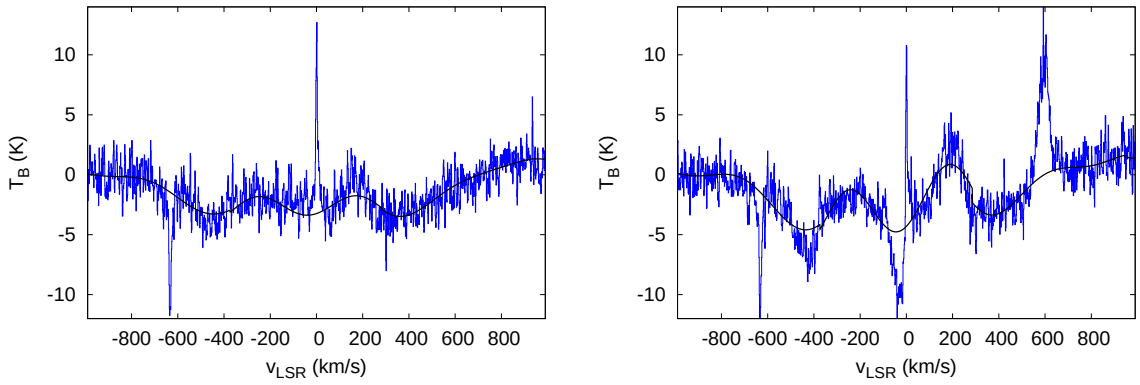


Fig. 5.11: During the second observing session several strong bandpass fluctuations occurred, which can not be described by the polynomials used for baseline fitting. The plots show exemplarily spectra without (left panel) and with (right panel) such a distortion. The frequency switching scheme was already applied, as well as intensity calibration.

Considering the amount of problems with the multi-feed system in the following (Section 5.6) only the single-beam data are used for the astronomical analysis of the sources. They not only provide a better overall data quality but also have a better spectral resolution and provide better constraints on the derived physical parameters, as all polarization channels and switching phases could be added reducing noise.

5.6 Physical properties of the observed sources

The physical parameters were derived using the single-beam measurements, as they provide not only a better overall data quality but also a much better spectral resolution. As shown in the previous Section, the multi-beam receiver still is affected by complex gain curve shapes making the baseline calibration very error-prone.

5.6.1 The dwarf galaxy Leo T

Leo T was detected recently as a stellar overdensity in the Sloan Digital Sky Survey Data Release 5 (SDSS DR5 Adelman-McCarthy et al. 2007). Irwin et al. (2007) report two types of stellar populations, a red giant branch and a sequence of young, massive stars. An H I counterpart is present in the HIPASS data and Ryan-Weber et al. (2008) observed the source with the WSRT and GMRT. Leo T reveals a stellar morphology similar to typical dwarf spheroidal (dSph) galaxies. However, the presence of the younger blue population, which is more commonly in dwarf irregulars (dIrr) led to the classification as *transitional* dwarf (hence the 'T' in its name). Fig. 5.12 shows an optical image superposed by H I column density contours. The optical center of the galaxy is RA $09^{\mathrm{h}}34^{\mathrm{m}}53.4^{\mathrm{s}}$, DEC $17°03'05''$.

From our observations a datacube was compiled using the full velocity resolution of $0.25\,\mathrm{km\,s^{-1}}$ and using a spatial pixel size of $1'$. The raster map was fully sampled ($3'$ grid) with one minute of integration time per position. The final data cube has a RMS noise

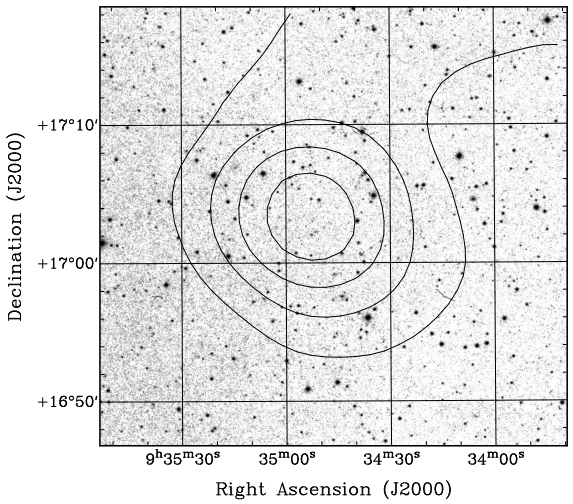


Fig. 5.12: The plot shows the dwarf galaxy Leo T. The optical image from SDSS[4] (Adelman-McCarthy et al. 2008) is overlaid with H I column density contours (starting at $5 \cdot 10^{18}\,\mathrm{cm}^{-2}$ in steps of $5 \cdot 10^{18}\,\mathrm{cm}^{-2}$; selected velocity range: $23.75 \ldots 43.59\,\mathrm{km\,s}^{-1}$) as observed during our observations.

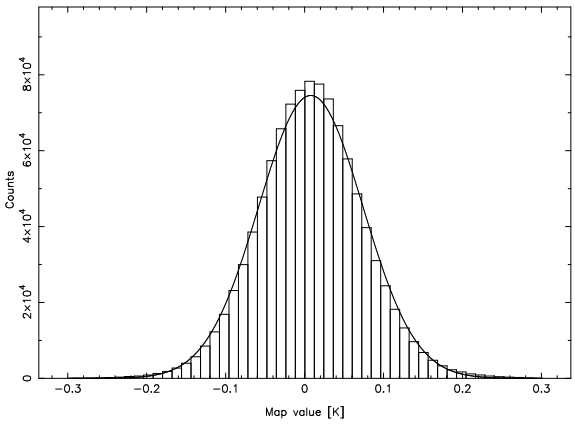


Fig. 5.13: Noise statistics calculated with the `miriad` task `imhist`. The resulting noise value is 66 mK.

level of 66 mK. The noise statistics were determined with the `miriad` task `imhist`; see Fig. 5.13. The resulting histogram is well described by a Gaussian.

Fig 5.14 shows the H I column density (moment 0) and velocity (moment 1) maps of Leo T, calculated using the `miriad` software package. The column densities follow from the integrated intensities using

$$N_{\mathrm{HI}}[\mathrm{cm}^{-2}] = 1.823 \cdot 10^{18} \int T_{\mathrm{B}}\,\mathrm{d}v \;\left[\mathrm{K\,km\,s}^{-1}\right] \tag{5.2}$$

assuming optically thin gas ($\tau \ll 1$). Applying a radial profile fit to the data (using `Kvis`) provides the spatial position (RA $09^{\mathrm{h}}34^{\mathrm{m}}52.4^{\mathrm{s}}$, DEC $17°03'10''$) and the total flux of the source. The coordinates are consistent with the optical and H I center reported by Irwin et al. (2007) and Ryan-Weber et al. (2008). A peak column density of $N_{\mathrm{HI}}^{\mathrm{peak}}[\mathrm{cm}^{-2}] = 2.1 \cdot 10^{19}\,\mathrm{cm}^{-2}$ is found.

[4] Funding for the Sloan Digital Sky Survey (SDSS) and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, and the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web site is http://www.sdss.org/.

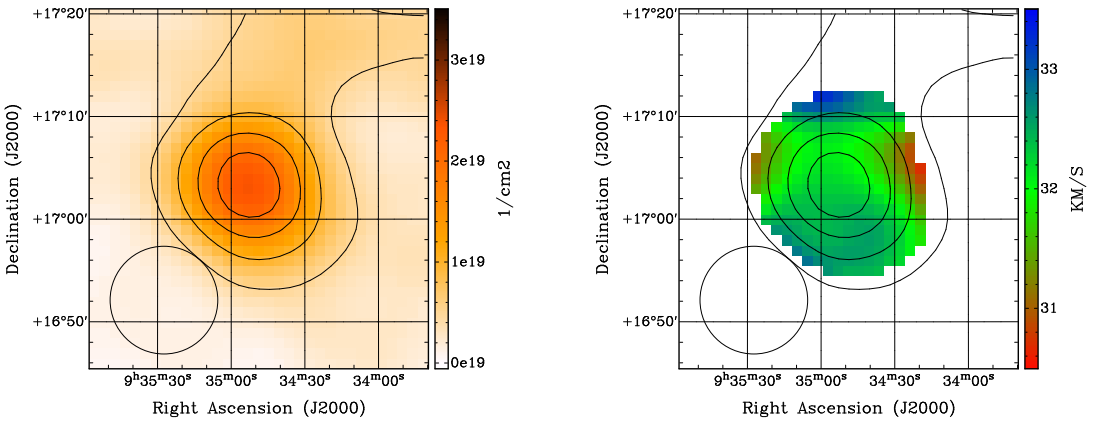


Fig. 5.14: Column density (moment 0, left panel) and velocity (moment 1, right panel) map of Leo T. Both plots show N_{HI} contours starting at $5 \cdot 10^{18}\,\mathrm{cm}^{-2}$ in steps of $5 \cdot 10^{18}\,\mathrm{cm}^{-2}$. ; The selected velocity range was $23.75 \ldots 43.59\,\mathrm{km\,s}^{-1}$. Apparently there is a diffuse contribution in the northern part of the map which can be attributed to Milky Way emission; see the position–velocity plots in Fig. 5.15 and 5.16. From the observations a tiny velocity gradient of about $0.5\,\mathrm{km\,s}^{-1}$ in North-South direction is inferred hardly indicating a rotation of the source. The circle in the lower left marks the effective H I beam size (after gridding) of about $10'$.

The column density map shows excess emission towards the North of Leo T, which can be attributed to the Milky Way. In Fig. 5.15 and Fig. 5.16 position–velocity slices are plotted along constant right ascension and declination, respectively. As the signal-to-noise ratio is low the moment-1 calculation in the outer parts of Leo T is poorly confined. Hence the mean velocities in Fig. 5.14 (right panel) should be treated carefully. A slight velocity gradient of at most $0.5\,\mathrm{km\,s}^{-1}$ is found. Although the p–v diagrams visually suggest a physical connection between MW emission and Leo T this can be ruled out based on the distance of 420 kpc of the latter. The low radial velocity of Leo T is peculiar leading to a superposition of emission. Also, there is slight increase of the MW flux towards the north–western part, which produces the feature in the column density map.

Using the resulting basic physical properties can be derived. The source has its peak intensity at around $v_{\mathrm{lsr}} = 30\,\mathrm{km\,s}^{-1}$. To obtain line parameters in the peak spectrum of Leo T in the presence of diffuse emission from the Milky Way a multi-component fit was performed, utilizing six Gaussian components

$$g_i(v) = A_i \exp\left[-\frac{(v - v_i)^2}{2\sigma_i^2}\right] . \tag{5.3}$$

The fit is shown in Fig. 5.17. The resulting the peak brightness temperature is $T_{\mathrm{B}}^{\mathrm{peak}} = 0.57 \pm 0.02\,\mathrm{K}$, centered around $v_{\mathrm{lsr}}^{\mathrm{peak}} = 31.6 \pm 0.2\,\mathrm{km\,s}^{-1}$ with a linewidth of $\sigma^{\mathrm{peak}} = 5.1 \pm 0.2\,\mathrm{km\,s}^{-1}$. The latter converts to an upper (physical) temperature limit of $T_{\mathrm{phys}} < 3200 \pm 300\,\mathrm{K}$ using

$$T[K] = 21.8 \cdot \Delta v_{\mathrm{fwhm}}[\mathrm{km/s}]^2\,; \qquad \Delta v_{\mathrm{fwhm}} = \sqrt{8 \ln 2}\,\sigma\,, \tag{5.4}$$

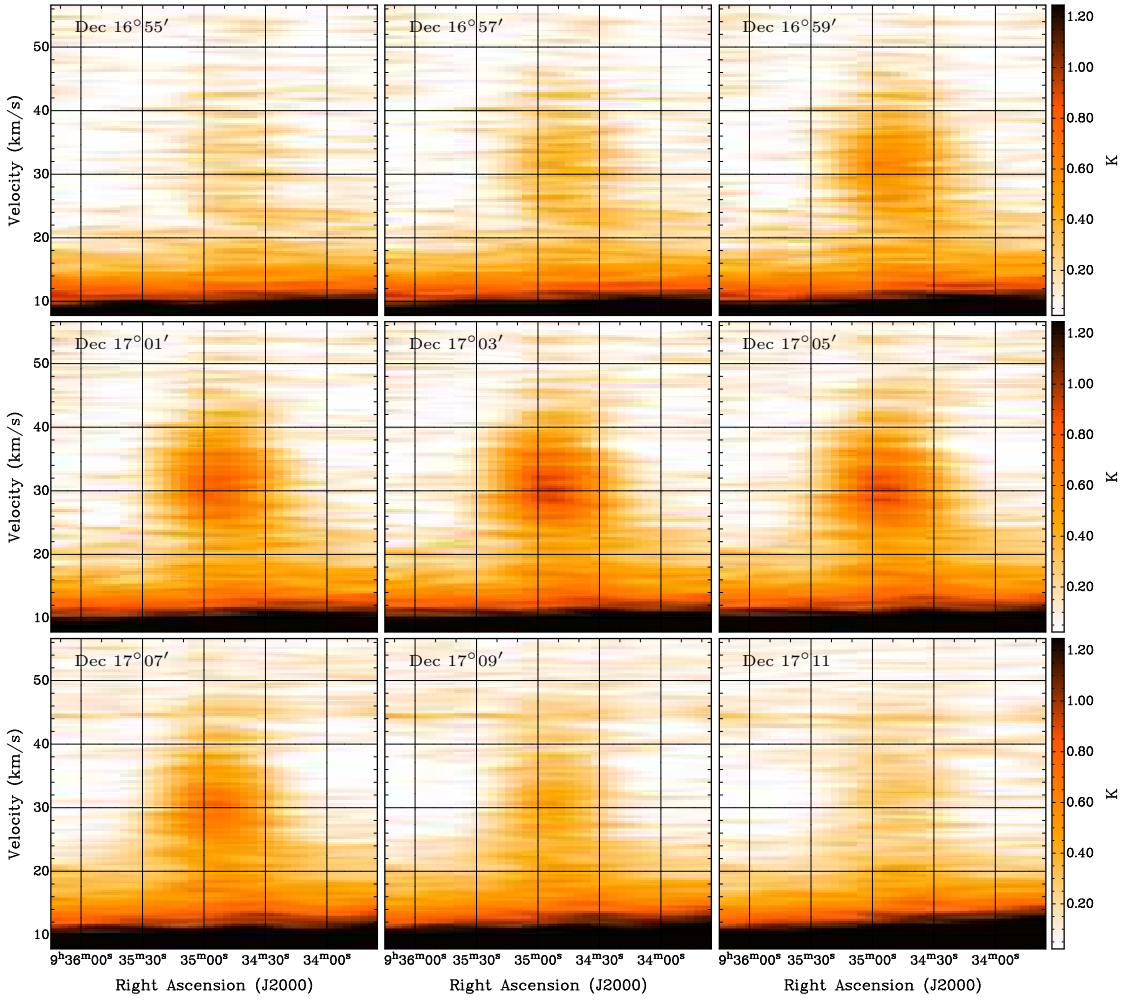


Fig. 5.15: Position–velocity maps of Leo T along right ascension. The high intensity around $v_{\mathrm{lsr}} \approx 0\,\mathrm{km\,s^{-1}}$ is due to Milky Way emission partly overlapping with NGC 2403.

while the peak brightness temperature T_{B} itself is a lower limit for the physical temperature. The given errorbars are only the statistical uncertainties from the fit itself. The spatial resolution and noise level are unfortunately not good enough to confirm here the co-existence of a cold and warm gas phase (Ryan-Weber et al. 2008). The final survey data will be about three times more sensitive and might address this issue.

The total H I mass M_{HI} can be estimated by assuming a certain distance d of the source and considering the spatial size of each pixel leading to

$$M_{\mathrm{HI}} = m_{\mathrm{HI}} d^2 \tan^2 \varphi \sum_i N_{\mathrm{HI}}^{(i)} \tag{5.5}$$

or equivalently

$$M_{\mathrm{HI}}\,[\mathrm{M}_\odot] = 7.95 \cdot 10^{-9}\,(d[\mathrm{Mpc}])^2 \tan^2 \varphi \sum_i N_{\mathrm{HI}}^{(i)}[\mathrm{cm}^{-2}]. \tag{5.6}$$

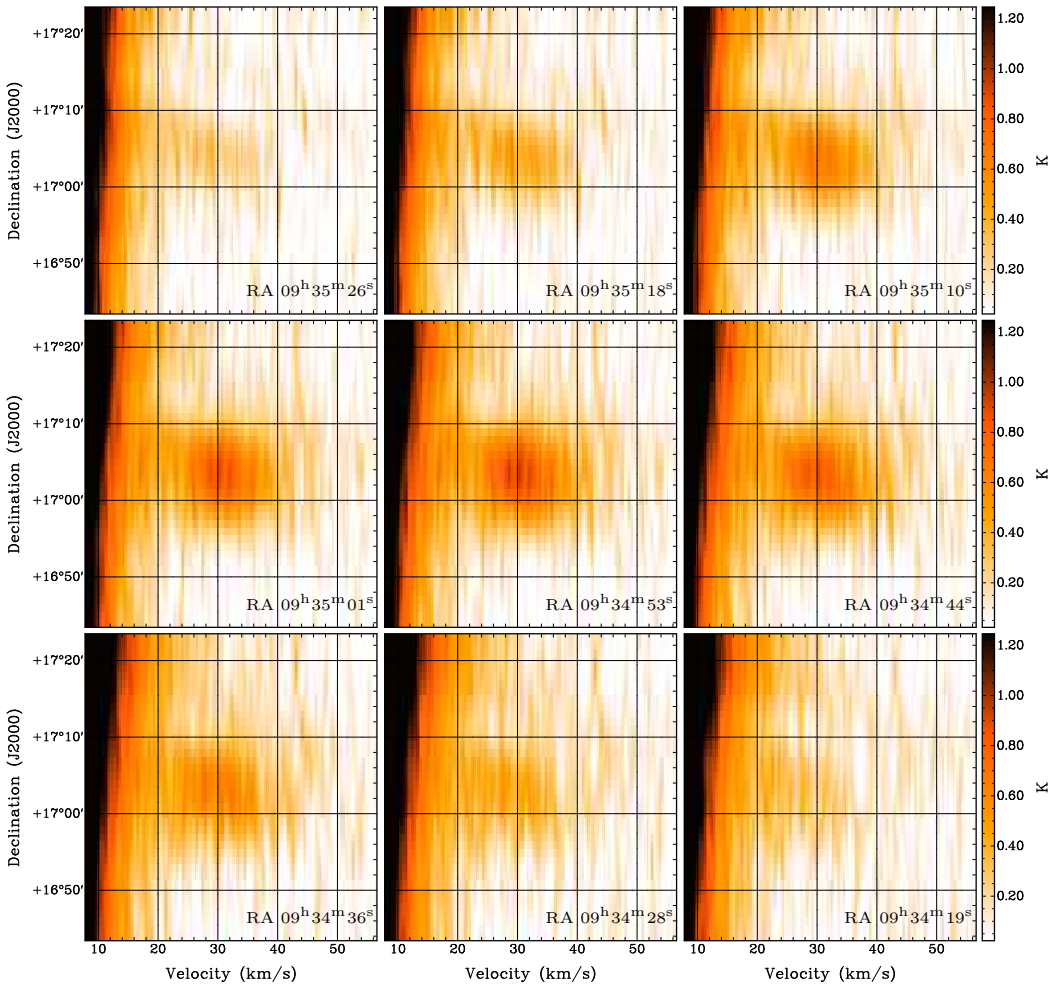


Fig. 5.16: Position–velocity maps of Leo T along declination. The high intensity around $v_{\mathrm{lsr}} \approx 0\,\mathrm{km\,s^{-1}}$ is due to Milky Way emission partly overlapping with NGC 2403.

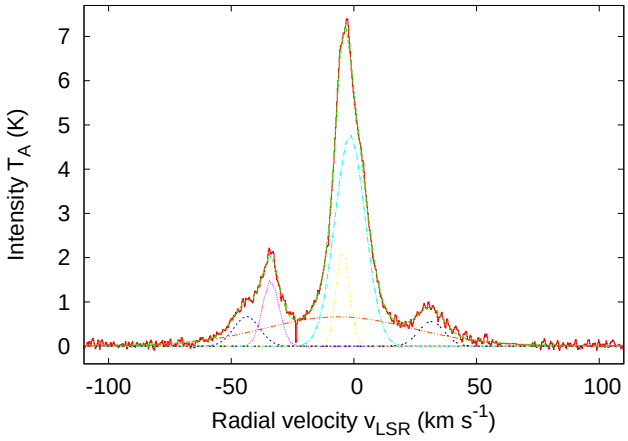


Fig. 5.17: To obtain line parameters of Leo T in the presence of diffuse emission from the Milky Way a multi-component fit was performed, utilizing six Gaussian components. The resulting parameters for the Leo T emission are presented in the text.

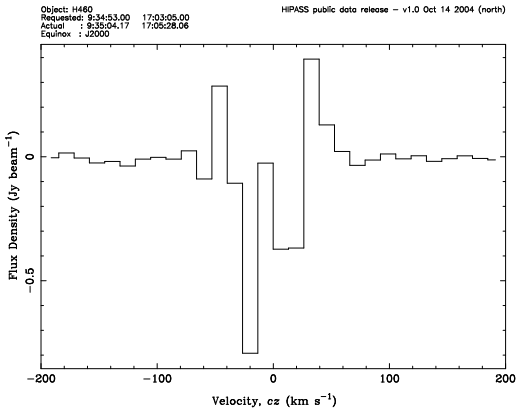


Fig. 5.18: HIPASS spectrum in the direction of Leo T. The HIPASS data are not very reliable in the vicinity of the Milky Way due to ringing effects caused by the autocorrelator. This data is provided under the auspices of the Multibeam Survey Working Group. The Parkes telescope is part of the Australia Telescope which is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO.

Using the above equations the masses per pixel can be derived from the data cube. However, by using a radial profile fit to the column density map the base level is taken into account, providing better estimates on the total flux (and hence mass) as the superposed MW emission is not entering the calculation.

Assuming a distance of $d = 420\,\mathrm{kpc}$ (as found by the optical measurements by Irwin et al. 2007) we estimate the H I mass to be $M_{\mathrm{HI}} = 3.6 \cdot 10^5\,\mathrm{M}_\odot$. Ryan-Weber et al. (2008) calculated a value of $2.8 \cdot 10^5\,\mathrm{M}_\odot$ based on their interferometric measurements with the GMRT and WSRT which are not sensitive to diffuse emission (having low spatial frequencies). With respect to this selection effect, both observations lead to a consistent mass estimation. It is also important to note, that Irwin et al. (2007) give an H I mass estimate of $M_{\mathrm{HIPASS}} = 2.0 \cdot 10^5\,\mathrm{M}_\odot$ based on the HIPASS (Barnes et al. 2001) data which they claim to be consistent with their results. Nevertheless, it is inconsistent with our results, which is not surprising as HIPASS can provide only a rough estimate in the vicinity of the Milky Way due to ringing effects caused by the autocorrelator; see Fig. 5.18.

The overall linewidth, or velocity dispersion, σ_v could also be used to estimate a dynamical mass

$$M_{\mathrm{dyn}} = \frac{r_{\mathrm{g}}\sigma_{\mathrm{v}}^2}{G} \tag{5.7}$$

with r_{g} being the gravitational radius and G the gravitational constant. This equation is only applicable assuming an isotropic velocity distribution and provides only a lower limit, since the dark matter might be more extended than the gas distribution, thus having a larger $r_{\mathrm{g}} \gtrsim 300\,\mathrm{pc}$. Our velocity dispersion of $\sigma_{\mathrm{v}} = 5.1\,\mathrm{km\,s^{-1}}$ is slightly smaller than reported by Ryan-Weber et al. (2008) ($\sigma_{\mathrm{v}} = 5.1\,\mathrm{km\,s^{-1}}$) leading to $M_{\mathrm{dyn}} \gtrsim 1.8 \cdot 10^6\,\mathrm{M}_\odot$. This makes the mass-to-luminosity ratio of Leo T rather high ($L_V \sim 6 \cdot 10^4\,\mathrm{M}_\odot$).

5.6.2 The spiral NGC 2403

The second observed source was the nearby spiral (SBc) galaxy NGC 2403. Its inclination angle of about 60° offers the possibility to study both the density structure and kinematics. Fig. 5.19 shows an optical image as obtained from SDSS (Adelman-McCarthy et al. 2008) superposed by the H I column density contour lines obtained from our datacube. The size of the gaseous disk is about twice the size of the optical disk.

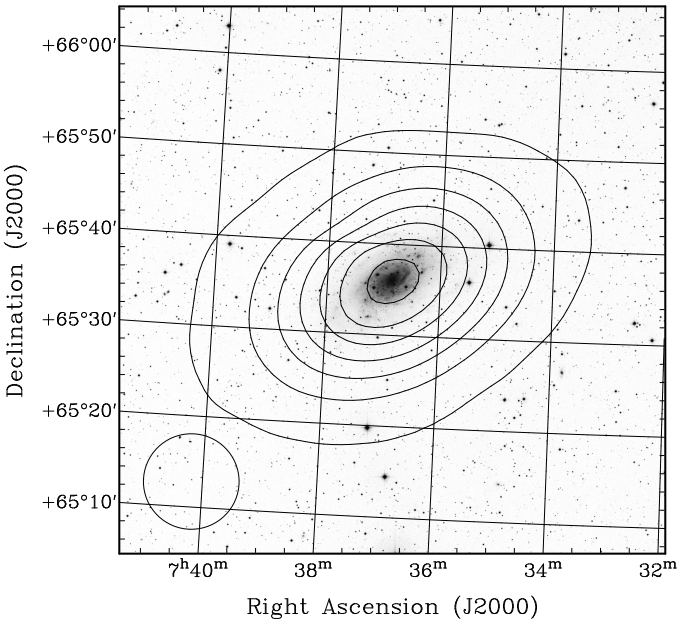


Fig. 5.19: The plot shows the nearby spiral galaxy NGC 2403. The optical image from SDSS (Adelman-McCarthy et al. 2008) is overlaid with H I column density contours (starting at $1 \cdot 10^{20}\,\mathrm{cm}^{-2}$ in steps of $2 \cdot 10^{20}\,\mathrm{cm}^{-2}$; selected velocity range: $13.39 \ldots 271.36\,\mathrm{km\,s}^{-1}$) as observed during our observations. The extend of the gaseous disk is about a factor of two larger than the optical size. The circle in the lower left marks the effective H I beam size (after gridding) of about $10'$.

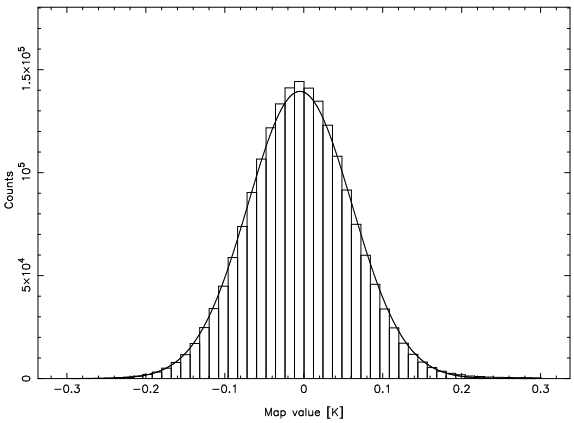


Fig. 5.20: Noise statistics calculated with the `miriad` task `imhist`. The resulting noise value is about 90 mK. Apart from the excess towards large brightness temperatures, caused by the source emission, there is a slight increase of the mean value of about 20 mK. This is due to the imperfect baseline fitting; see text for details.

The resulting datacube has the same properties as in case of Leo T ($\delta v = 0.25\,\mathrm{km\,s}^{-1}$, pixel size of $1'$, fully sampled $3'$ grid, RMS noise level $\Delta T = 66\,\mathrm{mK}$). The noise statistics are shown in Fig. 5.20. As before, the resulting histogram is well described by a Gaussian.

Fig. 5.21 contains the column density map, as well as velocity field (i.e., the 0^{th} and 1^{st} moment). The 1^{st} moment map is typical for a rotating disk and reveals a global velocity gradient of about $200 \ldots 240\,\mathrm{km\,s}^{-1}$. At low velocities $v_{\mathrm{lsr}} \approx 0 \ldots 30\,\mathrm{km\,s}^{-1}$ emission from NGC 2403 overlaps with the spectral lines of Milky Way gas, though only its outer spectral wing is affected. The moment maps, therefore, where computed only for the velocity planes from $30\,\mathrm{km\,s}^{-1}$ to $280\,\mathrm{km\,s}^{-1}$. As a consequence, the total column density is probably slightly underestimated. On the other hand, residual MW emission could increase the measured flux. In the moment-0 map there is a small overall offset visible. During the determination of the total source flux, this offset was considered.

A careful inspection of the datacube revealed that at small to intermediate positive velocities the mean flux (off-source) is slightly above zero. While for the lowest velocities this is due to Milky Way emission, for the regime between 25 and $100\,\mathrm{km\,s}^{-1}$ it is probably

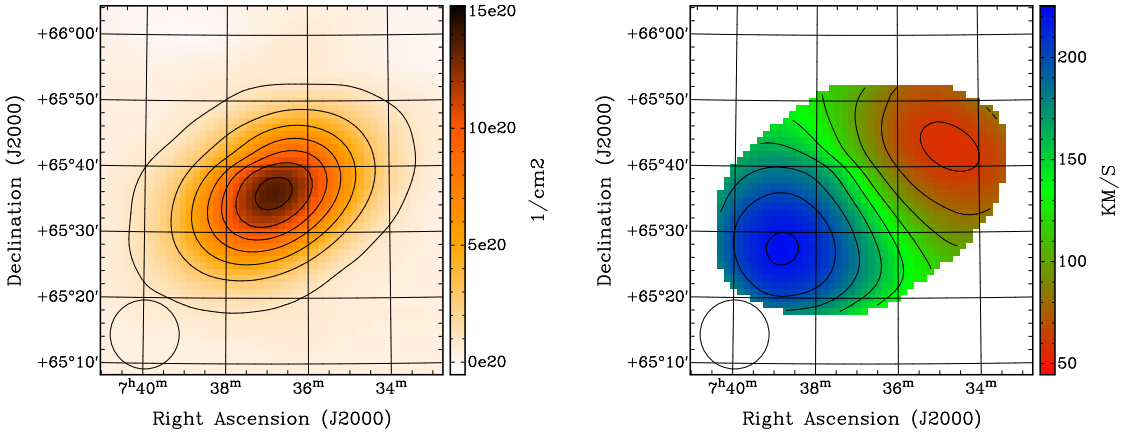


Fig. 5.21: $N_{\rm HI}$ column density and velocity map of NGC 2403. $N_{\rm HI}$ contours start at $10^{20}\,{\rm cm}^{-2}$ and increase in steps of $2 \cdot 10^{20}\,{\rm cm}^{-2}$ (selected velocity range: $13.39 \ldots 271.36\,{\rm km\,s}^{-1}$). The $v_{\rm lsr}$ contours range from 60 to $220\,{\rm km\,s}^{-1}$ in steps of $20\,{\rm km\,s}^{-1}$ Regions containing negative flux as a result of overfitted baselines were manually set to zero intensity before computation. The velocity plot reveals the typical double-horn profile of rotating disks. A global velocity gradient of about $200 \ldots 240\,{\rm km\,s}^{-1}$ is observed.

a baseline effect. Due to the fact that for the chosen bandwidth of 20 MHz the emission line is rather broad (about 1200 spectral channels out of the total of 16k) the gain curve is not perfectly confined within the spectral window. However, compared to the line strength the effect is small. Furthermore, the residual baseline is equal for all positions, which makes it possible to determine the true source flux.

In Fig. 5.22 channel maps of WSRT high-resolution data of NGC 2403 are plotted for $v_{\rm lsr} = 30 \ldots 250\,{\rm km\,s}^{-1}$ in steps of $20\,{\rm km\,s}^{-1}$. The observed single dish H I flux as obtained by the 100-m telescope is overlaid as contours.

To further investigate the kinematic structure several slices were computed along the major and minor axis as shown in Fig. 5.23. The `Karma` task `kpvslice` was used to calculate the position–velocity maps (Fig. 5.24 and Fig. 5.25) along each of the slices[5].

Using such p–v diagrams can provide a wealth of information. Fraternali et al. (2002) performed high-resolution WSRT observations. Comparing the datacube with a thin-disk model shows residual emission, which they call anomalous gas, which can be attributed to the halo of NGC 2403. In a second step also a halo component was considered in the model. The outcome is only consistent with the measured data if its rotation is slower than that of the thin disk. This phenomenon is widely observed for spiral galaxies also at different wavelengths (Oosterloo et al. 2007a; Kamphuis et al. 2007; Heald et al. 2007). The reason for a lagging halo is not yet clear, most likely it is due to outflows and/or accretion. The main drivers for outflows are probably supernovae (associated shells and bubbles), cosmic ray electrons, and galactic winds. The hot ionized gas would cool down in the outer halo and fall back to the plane. Such a galactic fountain provides momentum which can

[5] Please note, that due to the curvilinear grid the `kpvslice` slicing algorithm gives only approximative results, as it is working on the (rectangular) pixel grid; see the `kpvslice` documentation `http://www.atnf.csiro.au/computing/software/karma/user-manual/`.

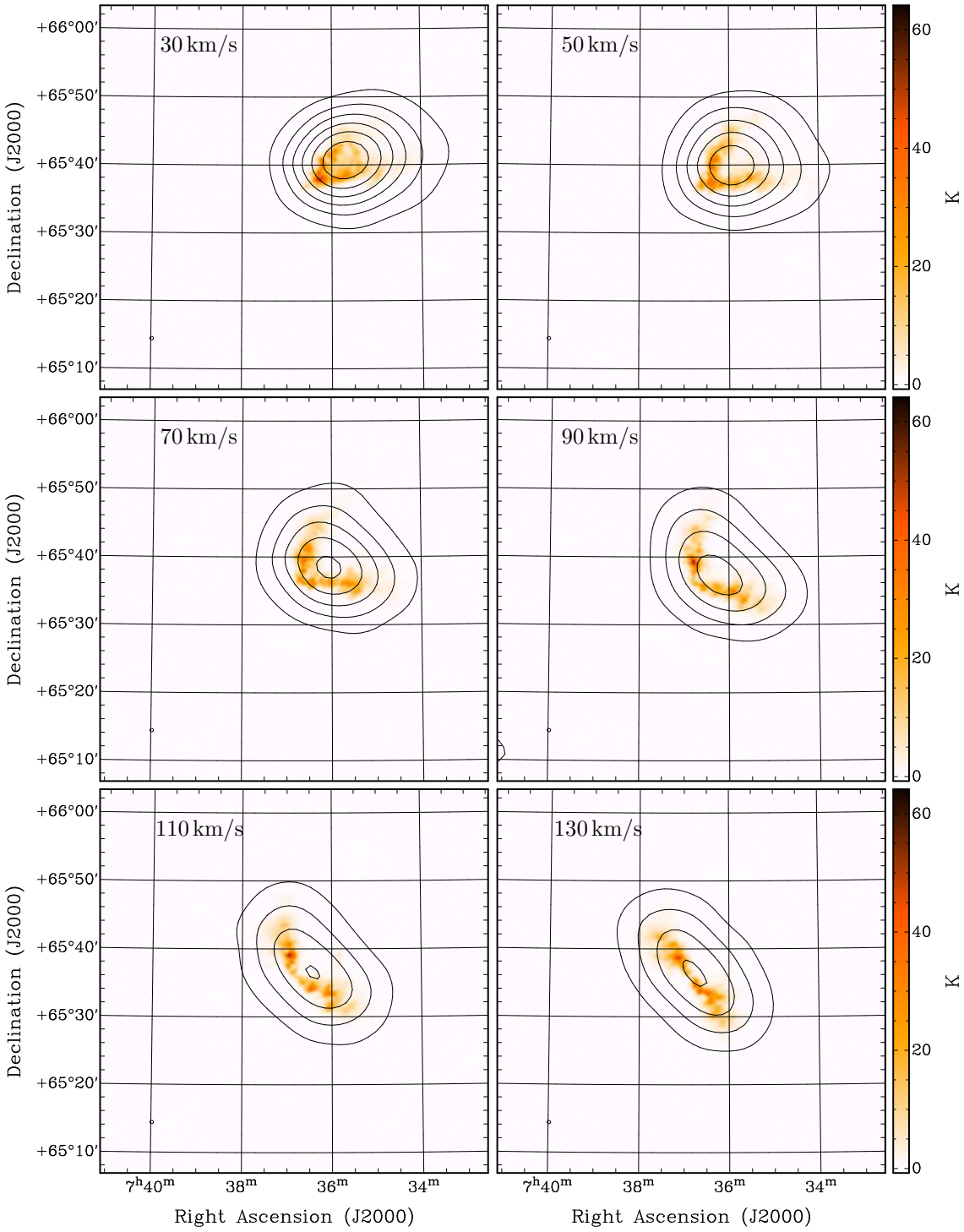


Fig. 5.22: Velocity channel maps of NGC 2403 showing the rotating disk of the spiral galaxy. The colored plot displays high-resolution WSRT data (by courtesy of T. Oosterloo), the contours (starting at $T_B = 1\,K$ in steps of 1 K) mark the single dish data as observed with the 100-m telescope.

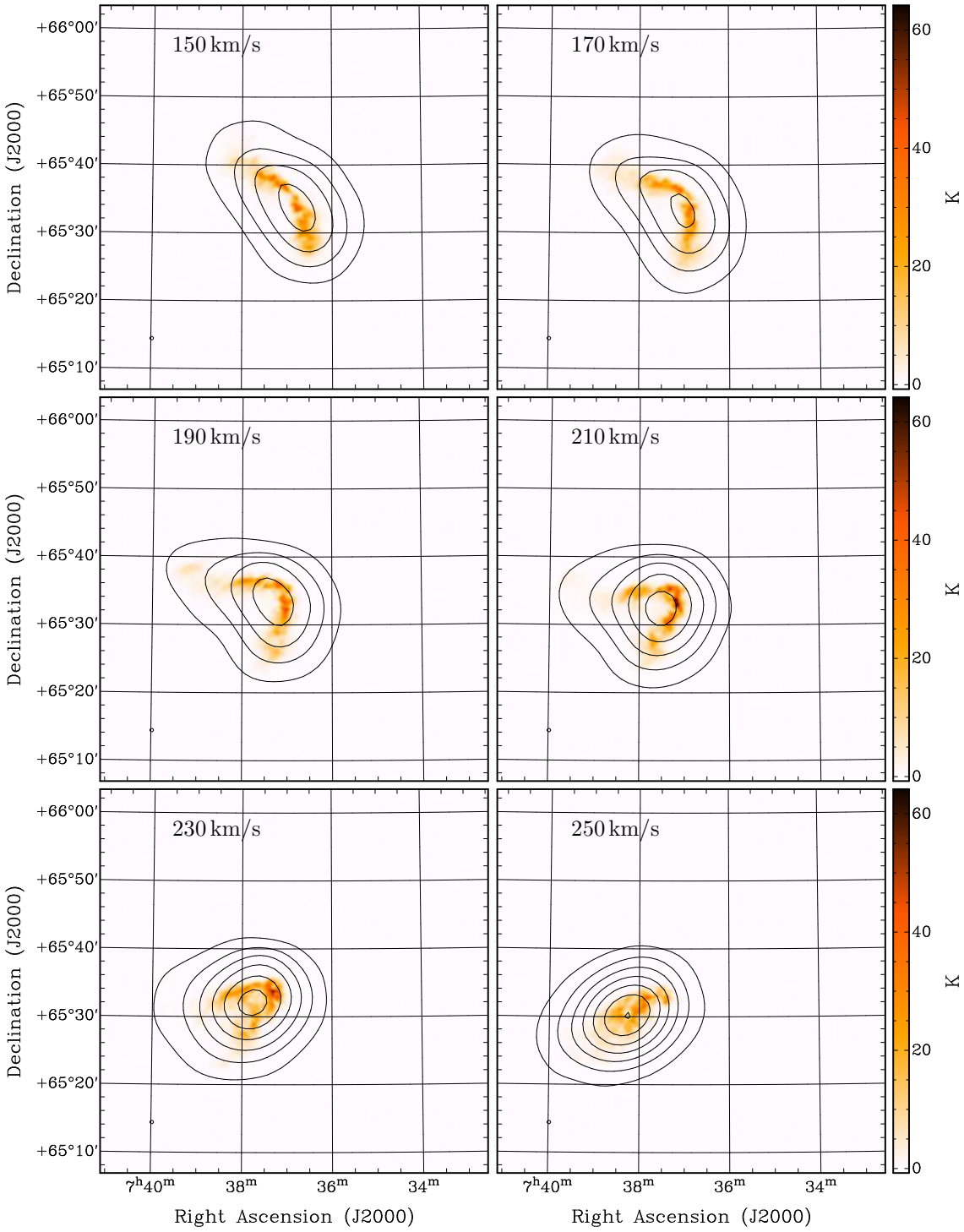


Fig. 5.22: *Continued.*

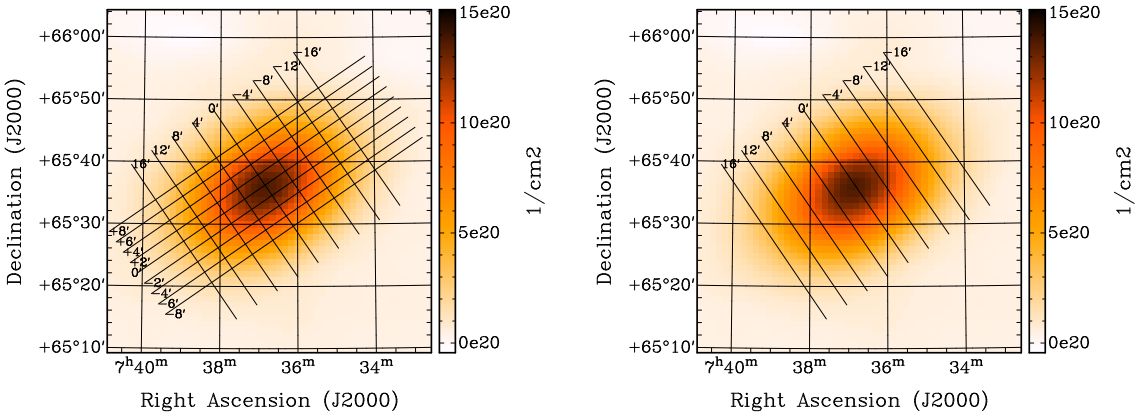


Fig. 5.23: To investigate the kinematics of NGC 2403 position–velocity maps (see Fig. 5.24 and Fig. 5.25) were computed along slices parallel to the major (left panel) and minor axis (right panel).

slow down the halo (Fraternali & Binney 2006). However, a small amount of accretion is probably needed to explain all observational quantities (e.g., the exact rotational velocity gradient).

Unfortunately, single dish observations have too low angular resolution for this kind of analysis, but they can provide short spacings for interferometric measurements making flux corrections possible.

Using the H I column densities and assuming a distance of $d = 3.18\,\mathrm{Mpc}$ (Madore & Freedman 1991) the total mass $M_{\mathrm{HI}} = 3.8{\cdot}10^9\,\mathrm{M}_\odot$ was computed using Eq. (5.5). The mass might be slightly overestimated due to the baseline fitting problems, as well as the fact that (some) emission from the lower velocity regions ($\leq 30\,\mathrm{km\,s^{-1}}$) might contain a Milky Way contribution. As both effects are more or less independent on position they can be accounted for by subtracting the offset ($4{\cdot}10^{19}\,\mathrm{cm^{-2}}$) in the column densities (see Fig. 5.19). Considering this offset leads to a corrected mass estimate of $M_{\mathrm{HI}} = 3.35 \pm 0.01 \cdot 10^9\,\mathrm{M}_\odot$ which is consistent with results from Fraternali et al. (2002) who stated a total mass of $M_{\mathrm{HI}} = 3.24 \cdot 10^9\,\mathrm{M}_\odot$ using interferometric data (which lacks the short-spacings and, hence, does not include all diffuse gas) and Rots (1980, $M_{\mathrm{HI}} = 3.31{\cdot}10^9\,\mathrm{M}_\odot$) for single-dish observations[6]. The single-dish fluxes match remarkably well.

5.7 Conclusions — quality of the new receiving system

During several test observations the new receiver as well as the FFTS backend were inspected. While the system temperature was acceptable for the offset feeds, the central horn showed strong baseline instabilities, leading to an increased value of T_{sys}. Furthermore, its stability, i.e., the Allan time, was far from being usable. Both problems most likely are associated with strong RFI signals entering the system — although being outside the observed spectral band. Here, the filters in use seem to have to low stopband attenuation. As a consequence new IF filters have been designed and were implemented, but not yet

[6] The errorbars take only statistical errors into account.

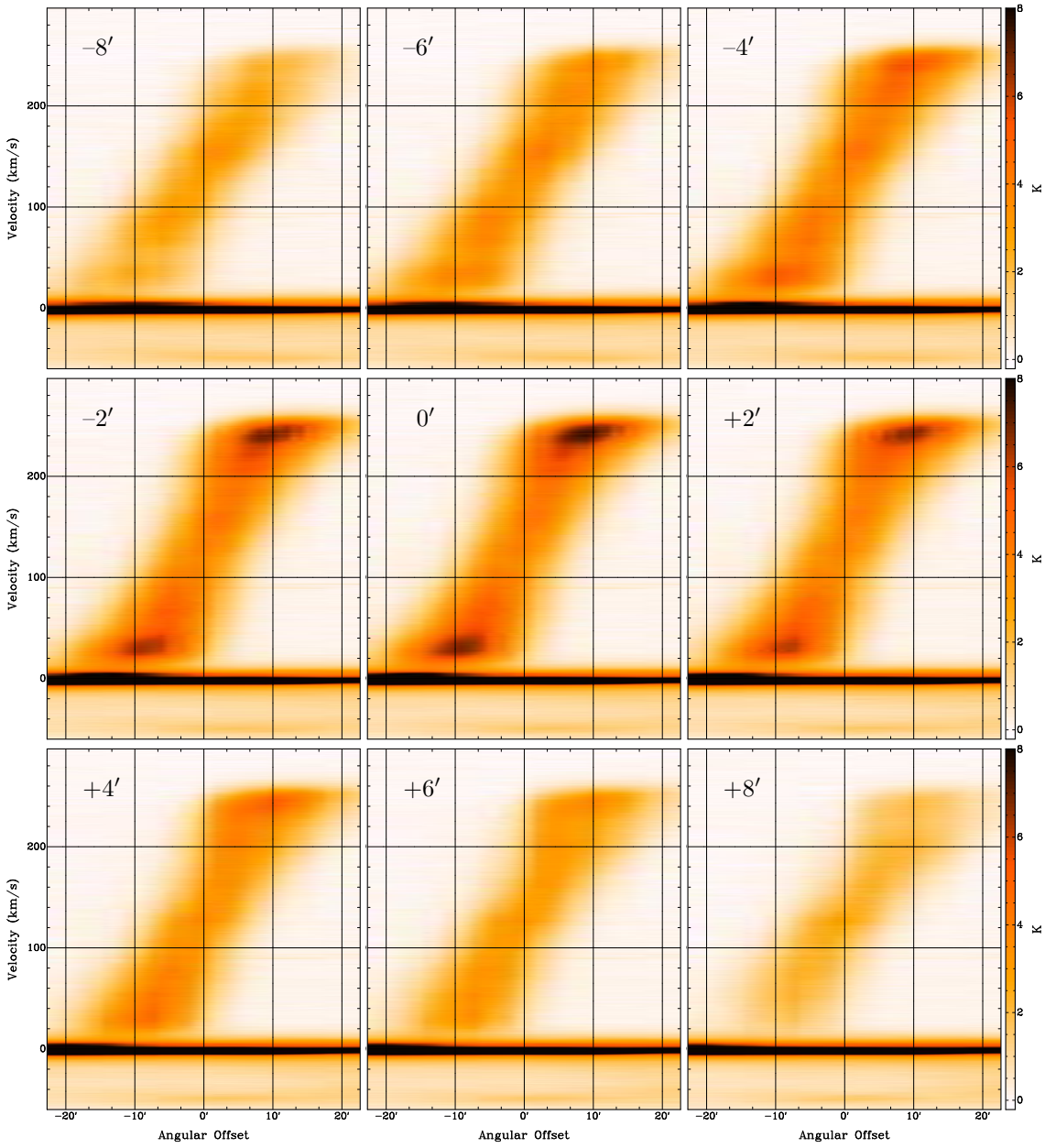


Fig. 5.24: Position–velocity maps of our observations along the major axis of NGC 2403. Fig. 5.23 shows the position of each of the slices. The high intensity around $v_{\mathrm{lsr}} \approx 0\,\mathrm{km\,s^{-1}}$ is due to Milky Way emission partly overlapping with NGC 2403.

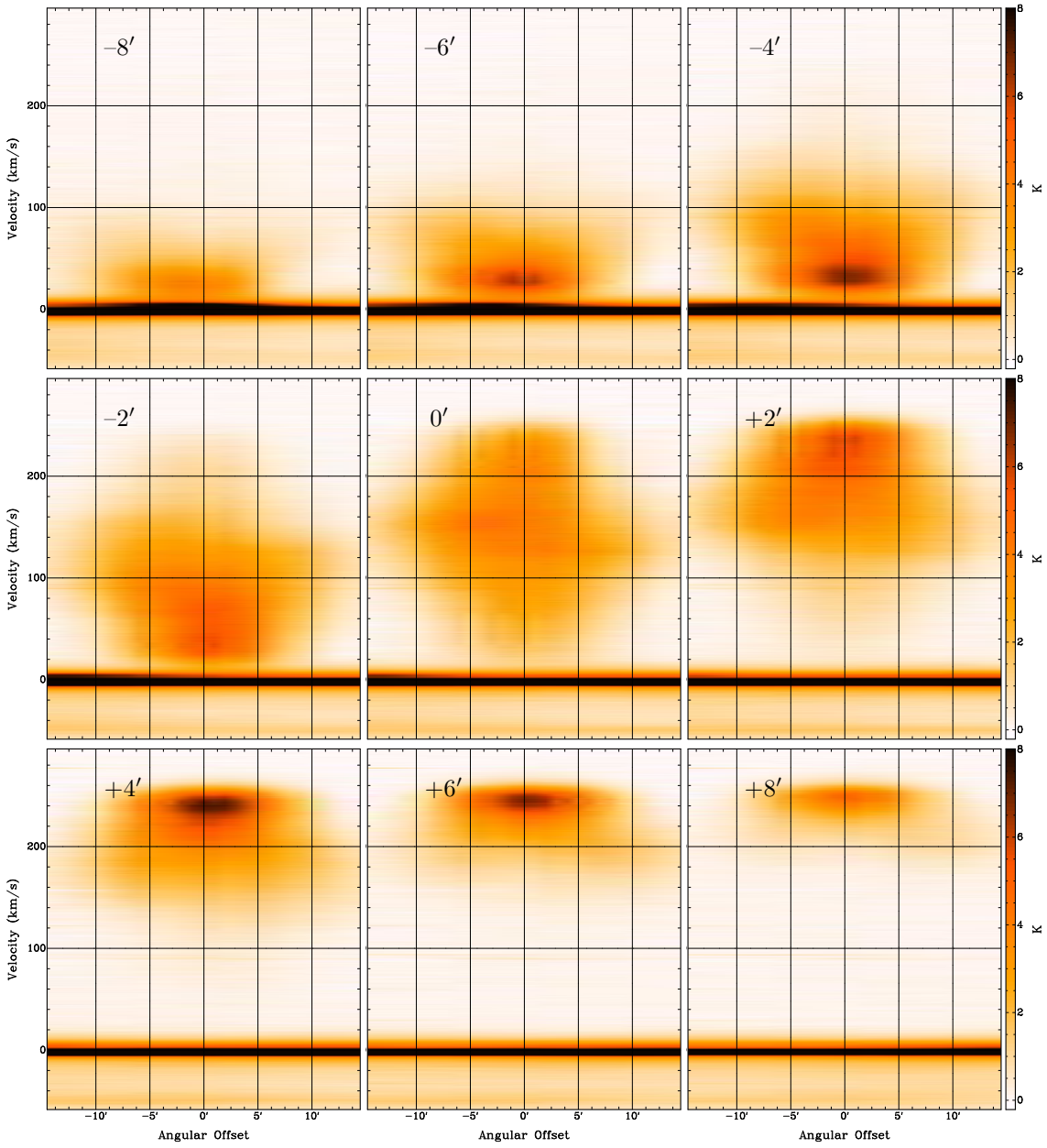


Fig. 5.25: Position–velocity maps of our observations along the minor axis of NGC 2403. Fig. 5.23 shows the position of each of the slices. The high intensity around $v_{\mathrm{lsr}} \approx 0\,\mathrm{km\,s^{-1}}$ is due to Milky Way emission partly overlapping with NGC 2403.

tested. Interestingly, only the central feed suffered strongly from the RFI emission. This is likely due to the circular polarization which is measured with the central feed, while the offset feeds provide linear polarization. As RFI is mostly linearly polarized, the signals can couple into the feed with arbitrary polarization angles.

A second major issue are multi-mode standing waves (between the primary and secondary antenna) within the 100 MHz of bandwidth. Usually, one expects this phenomenon for observations using the secondary focus. As all test measurements were performed during night, also the Sun, being a strong continuum emitter, can not be the reason for these waves, which are known to appear occasionally as "solar ripples" during daylight. The origin of the waves, producing strong baseline distortions is still not clear and needs further investigation. These baseline deficiencies remain in the residual baseline after applying frequency switching. As a consequence baseline fitting had to be done for smaller spectral portions using high-order baselines increasing the uncertainty. This could be in principle problematic for broad emission lines, which might partly be treated as baseline feature.

The long-term tests, i.e., the mapping of two sources (Leo T and NGC 2403) were partly successful, though affected by the problems described in the previous paragraphs. Datacubes were compiled and analyzed. In case of NGC 2403 unfortunately parts of the data were rendered completely useless due to a strong distortion of the bandpass the reason for which is unknown. The Leo T datacube led to similar results as one from the comparison measurement with the old 18/21-cm receiver, but having slightly lower quality, due to the higher polynomial order needed to describe the baseline. Consequently, the scientific analysis was done using the data from the old receiver.

CHAPTER 6

Summary

The presented work lays the foundation for the Effelsberg–Bonn H I survey (EBHIS), a new database of the neutral hydrogen covering the complete northern hemisphere of the galactic and extragalactic (local) universe. The EBHIS will be an order of magnititude more sensitive than any previously performed survey (i.e., the LAB survey; Kalberla et al. 2005) and have a much higher angular resolution. It will complement the recently performed HIPASS (Barnes et al. 2001; Meyer et al. 2004) and GASS surveys (McClure-Griffiths et al. 2006) in the southern hemisphere. For completeness it should be noted, that currently two surveys, ALFALFA (extragalactic regime; Giovanelli 2005) and GALFA (galactic sky; Goldsmith 2004; Heiles et al. 2004), are carried out at the Arecibo telescope, which will have comparable sensitivity to the EBHIS but are restricted to a smaller fraction of the sky.

Such a deep and complete measurement is only possible due to the development of a new 21-cm seven-beam receiver (Keller et al. 2006) allowing increased mapping speed. State-of-the-art FPGA-based FFT spectrometers provide the large number of spectral channels (16k) at high bandwidth (100 MHz) which is necessary to reach the required redshift range out to $z \sim 0.07$ maintaining at the same time a high velocity resolution of $1.25\,\mathrm{km\,s^{-1}}$. Furthermore, only this kind of (digital) backend has the properties, i.e., high dynamic range and the ability to dump spectra fast, which are needed for a sophisticated RFI detection scheme.

There is a wide range of scientific questions to be addressed using the EBHIS. These were discussed in detail in Chapter 2. The most prominent drivers are:

1. The measurement of the H I mass function down to H I masses of dwarf galaxies ($M_{\mathrm{HI}} \sim 10^7\,\mathrm{M}_\odot$) and its dependence on environmental conditions and evolution of the universe.

2. The search for tidal debris produced by interaction of galaxies, being an important measure for the (ongoing) structure formation in the universe and helping to understand the complex systems of galaxies and their gaseous halos.

3. The latter point will be supplemented by measurements of our own galaxy, the Milky Way. Many interesting features like the low-, intermediate-, and high-velocity clouds, galactic (super-)shells, and outflows are tracers of the flow of matter between the halo and the disk. It is still unclear, whether the accreted material is mainly produced by the various outflow processes of the Milky Way itself or is of external origin, e.g., being tidally disrupted gas from interactions with satellite galaxies, relics of merging processes in the past, or even primordial gas.

4. The EBHIS database will also be of big interest for future surveys in other wavelength regimes, which have to correct their data for foreground (Milky Way H I) emission to reach the desired sensitivity.

The thesis consists of three main parts, the development of the data reduction pipeline and testing it by the use of simulations and various test measurements. They will be summarized briefly and are followed by an outlook.

6.1 Data reduction software

The basis of this thesis certainly was the development of a sophisticated data reduction pipeline. The huge amount of data which will be collected during the prospected five years of observing needs reduction software which allows high data throughput. Apart from using a "fast" programming language (C/C++) all time-critical procedures were implemented using multithreading techniques, improving computational speed on multi-core or -processor platforms, which are widely in use today. Furthermore, various calibration factors, flags, or other parameters computed during the data reduction will be stored separately in a database. This way in-between stages do not use more disk space than absolutely necessary — usually the processed data are stored after each step in the reduction pipeline.

For the EBHIS a decent RFI detection software is needed. The impact, interferences can have on the quality and sensitivity of a measurement, is severe (Winkel 2005). The algorithm which is proposed, first, performs a feature extraction with respect to the baseline, and second, analyzes the resulting signals statistically to distinct between RFI and astronomical emission. Several tests were carried out revealing that interferences are found down to the $3 \ldots 4\sigma$ level (Winkel et al. 2007). Furthermore, the RFI mitigation was applied successfully to the artificial data generated for the survey simulations. The results showed no significant difference to the case where no RFI was added to the data. All parts of the reduction pipeline can make use of the *flag* database compiled by the detection software. This way, all subsequent tasks can take these bad data points into account rendering the complete data processing more robust against RFI.

A second major issue is the estimation of the systems gain curves. Traditionally, this is done via position- and/or (inband) frequency switching. In this work a new method, *Least-squares frequency switching* (Heiles 2007), is presented. Instead of using only two LO frequencies, a set of different (much smaller) frequency shifts is introduced. The true baseline is calculated (in linear approximation) via an iterative least-squares method. One big advantage is the much smaller fraction of the bandwidth which is "lost", compared to inband frequency switching where only half of the available bandwidth is actually

used. The robustness of the method against instabilities of the system, strong continuum emitters, and RFI was investigated. While LSFS was intrinsically robust against the former issue, the latter two needed improvements (Winkel & Kerp 2007). The proposed modifications make the LSFS especially well-suited for the EBHIS, although it is not finally clear, whether the hardware implementation at the 100-m telescope is easily possible.

For the galactic part of the EBHIS the stray–radiation correction (see Kalberla 1978; Hartmann et al. 1996; Kalberla et al. 2005) is crucial, considering radiation from the MW disk entering the receiver via the sidelobes of the antenna pattern. Mainly, but not exclusively, the Milky Way gaseous disk emission has a large impact on the measured intensity in the affected regions. The algorithms and software to calculate the SR correction were already available. Hence, only a brief summary of the main ideas behind was given. The SR correction was successfully applied to the data recorded during various test measurements, although only the LAB survey data could be used to approximate the true sky brightness distribution, lacking the complete EBHIS survey data at this point. Nevertheless, the results have acceptable precision.

Finally, a source finder algorithm based on the Gamma test (Evans 2002; Boyce 2003) was presented. It is integrated into a graphical user interface, which allows for visualization of datacubes, manual and automatic finding and parametrization of sources, as well as enabling to compute meaningful (statistical) errors. The source finding was utilized in the simulations, showing good results.

6.2 Galaxy simulations

An important part of this thesis was the simulation of an extragalactic dataset. Generating artificial spectra made it possible to investigate the influence of RFI on the completeness and quality of the source catalog compiled. Furthermore, these simulations allowed to test the data reduction pipeline under nearly realistic circumstances. Three different datasets were compiled, one without RFI signals added, one with interferences but no mitigation, and a third dataset incorporating RFI mitigation. The tests revealed that interferences have a disastrous influence on the overall detection rate. However, using the proposed RFI detection scheme, the compiled galaxy catalog shows no significant difference to the ideal case.

From the comparison of the generated and recovered sources, the completeness function could be calculated, which is the basis for computing the HIMF. The completeness properties of the simulations are very similar to the results found by Zwaan et al. (2004). However, as only Gaussian profiles were included (for simplicity) the resulting completeness is not yet applicable to the final EBHIS data. Nevertheless, to implement different velocity profile shapes is not a principle problem. Eventually, it might also be reasonable to seed the artificial sources directly into real data, which would take the complete data reduction pipeline into account.

An interesting aspect regarding the completeness function was that from analyzing only the one-dimensional case, the completeness is much worse, than in the two-dimensional case, which is due to the marginalization over the second (correlated) parameter.

As a sideeffect, having a large sample of sources enabled to analyze potential bias and selection effects, systematically changing the recovered source parameters compared

to the generated ones. Several issues were identified and could be explained. First, the HICAT parametrization scheme is suboptimal in the presence of noise. The peak flux is overestimated, while the profile widths are underestimated, especially for weak sources. Furthermore, filtering the spectra to decrease the noise level changes the determined source parameters (peak flux and width). Second, there is a selection effect due to noise. For sources close to the detection limit (regarding the integrated flux) the flux is overestimated. Finally, the data reduction itself has (small) impact on the parameters. Due to the gridding the effective beam width is slightly enlarged, leading to smaller peak flux. This is not a real issue, since the peak flux does not convert to a physical relevant quantity. Also, very broad and weak profiles show a tendency to underestimate the total flux. This is due to problems to find the correct (sufficiently large) subcube size in the Galaxy Parametrizer task.

Having identified the potential bias due to the HICAT parametrization scheme the next step was to analyze this systematic effect more quantitatively. Artificial spectra were simulated covering (uniformly) a certain parameter space with respect to the peak flux and velocity profile width. By using thousands (100 per parameter space grid cell) of spectra, the statistical error became small leaving only the systematic influence of the parametrization scheme. The smaller the peak flux and the larger the width, the larger the underestimation of the profile width becomes. The overestimation of the peak flux is only slightly dependent on profile width, but gets stronger for lower values of the peak fluxes. For completeness, the analysis was repeated for smoothed (low-pass filtered) spectra, using a Gaussian and a Hanning filter. This has additional impact on the results. For the Gaussian filtering — and using only Gaussian velocity profiles (which is an unrealistic assumption) — a scaling method was developed, which allows to correct for the systematic effects.

Finally, the Kolmogorov-Smirnov (KS) test was applied to address the question whether the recovered parameter distributions (i.e., integrated and peak flux, profile width, and mass) are compatible with the simulated ones. This is the case when considering the completeness function (thresholding above the completeness limits).

6.3 Test observations

In 2007/2008 several test measurements were carried out, using the new multi-feed receiver as well as the FFT spectrometer backend which will be the standard backend for the EBHIS. For comparison also the old single-beam 18/21-cm frontend was utilized for several map observations. The latter was important, to have an independent dataset not influenced by potential issues with the frontend allowing to analyze the performance and quality of the data reduction pipeline.

First tests were performed in November 2007. These allowed to determine the stability and system temperature of the receiver. Especially, the central feed showed strong baseline instabilities leading to an increased system temperature and poor Allan times. The problems can likely be attributed to strong (out-of-band) RFI signals which can more easily enter the central horn, measuring the circular polarization of the incoming radiation. In the meantime new stop-band filters have been implemented into the IF chain, which improved the situation, though no astronomical measurements have been carried

out yet. The offset feeds were measured to have good system temperatures in the range of 20 . . . 25 K. Also their stability and Allan times were acceptable. Another big issue are standing waves between the primary and secondary foci. Their origin is not clear at the moment and needs investigation.

During the following observations in December 2007 two sources of astronomical interest, Leo T and NGC 2403 were mapped. These measurements can be considered successfull, although for NGC 2403 one of the datasets was strongly disturbed by baseline fluctuations, rendering the final result, i.e., the datacube, useless. Nevertheless, the receiver as well as the backend provided data which lead to a clear detection in both cases. Due to the standing waves the baselines had to be fitted piecewise with high-order polynomials which is not very robust against potential broad and weak emission lines in the spectra. That is one of the reasons why the scientific analysis of the two sources was based on the datasets recorded with the old single-beam instrument (in June/July 2008). Apart from that, these comparison measurements were done using a bandwidth of 20 MHz, which led to increased velocity resolution of the spectra.

Using the data reduction pipeline two high-quality data cubes were compiled. Basic physical parameters were deduced from the column densities maps and spectral profiles. As the distance to both objects is known, the H I masses could be determined, being $M_{\mathrm{HI}} = 3.35 \pm 0.01 \cdot 10^9$ $\mathrm{M}_\odot$ for NGC 2403 and $M_{\mathrm{HI}} = 3.6 \cdot 10^5$ $\mathrm{M}_\odot$ for Leo T. Both values are in good agreement with previous single-dish and high-resolution interferometric measurements. In case of Leo T no clear evidence for an outflow was observed, though the spectral lines partly overlap with Milky Way emission and the total integration time was only 1 min limiting the potential to detect such features.

Bibliography

Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, ApJS, 175, 297

Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2007, ApJS, 172, 634

Bach, U., Kraus, A., Fürst, E., & Polatidis, A. 2007, First report about the commissioning of the new Effelsberg sub-reflector, Tech. rep., Max-Planck-Institut für Radioastronomie

Bachiller, R. & Cernicharo, J. 2008, Science with the Atacama Large Millimeter Array: A New Era for Astrophysics (Astrophysics and Space Science, Vol.313, Springer)

Bahcall, J. N. 1966, ApJ, 145, 684

Bahcall, J. N. & Salpeter, E. E. 1965, ApJ, 142, 1677

Barnes, D. G. 1998, in ASP Conf. Ser. 145: Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse, 32–+

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

Basilakos, S., Plionis, M., Kovač, K., & Voglis, N. 2007, MNRAS, 378, 301

Ben Bekhti, N., Richter, P., Westmeier, T., & Murphy, M. T. 2008, A&A, 487, 583

Benz, A. O., Grigis, P. C., Hungerbühler, V., et al. 2005, A&A, 442, 767

Bhat, N. D. R., Cordes, J. M., Chatterjee, S., & Lazio, T. J. W. 2005, Radio Science, 40, 5

Blitz, L., Spergel, D. N., Teuben, P. J., Hartmann, D., & Burton, W. B. 1999, ApJ, 514, 818

Bonamente, M., Swartz, D. A., Weisskopf, M. C., & Murray, S. S. 2008, ApJ, 686, L71

Boyce, P. J. 2003, Master's thesis, School of Computer Science, Cardiff University

Bradley, R. & Barnbaum, C. 1996a, in Bulletin of the American Astronomical Society, Vol. 28, Bulletin of the American Astronomical Society, 1418–+

Bradley, R. & Barnbaum, C. 1996b, Bulletin of the American Astronomical Society, 28

Braun, R. & Burton, W. B. 1999, A&A, 341, 437

Braun, R. & Thilker, D. A. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 331, Extra-Planar Gas, ed. R. Braun, 121–+

Bregman, J. N. 1980, ApJ, 236, 577

Briggs, F. H., Bell, J. F., & Kesteven, M. J. 2000, AJ, 120, 3351

Brüns, C., Kerp, J., Kalberla, P. M. W., & Mebold, U. 2000, A&A, 357, 120

Brüns, C., Kerp, J., & Pagels, A. 2001, A&A, 370, L26

Brüns, C., Kerp, J., Staveley-Smith, L., et al. 2005, A&A, 432, 45

Brüns, C. & Westmeier, T. 2004, A&A, 426, L9

Calabretta, M. R. & Greisen, E. W. 2002, A&A, 395, 1077

Carignan, C., Chemin, L., Huchtmeier, W. K., & Lockman, F. J. 2006, ApJ, 641, L109

Carilli, C. & Rawlings, S. 2004, Science with the Square Kilometre Array (New Astronomy Reviews, Vol.48, Elsevier)

Cerny, V. 1985, Journal of Optimization Theory and Applications, 45, 41

Chevalier, R. A. & Clegg, A. W. 1985, Nature, 317, 44

Chynoweth, K. M., Langston, G. I., Yun, M. S., et al. 2008, AJ, 135, 1983

Cooper, J. L., Bicknell, G. V., Sutherland, R. S., & Bland-Hawthorn, J. 2008, ApJ, 674, 157

Cunha, J. V., Marassi, L., & Lima, J. A. S. 2007, MNRAS, 379, L1

Dale, D. A., Barlow, R. J., Cohen, S. A., et al. 2008, AJ, 135, 1412

Dame, T. M., Hartmann, D., & Thaddeus, P. 2001, ApJ, 547, 792

de Blok, W. J. G., Walter, F., Brinks, E., Thornley, M. D., & Kennicutt, Jr., R. C. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 329, Nearby Large-Scale Structures and the Zone of Avoidance, ed. A. P. Fairall & P. A. Woudt, 265–+

di Serego Alighieri, S., Gavazzi, G., Giovanardi, C., et al. 2007, A&A, 474, 851

Dobbs, M., Halverson, N. W., Ade, P. A. R., et al. 2006, New Astronomy Review, 50, 960

Duda, R. O., Hart, P. E., & Stork, D. G. 2001, Pattern classification (Wiley, New York)

Evans, D. 2002, PhD thesis, Department of Computer Science, Cardiff University

Falcke, H. D., van Haarlem, M. P., de Bruyn, A. G., et al. 2007, Highlights of Astronomy, 14, 386

Ferrando, P. 2002, in SF2A-2002: Semaine de l'Astrophysique Francaise, ed. F. Combes & D. Barret, 271–+

Ferrière, K. M. 2001, Reviews of Modern Physics, 73, 1031

Fisher, J. R. 2002, in ASP Conf. Ser. 278: Single-Dish Radio Astronomy: Techniques and Applications, 433–445

Fraternali, F. & Binney, J. J. 2006, MNRAS, 366, 449

Fraternali, F., van Moorsel, G., Sancisi, R., & Oosterloo, T. 2002, AJ, 123, 3124

Fridman, P. A. 2001, A&A, 368, 369

Garcia-Appadoo, D. A. 2005, PhD thesis, Cardiff University

Giovanelli, R. 2005, in Bulletin of the American Astronomical Society, Vol. 37, Bulletin of the American Astronomical Society, 1488–+

Goldsmith, P. F. 2004, in Bulletin of the American Astronomical Society, Vol. 36, Bulletin of the American Astronomical Society, 1475–+

Greisen, E. W. & Calabretta, M. R. 2002, A&A, 395, 1061

Greisen, E. W., Calabretta, M. R., Valdes, F. G., & Allen, S. L. 2006, A&A, 446, 747

Grossi, M., Giovanardi, C., Corbelli, E., et al. 2008, A&A, 487, 161

Güsten, R., Booth, R. S., Cesarsky, C., et al. 2006, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Vol. 6267, Ground-based and Airborne Telescopes. Edited by Stepp, Larry M.. Proceedings of the SPIE, Volume 6267, pp. 626714 (2006).

Haffner, L. M., Reynolds, R. J., Tufte, S. L., et al. 2003, ApJS, 149, 405

Hall, P. 2005, The SKA: an engineering perspective (Springer)

Hanisch, R. J., Farris, A., Greisen, E. W., et al. 2001, A&A, 376, 359

Hartmann, D., Kalberla, P. M. W., Burton, W. B., & Mebold, U. 1996, A&AS, 119, 115

Haud, U. & Kalberla, P. M. W. 2007, A&A, 466, 555

Haynes, M. P., Giovanelli, R., & Kent, B. R. 2007, ApJ, 665, L19

Heald, G. H., Rand, R. J., Benjamin, R. A., & Bershady, M. A. 2007, ApJ, 663, 933

Heiles, C. 2007, PASP, 119, 643

Heiles, C., Goldston, J., Mock, J., et al. 2004, in Bulletin of the American Astronomical Society, Vol. 36, Bulletin of the American Astronomical Society, 1476–+

Hochgürtel, S., Bertoldi, F., & Klein, B. 2008, in prep.

Hoffman, G. L., Salpeter, E. E., & Hirani, A. 2004, AJ, 128, 2932

Irwin, M. J., Belokurov, V., Evans, N. W., et al. 2007, ApJ, 656, L13

Jansky, K. G. 1932, in Proceedings of the IRE, Vol. 20, 1920–+

Kalberla, P. M. W. 1978, PhD thesis, University of Bonn

Kalberla, P. M. W., Burton, W. B., Hartmann, D., et al. 2005, A&A, 440, 775

Kalberla, P. M. W., Dedes, L., Kerp, J., & Haud, U. 2007, A&A, 469, 511

Kamphuis, P., Peletier, R. F., Dettmar, R.-J., et al. 2007, A&A, 468, 951

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

Kennicutt, Jr., R. C., Armus, L., Bendo, G., et al. 2003, PASP, 115, 928

Kerp, J. 2003, Astronomische Nachrichten, 324, 69

Kerp, J., Burton, W. B., Egger, R., et al. 1999, A&A, 342, 213

Kerp, J., Lesch, H., & Mack, K.-H. 1994, A&A, 286, L13

Kerp, J. & Pietz, J. 1998, in Lecture Notes in Physics, Berlin Springer Verlag, Vol. 506, IAU Colloq. 166: The Local Bubble and Beyond, ed. D. Breitschwerdt, M. J. Freyberg, & J. Truemper, 337–340

Kilborn, V. A. 2002, in Astronomical Society of the Pacific Conference Series, Vol. 276, Seeing Through the Dust: The Detection of HI and the Exploration of the ISM in Galaxies, ed. A. R. Taylor, T. L. Landecker, & A. G. Willis, 80–+

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. 1983, Science, 220, 671

Klein, B., Krämer, I., & Wielebinski, R. 2005, URSI General Assembly

Klein, B., Philipp, S. D., Krämer, I., et al. 2006, A&A, 454, L29

Klypin, A., Kravtsov, A. V., Valenzuela, O., & Prada, F. 1999, ApJ, 522, 82

Koribalski, B. S., Staveley-Smith, L., Kilborn, V. A., et al. 2004, AJ, 128, 16

Levenberg, K. 1944, Quart. Appl. Math., 2, 164

Levine, E. S., Blitz, L., & Heiles, C. 2006a, Science, 312, 1773

Levine, E. S., Blitz, L., & Heiles, C. 2006b, ApJ, 643, 881

Levine, E. S., Heiles, C., & Blitz, L. 2008, ApJ, 679, 1288

Liszt, H. 1997, A&AS, 124, 183

Lockman, F. J. 2002, ApJ, 580, L47

Lockman, F. J., Jahoda, K., & McCammon, D. 1986, ApJ, 302, 432

Madore, B. F. & Freedman, W. L. 1991, PASP, 103, 933

Marquardt, D. 1963, SIAM J. Appl. Math., 11, 431

Mayer, L., Governato, F., & Kaufmann, T. 2008, ArXiv e-prints, 801

McClure-Griffiths, N. M., Dickey, J. M., Gaensler, B. M., et al. 2005, ApJS, 158, 178

McClure-Griffiths, N. M., Ford, A., Pisano, D. J., et al. 2006, ApJ, 638, 196

McClure-Griffiths, N. M., Staveley-Smith, L., Lockman, F. J., et al. 2008, ApJ, 673, L143

Melioli, C., Brighenti, F., D'Ercole, A., & de Gouveia Dal Pino, E. M. 2008, MNRAS, 388, 573

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

Milgrom, M. 1983, ApJ, 270, 365

Minchin, R., Davies, J., Disney, M., et al. 2007, ApJ, 670, 1056

Miville-Deschênes, M.-A., Boulanger, F., Martin, P. G., et al. 2006, ArXiv Astrophysics e-prints

Moore, B., Ghigna, S., Governato, F., et al. 1999, ApJ, 524, L19

Morandi, A., Ettori, S., & Moscardini, L. 2007, MNRAS, 379, 518

Morganti, R., Tadhunter, C., Oosterloo, T., Holt, J., & Emonts, B. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 373, The Central Engine of Active Galactic Nuclei, ed. L. C. Ho & J.-W. Wang, 343–+

Muders, D. 2007, Multi-Beam FITS (MBFITS) Raw Data Format Summary, Tech. rep., Max-Planck-Institute für Radioastronomie, Bonn

Nakanishi, H., Sofue, Y., & Koda, J. 2005, PASJ, 57, 905

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Nidever, D. L., Majewski, S. R., & Burton, W. B. 2008, ApJ, 679, 432

Nyquist, H. 1928, Trans. AIEE, 47, 617

Oosterloo, T., Fraternali, F., & Sancisi, R. 2007a, AJ, 134, 1019

Oosterloo, T. A., Morganti, R., Sadler, E. M., van der Hulst, T., & Serra, P. 2007b, A&A, 465, 787

Pietz, J., Kerp, J., Kalberla, P. M. W., et al. 1996, A&A, 308, L37

Pradas, J., Kerp, J., Kalberla, P. M. W., et al. 2004, in Astronomical Society of the Pacific Conference Series, Vol. 317, Milky Way Surveys: The Structure and Evolution of our Galaxy, ed. D. Clemens, R. Shah, & T. Brainerd, 29–+

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical recipes in C. The art of scientific computing (Cambridge: University Press, 1992, 2nd ed.)

Putman, M. E., Grcevich, J., & Peek, J. E. G. 2008, in American Institute of Physics Conference Series, Vol. 1035, The Evolution of Galaxies Through the Neutral Hydrogen Window, ed. R. Minchin & E. Momjian, 141–146

Reber, G. 1940, ApJ, 91, 621

Regulierungsbehörde für Post und Telekommunikation. 2004, Frequenznutzungsplan der Bundesrepublik Deutschland

Reich, R. 2006, RFI Abschirmungsmassnahmen der LOFAR Elektronik in Effelsberg, Tech. rep., Max-Planck-Institut für Radioastronomie

Roberts, M. S. 1962, AJ, 67, 437

Rohlfs, K. & Wilson, T. L. 1996, Tools of Radio Astronomy (Tools of Radio Astronomy, XVI, 423 pp. 127 figs., 20 tabs.. Springer-Verlag Berlin Heidelberg New York. Also Astronomy and Astrophysics Library)

Rots, A. H. 1980, A&AS, 41, 189

Ryan-Weber, E. V., Begum, A., Oosterloo, T., et al. 2008, MNRAS, 384, 535

Saintonge, A. 2007, AJ, 133, 2087

Sancisi, R., Fraternali, F., Oosterloo, T., & van der Hulst, T. 2008, A&A Rev., 15, 189

Schechter, P. 1976, ApJ, 203, 297

Schneider, P. 2006, Extragalactic Astronomy and Cosmology (Extragalactic Astronomy and Cosmology, by Peter Schneider. Berlin: Springer, 2006.)

Shannon, C. E. 1949, Proc. Institute of Radio Engineers, 37, 10

Shapiro, P. R. & Field, G. B. 1976, ApJ, 205, 762

Slane, P. O., Romaine, S., Murray, S. S., et al. 2008, in AAS/High Energy Astrophysics Division, Vol. 10, AAS/High Energy Astrophysics Division, 28.01–+

Sofue, Y., Koda, J., Nakanishi, H., et al. 2003, PASJ, 55, 17

Sormann, H. 2008, Skript zur Vorlesung Numerische Methoden in der Physik, TU Graz

Stanimirović, S., Hoffman, S., Heiles, C., et al. 2008, ApJ, 680, 276

Stanimirović, S., Putman, M., Heiles, C., et al. 2006, ApJ, 653, 1210

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

Stil, J. M., Lockman, F. J., Taylor, A. R., et al. 2006a, ApJ, 637, 366

Stil, J. M., Taylor, A. R., Dickey, J. M., et al. 2006b, AJ, 132, 1158

Stil, J. M., Taylor, A. R., Martin, P. G., et al. 2004, ApJ, 608, 297

Taylor, A. R., Gibson, S. J., Peracaula, M., et al. 2003, AJ, 125, 3145

Taylor, A. R., Stil, J. M., Dickey, J. M., et al. 2002, in Astronomical Society of the Pacific Conference Series, Vol. 276, Seeing Through the Dust: The Detection of HI and the Exploration of the ISM in Galaxies, ed. A. R. Taylor, T. L. Landecker, & A. G. Willis, 68–+

Thilker, D. A., Braun, R., Walterbos, R. A. M., et al. 2004, ApJ, 601, L39

Thom, C., Peek, J. E. G., Putman, M. E., et al. 2008, ApJ, 684, 364

Tufte, S. L., Wilson, J. D., Madsen, G. J., Haffner, L. M., & Reynolds, R. J. 2002, ApJ, 572, L153

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

van Driel, W. 2007, in The First MCCT-SKADS Training School, Medicina Bologna, Italy

van Driel, W., Schneider, S. E., Lehnert, M., & Minchin, R. 2008, in American Institute of Physics Conference Series, Vol. 1035, The Evolution of Galaxies Through the Neutral Hydrogen Window, ed. R. Minchin & E. Momjian, 256–258

van Woerden, H. 1962, PhD thesis, University of Groningen

Weijmans, A.-M., Krajnović, D., van de Ven, G., et al. 2008, MNRAS, 383, 1343

Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, A&AS, 44, 363

Westmeier, T. 2007, PhD thesis, Argelander-Institut für Astronomie, Bonn University

Westmeier, T., Braun, R., Brüns, C., Kerp, J., & Thilker, D. A. 2007, New Astronomy Review, 51, 108

Westmeier, T., Braun, R., & Thilker, D. 2005, A&A, 436, 101

Whittaker, E. T. 1915, Proc. Royal Soc. Edinburgh, 35, 181

Winkel, B. 2005, Master's thesis, Radioastronomisches Institut, University of Bonn

Winkel, B. & Kerp, J. 2007, ApJS, 173, 166

Winkel, B., Kerp, J., Kalberla, P. M. W., & Keller, R. 2008, in American Institute of Physics Conference Series, Vol. 1035, The Evolution of Galaxies Through the Neutral Hydrogen Window, ed. R. Minchin & E. Momjian, 259–261

Winkel, B., Kerp, J., & Stanko, S. 2007, Astronomical Notes, 328, 68

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

Zwaan, M. A., Meyer, M. J., Staveley-Smith, L., & Webster, R. L. 2005, MNRAS, 359, L30

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, AJ, 125, 2842