rijksuniversiteit
groningen

Algorithms for Radio Interference Detection and Removal

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 22 juni 2012
om 11.00 uur

door

Anne René Offringa

geboren op 8 mei 1982
te Hardenberg

Promotores:	Prof. dr. A.G. de Bruyn Prof. dr. S. Zaroubi Prof. dr. M. Biehl
Beoordelingscommissie:	Prof. dr. J.M. van der Hulst Prof. dr. F.H. Briggs Prof. dr. W.N. Brouw

ISBN 978-90-367-5545-0
ISBN 978-90-367-5544-3 (electronic version)

Contents

Chapter 1

Introduction

WHILE ASTRONOMY is one of the oldest sciences in existence, the *radio* sky was not discovered before 1931, the year when Karl Jansky first detected the radio signals from our Galaxy (Jansky, 1933). This discovery revealed a mysterious new part of our Universe that had been hidden ever before. A lot of exciting discoveries followed, many of which had a large impact on the field of astronomy and our knowledge of the Universe, such as the detection of the hydrogen spectral line at a wavelength of 21 cm (Ewen and Purcell, 1951), the detection of radiation of the cosmological microwave background (CMB) (Dicke et al., 1965) and the discovery of an entirely new class of sources: the pulsar (Hewish et al., 1968), a class of neutron stars that sends radio wave pulses that are similar to the beam of a lighthouse. The field of radio astronomy has been flourishing for decades, and scientists continue to stretch the limits on the possibilities of observing in the radio domain. With better telescopes, faster computers and more knowledge, the future of radio astronomy is (radio) bright.

Modern telescopes in radio astronomy, such as the Expanded Very Large Array (EVLA) in New Mexico, the Low-Frequency Array (LOFAR) in the Netherlands and the Murchison Widefield Array (MWA) in Australia, are incredible sensitive devices that observe the sky with enormous depth and detail. The observed bandwidth of telescopes has dramatically increased over the last decades, and often overlaps with parts of the radio spectrum that have not been reserved for radio astronomy. Simultaneously, the radio spectrum is becoming more crowded because of technological advancement. Therefore, radio observations are affected by man-made radio transmitters, which can be several orders of magnitude stronger than the weak celestial signals of interest. This kind of interference, which seriously disturbs radio observations, is called radio-frequency interference (RFI).

If the frequency at which is observed contains RFI, thus overlaps with the frequency at which other devices transmit, the recorded data will be corrupted and can not be used directly. For example, because RFI can be many orders of magnitude stronger than the signal of interest, it might not be possible to calibrate contaminated data. Moreover, because of the difference in strength, the signal of interest will be overshadowed by the RFI, and the signal can not be extracted.

Around 1980 when the radio spectrum was becoming more and more occupied as a result of technical advancements (Pankonin and Price, 1981), radio observers started to notice RFI caused

by electronic equipment (Thompson et al., 1991), and started to develop methods to mitigate it. The first techniques to deal with contaminated data were performed by the data reducing scientist, and consisted of manual selection of good data. Examples are to manually remove data from particular antennae or time and frequency ranges at which the interference was received. Although these techniques can be tedious, they have been sufficient in most situations during the last few decades. Building telescopes at radio quiet sites would improve this situation, but this is not always feasible.

Now that the radio telescopes of the next generation are coming into operation, the dawn of software-driven telescopes producing terabyte sized data sets has begun. Because of that, data reduction in radio astronomy is entering a new era in which more emphasis is put on automated data processing and pipelining the various steps in the data reduction. Dealing with RFI is an important step in the reduction process. As the volume of data and the required sensitivity of observations increases significantly, and the contamination of RFI through an increased usage of electronic equipment grows, more sophisticated automated flagging procedures are required for the next generation of telescopes.

1.1 Radio astronomy & its instruments

Radio astronomy concerns the observation and analysis of extraterrestrial electromagnetic radiation at radio frequency. Electromagnetic radiation is defined to be of radio frequency when it has a frequency up to 300 GHz, which corresponds with a wavelength $\lambda \geq 1$ mm. An impression of the electromagnetic spectrum is given in Fig. 1.1. Frequencies between 30 MHz and 30 GHz can easily penetrate the atmosphere (including troposphere and ionosphere) of the Earth, and observing the sky at these frequencies is therefore possible from the ground.

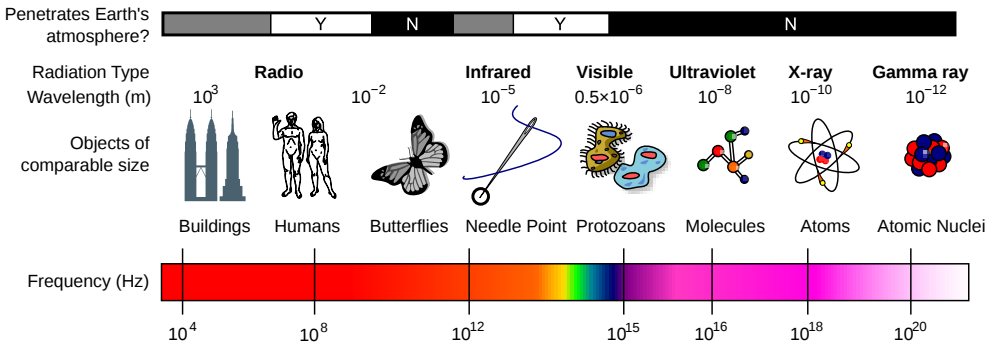


Figure 1.1: *An impression of the electromagnetic spectrum and the wavelength scale, along with whether a given frequency penetrates Earth's atmosphere. (Source: derivation of a Wikipedia and NASA image.)*

Radio waves can be received with an antenna, that converts electromagnetic radiation into an electric current. The same antenna can also work as a transmitter by feeding current into the antenna. Since celestial objects are far away and its radiation is often faint, for many radio-astronomical observations it is necessary to collect large amounts of radiation. Traditionally, this

is performed with large dishes, which optics are optimized to reflect the radiation towards a feed. The increased size of dishes not only increases the sensitivity of such a radio telescope, but simultaneously increases the resolution of the instrument, thereby revealing finer details of the celestial objects that are observed. Large telescopes are therefore common in radio astronomy. Well known examples are the 305-m dish of the Arecibo Observatory near the city of Arecibo (Puerto Rico), the 100-m Robert C. Byrd Green Bank Telescope (GBT) of the National Radio Astronomy Observatory (NRAO) at Green Bank (USA), the Effelsberg 100-m Radio Telescope, named after the nearby village of Effelsberg (Germany), the 76-m dish of the Jodrell Bank Observatory near Goostrey (UK) and the 64-m dish of the Parkes Observatory in New South Wales (Australia).

Around 1946 it was discovered that interferometry could be used to perform radio aperture synthesis. Initially this was performed with a sea cliff-based observatory that consisted of a single antenna, called a sea interferometer, to analyse the sun (McCready et al., 1947). Many modern observatories have multiple dishes to utilize interferometry and synthesis a large aperture to create high resolution images. The best known instruments of this kind include the Karl G. Jansky Very Large Array (VLA) near Socorro (New Mexico), the Westerbork Synthesis Radio Telescope (WSRT) in Westerbork (the Netherlands), the Australia Telescope Compact Array (ATCA) near Narrabri (Australia), the Giant Metrewave Radio Telescope (GMRT) near Pune (India) and the Atacama Large Millimeter/sub-millimeter Array (ALMA) in the Atacama desert (Chile).

For low-frequency observations, e.g. around 150 MHz, dishes are not very cost effective. This is because at low frequencies, large dishes do not provide as much sensitivity benefit as at higher frequencies, because their gain is inverse proportional to λ^2. However, advances in technology caused the receiver chain of an antenna to become cheaper, and it was found that connecting many small and cheap antennae together is a cost effective method to build a high resolution, high sensitivity telescope. The Low-Frequency Array is such a next generation telescope, and is currently the largest connected interferometer in the world. The Low-Frequency Array is the key instrument in this thesis and will be briefly described in the next section. Other instruments that consist of many small antennae include the Murchison Widefield Array (MWA) in Western Australia, the Precision Array for Probing the Epoch of Reionization (PAPER) whose (primary) location will be the Karoo Desert of South Africa, and the Long Wavelength Array (LWA) that is build near the VLA in New Mexico. Finally, an important future instrument will be the Square Kilometre Array (SKA), that is to be built in either Australia or South Africa. This telescope is an international collaboration of many countries. Its planned sensitivity and spatial resolution will outperform LOFAR by an order of magnitude at the same frequency. Its first operation is scheduled to start around 2019.

1.2 The Low-Frequency Array

The Low-Frequency Array (LOFAR) is a new antenna array that observes the sky from 10–90 and 110–240 MHz. It consists currently of 41 (validated) stations, while 7 more are planned and more might follow. Of the validated stations, 33 stations are located in the Netherlands and 5 in Germany. Sweden, the UK and France contain one station each. A Dutch station consists of a field of 96 dipole low-band antennae (LBA) that provide the 10–90 MHz range, and one or two fields of in total 48 tiles of 4x4 dipole high-band antennae (HBA) for the 110–240 MHz. The international stations have an equal amount of LBA antennae, but 96 HBA tiles. Deployed antennae of both kinds are displayed in Fig. 1.2. For the latest information about LOFAR, we refer the reader to

the LOFAR website[1].

Figure 1.2: *Antenna types of the Low-Frequency Array. Left image: A low-band antenna with a cabin in the background. Right image: Part of a high-band antenna station, consisting of 24 tiles of 4 × 4 high-band antennae.*

The core area of LOFAR is located near the village of Exloo in the Netherlands, where the density of the stations is higher. The six most densely packed stations are on the Superterp, an elevated area surrounded by water. It is an artificial peninsula of about 350 m in diameter that is situated about 3 km North of Exloo. A map of LOFAR's surroundings is given in Fig. 1.3. Exloo is a village in the municipality of Borger-Odoorn in the province of Drenthe. Drenthe is mostly a rural area, and is, relative to the rest of the Netherlands, sparsely populated, with an average density of 183 persons/km^2 over 2,680 km^2 in 2011[2]. Nevertheless, the radio quiet zone of 2 km around the Superterp is relatively small and households live within 1 km of the Superterp. The distance from households to the other stations is even smaller in certain cases. Therefore, contamination of the radio environment by man-made electromagnetic radiation was a major concern for LOFAR (Bregman, 2000; Bentum et al., 2008). Because this radiation interferes with the celestial signal of interest, it is referred to as radio-frequency interference (RFI). Such radiation can originate from equipment that radiates deliberately, such as citizens' band (CB) radio devices and digital video or audio broadcasting (DVB or DAB), but can also be due to unintentionally radiating devices such as cars, electrical fences, power lines or wind turbines (Bentum et al., 2010).

While RFI mitigation methods in this work are generic methods that can be applied to any interferometer — and some also on single dish receivers — they were largely developed for and

[1]The website of LOFAR is `http://www.lofar.org/`

[2]From the website of the province of Drenthe, `http://www.provincie.drenthe.nl/`.

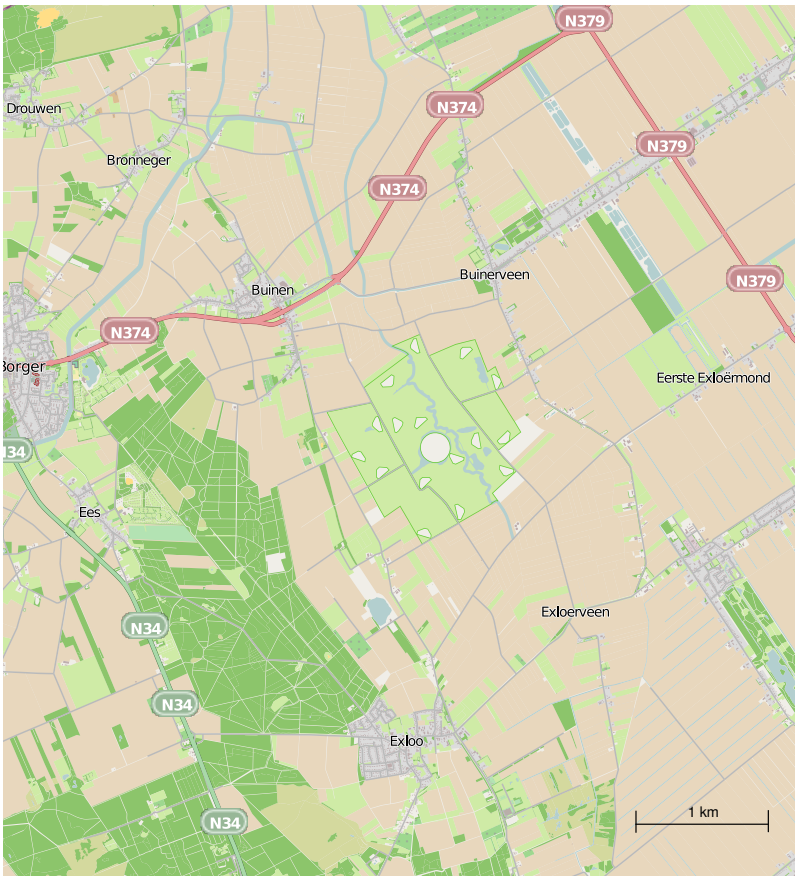


Figure 1.3: *Map of the LOFAR core and its surroundings. The circular peninsula in the centre is the Superterp. Several other stations are visible as well. (source: OpenStreetMap)*

tested with the Low-Frequency Array. When this project started in 2008, some LOFAR test-stations were ready. However, the final stations were not ready, and consequently no representable data was yet available. Over the course of the project, more and more stations became available, and the required representable data became available as well. Fortunately, the low-frequency front-end (LFFE) of the Westerbork Synthesis Radio Telescope provides the frequency range 115–180 MHz, and since it is situated near LOFAR, its radio environment is similar. Therefore, data from the WSRT LFFE was also used in this work to analyse RFI methods and the low-frequency radio environment. For details about the WSRT LFFE system we refer to van der Marel et al. (2005).

1.2.1 LOFAR's key science projects

When the LOFAR project started, four projects were formulated that would drive its design. Later, two more projects were added, led by German groups. These projects are called the key science projects (KSPs) of LOFAR. The following key projects have been formed:

- *The Epoch of Reionisation project* — This project tries to achieve the first statistical detection of the highly redshifted 21-cm signal from neutral hydrogen, from a pivotal period of our Universe named the Epoch of Reionisation. During this era after the Big Bang, the first stars were formed. These stars heated the surrounding gas, which went from neutral into ionized state. An introduction of the Epoch of Reionization will be given in §1.3.
- *Deep extragalactic surveys* — Because of LOFAR's wide field-of-view and very wide frequency coverage, it will be an excellent instrument to survey the low-frequency sky. It is expected that these surveys will find high redshift ($z \geq 6$) galaxies that might provide the missing information on the forming of the first objects, such as galaxies, clusters and black holes. Moreover, the surveys will target clusters, galaxies and early star formation processes.
- *Searching for transient sources* — LOFAR's wide field-of-view also allows to periodically observe large parts of the sky to search for explosive events or other variability. Hereby, one hopes to discover new phenomena, that will be quickly followed up with observations from other instruments. An important class of transient sources are pulsars, and their discovery and analysis will also be one of the targets in this project (Stappers et al., 2011).
- *Detection of ultra-high energy cosmic rays* — Cosmic rays are mysterious pulses from space of which both the origin and cause are mostly unknown. In this project, LOFAR will be used to detect the radio emission that is caused when ultra-high energy cosmic rays (UHECRs) hit Earth's atmosphere. Ones hopes to gain more information from this and explain some of these events.
- *Solar science and space weather* — Although the Sun is a relative close neighbour, still a lot can be learned from it. In this project, the Sun will be studied to gain more information about it, for example to provide forecasts of the space weather, i.e., the activity of the Sun and its influence on devices on and around Earth.
- *Cosmic magnetism* — Magnetism is an important process in our Universe, and plays a role in the evolution of galaxies and clusters. However, what this role is, is not yet very well understood. In this project, LOFAR will be used to detect synchrotron radio waves, that are caused by magnetic fields in our cosmos. This will provide important information about the role of magnetism.

Although these key projects have been formed, LOFAR is a very generic instrument and will explore many more fields. Because LOFAR is an instruments that will explore a large and mostly unexplored parameter space, one can also expect some serendipitous discoveries.

1.3 The Epoch of Reionisation

More than thirteen billion years ago, the Universe as we know it was created during the Big Bang. After this starting point of our Universe, the Universe underwent several stages. Fig. 1.4 shows

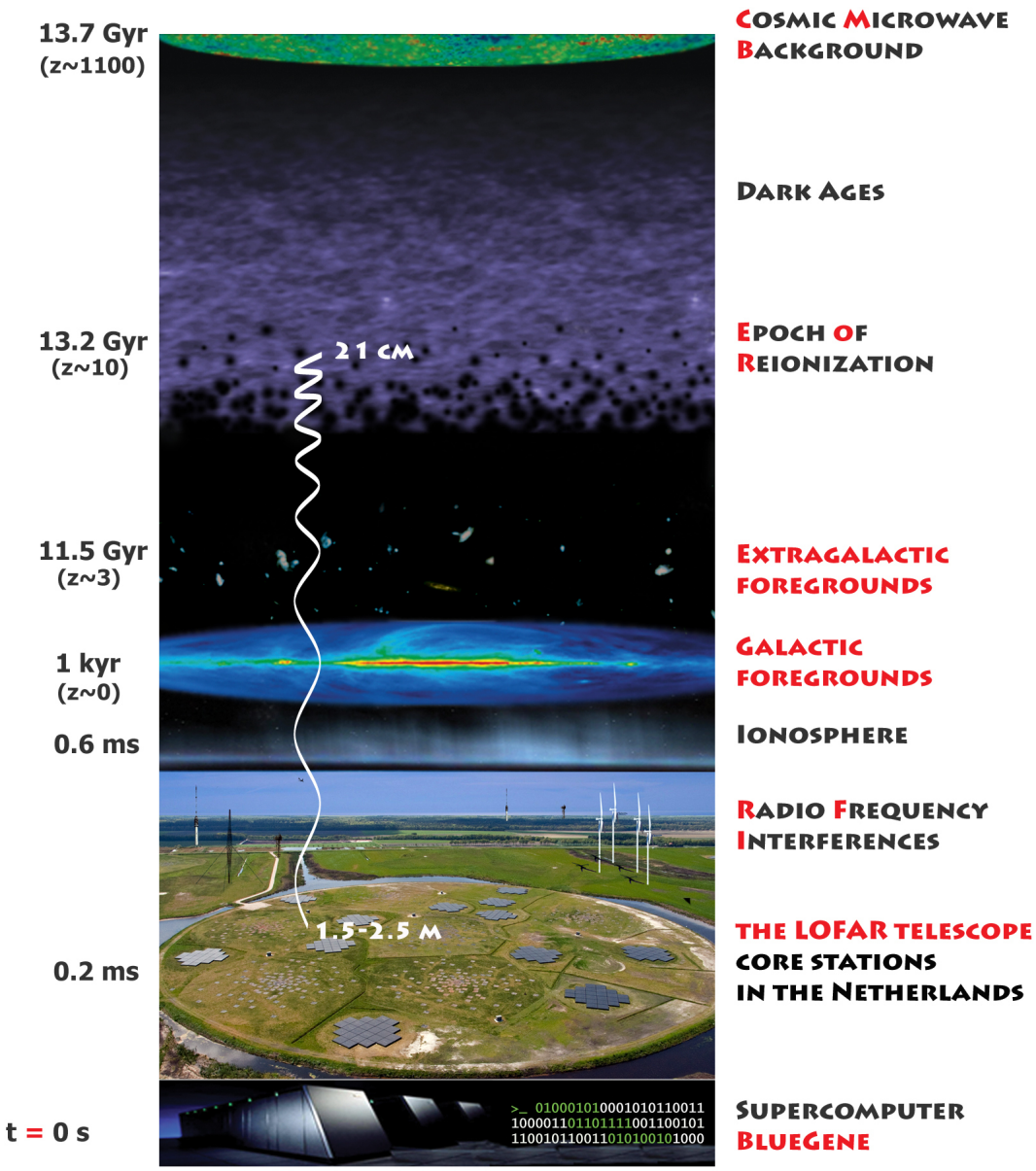


Figure 1.4: *The path of the signal of the Epoch of Reionisation. The RFI generating towers are actually further from the core of LOFAR. Figure by Vibor Jelić.*

the emitted light during the various stages, which will now be explained one by one.

Following the Big Bang, the resulting hot Universe went through a process called inflation, a quick and exponential expansion of the Universe. Not more than 10^{-3} s after the Big Bang, the Universe consisted of expanding matter that gradually started to cool down. Matter in the Universe evolved, and about 380 000 years after the Big Bang, the Universe recombined allowing radiation to travel freely. The following era is referred to as the Dark Ages, because very little radiation was emitted during this period. However, the radiation that was emitted before this era can travel through this transparent Universe, allowing us to observe radiation from before the Dark Ages. Because of the expansion of the Universe, this early radiation has been redshifted. Although the emitted wavelength of this light dominates in the ultraviolet region, it will be redshifted towards mm scale, which is referred to as microwave radiation. The observable radiation is therefore called the cosmological microwave background (CMB) radiation, and has first been observed in 1964 (Dicke et al., 1965). After that, new experiments followed that measured the CMB radiation to an extreme precision, such as the Cosmic Background Explorer (COBE) satellite project (Mather et al., 1990). This directly observed relic radiation is consequently one of the best proves for the Big Bang theory.

About 400 million years after the Big Bang, objects started to form because of the effects of gravity. The objects, including stars and black holes, started to heat their surroundings. As a result, the warmed matter around these objects — consisting mainly of hydrogen — started to ionize. Initially, only spherical "bubbles" around these objects were ionized, but these bubbles expanded until finally the Universe would have been fully ionized. This important era of our Universe is called the Epoch of Reionisation (EoR). It is that era that we try to detect in the LOFAR EoR project. It is detectable because the neutral hydrogen will emit photons as a result of its spin-flip transition. Ionized hydrogen, however, will not emit such photons, and this therefore allows us to detect the transition of the epoch. These Epoch-of-Reionisation photons will be redshifted from their original 21 cm wavelength or 1420 MHz to frequencies of around 150 MHz, thus can be observed with LOFAR.

As shown in Fig. 1.4, this signal is contaminated by several foregrounds, which makes detection very challenging. In fact, in Jelić (2010) it is compared with finding a needle in a haystack. A first step will be to simulate the involved processes, which includes both the reionisation (Thomas et al., 2009) and the extra-galactic and Galactic foregrounds (Jelić et al., 2008). These need to be simulated together with the instrumental response of LOFAR (Labropoulos, 2010). From there, one can test whether the signal will be detectable. To this end, various detection strategies have been proposed, and as it turns out, perhaps the first detection of the Epoch of Reionisation with LOFAR will be a statistical detection (Harker et al., 2009b). Many different techniques have been proposed for this task (Santos et al., 2005; McQuinn et al., 2006; Harker et al., 2009a, 2010; Chapman et al., 2012).

To explore the low-frequency sky and the foregrounds, and to select appropriate fields in the sky to observe the signal with LOFAR, initial experiments at the relevant frequencies have already been performed with the WSRT (Bernardi et al., 2009, 2010; de Bruyn and Bernardi, 2009). Quite recently, the first few LOFAR observations on the road to EoR detection have started (de Bruyn et al., 2011), and at the time of writing, regular LOFAR EoR observations to accumulate the required 100 nights of 6 hour observations are planned to start late 2012. An important step in the extraction of the signal from the data, is the removal of man-made interference. This step will be the main focus of this thesis.

The LOFAR EoR project is not the only project that is trying to achieve detection of the

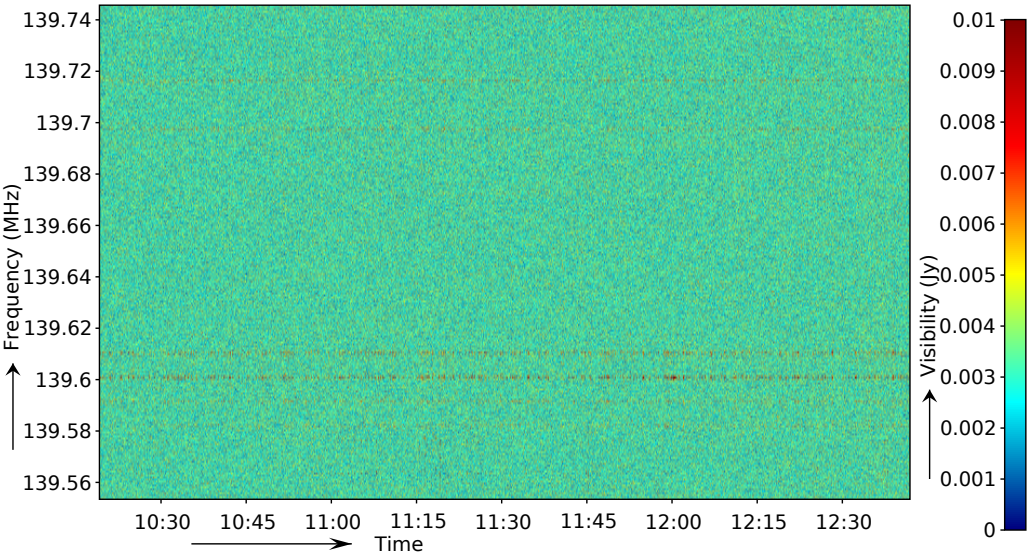


Figure 1.5: *A typical example of RFI in a sub-band of a LOFAR observation. While most of the time-frequency diagram is noise-like, the repetitive higher (red) values at constant frequencies are due to narrow-band RFI.*

EoR using the 21-cm radiation. Other projects use instruments such as the GMRT (Paciga et al., 2011), the MWA (Ord et al., 2010), PAPER (Jacobs et al., 2011) and the Experiment to Detect the Global EoR Step (EDGES) (Bowman and Rogers, 2010). Nevertheless, LOFAR is progressing very quickly, and it may well be that LOFAR will be the first instrument to detect signals from the EoR.

1.4 RFI mitigation techniques

Numerous techniques have been suggested to perform the challenging task of the excision or mitigation of radio-frequency interference from the data. They include using spatial information to null directions, provided in interferometers or multi-feed systems (Leshem et al., 2000; Ellingson and Hampson, 2002; Smolders and Hampson, 2002; Kocz et al., 2010); removing the RFI by using reference antennae (Barnbaum and Bradley, 1998); and blanking out unlikely high values at high time resolutions with the `CUSUM` method (Baan et al., 2004) or other thresholding techniques (Weber et al., 1997; Leshem et al., 2000; Niamsuwan et al., 2005). The following subsections will elaborate on some of the methods that are regularly used in the field of radio-astronomical interference mitigation.

RFI comes in many forms (Lemmon, 1997; Fridman and Baan, 2001). The strong RFI that is problematic is often either local in frequency or in time. An example of RFI that is local in frequency is shown in Figure 1.5. Such RFI can for example be caused by television stations, aeroplanes and radar (at low frequency resolution), while for example broadband RFI caused by

phenomena such as lightning, high-voltage power cables and sparking electrical fences are often local in time. Sometimes, the frequency of RFI drifts with time. This can be caused by Doppler shifting of a satellite signal, by imperfect transmitters or because the transmitter is intrinsically changing its frequency, such as with certain radar signals when observed at high frequency resolution. A different class of RFI is caused by weakly transmitting but stationary — and therefore systematic — devices on site. This class of RFI is hard to recognize, as it might contaminate all the channels in a spectral band. In fringe stopping interferometers, the fringe rotation causes this type of RFI to have a sinusoidal response in the time-frequency domain (Thompson, 1982). It can be recognized and subtracted in various ways, as for example described recently by Athreya (2009).

1.4.1 RFI excision during off-line processing

Despite the numerous possible techniques, during the off-line processing phase — i.e., after the data has been recorded to disk — almost any observation requires additional processing steps because of RFI contamination. In the post-correlation phase, the use of an independent RFI reference signal to subtract the RFI (Briggs, Bell, and Kesteven, 2000), fringe fitting (Athreya, 2009) and post-correlation spatial filtering are possible. However, none of the above are applicable or sufficient in all cases or for all types of RFI. Therefore, the most used technique in the final processing step consists of detecting the RFI in time, frequency and antenna space, and ignoring the contaminated data in further data processing. This step is often referred to as "data flagging".

Historically, this step was performed by the astronomer. However, in modern observatories that operate at low frequencies, such as the Westerbork Synthesis Radio Telescope (WSRT), the Giant Metrewave Radio Telescope (GMRT), the Low Frequency Array (LOFAR), and the Expanded Very Large Array (EVLA), RFI mitigation is an essential component in the signal processing. In the case of LOFAR, there are high sensitivity requirements, especially for the Epoch of Reionisation project (Jelić et al., 2008; Thomas et al., 2009), with data sets up to a petabyte in size. RFI mitigation before correlation remains important (Boonstra et al., 2005), yet the amount of data will be too large for manual post-correlation flagging, implying the need for automated flagging strategies.

1.4.2 Detection

Probably the easiest and most used method to deal with RFI is by detecting its presence. After all, if a signal is clean of interference, doing nothing is the best an RFI excising method can do. Detection involves standard methods from signal processing theory, such as methods to detect changes in a signal feed, e.g. Basseville and Nikiforov (1993). However, the generic methods often need to be adapted to perform well in the context of radio astronomy. This is mainly because of the large data rates and/or their application in combination with data correcting methods that are described in the next subsections.

Detection almost always involves *thresholding* (Leshem et al., 2000; Niamsuwan et al., 2005; Weber et al., 1997). This means that a sample, channel, timestep or other region is marked as RFI if a specific quantity, such as the signal strength, exceeds some either pre-determined or variable limit. If this limit is adaptively determined, the technique is often referred to as *adaptive thresholding*. Thresholding will be a recurring concept in this work.

Detection can be performed at several stages and its results can be used in several ways. It might for example be applied at high time resolution, and if interference is detected, the affected samples are set to zero, e.g.,, as in Weber et al. (1997). This is often referred to as (time) blanking or nulling. When detection is applied later in the path, for example after correlation and/or time integration, the real-time requirements are relaxed and detection methods can be evaluated and selected after observing. If the data is integrated over time, the interference-to-noise ratio will normally be higher and allow more accurate detection. On the down-side, more data might be lost because of the lower resolution.

If RFI contaminates large parts of the data, simple detection might not be sufficient, and one needs to refer to data correcting techniques that try to subtract the RFI from the data. A few of those are described in the following subsections. Nevertheless, detection might be a valuable tool to determine which parts to subtract the RFI from, thereby making sure not to alter unaffected data.

In the case of LOFAR, detection needs to be both very fast and very accurate, and the possible effects of leaked interference need to be well understood for projects such as the LOFAR Epoch of Reionisation project. To this end, several existing detection methods will be analysed and new high-speed algorithms will be introduced and tested in Chapter 2.

1.4.3 Using reference antennae

Another way of dealing with RFI during observations, is by using antennae that are dedicated to observe the RFI source(s) with maximum sensitivity. The technique relates to many other adaptive cancellation techniques, such as the recent advances that allow noise cancelling headphones to cancel out the interfering environment. For this to work for radio observations, at least one extra antenna — one that is often much smaller than the main dish(es) — needs to be set up that provides a high gain towards the source of interference. Simultaneously, this antenna needs to have a low gain towards the signal of interest, so that the RFI signal can be subtracted from the observation without modifying the signal of interest. This kind of cancellation was first applied in contexts other than radio astronomy, such as radar and interference excision in communication systems (Ghose, 1996). Later, it was also successfully introduced in the context of radio astronomy (Barnbaum and Bradley, 1998).

The first adaptive cancellation results as presented by Barnbaum and Bradley were laboratory based, but later experiments were conducted with a reference antenna at the single dish Parkes Observatory (Briggs et al., 2000) and the Australia Telescope Compact Array (Mitchell et al., 2005). In the latter, it was found that the use of two reference antennae can provide even better results, with complete removal of the RFI source, although the system temperature is somewhat increased. At the Westerbork Synthesis Telescope Array, it was not possible to install reference antennae near the focal region due to space constraints. A scheme that used the array neighbours of a particular dish together with spatial filtering and adaptive nulling was implemented instead (Baan et al., 2004). Although the resulting WSRT RFI mitigation system (RFIMS) proved to be an effective way of reducing RFI, the system was not very popular amongst astronomers due to concerns about its impact on the signal. Another reason for its low popularity was the lack of strong incentives, as at that time the need to observe RFI contaminated bands was not as pressing (Baan et al., 2010).

Since its introduction the usage of reference antennae has evolved, and other successes include its usage to excise moving objects. Examples are the removal of the GLONASS satellites from

observations at the Parkes observatory (Mitchell and Robertson, 2005) and Green Bank Telescope (Poulsen et al., 2005).

1.4.4 Spatial filtering

Spatial filtering is a technique that makes use of multiple cross-correlated sensors to disentangle sources that are received with different power levels and with different delays. Signals from different geometrical locations are received with different power levels because of the different distance between the transmitter and the receiver, different propagation paths and/or different antenna response towards the source. Sensors that are at an increased distance from the transmitter receive the signal with a longer delay, causing the signal to be phase shifted. Together, the change in power and phase will cause each transmitter to have a specific *spatial signature* at each sensor.

The technique then forms a correlation matrix consisting of the cross-correlated sensor values; matrix element i, j consists of sensor value i times sensor value j. The diagonal elements contain the auto-correlated values. Normally, these matrices are constructed for small time periods of the order of milliseconds, but it is possible to integrate over longer time spans as long as the power and direction of the signals remain reasonably constant over the time interval. By decomposing the correlation matrix with an eigenvalue decomposition, several new correlation matrices are formed that represent the contribution from individual directions. The use of the eigenvalue decomposition to separate the different contributions assumes that the contributions produce orthogonal additions to the matrix. In practice, the addition of noise and the possibility that the correlation matrices of the RFI sources and the signal of interest are not completely orthogonal complicate things slightly, but these can be corrected for as well.

Spatial filtering had been a generic signal processing technique in the literature for some time (e.g., Widrow and Stearns (1985)). Around 2000, it was realized this method might be useful in the context of radio-astronomical interference mitigation (e.g., Fisher (2001)). The first radio astronomical simulations with spatial filters were performed by Leshem and van der Veen (2000a,b). The first observatory which was used to test the method (and was found in the literature) was the Westerbork Synthesis Radio Telescope (Raza et al., 2002). Later, it was combined with several other techniques in the WSRT RFIMS system (Baan et al., 2004), including adaptive cancellation as discussed in §1.4.3. To get good results, the deconvolution method that is used during imaging needs to be adapted for this. An extensive analysis of spatial filtering and the required imaging steps was performed by Boonstra (2005).

So far, it was assumed that the dishes observe the same target of interest, and RFI signals produce a different spatial signature at each telescope. The situation is slightly different for multi-beam systems, as in multi-beam observations each beam aims at a different target. Spatial filtering was shown to be very useful for the multi-feed system of the single dish Parkes observatory (Kocz et al., 2010), where it was very recently also successfully used to distinguish RFI from temporal signals with celestial origin (Kocz et al., 2012).

As a side note, one can argue that the use of reference antennae is a special case of spatial filtering, as the additional spatial information provided by the extra antenna is used to remove the interference. In my opinion, this is just a matter of definition. If the signal of a — often dedicated — reference antennae is normalized (adapted) and subtracted from the feed, it is common to refer to the technique as being an adaptive noise cancelling (ANC) filter, as it was initially introduced. The term spatial filtering on the other hand is most commonly referring to techniques that involve subspace projection and do not have dedicated reference antennae to increase the spatial

information. Therefore, in this work we will use this terminology as well.

1.5 Scope and aim of the thesis

In this project we will try to answer a few questions that are key to observing with LOFAR in its populated environment:

- *What are the observational consequences of building LOFAR in a populated area?*
  A populated environment is an unusual choice for building a radio telescope. Although care has been taken to make sure that LOFAR will perform well, an instrument like LOFAR is extremely complex. The harmful effects of interfering noise from urban areas are hard to accurately predict beforehand, especially when different RFI mitigation strategies are to be combined. Therefore, an important goal of this work is to extensively analyse and evaluate the effects of interference on LOFAR. Quantizing the interference occupancy over frequency and hour of the day will provide information that is very relevant for the key science projects of LOFAR and their observing strategy, as well as for dynamic scheduling purposes.

- *What existing methods can one use to excise radio-frequency interference in LOFAR observations?*
  In §1.4, we have summarized a few existing methods that are currently being used for RFI excision. Many of these methods were originally developed at GHz frequencies and higher. Because LOFAR is a low-frequency instrument, some of these methods might work well, while others might not. Moreover, some of these methods require special hardware in the field, e.g. reference antennae, or they might require on-line computing power. It is to be seen whether LOFAR can provide this.

- *Can the accuracy and performance of currently available interference excision methods be improved?*
  LOFAR is going to be among the first interferometers that will cross-correlate over 50 stations. Each station will provide 0.76 kHz of spectral resolution over 48 MHz of bandwidth with two linear polarizations, and 96 and 192 MHz modes are planned. This yields enormous data rates, and to cope with these rates the RFI algorithms need to be extremely fast. On the other hand, the high spectral and temporal resolution might improve RFI mitigation techniques due to the higher amount of information that is available to distinguish RFI from the signal of interest. In this project, we will investigate the performance requirements for RFI algorithms, and try to improve existing techniques to work as accurately as possible for the LOFAR case.

- *Will RFI cause a limit on the sensitivity with which LOFAR can observe?*
  Observations from an interferometer are fundamentally noise limited because of noise from the sky and receiver electronics. Nevertheless, it is expected that the noise in a LOFAR observation is inversely proportional to the square root of the integrated time, and therefore it is theoretically possible to reach any noise level, as long as the duration of integration is long enough. This is an assumption in LOFAR's Epoch of Reionisation project, in which 100 nights of 6 hour observations will be integrated to achieve enough signal-to-noise to statistically detect the extremely feeble signals from this era of our Universe.

However, radio interference from stationary sources might break this assumption. Unlike Gaussian noise, such sources could add up coherently. This could mean that, even when all detected RFI is successfully excised in observations of a few nights, RFI might exist under the noise that only shows up after longer integration. In this project we will analyse this possibility and investigate possible measures to prevent low-level RFI from causing a sensitivity limit or false detection.

While these questions are asked specifically for the LOFAR case, their answers will be very useful for both existing and future radio-astronomical instruments. To this end, care will be taken to provide analysis and algorithms that are as generic as possible, with the ultimate goal of developing RFI techniques that are useful for any radio observatory. Moreover, the astronomers that use the instruments and try to reduce astronomical data will also benefit from improved post-processing algorithms and understanding of interference.

A major future endeavour will be the Square Kilometre Array (SKA). Its planned observing bandwidth will partly overlap with LOFAR's frequency range. Although LOFAR's collecting area is currently unprecedented, the SKA will be tens of times more sensitive compared to LOFAR, and interference is therefore a large concern. It will therefore be located in an area with minimal RFI in either South Africa or Australia (decision to be made on 4 April 2012). The knowledge that will be gained from LOFAR's RFI environment and RFI strategy will be highly relevant for SKA's design, implementation and operation.

Summa summarum, this thesis addresses the aspects of interference in radio astronomy, in special for the LOFAR telescope and the LOFAR Epoch of Reionisation project.

1.6 Thesis layout

We will now briefly describe the layout of this thesis. In Chapter 2, we will start analysing existing and designing new methods that can be used for interference detection. The steps necessary for accurate detection can be classified in (i) estimating the astronomical signal; (ii) adaptive (combinatorial) thresholding; and (iii) applying morphological detection. For each of these steps, methods will be compared and the best method for LOFAR will be picked. Next, in Chapter 3, these methods will be combined to form a fully automated iterative pipeline, that is currently the recommended way to remove RFI from LOFAR data. Chapter 4 will present filters that can remove interference from terrestrial sources and off-axis celestial sources. The fundamental concept of fringe speed is discussed, and using this theory novel filters are constructed. The methods that have been presented so far will then be used in the next chapter, Chapter 5, to analyse the radio environment of LOFAR. Using RFI surveys, we present spectral occupancy statistics, determine the difference between observing at day and night and analyse the effectiveness of the methods and see if any leaked interference is visible. Then, in Chapter 6 the spatial distribution of interfering sources is analysed using statistical derivations of the RFI surveys. From this, we try to foresee what the interference effects might be for detection of the Epoch of Reionisation. Finally, in Chapter 7, we will evaluate the current interference situation and methods to deal with it, and look forward to possible related future developments.

Chapter 2

Detection of Radio-Frequency Interference

Based on:
"*Post-correlation radio frequency interference classification methods*"
(Offringa et al., 2010, MNRAS, 405, 155–167)
"*A morphological algorithm for improving radio-frequency interference detection*"
(Offringa et al., 2012, A&A, 539, A95)

THE SITUATION for RFI flagging strategies in modern observatories such as the Westerbork Synthesis Radio Telescope (WSRT), LOFAR and the Giant Metrewave Radio Telescope (GMRT) has changed. On one hand, time and frequency resolutions have improved considerably over the last decade. Because of this, the detection of RFI can also be performed at higher resolutions, and the accuracy of flagging of contaminated samples improves, resulting in smaller loss of data. On the other hand, radio quiet zones are harder to achieve, and all of the above mentioned telescopes are situated in populated areas. Moreover, sensitivity requirements for telescopes are growing. For example, one of the LOFAR key science projects is the LOFAR Epoch of Reionization (EoR) project (Jelić et al., 2008; Thomas et al., 2009), a very ambitious project with high demands on sensitivity and noise behaviour. These new constraints require new techniques with different requirements for the excision of RFI.

In this chapter, we will introduce several fundamental automated detection methods. These detection methods can be compared on accuracy, i.e., the true/false-positive ratio; the speed of the algorithm; robustness; and technical requirements that they impose. Constructing a detection mechanism that performs good on all aspects is challenging. During this chapter, existing methods will be described and new methods will be introduced that are designed to take this challenge. Some of these have been implemented in the LOFAR observatory pipeline, that will be described in Chapter 3. This pipeline consists of scripted iterations in which the methods from this chapter are taken as building blocks, and are combined in a way to optimize performance and accuracy.

We will evaluate the effectiveness of several automatic RFI mitigation methods. The methods

will be compared with each other in order to be able to pick a general optimal RFI strategy for a specific detection step. We will do this by testing the methods on both artificial data and data from WSRT, most of which has been observed in the frequency range of LOFAR. Testing the methods on WSRT data will also provide an indication of the effects of the RFI environment on future LOFAR observations.

While most of the methods are tested post-correlation (off-line), the detection schemes are not limited to application after correlation. Some of the methods are currently being tested before correlation (on-line) at the LOFAR observatory.

The upcoming section will explain the difference between the pre-correlation and post-correlation application of detection methods. In the sections that follow, we describe several new methods for the detection of RFI. These methods are categorized in signal estimation methods, thresholding methods and morphological methods. The signal estimation methods and thresholding methods are depending on each other, and will therefore be described together in section 2.2. We present our results on the signal estimation and thresholding methods, including the comparative study, in section 2.3. The application of morphological operators will be discussed in section 2.4.

2.1 Detection stage

RFI mitigation can be applied at two different stages: a pre-correlation stage and a post-correlation stage. The pre-correlation mitigation stage is very powerful as the observational data is still available at its highest time resolution. For example, there are methods that blank or subtract short periodic radar RFI bursts on-line (Niamsuwan et al., 2005), leaving the astronomical signal intact with only a very slightly increased signal to noise ratio. Any residual RFI has to be removed during the data reduction or imaging stage, often manually, for example by finding appropriate clipping levels for contaminated baselines until the reduced data is free of artefacts.

Pre-correlation methods have to handle large amounts of data in a very short time and, because of hardware constraints, they can often only access limited dimensions of the data, such as the data from a single dish or station, or the data from a small time range. Examples of pre-correlation methods are based on thresholding (Weber et al., 1997; Leshem et al., 2000; Baan et al., 2004; Niamsuwan et al., 2005); spatial filtering with eigenvalue decomposition (Leshem et al., 2000; Smolders and Hampson, 2002; Ellingson and Hampson, 2002); and adaptive cancellation with a reference antenna (Barnbaum and Bradley, 1998).

To deal with RFI, the post-correlation phase is the final resort. Demonstrated techniques include the use of an independent RFI reference signal to subtract RFI (Briggs et al., 2000); an approach using singular value decomposition (Offringa et al., 2010a; Pen et al., 2009); and fringe fitting (Athreya, 2009). Since RFI comes in many forms (Fridman and Baan, 2001; Lemmon, 1997) not all contaminated samples can be recovered, despite the numerous existing techniques. Therefore, flagging remains an important final step (Offringa et al., 2010a; Winkel et al., 2006, 2009).

Pre-correlation and post-correlation techniques are mostly complimentary: they find or remove different kinds of RFI. Hence, the implementation of one does not make the other obsolete. However, a huge advantage of off-line detection, is that it allows one to easily experiment with different settings of the detection method parameters.

2.2 Thresholding & signal estimation methods

Radio astronomers have developed their own ways of dealing with RFI during data reduction using numerous astronomical software packages. In many cases, this implies flagging by hand – a tedious and time consuming job. Many toolkits, such as AIPS[1], AIPS++[2], MIRIAD[3] and NEWSTAR[4], provide specific features to perform flagging, such as the `FLAGR` task in AIPS++. Astronomers have automated the process further by designing scripts in which common signal processing techniques such as thresholding, smoothing, line detection and curve fitting are combined (Winkel et al., 2006; Bhat et al., 2005). Another common signal processing technique known as Singular Value Decomposition has recently been used for the automatic removal of broadband RFI in GMRT observations (Pen et al., 2009). In this section we will describe some of the techniques available that relate to a new method of interference mitigation that we will introduce, and finally we will explain the new method itself.

2.2.1 Post-correlation thresholding

Since RFI increases the measured absolute amplitude of a signal, thresholding is an effective method that is often used to remove strong RFI. The threshold level is often globally determined, or sometimes set relative to the variance or mode distribution parameters per baseline. These can be stably estimated using, for example, the Winsorized variance or mode (Fridman, 2008). All values that are outside a certain range around the mean or median are flagged as bad data and not used in subsequent data reduction. Sometimes a number of samples around a bad data sample are flagged as well. Most astronomical reduction toolkits provide options to threshold part of a data cube, allowing different thresholds at the cost of an increased effort for the astronomer. An important consequence of thresholding is that good data is selected with a bias. When many non-contaminated samples are above the threshold, they will be flagged and not used in subsequent data reduction. As a result, artefacts such as incorrect flux densities might be caused in the image plane. It is therefore important to have a low false-probability rate of RFI detection.

2.2.2 Surface fitting and smoothing

A surface fit to the correlated visibilities $V(\nu, t)$ as a function of frequency ν and time t can produce a surface $\hat{V}(\nu, t)$ that represents the astronomical information in the signal. Requiring $\hat{V}(\nu, t)$ to be a smooth surface is a good assumption for most astronomical continuum sources, as their observed amplitudes tend not to change rapidly with time and frequency, whereas specific types of RFI can create sharp edges in the time-frequency domain. Because of the smoothing in both time and frequency direction, this method is not directly usable when observing strong line sources or strong pulsars. The residuals between the fit and the data contain the system noise $N_{\text{noise}}(\nu, t)$ and the RFI, $N_{\text{RFi}}(\nu, t)$, which can then be thresholded without the chance of flagging astronomical sources that have visibilities with high amplitude.

[1] AIPS: Astronomical Image Processing System, `http://aips.nrao.edu/`.

[2] AIPS++, `http://aips2.nrao.edu/`.

[3] MIRIAD, a data reduction package tailored for the Australia Telescope Compact Array (ATCA), `http://www.atnf.csiro.au/computing/software/miriad/`.

[4] NEWSTAR, a data reduction package tailored for the Westerbork Synthesis Radio Telescope (Noordam, 1994).

Several suitable surface fitting methods exist. As an example, in Winkel et al. (2006) a pipeline is described in which a two-dimensional, low order, dimensional independent polynomial is iteratively fitted to time-frequency tiles in the data using a least square fit:

$$\hat{V}_k(\nu,t) = \sum_{i=1}^{N_\nu} a_{k,i}\nu^i + \sum_{i=1}^{N_t} b_{k,i}t^i + c_k, \tag{2.1}$$

where $\hat{V}_k$ is the fitted surface that represents the astronomical information in the k-th tile, N_ν, N_t are the polynomial order for the frequency and the time, respectively, and $a_{k,i}, b_{k,i}, c_k$ are the coefficients of the fit for tile k.

The fit is performed iteratively, and values which have been flagged in previous iterations are excluded from the fit. This can be done by introducing a weight function $W_F(\nu,t)$, where $W_F(\nu,t) = 0$ indicates that the value is flagged or outside the boundaries of the measured time or frequency range, and $W_F(\nu,t) = 1$ means the value is accepted. The fit is performed by minimizing an error function E_k for each tile:

$$E_k = \sum_\nu \sum_t W_F(\nu,t) f(\hat{V}_k(\nu,t), V(\nu,t)) \tag{2.2}$$

where $f(a,b) = (a-b)^2$ for a least squares fit or $f(a,b) = |a-b|$ for a fit with a minimal absolute error.

An example of this approach after a few iterations can be seen in Figure 2.1. In simple cases, the surfaces that are created with this approach represent the astronomical information reasonably well, and the method is also quite fast. However, as polynomial fits tend to show deviations near boundaries, the method is inaccurate near the boundaries of each tile.

Compared with tile-based approaches, sliding window methods tend to be more accurate. A simple example of a sliding window approach is to calculate the average of a window of size $N \times M$ around each data value:

$$\hat{V}(\nu,t) = \frac{1}{\text{count}} \sum_{i=-\frac{1}{2}N}^{\frac{1}{2}N} \sum_{j=-\frac{1}{2}M}^{\frac{1}{2}M} W_F \cdot V(\nu + i\Delta\nu, t + j\Delta t), \tag{2.3}$$

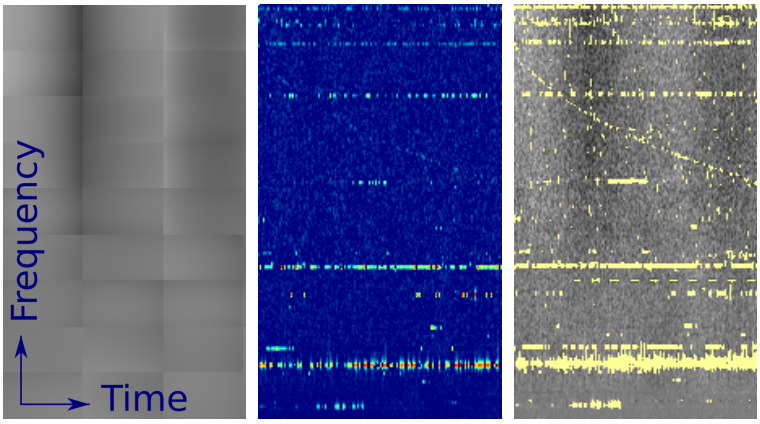
with

$$\text{count} = \sum_{i=-\frac{1}{2}N}^{\frac{1}{2}N} \sum_{j=-\frac{1}{2}M}^{\frac{1}{2}M} W_F(\nu + i\Delta\nu, t + j\Delta t) \tag{2.4}$$

This method is still fast and creates a surface without tile edges. However, the sliding window average represents the astronomical signal less well. For example, peaks in the original function cause square-shaped edges in the fit, which in the end cause detection inaccuracies.

One way to overcome this problem is to calculate the local median instead of the local average. Values that have been flagged in a previous iteration should be ignored by the median calculation. The median is insensitive to peaks and the surface created by the local median remains smooth when the window is slid over the data. The median however is not always a good estimate of the sliding window centre sample specifically, as all samples have equal weight.

Another way to overcome the problem is to calculate a weighted average. Consider a weight function $W_d(i,j)$ that depends on the two components i, j that represent the distances from the



(a) $\hat{V}(\nu, t)$ **(b)** $\hat{V}(\nu, t) - V(\nu, t)$ **(c)** Thresholded

Figure 2.1: *Tile-based polynomial fitting applied to the raw visibilities from an observation of 3C196 at 140 MHz using a 144m WSRT baseline (see §2.3.3). Panel (a) shows the tiled fit of the astronomical signal. Panel (b) shows the difference between the fitted astronomical signal and the observed signal used for thresholding. Panel (c) shows the flags on top of the original signal. The flags established by single pixel thresholding cover the RFI when verified by eye, although many false-positives can be seen which are caused by ("normal") noise. The tile size used for this image is 30 frequency channels with 10 kHz width × 50 time scans with 10s integration time.*

centre of the window in time and frequency respectively. Then

$$\hat{V}(\nu, t) = \frac{\sum_{i=-\frac{1}{2}N}^{\frac{1}{2}N} \sum_{j=-\frac{1}{2}M}^{\frac{1}{2}M} W_d(i,j)\,(W_F \odot V)\,(\nu_i, t_j)}{\text{weight}} \tag{2.5}$$

where

$$\text{weight} = \sum_{i=-\frac{1}{2}N}^{\frac{1}{2}N} \sum_{j=-\frac{1}{2}M}^{\frac{1}{2}M} W_d(i,j) W_F(\nu + i\Delta\nu, t + j\Delta t) \tag{2.6}$$

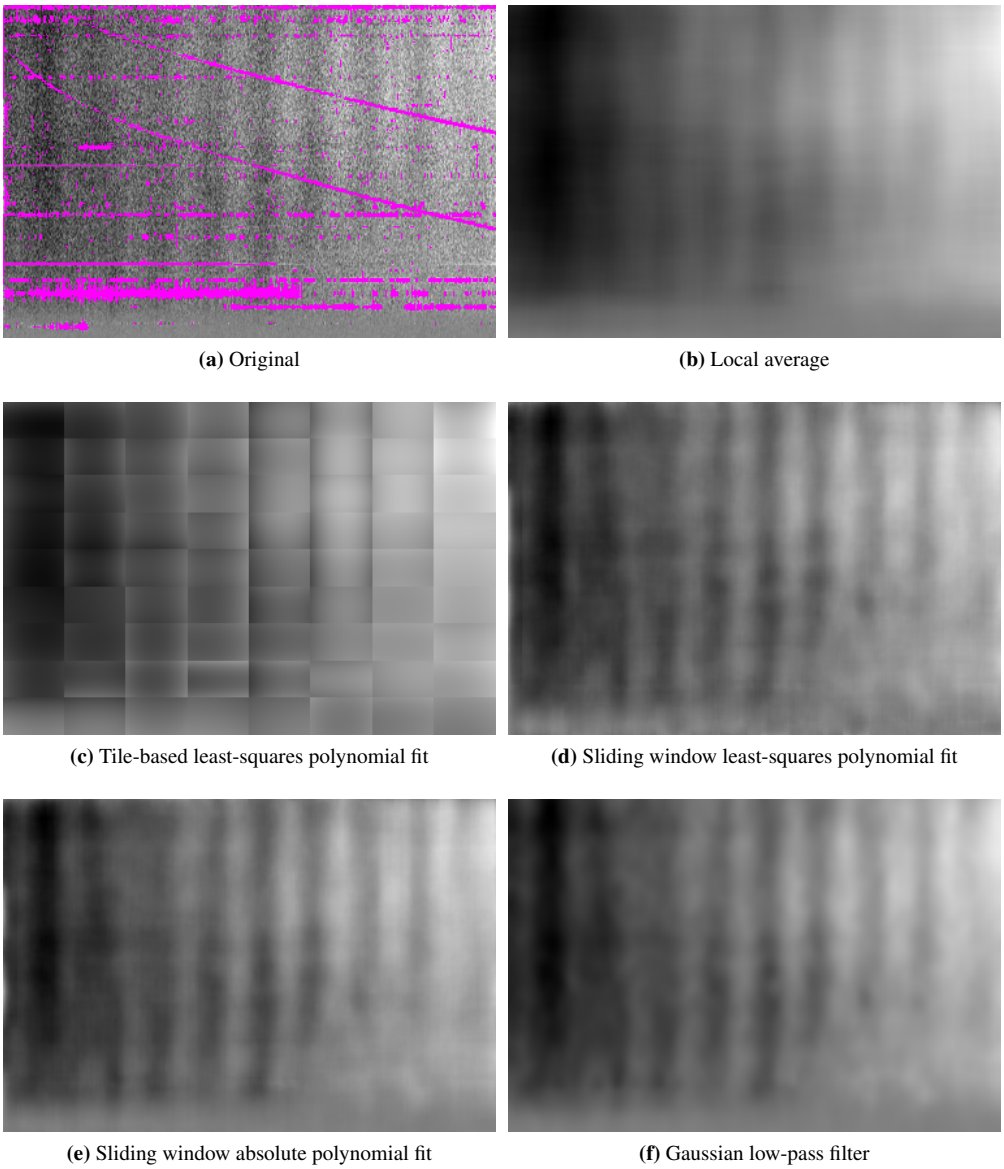
This can be calculated very fast, since (2.5) is the convolution operation $W_d * (W_F \odot V)$ and (2.6) is another convolution $W_d * W_F$, giving:

$$\hat{V} = ((W_F \odot V) * W_d) \oslash (W_F * W_d) \tag{2.7}$$

where $\odot$ and $\oslash$ are the elementwise multiplication and division operators. A good choice for W_d is the two-dimensional (dimensional independent) Gaussian function:

$$W_d(i,j) = \exp\left(-\frac{i^2}{2\sigma_\nu^2} - \frac{j^2}{2\sigma_t^2}\right) \tag{2.8}$$

Together, equations (2.7) and (2.8) essentially describe a weighted Gaussian smoothing operation, or more specifically, a Gaussian smoothing operation with missing data. The parameters σ_ν and



(a) Original

(b) Local average

(c) Tile-based least-squares polynomial fit

(d) Sliding window least-squares polynomial fit

(e) Sliding window absolute polynomial fit

(f) Gaussian low-pass filter

Figure 2.2: *Overview of various fitting methods*

σ_t can be used to specify the level of smoothing in frequency and time respectively. Since the weight function is dimensionally separable, the convolutions can be dimensionally separated:

$$\hat{V} = \frac{(W_F \odot V) * U_\nu * U_t}{W_F * U_\nu * U_t} \tag{2.9}$$

with $U_\nu(i) = W_d(i, 0)$ and $U_t(j) = W_d(0, j)$. Each of the convolutions in (2.9) is a one dimensional convolution, and this is therefore a fast operation.

An overview of various fitting methods is given in Figure 2.2.

2.2.3 The cumulative sum method

The cumulative sum (`CUSUM`) method is a well known analysis method used to detect changes in distribution parameters (Page, 1954; Basseville and Nikiforov, 1993), such as in quality control in production environments. If the cumulative sum of sequential samples exceeds an adaptive threshold, the system enters an alarmed state and changes can be made to correct the quality. In its common form, the likelihood for two distribution parameters is used to compute the threshold.

To turn this method into an RFI mitigation strategy, the total observed power or power received at a certain frequency by a single dish can be used as the sequential input values to the `CUSUM` method. The likelihoods of either variance or mean of RFI can be estimated using the variance of the signal (Friedman, 1996; Baan et al., 2004). Observing can be stopped as soon as RFI is detected, and can continue when reception has returned to normal. This method can be easily implemented for on-line RFI detection, as it is simple and fast. However, the `CUSUM` method does not estimate the start time of the change, it only detects the change quickly, which nevertheless may cost time and thus some bad data may leak through before the method detects faint RFI. Hence, the method is more applicable to a first check on the data than to actually perform flagging. The subsequent sections will describe a method that combines the detection strength of the `CUSUM` method with the possibility of performing flagging off-line.

2.2.4 Combinatorial thresholding

RFI bursts often affect multiple samples which are connected either in frequency or time. We will now introduce a new threshold mechanism that makes use of this knowledge: we will flag a combination of samples when a property of this combination exceeds some limit. Assume that A and B are neighbouring samples. In normal thresholding, we will look at each of the samples A and B individually and flag one of them if it exceeds some "single sample" threshold χ_1. For combinatorial thresholding, a new flagging criterion is added: if A and B do not exceed the single sample threshold χ_1 individually, they can still be flagged when A and B *both* exceed a somewhat lower threshold χ_2. If not, they can be combined with a third neighbour, C, and thresholded at χ_3, etc. The more connected samples are combined, the lower the sample threshold.

2.2.5 The `VarThreshold` method

Given a set of strictly decreasing thresholds, $\{\chi_i\}_{i=1}^N$, a value will be classified as RFI if it belongs to a combination of i values in either the time or frequency direction in which all absolute values are above the threshold χ_i. To determine whether a single sample $R(\nu, t)$ should be flagged

because of an RFI sequence in the frequency direction, the following rule is applied:

$$\begin{aligned}\text{flag}\nu_M(\nu,t) = \exists i \in \{0\ldots M-1\} : \forall j \in \{0\ldots M-1\} : \\ |R\left(\nu + (i-j)\,\Delta\nu, t\right)| > \chi_M\end{aligned} \tag{2.10}$$

where M is the number of samples in a combination. The flagging rules for the time direction are correspondingly determined. Finally, a sample is flagged if any of the two rules is satisfied. We will call this method the `VarThreshold` method.

We will show a simple example to demonstrate the method. Consider the following values:

$$R = \begin{pmatrix} 1 & 2 & 1 & 4 \\ 4 & 1 & 1 & 4 \\ 2 & 2 & 1 & 4 \end{pmatrix} \tag{2.11}$$

Each row represents a frequency channel and each column represents a time scan. Assume the high values in the fourth column were caused by broadband RFI. When using a normal threshold $\chi = 3$, all samples with value 4 would be thresholded, including one false-positive. However, if we used combinatorial thresholding, with $\chi_1 = 5$ and $\chi_2 = 3$, we would threshold only the three broadband RFI samples.

The above text suggests an implementation of this method by a procedure which iterates over all samples and, for each sample, checks if it and its $M \in \mathcal{M}$ neighbours form an RFI sequence in one of the directions. Alternatively, an implementation can start by marking all samples above a certain χ_M as candidates. Subsequently, only the marked candidates that form a connected segment with more than M connected samples in an orthogonal line in one of the directions are flagged. This procedure is repeated for all $M \in \mathcal{M}$. From this perspective, it is easy to add other morphological constraints. Instead of looking for straight lines in the time and frequency direction, an enhanced version might flag connected shapes covering a specific area, or shapes that form a line or curve in the plane, possibly not connected, that are likely to be caused by RFI.

`VarThreshold` parameters

The following list of parameters need to be optimised to make efficient use of this approach:

- The false-positive rate on uncontaminated samples. The lower the value, the more RFI remains. The higher the value, the more uncontaminated samples will be flagged. We will discuss this in §2.2.5.
- A set that defines which samples are combined. For this we define $\mathcal{M}$, a set containing the number of samples that will be combined in each of the four directions. Ideally, each sample will be combined with all samples of either the same frequency or the same time, i.e., $\mathcal{M} = \{i \in \mathbb{Z} : 1 \leq i \leq \max(N_\nu, N_t)\}$, with $\mathbb{Z}$ the set of integers. Empirically, a small subset $\mathcal{M} = \{1, 2, 4, 8, 16, 32, 64\}$ works almost as well and saves summing and comparing many samples.
- The set of thresholds $\{\chi_M : M \in \mathcal{M}\}$ for the different number of combinations M. The total set of thresholds is expressed by two parameters, χ_1 (the threshold on a single sample) and ρ, using the following formula:

$$\chi_i = \frac{\chi_1}{\rho^{\log_2 i}} \tag{2.12}$$

A value of $\rho = 1.5$ empirically seems to work well for the `VarThreshold` and the below defined `SumThreshold` method. To find χ_1 for a desired false probability rate, ρ is kept constant and the χ_1 value is binary searched by performing mitigation on data selected from the distribution of the noise, with the values $\{\chi_i\}_{i\in\mathcal{M}}$ computed as in (2.12), until the false probability rate is close to the desired rate.

Since the method is combined with a surface fitting strategy, the following parameters are added:

- The number of iterations to be performed. The resulting accuracies are good with about 5 iterations.
- The iteration sensitivity as a function of the iteration number, $\eta(i)$. In each iteration, the threshold sensitivity is increased (more samples are flagged). To accomplish this, all the thresholds $\{\chi_i\}_{i\in\mathcal{M}}$ are decreased by dividing them by a factor of $\eta(i)$. Only during the last iteration will a sensitivity of 100% be used. By slowly increasing the sensitivity a first bad fit to the background won't have much effect, since only the very strongly RFI contaminated samples are removed. Using an exponential function for $\eta(i)$ was found to work well.

The `VarThreshold` false-positive ratio

Assume that $R \sim \mathcal{D}(\sigma_{N_s})$, where R is the residual of the complex correlated visibilities V and the surface fit $\hat{V}$, and $\mathcal{D}$ is a distribution with parameter σ. The probability that a non-RFI contaminated sample from the residual is larger than χ can be determined with:

$$\forall\nu\forall t : P\left(|R(\nu,t)| \geq \chi\right) = \int_{-\infty}^{-\chi} \varphi_\sigma(x)dx + \int_{\chi}^{\infty} \varphi_\sigma(x)dx, \tag{2.13}$$

where $\varphi(x)$ is the probability density function of the distribution $\mathcal{D}(\sigma_{\mathcal{N}_f})$. Note that the term $\int_{-\infty}^{-\chi} \varphi_\sigma(x)dx$ is only relevant when the distribution contains negative values – unlike the Rayleigh distribution – and the values are thresholded above χ as well as below $-\chi$.

The combined threshold false-positive rates can best be calculated numerically, since an analytical calculation is rather complex, even for $\mathcal{M}$ with a single combined threshold χ_M. This analytical calculation will be demonstrated for $M = 2$. First it is assumed that any two samples, $R(\nu_1, t_1)$ and $R(\nu_2, t_2)$, are independent when they are not RFI contaminated. This is the case if the fit represents the astronomical data and system noise is uncorrelated. With this assumption, the probability P_{false} for a single non-contaminated sample R_1 with $M = 2$ to be flagged in one of the four combinations with its neighbours $R_{2\ldots5}$ can be calculated with:

$$\begin{aligned} P_{\text{false}}(\nu,t) & \\ &= P\big(\text{flag}\nu_{M=2}(\nu,t) \ \vee \ \text{flag}t_{M=2}(\nu,t)\big) \\ &= P(|R_1| > \chi \wedge \exists i \in [2\ldots5] : |R_i| > \chi) \\ &= P(|R| > \chi) - P(|R| > \chi)\left(1 - P(|R| > \chi)\right)^4. \end{aligned} \tag{2.14}$$

The corresponding formulae for larger M are more complex. When $\mathcal{M}$ contains more than one element, the false-positive ratios for the elements M_i can not be simply added to obtain the combined false-positive ratio, as $P(\text{flag}\nu_{M_i}(\nu,t))$ and $P(\text{flag}\nu_{M_j}(\nu,t))$ are not statistically

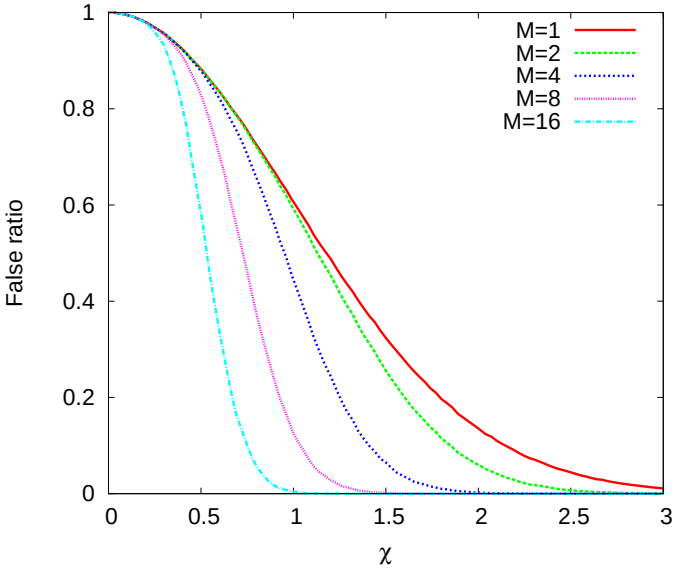


Figure 2.3: *The false-positives of the* `VarThreshold` *method when flagging with a single combination $\mathcal{M} = \{M\}$ without surface fitting. Samples were selected from a Rayleigh distribution, which is the distribution of the visibility amplitudes. χ is relative to the mode of the distribution.*

independent: both will at least make use of sample $R(\nu, t)$. Given this, the analytical expression becomes rather complex and the probability is evaluated numerically.

Figure 2.3 shows the result of calculating the total false-positive ratio numerically, for several values of M.

2.2.6 The `SumThreshold` method

Now we will present a variation on the `VarThreshold` method that improves the detection performance. This method, named the `SumThreshold` method, is a flagging method that combines samples as in the `VarThreshold` method. In this alternative case, the sum of a combination of one or more other samples is used as a threshold criterion. As in the `VarThreshold` method, the threshold χ_M is variable and depends on M, the number of samples that are summed.

Unlike the `VarThreshold` method, this approach allows the flagging of a sequence of samples when it contains samples with values below the sequence threshold value. However, without an additional rule, there are situations in which this method might flag too many samples. For example, consider the sequence $[0, 0, 5, 6, 0, 0]$ that represents a strong RFI contamination in two samples. When the `SumThreshold` method without a second rule is applied with average threshold values $\chi_1 = 7$, $\chi_2 = 5$, $\chi_3 = 4$, …, $\chi_6 = 1.8$, all six values would be thresholded, as their average exceeds $6\chi_6$. The following rule is therefore added: the values are thresholded in the increasing order χ_1, χ_2, …, χ_M. When a lower threshold has already classified samples as RFI contaminated, the samples will be left out of the sum and replaced by the average threshold

level. In the example case, the values 5 and 6 will be classified as RFI by the second threshold, and therefore will be replaced by χ_6 when combining all the six samples. The average of the sequence for the sixth threshold is therefore calculated as $(0+0+\chi_6+\chi_6+0+0)/6 = \frac{2}{6}\chi_6$. As a consequence, only the samples with values 5 and 6 are flagged.

Implementation details of the `SumThreshold` method are given in Appendix A, which includes a vectorized algorithm that uses the SSE instruction set.

The `SumThreshold` false-positive ratio

We calculate the theoretical false-positive ratio for $M = 2$ as for the `VarThreshold` method. The probability $P(T_{\chi,1,2})$ that the sum of two independent random samples exceeds a certain value χ is given by:

$$\begin{aligned} \forall \nu_1 \nu_2 t_1 t_2 : P(T_{\chi,1,2}) &= P\left[R(\nu_1,t_1) + R(\nu_2,t_2) \geq \chi\right] \\ &= P\left(\mathcal{D}(2\sigma_{N_s}) \geq \chi\right) \\ &= \int_{\chi}^{\infty} \varphi_{2\sigma}(x)dx \end{aligned} \tag{2.15}$$

When thresholding the average of a combination of two samples, each sample will occur four times in a hypothesis test, once with each of its neighbours. On uncontaminated samples, the probability of a false-positive for each of these tests is given by (2.15). The probability for a false-positive with the four tests applied on each sample becomes:

$$P(T_{\chi,1\times4}) = P(T_{\chi,1,2} \vee T_{\chi,1,3} \vee T_{\chi,1,4} \vee T_{\chi,1,5})$$

Because the tests $\{T_{\chi,1,i}\}_{i=2}^{5}$ are dependent on each other, it is much easier to calculate the false-positive rates numerically. This can be performed by applying the `SumThreshold` on a large amount of data selected from the distribution $\mathcal{D}$. The result of such a simulation is in Figure 2.4.

2.2.7 The Singular Value Decomposition

Singular value decomposition (SVD) is a mathematical tool for finding the singular values of a matrix, which can exhibit certain properties of the matrix.

A singular value decomposition consists of finding the complex unitary $M \times M$ and $N \times N$ dimensional matrices U and V containing respectively a left and right singular vector in each row, and the diagonal, $M \times N$ dimensional real matrix Σ containing the singular values, such that:

$$A = U\Sigma V^T \tag{2.16}$$

RFI is mitigated from the data set by performing this decomposition on a matrix A. Each element A_{ij} represents the measured flux, where i is a baseline-frequency index and j a time index. Each given matrix A has a unique solution for the singular values Σ, if the singular values are sorted, but there is no unique solution for U and V (for example, A remains equal when all values in U and V are negated). It is assumed that the highest singular values represent the singular values of the RFI data. To mitigate the RFI, the highest singular values in Σ are set to zero and the new matrix $\hat{A}$ is recomposed from U, Σ and V.

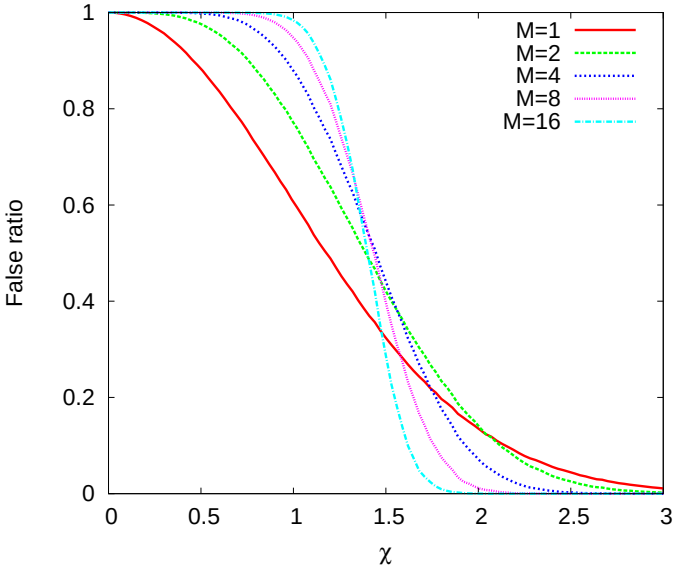


Figure 2.4: *The probability of a false-positive when thresholding with a single combination $\mathcal{M} = \{M\}$ using the* `SumThreshold` *method without surface fitting. The Rayleigh distribution was used for the simulation. χ is the average threshold relative to the distribution mode. Thus a combination of samples was thresholded when their sum exceeds $\chi \times M \times \sigma$. The false ratio for $M \geq 2$ is different from the* `VarThreshold` *method (Figure 2.3). Because of this difference, the parameter ρ used to calculate the set of thresholds as in* (2.12) *needs to be optimised for the methods individually. Although the false ratio is not smaller than the* `VarThreshold` *false positive method, the true ratio is often increased (Figure 2.9).*

The number of singular values to be removed or set to zero has to be chosen in such a way that only the RFI is removed. The singular values that correspond to RFI are often strong outliers, whereas the singular values of Gaussian noise form a smooth curve. The position of the abrupt change in the curve of the singular values is used as the number of singular values to be removed, as is shown in Figure 2.5.

Properties

Let $L = \min(M, N)$, then:

$$A_{ij} = \sum_{k=1}^{L} U_{ik} \Sigma_{kk} V_{jk}. \tag{2.17}$$

U and V are unitary, $U\overline{U} = I$ with $\overline{U}$ the Hermitian transpose, and the rows and columns of the matrices form by definition a complex orthonormal basis. This implies:

$$\forall i \in [1..M] : \sum_{j=1}^{L} U_{ij}^2 = 1. \tag{2.18}$$

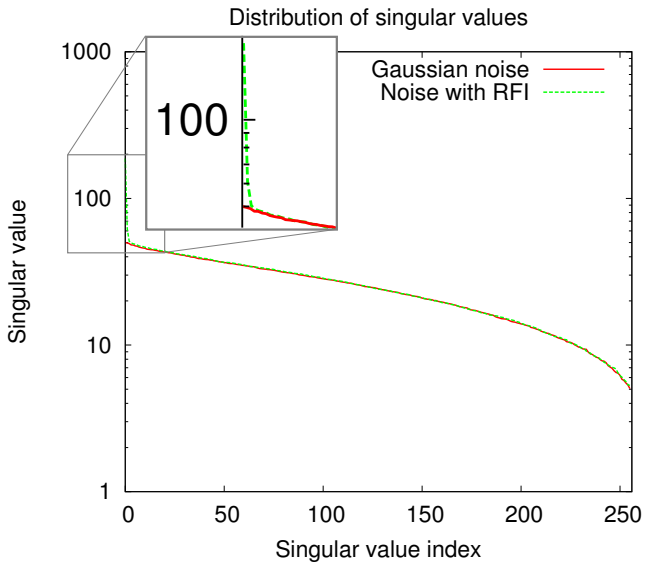


Figure 2.5: *The distribution of the singular values of two artificial measurements: one containing Gaussian noise only, the other containing Gaussian noise poluted by broadband RFI. In this example, the first five singular values are affected by the broadband RFI. In general, the number of singular values that are affected by RFI and the possibility to recognize them varies depending on the orthogonality properties of the RFI.*

Hence there is at least one non-zero value in each row and column of the matrices U and V, and setting a non-zero singular value to zero changes A. If A contains real values only, equation (2.18) implies that all values in U and V are between -1 and 1, and removing a singular value Σ_{ii} can alter each value in A by at most Σ_{ii}. In the complex case, removing a singular value can alter the absolute value of each value in A at most by Σ_{ii}. In general, setting Σ_{ii} to zero subtracts a matrix Γ_i with rank 1 from A, as $(\Gamma_i)_{jk} = U_{ji}\Sigma_{ii}V_{ki}$, and thus all columns are linearly dependent.

The orthogonality properties imply that the order of the rows and columns in the original matrix A do not change the singular values: the order of the rows and columns is irrelevant for the SVD method to detect RFI. Intuitively, the SVD method does not "distinguish" between a smoothly increasing amplitude, caused by astronomical sources, and RFI, and might fail to correctly subtract or detect RFI because of the astronomical signal.

If RFI is to be separated from the signal, the RFI and the signal have to adhere to the following properties:

- All columns containing RFI (and consequently all rows) have to be orthogonal to the astronomical signal. In other words, for any column or row $\boldsymbol{a}$ in the matrix, $\boldsymbol{a}_{\text{RFI}} \cdot \boldsymbol{a}_{\text{signal}} = 0$, with $\boldsymbol{a}_{\text{RFI}}$ the RFI component and $\boldsymbol{a}_{\text{signal}}$ the signal component in the data.

- The singular values of the RFI are substantially higher than the singular values of the astronomical signal. This requires the RFI to be strong.

- The individual RFI columns are either fully linearly dependent on or fully orthogonal to each other. If the individual RFI components are partially dependent, the largest part of the RFI is removed and the singular value of what is left of the RFI might not have enough 'gain' to be removed or flagged.

Iteratively fitting a surface and subtracting the surface, as in §2.2.2, might improve the compliance to the first requirement, although it increases the execution time of the method. Another way to improve compliance to the requirement is to remove the astronomical signal by subtracting a good model beforehand.

It is useful to note that unitary transformations do not change the singular values of a matrix, although they might change the singular vectors. Since the Fourier transform is a unitary transformation according to Parseval's theorem, the following equation holds:

$$A = USV \Leftrightarrow \mathcal{F}(A) = U'SV' \tag{2.19}$$

The consequence of this is that it does not matter whether the SVD method is executed in the time-frequency domain, the time-lag domain, or another Fourier domain, since setting singular values to zero in the Fourier domain would set the singular value to zero in the original domain.

2.2.8 Input data types

The combined thresholding methods described in this paper can be applied to several types of data: auto-correlated or cross-correlated, to the cross-correlations of specific polarized feeds, to Stokes parameters, to amplitude or to phase, etc.

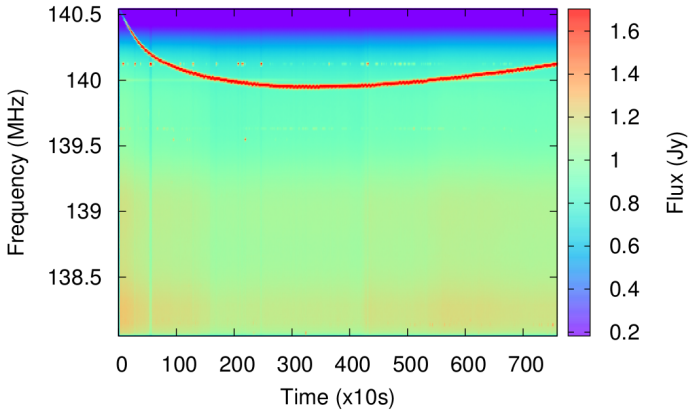
We have compared flagging on cross-correlations and auto-correlations. The cross-correlations of each baseline can be processed with one of the flagging methods, resulting in $N(N-1)/2$ correlations to be processed. Alternatively, every antenna can be individually flagged by processing the auto-correlations, and samples in a baseline might be flagged if either of the corresponding samples in the individual antenna auto-correlations have been flagged. Only N correlations need to be searched for RFI in this case. In addition to the benefit of speed, RFI is strongest in auto-correlations and the data contain no fringes from astronomical sources, as auto-correlations do not have interference patterns, thus offering an improved accuracy in RFI detection. On the down side, some RFI might be present in auto-correlations that would have been mitigated by cross-correlation, and detecting RFI in auto-correlations potentially throws away some usable data in the cross-correlations.

In cases where the polarization of the observed electromagnetic waves is measured, the polarization might contain valuable information for RFI detection. For now, we have processed the cross-correlations from differently polarized feeds individually, without exploiting relationships between these cross-correlations.

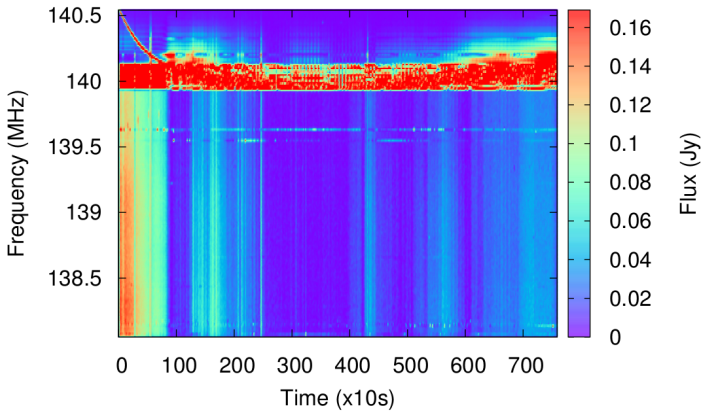
2.3 Thresholding & smoothing results

2.3.1 Smoothing results

In §2.2.2 we described several surface fitting methods to estimate the astronomical signal in the frequency-time domain. We found that the surface fitting methods when combined with one of the detection methods do not differ much in accuracy. A sliding window approach was found to be

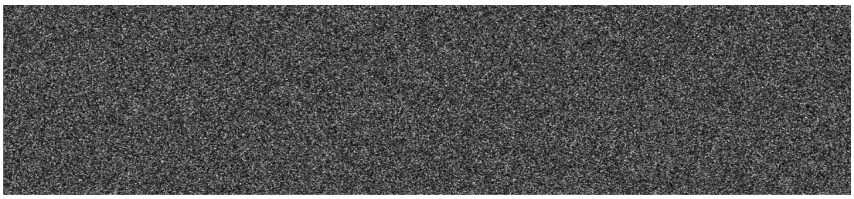


(a) Original observation

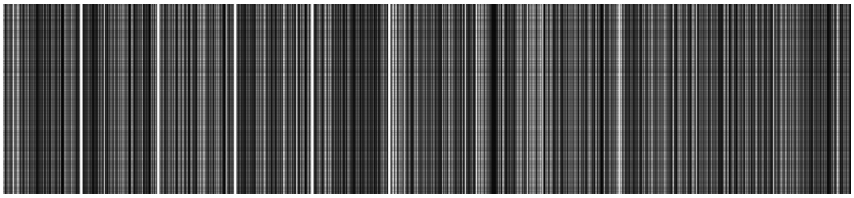


(b) After removing the highest singular values from the image (note the different flux scale).

Figure 2.6: *The auto-correlated data in this image demonstrate the inability of the SVD method to remove sources that slowly change frequency over time (e.g., because the source has a changing velocity in the direction of the antenna). This type of RFI seems to be relatively common in low-frequency WSRT data. The RFI in this particular example is so strong that it can be easily removed by thresholding, but this plot is to illustrate the effects of such RFI. When the frequency-changing signal is faint and cannot be removed by thresholding, applying SVD will, as in this example, change the astronomical information in the data in an unpredictable way.*



(a) Recomposed image from the low singular values.



(b) Recomposed image from the high singular values.

Figure 2.7: *SV decomposition of test set A (Figure 2.8a): noise with broad band RFI covering all channels homogeneously. The recomposed image from the low singular values (top panel) looks very promising: none of the RFI is left and the noise* seems *to be untouched. However, a recomposition of the matrix with only high singular values (bottom panel), i.e., the part that has been subtracted from the image, shows that the noise is affected in an unpredictable way by the decomposition. This is the best case for the SVD method; in more realistic scenario's, the data should include a residual astronomical signal and broadband RFI that might not be linearly dependent.*

more stable compared with a tile based approach. The Gaussian weighted average, a polynomial fit and the window median for the subtracted surface were found to be approximately equal in their accuracies after optimising their parameters such as the window size, the Gaussian kernel size and the order of the polynomial, although their parameters do influence the accuracy.

Finding global parameters that always work well (or automatic procedures to find the parameters) is not trivial. The algorithm can handle data with very different characteristics: it can be applied to XX, XY, YX or YY cross-correlations, to auto-correlations, to either long or short baselines, to LOFAR or for WSRT data, before or after calibration, etc. To use the same surface fitting parameters in all these different situations, the window size, and if applicable the Gaussian kernel size, needs to be rather small. The expected amplitude changes of celestial signals are usually much less in the frequency direction, and setting the window size larger in the frequency direction improves stability. We used a typical size of the sliding window of 40 frequency channels $\times$ 20 time scans and Gaussian kernel parameters of $\sigma_\nu = 15$ and $\sigma_t = 7.5$. The numbers are based on trials using different observed and artificial data sets. The parameters are relative to the number of channels and number of time steps. For WSRT data, a channel is 10 kHz wide and a time scan is 10 seconds long. LOFAR will have a 1 kHz $\times$ 1 second correlation output resolution. For best results, the length and width of the window should be about three times the Gaussian kernel size or larger.

2.3.2 RFI detection results

Both the SVD and threshold methods show accurate results on removing line RFI and broadband RFI. The SVD method is not suitable for removing frequency-varying RFI, as demonstrated in Figure 2.6, and thus has to be complemented with other techniques to remove all RFI. However, the SVD method can be used to subtract and remove the RFI from the image, leaving the astronomical signal intact. For this to be succesful, considerable assumptions about the mathematical properties of RFI and the astronomical signal have to be true: the time-frequency matrix with the RFI components has to be orthogonal to the time-frequency matrix of the astronomical signal, and the different RFI components have to be either orthogonal to each other or linearly dependent on each other. Figure 2.7 shows the SVD decomposition of test set A that consists of uncorrelated noise and linear RFI.

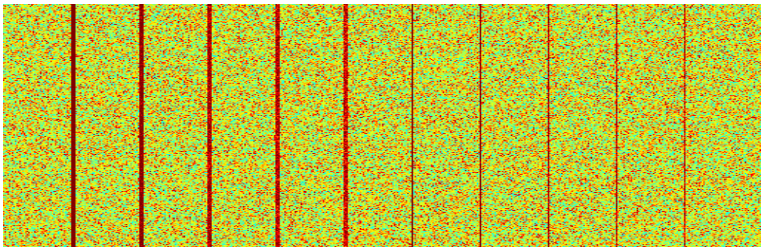
As it is hard to quantitatively compare RFI mitigation methods based on data sets of which the characteristics of the RFI cannot be known for certain, several artificial test sets were created. These sets are shown in Figure 2.8 and contain broadband RFI only. Since the RFI was added artificially, the location of the RFI in the time-frequency domain is known, and the accuracy of the methods can be tested quantitatively. The results are drawn as receiver operating characteristic (ROC) curves in Figure 2.9. ROC curves show the true probability rate against the false probability rate. The different accuracies and characteristics of the methods can easily be compared in ROC graphs.

The `SumThreshold` method shows a considerably better accuracy in all the test sets. Test sets A and B contain RFI that is completely linear dependent, and the SVD method also works very well in these sets. The SVD method could actually be used to subtract the RFI instead of flagging and not using the data. However, to mitigate the RFI in test set C, the methods have to deal with RFI that is neither orthogonal nor completely dependent on each other, and thus the accuracy of the SVD method decreases.

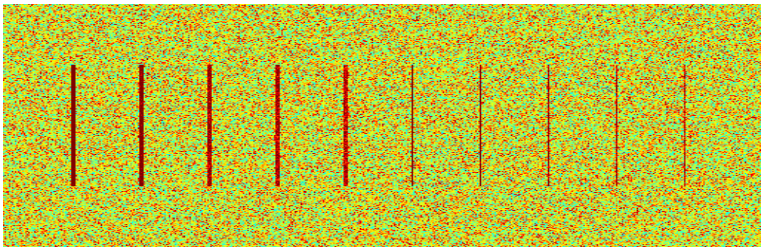
A normal thresholding strategy was also tested to compare the results. When performing normal thresholding with a surface fit as in the `SumThreshold` method, the accuracy for thresholding actually decreases in the test cases without an astronomical signal (see the curves labelled "Fit + simple threshold" in Figure 2.9). This is partially because the surface fit was optimised for the `SumThreshold` method. Furthermore, since the accuracy of the thresholding is not very good, the fit is influenced by the undetected RFI, causing more errors.

When astronomical information is added as in test set E and a more complex background is added as in test set F, the SVD method shows a decreased accuracy in mitigating the RFI, as can also be seen in Figure 2.10. However, in test set G, the background of test set F is Gaussian smoothed and subtracted, as is done before thresholding. The SVD method now shows an improved accuracy, though still not as good as the `SumThreshold` method. Test set H shows that the linear dependency of the RFI is not the only requirement for succesful mitigation with the SVD method: the added RFI is completely linearly dependent in this test set, but the background is still causing low accuracies in the SVD method.

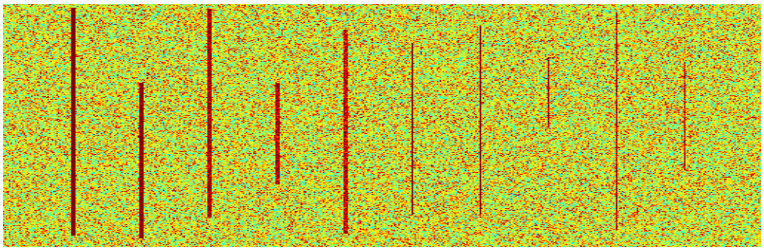
It should be noted that some of these test sets are measuring the theoretical accuracy of non-orthogonal, but not completely independent RFI contamination. As shown in §2.2.7, this was the hardest case for the SVD method. When in practice the RFI does behave in an orthogonal or dependent manner, the results might be quite different. Nevertheless, it is unlikely that all RFI contaminations that are measured by different antennae at different times are always either linearly dependent or orthogonal.



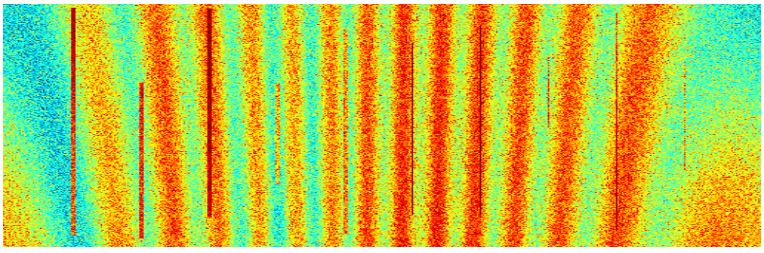
(a) Test set A: noise with broadband RFI contaminating all channels, ordered from strong (left) to weak (right).



(b) Test set B: broadband RFI contaminating a part of the channels

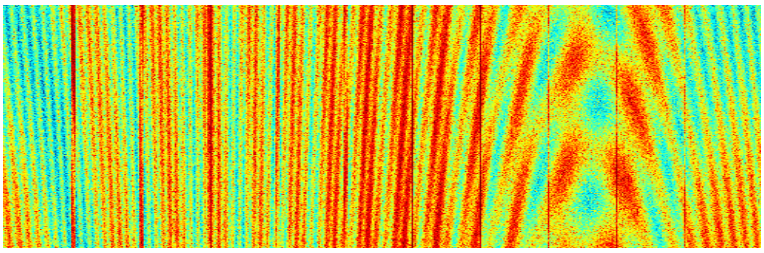


(c) Test set C: broadband RFI contaminating different channels

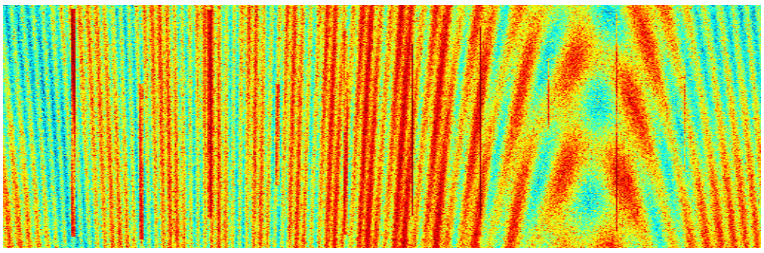


(d) Test set D: a simulated observation of the cross-correlation of three point sources being close together added to test set C

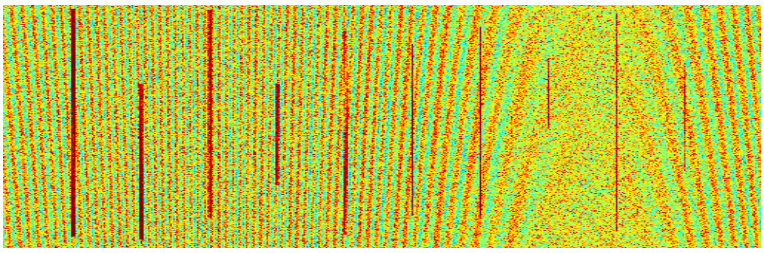
Figure 2.8: *(continued on next page)*



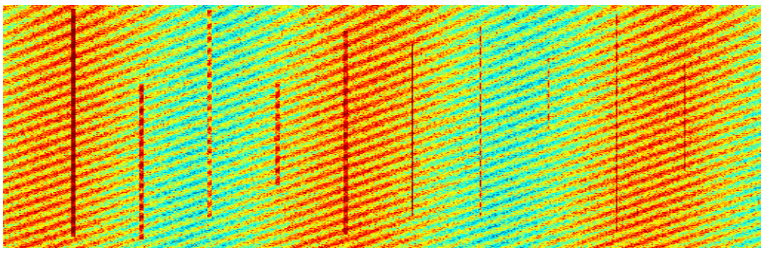
(e) Test set E: a simulated observation of the cross-correlation of five distant sources added to test set A



(f) Test set F: as E, but RFI as in test set C



(g) Test set G: as F, but Gaussian smoothed before adding RFI



(h) Test set H: a high frequency background signal added to test set C

Figure 2.8: *The artificial test sets containing broad band RFI, used for testing and parameter optimisation. In all images, time is along the horizontal and frequency along the vertical axis. All test sets simulate a similar baseline.*

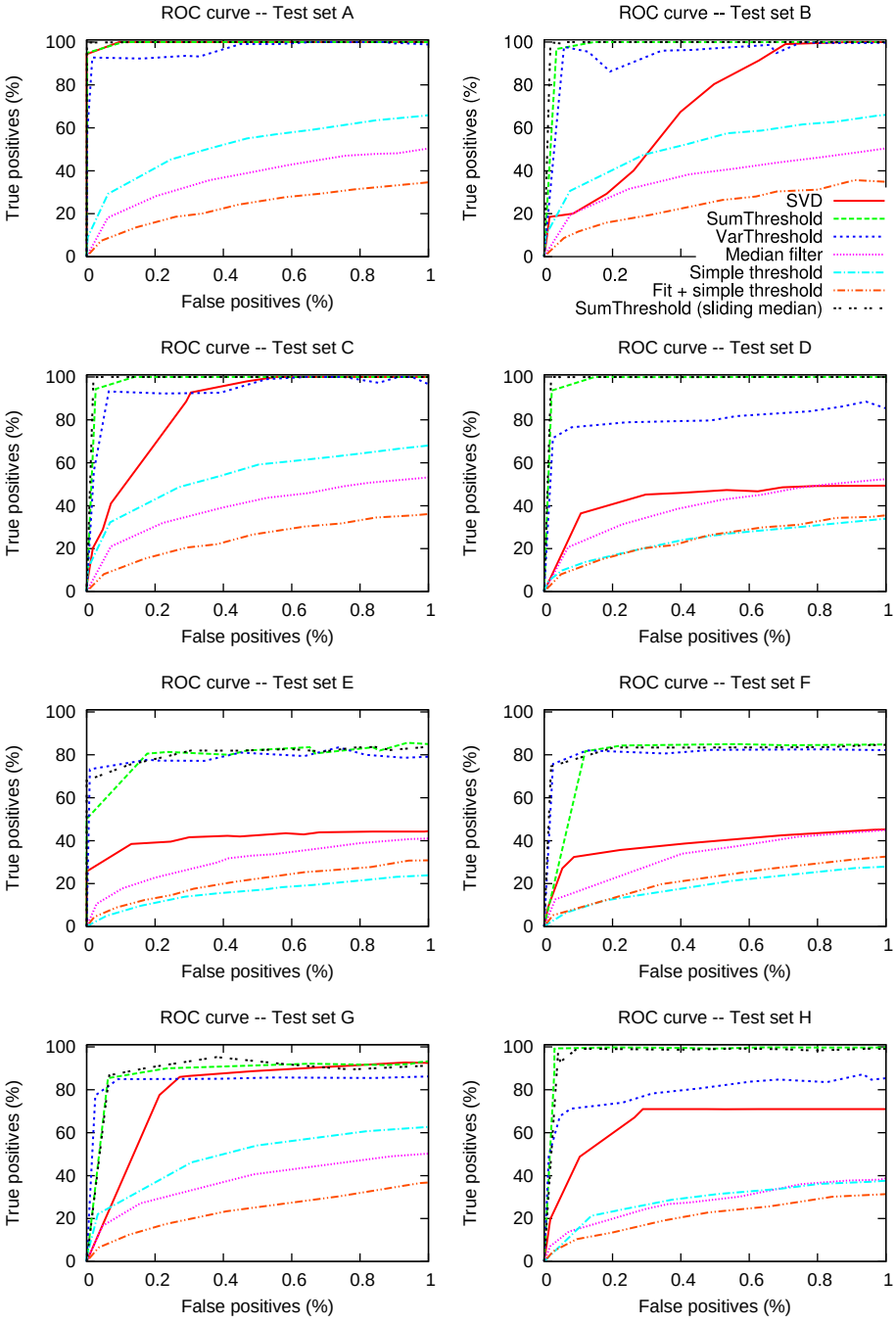
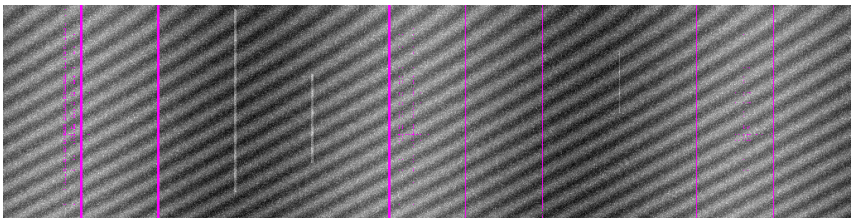
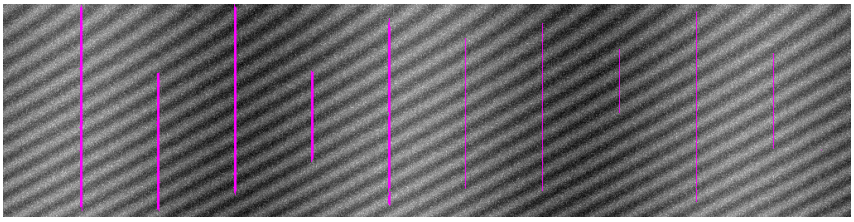


Figure 2.9: *The ROC curves produced by applying various RFI detection methods to the test sets. The closer an ROC curve passes the top-left of the graph at 100% true-positives with 0% false-positives, the more accurate the method is.*



(a) SVD performed on test set H (71.0% recognized, 0.6% false).



(b) `SumThreshold` performed on test set H (99.4% recognized, 0% false).

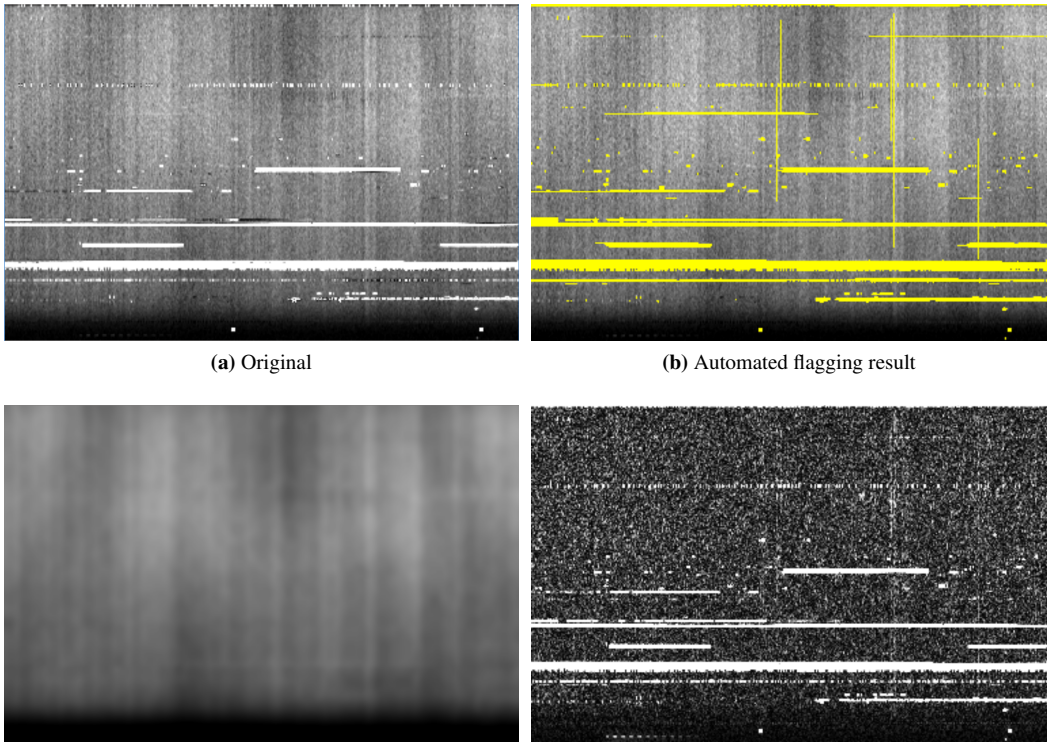
Figure 2.10: *The results of two mitigation methods applied to test set H.*

The presented test sets simulate a single baseline, whereas in a real measurement, the SVD method will exploit the correlation of RFI between different antennae. This will, however, also decrease the probability that all RFI is either orthogonal or linearly dependent.

2.3.3 Automatic flagging of WSRT data

To test the various RFI flagging algorithms we have used WSRT data in the low-frequency front-end (LFFE) band from 138-157 MHz obtained in November and December 2007. The observations have been described and analysed by Bernardi et al. (2009, 2010) to which we refer for details of the astrophysical motivation and calibration. For our analysis, however, we used the raw uncalibrated visibilities. The correlator integration time for the data was 10s. A total of 8 bands of 2.5 MHz width were available. The central frequencies of these bands were located at 139.3, 141.5, 143.7, 145.9, 148.1, 150.3, 152.5 and 154.7 MHz. Each band was divided into 512 spectral channels. The data were Hann tapered, yielding an effective spectral resolution of 9.8 kHz. Therefore, adjacent spectral channels are highly correlated. A total of 13 telescopes participated in the observations providing a total of 78 interferometers with baselines from 36 to 2736 meters. All four cross-correlations between the orthogonal, linearly polarized feeds were used in the analysis.

We have tested the various methods on several data sets. The `SumThreshold` method in combination with Gaussian smoothing shows especially excellent results. Figure 2.11 shows a typical time-frequency diagram of WSRT data at $\sim$140 MHz and the application of the `SumThreshold` method. Although the smoothed surface is slightly affected by the RFI after five iterations, as faint artefacts are visible in the smoothed surface around places where RFI used to be, the effect is so small that it does not pose a problem for the `SumThresholding` method. However, it makes the calculated false probability rate inaccurate, as the false probability calculations assume independence between the residual samples. When validating the results by visual



(a) Original

(b) Automated flagging result

(c) Smoothed

(d) Difference

Figure 2.11: *Time (horizontal) vs. frequency (vertical) plots of uncalibrated WSRT data, cross-correlations of antenna C vs. D. and the application of the* `SumThreshold` *automated flagging procedure. Panel (a) shows one hour of the amplitude of a 3C196 observation, panel (b) shows the result of the flagger, panel (c) shows the fitted surface after 5 iterations, and panel (d) shows the difference between panel (a) and panel (c).*

inspection, we see far less false detections than the calculated false probability rate.

We were able to use the same parameters for WSRT data from different baseline configurations and different target fields, and therefore were able to completely automate the flagging process. Even at baselines and frequencies with dramatic RFI contamination of up to 50%, the `SumThreshold` flagging method remained stable and accurate. Figure 2.12 shows, for example, a badly contaminated band of WSRT data that is almost perfectly RFI flagged.

2.3.4 Conclusion and discussion

In this chapter we have shown several approaches to deal with RFI that is left after correlation. The results show that automated flagging with the `SumThreshold` method works well for broadband and peak RFI. In all cases, the default parameters for the method work well, although parameter tweaking might in some cases improve the detection. In the artificial broadband RFI situations, it detects 80% of the artificially inserted RFI with less than 0.1% error, and often approaches a 99% recognition almost without error. The accuracy of this method is therefore as good as can be expected from manual flagging. In the case of WSRT, the new method does not improve the dynamic range of the data compared with manual flagging, but the method saves a considerable amount of work.

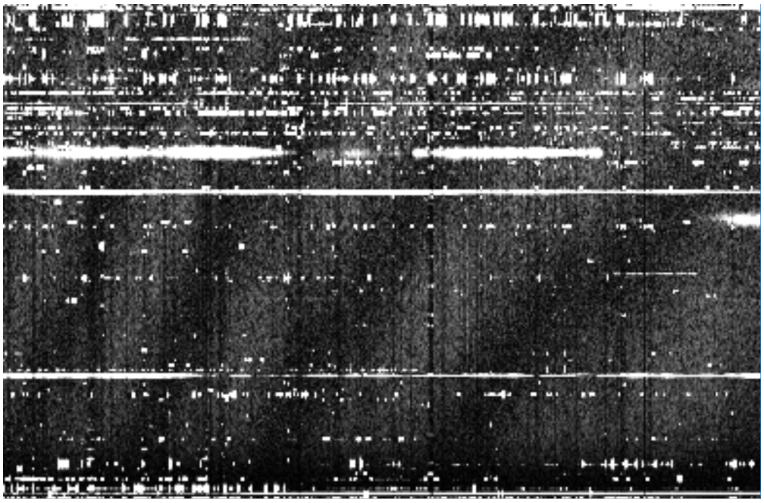
New telescopes such as LOFAR and SKA require robust automatic procedures, as these telescopes will produce data sets that exceed current measurements in volume by orders of magnitude. The ability to flag or check baselines or subbands individually will be lost.

The ROC analysis shows that the `SumThreshold` method is to be preferred above the `VarThreshold` and SVD methods. The SVD method can be used in some respects to detect RFI, but is less accurate. It can either be used to detect the RFI or to correct samples. If it is used to correct samples by filtering the RFI out, rather than only detecting and flagging it, artefacts with unknown characteristics could remain in the data. For WSRT data, these artefacts look as bad as the broadband RFI itself.

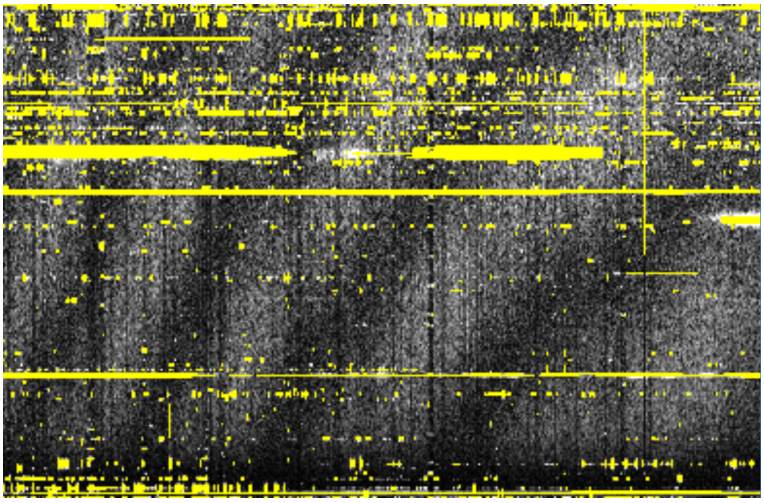
All methods have been tested without assuming a data model. Subtracting the model before RFI detection might improve the detection further. Nevertheless, the detection accuracy with and without a model do not differ much. As such, going back and forth between flagging data and creating a model is not necessary in most cases.

2.3.5 Further work

RFI with a moderate strength that can be detected by eye was found to be of no concern for automatic flagging methods in sensitive telescopes such as WSRT. However, a different kind of RFI might still pose problems. Certain weak RFI, such as radiation that leaks from cabins in situ, might be present in many channels for a substantial duration of the observation. This might pose problems for observations that require long integration times to achieve their required signal-to-noise ratios, such as the LOFAR-EoR project. If the RFI is persistent in time, systematic errors could result. There are some interesting ways to remove these, and one of them is the fringe-fitting RFI mitigation method described by Athreya (2009). Although this technique works at the GMRT, preliminary tests with the fringe-fitting RFI mitigation method on WSRT and LOFAR data do not show a strong presence of this type of RFI, and removing very weak RFI with a similar method requires more work. Therefore, to determine whether this type of RFI is really present, and whether it might be removable is yet to be seen.



(a) Original



(b) Automated flagging result

Figure 2.12: *Time (horizontal) vs. frequency (vertical) plots of WSRT data, cross-correlations of antenna 1 vs. 2: a particular bad band at 121.3 MHz - 123.7 MHz of an observation of 3C147, showing that the method remains robust in one of the worst cases at the WSRT.*

An important next step is to consider practical issues in RFI mitigation techniques. For example, the effects of many RFI mitigation methods, post-correlation as well as pre-correlation, need to be simulated, since we never know what the image plane ought to look like. Also, which post-correlation and pre-correlation methods can be combined? Under which practical circumstances do RFI mitigation methods fail? How can we be sure that astronomical detections are not caused by RFI, or by the methods that try to reduce RFI? Answering these questions is important for establishing the reliability of new RFI mitigation methods and for their regular use by astronomers.

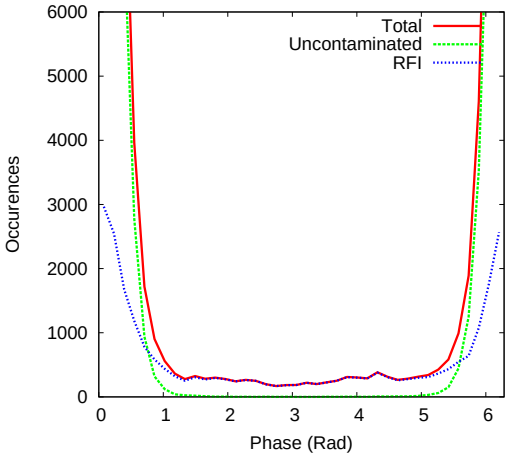


Figure 2.13: *Typical histogram of the phase in a short baseline of a WSRT observation. The RFI was detected by using the* `SumThreshold` *method. The plot implies that RFI-contaminated samples have a much higher probability to have a phase deviating from zero, and the phase thus contains valuable information for RFI detection.*

Although, at this point, it seems to be of little concern to improve the `SumThreshold` automatic flagging method any further, it might be interesting to improve it by combining more information for detection and by using fuzzy logic to decide the sample classification. An interesting example would be to include phase information in the recognition, as only the amplitude information has been used so far by the threshold methods. For example, Figure 2.13 shows that the phase contains valuable information about a sample: in uncontaminated samples, the phase is likely to be near zero rotation, whereas many contaminated samples do have a phase deviating from zero. Other distinguishing information could be contained in the polarization information per sample and in the combination of different baselines.

Based on the low-frequency observations with the WSRT, it can be expected that the radio environment of LOFAR is sufficiently clean for sensitive astronomical experiments. In upcoming chapters we will fully analyse and describe the LOFAR environment and the effectiveness of the RFI strategies.

Finally, we would like to emphasise that the methodology of RFI flagging, or any kind of error detection, needs to change because of the introduction of telescopes such as LOFAR, that generate so much data that it is not possible for astronomers to browse through the data for “the

baseline that was producing this artefact" or "the timestep that corresponds to these stripes in my image". Therefore, another important next step is to be able to automatically detect errors that are caused by RFI, calibration issues, broken hardware, faulty software or any step in the complicated pipeline of a radio observatory.

2.4 Morphological detection

In the previous sections, we have seen that RFI detection often involves thresholding. Often, detection is based on amplitude thresholding (Winkel et al., 2006; Offringa et al., 2010a), although also higher order statistics such as the kurtosis have been used (Gary et al., 2010). The latter requires storing both the mean powers and the squared powers, thereby doubling the data rate, and hence is not always usable. Most interfering sources radiate either in a constant small frequency range, or produce a broadband peak over a short time interval. Examples of such interferences are respectively air traffic communication and lightning. Consequently, an interfering source tends to affect multiple neighbouring samples in the time-frequency domain. These samples form straight lines, parallel to the time and frequency axes. An example is given in Fig. 2.14(a), which shows data from the Westerbork Synthesis Radio Telescope (WSRT). This line-shaped behaviour of RFI can be used to improve the accuracy of detection algorithms. In §2.2.6 thresholding algorithms were described that use this information, such as the `SumThreshold` method, which shows a very high detection accuracy compared to other methods (Offringa et al., 2010a). This method is used in one of the steps in the AOFlagger pipeline (Offringa et al., 2010b). An important consideration for succesful application of automated feature detection algorithms such as these, is that the signal of interest should not contain significant line-shaped features, as is the case with spectral line observations. Also, methods that assume straight, one-dimensional features in the time-frequency domain, might not work well in situations where the features are curved. This can occur when both the frequency and the time resolutions are high enough to resolve frequency variation in sources, for example when sources are Doppler shifted or vary intrinsically, such as with certain radar signals. With LOFAR, we see very few such sources.

Typically, the received power of interfering sources varies over time and frequency. This happens because of several effects, such as intrinsic variation of the source; changing ionosphere; and because of instrumental effects. A typical example of the latter, which is present in almost every observation, is the change of the telescope's gain towards a terrestrial source as the telescope tracks a field in the sky. Like time variation, frequency variation can be caused intrinsically by the source. The instrument also adds frequency-dependent gain, for example due to imperfect band-pass filters. Even though radiation from a source might be continuously received by the telescope, thresholding detection methods might fail to detect the interferer over its full range due to the variation in received power. Figure 2.14(b) shows an example where this is likely the case. Increasing the sensitivity of the thresholding method might help somewhat, but will also cause an increase of false positives. While some falsely detected samples are tolerable, they should be kept minimal in order to avoid data bias and insufficient uv-coverage.

In the following section we will use the mathematical morphology of RFI for increasing detection accuracy. Using mathematical morphology for this purpose is not a new idea; a dilation is often used during RFI processing to flag areas near high values in the time-frequency domain. An example of this can be found in Winkel et al. (2006), where windows of 5 time steps $\times$ 5 frequency channels around detected samples are flagged. However, standard morphological techniques are not scale invariant. An operator is called scale invariant if scaling its input results in the same scaling of its output. An ordinary dilation will cause sharp RFI features to create a high amount of false positives, while flagging smooth RFI features requires a very large dilation kernel. Another scale-dependent technique used for RFI detection is to consider the statistics of time steps and frequency channels. In the upcoming sections, we will show that scale invariance is a desirable property of RFI detection algorithms. In these sections we will provide:

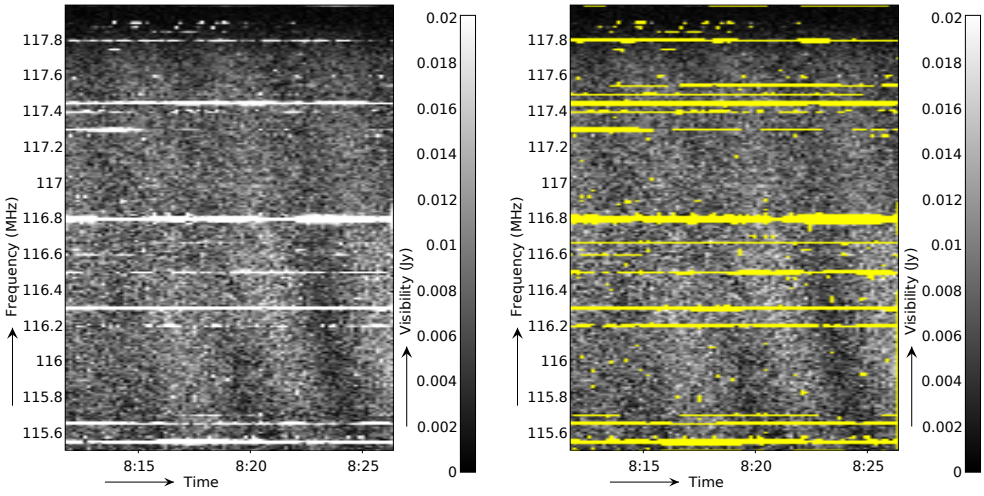


Figure 2.14: *Typical spectral line RFI received in a short period of WSRT data around 117 MHz. It is likely that such RFI sources transmit continuously within a small bandwidth. Panel (a) shows the original observation, while panel (b) shows what the AOFlagger with default settings would flag without morphology-based flagging. Detection is quite accurate, but some of the detected lines in panel (b) are not continuous. It is likely that those RFI sources were active in the gaps as well. Morphology-based detection will help in such cases. The plot shows Stokes I amplitudes of the cross-correlation of antennas RT0 × RT1, which is a 144m East-West baseline. A single pixel is 10 seconds × 10 kHz of data.*

- A detailed description of a morphological technique for RFI detection introduced in Offringa et al. (2010b);
- Analysis of the technique and a comparison with an ordinary dilation, using simulations and real data from two different radio-observatories;
- A novel fast algorithm with linear time complexity to implement the technique.

The method that will be discussed flags additional samples that are likely to be contaminated with RFI, based on the morphology of the flag mask output of a thresholding stage in the pipeline. In sect. 2.4.1 we describe the technique and show a fast algorithm to implement it. We present some results of the method on simulated data and real data in Sect. 2.4.2. Finally, we summarize and discuss the results in Sect. 2.4.3.

2.4.1 The scale-invariant rank operator

RFI features such as in Fig. 2.15(a) are common in radio observations, and can occur at different scales. However, a morphological dilation is not *scale invariant*, and will thus necessarily work

better for some RFI features than others. To overcome this problem, we will describe and analyse a morphological rank operator that is scale invariant[5]. Scale invariance is a desirable property of RFI detection algorithms, because (a) it implies the method can be applied on data with different resolutions without changing parameters and (b) the time and frequency scale of RFI itself can be arbitrary, so any method to detect RFI should work equally well for RFI at different scales. In practice, RFI seems to behave in a more or less scale-invariant manner at the resolution of LOFAR, as for example can be seen in Fig. 2.14, so we should also use a scale-invariant method to detect it. This scale-invariant behaviour of RFI breaks down at high time and frequency resolutions, at which many features become diagonal in the time-frequency plane.

The proposed technique was first mentioned in Offringa et al. (2010b), as it is part of the AOFlagger, which is the default LOFAR RFI detection pipeline. In that article the operation was referred to as a dilation, however, it does not strictly adhere to all the properties of a morphological dilation. For example, we will see that the operator ρ is not distributive over the union set operator: $\rho(X \cup Y) \neq \rho(X) \cup \rho(Y)$ for some X and Y. Because a rank operator flags points for which the number of flagged points in a neighbourhood exceeds a threshold (Goutsias and Heijmans, 2000, §3.4, Soille, 2002), we will refer to the operator ρ as the scale-invariant rank (SIR) operator.

We will now describe the method in-depth and analyse its effectiveness. In Offringa et al. (2010b), it was mentioned that the full algorithm has a time complexity of $\mathcal{O}(N^2)$, N being the input size of the SIR operator, but by making the algorithm less accurate, an implementation of $\mathcal{O}(N \times \log N)$ was mentioned to be possible. Here, we will introduce a faster algorithm with linear time complexity, which is also an *exact* implementation of the SIR operator.

Description

Consider F, a set of positions in the time-frequency domain, such that a sample at time t and frequency ν has been flagged when $(t, \nu) \in F$. Assume F is the result of a statistical detection algorithm, such as the `SumThreshold` algorithm. We will apply the SIR operator in time and frequency directions separately, and define the sets Θ_t and Φ_ν to contain the flags of a slice in time and frequency direction:

$$\Theta_t \equiv \{(s, \nu) \in F \mid s = t\}, \tag{2.20}$$

$$\Phi_\nu \equiv \{(t, \mu) \in F \mid \mu = \nu\}. \tag{2.21}$$

A single one-dimensional set Θ_t or Φ_ν is the input for the SIR operator. The operator considers a sample to be contaminated with RFI when the sample is in a subsequence of mostly flagged samples. To be more precise, it will flag a subsequence when more than $(1 - \eta)N$ of its samples are flagged, with N the number of samples in the subsequence and η a constant, $0 \leq \eta \leq 1$. Using ρ to denote the operator, the output $\rho(X)$ can be formally defined as

$$\begin{aligned} \rho(X) \quad \equiv \quad & \bigcup \{[Y1, Y2) \mid \\ & \#\left(X \cap [Y1, Y2)\right) \geq (1 - \eta)(Y2 - Y1)\}, \end{aligned} \tag{2.22}$$

with $[Y_1, Y_2)$ a half-open interval of Θ_t or Φ_ν, and the hash symbol $\#$ denoting the count-operator that returns the number of elements in the set. In words, Equation 2.22 defines $\rho(X)$ to consist

[5]The mathematical properties of this technique will be analysed in more detail in van de Gronde et al., 2012, in preparation.

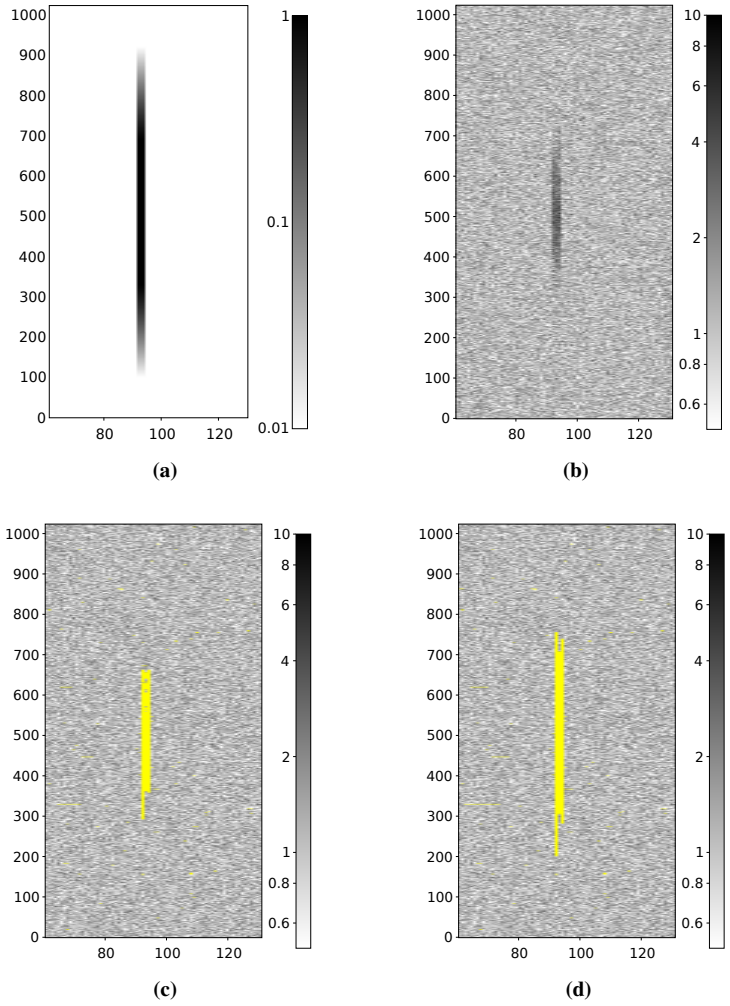


Figure 2.15: *Simulation of a typical broadband RFI feature with Gaussian frequency profile as used in the ROC analysis. Panel (a): isolated RFI feature; panel (b): when noise is added, a part of the feature becomes undetectable; panel (c): flagged with the* `SumThreshold` *method; panel (d): with SIR operator applied, parameter* $\eta = 0.2$.

of all the samples that are in an interval $[Y1, Y2)$, in which the ratio of samples in the input X is greater or equal than $(1 - \eta)$. Parameter η represents the aggressiveness of the method: with $\eta = 0$, no additional samples are flagged and $\rho(X) = X$. On the other hand, $\eta = 1$ implies all samples will be flagged. Figure 2.15 shows an example of a simulated Gaussian broadband RFI feature, and the input and output of the SIR-operator.

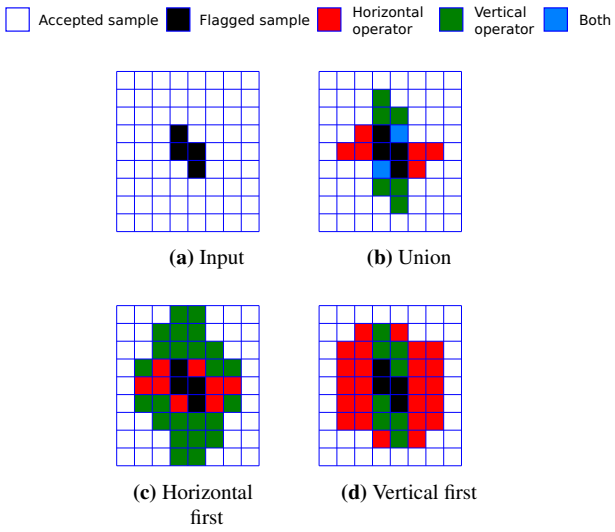
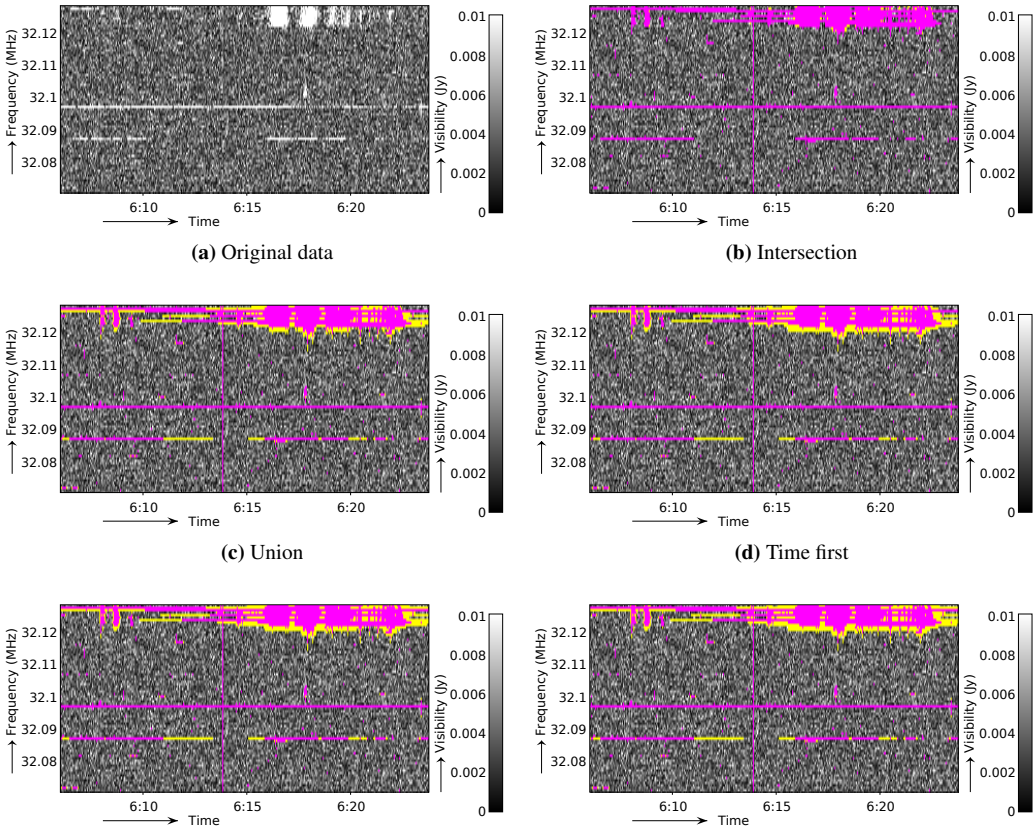


Figure 2.16: *Example outputs of the SIR operator in which the one-dimensional output has been combined in three different ways. Panel (a) is the input, panel (b) shows the result of performing a union on the outputs of both directions, and in panels (c) and (d), the SIR operator was first applied in, respectively, the horizontal and vertical direction. Parameter η was 0.5 in this example.*

The one-dimensional outputs can be remapped to the original two-dimensional domain in various ways. A simple and useful way is to perform a logical union of $\Theta'_t = \rho(\Theta_t)$ and $\Phi'_\nu = \rho(\Phi_\nu)$, the flags on respectively the time and frequency outputs:

$$F' = \left(\bigcup_t \Theta'_t\right) \cup \left(\bigcup_\nu \Phi'_\nu\right). \tag{2.23}$$

An alternative is to initially apply the SIR operator only in one direction, i.e., on the sets that correspond with either the time or frequency direction, and subsequently applying the SIR operator on the outputs of the first in the other direction. The latter is more aggressive than the former. The result also depends on which direction is processed first. The difference is demonstrated in Fig. 2.16, and an example of how that would work out on actual data is given in Fig. 2.17. Optionally, the operator can be applied in frequency and time directions with different η, if one suspects that RFI acts differently in either direction.



(a) Original data

(b) Intersection

(c) Union

(d) Time first

(e) Frequency first

(f) Union of (d) and (e)

Figure 2.17: *Example of the SIR operator applied on a LOFAR observation, displaying five different methods to make the SIR operator two-dimensional. The visibilities shown are from baseline CS003 $\times$ CS007 of a LOFAR low-band-antenna (LBA) observation with 3s $\times$ 0.8 kHz resolution. This observation part was selected as an example because it has a two-dimensional RFI structure. Such RFI is less common, hence this is not a typical case. With the exception of the intersection, there is no difference between the different methods on the thin lines below 32.1 MHz. Applying the operator sequentially (panels d, e and f) is more aggressive for the two-dimensional structures, as it will flag samples that have diagonal neighbours that are flagged. Intersecting the two methods (panel b) will only flag concave samples. Pink is pre-flagged by the* `SumThreshold` *method, yellow is added by the SIR operator. A value of $\eta = 0.2$ was used in this example.*

Properties & parameters

Consider the case in Equation (2.22) when a subsequence of arbitrary length is flagged. Since the fraction of flagged samples within the subsequence is explicitly used to define its output, the operator is *scale invariant*. Formally, an operator ρ is scale invariant if and only if $\rho(\lambda X) = \lambda\rho(X)$, i.e., scaling the input X with λ followed by ρ is equal to scaling the output $\rho(X)$ with λ. We will now give a formal proof of the scale invariance of the SIR operator.

Proof. With ρ the SIR operator, we will scale the input X with factor λ. If $\lambda = 0$ we trivially have that $\rho(\lambda X) = \lambda\rho(X)$. Also, if $\rho(\lambda X) = \lambda\rho(X)$ for $\lambda > 0$, it is not difficult to see that we also have $\rho(-\lambda X) = -\lambda\rho(X)$, as mirroring the input will mirror the output. Therefore, assume without loss of generality that $\lambda > 0$. Now, substituting X with λX in Equation (2.22) results in

$$\begin{aligned}\rho(\lambda X) &= \bigcup\{[Y1, Y2) \mid \\ &\quad \#\left(\lambda X \cap [Y1, Y2)\right) \geq (1-\eta)(Y2 - Y1)\}\,.\end{aligned}$$

By using Z_1=Y_1/λ and Z_2=Y_2/λ, this can be rewritten to

$$\begin{aligned}\rho(\lambda X) &= \bigcup\{[\lambda Z1, \lambda Z2) \mid \\ &\quad \#\left(\lambda X \cap [\lambda Z1, \lambda Z2)\right) \geq (1-\eta)(\lambda Z2 - \lambda Z1)\}\,.\end{aligned}$$

If we assume continuous positions, both the left side and the right side of the comparison can be scaled by $1/\lambda$:

$$\begin{aligned}\rho(\lambda X) &= \bigcup\{[\lambda Z1, \lambda Z2) \mid \\ &\quad \#\left(X \cap [Z1, Z2)\right) \geq (1-\eta)(Z2 - Z1)\}\,,\end{aligned}$$

and by using $[\lambda Z_1, \lambda Z_2) = \lambda[Z_1, Z_2)$ and the definition in Equation (2.22), this is equivalent to $\rho(\lambda X) = \lambda\rho(X)$. □

Because the time and frequency dimensions are obviously discrete and finite when applied on radio observations, in practice the scale invariance is limited by the resolution and size of the data.

The aggressiveness of the SIR operator can be controlled with the η parameter, which can be chosen differently for the time and frequency directions. Because the method is scale invariant, the choice of η can be made independent of the time and frequency resolutions of the input. The default η parameter in the LOFAR pipeline is currently $\eta = 0.2$ and is equal in both directions. This value has been determined by tweaking of the parameter and data inspection, e.g. by looking at the resulting time-frequency diagram and projections of the data variances. The results were checked for many observations. Higher values seem to remove too much data without much benefit, while some RFI is left undetected with lower values. The value works well for various telescopes and on different time and frequency resolutions. We will evaluate this setting in Section 2.4.2.

Since most telescopes observe with two linear or circular polarized feeds, RFI detection can consider each cross-correlated polarization individually, and the operator can be applied on each produced mask independently. However, the flag masks are often kept equal between the different cross-correlated polarizations, because calibration might become unstable when, for a particular sample, part of the polarization information is missing. Moreover, if one of the polarization feeds of the telescope has been affected by RFI, it is likely that the others also have been affected. For

these reasons, the approach taken in the LOFAR pipeline is to use the `SumThreshold` method on all cross-correlated polarizations (XX, XY, YX and YY) individually, then flag any sample for which at least one cross-correlation has been flagged, and finally apply the SIR operator once on the combined mask.

The algorithm

A straightforward implementation of the operator in Equation (2.22) is to test each possible contiguous subsequence. In this case, if N is the number of samples in the sequence Θ_t or Φ_ν, $\mathcal{O}(N^2)$ sums of subsequences have to be tested. Since the sums of all subsets can also be constructed in quadratic time complexity, the total time complexity of a straightforward implementation is $\mathcal{O}(N^2)$. We will now show an algorithm that solves the problem with linear time complexity. The algorithm is somewhat similar to the maximum contiguous subsequence sum algorithm.

Listing 1: *Linear time complexity algorithm for the scale-invariant rank operator*

```
function ScaleInvariantRankOperator
  Input:
    N       :  Size
    Ω       :  Input array of size N
               (Ω[i] = 1 ⟹ i is flagged,
                Ω[i] = 0 otherwise)
    η       :  Aggressiveness parameter
  Output:
    Ω'      :  Output flag array of size N

 1:begin
     // Initialize Ψ
 2:  for x = 0...N-1 do Ψ[x] ← η - 1 + Ω[x]

     // Construct an array M such that:
     // M(x) = ∑ j ∈ {0...x-1}: Ψ[j]
 3:  M[0] ← 0
 4:  for x = 0...N-1 do M[x+1] ← M[x] + Ψ[x]

     // Construct array P such that:
     // M[P[x]] = min M[j]: 0 ≤ j ≤ x
 5:  P[0] ← 0
 6:  for x = 1...N-1 do
 7:    P[x] ← P[x-1]
 8:    if M[P[x]] > M[x] then P[x] ← x
 9:  end for

     // Construct array Q such that:
     // M[Q[x]] = max M[j]: x < j < N
10:  Q[N-1] ← N
11:  for x = N-2...0 do
12:    Q[x] ← Q[x+1]
13:    if M[Q[x]] < M[x+1] then Q[x] ← x+1
14:  end for
```

```
    // Flag sample x if M[Q[x]] − M[P[x]] ≥ 0
15: for x = 0...N − 1 do
16:   if M[Q[x]]-M[P[x]]≥ 0 then
17:     Ω′[x]← 1
18:   else
19:     Ω′[x]← 0
20:   end if
21: end for

22: return Ω′;
23:end
```

Listing 1 shows a direct algorithm to solve the SIR operator problem.

Proof. Using the definition of $\Omega(x)$ and $\Omega'(x)$, such that 1 indicates that x is flagged and 0 that it is not, we can rewrite Equation (2.22) as

$$\Omega'(x) = \begin{cases} 1 & \text{if } \exists Y_1 \leq x, Y_2 > x, \text{ such that} \\ & \sum_{y=Y_1}^{Y_2-1} \Omega(y) \geq (1-\eta)(Y_2 - Y_1) \\ 0 & \text{otherwise.} \end{cases} \tag{2.24}$$

In line 2, the array $\Psi(y)$ is initialized such that $\Psi(y) = \eta$ in case y is flagged, and $\Psi(y) = \eta - 1$ otherwise. Equation (2.24) can now be rewritten to the following test:

$$\Omega'(x) = \begin{cases} 1 & \text{if } \exists Y_1 \leq x,\ \exists Y_2 > x : \sum_{y=Y_1}^{Y_2-1} \Psi(y) \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

Line 3-4 initialize $M(x)$ for $0 \leq x \leq N$ to

$$M(x) = \sum_{j=0}^{x-1} \Psi(j),$$

so that Equation (2.25) can be rewritten as

$$\Omega'(x) = \begin{cases} 1 & \text{if } \exists Y_1 \leq x,\ \exists Y_2 > x : \\ & \quad M(Y_2) - M(Y_1) \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.26}$$

Because we are only interested in $\Omega'(x)$ in the range $0 \leq x < N$, we can limit the search for Y_1 and Y_2 to $0 \leq Y_1 \leq x < Y_2 \leq N$. There exists Y_1 and Y_2 in this range such that $M(Y_2) - M(Y_1) \geq 0$, if and only if

$$\max_{y:x<y\leq N} M(y) - \min_{y:0\leq y\leq x} M(y) \geq 0.$$

Lines 5-14 make sure that P and Q are initialized for $0 \leq x < N$, such that

$$\begin{aligned} P(x) &= \operatorname*{argmin}_{y \in 0 \ldots x} M(y), \\ Q(x) &= \operatorname*{argmax}_{y \in x+1 \ldots N} M(y). \end{aligned}$$

Finally, this allows Equation (2.26) to be rewritten as

$$\Omega'(x) = \begin{cases} 1 & \text{if } M(Q(x)) - M(P(x)) \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.27}$$

which is performed and returned in lines 15-23. □

The algorithm is $\Theta(N)$, and performs $3N$ additions or subtractions and $3N - 2$ comparisons on floating point numbers. The algorithm uses the temporary arrays Ψ, M, P and Q, each of size N, with the exception of M which is of size $N + 1$. Array Ψ can be optimized away and the input Ω can be reused for output by assigning directly to it in lines 17 and 19. The total amount of temporary storage required is thus about N floating point values and $2N$ index values, thus $\mathcal{O}(N)$. When the function is applied on a two-dimensional image, as in the case of RFI detection, the temporary storage is negligible, as the number of processed slices is usually much larger than one or two. If η is expressed as a ratio of two integer values, it is possible to scale all values and only use integer math.

The algorithm has been implemented in C++ and takes around 40 lines of code[6].

Because the problem is somewhat similar to the maximum contiguous subsequence sum (Bentley, 1984) and the all maximal contiguous subsequence sum problems, it might be possible to parallelize the algorithm by similar means, e.g. as in Alves et al. (2005). Moreover, parallel algorithms exist for the prefix sum/min/max calculations. For the specific application of RFI detection for LOFAR, the pipeline has already been maximally parallelized by flagging different baselines and/or sub-bands concurrently. Unlike parallelizing on the algorithm level, this requires no communication between the different processes.

2.4.2 Analysis & results

In the following sections, the performance and the accuracy will be analyzed.

Performance

Figure 2.18 displays the performance of the C++ implementations and compares the linear algorithm with the approximate $\mathcal{O}(N \log N)$ algorithm and the full quadratic algorithm. The measurements have been performed on a regular desktop with a 3.07 GHz Intel Core i7 CPU, using only one of its cores. The time complexities of the three algorithms for increasing N behave as expected. The linear algorithm is faster in all cases, even for small N. The $\mathcal{O}(N \log N)$ time complexity algorithm is more than one order of magnitude slower at both small and large N. The linear algorithm has been executed with different values for η. Except for some slight variations — especially for $\eta = 0$ — the algorithm's speed is independent of η.

[6]The implementation is part of the AOFlagger and can be downloaded from `http://www.astro.rug.nl/rfi-software`.

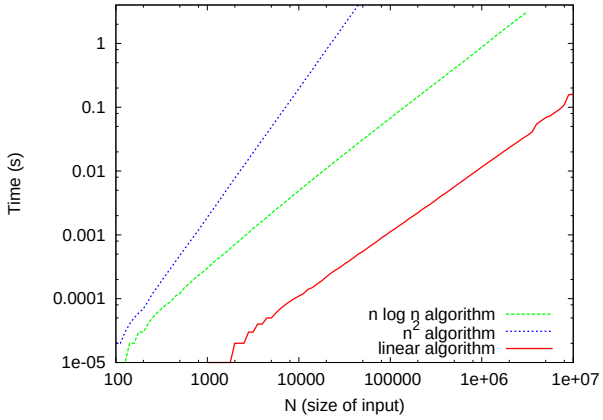


Figure 2.18: *Computation time versus input size with the different algorithms and fixed $\eta = 0.2$. The average over 1000 runs was taken for each different configuration.*

In the LOFAR pipeline, it takes 3.8 seconds to process a single sub-band for a single baseline, assuming 100,000 time steps and 256 channels (which is common). Of these 3.8 seconds, only 49 milliseconds (1.3%) are spent applying the SIR operator. In common applications, an observation contains on the order of a 1,000 baselines and 250 sub-bands. The pipeline is heavily parallelized by concurrently flagging baselines over multiple cores and sub-bands over multiple cluster nodes. In this case, the pipeline's performance is dominated by disk access, and the relative contribution of the SIR operator is even smaller.

Accuracy analysis

The performance of the SIR operator was tested by using receiver operating characteristics (ROC) analysis. To do so, a ground truth needs to be available, which can only be accurately acquired in a simulated environment. As discussed previously, a very large fraction of RFI is line-like. The samples on such a line are not uniform due to intrinsic effects or instrumental gain variations. Therefore, we have used simulations of four line-shaped RFI features as displayed in Fig. 2.19: (a) a single Gaussian that reaches its 3σ point at both borders and is 1 in the centre; (b) three periods of a sinusoidal function which is scaled between zero and one; (c) the Gaussian feature, but slanted by 1/50 fraction; and (d) a burst-like signal in which the amplitude levels are drawn from a Rayleigh distribution with mode $\sigma = 0.6$. All features are three samples wide. Complex Gaussian distributed noise with $\sigma = 1$ was added to the image, such that the amplitudes are Rayleigh distributed. The created two dimensional image of size 180×1024 was subsequently flagged by the `SumThreshold` method with settings as in the LOFAR AOFlagger pipeline, and the created flag mask was used as input.

To estimate the performances, the true and false positives ratios (TP and FP ratios respectively) were calculated after detection. We created a fuzzy ground truth mask in which a value of one corresponds with maximal RFI, zero corresponds with samples not contaminated by RFI, and values in between correspond with lower levels of RFI contamination. Fig. 2.15(a) shows for

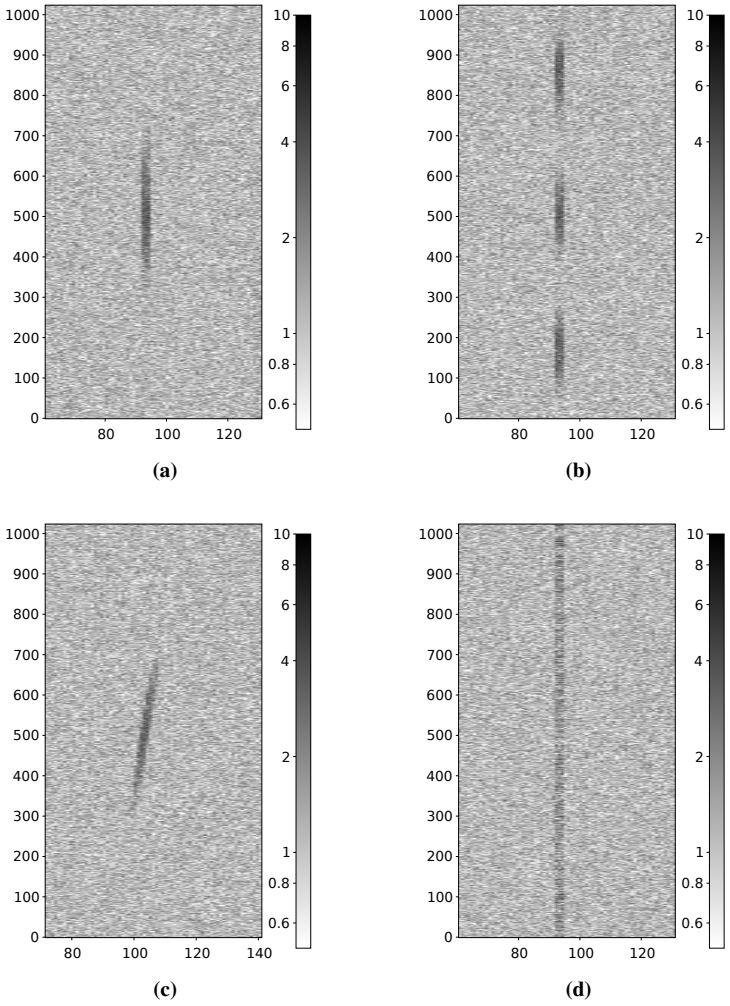
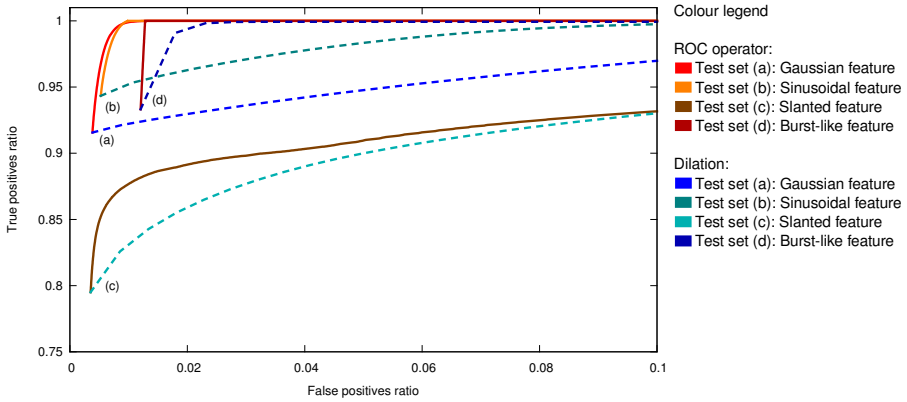
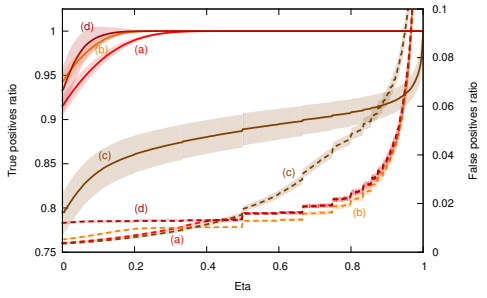


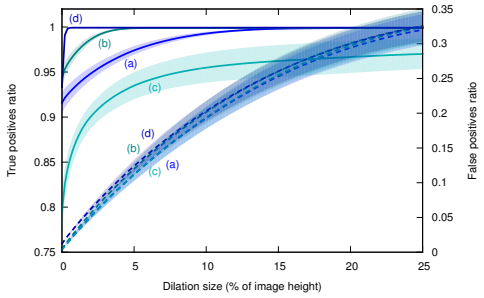
Figure 2.19: *The features used for the accuracy analysis. Panel (a): Feature with Gaussian slope; panel (b): Sinusoidal feature; panel (c): Slanted feature with Gaussian slope; panel (d): Burst feature with samples drawn from a Rayleigh distribution.*



(a) ROC curves. Solid lines: rank operator; dashed lines: dilation.



(b) Influence of η on the SIR operator. Solid lines: true positives (left axis); dashed lines: false positives (right axis).



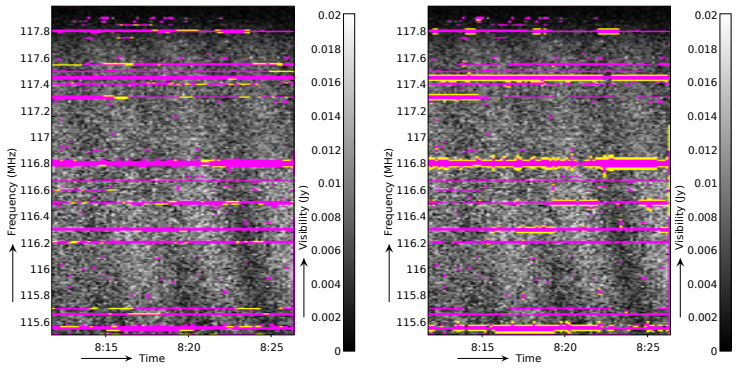
(c) Influence of kernel size on the dilation. Solid lines: true positives (left axis); dashed lines: false positives (right axis).

Figure 2.20: *Analysis of the receiver operating characteristics of the SIR operator and a standard dilation on simulated data. Marks (a)–(d) correspond with the features shown in Fig. 2.19, respectively a Gaussian broadband feature, a sinusoidal feature, a slightly slanted Gaussian feature and a burst-like feature. The shadowed areas show* 1σ *levels over 100 runs.*

example the ground-truth mask of the Gaussian feature. Given a sample with ground truth value β, if the corresponding sample was flagged by the method, it would be counted with ratio β as a true positive and $1 - \beta$ as a false positive. The total TP and FP ratios are the sum of all the TP and FP values, divided by the total sum of positives and negatives in the test set, respectively.

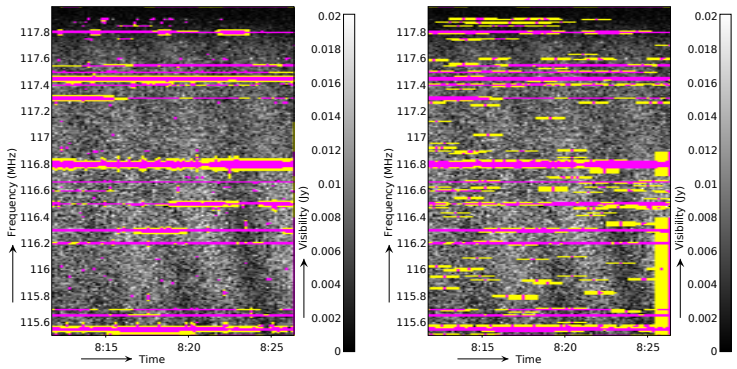
The SIR operator and a standard morphological dilation have been applied in the direction of the feature, i.e., vertical/frequency direction. The true and false positives were varied by changing the parameter η or the dilation size for respectively the SIR operator and the dilation. Different runs gave slightly different results because of the introduced Gaussian noise, hence the simulation was repeated 100 times and the results were averaged.

Figure 2.20 shows the average results. The shadowed areas in panels (b) and (c) show the standard deviation over the 100 runs. In the case of the Gaussian RFI feature, the `SumThreshold` pre-detection removes on average 91.3% RFI power, while simultaneously falsely flagging



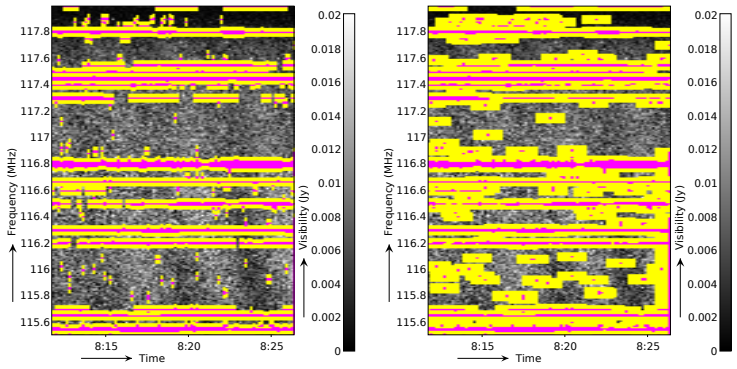
(a) Horizontal SIR operator

(b) Vertical SIR operator



(c) SIR operator in both directions

(d) Horizontal dilation



(e) Vertical dilation

(f) Dilation with a square kernel

Figure 2.21: *Gray-scale plots showing examples of the effectiveness of two morphological techniques on the data from Fig. 2.14. The pink samples have been set by the* `SumThreshold` *algorithm and the yellow samples have additionally been detected with the morphological techniques. Panels a–c show results of the SIR operator with* $\eta = 0.2$ *in the time direction and/or* $\eta = 0.3$ *in frequency direction. Panels d–f show an ordinary dilation with a horizontal kernel of five pixels and/or a vertical kernel of three pixels.*

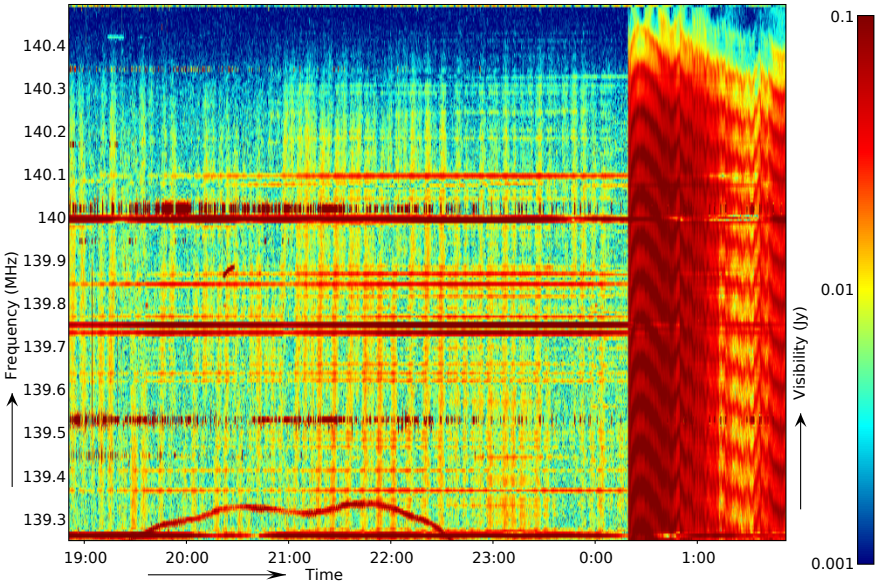
a ratio of 0.38%. Hence, if the methods do not flag any additional samples, they have a TP/FP ratio of 91.3%/0.38%, and this is therefore the start of both ROC curves for this RFI feature. With $\eta = 0.48$, the SIR operator flags all the RFI features with 100% TP with a FP ratio of 1.36%, with the exception of the slanted feature. The SumThreshold pre-detection, dilation and SIR-operator work less well on the slanted feature, and fail to detect it with 100% even at very high sensitivity. The dilation operator needs a size of 32.8% of the height of the image to detect the vertical RFI features. Since it will dilate any falsely detected input sample equally, its FP ratio is 46% with this setting. Changing the signal-to-noise ratio (SNR) of the features changes the scaling of the ROC curves, but the relative difference between the two methods remains the same.

Given the various types of RFI, Fig. 2.20 shows that (I) the SIR operator is extremely accurate, by detecting all previously undetected samples with only a very slight increase in false positives; (II) the SIR operator is superior to the dilation in all tested situations; and (III) a setting of $\eta \sim 0.2$–0.4 seems to be a good compromise between FP and TP ratios.

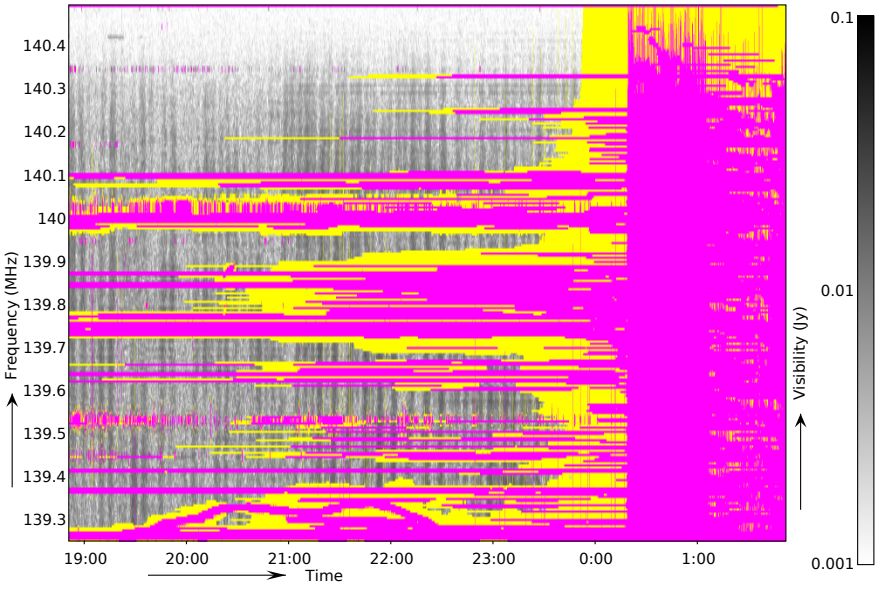
These tests have been performed by applying the operators in one dimension. When applied in two dimensions by using the output of the first dimension as input for the second, the comparison between the dilation and the SIR operator will diverge even more, because the false positives created by the first dimension will be multiplied by the repeated application towards the second dimension. An example of a two-dimensional application is shown in Fig. 2.21. Certain RFI sources create more complex shapes in the time-frequency domain, and contaminate larger non-line like areas. These RFI sources cause higher values in the output of the Gaussian smoothing, which is commonly part of the earlier RFI detection stage, and consequently some of the lower RFI levels of the RFI feature are not flagged. We have seen that the SIR operator will work very well on such features, because it fills the feature and slightly extends the flags in all directions.

It should be noted that one of the assumptions made for the SIR operator to improve detection accuracy, is that parts of the RFI features are not detectable by amplitude thresholding. In practice, however, a small subset of received RFI sources does contribute to an observation with sufficient strength to detect the entire feature with amplitude thresholding. Such transmitters are the worst-case situation for the SIR operator, as the operator will enlarge the flag mask relative to its length, but any samples it flags extra are false positives. Note that it is not useful to perform ROC analysis of such a situation, as the true positives will be constant. The number of false positives can easily be calculated, and scales linearly with η and the duration of the transmitter. For example, when applying the SIR operator with $\eta = 0.2$ on a strong RFI transmitter that occupies ten minutes of data in one channel, the operator will falsely flag two minutes of the channel before and after it. An example of a band that contains intermittent transmitters is the air traffic communication band of 118–137 MHz. Nevertheless, while some of these transmitters are indeed strong, e.g., when they fly through a beam sidelobe, there are also many transmitters at this frequency that are too weakly received to be detected all of the time. Consequently, some of them are only partially detected with amplitude thresholding. This is why we expect better results using the SIR operator even in these bands, compared to using a dilation.

Figure 2.22 shows a WSRT example that contains many different RFI kinds. The initial `SumThreshold` method detects the RFI quite accurately, but it leaves some parts of the last 1.5 hours unflagged. This is solved by the SIR operator, although, because of the sudden start of the RFI, it falsely flags about 20 minutes of data before the start of the RFI. The strong RFI produced by the sporadic transmitter around 140 MHz is flagged by the `SumThreshold` method, but in this case it is likely that these channels have been occupied all of the time. Therefore, the SIR operator gives the desired result by increasing the flags in those channels. All in all, the baseline



(a) Original data



(b) Flagged with `SumThreshold` (pink) followed by the SIR operator (yellow)

Figure 2.22: *An interesting but uncommon WSRT case: part of a baseline of an observation at 140 MHz that suffered unusually strong broadband RFI during the last 1.5 hrs. It also contains many different kinds of transmitters that mostly occupy constant channels. The vertical stripes are fringes of celestial sources, hence contain the information of interest. The image shown is* 2000×250 *samples in size.*

might be somewhat overflagged. Nevertheless, it does allow further data reduction without manual intervention, and without thresholding part of the noise. Moreover, this case is exceptional, and the sudden start of very strong broadband RFI is (fortunately) seldomly seen, while the sporadic transmitters such as the one at 140 MHz are seen very often.

A final remark on the ROC analysis performed here is that the given absolute true/false positive ratios are not an accurate representation of actual RFI detection, because our two models are very simplistic and based on the assumption that RFI behaves in a well defined manner. Establishing absolute true/false positive ratios would require a detailed statistical model of the behaviour of RFI. A realistic estimate for the number of samples occupied by RFI with LOFAR is in the order of a few percent (Offringa and de Bruyn, 2011).

2.4.3 Conclusions and discussion

From panel 2.20(a) it is clear that the SIR operator is much more suitable to detect the tested kinds of RFI than an ordinary morphological dilation. A value of $\eta = 0.2$ was determined by tweaking and validating the results to be a reasonable setting for the LOFAR RFI pipeline, and has been used in the default LOFAR pipeline for over a year. Panel 2.20(b) shows that this value of η agrees approximately with what is found in the simulations: at $\eta = 0.2$, the vertical features have almost been completely detected by the SIR operator (Gaussian: 98.9%, a 7.6% increase, sinusoidal: 99.9%, 5.8% increase, burst: 100%, 6.7% increase) with a minor increase in the false positive ratio (Gaussian: 0.69%, an 0.31% increase, sinusoidal: 0.95%, 0.42% increase, burst: 1.3%, 0.08% increase). The slanted feature is not as accurately detected (86%, 6.1% increase), but the method does enhance the detection. It is hard to give a similar optimal value for the dilation operation, since the false positives scale linearly with the size of the dilation kernel. Therefore, it depends on what is an acceptable loss in terms of the false positives.

In the case of simulated Gaussian broadband features, only 8.7% of the RFI power was not detected by the `SumThreshold` method. For the sinusoidal and burst features, the `SumThreshold` method performs even better. Therefore, the total benefit of the SIR operator might seem small. However, we think that there are strong reasons to use the method:

- The added false positives are almost negligible, and the chances of biasing your data are much smaller compared to using amplitude thresholding exclusively. For example, thresholding biases the final distribution of uncorrelated white noise, while morphologically extending a flag mask does not. For these reasons, it is preferable to use morphology to find the final few RFI samples, compared to lowering the threshold.

- The method is extremely fast and simple, and its processing time is almost negligible in a full RFI pipeline.

- We have seen situations in which even the low ratio of false negatives that are leaked through an amplitude-based RFI detection pipeline can cause calibration to fail. Empirically, we have seen an improvement of the calibratability of LOFAR observations by using the morphological method.

Section 2.4.2 describes that strong intermittent (on $\sim$minute scale) RFI transmitters are probably the worst case for the SIR operator, as in these cases the application of the operator with $\eta = 0.2$ could in theory yield 40% false positives. However, because of LOFAR's high resolution,

in combination with the `SumThreshold`'s unprecedented detection accuracy, the total percentage of flagged data in the case of LOFAR is only a few percent. This implies that even if a large ratio of these were strong intermittent transmitters – which is unlikely – the benefits of not having to manually consider data quality in cases where the technique does help, probably outweigh the $\sim$1% added false positives. If it turns out that some bands do have mostly strong transient transmitters, the η parameter could become a function of frequency. At the moment, application of the SIR operator seems to be helpful at any frequency.

In this paper we have assumed scale-invariant behaviour for RFI. In reality this might not be entirely accurate, so instead of using a threshold that grows linearly with the scale, as in our definition of the SIR operator, it might be better to have a threshold that depends on the scale in a non-linear fashion. Also, when looking at the problem from a statistical point of view, RFI might not be equally likely to occur on all scales. For example, RFI might be less likely to occur on a scale of days than on a scale of seconds. When it does occur on large scales, it is doubtful that we actually need to extend the detected intervals in a scale-invariant manner, because the signal would likely already be detected at a smaller scales and gaps would likely be filled. Such considerations would suggest that it might be better to have a threshold that grows less than linearly for large scales. Better RFI statistics and RFI modelling might provide the required information for assessing such considerations.

Several options are available to apply the SIR operator on a two-dimensional input. As shown in Fig. 2.17, the intersection of the results in both directions does not extend line RFI, thus is not useful in this context. A union does extend such RFI, but does not extend the flags diagonally. Processing the directions sequentially might therefore be beneficial for RFI that has structure in both frequency and time, as this kind of RFI does likely also slightly contribute in the diagonal direction. The difference between processing time first, frequency first or taking the union of both, is small. Taking the union overcomes the somewhat arbitrary decision of which direction to process first. In the case of LOFAR, we decided to only perform filtering time first, because taking the union of both time and frequency first is more expensive.

Morphology can be used in several image processing tasks, for example in feature detection. Often, generic morphological operations need to be applied on different resolutions. In such cases, the scale-invariant operation to extend binary masks as presented here might be generally useful.

So far, we have considered combinations of one-dimensional application of the SIR operator in order to use it for our two-dimensional application in the time-frequency domain. For this application, but also for more generic applications, it might be interesting to consider a true two-dimensional version of the SIR operator. While the one-dimensional operator selects all subsequences (lines) with a ratio $\geq \eta$ of flagged values, a two-dimensional operator would select all *rectangles* that have a ratio $\geq \eta$ of flagged values. It is however likely that such an operator can not be implemented with a linear time complexity, which makes it less attractive for the large data rate of LOFAR.

We have shown that even slightly slanted features are harder to detect accurately. Fortunately, in the case of LOFAR, such features are very rare. If the features to be detected have a known orientation that is not parallel to one of the axes, it might be an option to apply the operator in the direction of the features. While a trivial implementation can apply the operator along fixed lines, some work might be necessary to maintain translation invariance (Soille and Talbot, 2001).

Chapter 3

The LOFAR RFI pipeline

Based on:
"*A LOFAR RFI detection pipeline and its first results*"
(Offringa et al., 2010, Proc. of RFI2010)
"*Interference detection results with LOFAR*"
(Offringa and de Bruyn, 2011, Proc. of URSI General Assembly 2011)

MANUAL flagging by the data reducing astronomer used to be sufficient for dealing with RFI. However, because of the major increase in resolution and bandwidth of modern observatories, such as LOFAR, the GMRT, the EVLA and the MWA, that generate observations of tens of terabytes, this is no longer feasible. The tendency is therefore to implement automated RFI flagging pipelines in the observatory's pipeline. Examples of these are the RFI mitigation pipeline used for the Effelsberg Bonn HI Survey (Flöer et al., 2010) and the AOFlagger pipeline (Offringa et al., 2010b).

The LOFAR imaging pipeline (Heald et al., 2010) consists of automated steps to (1) flag interference contaminated data; (2) reduce the size of the observation by averaging in time or frequency; (3) calibrate the data; (4) deconvolve the data data with the point spread function; and (5) image the observation. Since flagging is the first step after correlation, flagging is normally performed on the highest resolution, and its performance is an important issue. Moreover, the pipeline needs to be robust and accurate.

One of the LOFAR key science project depending on a robust and accurate pipeline is the LOFAR epoch of reionization (EoR) project (Labropoulos, 2010; Jelić, 2010), a very ambitious project with high demands on calibration, sensitivity and noise behaviour. At the time of writing, the first EoR LOFAR data has been acquired (de Bruyn et al., 2011) and will be be used to test the EoR pipeline. Because this project observes the same fields in the sky repeatedly, it simultaneously allows effective analyses of the radio environment and its variability.

In this chapter we will explain how the automated LOFAR pipeline is formed from the methods discussed in Chapter 2.

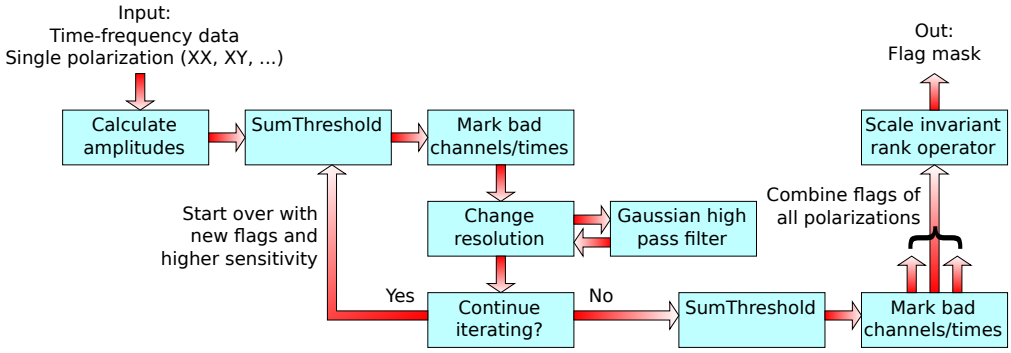


Figure 3.1: *Overview of the RFI flagging strategy*

3.1 Input data

For LOFAR, a typical resolution is one second time integration and 0.8 kHz frequency resolution. LOFAR can observe in two bands: the 10-80 MHz low band and the 110-240 MHz high band, which are observed by physically different antennae. It allows observing of 48 MHz of bandwidth concurrently. This bandwidth is currently limited by the transfer of the data from stations to correlator. At a later time, LOFAR will allow different quantization modes on station level, allowing even higher bandwidths. The 48 MHz is split into 244 sub-bands of 256 channels. Therefore, in this common mode of operation, the total output of the correlator when using 50 stations is 244 sub-bands $\times$ 256 channels $\times$ 4 polarizations $\times$ 1 Hz $\times \left(\frac{1}{2} \times 51 \times 50\right)$ baselines $\approx$ 319 million visibilities per second. Since a visibility consists of a real and imaginary floating point number of four bytes each, the total output rate of the correlator can reach 2.5 GiB/s. Although the data processing will be done on large off-line clusters, this data rate imposes high constraints on the efficiency of the flagger.

The flagger is executed on the amplitude information of one polarization of a single sub-band of a baseline. If speed is essential, the algorithm can be executed once on the Stokes-I values. Otherwise, if accuracy is more important than speed, the algorithm can be executed on the individual XX and YY or LL and RR polarizations, or on all polarizations individually. We do see some RFI that manifests in only one of the polarizations, or rotates through the polarizations, and some advantage is therefore seen when flagging all polarizations individually.

3.2 Processing steps

An overview of the flow of execution is given in Figure 3.1. We will describe each step in the following subsections.

3.2.1 Iterative approach

A part of the pipeline is iterated a few times, depicted in Figure 3.1 by the "Continue iterating" block. This is necessary for finding low-level RFI, as will be explained in the thresholding paragraph, §3.2.2. Iterations, however, are costly in terms of speed, and should be kept to a minimum. To do so, the fit should converge quickly. We do this by entirely ignoring channels and time steps in the first surface fit that superficially look bad, yet might only have been partially uncontaminated. The extra information that might have been added if the uncontaminated part of the channel or time step was added does not change the fit much, and therefore is not slowing down the convergence.

It was determined that performing the fit two times is enough for a stable, accurate fit. This is true for all data that was tested, in special for both WSRT and LOFAR data, and for both clean bands and strongly contaminated bands.

3.2.2 The `SumThreshold` method

The `SumThreshold` method detects series of samples with higher values than expected. In the previous study of Offringa et al. (2010a), the `SumThreshold` was introduced and was shown to produce the highest accuracy of current post-correlation RFI detection algorithms. We refer to §2.2.6 for detailed information about the `SumThreshold` method.

`SumThreshold` is performed in each iteration once, before the surface fit, in order to ignore RFI when fitting. It is performed one last time when the surface fit is expected to have been converged, to establish the actual flags. To increase the stability of the strategy, the sensitivity of the `SumThreshold` method starts low, i.e., it finds only the strongest RFI, and is exponentially increased each time it is executed.

3.2.3 Channel and time selection

After `SumThreshold` has found the contaminated samples, we observe especially after the first iteration, that some channels and time scans have not been flagged, even though they are mostly contaminated. As explained in §3.2.1, this might slow down convergence, which is why a second step was implemented in order to completely — hence inaccurately and quickly — flag these channels and time steps before smoothing.

In order to detect problematic channels and time steps, the values are compared based on their root mean square (RMS) values. The RMS series are Gaussian smoothed and if the difference exceeds 3.5 times the standard deviation of the sequence of differences, they are flagged completely. Optionally, this selection can be executed again as the last step in the algorithm.

3.2.4 Smoothing / sharpening

The signal of interest is assumed to be smooth, and a sharpening operation is executed to subtract fringes caused by strong sources. This is done to increase the accuracy. Several sharpening strategies and surface fitting methods have been tested, and all sliding window methods show similarly good results in terms of accuracy. In non-sliding window approaches such as the tiled dimensional-independent polynomial fit described in Winkel et al. (2006), we have observed instability near the borders of the fixed windows.

A Gaussian kernel was found to produce the best average between speed, accuracy and stability. Because the signal is estimated and subtracted by convolution with a Gaussian kernel, this is essentially a Gaussian high-pass filter. The accuracy is not significantly different from other sliding window fitting strategies, such as a dimensional-independent polynomial fit applied on a sliding window.

Since the fit is, relative to the other operations, a time-consuming operation, the input time-frequency matrix is rescaled before fitting. The time dimension and frequency dimensions are three times reduced before fitting, and the fitted Gaussians are interpolated to restore the original scale. No significant change in accuracy was observed, which underlines that the quality of the fit is, up to some point, not a crucial aspect of accurate detection.

We have implemented the Gaussian filter with a direct convolution of a truncated inverted Gaussian. Because the width of the Gaussian is in our case generally small — the σ parameter is on the order of 5–10 samples — a direct convolution is faster than a multiplication in the Fourier domain. A recursive Gaussian filter can however increase the performance somewhat. Two different methods for designing recursive Gaussian filters are described by Deriche (1992) and van Vliet et al. (1998). While the direct convolution with a finite impulse response filter is faster for Gaussians with $\sigma < 3$, for the ranges $3 \leq \sigma < 32$ and $32 \leq \sigma$, the recursive filters of Deriche and van Vliet are recommended respectively (Halen, 2006).

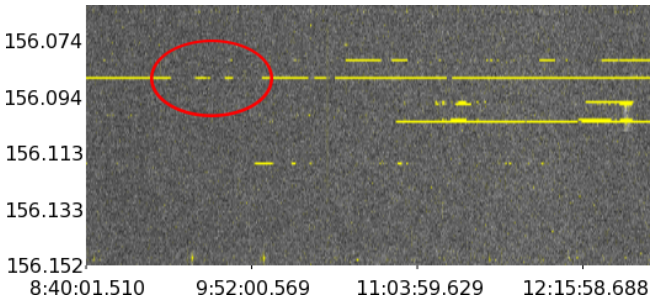
3.2.5 The scale-invariant rank operator

It may be desirable to flag samples that are up to a few channels away from strong, continuous RFI. Thresholding does not flag these samples, if they are not significantly different in amplitude. Likewise, it may be desirable to flag more of a partially flagged channel, because a continuous transmitter might be recorded at different amplitudes, either because of different propagation of the signal, because of the transmitter moving in respect with the beam or because of a transmitter's intrinsically changing strength, and this might cause the received RFI not to trigger the threshold in some samples. To overcome this problem, we enlarge the flag mask after the apparent RFI has been flagged by the iterative procedure.

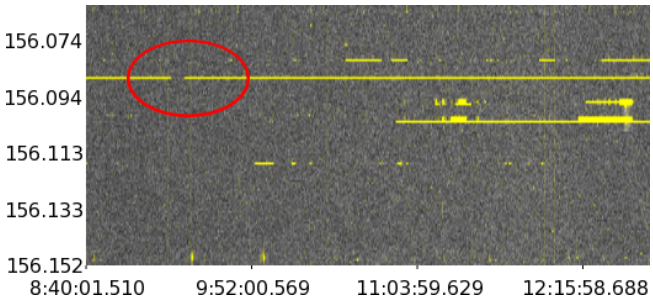
A typical approach in this problem is to perform a morphological dilation operation on the flag mask. For example, a dilation with a square mask of size $N \times N$ would enlarge each flag to a square of $N \times N$. Every sample, that has an orthogonal distance smaller than N samples from a flagged sample, would be flagged in this case. Although this technique is advantageous for its simplicity and establishment in the field of mathematical morphology, using this technique for the described purpose has the disadvantage of being inaccurate: it will typically flag too many samples when only a few samples are flagged in some area, while too few samples will be flagged when a channel or time step is almost completely flagged.

To correct for these problems, we have used the morphological scale-invariant rank (SIR) operator which mask size is related to the one dimensional flag density: the dilation mask is larger in dense areas and smaller in sparse areas, in respect to either the one dimensional time domain or frequency domain.

Consider an orthogonal slice $\Omega_d(x)$ through the flag mask as defined in §3.2.2. The following



(a) Time-frequency plot without the SIR operator



(b) Time-frequency plot after the SIR operator

Figure 3.2: *The result of the SIR operator with $\eta = 0.1$: the flags in panel (a) are established by the* `SumThreshold` *method and dilated based on the flag density. The result is shown in panel (b). Noticeable differences are the small gaps in orthogonal lines that have been filled by the dilation, such as the area within the red ellipse. While this diagram displays over 6000 time steps, the algorithm also fills many invisible small holes: its behaviour is scale invariant.*

decision rule is introduced:

$$\Omega'_d(x) = \begin{cases} 0 & \text{if } \exists Y_1 \leq x,\ \exists Y_2 > x : \sum_{y=Y_1}^{Y_2-1} \Omega_d(y) \leq \eta\,(Y_2 - Y_1) \\ 1 & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\eta \in [0, 1]$ is the density ratio threshold. In words, this rule flags the samples that are in any constructable area $[Y_1; Y_2\rangle$ with an unflagged sample ratio less or equal than η. Specifically, $\Omega'(x) = 0$ for all x if $\eta = 1$, while $\Omega'(x) = \Omega(x)$ for $\eta = 0$. Furthermore, since any element x with $\Omega(x) = 0$ will be in the single element area containing only itself, $\Omega(x) = 0 \implies \Omega'(x) = 0$. Consequently, the number of flags is increasing. Although a strict implementation of (3.1) will take $\mathcal{O}\left(n^2\right)$ operations for n samples in the orthogonal slice $\Omega_d(x)$, by putting extra constraints on Y_1 and Y_2, an $\mathcal{O}\left(n \log n\right)$ implementation is possible without much loss of its accuracy. After having used the $\mathcal{O}\left(n \log n\right)$ implementation for half a year, an *exact* and fast algorithm was found with $\mathcal{O}\left(n\right)$ time complexity (Offringa et al., 2012b), as described in Section 2.4. This was used thereafter. The remainder of this chapter assumes the $\mathcal{O}\left(n \log n\right)$ algorithm is used. Figure 3.2 shows the result of the operator on actual data.

Table 3.1: *Computational requirements of the RFI pipeline*

Step	F/smp[1]	Count	Total F/smp[1]
Calculating amplitudes	4	1	4
`SumThreshold`	20	3	60
Time/frequency selection	2	3	6
Change resolution	4	2	8
Surface fit	50	2	100
SIR operator ($\mathcal{O}(N \log N)$ version)	100[2]	1	100
Total			278

3.3 Computational requirements

Table 3.1 shows an estimate of the required floating point operations per sample for each individual step. The total number of operations required is on the order of 300 floating point operations (FLOP) per sample. In a typical full LOFAR observation, the correlator will output 4 polarizations $\times$ 256 channels/sub-band $\times$ 248 sub-bands $\times$ $\frac{50^2}{2}$ baselines $\times$ 1 sample/second $\approx$ 0.3 gigasamples per second, yielding a computational requirement of $\sim$0.1 TFLOP/s in the best flagging mode.

Although this is only a small fraction of the required computations for correlation, some simplifications can be made to lower the computational requirements. Techniques to improve the computational performance include: flagging on Stokes-I values; using a larger resizing factor before fitting; using a smaller window size; and determining the cross-correlation flag masks using auto-correlations.

The LOFAR flagging pipeline will be run on an off-line computing cluster. The flagging pipeline is parallellized by running each sub-band on a different computational node, and the flagging of the individual sub-bands is executed by a multi-threaded implementation. Concluding from the interpolation of the performance of the current implementation of the pipeline, which achieves processing 27 stations in a quarter of the observing time with its most computational expensive flagging strategy, real-time performance can be realised in a full 50 station LOFAR.

3.4 Input/output requirements

Processing baseline by baseline in a pipeline has implications for the software architecture of the observatory: since baselines are correlated simultaneously, the observed visibilities have to be written to disk before running the RFI pipeline, which is inefficient. After finishing observing, the flagging pipeline can read the data in the required order. However, flagging is normally followed by tasks such as calibration and source subtraction. These tasks expect time-sorted data, thereby requiring a second read of the data in its previously observed order. Since the architecture of LOFAR allows this flow of processing, and because of the advantages of baseline by baseline flagging in terms of accuracy and computational speed, the input/output-overhead caused by this

[1] Floating point operations per sample

[2] These are actually integer operations, since this step uses the masks.

deficiency is ignorable. This however might become a serious issue in even larger telescopes such as the SKA.

3.5 Flagging results

The implementation[3] of the algorithm was tested on several LOFAR observations. At the time of writing, 27 of the approximately 50 total LOFAR stations are ready. Flagging a single subband of a 6 hour observation with the 27 stations takes 90 minutes on a single cluster node. This implies real-time flagging speed for the full 50 station LOFAR that will produce four times more data. All RFI that can be found by visual inspection is typically flagged, thereby outperforming simpler methods such as a median absolute deviating (MAD) thresholding filter in both accuracy and speed. An example result can be found in Figure 3.3.

3.5.1 The flagging strategy

The AOFlagger is the recommended way of flagging LOFAR data (Pizzo, 2012), because it was found to be both the most accurate and the fastest flagging algorithm available. Figure 3.4 shows a comparison between the AOFlagger and the median absolute deviation (MAD) flagger. Many flaggers implement a method that is similar to the MAD flagger, i.e., a strategy that is based on the median of a sliding window and single sample thresholds. Examples are the AIPS `FLGIT` task and the `PIEFLAG` program (Middelberg, 2006). An important difference with these and the iterative AOFlagger algorithm is the combinatorial threshold `SumThreshold` step of the AOFlagger.

In some cases, the algorithms finds RFI which is invisible by eye on full scale time-frequency diagrams, but becomes only apparent when zooming in on the data and integrating certain cuts of the data cube. In the band shown in figure 3.3 an interferer is visible at approximately 156.03 MHz. Although it is visible as a small bump in the time integrated spectrum in Figure 3.3e, it is not apparent in the time frequency plot of Figure 3.3a. Nevertheless, the algorithm finds the samples that are contaminated by the interferer, and the particular bump at 156.03 MHz in Figure 3.3e is flattened.

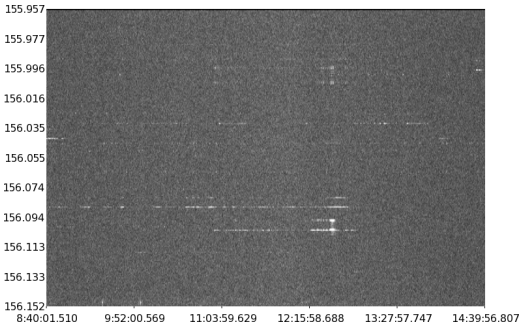
On the other hand, if an interferer has a smooth time-frequency profile, it will be mistaken for astronomical data and will not be flagged. In these situations it might help to subtract a rough model for the celestial signal and increase the flagger's sensitivity.

3.6 LOFAR RFI environment: preliminary results

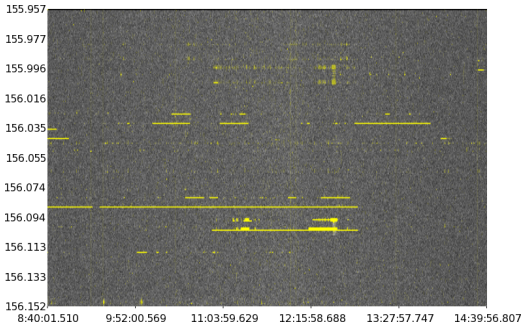
LOFAR breaks the tradition of building telescopes in sparsely populated areas, with its core being installed in the North-East of the Netherlands. Although the core is in a nature reserve, and therefore in a sparser populated part of the Netherlands, all the stations are relatively close to farms, roads and some nearby municipalities. Now that LOFAR is half-way ready and performing representable observations, we can start to evaluate the dynamic radio environment.

The first results of RFI mitigation show several promising characteristics of the LOFAR site. First of all, hardly any broadband RFI is observed. If observed, it is typically caused by electrical

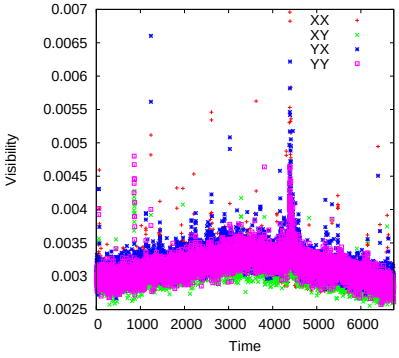
[3]The software implementation of the presented RFI pipeline has been made publicly available and can be downloaded from the following location: `http://www.astro.rug.nl/rfi-software/`



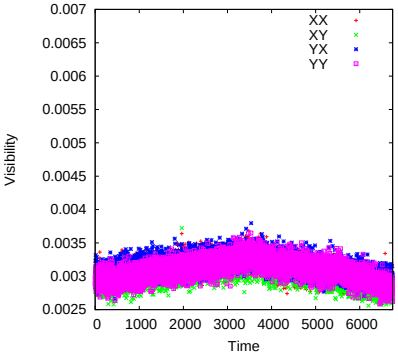
(a) Time-frequency plot before flagging



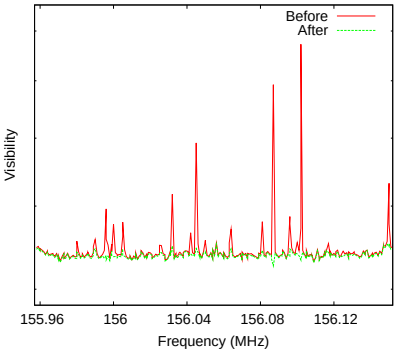
(b) Time-frequency plot after flagging



(c) Amplitude plot before flagging

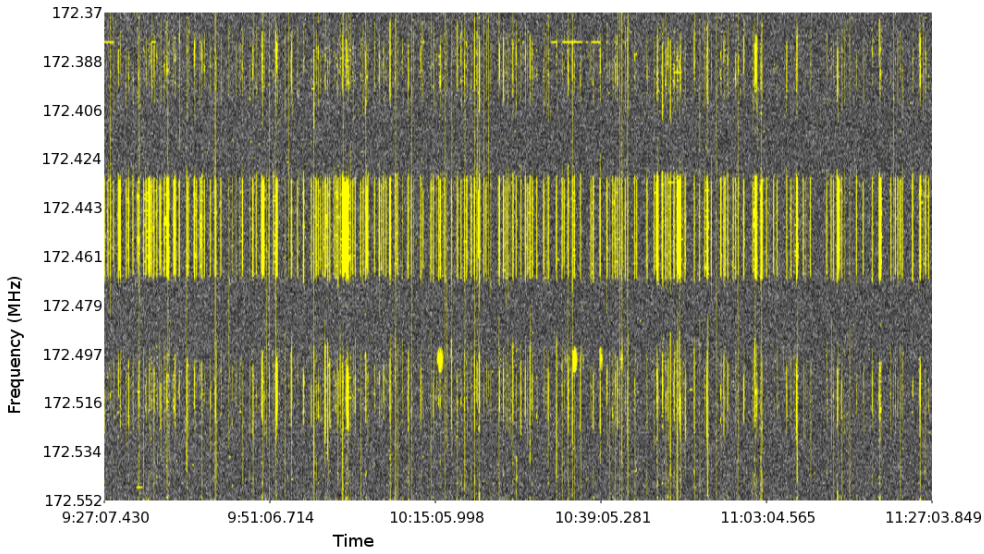


(d) Amplitude plot after flagging

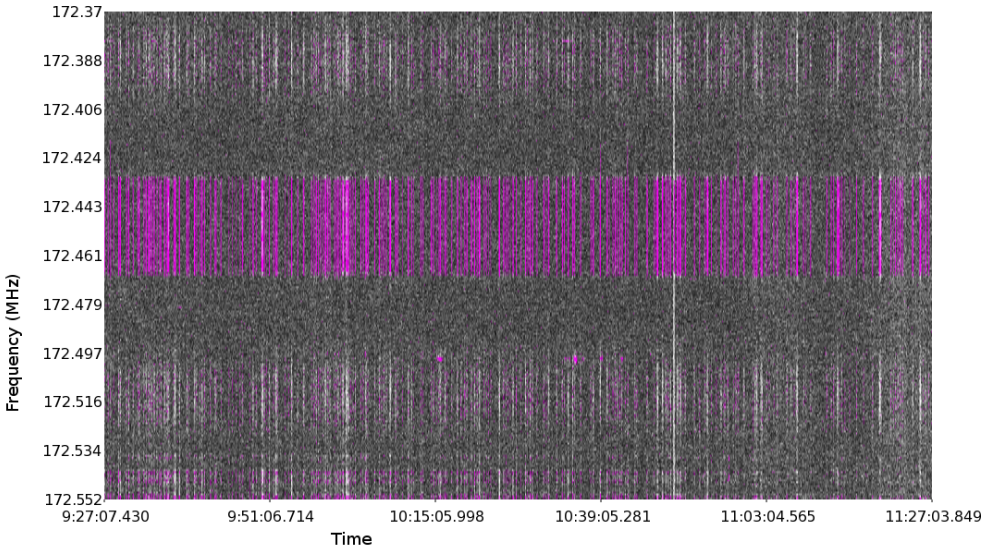


(e) Power spectrum

Figure 3.3: *Flagging results of the 6 hour LOFAR observation L2010_07096 of April 24, 2010. All plots show the same randomly chosen sub-band around 156 MHz for a 1.5-km baseline (CS302_HBA1 × CS005_HBA0) with three second integration time. The flagging pipeline was run with its default settings, and 1.8% of the data is flagged. As can be seen from panels (a), (c) and (e), this sub-band contains relatively many interfering transmitters, yet all of them are relatively weak. The panels (b), (d) and (e) show the cleaned band after flagging.*



(a) Result of the AOFlagger



(b) Result of the MAD flagger

Figure 3.4: *Comparison of the AOFlagger and the median absolute deviation (MAD) flagger on a badly contaminated LOFAR sub-band around 172 MHz. The plots show two hours of data. Both methods have been optimized to flag this particular baseline as accurate as possible. Evidently, the MAD flagger misses a lot of the RFI. Increasing its sensitivity helps very little, while this would increase the false positives considerably.*

fences, lightning, power cables, hardware in situ, cars and trains. It can be concluded that the site is sufficiently remote and hardware on site is sufficiently shielded to prevent these interferences. Only one of the stations is close to an electrical fence that surrounds a farming meadow, causing broadband spikes every two or three seconds. The flagging pipeline flags 40% of the data in this station, and this station is therefore currently not useful. Options include negotiation with the farmer to switch off the electrical fence during observations or implementing an RFI nulling method in the station that nulls spikes on a high time resolution.

A second class of interferers are constant transmitters at a fixed frequency, such as FM radio. The FM range lies between the physically separated low and high bands. Transmitters in this range are therefore effectively blocked by the bandpass filters. Other constant sources that do transmit within the observing frequency often occupy only one or a few 0.8-kHz channels, which, after the pipeline has flagged these transmitters, cause only a minimal amount of data loss. While many of the sub-bands of 256-channels are completely clean of such constant transmitters, others have a few of such transmitters, such as the one shown in Figure 3.3.

A third class of interferers are transient sources with variable frequency. These occur mostly at random and their exact origin is often unknown. Some of these can be caused by moving objects, such as meteors or aeroplanes, that reflect a distant signal for a short period.

Both the high-band antennae (HBA) and the low-band antennae (LBA) observations of the EoR project show a very promising radio environment for LOFAR. Considering all classes of interferers, typical observations with representable stations show only a few percent of data loss due to interference. We have not encountered problematic RFI in observations after flagging, which confirms the performance and stability of the flagger.

In some observations it is, at this point, still required to do some manual flagging due to other reasons than RFI. The most common reasons are issues with a whole station which cause them to produce erroneous data, e.g., because the station was not tracking correctly. Validating stations will be performed fully automatic in the future. Some false flagging is seen in short periods of strong atmospheric scintillation. During such periods, the amplitude can change very rapidly in time, such that the flagger marks these periods as broadband RFI. This occurs however very rarely ($<$1%). Such periods are very interesting to investigate ionospheric phase stability, however these are unfortunately lost after further averaging, which is done by default to reduce the size of the observation and make room for further observations. A solution is to change the default flagger parameters to be insensitive for broadband RFI.

Work is being done on improving calibration, sky models and beam models, which are currently the limiting factors in getting to high dynamic ranges. The RFI monitoring observations (Chapter 5 and 6) will provide further knowledge on the spatial distribution of RFI and the difference between day and night. We might also be able to estimate how much RFI is being missed by the flagger, and estimate the influence of missed RFI on further data extraction. In the end, we might be able to inject such artefacts in the EoR testing pipeline, to be able to test our signal extraction in the presence of RFI.

3.7 Conclusion and discussion

Radio astronomy is entering a new era with futuristic observatories such as LOFAR and the SKA. In this article we have presented a flagging technique that has shown the ability to operate accurately and efficiently on the LOFAR observations. Therefore, this technique is also a good basis

for future observatories.

Because the computational costs of the RFI pipeline are only a fraction of the correlation costs, efficiently ordering the data before presentation to an RFI algorithm is the largest challenge, rather than optimising the computational costs. The pipeline also stipulates the importance of flexibility in an observatories' architecture, which adds freedom to design decisions. The LOFAR architecture allows more sophisticated variations of interference stratagies that include RFI mitigation at station level and different pipelines based on the observation mode. With the example of a complicated pipeline as described in this paper, it can be concluded that other algorithms such as transient detection and other pattern recognition techniques can be implemented in a similar manner in the pipeline.

Both the software and hardware of LOFAR are still under construction at the time of writing. The first observations of LOFAR nevertheless show very good prospects for the telescope, with only a few percent lost data due to interferers and, highly important, neither broadband nor in situ interference is commonly seen. The next step in RFI mitigation is to produce and analyse images on a maximum dynamic range, in order to analyse the effects of possible weak RFI that is undetectable in post-correlated time frequency domains. Prevention of new transmitters remains very important, and establishment of a radio-quiet zone, especially around the core, is recommended.

In order to improve data quality further, pre-correlation techniques might be added at station level or during correlation. An interesting improvement to the robustness of a correlator might be to execute the `SumThreshold` method prior to correlation. Considering the accuracy gain of the `SumThreshold` compared to normal thresholding, and considering the correspondence of RFI on small and large timescales, implementing this pre-correlation method on the highest time resolution data might improve blanking accuracy further.

Chapter 4

Filter techniques

Based on:
"*Post-correlation filtering techniques for off-axis source and RFI removal*"
(Offringa et al., 2012, MNRAS, 422, 563–580)

For several decades, it has been a challenge to increase the dynamic range of images produced by interferometric radio telescopes. The raw sensitivity improvements and advanced understanding of calibration errors have pushed the limits on the dynamic range of modern telescopes to unprecedented levels (Smirnov, 2011). The final dynamic range is constrained by the celestial field being observed, the efficiency of the telescope's hardware and the time spent observing. However, this theoretical dynamic range is limited further by imprecise models of instrumental effects and celestial sources used in the data reduction process, as well as by the quality of the radio environment.

The noise level in the final result of an observation can be set by several phenomena. In the ideal case, the noise level equals the thermal sky noise level, and the detection of sources or other features is limited by this noise level only. An image can also be limited by confusion noise when it does not provide enough resolution to distinguish sources. Sidelobes provide a third type of noise. This noise is generated by the point spread function (PSF) of the instrument, that convolves strong sources that are in or outside the field of interest. Finally, radio-frequency interference (RFI) can add additional noise to the final result of an observation. In this paper, we will aim at suppressing noise coming from RFI and sidelobe noise coming from off-axis sources, using similar techniques based on fringe theory.

4.1 Introduction

Because we address two problems at once, we will introduce both problems individually. In the following subsection, we will introduce the problem of RFI and describe current techniques to deal with it. Thereafter, we will introduce the concerns of off-axis sources and approaches to deal with those as well.

4.1.1 Radio-frequency interference

While technical advances gave rise to better telescopes, different technical advances have ironically decreased the quality of the radio environment for radio astronomy. A potential problem that limits the effective dynamic range of modern telescopes such as LOFAR, the WSRT, the Giant Metrewave Radio Telescope (GMRT), the Australia Telescope Compact Array (ATCA) and the EVLA, is radio-frequency interference (RFI). Fortunately, practically all RFI interferes within a limited amount of time or frequency channels, and can be flagged automatically in post-correlation. In Offringa et al. (2010a), the SumThreshold algorithm is described and is proven to be very accurate for that purpose. Further implementation of the method into the LOFAR pipeline has shown excellent results (Offringa et al., 2010b).

Although reasonably strong temporal and spectral RFI can successfully be removed by flagging, it is not always a satisfactory solution. Sporadic continuous broad-band RFI for example poses a potential problem, since this type of RFI can not be removed by flagging. Doing so might affect considerable parts of the observation, potentially throwing away too much of the data. Athreya (2009) has shown that the GMRT suffers from this type of RFI at low frequencies, for example caused by high-voltage power lines. Athreya (2009) describes a method to remove this kind of RFI based on fringe fitting of RFI. This approach has been recently implemented in AIPS[1] (Kogan and Owen, 2010). This method will be analysed in §4.2. Most other telescopes do not report such severe broad-band RFI: LOFAR, although build in a populated area, shows very little of this kind of RFI in the currently finished stations (Offringa and de Bruyn, 2011) and (E)VLA interference reports also mention spectral RFI affecting a few channels, but no broad-band RFI, although low frequency causes more problems (Chandler and Perley, 2010, §4.6). Nevertheless, when approaching the thermal noise on low frequencies, such as LOFAR will do in the future, faint RFI might show up. The fringe fitting method is not so well applicable in these cases, because such RFI will be below the noise. By removing a spatial frequency component from (white) noise dominated data, a component from the noise will be removed instead of removing actual RFI. Work has been done to apply post-correlation RFI removal techniques for the (E)VLA, by ways of calibrating and removing the RFI source (Lane et al., 2005), but this method is tedious and requires the RFI to be reasonably stable.

Another solution for removing continuous RFI is spatial filtering by eigenvalue decomposition (Leshem et al., 2000; Smolders and Hampson, 2002; Ellingson and Hampson, 2002), which disentangles the contribution of sources from different directions, and subsequently removes the contributions from the direction of interference. Recently, this was implemented for the Parkes multibeam receiver (Kocz et al., 2010). However, the requirement of specialized hardware and/or having to configure the filter before correlation is a major disadvantage of spatial filtering techniques, in the context of interferometers. The latter requires the configuration to be fixed before the observation in most cases. This makes it hard to react to unanticipated RFI, and impossible to change the filter after observing if the filter has not worked correctly. RFI is often not stable enough to be removed during post-correlation processing.

Another technique for removing sporadic continuous RFI has been introduced in Pen et al. (2009), which decomposes the time frequency data with a singular value decomposition (SVD). This method however was shown in Offringa et al. (2010a) to potentially alter the astronomical data, making the method less attractive to use for data reduction without further research. In

[1] AIPS is the Astronomical Image Processing System (`http://aips.nrao.edu/`)

Briggs et al. (2000), the RFI is subtracted from the data after correlation by the use of a reference signal. Unfortunately, such a reference signal is not always available or practical to implement.

4.1.2 Off-axis sources

Signals from off-axis sources received in the sidelobes, like RFI, decrease the dynamic range of observations, or might even cause calibration to fail. New wide-field telescopes such as LOFAR see a large area of the full sky, and always have a few strong sources in their sidelobes. Examples of such sources are Cassiopeia A, Cygnus A and the Sun. These sources are often not of interest, but have to be removed accurately.

A common method to deal with off-axis sources is peeling (Noordam, 2004; Intema et al., 2009). Peeling is iterative, and changes the phase centre towards the source, optionally averages in time and frequency to suppress other sources, and self-calibrates and subtracts the source. This method has shown good results, but is very computational intensive — too intensive to use by default on high-resolution telescopes such as LOFAR. Demixed peeling is a variation on normal peeling, that is currently being tested for LOFAR observations. However, early results show similar computational requirements when the same removal quality is required (Jeffs et al., 2006).

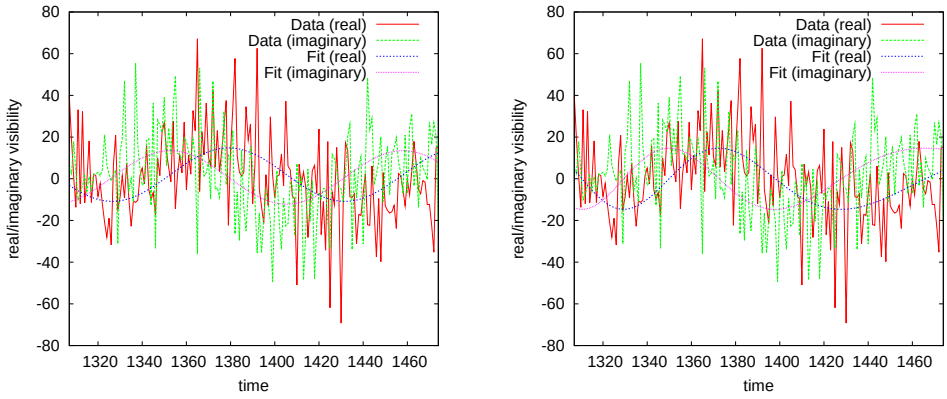
Finally, in Parsons and Backer (2009) a delay-delay rate (DDR) filter is proposed that disentangles the flux contribution into the different sky facets they originate from. The DDR-filter was used by Parsons & Backer for first order calibration, but the idea of such a filter is also attractive for application in a later stage and over longer timescales, because the filter can be applied on post-correlated data without additional hardware. It is however unclear how accurate the filter will be for off-axis source removal. We will propose related filters, while trying to increase its application and accuracy.

4.1.3 Outline

In this paper we will describe and analyse new methods for filtering both RFI and off-axis sources, with the ultimate goal of reaching lower noise levels. We will start by analysing Athreya's fringe fitting method in §4.2 and describe why it is insufficient for e.g. LOFAR observations. In §4.3, several new methods will be introduced and analysed with the help of simulations. We will test our filtering approaches in §4.4 on a WSRT dataset at a frequency of about 140 MHz of the field centred on the radio galaxy B1834+62 (Schoenmakers et al., 2000). At this low frequency, the WSRT is sensitive to very bright sources like Cygnus A and Cassiopeia A (de Bruyn and Bernardi, 2009), which despite their large angular distance are not sufficiently attenuated by the primary beam. They therefore generate intense spurious sidelobes across the target field of view. We will discuss the results in §4.5, where we will also discuss how time or frequency averaging and gridding may effect off-axis sources or RFI. Finally, we will draw conclusions based on our findings in §4.6.

4.2 Analysis of the fringe filtering method

Athreya (2009) describes how geometrically stationary RFI can be removed from an observation by fitting out a sinusoid with a frequency opposite to the natural fringe rate. A stationary earth-bound RFI source receives a fringe rate opposite to the applied fringe stopping rate. Therefore,



(a) Fit with constant fringe rate, amplitude found = 12.8

(b) Using fringe count, amplitude found = 14.7

Figure 4.1: *Comparison of fitting methods using simulated data: the original amplitude of the source is* 16*. Only the shown data is used for the fit. Using a constant fringe speed (left panel) over this range produces a somewhat less accurate fit compared to using the fringe count for each sample in the fit (right panel). The x-axis is in time steps of 15 seconds from the start of the (simulated) observation. At time step 1570, the simulated baseline is orthogonal to the direction linking the target source and the phase centre and ν_F=0. Hence, the fringe speed changes significantly over the displayed time range, which can be seen by the somewhat elongated fringes near the right.*

one can estimate its contribution. The natural fringe rate is given by:

$$\begin{aligned}\nu_F(t) &= \frac{dw(t)}{dt} \\ &= -\omega_E u(t) \cos\delta, \end{aligned} \tag{4.1}$$

with t the sidereal time, $\omega_E = 1$ rotation/day, the rotation speed of the earth, $u(t)$ the component representing the standard u position of the baseline in the uv-plane, $w(t)$ the standard w-component representing the applied phase delay and δ the declination of the phase centre. When a baseline is orthogonal to the direction of the phase centre, $\nu_F(t)$ is zero. A stationary source of RFI contributes to a correlation in the form of the complex function

$$\text{RFI}(t) = \mathcal{A}e^{-i\nu_F t}, \tag{4.2}$$

with $\mathcal{A}$ the complex amplitude of the RFI at time t. The 2π term is absorbed in ν_F, such that its value is in radians/time unit. This amplitude is initially assumed to be constant over some period $[t_0, t_E]$, and ν_F is assumed not to change over this time interval. It is then possible to estimate $\mathcal{A}$ by performing a least-squares fit between the complex function $V(t)$, representing the observed visibilities, and the RFI signal by minimizing the error function

$$\epsilon(\mathcal{A}) = \int_{t_0}^{t_E} \left(\mathcal{A}e^{-i\nu_F t} - V(t)\right)^2 dt. \tag{4.3}$$

Minimization of $\epsilon(\mathcal{A})$ results in

$$\mathcal{A} = \int_{t_0}^{t_E} V(t) e^{i\nu_F t} dt, \tag{4.4}$$

which corresponds to $\mathcal{A} = \mathcal{F}(\nu_F)$, the frequency component ν_F of the Fourier transform $\mathcal{F}$ of V over the time interval. Therefore, removing a Fourier component of a signal can be implemented as a standard frequency filter. Equation (4.2) corresponds to a single component of the delay-rate (DR) transform, creating a symmetry with the DDR filter proposed in Parsons and Backer (2009). An example of the application of Equation (4.4) on simulated data is given in Fig. 4.1a. The two plots show the result of fitting a sinusoidal function to simulated data. We simulated a WSRT interferometer, correlating antennae RT0 and RT5: a 720m baseline. A single channel is simulated with a frequency of 147 MHz. The simulated observation has eight sources, seven of which are faint and in the primary beam, while the last source simulates an interfering source that is four times stronger. This off-axis source generates a visibility amplitude of 16 and is a 40° from the phase centre, hence far from the other sources.

Since ν_F changes slowly with time, Equation (4.4) will become inaccurate when increasing the time interval. Additionally, it can not be calculated near $\nu_F = 0$. By observing that the number of wavelengths of delay caused by the geometrical delay corresponds to the number of rotations applied on the visibilities, we can replace $\nu_F t$ by $w(t) - w(t_0)$, where w is the applied phase delay in radians/time unit as function of time. As $w(t_0)$ causes a constant phase shift, it can be absorbed in $\mathcal{A}$. By substituting $\nu_F t$ with $w(t)$ in Equation (4.4), we get a more accurate solution for $\mathcal{A}$:

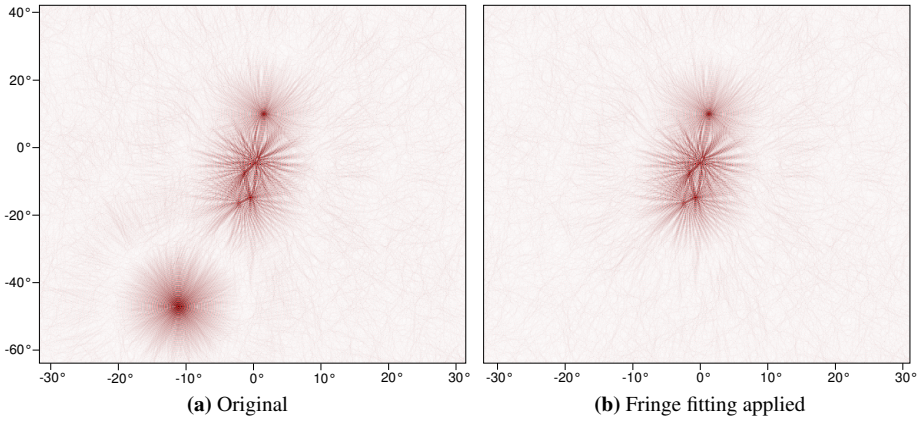
$$\mathcal{A} = \int_{t_0}^{t_E} V(t) e^{iw(t)} dt. \tag{4.5}$$

An example of such a fit is given in Fig. 4.1b. As long as the amplitude of the RFI source remains constant, this allows successful removal of the source when $\nu_F \gg 0$. As is visualized by Fig. 4.2, it removes the strong source in the example without unwanted side effects on the area of interest.

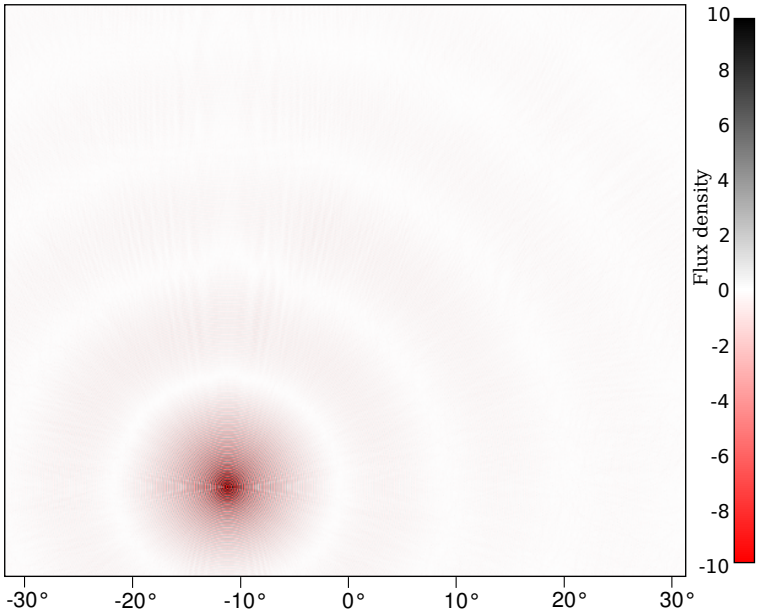
4.2.1 Removing variable RFI

With the algorithm presented by Athreya, the received strength of the RFI source is not only assumed to be different for different baselines, but also in time. Since the beam rarely follows the RFI source, it is likely that the gain towards the RFI source will change. Athreya proposes tiling of the data, making separate fits on each tile, where each tile is approximately the size of a fringe. However, tiling the data and performing fits on each tile causes instability near the borders of the tiles.

A more accurate way is to perform individual fits for each sample, sliding a kernel of weights over the data that are used to perform the fit. Two trivial suggestions for a weighting function are the rectangular function and the Gaussian function. A rectangular function would result in a sliding window method, which has implementational advantages. However, a rectangular function produces a sinc response in delay space. Therefore, the fit will be affected by any other frequency in the data set that corresponds to non-zero values in the sinc function, which undesirably would remove part of the signal of interest. A Gaussian kernel would localize the frequency response somewhat better. A larger kernel or tile size would decrease the frequency response to other

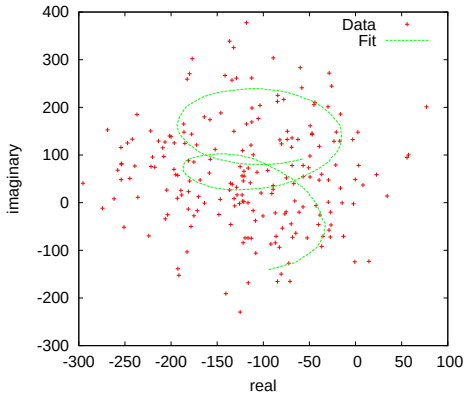


(a) Original **(b)** Fringe fitting applied

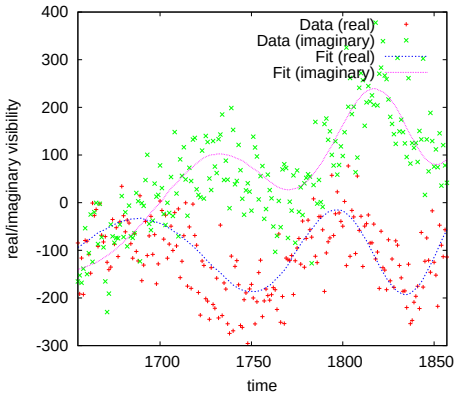


(c) Difference

Figure 4.2: *These images show the application of a fringe filter that takes out a hypothetical source with a constant amplitude (Equation* (4.5)*). The same 720m WSRT baseline and set-up as in Fig. 4.1 was simulated and imaged without deconvolving. The image in the left panel is the result of imaging without any filtering. The middle panel shows the result after application of the filter, while the right image shows the difference. The filter removes the source up to the sidelobe confusion noise of the other sources, which is over three orders of magnitude. The residual shows that it does not affect the sources of interest, again up to at least three orders of magnitude. This simulated situation is only hypothetical, since it is unlikely that the received power of distant sources remains constant.*



(a) Fit visualization of the complex plane



(b) Fit visualization of the individual complex components over time

Figure 4.3: *Visualization of the sliding window fringe filter applied on data of a simulated baseline. In the complex plane (top panel), such a fit produces spirals. Since the mean of the sliding window was added to the fit in this figure, the difference between the fit and the moving centre of the ellipses is the actual value that will be subtracted from the data by the filter. A window size of two fringes was used. Only a small part of the baseline track is shown here.*

frequencies, but in order to remove the RFI it would be required that the received gain of the RFI changes less quickly.

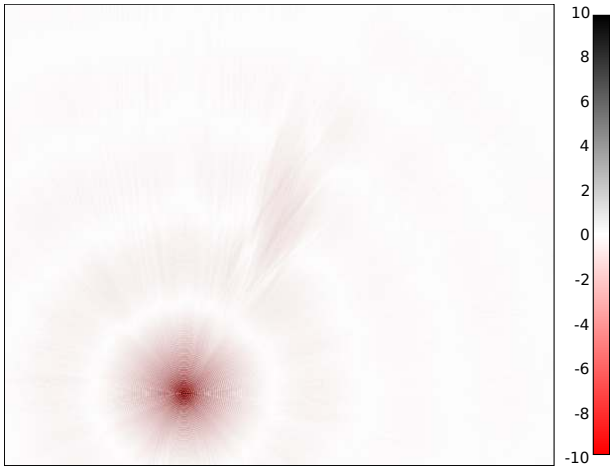
Allowing the amplitude to change in time creates spirals in the complex plane, as is visualized in Fig. 4.3a. This kind of fitting has recently been implemented in the AIPS astronomical package as described by Kogan and Owen (2010).

4.2.2 Generalization of the fringe fitting method

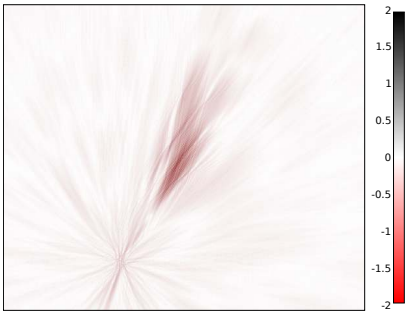
Up to now, the use of the method has been limited to the removal of a single (RFI) source that behaves like a point source at the celestial pole. It is common practice to peel and/or calibrate for sources that are outside the area of interest, because they need to be taken out carefully in order to avoid additional sidelobe confusion noise. In such a case, the off-axis source is similar to static RFI: the source itself is not of interest, but has to be taken out for calibration and imaging the field accurately. For this purpose, the fringe fitting method can be generalized to remove any point source. This requires a small change to Equation (4.5), which now becomes:

$$\mathcal{A} = \int_{t_0}^{t_E} V(t) e^{i(w(t) - w_S(t))} dt. \tag{4.6}$$

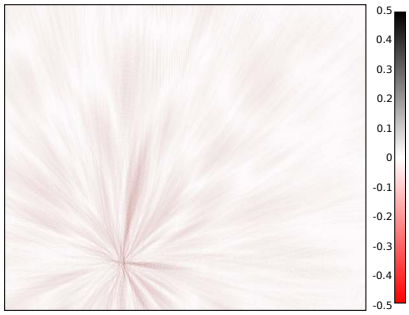
Here, $w(t)$ is the standard w-component in the uvw domain as before, while $w_S(t)$ is the w-component for an observation phase centred on S, the source to be removed. While the process



(a) Removed by filter



(b) Difference with the model



(c) Removed by filter when applied on white noise

Figure 4.4: *Results of performing a sliding window fringe filter. The configuration for the first and second panels are equal to Fig. 4.2, except that a sliding window fit is used. The window size was 128 time steps of 15 seconds integration, which corresponds to at most six fringes in one window. Most of the source has successfully been taken out. However, the middle panel shows two artefacts: The sidelobes of the removed source have not been taken out completely. The error is about 10 per cent at maximum, but the effectivity of the removal varies with direction. Second, artefacts are caused near the position of other sources, with errors up to 20 per cent of the sources at that position. This is the result of fitting the RFI on smaller parts of the data, causing the fit to respond as a sinc to other positions (as described in the text). The last panel shows the result that white noise with the same baseline settings would produce. The maximum error of the fit is about equal to the RMS of the noise, 0.4 in the image. Altogether, these simulations, with reasonable practical settings, show that a sliding window fit might be too inaccurate for practical applications.*

is easier and faster than normal off-axis source calibration or peeling, in practice it will be of little use: it neglects information present in polarizations, as defined by the measurement equation (Hamaker et al., 1996), and neglects the relations between baselines. Advanced calibration algorithms such as the space alternating generalized expectation-maximization (SAGE) calibration technique (Yatawatta et al., 2009; Kazemi et al., 2011) solve for source parameters by combining this information at once, and will in general be more accurate, as long as the source is (coherently) seen in multiple polarizations or antennas.

4.3 Novel filtering techniques

For high dynamic range, the source removal techniques as analysed in the previous section might not always suffice: the fringe fitting procedure can only remove a single unresolved source at a time. Also, since the fitting window has to be reasonably small, the fit will be slightly affected by the contribution of other sources. Therefore, the source has to be strong to be able to remove it, although the absolute error made will not depend on the strength of the source.

In the following sections, we will present several filters that are aimed to work when the fringe filter does not suffice. The key issues that these filter techniques share, are that they do not perform fitting on windows, but use the full data at once. They also remove high-frequency Fourier components that do not correspond with the fringe frequencies of sources of interest.

4.3.1 A low-pass filter in time domain

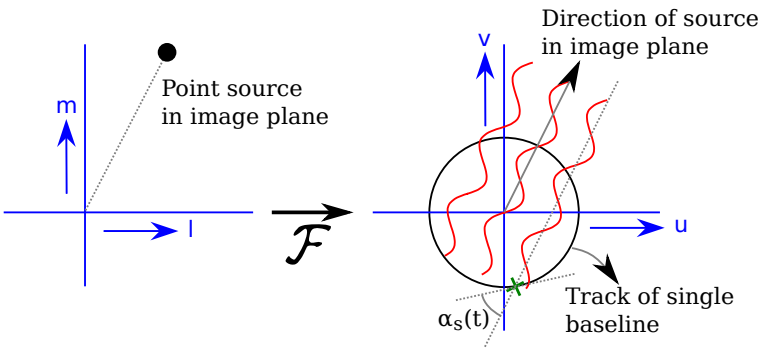


Figure 4.5: *Cartoon showing how a source in the image plane contributes fringes in the uv-plane. The further the source is from the phase centre (origin), the faster the fringe. Function $\alpha_S(t)$ is the angle between the direction of the source and the direction of a specific point in the uv-track as a function of time. The smaller α_S, the faster the fringe speed in the track at that point.*

The visibility of a single point source with strength I_{lm} and coordinates (l, m) is given by

$$V(u, v, w) = I_{lm} e^{i2\pi(ul+vm+wn)}. \tag{4.7}$$

Define $\boldsymbol{d} = (u, v, w)$ and $\boldsymbol{l} = (l, m, n)$. Since the source I_{lm} is real, the phase ϕ of V is given by

$$\phi(\boldsymbol{d}) = 2\pi \boldsymbol{d} \cdot \boldsymbol{l}. \tag{4.8}$$

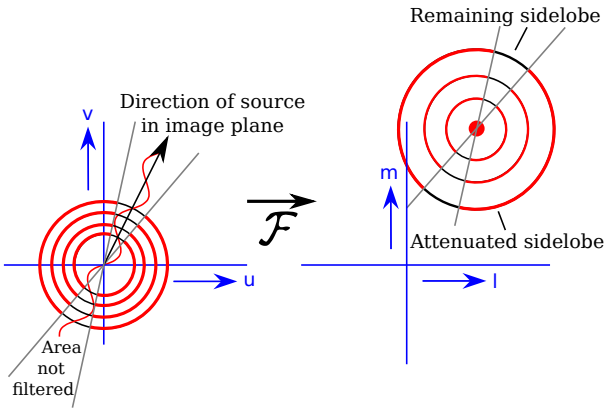


Figure 4.6: *Applying the low-pass filter on several baselines will filter parts of sources that exceed the frequency limit. For a particular source, this corresponds with multiplying the source with a hourglass shape in the uv-plane (left panel). Because of this multiplication, the sidelobes of the source in image plane (right panel) will be, relative to the phase centre, filtered in tangential direction. Sidelobes in radial direction will remain.*

The property that will be used in the filtering technique, is the implication of this formula that sources with large $|\boldsymbol{l}|$, i.e., that are far away from the phase centre, have a high fringe speed in the uv-plane.

Without loss of generality, we assume that our interferometer has a configuration such that its corresponding uv-track is a circle that is centred on the uvw-origin. This only occurs for an East-West Interferometer such as the WSRT. However, the technique can be straight forwardly extended to other interferometers that create possible elliptic tracks that might not be centred on the origin. In the assumed case, the uv-plane position $\boldsymbol{d}$ will be a function of time but have a constant size. If a time-sorted sequence of observed samples of a single correlation is considered, its fringe frequency is given by

$$\nu_S(t) = \frac{d\phi}{dt} = |\boldsymbol{d}|\,|\boldsymbol{l}_S| \cos\alpha_S(t), \tag{4.9}$$

where $\nu_S(t)$ is the fringe speed in fringes per second at time t for source S, $|\boldsymbol{d}|$ is the radius of the uv-track, $|\boldsymbol{l}_S|$ is the distance of S to the phase centre and $\alpha_S(t)$ is the angle between the uv-track and the line through S and the phase centre as drawn in Fig. 4.5. The fringe speed will be maximal at points where the corresponding uv-track is parallel to the direction of the source, and zero when the source direction and uv-track are orthogonal. The maximal fringe speed produced by a source is proportional to the distance between the source and the phase centre: $\nu_S(t) \propto |\boldsymbol{l}_S|$.

We will now consider low-pass filtering of the time-sorted visibility data with a filter frequency ν_F, specified in fringes per wavelength. Such a filter will have the following two properties: First, sources with $\forall t : \nu_S(t)/|\boldsymbol{d}| < \nu_F$, will never be filtered. In image plane, the area corresponding to $\nu_S(t)/|\boldsymbol{d}| < \nu_F$ is a circle that is centred on the phase centre. The fringe speed in the uv-track is translation independent, hence it is not necessary for the track to be centred on the origin. In case the uv-track is an ellipse, the filtering area will be an ellipse as well, but we will continue to assume circularity. Second, sources outside the circle will be filtered during the periods in which

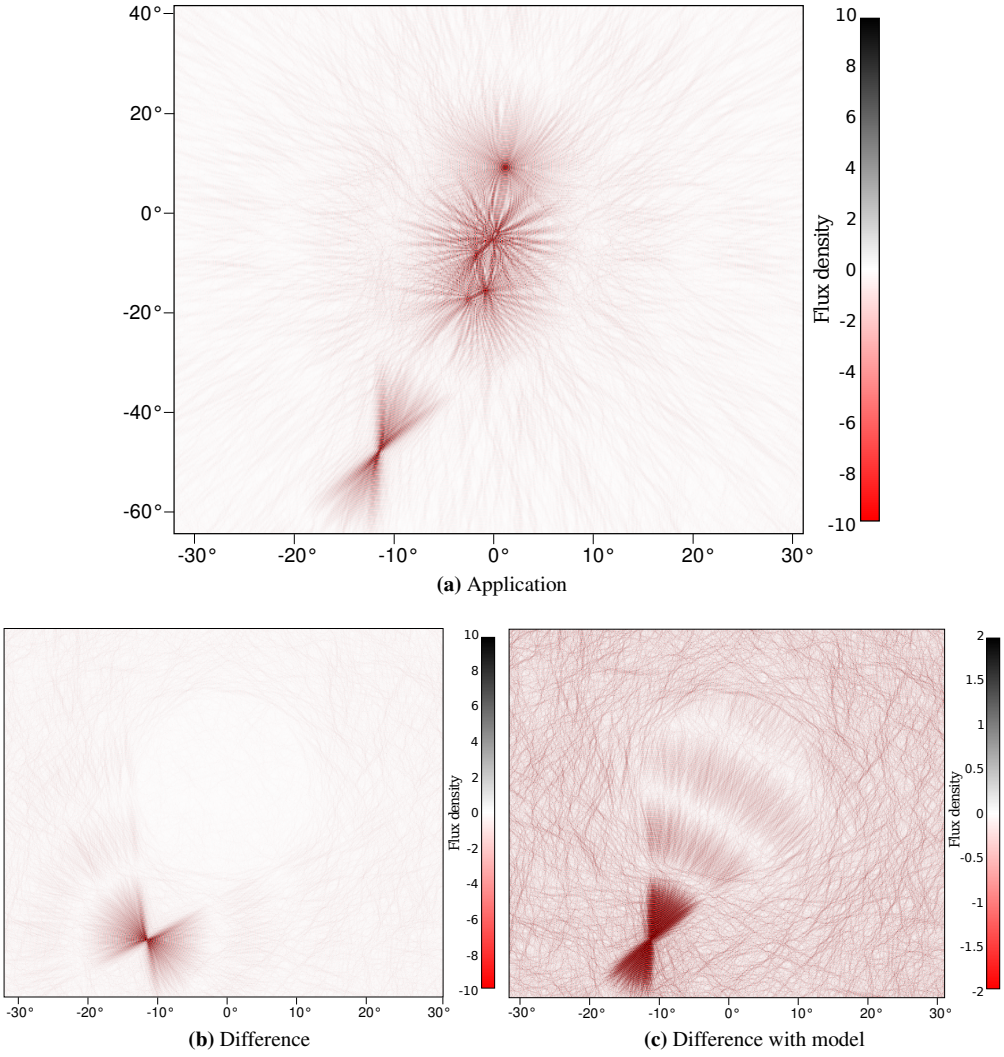


Figure 4.7: *Application of a low-pass filter in the time domain (§4.3.1). The source has been attenuated by filtering (first panel), but some of the sidelobes have not been removed. This is because the fringe rate of the source does not always exceed the filtering frequency. The second panel shows what has been removed and confirms that the sources of interest have not been attenuated (up to the 100 times lower noise level), the third panel shows with high contrast what has not been removed from the source. Note the different intensity scales.*

$\nu_S(t)/|\boldsymbol{d}| \geq \nu_F$. The differential start and end angle, respectively α_S^s and α_S^e, at which a source will enter the filtered area are given by

$$\begin{aligned} \alpha_S^s &= \arccos \frac{\nu_F}{|\boldsymbol{l}_S|}, \\ \alpha_S^e &= \pi - \arccos \frac{\nu_F}{|\boldsymbol{l}_S|}. \end{aligned} \tag{4.10}$$

The area filtered is independent of the baseline length because ν_F is specified in fringes per wavelength. For a single baseline, the filter ratio can be calculated with $\left(\alpha_S^e - \alpha_S^s\right)/\pi$. Consequently, in an array with N baselines with different sizes, the fraction of samples in which the source is filtered is given by

$$\begin{aligned} \rho_s &= \frac{1}{N} \sum_{i=0}^{N-1} \frac{\alpha_S^e - \alpha_S^s}{\pi} \\ &= 1 - \frac{2}{\pi} \arccos \frac{\nu_F}{|\boldsymbol{l}_S|}, \end{aligned} \tag{4.11}$$

which is therefore the total attenuation of the source by the filter.

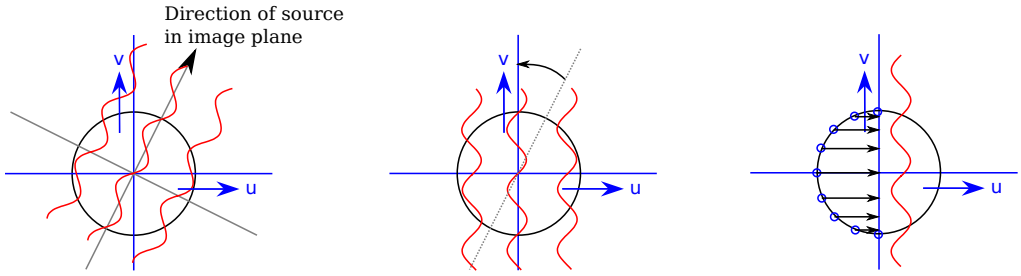
Although we have shown with Equation (4.11) that the total attenuation of a source is known, the shape of the area that is filtered is important as well, as that defines the shape of the sidelobes. The effect of low-pass filtering is sketched in Fig. 4.6: the filter removes the source fringes at two symmetric radial areas in the uv-plane. Subsequently, the application of this filter can be seen as an additional multiplication of the source in the uv-plane. Instead of a convolution with the nominal point spread function (PSF), sources in the image plane are convolved with a partly attenuated PSF. The side lobes that the source would normally have are not filtered in the direction of the phase centre, and can still increase the noise in the area of interest. This effect can be seen in Fig. 4.7.

Although this filter does not directly suppress confusion noise, it does filter high frequencies that can increase aliasing effects during averaging or gridding (§4.5.2). A more sophisticated filter will be presented in the next section, which utilizes the same theory about the fringe speed of sources.

4.3.2 A projected fringe low-pass filter in time domain

As was shown in Section 4.3.1, in order to remove the side lobes of an interfering source from the area of interest successfully, the interferer has to be filtered over the entire length of the observation. We will now introduce a filter with the purpose of filtering out all sources in a certain direction beyond a minimum distance from the phase centre.

The first step of the filter is to make the speed of fringes, coming from any source from a specific direction α_D, constant in the time direction. This is done by rotating the uv-plane such that the fringes are parallel to the v-axis, and subsequently projecting the samples from the track onto the v-axis, thereby stretching the high-frequency fringes and pushing together the low-frequency fringes from sources from direction α_D. Fig. 4.8 visualizes the transformation. At each point on the uv-track given by an angle $\alpha(t)$, the fringe frequency $\nu_S(t)$ of a source at time t is multiplied by a factor due to the projection, resulting in a new fringe frequency $\nu_{\text{projected}}$ at angle



(a) Fringe contribution of a source **(b)** Rotation **(c)** Projection

Figure 4.8: *Creating a constant fringe rate towards a single direction. Panel (a): A source with a certain direction from the origin in the image plane will cause a fringe in the uv-plane corresponding to that direction. Panel (b): Rotating the direction of the source onto the v-axis will align its fringe with that axis. Panel (c): Projection of the sample track onto the v-axis will make any source in the direction of rotation have a constant fringe rate.*

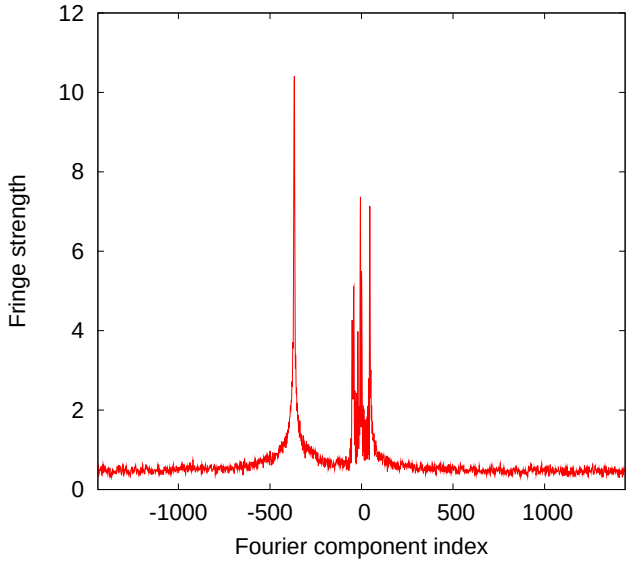
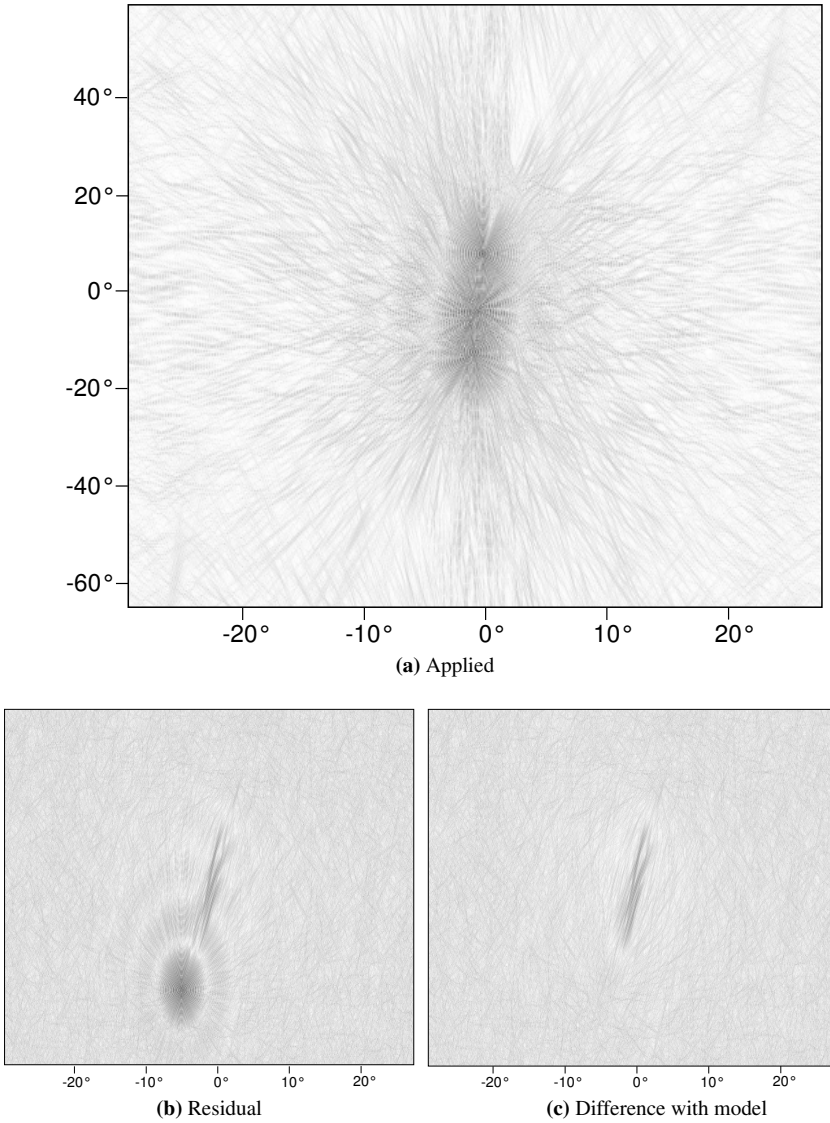


Figure 4.9: *Fourier transform of a uv-track that was rotated and projected, such that sources in a certain direction have a constant fringe speed. The model of Fig. 4.7 was used. Most of the contribution of sources near the centre collect near Fourier component index zero, while the contribution of the off-axis source shows up as a peak at an index away from zero.*



(a) Applied

(b) Residual

(c) Difference with model

Figure 4.10: *Application of the projected fringe low-pass filter (§4.3.2) on simulated data. The projected fringe low-pass filter nulls a single direction starting at a certain distance, but does not preserve the phase centre well. In this simulation, the off-axis source has been removed completely up to the noise, two orders of magnitude lower. In (a), the filter is applied and the top source is removed. Panel (b) shows what has been removed from the image, while (c) shows what has been removed from the area of interest.*

$\alpha(t)$ on the circle, given by

$$\nu_{\text{projected}} = \frac{\nu_S(t)}{\cos\left(\alpha(t) - \alpha_D\right)}. \tag{4.12}$$

By substituting the definition of $\nu_S(t)$ from Equation (4.9) into this equation for a single source in the direction of the filter, i.e., $\alpha_S(t) = \alpha(t) - \alpha_D$, the result is $\nu_{\text{projected}} = |\boldsymbol{d}|\,|\boldsymbol{l}_S|$. Hence, the fringe speed becomes independent of time. Sources from other directions, however, will not become constant.

An example of this effect is shown in Fig. 4.9, which shows the Fourier transform of a projected uv-track. The model of Fig. 4.2a was used as input. The projection is towards the direction of the strong source in the bottom. This source shows up as an isolated feature away from Fourier component index zero, because this source lies furthest away from the phase centre. Although the power of this source peaks in one component, it is distributed over several Fourier components, because the time series is finite. Therefore, the point is convolved with the Fourier transform of a windowing function. The sources near the phase centre collect at component indices around zero.

By performing a low-pass filter with frequency ν_F on the projected samples, we will remove fringes from sources at time $t \in [t_0; t_e]$ for which

$$|\boldsymbol{l}_S| \left| \frac{\cos\alpha_S(t)}{\cos\left(\alpha(t) - \alpha_D\right)} \right| > \nu_F \tag{4.13}$$

holds.

Fig. 4.10 visualizes the application of the filter. Its effect can be summarized by these three characteristics: (A) any sources at direction α_D that are further away than the limiting distance corresponding to ν_F will completely be removed; (B) sources at direction α_D within the limiting distance will not be removed at all; and (C) any sources from directions other than α_D will neither be removed completely nor stay untouched completely. The latter is because the denominator and the numerator in Equation (4.13) will have zero crossings at different t. Consequently, the left term in Equation (4.13) will become large when the denominator is near zero.

While incomplete filtering of sources in some directions that are not of interest is not very problematic, it is impractical that the only sources for which absolute preservation can be guaranteed, are sources that lie on the line going through the phase centre in the direction of the applied rotation. In the next subsection, we will present modifications that will solve this issue.

Despite this complication, this method might still be usable in practice. According to Equation (4.13), the fringes of sources will all be filtered around the same angle $\alpha(t)$ in the uv plane. This direction is known, and the area in the uv plane that is affected is therefore known. Samples in this area can be removed from the data, causing a small loss of data. However, the source will successfully be removed without side effects.

4.3.3 The iterative projected fringe filter in time domain

The projected fringe frequency of an on-axis source can exceed the filtering frequency when $\alpha_S(t) \approx \alpha_D$, i.e., when the uv-track is near parallel to the applied direction of the filter. To create an area of unfiltered sources in the image plane, one can leave this range out of the filter. This however, would create artefacts similar to the low-pass filter of §4.3.1, and would still not improve the dynamic range in the area of interest.

A solution is to perform a Fourier transform only on the part of the projected samples at which $|\alpha_S(t) - \alpha_D| > \eta_F$, for some small angle η_F, and use a deconvolution method to extrapolate

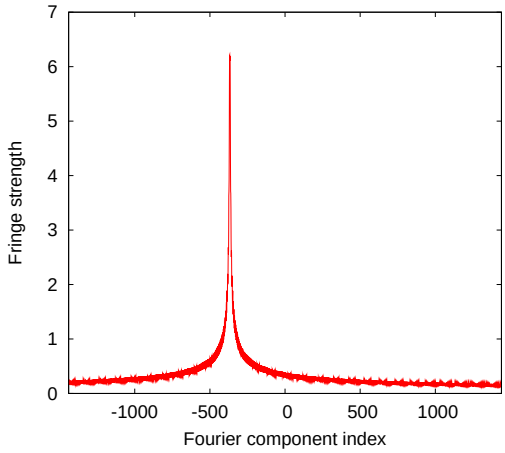


Figure 4.11: *Visualization of the first component in a one-dimensional CLEAN of the plot in Fig. 4.9.*

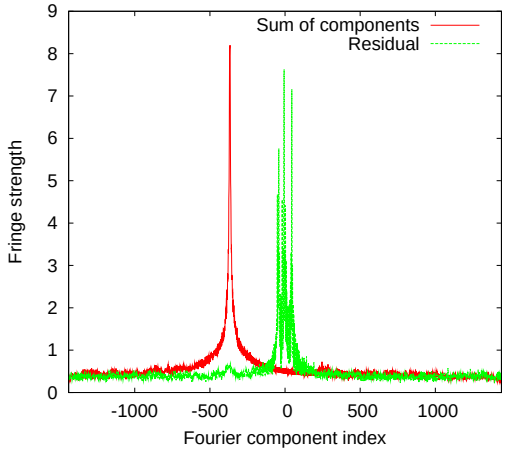
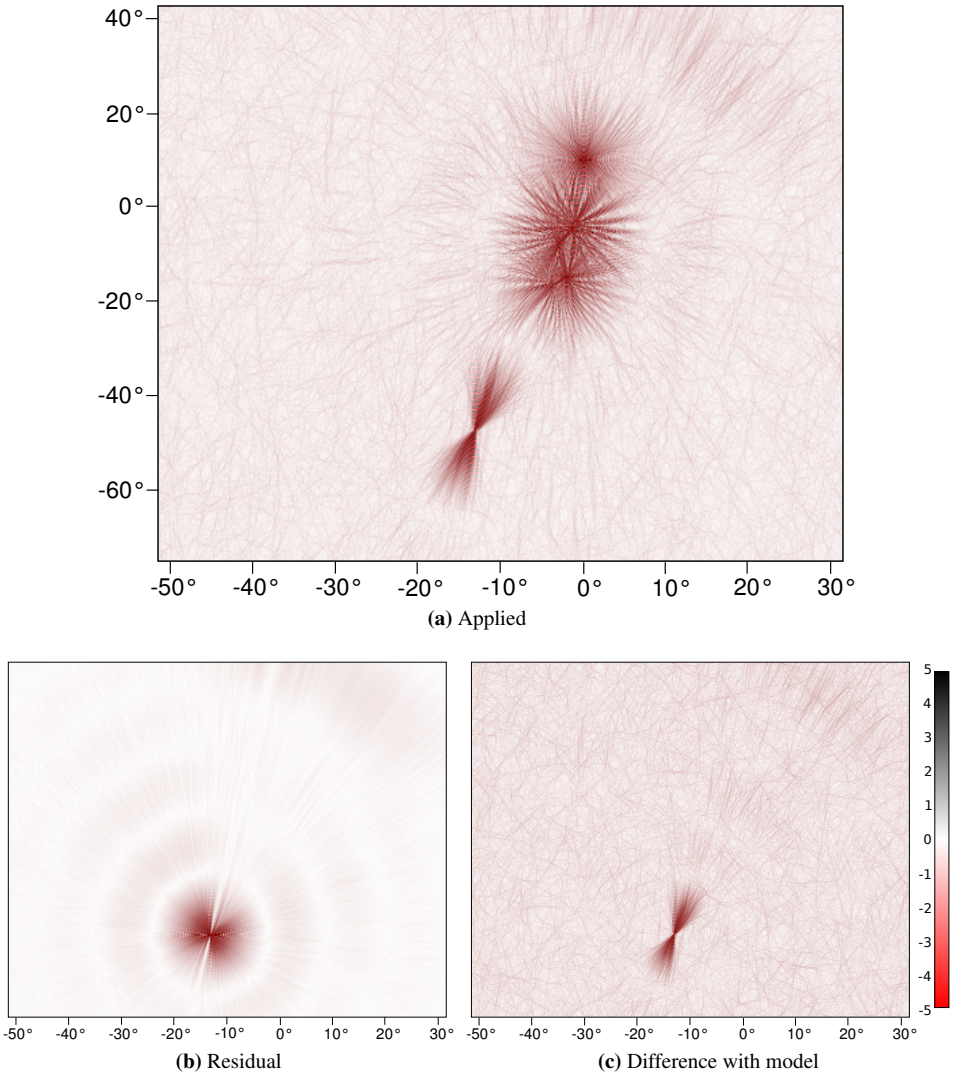


Figure 4.12: *In red, showing the sum of the first hundred components removed by the deconvolution and in green, showing the residuals that contain the data for the area of interest. In the Fourier transform similar to Fig. 4.9, η_{filter} part of the data around $\alpha_S(t) \approx \alpha_D$ was left out to make sure no sources in the area of interest map to higher components.*



(a) Applied

(b) Residual

(c) Difference with model

Figure 4.13: *Application of the iterative projected fringe filter (§4.3.3) on a single simulated baseline of 720 m as in Fig. 4.2. The filter was aimed at the source in the bottom and iteratively removes fringes with high frequency. A value of $\eta_{filter} = 0.2$ was used to preserve all of the centre sources, and 100 one-dimensional CLEAN iterations were performed in the projected fringe spectrum domain. Although this has attenuated the source without needing a model of the source, the sidelobes in the direction of the phase centre still remain.*

the found frequencies to the area that has been left out. A one-dimensional CLEAN on the fringe spectrum can be used to remove and extrapolate fringes, taking fringes out one by one. Altogether, such a filter removes sources from a single direction α_D at a distance corresponding to ν_F and create a rectangular area around the phase centre which will be preserved. The width of this area is given by

$$\kappa(\nu_F, \eta_F) = \frac{\nu_F}{|\boldsymbol{d}|} \left|\sin \eta_F\right| . \tag{4.14}$$

Off-axis sources from directions other than α_D will be partially removed and sources of interest will be fully preserved. We will discuss the results of practical application of this filter in §4.4.

Fig. 4.11 visualizes the Fourier transform of the first component that will be removed by a one-dimensional CLEAN on the plot in Fig. 4.9. In the Fourier transform, η_{filter} part of the data was left out. Because of the finite time domain, the power in a single component is convolved with a function formed by the windowing function, which also depends on the angle between the source and the filter direction. Intuitively, one can think of this as the shape of the PSF in the projected fringe spectrum domain of a single baseline. 75 per cent of the power in the highest component are selected for subtraction in each iteration. Figs. 4.12 and 4.13 show the resulting projected fringe domain and image domain respectively, after applying the iterative fringe filter with 100 iterations.

4.3.4 Filtering in frequency direction

The filters that have been presented so far, have been applied in the time domain of correlations from a single baseline. If an interferometer observes several frequency channels over some limited bandwidth, a logical extension is to filter in frequency direction. The samples from different frequencies in the same baseline at the same time form a straight line in the uv-plane. A source S produces a fringe speed μ_S in frequency direction given by

$$\mu_S(t, \lambda) = |\boldsymbol{d}(\lambda)||\boldsymbol{l}_S| \sin \alpha_S(t), \tag{4.15}$$

and $|\boldsymbol{d}(\lambda)| \sim \frac{1}{\lambda}$.

A low-pass filter in the frequency direction removes fringes of off-axis sources at which $\mu_s(t, \lambda) < \mu_f$. In contrast to filtering in time, the situation differs on some points:

- The use of the $\sin$ function in Equation (4.15) implies that sources produce a high fringe rate in frequency direction when the uv-track is orthogonal to the source direction in the image plane. The result is that the source sidelobes in direction of the phase centre, which is the area of interest, will be removed. Therefore, a low-pass filter in frequency direction would complement a filter in time direction, which depends on the cosine of the source angle and the uv-track (Equation (4.9)). Therefore, the part that is not filtered by the latter can be further attenuated with a frequency direction low-pass filter.

- While most radio sources are constant over the observation time, they vary over frequency. Low-pass filtering in frequency would low-pass filter the variation of the source over frequency. Because the primary beam is smaller at higher frequencies, an off-axis source can have a steep apparent spectral index.

- In the frequency direction, the number of fringes is limited by the observing bandwidth, and the bandwidth might be limited such that the fringes of a source rotate too little for

filtering. For example, if a bandwidth-frequency ratio of 2.5 MHz/100 MHz is assumed for a 100 m baseline (approximately the shortest WSRT baseline observing with a single band), a source needs to be at a distance of about 8° from the phase centre to create a single fringe within the bandwidth.

Due to these characteristics, the use of a frequency filter can complement a low-pass filter in time, but might be limited to the longer baselines or large filter radii. To be effective, sufficient bandwidth is required. The available bandwidth for filtering might be further limited if the apparent spectral indices of the off-sources are steep.

4.4 Practical applications

Several filters for off-axis sources were described in the previous chapters. Fig. 4.14 shows an overview of all the filters, applied on several classes of simulated off-axis sources. The fringe filter works well, as long as an accurate model of the source exists, and the received strength of the source does not change much in time. The low-pass filters in time and frequency direction together remove the off-axis source quite well. The projected iterative fringe filter in time direction can only attenuate the off-axis source moderately, even though it requires an accurate estimate of the source location. Application of the method on real data shows comparable results.

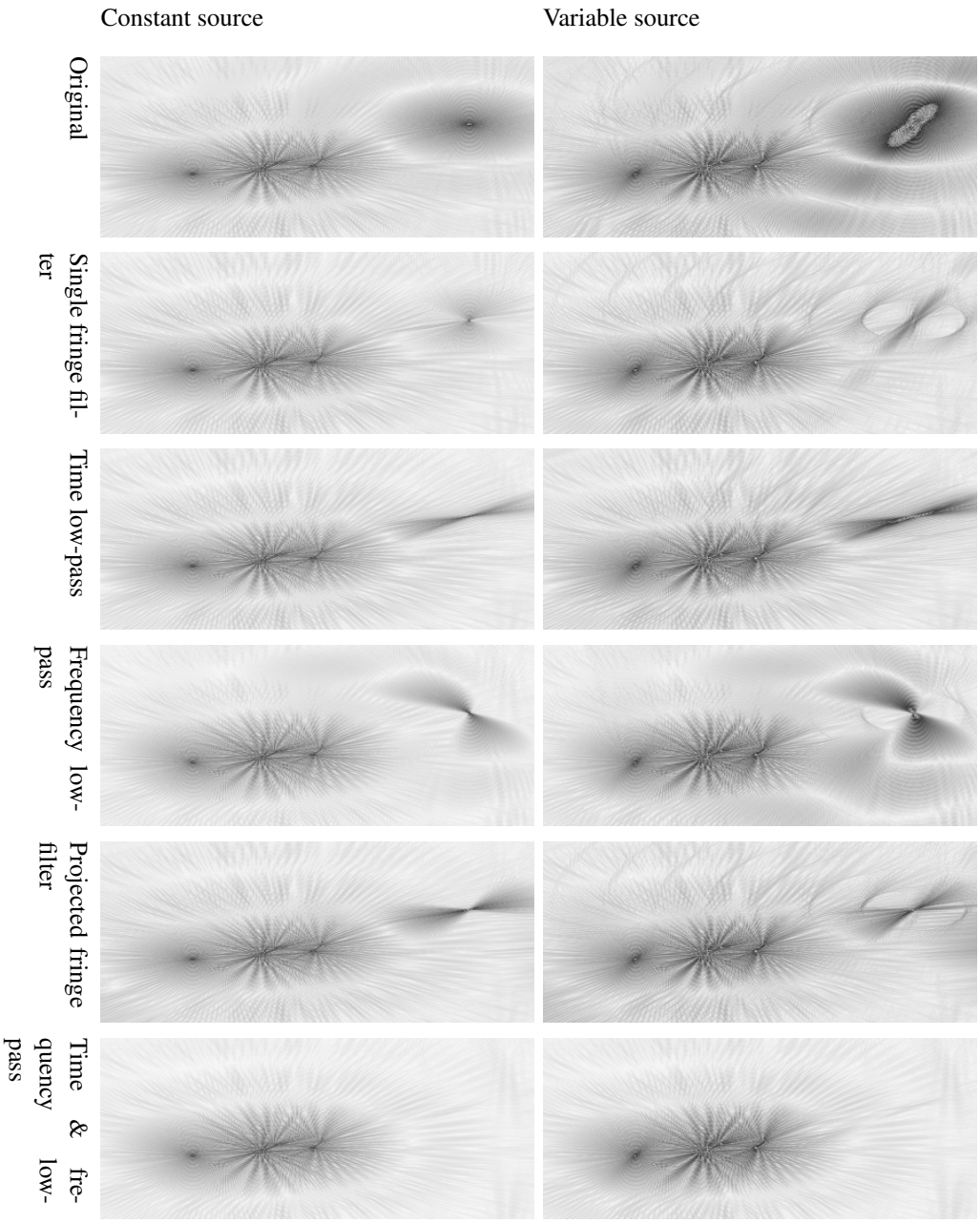
4.4.1 Attenuation efficiency

To test the level to which sources can be removed, we have simulated a single 40 degrees off-axis source in an otherwise empty field, i.e., without any on-axis sources, and also without noise. We simulated a single 2.5 MHz band at 130 MHz with a standard WSRT configuration and compared the level of the sidelobes before and after source filtering. The single fringe filter shows 40 dB of sidelobe attenuation on a constant source, but only attenuates up to 3 dB of a varying source, which provides a more realistic setting.

The frequency direction low-pass filter can remove 10 dB of a source, which can be varying. Because the low-pass filters are less effective near the borders of the band and the start and end of the observation, we have tried flagging 5 per cent of the border channels in the time frequency plane after filtering. This leads to 20 dB of attenuation. The low-pass filter in time direction does in theory not remove sidelobe noise in the direction of the source. However, in practice, it attenuates the RMS in areas around the phase centre by zero to 3 dB. This is because of a property of gridders: high fringe frequencies are mapped back to the area of interest, i.e., resampling causes aliasing effects. Therefore, removing the high frequencies before imaging lowers the noise as well. The RMS decrease in the radial direction due to low-pass filtering in time is around 25 dB. The large difference between attenuation of the tangential direction of time low-pass filtering versus the radial direction of frequency low-pass filtering is due to the limited bandwidth: in time, the observation contains lots of fringes which can be accurately filtered, but only a few fringes appear in frequency direction.

In the same test, the projected fringe low-pass filter shows 25 dB of attenuation around the phase centre. Finally, the projected iterative fringe filter attenuates only up to 3 dB.

Obviously, these results are highly dependent on many parameters, including the distance of the source to the phase centre, the amount of available bandwidth and its central frequency, the



(continued on right side)

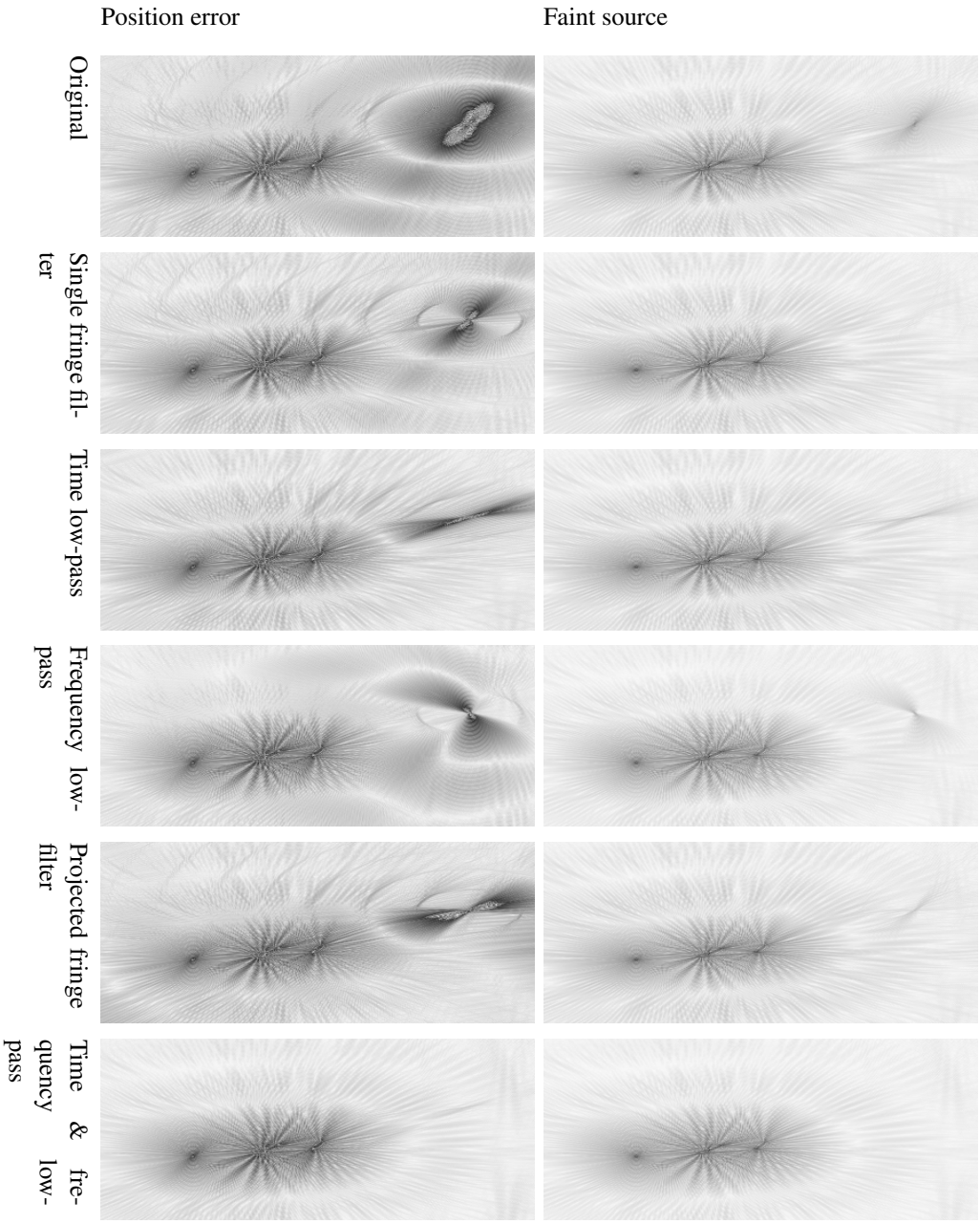


Figure 4.14: *Simulated test sets with various types of off-axis sources that need to be removed. On its own, the single fringe filter removes the largest part of the source and its sidelobes, and only becomes inaccurate when the source changes in time or when the the model is inaccurate. The time and frequency low-pass filter complement each other, and together can remove everything outside a certain radius, if bandwidth allows. The projected fringe filter seems not to work very well – it removes a part of the source, but leaves artefacts in the image in every test case.*

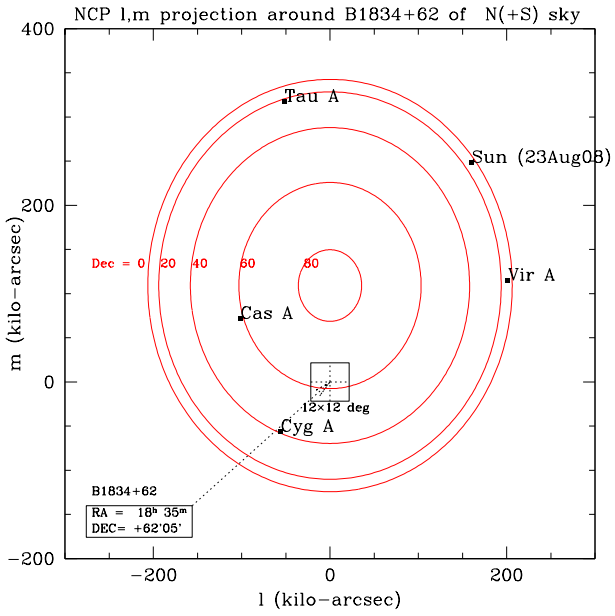


Figure 4.15: *Position in the sky of B1834 relative to other strong sources.*

time and frequency resolutions and, for the single fringe filter, the speed of change of the source due to instrumental effects and the number and size of the interferometers.

4.4.2 Low-pass filtering a WSRT observation

We will now apply the filtering approaches on a WSRT dataset of the field centred on the radio galaxy B1834+62. This field was observed to search for polarized emission in this double double radio galaxy (Schoenmakers et al., 2000) at very low frequencies. The observations were done in August 2008 and lasted for 12 h. The backend was configured to observe 8 frequency bands, each 2.5 MHz wide and covered in 512 spectral channels, at frequencies ranging from 115 to 163 MHz. Here we will only use data from the band at 139 MHz. The integration time was 10 s, the spectral resolution, after Hann tapering, was 10 kHz. At this time and spectral resolution even sources more than 1 radian from the phase tracking centre were not significantly smeared. The field was affected by sidelobes from Cygnus A, Cassiopeia A and the Sun (for about 8 hours). An image of the locations of these sources, in the North Celestial Pole (NCP) projection of the whole sky suitable for the WSRT — an East-West array — is shown in Fig. 4.15.

Although each of these three sources is not in the primary beam, each of them is strong enough to lower the dynamic range of the observation considerably because of their sidelobes in the image plane. It is hard to remove these sources from the observation, because they are in the sidelobes of the beam and, especially in the case of the Sun, they are complex and their apparent strength varies over time. Because we do not have accurate models of the sources in our observation, the low-pass filters are a good choice, and we will show that the low-pass filters prove to be quite effective for attenuating the three sources.

Table 4.1: *Fringe speed in time and frequency directions as a function of scale, looking at zenith with a 1 km baseline.*

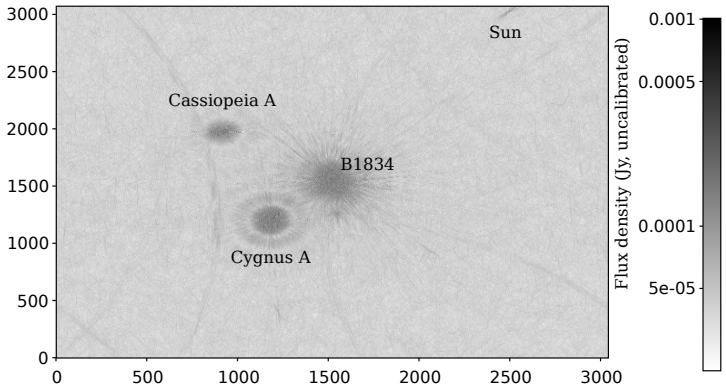
	1 km		λ =21 cm
Scale	Time	Freq	Time
	λ/h	MHz^{-1}	h^{-1}
45°	2.9	2.4	140
10°	0.72	0.58	34
1°	0.073	0.058	3.5
	λ/d	GHz^{-1}	d^{-1}
10 arcmin	0.29	87	14
1 arcmin	0.029	8.7	1.4

Fig. 4.16 shows a single baseline of the B1834 observation. The baseline used is RT0 $\times$ RTA, a 1.3 km East-West baseline, and only data from a single 2.5 MHz band at 140 MHz was used. The displayed images correspond to several tens of degrees of the sky. The observation is limited by confusion noise of the Sun (right top corner, also aliased to the bottom), Cassiopeia A (left top) and Cygnus A (left bottom). The observation takes 12 hours and the (resolved) contribution of the Sun moves through the image and sets halfway. Consequently, the Sun and its sidelobes would be very hard to remove with traditional methods. The two low-pass filters together remove the Sun down to the noise: in the filtered image, its peak value is 1 per cent of the original value. It is hard to remove more, i.e., make the filter circle smaller, since only a small bandwidth is available. Because of this, the edge of the filter border is blurred in the frequency filtering cases. For the same reason, Cassiopeia A should have been filtered but is removed only 95 per cent, and Cygnus A should not have been filtered, but is attenuated 25 per cent. These errors occur because these sources are too close to the filter border. Other sources within the filter radius have been attenuated less than 1 per cent.

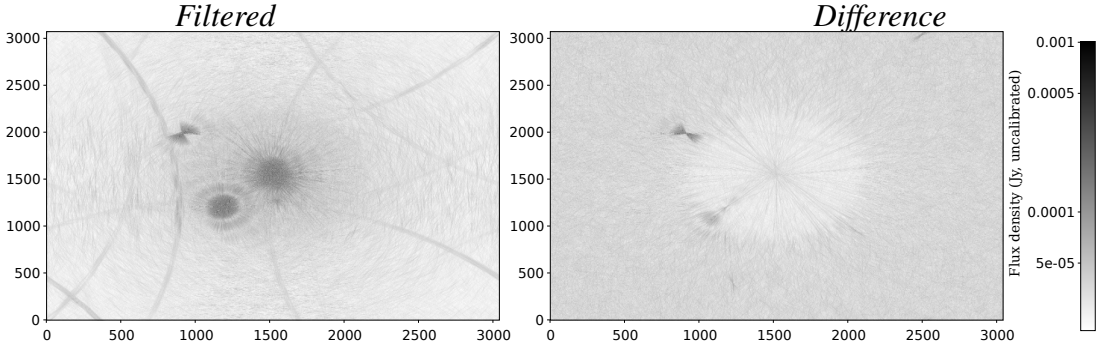
The application of the low-pass filters on this baseline shows the practical effectiveness of the filters: filtering in time direction removes the tangential components of the sources, while the frequency direction removes the radial components. The frequency filter is not as accurate as the time filter, because of the limited 2.5 MHz bandwidth available. This causes the circular "filtered" area not to have a sharp edge that a perfect sinc function would produce. Instead, the edge is somewhat blurred. As a consequence, a part of Cassiopeia A has been removed, although it did not exceed the theoretical cutting frequency.

In Fig. 4.17, a shorter baseline was processed with the filtering techniques. Baseline RT0 $\times$ RT2 was used, which is only 288 meters long. Because of the combination of a short baseline and the small available frequency bandwidth, the frequency filter is only able to filter out 80 per cent of Cassiopeia A on this baseline. The Sun is still successfully attenuated over 99 per cent, up to the noise. Cygnus A is 10 per cent attenuated. No other sources in the area of interest have been visibly attenuated. Because the off-axis sidelobe noise RMS is around 10 per cent of the peak of strong on-axis sources in the area of interest, one can conclude from this image only that the on-axis sources have been preserved for at least 90 per cent.

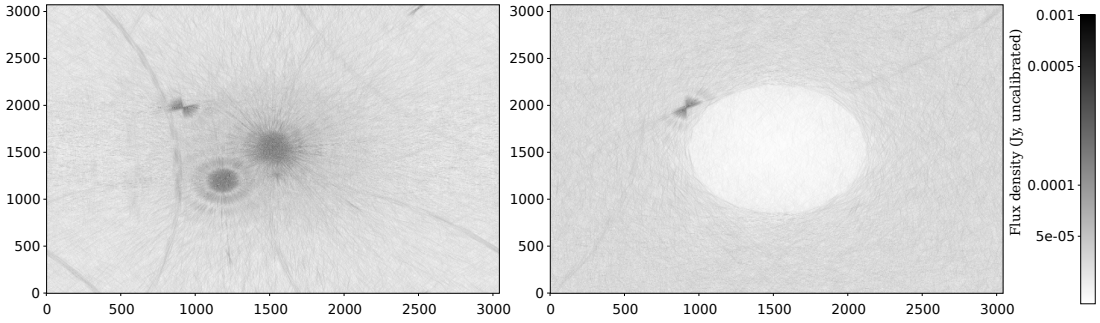
As discussed, the filter frequency scales linearly with the baseline size: on long baselines, the fringe speed of sources is fast in both the frequency direction and the time direction. On



(a) Original

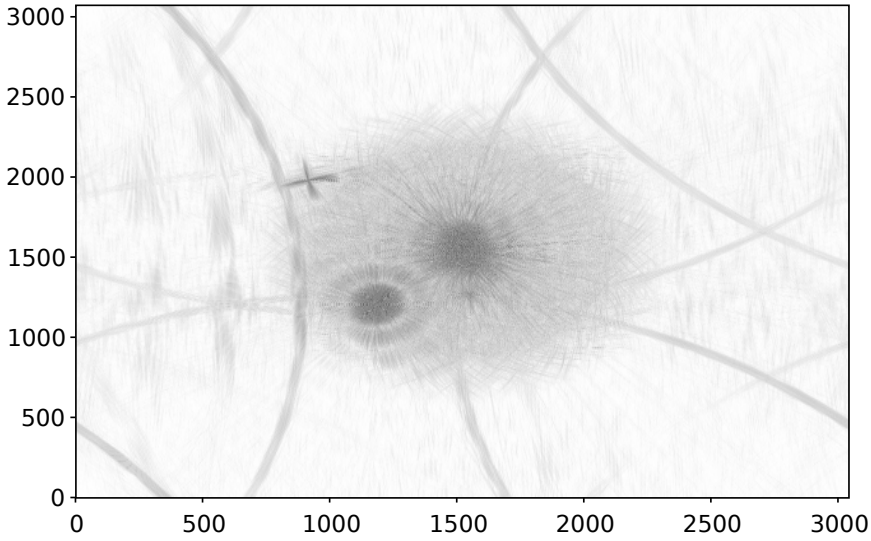


(b) Low-pass filter in frequency direction

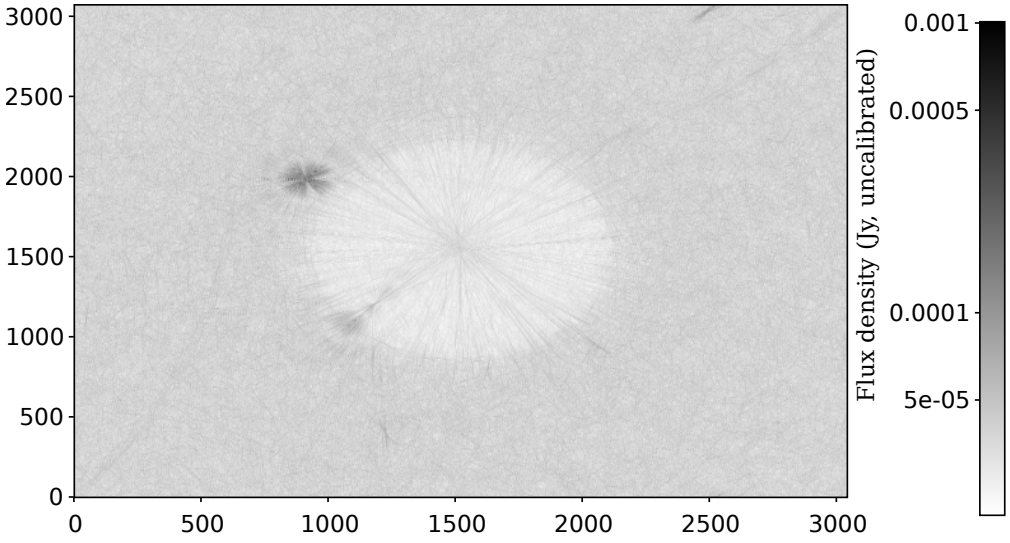


(c) Low-pass filter in time direction

Figure 4.16: *(continued on right side)*

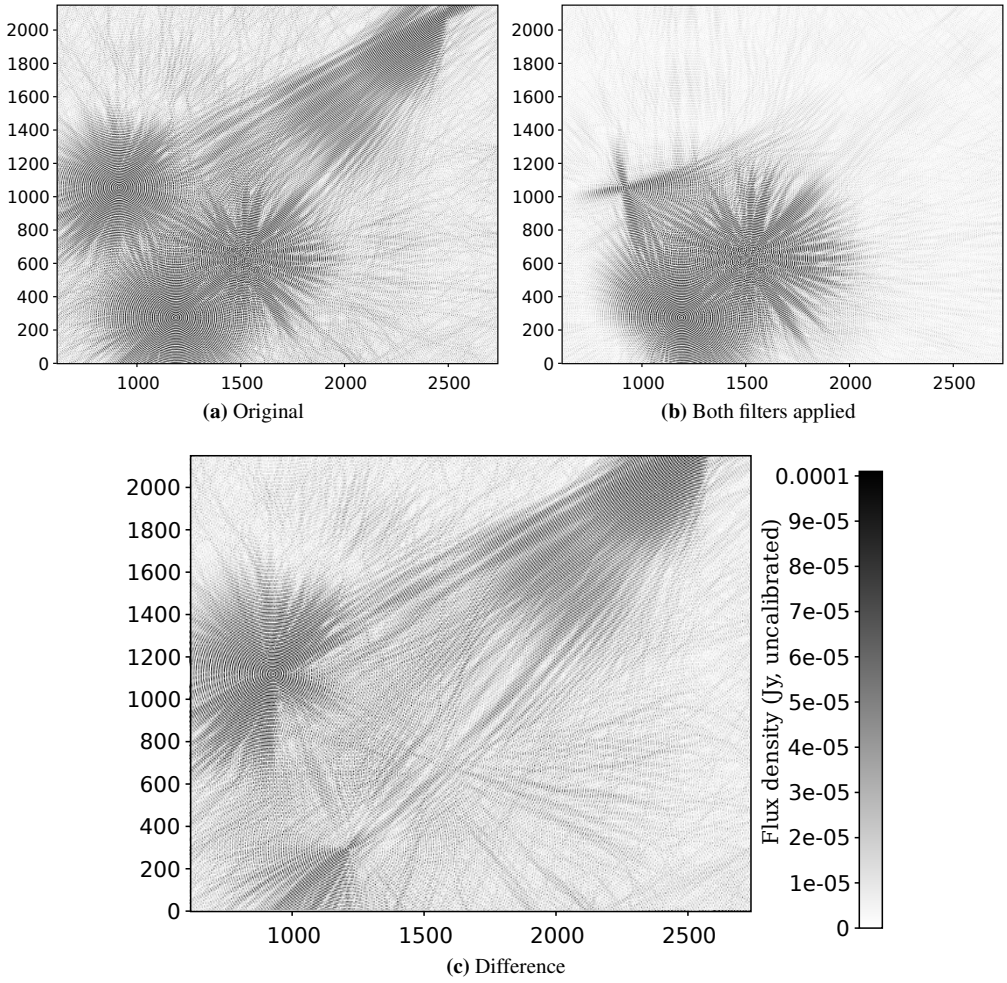


(d) Low-pass filter in both directions (filtered)



(e) Low-pass filter in both directions (difference)

Figure 4.16: *Application of the low-pass filters on a single 1.3 km baseline of an actual WSRT observation of the B1834 area, observed partially in daytime. Frequency filtering removes the Sun down to the noise, including its sidelobes in the area of interest. The filter is less effective near the circular filter edge. The rings are aliasing effects.*



(a) Original

(b) Both filters applied

(c) Difference

Figure 4.17: *Application of low-pass filters in both directions as in Fig. 4.16, but on a shorter baseline of 288 meters. The Sun is successfully attenuated, but the filter has been less effective on Cygnus A and Cassiopeia A.*

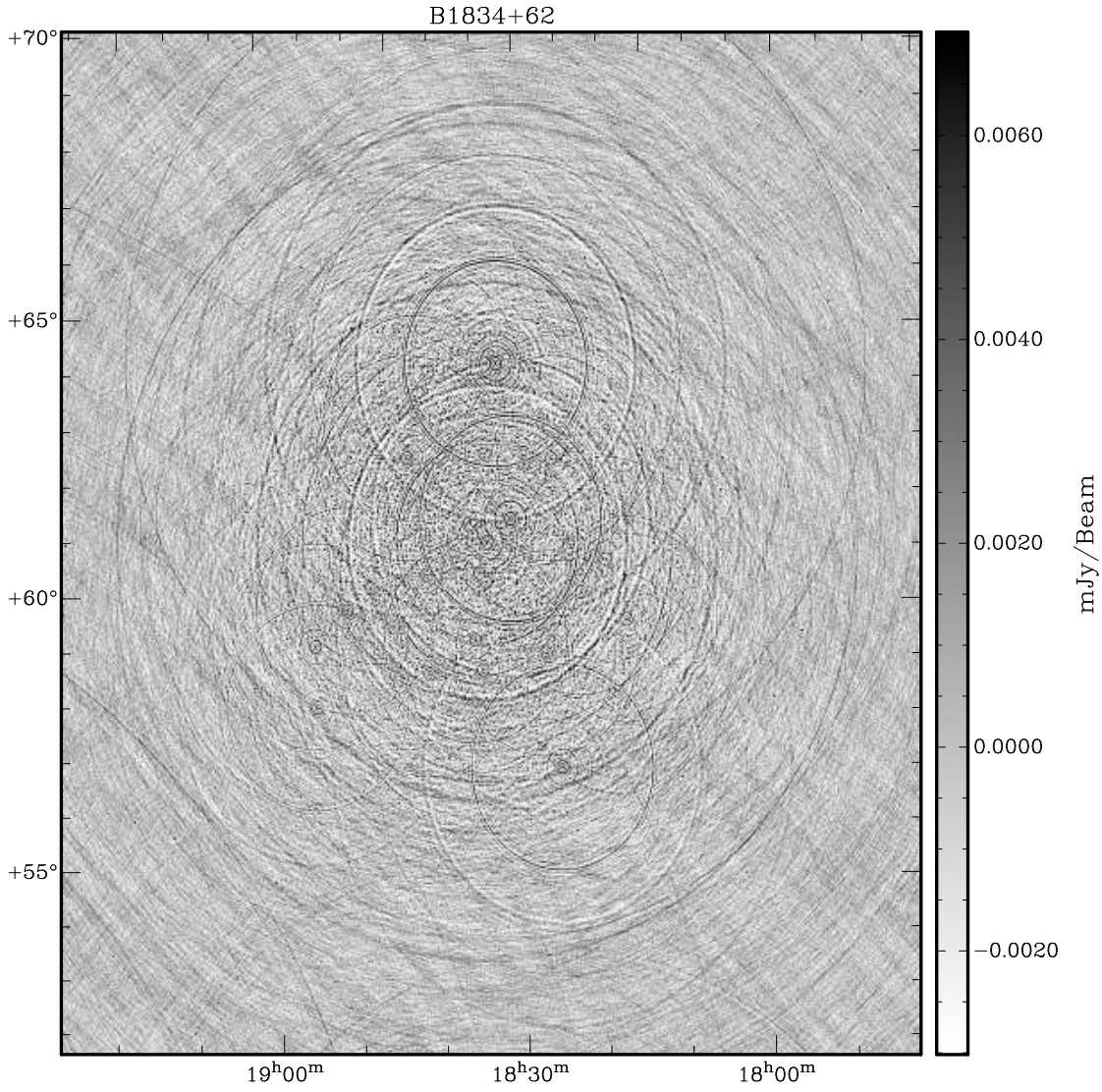
short baselines, a source might cause only a few fringes or less in the frequency direction. It is therefore more difficult to filter short baselines, and Fig. 4.17 visualizes this problem. While the tangential contribution of Cygnus A has been removed effectively in the figure, only a small part of its radial contribution has been removed. The filter was able to remove the Sun because it is further away. On very short baselines, the real and imaginary components produced by a source are almost constant, and applying a low-pass filter in frequency direction on such a baseline will perform similar to averaging the frequency channels. In such cases, the filter will not affect the astronomical data, but only average the noise out. If the fringe speed does not exceed the filtering frequency sufficiently on all baselines, the source will appear in the shorter baselines, hence the large scale structures of the source sidelobes will remain. In general, the combination of bandwidth, filter area and baseline length define the success of the frequency filter. Table 4.1 shows a few configurations and their corresponding fringe speed for a particular baseline size and distance to the phase centre.

In Fig. 4.18, all baselines were imaged together. The unfiltered Stokes I image is quite severely affected by sidelobes coming from off-axis sources. Moreover, because the off-axis sources come in through the far side of the primary beam, they appear in the polarized images as well. After filtering, the confusion noise is reduced significantly. Depending on the empty region over which the RMS has to be calculated, the noise goes down by a factor of 1.5–2 in Stokes I, while the polarized images show a factor of 2–3 decrease in noise. Because the short baselines could not be filtered correctly in the frequency direction due to the limited bandwidth, the low-frequency components of the sidelobes remain. With sufficient bandwidth, such as for LOFAR, the results will be even more significant. CLEANing the images of Fig. 4.18 removes some of the bright sources in the centre, but the strong sources in the sidelobes can not be removed by CLEANing. As one can expect, the CLEAN algorithm is able to CLEAN deeper and find more sources in the filtered image.

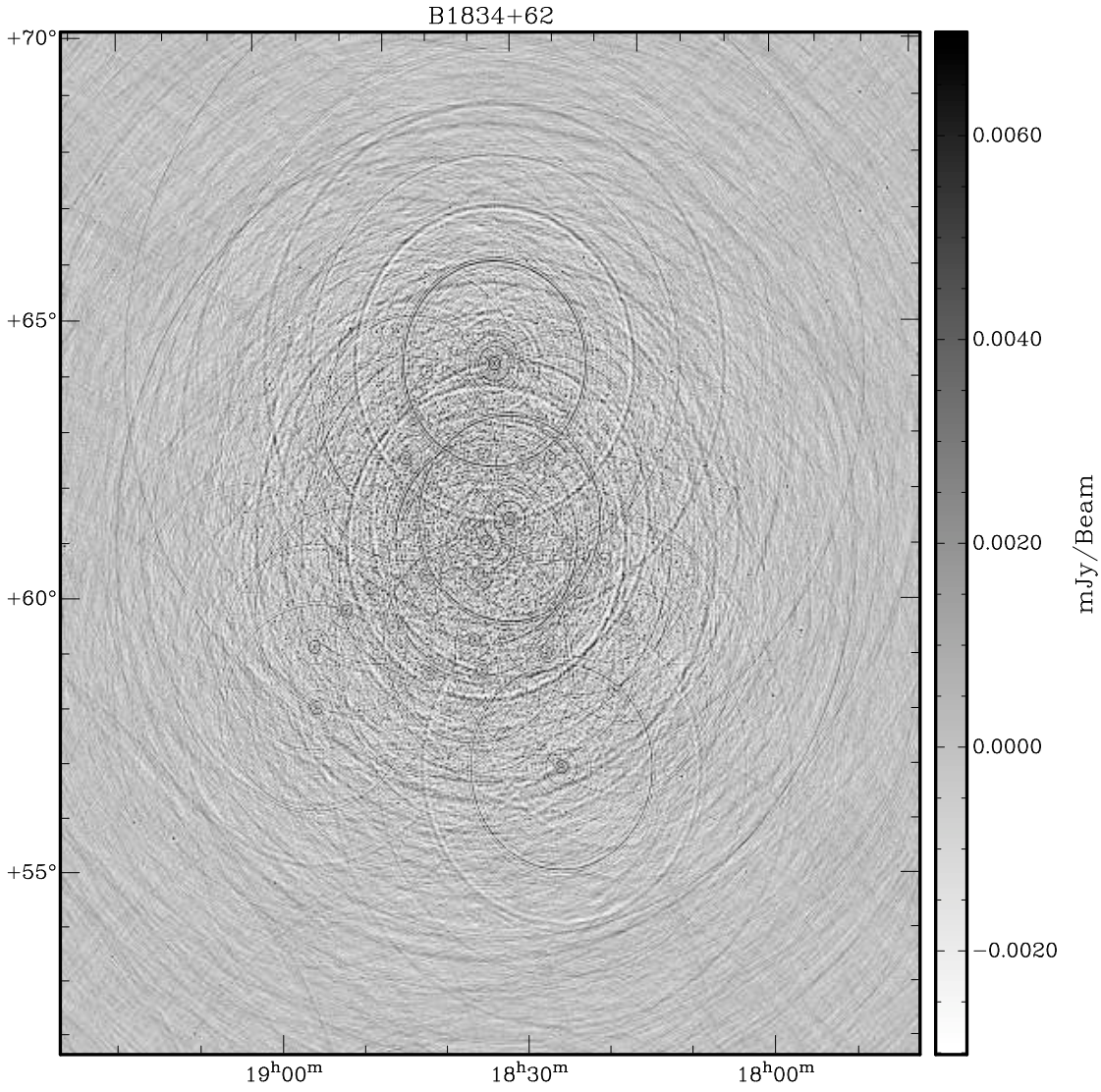
Another less obvious effect of the filter is suppression of ghost sources that are caused by aliasing of the off-axis sources. When looking at Fig. 4.18, it appears that there is one strong polarized source near the centre of the field. However, when performing the low-pass filters, the source disappears. The reason for this is that the source is not a real source, but a low frequency projection of an off-axis source: a ghost. Zooming in on this ghost as in Fig. 4.19 shows that the ghost is also present in Stokes I. This ghost is an aliasing artefact caused by the gridding in the imager. It appears as a normal source and contains regular sidelobes, as can be seen in Fig. 4.18. Low-pass filtering in time and frequency attenuates the ghost, as will any other method that attenuates the original off-axis source. The aliased ghost is caused by baselines which are gridded just below the Nyquist rate of the source. If the source is sampled correctly, its ghost will not appear at all. On the other hand, if the source is badly undersampled, its contribution will average out.

4.4.3 Dealing with flagged samples

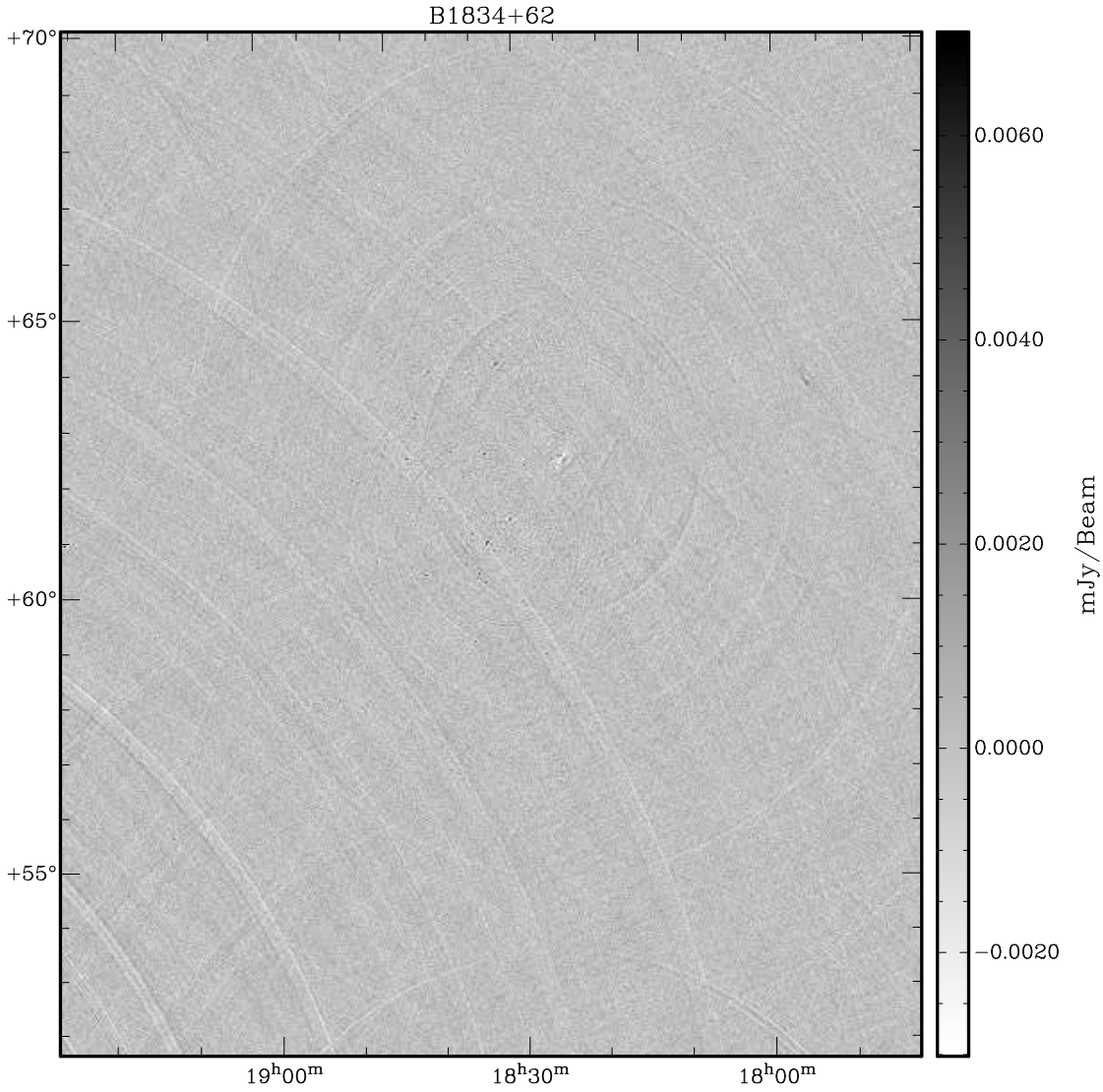
A complicating factor for low-pass filtering the time-frequency domain is the fact that the time-frequency plane contains flagged data due to RFI contamination. This has to be taken into account before convolving the data with a sinc function. To solve the problem, we will mimic how flags are handled during other stages of reduction. Two techniques for solving flagged samples are commonly used. The first is to set flagged samples to zero and account for the missing samples when deconvolving. The second is, if the samples are flagged before either correlation, further



Unfiltered Stokes I



Filtered Stokes I



Unfiltered Stokes Q

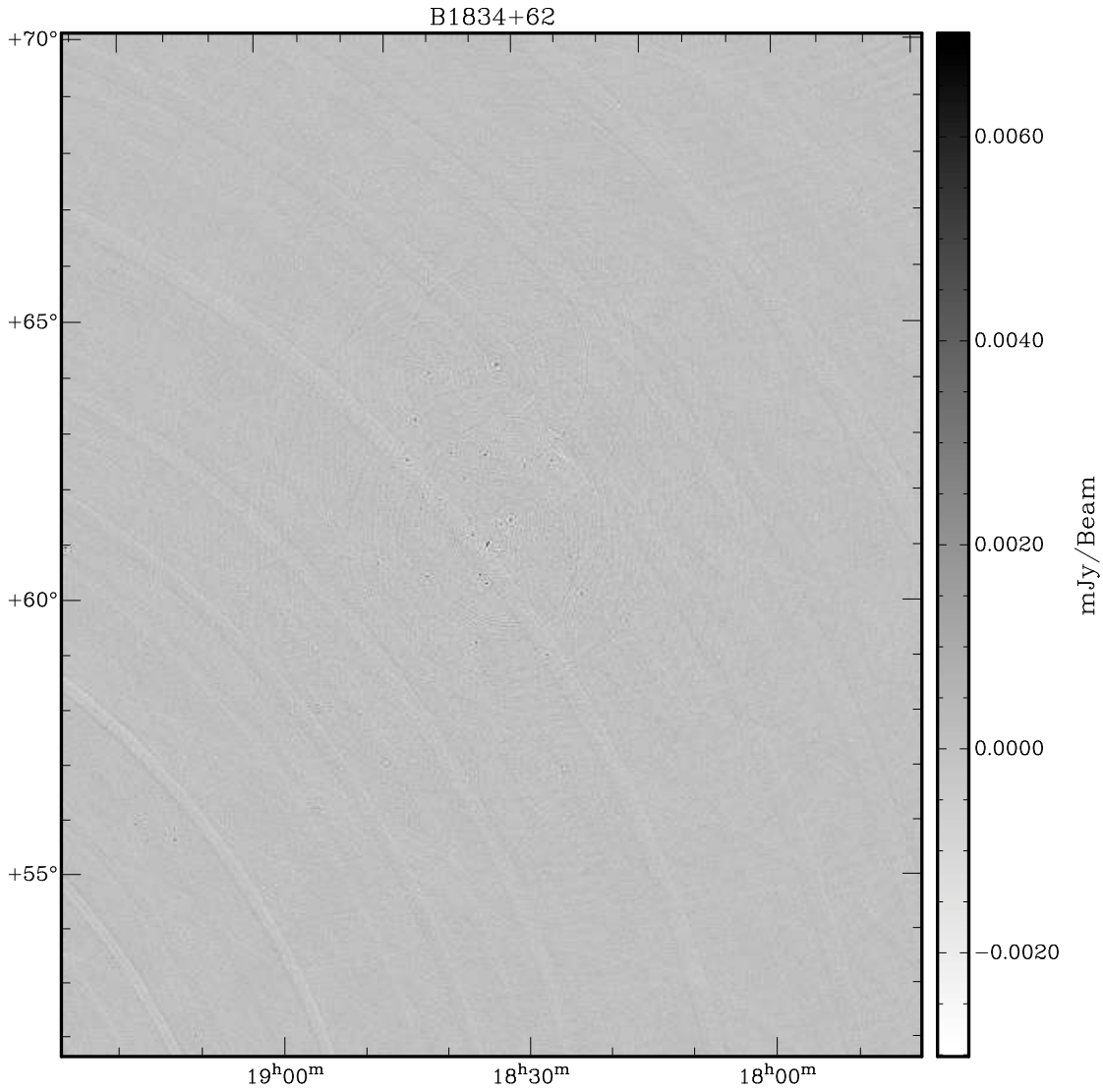
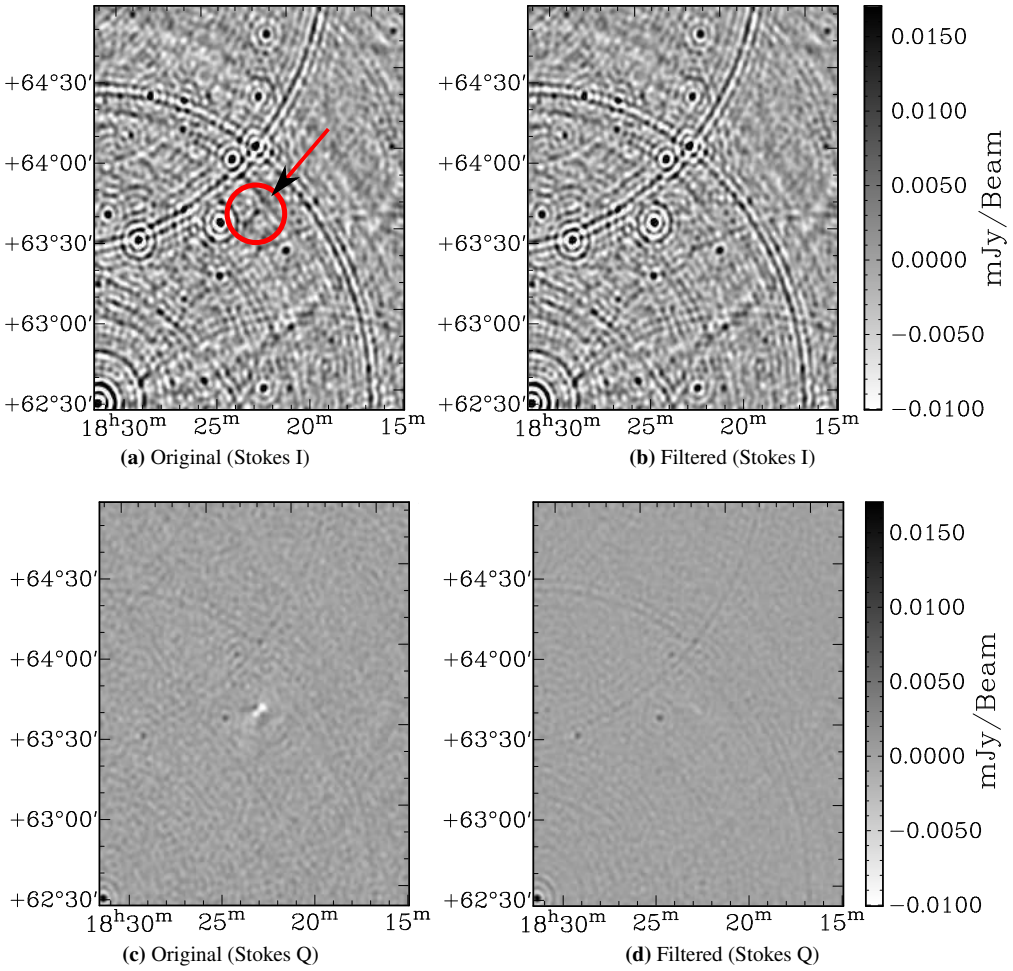


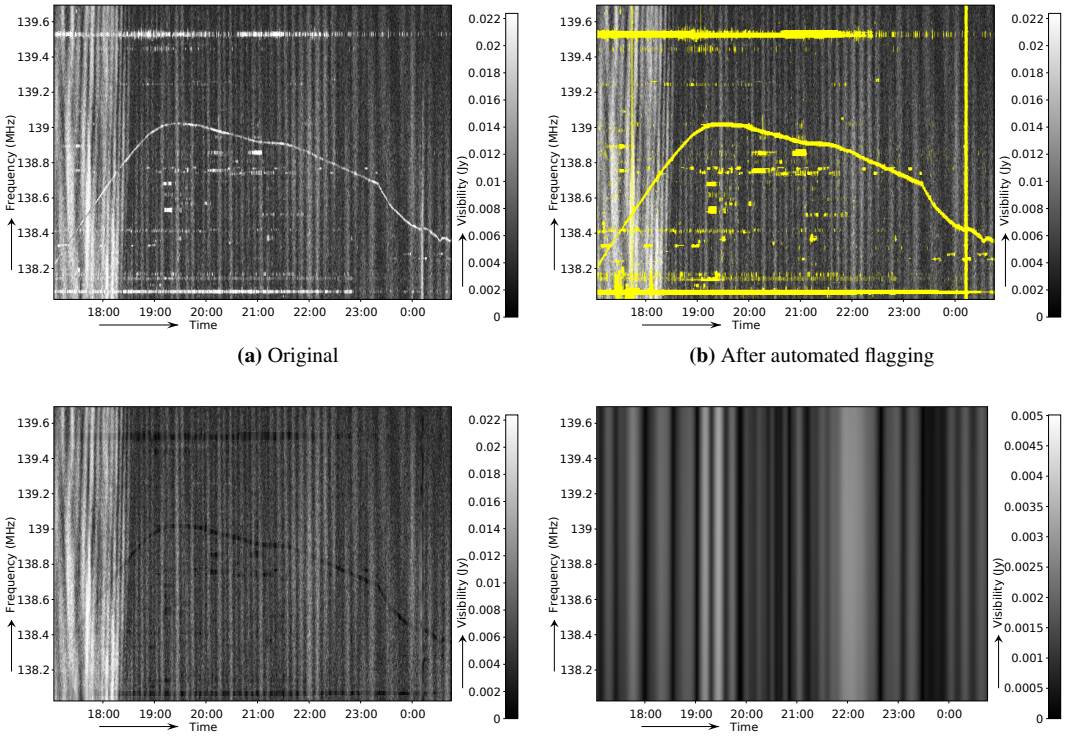
Figure 4.18: *A WSRT observation of field B1834 at 140 MHz containing three strong off-axis sources (see Fig. 4.15). WSRT can observe eight bands with 2.5 MHz bandwidth at this frequency, however, for this image, only one of the eight bands is used. The first two figures show Stokes I and the next two show Stokes Q. The second and fourth figures show the same data after low-pass filtering the set in both time and frequency directions. Even though the filter is limited by the small bandwidth, the suppression of the confusion noise of off-axis is significant. The effect is more detectable in the polarized images. Depending on which area is used for RMS calculation, the Stokes I and Q images show a noise reduction by a factor of 1.5–2 and 2–3 respectively. Moreover, a ghost of one of the off-axis sources (Cyg A) is strongly attenuated (see Fig. 4.19).*



(a) Original (Stokes I) **(b)** Filtered (Stokes I)

(c) Original (Stokes Q) **(d)** Filtered (Stokes Q)

Figure 4.19: *Enlargement of the central area of Fig. 4.18: Aliasing of off-axis source causes a ghost in the primary field, which is attenuated by the low-pass filter.*



(a) Original

(b) After automated flagging

(c) After interpolation of flagged areas

(d) After low-pass filtering in time and frequency directions

Figure 4.20: *A baseline for which the flags have been interpolated and filtered. Note that in panel (c), the RFI is still visible in the interpolated time-frequency plot by eye, because the interpolated area has a lower variance compared to the original.*

averaging or gridding, to only average over unflagged samples. The latter is similar to linear interpolation of the flagged samples, albeit the uv-position should be changed slightly because of the change of the centroid, to prevent bandwidth or time smearing. Before correlation or at high time and frequency resolutions, the difference between neighbouring samples is small enough that the error due to linear interpolation is small.

Since these methods have shown sufficient accuracy in practice, we have used a similar linear interpolation scheme: the data is indexinterpolating data interpolated by performing a Gaussian convolution on the unflagged data. The flagged samples in the original image are subsequently replaced with values from the convolved image. The result of this procedure on one of the WSRT B1834 set is given in Fig. 4.20. Normally, only data that are not flagged are used for imaging. These are the data from panel (b) in Fig. 4.20. To be able to filter the set, the flagged samples are interpolated as in panel (c). Tests using all baselines of the WSRT B1834 set show that the difference between imaging of the flagged set and the interpolated set in which all samples are used are small, as sources in the area of interest are changed less than 1 per cent. After low-pass filtering, we reapply the old flags. The rationale for this is to make sure that ranges that contain RFI are not used during further reduction, and the interpolated data is only used for filtering.

4.4.4 Computational requirements

For filtering the observation of B1834, we have used a regular desktop with a dual core Intel Core2 CPU running at 2.13 GHz and 2 GB of memory. Filtering the measurement set to create Fig. 4.18 in time and frequency direction, including interpolating the RFI samples, takes on the order of an hour on this machine, while we have been performing the filtering step with a non-optimized proof-of-concept script. This time is comparable with the time it takes to image the data set with the `lwimager`[2] that was used to create the images. The measurement set contains 91 baselines with 4 polarizations, 4300 time steps and 512 channels, and is 8 gigabytes in size. The IO takes about 15 per cent of the time. Hence, the computational requirements for filtering are not excessive. The method performs around an order of magnitude faster than demixed peeling as implemented in the LOFAR pipeline.

One complicating factor is that observations with a large number of frequency channels are often split up in many (sub-)bands. This is for example the case for LOFAR observations. Since the total data can become large, the sub-sequences are divided over several nodes on a cluster. Efficient synchronisation of the data between the nodes is not trivial, but by using a few nodes concurrently, we have been able to successfully filter a high resolution LOFAR observation within a few hours.

4.5 Discussion

4.5.1 Comparison of filter methods

The filters discussed were the single fringe filter (§4.2), the low-pass filter (§4.3.1 and §4.3.4) and the projected fringe filters (§4.3.2 and §4.3.3).

[2]The `lwimager` or Light Weight Imager is part of the casarest program, a subpackage of the Common Astronomy Software Applications package (`http://casa.nrao.edu/`)

The single fringe filter as proposed by Athreya and the introduced projected fringe filter can be applied before ionospheric calibration. We have shown that the single fringe filter is acceptable accurate for removing stable RFI sources, as long as the source to be removed is strong and reasonably constant. The filter should include the change in fringe frequency within the window as in Eq (4.5) for maximum accuracy. We do not observe stable, broadband RFI in LOFAR or WSRT that can be dealt with this method. To remove off-axis sources with the single fringe filter, an accurate model of the source is needed. In practical situations with non-constant sources, the fitting error exceeds 10 per cent and is therefore highly inaccurate in comparison to common ways to remove sources. It is therefore too inaccurate to be useful for off-axis source fitting.

One of the reasons for a projected fringe filter to be useful is that it requires no model, except for a direction to filter towards. However, the iterative projected fringe filter was shown not to be accurate enough and will in general remove little more than 50 per cent of the source's power. Hence, the iterative projected fringe filter provides little benefit when removing (celestial) off-axis sources. The projected fringe low-pass filter can remove a source completely, but has the unwanted effect of filtering part of the area of interest. However, this unwanted effect only occurs on a small part of the data; the further the source that is to be removed is from the area of interest, the smaller the area. A possible approach might therefore be to exclude the part of the data on which the fringe speed of the area of interest exceeds the filter speed. Subsequently, the data can be calibrated to first order, and the calibration solutions can be extrapolated to the excluded data. The method is about an order of magnitude faster than peeling and demixed peeling. This approach needs further research.

In contrast to the single and projected fringe filters, the use of the introduced low-pass filter lies mainly in removing off-axis sources. The low-pass filter in frequency will low-pass filter any structure in frequency direction, thus is probably only useful for multi-frequency synthesized imaging. In this situation, the frequency low-pass filter is an ideal tool to improve the signal to noise ratio of the area of interest after all calibration and subtraction of modelled sources has taken place, because it attenuates radial sidelobes. When structure in frequency direction is important, e.g., when performing spectrography, the method can not be applied. The frequency low-pass filter is not necessarily limited to application after calibration. Because the phases and amplitudes are reasonably stable in frequency direction, it can be assumed that filtering in frequency direction will not remove information essential for calibration – as long as all modelled sources are within the unfiltered area in image plane.

The low-pass filter in time might be less applicable for uncalibrated data, because it removes the high-frequency components introduced by quick phase or amplitude changes such as ionospheric changes. This problem is less relevant on longer baselines, because of the faster fringe speed: at λ=21 cm, a single degree off-axis source has a fringe duration of 17 min on a one kilometre baseline. The low-pass filter in time removes tangential sidelobes of off-axis sources, which implies that the sidelobe confusion noise in the area of interest is not directly attenuated. Nevertheless, this filter can be useful to reduce aliasing effects, such as removing an aliased ghost, where it is complementary to the frequency low-pass filter.

In case the low-pass filter in the time or frequency direction is applied before calibration, one should make sure that the filter does not introduce baseline-specific errors (closure errors), because these might cause self-calibration to fail. Since all presented filters are applied on individual baselines, this holds for all the filters. Although Athreya (2009) argues that fringe fitting does not introduce closure errors, that only holds if the fit is perfectly accurate. It is unclear if this is generally true, because the accuracy of the fit is dependent on the fringe rate, and therefore

baseline dependent. However, as long as the baseline-dependent error is small, self-calibration will benefit from the removal of the RFI source. We have not yet looked at calibrating filtered data, and this requires further research.

For low-pass filtering we have only looked at applying a rectangular windowed sinc convolution (truncated sinc), naturally imposed due to the finite time/frequency range. Especially when the window is small in comparison to the size of a fringe rotation, non-rectangular windows might improve efficiency. Different window functions can provide different trade-offs between the sidelobes and the steepness of the filter edge in the image plane: functions with a small resolution bandwidth, such as the rectangular function, will create a sharp edge that has ripples. On the other hand, functions with high sidelobe fall-off will create a smoother edge and will suppress the ripples better. An example of such a function is the Hann function (Harris, 1978).

It is harder to distinguish off-axis sources from on-axis sources in data that corresponds to specific areas in the uv-plane. The uv-areas for which this is the case, are areas at which the rotation angle of the uv-track is near the rotation angle of the off-axis source in the image plane. The reason for this is that the fringes of off-axis sources are slow in time direction in these uv-areas, and cannot be distinguished from the slow fringes of sources near the phase centre. Any method that tries to separate off-axis sources from on-axis sources, will consequently be less accurate in these areas. Unfortunately, off-axis sources cause sidelobes that interfere with the phase centre in these same areas, hence it is important to accurately remove the off-axis sources from these areas in order to achieve high dynamic ranges. Using frequency bandwidth to distinguish sources is necessary in these ranges. Many algorithms look at small bandwidths at a time. For example, most algorithms currently applied for LOFAR, such as demixed peeling or self-calibration, currently only use information from one or a few subbands at a time, while a LOFAR subband is only 200 kHz. To accurately separate off-axis sources with these algorithms, multiple subbands have to be combined together.

Low-pass filtering is an implicit effect of integrating and averaging that occurs in the standard pipeline of interferometers. The implications of that will be discussed in the next section.

4.5.2 Adverse effects of time and frequency averaging

To reduce the data volume, the correlation coefficients are integrated over time directly after correlation, and are sometimes further time averaged, for example after a RFI flagging procedure has detected corrupted samples, as is the default for LOFAR. When imaging, the visibilities are once more averaged for gridding, to be able to apply a fast Fourier transform (FFT). Nyquist's theory states that the original signal can be reconstructed as long as the sampling frequency is at least two times the highest frequency. Hence, in order not to lose information, the sampling frequency in time and frequency should be twice the fringe frequency of the source given by respectively Equation (4.9) and (4.15). In this section we will discuss two side effects of averaging: (1) the effect of low-pass filtering and (2) the effect due to aliasing.

When data is averaged, the highest frequency components can no longer be presented, and high frequencies are therefore removed from the data. The corresponding side effects of time and frequency averaging can be deducted from the low-pass filtering results. Since the amount of averaging is normally independent of the baseline size, i.e., all baselines will be averaged equally, an off-axis source will only be filtered on long baselines. This has been sketched in Fig. 4.21 for over-averaging in time direction. Over-averaging in frequency is similar, but in radial direction. For these reasons, the effect of time and frequency averaging is baseline dependent and

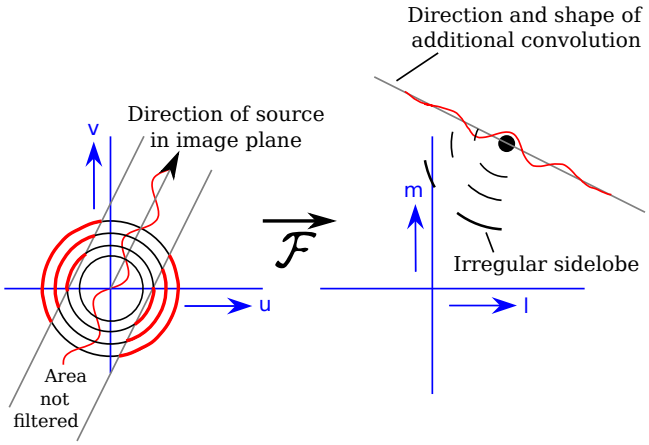


Figure 4.21: *The effect of over-averaging an observation in time direction, causing off-axis sources to be partly filtered on the long baselines. The left and right panel show respectively the uv-plane and the image domain.*

will contribute to closure errors. It is also a direction-dependent effect (DDE), since the distance of the source to the phase centre defines its fringe speed, and therefore the amount of attenuation. Therefore, different positions on the sky will be differently attenuated. Finally, averaging in time and frequency directions only complement each other partly: even by over-averaging the time and frequency directions significantly, the shorter baselines will still contain the source.

In an over-averaged set, a source will appear at its original location, but the source is fully present only in a subset of the baselines, which will cause it to have irregular sidelobes. Therefore, the source can not perfectly be removed with CLEAN, unless CLEAN is performed baseline by baseline or on smaller ranges of baselines, which is harder due to the low signal-to-noise ratio and dirtier point spread function of fewer baselines. Direction-dependent calibration might help, but directions that have been attenuated might still cause problems, e.g., in some antennas they will generate high gain solutions and therefore introduce noise. For these reasons, it is important to remove strong sources with fast fringe rates before time or frequency averaging in order to avoid their side lobes or added noise in the area of interest. This effect is most prominent in interferometric elements with a large field of view — a small element beam will naturally attenuate off-axis sources.

A second side effect of averaging comes from the fact that averaging is not a perfect low-pass filter, and will cause aliasing effects of high frequencies in the lower fringe frequencies. This will increase the noise generated by off-axis sources because they will not be filtered as much as possible. Time averaging can also distort sources of interest and can even generate ghost sources if off-axis sources have not been removed beforehand, as was seen in Fig. 4.19. To remove these effects, a low-pass filter can be used before down sampling the visibilities.

Fig. 4.22 shows the difference on a simulated observation between these two methods of changing the time resolution: (A) averaging the data; and (B) low-pass filtering the data followed by nearest neighbour interpolation. The down sampling factor was 3 and 25 for respectively the top and the bottom lines. The source is 30° from the phase centre and the simulated WSRT base-

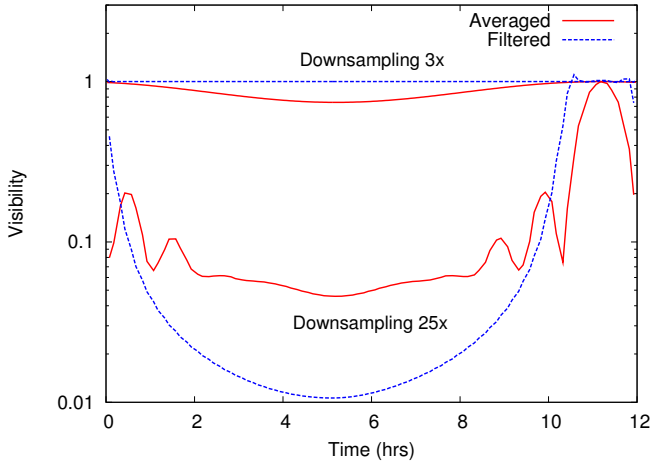


Figure 4.22: *Simulated effect of decreasing the time resolution with a factor of 3 and 25, on one single baseline with a single source, using two different methods: (A) averaging the data; and (B) low-pass filtering the data followed by nearest neighbour interpolation.*

line is 720 m, observing at 140 MHz and 62° declination. The maximum fringe speed is 30 Hz and the correlator integration time was 5 seconds. The figure demonstrates the non-ideal effect of averaging: sources which fringes beat with half the (new) Nyquist speed are attenuated up to 25 per cent, which does not occur in the filtered case. Moreover, a source that beats faster than the Nyquist speed (bottom lines) is better attenuated with less aliased sidelobes by the filtering compared to averaging. The attenuation effect of averaging quickly decreases when the source is closer to the phase centre, but is still on the order of one percent at one degree when three times averaged.

Time averaging has been used to average out RFI or other sources that have a high fringe rate. Athreya (2009) describes that RFI can be attenuated because of fringe stopping, although it is said that this is less effective at low frequencies. In Kogan and Owen (2010), the authors also describe averaging out RFI. As this article has shown, although the source itself is attenuated by averaging, and therefore helps calibration, we have shown it is better to perform an explicit low-pass filter before downsampling. The fringe frequency expressed in fringes/sample is almost always higher in time direction compared to frequency direction. Hence, if one relies on fringe stopping and correlator averaging to suppress RFI or off-axis sources, the noise in the area of interest is still affected by the source, since time averaging does not remove sidelobes in the direction of the phase centre (Fig. 4.7).

Time and frequency averaging are also part of the peeling algorithm, where it is used to filter off-axis sources. From the perspective of maximum attenuation, the baselines should be filtered with a filter size relative to the baseline length, instead of the de facto method of uniform averaging. This would suppress off-axis sources as much as possible, and all baselines would be filtered equally. However, care should be taken not to remove small temporal changes due to the ionosphere, that are needed for calibration. Fortunately, the ionosphere is typically stable in timescales

of several minutes.

It is well known that data averaging can cause tangential and radial smearing when averaging respectively the time and frequency dimension (Bridle and Schwab, 1999). The symptoms of bandwidth and time smearing can be intuitively explained with the results of this paper. As we have seen, the tangential and radial smearing happens because the longer baselines attenuate the source in a particular area of the uv-plane.

By using appropriate resampling techniques such as described in the paper, instead of time or frequency averaging which is used de facto, it is possible to reduce a data set to a smaller size with fewer artefacts. This might especially become important for arrays with a large field of view, long baselines and high data rates, such as LOFAR and the Square Kilometre Array (SKA), or high frequency interferometers such as the Atacama Large Millimeter Array (ALMA). In the future, it might be interesting to resample short baselines to lower resolutions, as these baselines contain the slowest fringe rates. This could further reduce the size of a measurement. However, operations such as calibration currently can not handle irregularly sampled data.

4.5.3 Relation to gridding

To perform the two-dimensional FFT transform used for imaging the data, the uv-tracks are normally gridded onto a uniform grid. Like averaging, this has the side effect of low-pass filtering the data: the maximal fringe speed in any direction is defined by the grid resolution. In contrast to time or frequency averaging, the filter size is relative to the length of the baseline: long baselines are gridded with a finer resolution compared to short baselines. The filtering effect of gridding is therefore equal to low-pass filtering in time and frequency: off-axis sources will be attenuated equally in all baselines. The somewhat counter-intuitive fact is that coarsely gridding the uv-plane will suppress sidelobes of off-axis (RFI) sources in the image plane, and might increase the signal-to-noise in the area of interest. Furthermore, frequencies that can not be represented in the UV-plane, correspond with sources that fall outside the image plane. Therefore, imaging only the area of interest is an efficient way of filtering off-axis sources not of interest.

Analogues to time and frequency averaging, the down-sampling before gridding is performed in a non-ideal way, for example by averaging[3]. From the conclusions in this work, we think aliasing effects are the reason why off-axis source that are not visible in the image plane, still produce sidelobes when performing regular gridding. The side effects are similar to the effects presented in Fig. 4.22, which shows that sources both faster and slower than the Nyquist frequency are not effectively attenuated. To solve this, the high fringe frequencies should be removed before gridding the data on the uv-plane. Again, the best way to do this is to low-pass filter the time and frequency directions before gridding.

4.5.4 Relation to other techniques

Although we have not tried combining this method with techniques such as (demixed) peeling, it is likely that the presented low-pass filters can complement these. There are two reasons for this:

- During calibration, the solutions are constrained by solving for antenna gains and by using the measurement equation. Calibration normally assumes solution constantness over short

[3]Most software packages do use more elaborate ways of sampling the data on the grid, for example by using prolate spheroidals.

time intervals and small bandwidths, and does not assume relations over the full time or frequency range. The low-pass filter uses the full time-frequency domain of a single baseline to disentangle sources. Therefore, it uses information that is complementary to the information used in standard removal techniques.

- The low-pass filtering techniques are not model-based. On the one hand, this allows direct and unbiased removal with less chance of inadvertently biasing towards an incorrect model, but on the other hand implies that there might not be enough data to separate sources in certain cases. Another difference with model based fitting, is that model based fitting can fail to converge due to an insufficient signal to noise level. Low-pass filtering is not limited by the signal to noise: due to the linearity of the Fourier transform, the result of low-pass filtering two time or frequency streams separately followed by averaging is equal to filtering the average of the two streams.

Because the low-pass filtering techniques do not involve non-linear fitting, they are much faster. If the filter techniques can be used for first order removal of off-axis sources, they might save a considerable amount of processing time. Investigation of the relation between the filter methods and other techniques will be the focus of further research. The LOFAR telescope provides a good test case for further research. Because of its large data volumes, its processing power is a considerable limitation, and it could potentially benefit a lot from faster source subtraction algorithms.

4.6 Conclusions & Outlook

We have shown that several filters can be used on individual baseline correlations to attenuate both off-axis sources and RFI sources in radio observations, thereby increasing the dynamic range of the observation. Because of the high performance of the filters, they are suitable for modern high-resolution observatories and can offer a complementary or alternative way to remove the sources. Especially the low-pass filter in the time and frequency directions are attractive, as they effectively attenuate all sources and their sidelobes outside a certain radius from the phase centre. However, they work less well on shorter baselines, and need a considerable bandwidth to remove sources effectively.

The next step is to further test the methods on other data, preferably with larger bandwidths, to see if the methods work in practice as well as in theory in other cases as well. Applying the filter on LOFAR data is attractive, because the off-axis source removal methods currently used are computationally intensive. With the large bandwidth of LOFAR, it would in theory be possible to, e.g., filter all sources outside 10 degrees even on baselines as short as 100 meters.

Chapter 5

The LOFAR radio environment

Based on:
"*The LOFAR radio environment*"
(Offringa et al., in preparation)

THE LOW-FREQUENCY ARRAY (LOFAR) is a new antenna array that observes the sky from 10–90 and 110–240 MHz. It consists currently of 41 (validated) stations, while 7 more are planned and more might follow. Of the validated stations, 33 stations are located in the Netherlands and 5 in Germany. Sweden, the UK and France contain one station each. A Dutch station consists of a field of 96 dipole low-band antennae (LBA) that provide the 10–90 MHz range, and one or two fields of in total 48 tiles of 4x4 dipole high-band antennae (HBA) for the frequency range 110-270 MHz. The international stations have an equal amount of LBA antennae, but 96 HBA tiles. For the latest information about LOFAR, we refer the reader to the LOFAR website[1].

The core area of LOFAR is located near the village of Exloo in the Netherlands, where the density of the stations is higher. The six most densely packed stations are on the Superterp, an elevated area surrounded by water. It is an artificial peninsula of about 350 m in diameter that is situated about 3 km North of Exloo. A map of LOFAR's surroundings is given in Fig. 5.1. Exloo is a village in the municipality of Borger-Odoorn in the province of Drenthe. Drenthe is mostly a rural area, and is, relative to the rest of the Netherlands, sparsely populated, with an average density of 183 persons/km^2 over 2,680 km^2 in 2011[2]. Nevertheless, the radio-quiet zone of 2 km around the Superterp is relatively small and households live within 1 km of the Superterp. The distance from households to the other stations is even smaller in certain cases. Therefore, contamination of the radio environment by man-made electromagnetic radiation was a major concern for LOFAR (Bregman, 2000; Bentum et al., 2008). Because this radiation interferes with the celestial signal of interest, it is referred to as radio-frequency interference (RFI). Such radiation can originate from equipment that radiates deliberately, such as citizens' band (CB)

[1]The website of LOFAR is `http://www.lofar.org/`.
[2]From the website of the province of Drenthe, `http://www.provincie.drenthe.nl/`.

radio devices and digital video or audio broadcasting (DVB or DAB), but can also be due to unintentionally radiating devices such as cars, electrical fences, power lines or wind turbines (Bentum et al., 2010).

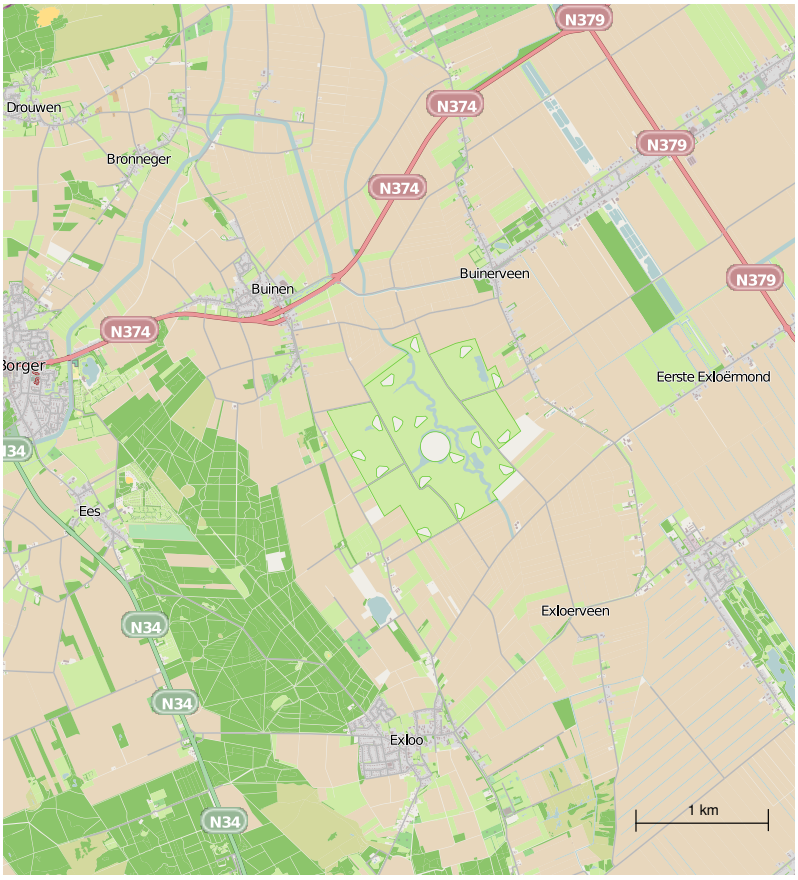


Figure 5.1: *Map of the LOFAR core and its surroundings. The circular peninsula in the centre is the Superterp. Several other stations are visible as well. (source: OpenStreetMap)*

During the hardware design phase of LOFAR, care was taken to make sure the signal would be dominated by the sky noise (Bentum et al., 2008). This included making sure that RFI would not drive the analogue-digital converters (ADCs) into the non-linear regime; applying steep analogue filters to suppress the FM bands and frequencies below 10 MHz; and applying strong digital sub-band filters to localize RFI in frequency. Optionally, an additional analogue filter can be turned on to filter frequencies below 30 MHz.

Now that LOFAR is largely finished, commissioning observations have started and preliminary results show that the LOFAR RFI strategy has worked out very well. For example, both the LOFAR EoR project (de Bruyn et al., 2011) and the LOFAR project on pulsars and fast transients (Stappers et al., 2011) report an excellent data quality. Moreover, new algorithms and a pipeline

have been implemented to automatically detect RFI with unprecedented accuracy (Offringa et al., 2010b,a). Preliminary results have shown that using these algorithms, only a few percent of the data is lost due to RFI (Offringa et al., 2010b).

In this chapter, we will study two 24 hr RFI surveys: one for the 30–78 MHz low-band regime and one for the 115–163 MHz high-band regime. We describe our general methods for analysing LOFAR data, and perform an extensive analysis of the two RFI observations. In Sect. 5.1, we start by describing the relevant technical details of the LOFAR observatory. In Sect. 5.2, we will describe the methods that are used to process and analyse the two sets. Sect. 5.3 describes the details of the RFI observations that are used in this chapter. In Sect. 5.4, a brief analysis of the spectrum allocation situation that is relevant for LOFAR follows. In Sect. 5.5 we will describe the observational results of the two RFI surveys. Those will be compared to other observations to assess whether they are representative in Sect. 5.6. In Sect. 5.7, we finish by discussing the results and making conclusions about the LOFAR RFI environment.

5.1 LOFAR

In this section, we will briefly describe the design details of LOFAR that are relevant for the impact of RFI. For further technical details, we refer the reader to van Haarlem et al., in preparation.

LOFAR consists of stations of clustered low-band and high-band antennae (LBA and HBA). Inside a station, the signal from dual polarization LBA antennae are amplified with a low-noise amplifier (LNA), and are subsequently transported over a coax cable to cabinet, which contains the receiver electronics. Here, the signal is band-pass filtered, digitized with a 12-bit ADC and one or more station beams are formed. The HBA antennae are processed by an analogue beamformer, which form the beams for a tile of four times four antennae. At the cabinet, digitized HBA station beams are subsequently formed from the analogue tile beams.

After beams have been formed, the HBA or LBA signals are split into 244 sub-bands of 195 kHz of bandwidth in standard imaging mode. Other modes can optionally be processed through different signal paths. The sub-bands are formed by using a poly-phase filter (PPF) that is implemented inside the station cabinet by using field-programmable gate array's (FPGA's). This allows for very flexible observing configurations (Romein et al., 2011). The 244 sub-bands are transported over a dedicated wide-area network (WAN) to a Blue Gene/P supercomputer in the city of Groningen. Currently, the samples are send in 16 bits. However, because the transfer rate is limited to about 3 Gbit/s, the transport limits the total observed bandwidth to 48 MHz. An eight and four bit mode are scheduled to be implemented in late 2012, which would allow the transfer of 96-MHz beams.

Once in Groningen, the BG/P supercomputer applies a second PPF that increases the frequency resolution with a factor of 256, yielding a resolution of 0.76 KHz. During this stage, the first of the 256 channels is lost for each sub-band, due to the way the PPF is implemented. Next, the BG/P supercomputer correlates each pair of stations, integrates the signal over time and a preliminary pass-band correction is applied (Romein, 2008), that corrects for the first (station level) poly-phase filter. Finally, the correlation coefficients are written to the disks of the LOFAR Central Processing II (CEP2) cluster.

The separation in sub-bands is used to distribute observations over the hard disks of the computing nodes on the CEP2 cluster. For storage of observations in imaging mode, LOFAR uses

the CASA[3] measurement set (MS) format. The first step of post-processing once the observation has been stored, is the RFI detection step. This step is performed by the AOFlagger pipeline that flags detected RFI, so that further processing steps such as the calibration step, will ignore RFI contaminated data. This step will be described in the next section. Following RFI mitigation, the steps that normally follow are (i) further averaging of the correlations to reduce the data volume; (ii) calibration; and (iii) finally the imaging.

5.2 Processing strategy

Processing an observation and acquiring an overview of the radio environment requires the detection of the RFI; collecting of the RFI statistics; and assessment of the quality of the remaining data. In the following subsection, we will address the detection strategy and the tools that we use for the detection. This will be followed by a description of the methods used to collect the statistics of the RFI and the data.

5.2.1 Detection strategy

For RFI detection, LOFAR uses the LOFAR AOFlagger pipeline that was described in Offringa et al. (2010b). At the time of writing, no changes were found necessary to alter the accuracy or sensitivity of the pipeline, but several optimizations were made to increase the speed of the flagger further. One of the changes was to use a more stable and faster algorithm for the morphological scale-invariant rank (SIR) operator (Offringa et al., 2012b), that finds samples that are likely contaminated by looking at their neighbouring samples. Another change was to implement several algorithms using the "streaming single-instruction-multiple-data extensions" (SSE) instruction set extension. The combined optimizations led to a decrease in the computational requirements of approximately a factor of 3, and the pipeline is now highly input-output (IO) dominated. To decrease the IO requirements, the pipeline was embedded in the next default processing step: the averaging step. Averaging is performed by the "New default pre-processing pipeline" (NDPPP) (Pizzo, 2012, §5), which can also carry out a few other steps, such as changing data alignment and changing the phase centre. The integration of the AOFlagger pipeline in NDPPP allows to read the raw data from disk only once.

The AOFlagger package[4] consists of three parts: (i) the library that implements the detection pipeline, to allow its integration in pipelines of other observatories and NDPPP; (ii) a stand-alone executable that runs the standard pipeline or a customized version; and (iii) a graphical user interface (GUI) that can be used to analyse the flagging results on a baseline-by-baseline basis and optimize the various parameters of the pipeline. The GUI was used intensively to optimize the accuracy of the pipeline. The GUI is also useful for adapting the strategy for data from other observatories. Once a strategy has been derived that works well on several individual baselines, the strategy can be exported and used with the library or the stand-alone flagger. This has led to the successful flagging of data from at least the Westerbork Synthesized Radio Telescope (WSRT)

[3]CASA is the Common Astronomy Software Applications package, developed by an international consortium of scientists under the guidance of NRAO.
Website: `http://casa.nrao.edu/`

[4]The AOFlagger package is distributed under the GNU General Public License version 3.0, and can be downloaded from `http://www.astro.rug.nl/rfisoftware`.

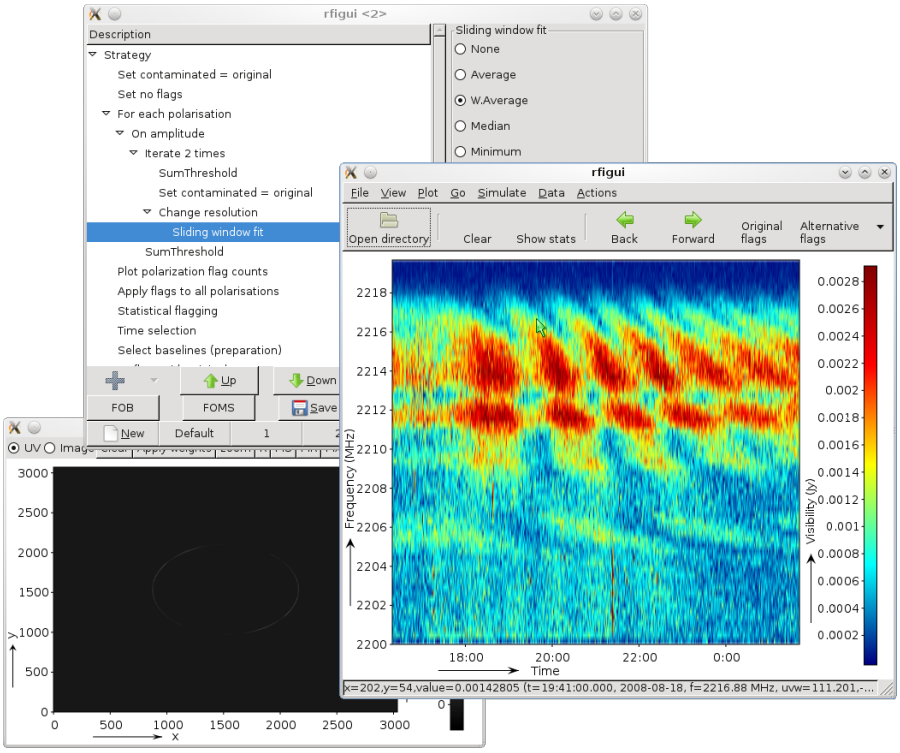


Figure 5.2: *The* `rfigui` *that can be used to optimize the pipeline steps and their parameters. The right window is the main window showing the spectrum of the selected baseline (in this case a WSRT S-band data set). The left bottom window shows the* uv *track that this baseline covers. The upper left window holds the script with the actions that are performed, which can be edited interactively.*

(Offringa et al., 2010a), the Giant Metrewave Radio Telescope (GMRT) (A. D. Biggs, personal communication, Sept. 2011) and the Australia Telescope Compact Array (ATCA).

For the data processing in this paper, we have not used NDPPP to average and/or process the data, but used the original full resolution sets and applied the stand-alone flagger.

5.2.2 RFI and quality statistics

Assessing the quality of observations that have a volume of tens of terabyte is not a trivial task. For example, if one wants to calculate the root mean square (RMS) of the data, all data has to be read from disk, and although this task can be distributed over the nodes, it still takes on the order of a few hours for large observations.

Our first effort to assess the RFI environment, was to produce a single informative sheet for each observation that summarizes the observation. This sheet contains a description of settings of the observation and four plots: (i) the amount of detected RFI over frequency; (ii) the amount of

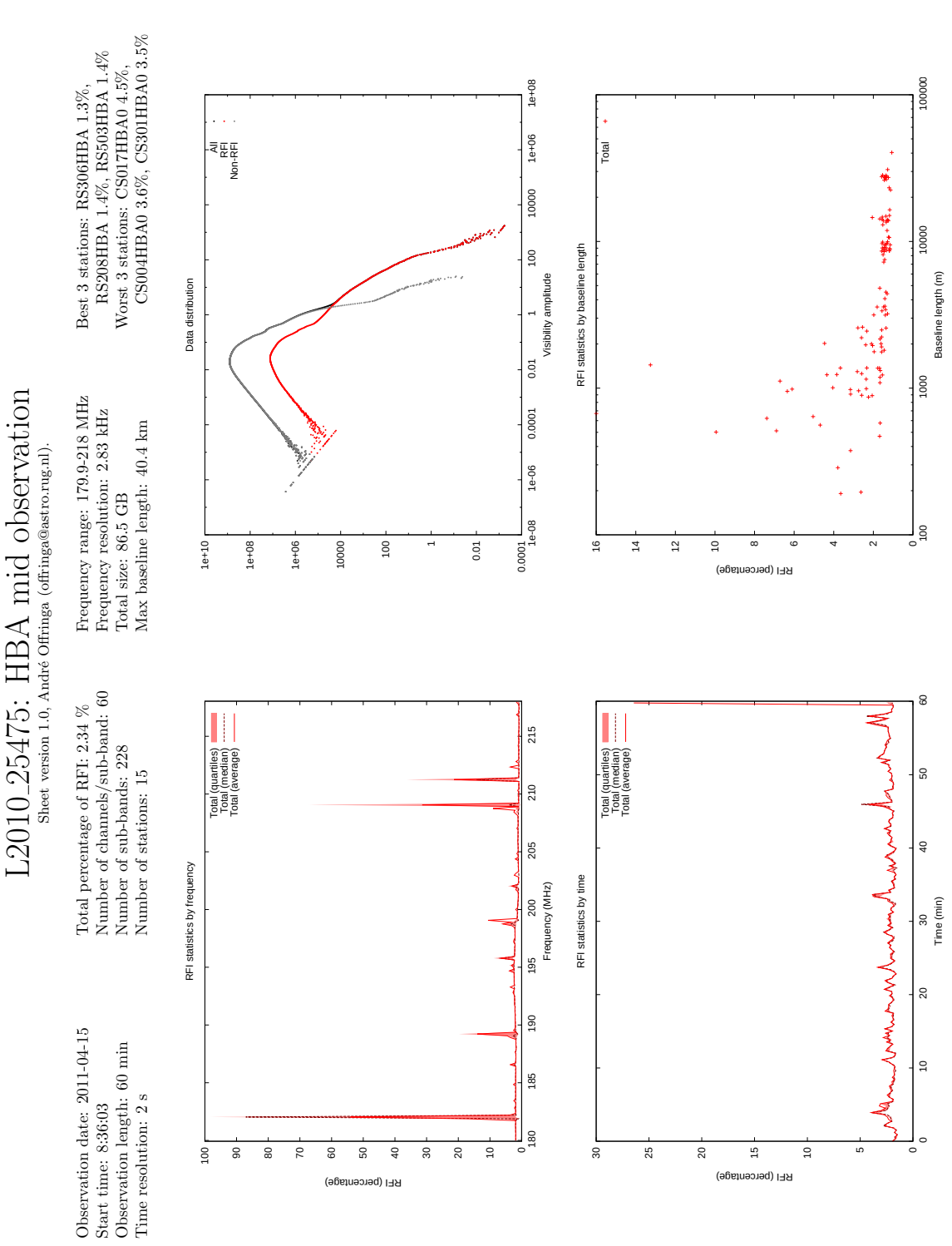


Figure 5.3: *An example of the RFI sheet as it was initially used to assess the RFI environment. The sheet describes a random observation.*

detected RFI over time; (iii) the amount of detected RFI as a function of the baseline length in which the RFI was found; and (iv) histograms of the flagged data, the non-flagged data and the sum of the two with logarithmic scales for both axes. A typical sheet of a random observation is given in Fig. 5.3. The information that is required to produce the plots is collected in the stand-alone flagger by adding an optional step to the detection pipeline. This allows to both flag the set and produce the plots with a single pass over the data.

The sheet provides a lot of useful data to assess the quality of the observation. The total percentage of RFI is a first indicator for a successful observation: percentages away from 2-5% denote some problem with the system during the observation. Faulty stations can be recognized from the baseline-length plot, while faulty sub-bands are displayed in the frequency plot. RFI can produce outlying sub-band statistics, though a failing cluster node can also produce this. If something significantly changes during the observation, the time plot will show this. Finally, the data histograms should show a Rayleigh distribution as long as the noise dominates the signal. Moreover, there should be a distinction between the RFI curves and the data curves. If something unexpected is seen, it might need to be followed up by looking at the full data sets to determine the cause.

The RFI sheet contains enough information to assess the RFI environment to first order — which was its purpose — but it holds no information about, e.g., the achieved signal to noise ratio and station or system temperature. These types of information are however closely related to the RFI statistics, and together they define the overall quality of the observation. The implementation of the sheet required manually gathering of the files produced by the RFI pipeline that hold the statistics. Because the observation statistics consists of information from 244 measurement sets, each measurement set was described in a few files. These files are subsequently fed to a script that combines them and produces the sheet.

Although the creation of the sheets can in principal be automated, a more generic solution was desired, that (i) combines the RFI statistics with other system statistics; and (ii) allows a standardized solution to read and display the statistics. Our solution consists of the following three parts: (1) a standardized storage format for the statistics; (2) software to collect the statistics; and (3) software to interpret the statistics. We will briefly describe each of these.

1. **The standardized storage format**: the format description of the so-called "quality tables" extension to the measurement set format (Offringa, 2011). The CASA measurement set format allows adding custom tables, and we used this possibility to add the statistics to the set. These statistics can be retrieved quickly without having to read the main data.

   Three statistics tables and one meta table are added to the measurement set. These tables contain the statistics as a function of frequency, time and baseline index. For LOFAR, the default is to add the total number of samples, the number of samples in which RFI has been detected, the sum of the samples and the sum of the squares of the samples. Together, these allow calculating the RFI ratio, the mean (signal strength) and the standard deviation as a function of time, frequency and baseline parameter. There are also statistics that describe the standard deviation of the noise, by subtracting adjacent channels. Since channels are only 0.76 kHz wide, the difference between adjacent channels should contain no significant contribution of the celestial signal, and this noise thereefore is a good measure of the celestial and receiver noise.

2. **Software to collect the statistics**: We have implemented software that collects the statistics and writes them in the described format to the measurement set. Since December 2011,

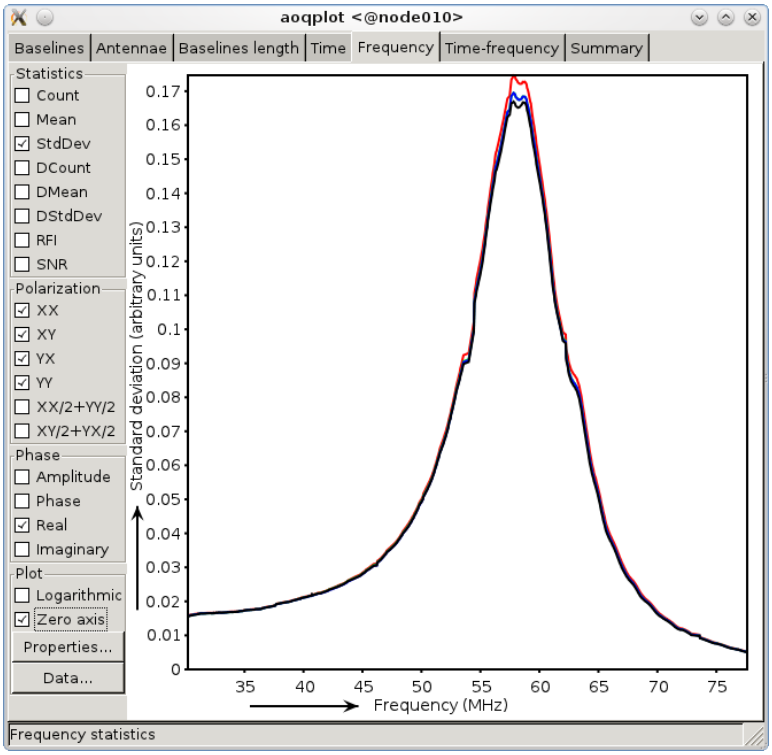


Figure 5.4: *The* `aoqplot` *tool that displays the statistics interactively. In this case it shows the standard deviation over frequency for a LBA observation.*

a statistics collector was added to the NDPPP averaging step. Because NDPPP performs various tasks that are required before further processing, NDPPP will be performed on most LOFAR imaging observations, and all observations will thereafter have these quality tables. NDPPP is slowed down by a few per cent because the statistics have to be calculated, which is acceptable. A stand-alone tool ("`aoquality`") is available in the AOFlagger package that can collect the statistics without having to run NDPPP.

3. **Software to interpret the statistics**: Finally, once the statistics are in the described format in the tables, tools are required to read and display the quality tables. Inside the AOFlagger package is an executable ("`aoqplot`") that performs this task: it takes either a single measurement set or an observation file that specifies where the measurement sets are located, and opens a window in which various plots can be shown and the selection can be interactively changed. An example of the plotting tool is shown in Fig. 5.4.

Table 5.1: *Survey data set specifications*

	LBA set	**HBA set**
Observation date	2011-10-09	2010-12-27
Start time	06:50 UTC	0:00 UTC
Length	24 hr	24 hr
Time resolution	1 s	1 s
Frequency range	30.1–77.5 MHz	115.0–163.3 MHz
Frequency resolution	0.76 kHz	0.76 kHz
Number of stations	33	13
Core	24	8
Remote	9	6
Total size	96.3 TiB	18.6 TiB
Field	NCP	NCP

5.3 Description of survey data

Table 5.1 lists the specifications of the two 24-h RFI surveys. The number of stations that were used in the HBA observation was limited to reduce the volume of the data. More stations were included in the LBA observation. The sets were observed at 0.76 kHz / 1 s resolution. Although this is the standard resolution at which LOFAR will observe in the future, the current commissioning observations are typically performed at a four times lower frequency resolution and two or three times lower time resolution to reduce their size. The observed field was the North Celestial Pole (NCP) in both sets. This field does not have a radio bright source and is therefore relatively easy to flag due to the absence of strong rapidly oscillating fringes.

Fig. 5.5 shows the locations of the stations that have been used in the two surveys. For the HBA set, the stations were selected to make sure that various baseline lengths were covered and the stations had geometrically a representative coverage. Due to the inclusion of additional core stations in the LBA set, the LBA set includes more baselines that are shorter.

We have used the LOFAR Epoch of Reionization (EoR) cluster (see Labropoulos et al., in prep.) to perform the data analysis. The first-time transfer of such large sets was challenging and helped us to develop the infrastructure further. In the LBA set, 6 sub-bands were corrupted due to two nodes on the LOFAR CEP2 cluster that failed during observing, causing six gaps of 0.2 MHz in the 48-MHz frequency span of the observation. At the time of these observations, the LOFAR CEP2 cluster was fairly new, and work is under way to fix its stability. Consequently, it is expected that such losses will be less common in future observations.

5.4 Spectrum management

In the Netherlands, the use of the radio spectrum is regulated by the government agency “Agentschap Telecom”, that falls under the Dutch Ministry of Economic Affairs, Agriculture and Innovation. This body maintains the registry of the Dutch spectrum users, which can be obtained from

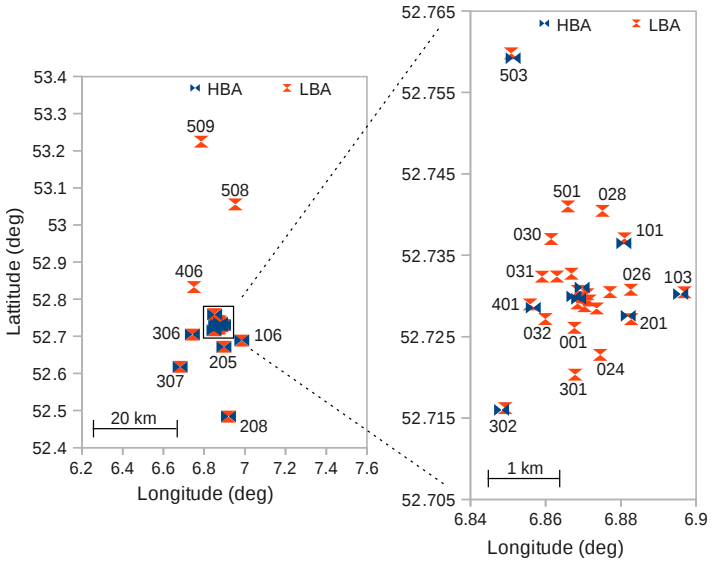


Figure 5.5: *Overview of the geometric distribution of the stations used for the RFI survey. Numbers next to the station symbols denote the station numbers.*

their website.[5]

The other countries that participate in the International LOFAR Telescope have similar bodies, and the Electronic Communications Committee[6] (ECC), a component of the European Conference of Postal and Telecommunications Administrations (CEPT), registers the use of the spectrum on European level. Most of the strong and harmful transmitters are allocated in fixed bands for all European countries, such as the FM radio bands, satellite communication, weather radars and air traffic communication. However, even though the allocations of the countries are equivalent, the usage of the allocated bands can differ. For example, several ranges of 1.792 MHz in the range 174–195 MHz are registered as terrestrial digital audio broadcasting (T-DAB) bands by the ECC. This range is correspondingly allocated to T-DAB both in the Netherlands and in Germany. However, these bands are currently used in Germany, yet not in the Netherlands. The range of 216–230 MHz is however actively used for T-DAB in the Netherlands. This range corresponds with T-DAB bands 11A–11D and 12A–12D, each of which is 1.792 MHz. These transmitters are extremely harmful for radio astronomy. Because they are wideband and have a 100% duty cycle and band usage, they do not permit radio observations. Digital video broadcasts (DVB) are similar, but occupy the range 482–834 MHz (UHF channels 21–66). They are therefore outside the observing frequency range of LOFAR.

A short list of services with their corresponding frequencies is given in Table 5.2. Only two small ranges are protected for radio-astronomy. The first range is the 37.5–38.25 MHz range. This

[5]The website of the Agentschap Telecom from which the spectrum registry can be obtained is `http://www.agentschaptelecom.nl/`.

[6]The website of the Electronic Communications Committee, which registers spectrum usage on European level, is `http://www.cept.org/ecc`, office: `http://www.ero.dk/`.

Table 5.2: *Short list of allocated frequencies in the Netherlands in the range 10–250 MHz (source: Agentschap Telecom)*

Service type	Frequency range(s) in MHz
Time signal	10, 15, 20
Air traffic	10–22, 118–137, 138–144
Short-wave radio broadcasting	11–26
Military, maritime, mobile	12–26, 27–61, 68–88, 138–179
Amateur	14, 50–52, 144–146
CB radio	27-28
Modelling control	27–30, 35, 40–41
Microphones	36–38, 173–175
Radio astronomy	38, 150–153
Baby monitor (portophone)	39–40
Broadcasting	61–88
Emergency	74, 169–170
Air navigation	75, 108–118
FM radio	87–108
Satellites	137–138, 148–150
Navigation	150
Remote control	154
T-DAB	174–230
Intercom	202–209

range is e.g. useful for observing the Sun and the Jupiter atmosphere. The second range is the 150–153 MHz range. Although the 10–200 MHz range is mostly allocated to other services, many of these — such as baby monitors — are used for short distance communication, and are therefore of low-power. In addition, services such as the Citizens' Band (CB) radio transmitters have a low duty cycle (especially during the night) and individual transmissions are of limited bandwidth. The most problematic services for radio astronomy are therefore the FM radio (87.5–108 MHz), T-DAB (174–230 MHz) and the emergency pager (169.475–169.4875 and 169.5875–169.6 MHz) services. The FM radio range is excised by analogue filters. The emergency pager was found to be the strongest source in the spectrum, and the LOFAR signal path was designed to be able to digitize its signals correctly.

Around the LOFAR core, a radio-quiet zone has been established that is enforced by the province of Drenthe. The area is split into two zones. The inner zone of 2 km diameter around the core enforces full radio quietness. A "negotiation zone" with a diameter of about 10 km around the core requires negotiation before transmitters can be placed.[7]

[7]The radio quiet zones are marked on "Kaart 12 — overige aanduidingen" of the environment plan of Drenthe.

5.5 Results

In this section, we will discuss the achieved performance of the flagger, look at the RFI implications of the surveys individually and analyse their common results.

5.5.1 Performance

The EoR cluster that was used for flagging consists of 80 nodes with two hyperthreaded quatro-core cpu's, 12 GB memory per node and 2 or 3 disks of approximately 2 TB size each. The cluster is optimized for computational intensive (GPU) tasks, such as advanced calibration and data inversion. Because it has relatively slow disks that are not in a redundant configuration (such as RAID), the cluster is not ideal for flagging, as flagging is computationally conservative but IO dominating. To make sure the flagging would not interfere with computational tasks that were running on the cluster at that time, we chose to use only 3 cpu cores, thus a ratio of 3/16 of the entire computational power of the cluster. Flagging the 96 TiB observation took 40 hours, of which 32 hours were spend on reordering the observation, which consists only of reading and writing to the hard disks necessary for flagging, and the remaining 8 hours were the actual flagging.

5.5.2 LBA survey

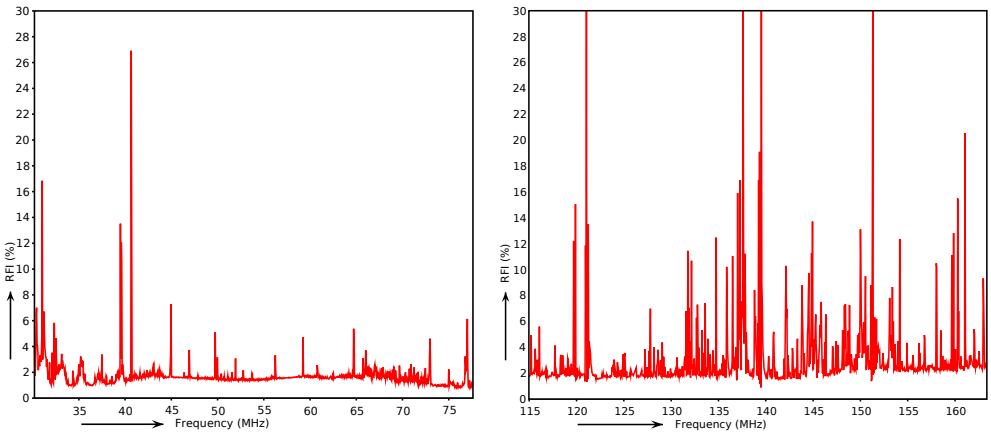


Figure 5.6: *The detected RFI occupancy spectra for both RFI surveys. Each data sample in the plot contains 48 kHz of data.*

The default flagging pipeline found a total RFI occupancy of 2.24% in the LBA survey. However, we found the flagger had a small bias. Because the sky temperature changes due to the setting of the Milky Way, the standard deviation of the data changes over time. The flagger applies a fixed sensitivity per sub-band and per baseline, and therefore does not take into account changes over time. This is not an issue for short observations of less than two hours, because then the sky temperature does not significantly change. However, on long observations in which

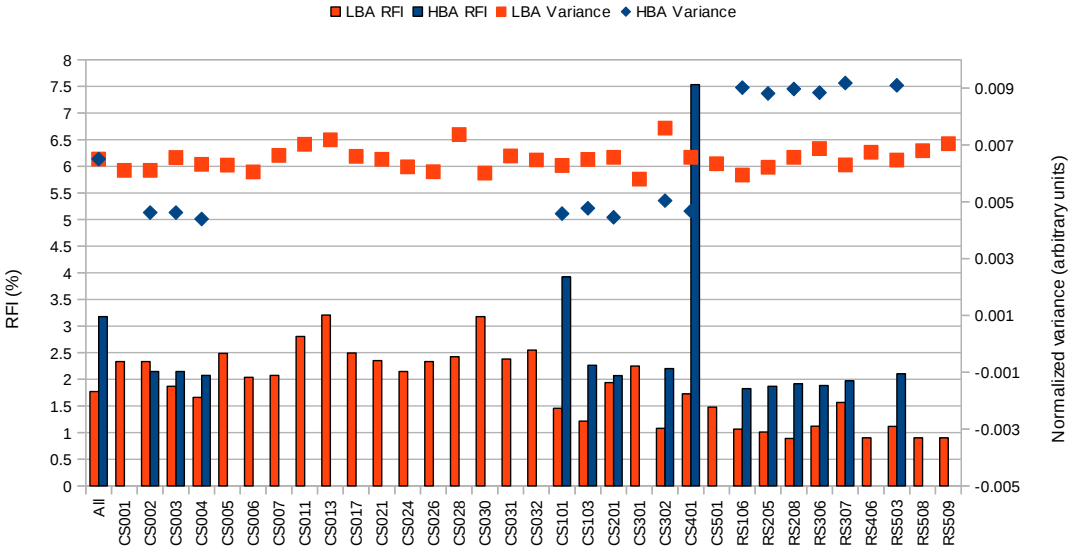


Figure 5.7: *The detected RFI percentages and the data variances per station.*

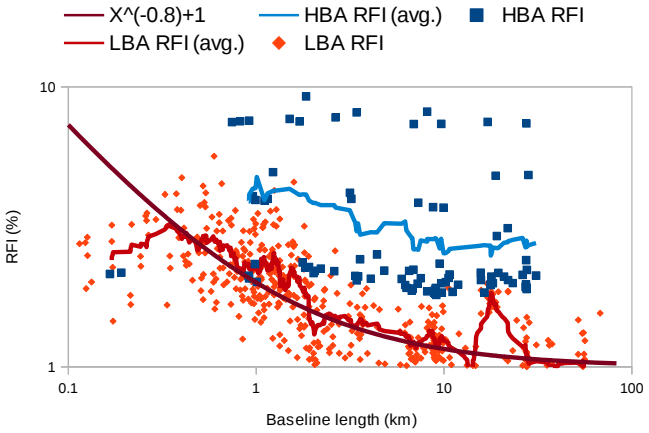


Figure 5.8: *RFI levels as a function of baseline length. Both axes are logarithmic. The local average shows the trend of the points.*

the sky temperature dominates the noise level, the flagger produces more false positives when sky temperature is higher and more false negatives when the sky temperature is lower.

Unfortunately, correcting for this effect requires an accurate estimate of the sky temperature, which in turn requires the interference to be flagged. Therefore, after the first flagging run, we have applied a second run of the flagger on normalized data. In the normalized data, each time step was divided by the standard deviation of the median timestep in a window of 15 minutes of data, thereby assuming that the first run has removed the RFI. The calculation of the standard deviation per timestep was performed on the data from all cross-correlations. Therefore, this procedure results in a very stable estimate. It is also possible to calculate the standard deviation or median of differences over a sliding window during the first run and base the detection thresholds on this quantity, but this does not match well with the SumThreshold method, which is crucial for the accuracy of the flagger.

After having corrected for the changing sky temperature, the detected RFI occupancy is 1.77%. The RFI occupancy over frequency is plotted in Fig. 5.6, while Fig. 5.7 shows the percentages of flagged data per station. The stations with higher station numbers are generally further away from the core, and therefore provide longer baselines. The remote stations (RS) are furthest away and additionally have more high-band antennae. Fig. 5.7 shows that the stations closer to the core generally have a lower amount of RFI, and by plotting the RFI as a function of baseline length as in Fig. 5.8, it can be seen that the RFI decreases as a function of baseline length for lengths > 300 m, and closely follows a power law that asymptotically reaches $\sim$1.0%.

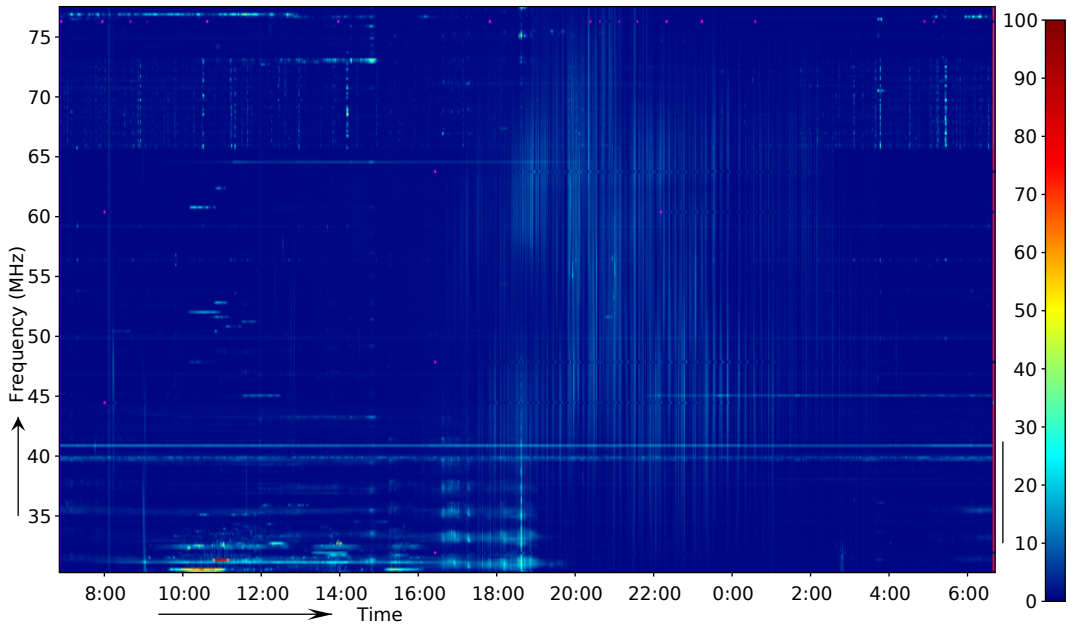


Figure 5.9: *The dynamic spectrum of RFI occupancy during the LBA survey*

The LBA set contains many broadband spikes between 18:00–0.00 hr. These are detected by the flagger as RFI, and therefore visible in the dynamic RFI occupancy spectrum of Fig. 5.9. An

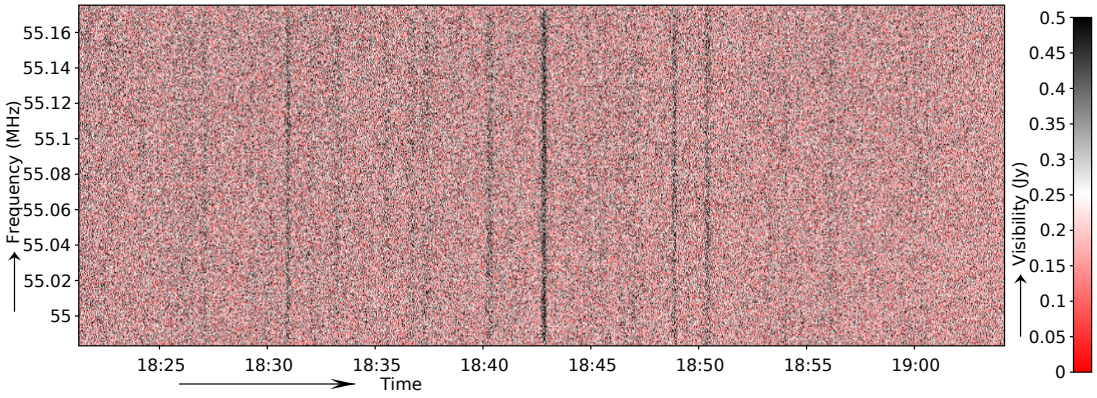


Figure 5.10: *Data from the LBA 4 km baseline CS001 × RS503 at high frequency resolution, showing strong fluctuations of 1–10 s. The flagger detects these as RFI.*

example of the spikes at high resolution on a 4 km baseline is shown in Fig. 5.10. Individual spikes affect all samples for 1–10 seconds. Despite the already relative long baseline of 4 km, these spikes have evidently not yet become incoherent. On the 56 km baseline CS001 × RS509, the spikes can not be seen in the time-frequency plot, but some of them are still detected by the flagger because of an increase in signal to noise in these time steps. It is assumed that they are strong ionospheric scintillations of signals from Cassiopeia A, because they correlate with its apparent position. Cas. A is 32° away from the NCP, which is the phase centre. Cygnus A might also cause such artefacts, but is 50° from the phase centre.

At the very low frequencies, around 30 MHz and 17:00–18:00 hrs, a source is visible that shows many harmonics. A high resolution dynamic spectrum is shown in Fig. 5.11. It is likely that this source has saturated the ADC. Nevertheless, its harmonics are flagged accurately, and it causes no visible effects in the cleaned data.

5.5.3 HBA survey

The analysis of the HBA survey shows a slightly higher RFI ratio with a total detected amount of 3.18%. The noisier RFI occupancy spectrum of the HBA in Figs. 5.6 and 5.12 also confirms that the RFI is more contaminated by interference than the LBA. However, as can be seen in Fig. 5.7, almost all stations have less than 2.5% RFI. Stations CS101HBA0 and CS401HBA0 are the only two exceptions, with respectively 3.9% and 7.5% RFI, and are also a cause of the higher level of RFI compared to the LBA survey. Despite the larger fraction of RFI in stations CS101HBA0 and CS401HBA0, the data variances of these are similar to the other stations. This suggests therefore the presence of local RFI sources near these two stations, which have successfully been taken out by the flagger. This is incidental: recent observations show normal detected RFI ratios of less than 3%. Fig. 5.7 also shows that the variances of the remote stations is higher. This is because the data is not calibrated, and these stations contain twice as many antennae.

As in the LBA case, the HBA RFI detected ratios are similarly affected by the changing sky temperature. Because of the computational costs involved with flagging a set of this size, we

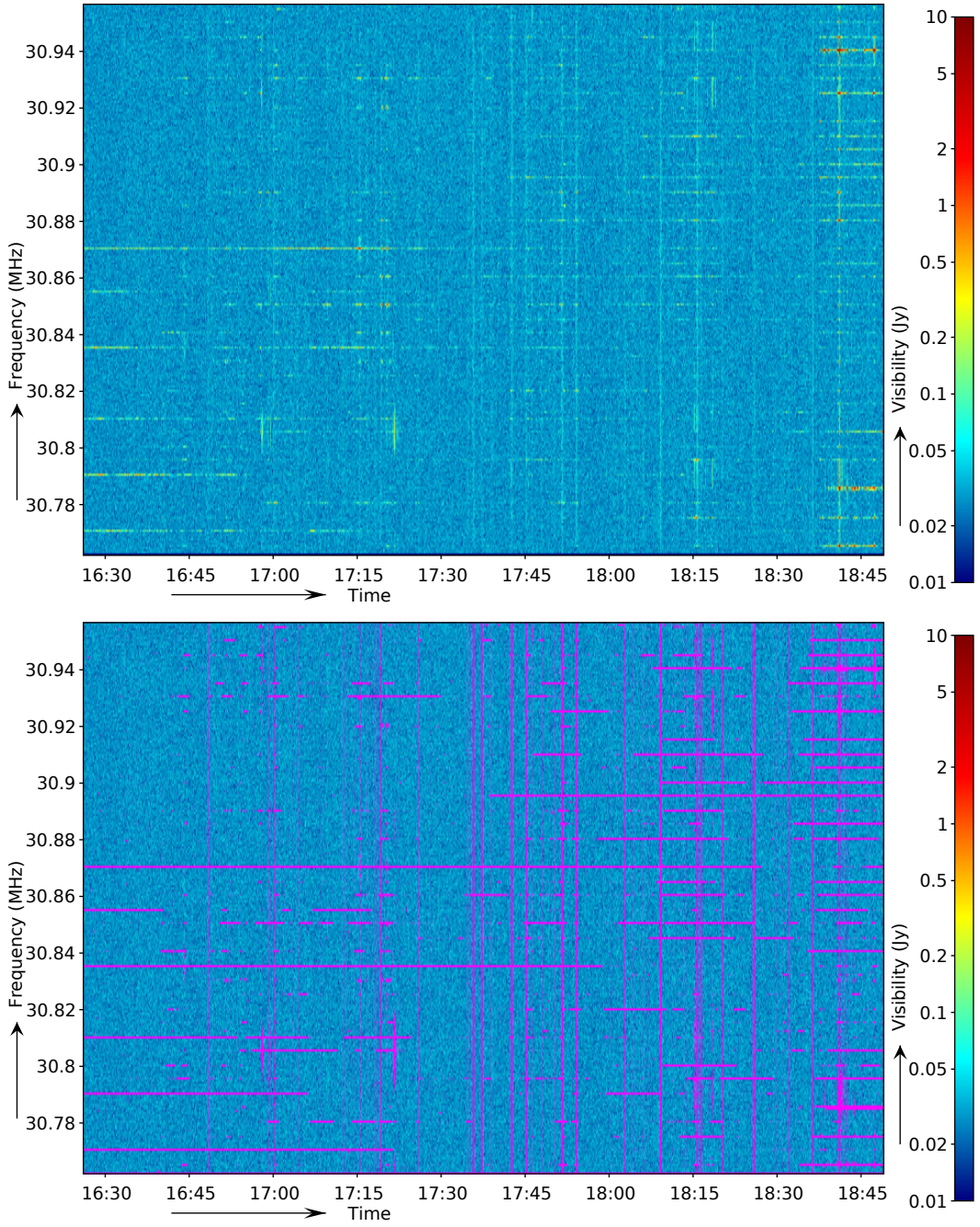


Figure 5.11: *A dynamic spectrum of data from one sub-band of the LBA survey, formed by the correlation coefficients of baseline CS001 × CS002 at the original frequency resolution of 0.76 kHz. The displayed sub-band is one of the worst sub-bands in terms of the detected level of RFI. The top image shows the original spectrum, while the bottom image shows with purple what has been detected as interference.*

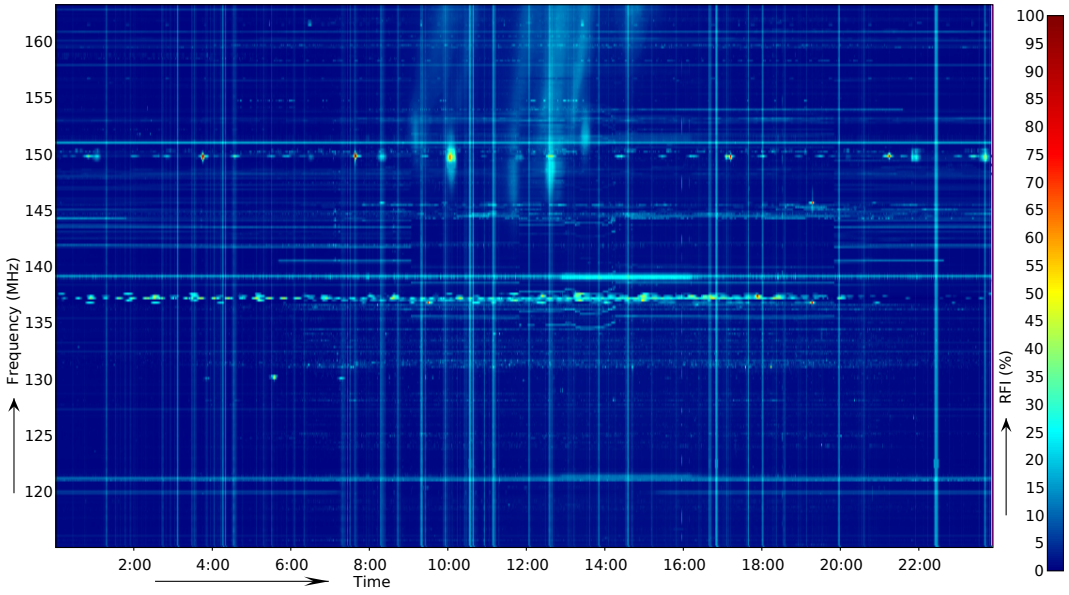


Figure 5.12: *The dynamic spectrum of RFI occupancy during the HBA survey*

have not corrected the bias with a second run. Flagging an individual sub-band shows a similar decrease of about 0.5% in detected RFI.

It is harder to assess whether the level of RFI decreases significantly on longer baselines, as the fewer number of baselines cause a rather noisy estimate of the curve in Fig. 5.8, but the general trend of the average curve follows the trend of the LBA reasonably.

5.5.4 Overall results

After the automated RFI detection, there are generally no harmful interference artefacts left in the data. The variance over frequency and time are displayed in respectively Fig. 5.13 and Fig. 5.14. While the HBA variances look clean in most frequencies, there are a few spikes of RFI that evidently have not been detected. These look like sharp features in the full spectrum, but are in fact smooth features when looking at full resolution. Because they are smooth at the raw sub-band resolution, the flagger does not detect them as RFI. Although there are interference artefacts visible in the HBA spectrum, after detection the data can be successfully calibrated and imaged. Nevertheless, a possible second stage flagger to remove any residual artefacts will be discussed in §5.7. The LBA variances over frequency contain no visible RFI artefacts at all.

The HBA spectrum contains a clearly visible ripple of about 1 MHz. This has been identified as the result of reflection over the cables, resulting from an impedance mismatch in the receiver unit. In fact, a similar phenomenon occurs in LBA observations, but because of the steeper frequency response and because not all LBA cables are of the same length, it is less apparent. The reflection is also less strong in the LBA, due to the better receiver design. Nevertheless, a Fourier transform of the LBA variance over frequency shows slight peaks at twice the delays of the cables.

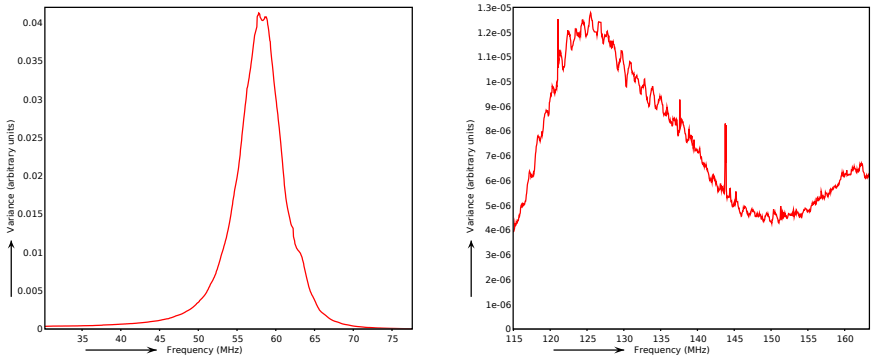


Figure 5.13: *The post-flagging spectra of data variances for both RFI surveys. The dominating effects are the antenna frequency response and sky noise.*

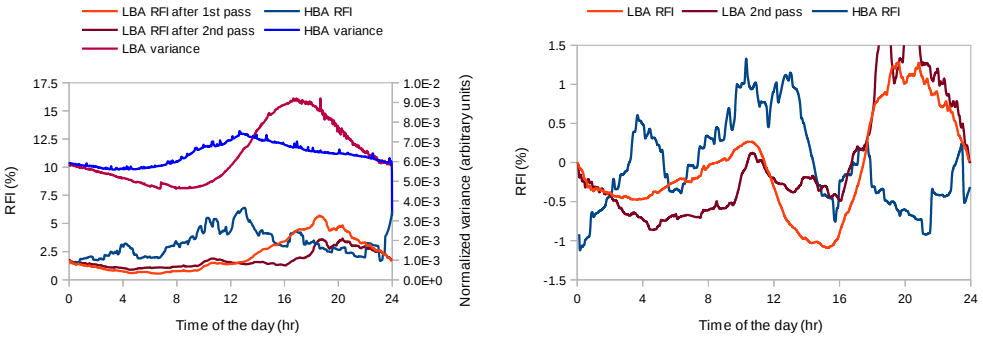


Figure 5.14: *RFI levels and variances as function of the time of the day. The RFI percentages are smoothed in both figures. In the figure on the right, the difference between day and night is enhanced: An estimate of the contribution of the sky noise is subtracted from the first runs. The LBA second pass is centred on the zero axis.*

5.5.5 Day and night differences

Fig. 5.14 shows variance and RFI occupancy as a function of the hour of the day in UTC. Local time is UTC+1. One might expect a lower RFI occupancy during the night, thus during 11.00–6.00 hrs UTC. However, after one flagging pass the data is highly dominated by the changing of the sky. Moreover, the LBA data contains artefacts of Cassiopeia A, which causes some peaks in the data due to strong ionospheric scintillation between 18.00–0.00 hrs. The second pass LBA data shows a small RFI occupancy decrease at night, especially between 2.00–8.00 hrs UTC of about 0.5%.

The right plot in Fig. 5.14 shows RFI occupancies in which the effect of the changed variance on the first-pass statistics has been estimated and taken out. This plot is derived by subtracting a linearly scaled version of the variance curve from the RFI occupancies, such that the mean and variance of the residual are minimized. If the interference occupancy increases during daytime, the effect should be enhanced by this method. However, the biasing effect of the sky temperature is not removed completely, because the detection rate is not completely linearly dependent on the variance of the data. After applying this method, the first pass residuals are relatively small with a total range of variation of about 2%, and this variation is likely the result of the changing of the sky that has not been subtracted out correctly. There is no obvious other relation visible. This implies that there is no significant relation between the hour of the day and the RFI occupancy due to less activity at night. This is also evident in the dynamic spectra of RFI in Figs. 5.9 and 5.12, which show no obvious increase or decrease of transmitters during some part of the day, although some transmitters start and end at random times. In a few cases, the starting of a transmitter at a certain frequency coincides with the termination of a transmitter at a different frequency, suggesting that some transmitters hop to another frequency. An example can be seen in Fig. 5.12, where several transmissions between 140–145 MHz end at 9 AM UTC, while at the same time several transmissions around 135–140 MHz start.

To further explore the possibility of increased RFI during daytime of the HBA set, we have performed the same analysis on a 123–137 MHz subset of the HBA observation. There are two reasons that the difference between day and night might be better visible in this frequency bandwidth: (i) the visual peaks of detected RFI that correspond to the Sun all have a frequency higher than 145 MHz; and (ii) this band corresponds to air traffic communication, which is less used during the night. Nevertheless, we still do not see an increase of RFI in this subset of the data, apart from the rise of detected RFI due to the fact that the flagger finds more RFI during time steps with higher variance.

In summary, any effect of increased activity during the day is not significant enough to be identifiable in the detected occupancies of either the LBA or the HBA data set. The post-flagging data variances are dominated by celestial effects, i.e., the Sun, the Milky Way or Cassiopeia A, and contain no clear signs of a relation between day and night time either.

5.5.6 Resolution & flagging accuracy

The frequency and time resolution of observations affect the accuracy of the interference detection. What the size of this effect is, is however not known. To quantify this, we have decreased the frequency resolution of the HBA RFI survey and reflagged the set. Subsequently, the resulting flags were compared with the flags that were found at high resolution. The original high resolution flags were used as ground truth.

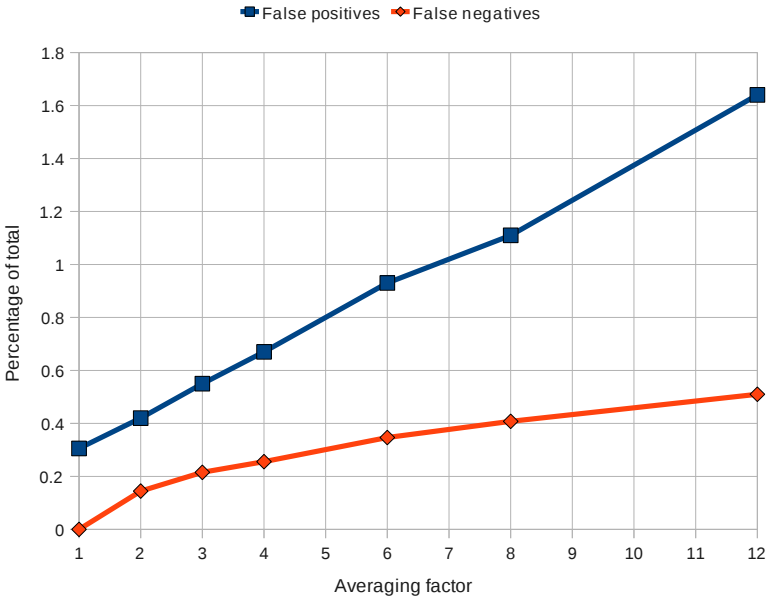


Figure 5.15: *The effect of frequency resolution on detection accuracy*

We found that the level of false positives is approximately linearly correlated with the resolution decrease factor. Unfortunately, false positives in our ground truth will likely show up as false negatives in the lower resolution detections. Therefore, the false positives for the ground truth data were determined by extrapolating the false-positives curve of the sets with decreased resolution. This yields a false-positives rate of 0.3%, which subsequently has been subtracted from the false negatives. The resulting curves after these corrections are plotted in Fig. 5.15.

Because the test is very computationally expensive, we have not performed the same test on the LBA survey or for the time direction. Tests on small parts of the data show that decreasing the time resolution results in similar false-negatives curves compared with decreasing the frequency resolution, although it causes about 20% less false positives. Therefore, from the RFI detection perspective, it is slightly better to have higher frequency resolution compared to higher time resolution at LOFAR resolutions. It should be further investigated whether the small amount of data was representative enough to draw generic conclusions.

5.5.7 False-positives ratio

If we assume that the least contaminated sub-bands in Figs. 5.9 and 5.12 are completely free of RFI on the long baselines, they can be used to determine the false-positive ratio of the flagger. For the LBA set, we selected the 4 km baseline CS001 $\times$ RS503 and the 56 km baseline CS001 $\times$ RS509 of one the best centre sub-bands at 55 MHz. In the 4 km baseline a total RFI ratio of 0.75% was detected, while the 56 km baseline shows 0.73% RFI. However, the 4 km baseline contains some

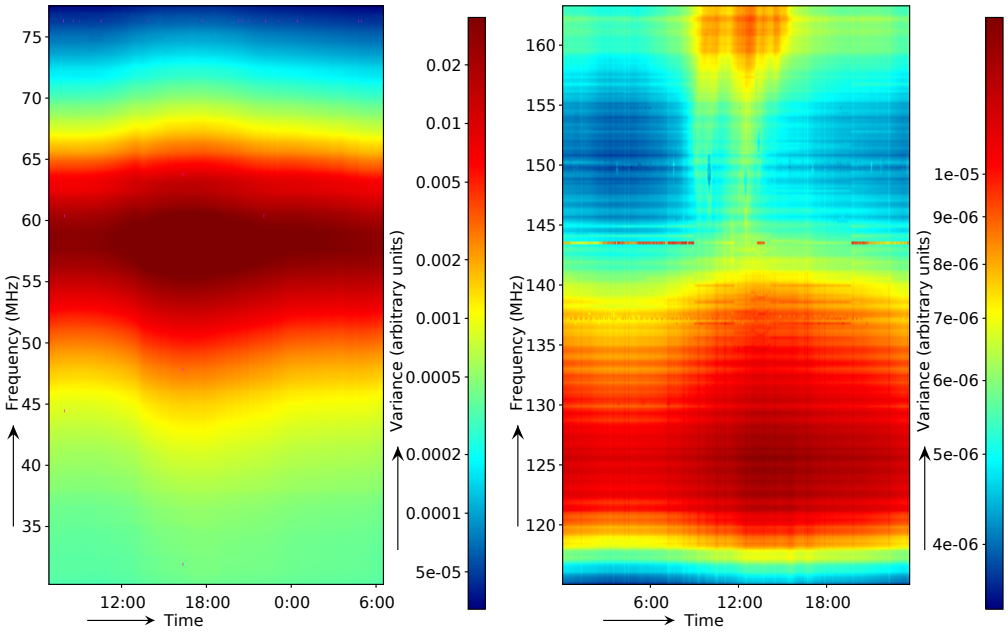


Figure 5.16: *The variance over time and frequency during the surveys. In the LBA set, no residual RFI is visible, also not when inspection the data at higher resolutions. A few purple dots can be seen in the data, which denotes missing data. The HBA set does show a few weak RFI residuals.*

broadband spikes around 18:40 hrs, as shown in Fig. 5.10. On the 56 km baseline CS001 × RS509, the spikes can not be seen in the time-frequency plot, but some of them are still detected by the flagger because of an increase in signal to noise in these time steps.

In the next step, we used only the last 50 minutes of the sub-bands to calculate the false-positives ratio. Visual inspection of this data shows indeed no RFI, except for two time steps in the 4 km baseline that might have been affected, but these can not be assessed with certainty. The flagger does flag those time steps, hence we ignore them in the analysis. When flagging only the 50 minutes of 4 km baseline data, thereby making sure that the threshold is based only on this 50 minutes of data, a fraction of 0.6% was flagged. If one assumes that the selected data contains no other RFI, then this value is the ratio of falsely flagged samples. In the 56 km baseline, the same analysis leads to a slightly lower ratio of false-positives of 0.5%.

The 0.6 and 0.5% detection rates are the result of flagging on all four cross-correlations (XX, XY, YX and YY). In the samples that have been detected as RFI, we observe that there are zero samples flagged in more than one cross-correlation, thus they are completely uncorrelated. Each cross-correlation adds about 0.13–0.15% of falsely detected samples.

5.6 Comparison with other observations

Although we have analysed a substantial amount of survey time, it is useful to validate whether the two observations are representable samples for determining the LOFAR interference environment. Unfortunately, comparing the surveys with other observations is hard at this point, because LOFAR is commissioning with lower frequency and time resolutions, and the analysed 24 hr surveys are the only substantial observations performed at the targeted LOFAR resolution. Also, there are no strong sources in the targeted NCP field, but fields that do have strong sources might trigger the flagger more easily, yielding higher detection rates.

Table 5.3: *Observations and their RFI occupancy as reported by automated detection.*

Date	Start (UTC)	Duration	Id	Target	$\Delta\nu$ (kHz)	RFI[1]
LBA observations (frequency range ≈ 30 – 78 MHz)						
2010-11-20	19.33	5 min	L21478	Moon	3.0	4.6%
2010-11-20	19.43	6 h	L21479	Moon	3.0	10.3%
2011-04-14	19.00	8 h	L25455	Moon	0.76	4.3%
2011-10-09	**6.50**	**24 h**	**L31614**	**NCP**	**0.76**	**1.8%**
HBA observations (frequency range ≈ 115 – 163 MHz)						
2010-11-21	20.26	5 min	L21480	Moon	3.0	5.6%
2010-12-27	**0.00**	**24 h**	**L22174**	**NCP**	**0.76**	**3.2%**
2011-03-27	20.00	6 h	L24560	NCP	3.0	1.5%
2011-04-01	16.08	6 h	L24837	3C196	3.0	2.6%
2011-06-11	11.30	1.30 h[2]	L28322	3C196	3.0	6.5%
2011-11-17	18.00	12 h	L35008	NCP	3.0	3.6%
2011-12-06	2.36	25 min[3]	L36691	3C196	3.0	5.5%
2011-12-06	8.34	25 min[3]	L36692	3C295	3.0	8.0%
2011-12-20	7.39	30 min	L39562	3C295	3.0	2.5%
2012-01-26	2.00	5.30 h	L43786	3C295	3.0	3.6%

Notes:

[1] RFI occupancy as found by automated detection. For some targets, this is too high because of the band-edge issues that are discussed in the text, leading to approximately a 1–2% increase in 3-kHz channel observations.

[2] This observation was originally 6 hrs, but failed after 1.30 h.

[3] These observations were originally 30 min, but the first 5 min failed.

To assess the differences between different observations, we have performed detection ratio analysis of several other observations. For this purpose, we collected several LOFAR Epoch of Reionisation test observations and a few observations that were used for quality assessment. These were subsequently processed similar to how we processed the surveys. The observations were selected independent of their quality, thus they sample the RFI situation randomly. Important to note is however, that in our experience the data quality is quite independent of the detected RFI occupancy. Much more relevant is the position of the Sun in the sky, the state of the ionosphere and the stability of the station beam. These have very little effect on the detected RFI occupancy.

Table 5.3 lists the observations and shows their statistics. The number of involved stations varies between the observations, but as many as possible core stations were used in all observa-

tions.

Currently, there is an issue with some LOFAR observations that causes a higher RFI detection rates in fields with strong sources. This is caused by the edges of sub-bands in some cross-correlated baselines. These edges are flagged because they show time-variable changes that are very steep in the frequency direction. This effect is only observed in cross-correlations that involve exactly one superterp station, so it is assumed that this is a bug in the station beamformer or correlator, but this has not been fixed or attributed at the time of writing. In 64 channel observations that show this issue, the highest and lowest sub-band channels get flagged in about half of the baselines, leading to about a 1–2% higher detected RFI occupancy. The issue only arises in fields that contain strong sources, and is consequently not affecting the 24 hr RFI surveys, because there are no such sources in the NCP field. All 3C196, 3C295 and Moon observations do show the issue.

The average detected RFI ratios are 5.4 and 4.3% with standard deviations 3.5 and 2.0% for the LBA and HBA observations respectively. Therefore, it appears that the analysed 24 hr RFI surveys, with 2.4 and 3.3% RFI occupancy in the low and high bands respectively, are of better quality than the average observation. If one however assumes that the observations with lower time and frequency resolutions have an approximately 1.0% RFI increase, which seems to be a reasonable estimate according to Fig. 5.15, and taking into account that the subband-edge issue causes in the fields with strong sources another 1.5% RFI increase on average, the averages after correction for these effects become 3.7% and 2.4%. Therefore, the RFI occupancies of the 24 hr surveys seem to be reasonably representative for the RFI occupancy of LOFAR at its nominal resolution of 0.76 kHz with 1 s integration time. On the other hand, it also shows that 3 kHz channels may well suffice for regular LOFAR observations.

Manual inspection of the same data agreed with this observation: the RFI environment is not significantly different between different observations. The only exception was the Moon observation of 2010-11-20, which seems to contain unusual broadband interference over the entire duration of the observation. The shape and frequency at which the interference occurred is not like in any other observation. Therefore, we suspect that either something went wrong during this particular observation or ionospheric conditions were exceptional. According to weather reports, it was observed at the day of the year with highest humidity, although we have no direct explanation why this would influence the RFI detection.

5.7 Discussion & conclusions

We have analysed 24 hour RFI surveys for both the high-band and low-band frequency range of LOFAR. Both sets show a very low contamination of detectable interference of 1.8 and 3.2% for the LBA and HBA respectively. These are predicted to be representative quantities for what can be expected when LOFAR starts its regular observing with resolutions of 1 kHz and 1 s. Therefore, the LOFAR radio environment is relatively benign, and is not expected to be the limiting factor for deep field observing.

Almost all interference is detected after the single flagging step at highest resolution, and RFI that leaks through is very weak. This agrees with the first imaging results, which are thought to be limited by ionospheric calibration issues and system temperature, but not by interference. However, whether this will be the case for long integration times of tens of night, as will be done in the Epoch of Reionization project, is still to be seen. In that case, one might find that weak,

stationary RFI sources add up coherently, and might at one point become the limiting factor. Nevertheless, the situation looks promising: our first-order flagging routines use only per-baseline information, but remove in most cases all RFI that is visible. The resulting integrated statistics of 24 hours show very few artefacts of interference, which are causing no apparent issues when calibrating and imaging the data.

Once RFI does becomes a problem, there are many methods at hand to further excise it. The interference artefacts that are currently present can be flagged in a second stage flagger. In such a stage, the flagger could use the information from the full observation at once, and such a strategy would have a higher sensitive towards the weak stationary sources. Moreover, the involved Fourier transform is a natural filter of stationary interference, that would place the contribution of the stationary sources near the North Pole. With sufficient uv-coverage, their sidelobes will be benign, and if necessary, can even be further attenuated with techniques such as the low-pass filters presented in Offringa et al. (2012a).

An unexpected result was to find that the RFI occupancy is not significantly different between day and night. The setting of the Galaxy and the Sun are dominating the fluctuations in both the system temperature of the instrument and the RFI detection ratio, and this is the only structured variation over time that is apparent in the data. Therefore, we think that RFI is not an argument for deciding whether to observe at day or night. Of course, there are other arguments to conduct low-frequency observations at night, especially because of the stronger effect of the ionosphere and the presence of the Sun during the day, which both make successful calibration more challenging.

An estimate of the false-positives ratio of the AOFlagger pipeline of 0.5–0.6% was given based on the amount of falsely detected samples in clean-appearing data. We have seen that during long observations, in which the system temperature changes due to the setting of the Galaxy and the Sun, time ranges with increased variance result in higher amounts of false detections. Therefore, it would be a good practice to apply the correction method that was used for the LBA set: by (temporarily) dividing the samples by an accurate estimate of the standard deviation before flagging the data, the ratio of false-positives remains constant. This requires two runs of the flagger: one run to be able to estimate the variance on clean data, and one more to flag the data with normalized standard deviation. This decreases the amount of false-positives with about 0.5%. However, it is also computationally expensive, and is not necessary for short observations that do not show a significant change in sky temperature.

Up to now, interference detection was often performed manually and ad-hoc by the observer. Consequently, few statistics are available in the literature that describe the amount of data loss in cross-correlated data due to interference for a particular observatory and frequency range. However, when compared with losses achieved with common strategies, the amount of data loss in LOFAR is very low. This is especially impressive considering the fact that LOFAR is build in a populated area and operating at low frequency. Several reasons can be given for the small impact of RFI on LOFAR:

- Many interfering sources contaminate a narrow frequency range or short duration. LOFAR's high time and frequency resolutions, of 1 s and 0.76 kHz respectively, minimize the amount of data loss caused by such interfering sources. Since the current loss of data is tiny, it seems not necessary to go to even higher resolutions.

- LOFAR is the first telescope to use many novel post-correlation detection methods, such as the scale-invariant rank operator and the `SumThreshold` techniques. The AOFlagger interference detection pipeline shows an unprecedented accuracy (Offringa et al., 2010a,

2012b).

- LOFAR's hardware is designed to deal with the strong interfering sources that are found in its environment. The receiver units remain in linear state in the neighbourhood of such sources, and the strong band-pass filters spectrally localize the sources. Consequently, almost no interfering source will cause ramifications in bands that are adjacent to their transmitting frequency. The only exception is at very low frequencies, where we do see a very strong source saturate the ADCs when ionospheric conditions are bad. This source and its harmonics are successfully removed during flagging.

- Propagation models for Earth-bound signals show a strong dependence on the height of the receiver (e.g., Hata (1980)). In contrast to dishes with feeds in the focal point, the receiving elements of LOFAR are close to the ground.

- LOFAR is remotely controlled, and the in situ cabins with electronics are shielded. We have found no post-correlation contamination that is caused by self-generated interference. This is in contrast with for example the WSRT, where the dishes close to the control room (which contains the correlator, but it is operated from elsewhere) are known to observe more interference. In the LOFAR auto-correlations, every now and then we do see some artefacts that suggest local interference, but these do not visibly contaminate cross-correlations.

Chapter 6

The brightness and spatial distributions of RFI

Based on:
"*The brightness and spatial distributions of terrestrial radio sources*"
(Offringa et al., in preparation)

RADIO ASTRONOMY concerns itself with the observation of radiation from celestial sources at radio wavelengths. However, astronomical radio observations can be affected by radio-frequency interference (RFI), which makes it difficult to calibrate the instrument and achieve high sensitivities. Many techniques have been designed to mitigate its effect, such as detection and flagging of the data, spatial filtering or adaptive cancellation techniques. In most cases, these methods excise enough of the interference to calibrate and image the data.

Regular radio observations record one or a few dayparts and the results are combined. In these cases, it often suffices to only excise interference that is apparent and thus above the noise to reach the thermal noise limit of the instrument. A new challenge arises when one desires much deeper observations, and a large number of observations need to be averaged. In such a case, feeble interference caused by stationary sources might not manifest itself above the noise in individual observations, but is coherently present in the data. Subsequently, when averaging these data, the interference might become apparent and occlude the signal of interest. One experiment involving long integration times is the Epoch of Reionisation experiment using the Low-Frequency Array (de Bruyn et al., 2011). For this experiment, it is important to know the possible effect of low-level interferences on the data, as these might overshadow or alter the signal of interest.

In this chapter, we will explore the information that is present in interference distributions, in order to analyse possible low-level interference that is not detectable by standard detection methods. Our approach will be similar to the radio source count analysis that is used in cosmology (von Hoerner, 1973), also named $\log N - \log S$ analysis, where N and S refer to the celestial source count and brightness respectively. The slope in such a plot contains information about source populations, their luminosity functions and the geometry of the Universe. We will analyse

such a double-logarithmic plot for the case of terrestrial sources, with the ultimate goal of predicting their full spatial and brightness distributions. This will give a better insight in the effects of low-level interference and allows one to simulate the effects of interference more accurately. A distribution function can be differential or cumulative. This work uses differential distributions, because they are better suited for visualizing the parameters we are interested in, although the two basically convey the same information.

In Sect. 6.1, we will predict the terrestrial interfering source count based on various assumptions. Sect. 6.2 presents the methods that we use to generate and analyse the histograms. This is followed in Sect. 6.3 by a short description of the two LOFAR data sets that have been used to perform the experiment. The results of analysing the sets will be presented in Sect. 6.4. Finally, in Sect. 6.5 the results will be discussed and conclusions will be drawn.

6.1 Prediction of the brightness distribution

Interference is generated by many different kind of transmitters, and these will have different spatial and brightness distributions ("spatial" refers here to the distribution on the Earth). For example, aeroplanes and satellites have widely different heights, while other sources are ground-based. Even ground-based sources might be spread differently. For example, it can be expected that citizens' band (CB) devices, that are often used in cars, are differently distributed from broadcasting transmitters. The frequencies of deliberately transmitting sources of a particular class of devices are constrained by the bands that have been allocated for the given class, which might allow one to distinguish different sources to some extent.

In a time-frequency plot, interfering sources can have complex structures. They can also be intermittent and different sources might overlap in time-frequency space. An example of interfering sources can be seen in Fig. 6.1, which shows raw visibility data of one baseline of a LOFAR observation in a time-frequency diagram (also called a dynamic spectrum). Because many sources change over time, are repetitive or affect multiple channels, many sources produce multiple unconnected features in the time-frequency diagram. It is often not clear what constitutes a single interfering source, hence it is hard to count individual sources. Instead, we will count the number of times a given brightness occurs in time-frequency space. This — as well as many other effects — will of course influence the distribution. First we will derive the expected intrinsic source distribution. After that, we will analyse the issues that arise when measuring the distribution by counting samples.

In every dynamic spectrum we can measure the number of times that the flux density is within a particular range. Dividing this quantity by the total number of samples yields the relative number of events as a function of intensity. We will refer to this quantity with the statistical term "frequency density" (not to be confused with the physical radiation frequency). We will now start by deriving a prediction of the frequency density function of ground-based interfering sources. Consider an interfering source of strength I. This source is observed by an interferometer that consists of two antennae or stations with gains g_1, g_2, which include all instrumental effects. The antennae are located at distances r_1, r_2 from the source. The interferometer will record an apparent instantaneous strength S of

$$S(r_1, r_2) = I \frac{g_1 g_2}{r_1 r_2}, \tag{6.1}$$

with $0 < g_1, g_2$ and $0 < r_1, r_2 < r_{\text{max}}$. r_{max} is a maximum distance, which is practically limited

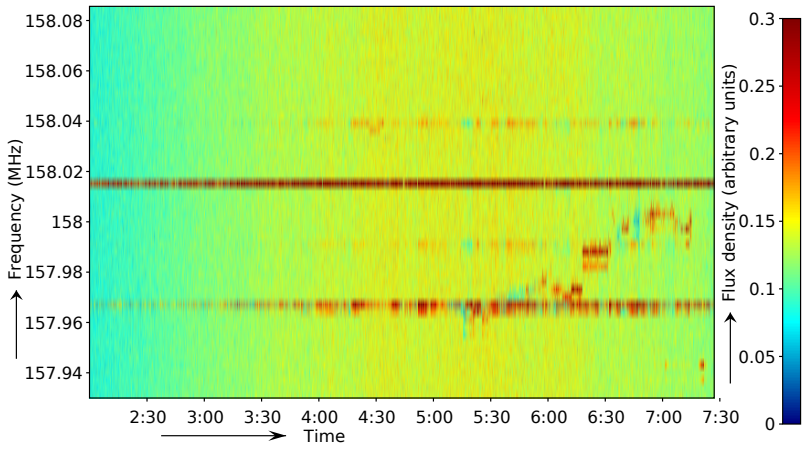


Figure 6.1: *A small part of an observation displayed in a time-frequency plot. The features with significantly higher values are caused by interference. Some of these have a constant frequency, while others are more erratic.*

to distances well below the diameter of the Earth. We will limit our analysis to cross-correlated antennae; the auto-correlations will be ignored.

Here we assume that the source is observed fully coherent, but a possible decoherence factor can be absorbed in the gains. Due to the small bandwidth of most interfering sources, most RFI will be received coherently, because of the narrow-band condition. For example, if the bandwidth of the signal $\Delta\nu = 1$ kHz, the narrow-band condition $\Delta\nu \ll (2\pi\tau)^{-1}$ with correlation delay τ will hold for baselines up to a few km, because it holds as long as the baseline length is significantly less than $\Delta x = c/2\pi\Delta\nu \approx 50$ km.

Now, we can treat the interferometer geometrically as a single point, as both antennae will see the same distribution. Then, we can express the received amplitude S for a given distance r and gain g as

$$S(r) = \frac{Ig}{r^2}. \tag{6.2}$$

Next, we assume that all sources have equal constant strength I. Moreover, we assume that the sources have a uniform spatial distribution in a two-dimensional plane. This is obviously a simplification, as the sources are actually distributed on the surface of a sphere. Beyond some distance, the Earth will also partly block the path between transmitter and receiver. Therefore, the assumption of uniformly distributed sources on a two-dimensional plane is only valid for a limited distance. Using these assumptions, we can express the expected cumulative frequency density of sources $f_I(r)$ at distance r as

$$F_{\text{distance}\leq r}(r) = c\pi r^2, \tag{6.3}$$

for some constant c that represents the number of sources per unit area. In other words, we will observe F sources that are at most a distance of r away.

We need the reverse of Eq. (6.3) to predict the amplitude distribution, because sources with an amplitude of *at most* a given strength will have a distance of *at least* some distance. Sources

with at least a distance of r are given by $F_{\text{distance}\geq r}(r) = N - c\pi r^2$, with N the total number of sources.

The cumulative number of sources $F_{\text{amplitude}\leq S}$ that have an amplitude of at most S can now be calculated with

$$\begin{aligned} F_{\text{amplitude}\leq S}(S) =& F_{\text{distance}\geq r}(S^{-1}(S)) \\ =& F_{\text{distance}\geq r}(\pm\sqrt{\frac{Ig}{S}}) \\ =& N - \frac{c\pi Ig}{S} \end{aligned} \tag{6.4}$$

where S^{-1} refers to the inverse of $S(r)$, i.e., the function that returns the distance r for a given amplitude S. Finally, the differential frequency density can be calculated by taking the derivative,

$$\begin{aligned} f_S(S) =& \frac{dF_{\text{amplitude}\leq r}}{dS} \\ =& \frac{c\pi Ig}{S^2}. \end{aligned} \tag{6.5}$$

Therefore, if we plot the histogram of the RFI amplitudes in a log-log plot, we predict to see a slope of -2 over the interval in which the RFI sources are spread like uniform sources on a two-dimensional plane.

6.1.1 Spherical case

If we do take into account the fact that the sources lie on a sphere, Eq. (6.3) needs to be adapted. In this case, the number of sources corresponds with the surface area of a hemisphere. By using the formula for the surface area of a hemisphere and some basic geometry, one finds that

$$F_{\text{distance}\leq r} = c2\pi R^2 \left(1 - \cos(2\arcsin\frac{r}{2R})\right). \qquad (\text{with } r \leq 2R), \tag{6.6}$$

with R the radius of the Earth. By following the same reasoning as above, one finds that $\tilde{f}_S(S)$, the frequency density function in the spherical case, can be calculated with

$$\tilde{f}_S(S) = \frac{d}{dS}\left[N - c2\pi R^2\left(1 - \cos(2\arcsin\frac{\sqrt{\frac{Ig}{S}}}{2R})\right)\right]. \qquad \left(\text{with } \frac{Ig}{S} \leq 4R^2\right) \tag{6.7}$$

This can be easily calculated with a computer algebra system. The resulting function itself has many terms, but we calculated what the slope of $\tilde{f}_S(S)$ would be in a log-log plot, and found that the spherical case also has a constant slope of -2. In fact, Eqs. (6.6) and (6.3) have the same value over the range they can be evaluated. Therefore, the spherical case equals the two-dimensional case in this respect. Note that $\tilde{f}_S(S)$ can not be evaluated for values of S for which $\frac{Ig}{S} > 4R^2$. This corresponds with sources that are at a distance beyond the diameter of the Earth. There are no such sources, since we only modelled sources on Earth, and $\tilde{f}_S(S) = 0$ in this case. In practice, the lowest amplitude of observed sources is further constrained because of the curvature of the Earth, causing distant sources to settle below the horizon. Nevertheless, some of those sources might still be visible due to reflection of the ionosphere and other (semi-)spherical propagation effects.

6.1.2 Propagation effects

So far, we have assumed that the electromagnetic radiation propagates through free space, resulting in a r^{-2} fall-off. In reality, the radiation will be affected by complicated propagational effects. Because of the irregular surface of the Earth (including urban areas) and the absence of a direct line of sight between transmitter and receiver, the propagation mode might be indirect. Multiple indirect paths might be formed by reflection, refraction or diffraction of the electromagnetic wave.

A commonly used propagation model is the empirical model determined by Okumura et al. (1968), which was further developed by Hata (1980). Hata gives the following analytical estimate for L_p, the electromagnetic propagation loss over land:

$$L_p = 69.55 + 26.16 \log_{10} f_c - 13.82 \log_{10} h_b - a(h_m) + (44.9 - 6.55 \log_{10} h_b) \log_{10} r, \quad (6.8)$$

where L_p the loss in dB; f_c the radiation frequency in MHz; h_b the height of the transmitting antenna in m; h_m the height of the receiving antenna in m; r the distance between the antennae in m; and $a(h_m)$ a correction factor in dB that corrects for the height of the receiving antenna and the urban density. Hata found this model to be representative for frequencies $f_c \sim$ 150–1500 MHz, with transmitter heights $h_b \sim$ 30–200 m, receiver heights $h_m \sim$ 1–10 m and over distances $r \sim$ 1–20 km.

Collecting the terms of Eq. (6.8) that are not depending on r in a single variable ζ, and converting from a subtrahend in decibels to a flux density factor L_S, results in

$$L_S = \zeta r^{4.49 - 0.655 \log_{10} h_b}, \quad (6.9)$$

with

$$\zeta = \frac{f_c^{2.616}}{h_b^{1.382}} - 10^{6.955 - \frac{1}{10} a(h_m)}. \quad (6.10)$$

The model is based on urban areas, but corrections are given by Hata for sub-urban areas with a lower population density and for open areas. These corrections are independent of r, thus would only change ζ. Note that according to Hata's model, the exponent of the power law in Eq. (6.9) depends only on the height of the transmitting antenna, that is, the exponent is independent of frequency, receiver height and urban density. Now, if in Eqs. (6.4) and (6.5) one replaces the definition of $S(r)$ from Eq. (6.2) with one that includes the propagation effects,

$$S(r) = \frac{\zeta I g}{r^\eta}, \quad (6.11)$$

with $\eta = 0.655 \log_{10} h_b - 4.49$, one finds the frequency density function f_p that considers propagation effects,

$$f_p(S) = \frac{d}{dS} c\pi \left(\frac{\zeta I g}{S} \right)^{2/\eta} = \frac{c2\pi}{\eta S} \left(\frac{\zeta I g}{S} \right)^{2/\eta}. \quad (6.12)$$

Consequently, due to non-free space propagation effects, the observed log-log histogram is predicted to have a $\frac{2}{\eta} - 1$ slope. By substituting η, one finds

$$\text{slope}(h_b) = \frac{1}{0.3275 \log_{10} h_b - 2.245} - 1. \quad (6.13)$$

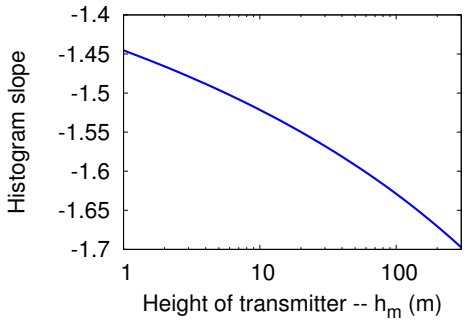


Figure 6.2: *Effect of transmitter height on the slope of a log-log histogram.*

This is valid for transmitters that have a height of 30–200 m, the range over which Hata's model was defined. This yields distribution slopes of approximately -1.57 and -1.67 for 30 m and 200 m high transmitters respectively. In Figure 6.2, the slope value is plotted as a function of the transmitter height, including extrapolated values for transmitter heights down to 1 m.

6.1.3 Inclusion of noise

The full measured distribution will consist of the power-law distribution combined with that of the noise and the celestial signal. For now, we will ignore the contribution of the celestial signal, as its contribution to the amplitude distribution will be minimal when observing fields without strong sources. Noise, however, will have a contribution. The real and imaginary components of receiver noise are independent and identically Gaussian distributed. Consequently, an amplitude x will be Rayleigh distributed:

$$f_{\text{noise}}(x) = \begin{cases} \frac{x}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}} & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.14}$$

Because most of the samples will be unaffected by RFI, this will be the dominating distribution. The Rayleigh distribution is plotted together with the -2 power-law distribution of Eq. (6.5) in Fig. 6.3.

So far, this is the expectation for histograms of pure noise and pure RFI that propagates through free space. However, the measured distribution is a mixture of the two. A perfect RFI detector would separate the samples in two distributions; one that is not affected by RFI, and therefore contains noise only, and one that is the sum of RFI and noise. Real RFI detectors can separate these distributions to some extent, but due to false positives and false negatives, the histograms will get mixed nevertheless. Even more, we still need to take into account that the RFI classified samples are also affected by noise. Although this effect is moderate, because 'most' RFI samples are of much higher amplitude than what is added because of the noise, it is the low-level RFI samples which we are interested in, and the relative effect of noise on those samples is large.

The real and imaginary components of samples that are affected by RFI, will follow a bivariate distribution: each complex component is the sum of the complex component of RFI and of a noise

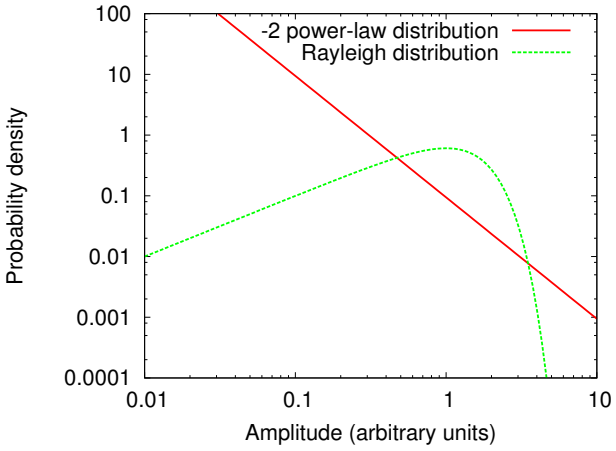


Figure 6.3: *The Rayleigh and power-law distributions in a log-log plot. The power-law distribution (Eq.* (6.5)*) has a constant slope of -2. The slope of the Rayleigh distribution in the limit of the origin is 1. Its maximum occurs where the amplitude value equals its mode σ, which is 1 in this example. For higher amplitudes, its slope decreases exponentially.*

sample. The noise comes from a Gaussian distribution with zero mean,

$$f_{\text{Gaussian}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}, \tag{6.15}$$

To calculate the corresponding distribution of the amplitudes, it is easier to perform numerical simulations, as algebraic calculation is not trivial. The histogram can be numerically estimated by drawing complex real and imaginary samples from the two distributions and calculating and counting the amplitudes. A sample can be drawn from the RFI distribution by normalizing and inverting the cumulative frequency function in Eq. (6.4) and evaluating it for a uniformly distributed variable. Note however that $F_{\text{amplitude}\leq S}(S)$ is not limited; when decreasing the amplitude S towards zero, the number of sources will go to infinity. Therefore, to draw a sample by using the inversion method, this would require a uniformly distributed variable $\sim U(0, \infty)$. By assuming that there are no samples beyond some limiting distance r_l, then $F_{\text{amplitude}\leq S}(S) = N - c\pi r_l^2$ for $S < Ig/r_l^2$. Consequently, by sampling a uniform value x_u with $0 \leq x_u \leq N - c\pi r_l^2$ and applying the inverse of Eq. (6.4), we will sample amplitudes S with $S \geq Ig/r_l^2$, and $S \sim$ the power-law distribution.

A single real or imaginary RFI contaminated amplitude S_R can then be sampled with

$$S_R = \frac{Ig}{x_u r_l^2}, \tag{6.16}$$

and the corresponding formula for drawing a sample S that is contaminated by both RFI and noise is

$$S = x_n + \frac{Ig}{x_u r_l^2}, \tag{6.17}$$

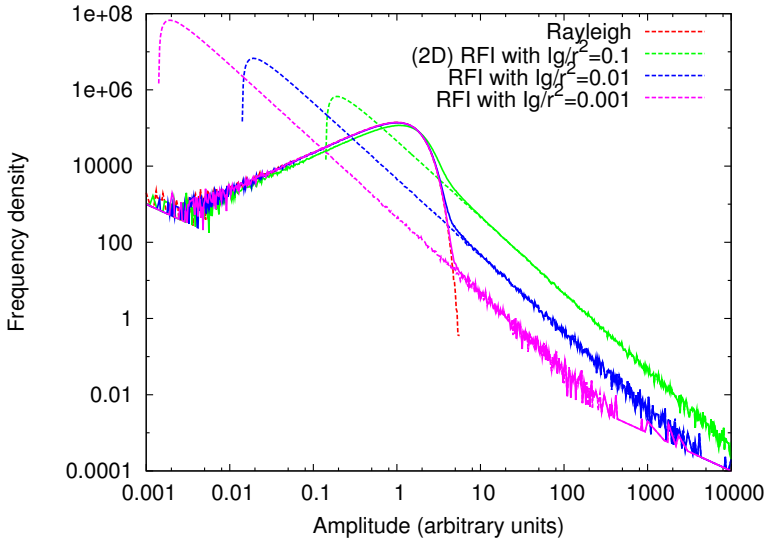


Figure 6.4: *Histograms of simulated samples that all have a contribution of noise and RFI. Various settings of the parameters were used, and samples were drawn as described in Eq.* (6.18)*. Solid lines: the combined distributions, dashed lines: the original distributions before mixing.*

with $x_n \sim N(\mu = 0; \sigma = 1)$ and $x_u \sim U(0, 1)$. The final amplitude sample S_A can then be calculated with

$$S_A = \sqrt{S_1^2 + S_2^2}, \tag{6.18}$$

where S_1 and S_2 have been independently drawn in accordance with Eq. (6.17). An example of distribution curves for various settings of Ig/r_l^2 is given in Fig. 6.4.

If we assume that I, the average source strength; g, the average instrumental gain; and r_l, the distance over which those sources are visible, are all unknown, it can be seen from Eq. (6.17) that given the histogram we can solve neither r_l, I or g, as they all have the same effect of scaling the -2 power-law part of the histogram. Although calibration could in theory solve g, almost all sources will come in through the edges of the beam, and finding the expected values for the gains is therefore hard. The effects of these parameters on the histogram will be further discussed in §6.1.4.

Fig. 6.4 shows that the left side of the graph corresponds with the original Rayleigh distribution. From the right side of the histogram, we are able to estimate Ig/r_l^2, as the placing of the power-law curve is independent of the Rayleigh distribution for large S. Methods to constrain the parameters of the RFI distribution will be discussed in §6.2.3. The curves in Fig. 6.4 are from histograms of samples that are all contaminated by both noise and RFI. In practice, we can not make this distinction 100% accurately, as RFI detectors have a limited accuracy. Consequently, the observed histogram will be the sum of two types of histograms: the first histogram being contaminated by both noise and RFI, the second one only by noise.

6.1.4 Parameter variability

In reality, the parameters c, I and g, which are the source density per unit area, source strength and instrumental gain respectively, will not be constant over time and frequency but have a stochastic nature. However, since each specific value for these parameters produces a power law, the combined distribution will still show a power law, as long as the parameters follow a distribution that is steep at high amplitudes, such as a Gaussian or uniform distribution.

One instrumental effect that is absorbed in g is the frequency response of the instrument, i.e., the antenna response in combination with the passband of the analogue and digital filters. Because the data that are analysed in Sect. 6.4 has initially not been band-pass calibrated, the instrumental response is not uniform over frequency. We determined that the variation due to the band-pass is about one order of magnitude in the LBA and about a factor of two in the HBA.

The effect of the band-pass on the data distribution is consequently limited to one order of magnitude or less. If the RFI sources have an apparent power-law distribution with a certain exponent, the distribution will have a feature at low amplitudes due to the frequency response, but a power law with equal exponent will still appear when the distribution is observed over a wide enough range.

Another effect that is absorbed in g, is the beam of the instrument. At the point of writing, LOFAR beam models are still being developed and are not yet well parametrized. However, since it is likely that most RFI sources are observed at the edges of the beam, it can be expected that the beam will have a benign effect on the histogram of an observation, comparable with the effect of the frequency response.

6.2 Methods

In this section we will briefly discuss how the histograms are created, how the slope of the underlying RFI distribution is estimated and show how to constrain some of the intrinsic RFI parameters.

6.2.1 Creating a histogram

While creating a histogram is trivial, it is important to note that it is necessary to have a variable bin size. This is mandated by the large dynamic range of the histogram that we are interested in. Therefore, we chose to have a bin size that increases linearly with the amplitude S, and the frequency distribution is normalized by the bin size after counting. Consequently, in parts of the histogram that have a sparse number of samples, the outlying samples will follow a $1/S$ curve, or a -1 slope in a log-log plot. This can be seen in the tails of the distributions of Fig. 6.4. This, however, is not an intrinsic feature of the data but a consequence of the binning method.

6.2.2 Estimating the slope

An automated procedure is used to calculate the slope of the observed log-log histogram. First, the mode $\hat{\sigma}$ of the Rayleigh distribution is estimated by finding the amplitude with the maximum occurrences, i.e., the amplitude corresponding to the peak of the histogram. Then, the maximum amplitude S_{max} is calculated, i.e., the sample with highest amplitude. Finally, the slope is esti-

mated using linear regression over the interval R,

$$R = \left[20\hat{\sigma}; 10^{\frac{1}{2}(\log_{10} 20\hat{\sigma} + \log_{10} S_{\max})}\right]. \tag{6.19}$$

In terms of the log-log histogram, the start of the interval corresponds with the point that is 1.3 units to the right of the peak in the histogram, and the interval end corresponds with the point halfway between the start point and the maximum amplitude. This interval starts near the start of the RFI slope but past where the Rayleigh curve dominates. It also does not include the noisy data at the end of the curve.

One should note that fitting straight lines to the distribution curve in a log-log plot is not the most accurate way of estimating the exponent of a power-law distribution (Clauset et al., 2009). However, because of our enormous sample size, which allows fitting the line over a large interval, the estimator will be sufficiently accurate for our purpose. Nevertheless, we will additionally calculate a maximum likelihood estimator for comparison. The maximum-likelihood estimator for the exponent in a power-law distribution is given by the Hill estimator (Hill, 1975; Clauset et al., 2009), defined as:

$$\alpha_H = 1 + N \left(\sum_{i=1}^{N} \ln \frac{x_i}{x_{\max}} \right)^{-1}. \tag{6.20}$$

However, this estimator assumes the distribution is not cut-off at a high point. In our case, the distribution is cut-off, for example because of the limit of the analogue-digital converter (see §6.2.3). Therefore, using this estimator will result in an estimate that is steeper (i.e., more negative) than the actual distribution. Moreover, the estimator requires an iteration over all data points, and because of the size of the data sets this is somewhat impractical. However, we can calculate Eq. (6.20) by adding the bin centre values N times to the calculation, where N the bin frequency. Because the bin size is rather small, the resulting estimation will be close to the normal Hill estimator and one does not have to iterate over all data.

6.2.3 Determining RFI distribution limits

In this section we will show methods to find S_U and S_L, the upper and lower flux limits of the power-law distribution at which the power law breaks down. Once the exponent of the power-law part of the distribution is estimated with the previously discussed techniques, the distribution can be extrapolated to find a lower flux limit. Assume that we have found a power law with exponent α and factor β over an amplitude region $[S_1; S_2]$, resulting in the following frequency density function g:

$$g(S) = \beta S^{\alpha}. \tag{6.21}$$

Assume also that the histogram contains N_{part} (RFI) samples with amplitude $> S_1$, as sketched in Fig. 6.5. The hypothetical upper limit S_U of the distribution can now be found, i.e., the highest amplitude that would be observed when the distribution follows the power law up to that point, by solving

$$\int_{S_1}^{S_U} g(S) dS = N_{\text{part}}. \tag{6.22}$$

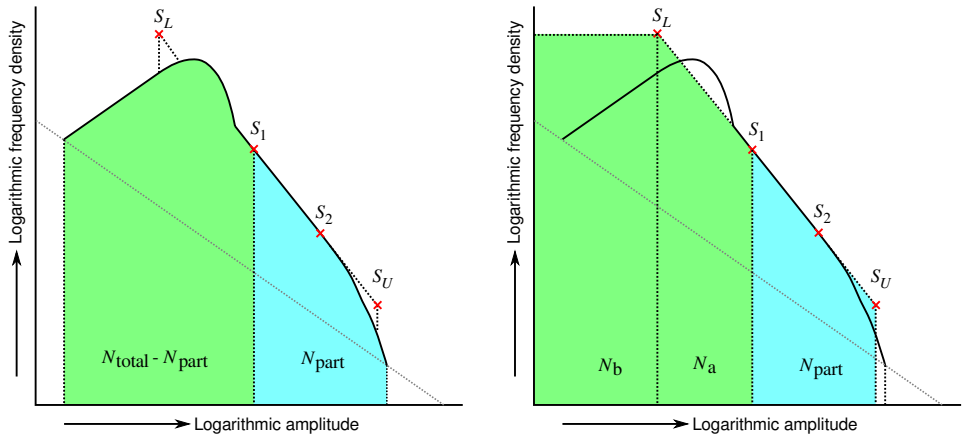


Figure 6.5: *Cartoon of how a constraint on the lower fall-off point of the power-law distribution can be determined. Note that the labelled areas are areas as occupied in a linear plot, i.e., the integration of the density function. Areas in a log-log plot are not linearly related to the integral. There are two ways to estimate the lower constraint S_L: (i) the areas N_a and $N_{total} - N_{part}$ are equal if Ig/r^2 is constant during the observation, and (ii) if one assumes $Ig/r^2 \sim$ uniform, then $N_a + N_b = N_{total} - N_{part}$.*

In practice, the observed histogram will break down beyond some amplitude because of several reasons. First and most importantly, the samples itself are digitized with an analogue-to-digital converter (ADC) with limited range. Second, we observe for a limited time and the frequency count is discrete. Because the chance of finding a sample with a very high intensity is low, it is unlikely to observe samples beyond some amplitude within the finite observing time. Finally, under the assumption of a uniform spatial distribution of RFI transmitters, samples with very high amplitude would have to be produced by transmitters that are very close to the telescope. However, it is likely that the uniform spatial distribution of transmitters will break down at closer distances.

Solving Eq. (6.22) results in

$$S_U = \sqrt[\alpha+1]{\frac{\alpha+1}{\beta} N_{\text{part}} + S_1^{\alpha+1}}. \tag{6.23}$$

In some cases there will be no solution for S_U. The reason for this is that the integral of the distribution function converges, and the total number of samples is finite even when $S \to \infty$. If N_{part} exceeds the integrated value of the fitted distribution function, there is no solution. This can happen when the empirical distribution is not limited at the high end, or when it contains features that are not in the model.

Similar to the calculation of S_U and by using N_{total} as an upper limit to the total number of samples that are affected by the power-law distribution, one can estimate the lower limit S_L. For Fig. 6.5 this means that the areas labelled N_a and $N_{\text{total}} - N_{\text{part}}$ are equal. Solving this equation, one finds that

$$S_L = \sqrt[\alpha+1]{\frac{\alpha+1}{\beta} \left(N_{\text{part}} - N_{\text{total}}\right) + S_1^{\alpha+1}}. \tag{6.24}$$

With the assumption that $Ig/r_l^2 \sim$ a uniform distribution, the area labelled in Fig. 6.5 as N_b is also part of the RFI distribution, and a stronger constraint $\tilde{S}_L$ can be found by solving

$$\int_{\tilde{S}_L}^{S_U} g(S)dS + \tilde{S}_L g(\tilde{S}_L) = N_{\text{total}} - N_{\text{part}}, \tag{6.25}$$

which yields

$$\tilde{S}_L = \sqrt[\alpha+1]{-\frac{1}{\alpha}\left(\frac{\alpha+1}{\beta}\left(N_{\text{part}} - N_{\text{total}}\right) + S_1^{\alpha+1}\right)}. \tag{6.26}$$

With estimates of α, β, S_L and S_U, one has obtained a parametrization of the RFI distribution. As was shown in §6.1.3, the left-most point where the power-law distribution falls off is Ig/r^2 – assuming free space propagation for the moment – and therefore $S_L = Ig/r^2$. For Hata's propagation model, the more generic solution $S_L = \zeta Ig/r^\alpha$ is found. This value represents the apparent brightness of the sources that are the furthest away from the telescope. With a fully parametrized distribution of the effect of RFI sources, an empirical model for RFI effects can be made. Moreover, one can calculate $E(S_R)$, the expected apparent strength of RFI:

$$E(S_R) = \frac{1}{N_{LU}} \int_{S_L}^{S_U} \beta S^\alpha S dS = \frac{\beta}{N_{LU}} \left[\frac{1}{\alpha+2} S^{\alpha+2}\right]_{S_L}^{S_U} = \frac{\beta\left(S_U^{\alpha+2} - S_L^{\alpha+2}\right)}{N_{LU}\left(\alpha+2\right)} \tag{6.27}$$

Here, N_{LU} is the number of samples between S_L and S_U after normalizing for the bin size:

$$N_{LU} = \int_{S_L}^{S_U} \beta S^\alpha dS. \tag{6.28}$$

Substitution and simplification of these two results in

$$E(S_R) = \frac{\left(S_U^{\alpha+2} - S_L^{\alpha+2}\right)\left(\alpha+1\right)}{\left(S_U^{\alpha+1} - S_L^{\alpha+1}\right)\left(\alpha+2\right)}. \tag{6.29}$$

This is essentially the flux density that is caused by RFI without using RFI detection or excision algorithms. $E(S_R)$ has the same units as S_L and S_U, thus for example in Jy after calibration (see §6.2.4). In practice, the increase of system noise after correlation is much less severe because of RFI flagging, which excises a part of the RFI. One can assume that all RFI above a certain power level is found by the detector. Since modern RFI detection algorithms can find all RFI that is detectable "by eye" (Offringa et al., 2010a), this power level will be near the level of the noise mode. In Chapter 5, the false-positives rate for the AOFlagger is estimated to be 0.5%. An estimate of S_d, the average lower limit of detected RFI, can be calculated by finding the point on the distribution where the area under the distribution to the right of S_d equals the "real number" (true positives) of RFI samples. Therefore, the limit is calculated similar to Eq. (6.24), where the term $N_{\text{part}} - N_{\text{total}}$ needs to be replaced with N_{RFI}, which equals the total number of samples detected as RFI minus the 0.5% false positives.

Finally, $E(S_{\text{leak}})$, which is the expected value of leaked RFI not detected by the flagger, can be calculated by replacing S_U with S_d in the numerator of Eq. (6.29) and subtracting the removed

number of samples from the total of number of samples. Assume that a fraction of $1-\eta$ samples have been detected as RFI, then

$$E(S_{\text{leak}}) = \frac{1}{\eta N_{LU}} \int_{S_L}^{S_d} \beta S^{\alpha} S dS = \frac{\left(S_d^{\alpha+2} - S_L^{\alpha+2}\right)(\alpha+1)}{\eta\left(S_U^{\alpha+1} - S_L^{\alpha+1}\right)(\alpha+2)}. \tag{6.30}$$

This is the average contribution that leaked RFI will have on a single sample. It has the same units as the parameters S_L, S_U and S_d. Typical values for η are 0.95–0.99.

6.2.4 Calibration

We can assign flux densities to the horizontal axis of the histogram by using the system equivalent flux density (SEFD) of a single station. The current LOFAR SEFD is found to be approximately 3000 Jy for the HBA core stations and 1500 Jy for the remote stations in the frequency range from 125–175 MHz. The SEFD is approximately 30000 Jy for all Dutch LBA stations in the frequency range 40–70 MHz. The standard deviation σ in the real and imaginary values is related to the SEFD with

$$\sigma = \frac{\text{SEFD}}{\sqrt{2\Delta\nu\Delta t}}, \tag{6.31}$$

where $\Delta\nu$ is the bandwidth and Δt is the correlator integration time. The standard deviation will appear as the mode of the Rayleigh distribution. By fitting a Rayleigh function with fitting parameter σ to the distribution, one finds the corresponding flux density scale.

RFI sources will enter through the distant sidelobes of the station beams from many unknown directions. Moreover, models for the full beam are often hard to construct. Therefore, we will not try to calibrate the beam, and the flux densities in the histogram are apparent quantities. Consequently, we will not be able to say something about the true intrinsic power levels of RFI sources.

6.2.5 Error analysis

An estimate for the standard deviation of the slope estimator $\hat{\alpha}$ can be found by calculating $\text{SE}(\hat{\alpha})$, the *standard error* of $\hat{\alpha}$. The standard error of the slope of a straight line (Acton, 1966, pp. 32–35) is given by

$$\text{SE}(\hat{\alpha}) = \sqrt{\frac{SS_{yy} - \hat{\alpha} SS_{xy}}{(n-2)\, SS_{xx}}}, \tag{6.32}$$

where SS_{xx}, SS_{xy} and SS_{yy} are the sums of squares, e.g., $SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and n is the number of samples. However, we found that this is not a representable error in our case, because the errors in the slope are not normally distributed. The noise in the part of the histogram over which the slope is calculated, is very small due to the large amount of samples. Consequently, the estimated standard deviation of $\hat{\alpha}$ is very low. Nevertheless, when the slope is calculated over subsections of the original range over which the slope is calculated, one finds the line is not completely straight and the slope is changing more than is predicted by the standard error. Therefore, we introduce an error estimate ϵ_{α} that quantifies a normalized standard deviation of the slope over the range. This error is formed by calculating the slope over n_{α} smaller subranges in the histogram, creating n_{α} estimates α_i. If the errors in α_i are normally distributed with mean

zero, the standard deviation over the full range will be $\sqrt{n_\alpha}$ times smaller. Therefore, an estimate of the variance of $\hat{\alpha}$ can be calculated with

$$\epsilon_{\hat{\alpha}} = \sqrt{\frac{\sum (\alpha_i - \hat{\alpha})^2}{n_\alpha^2 - n_\alpha}}. \tag{6.33}$$

This estimate is slightly depending on the number of subranges that is used, n_α, but we found that $\epsilon_{\hat{\alpha}}$ is more representative than the standard error of $\hat{\alpha}$.

The Hill estimator of Eq. (6.20) is a different estimation method for the exponent in the power-law distribution, and yields therefore also a different standard error. The standard error of the Hill estimator is (Clauset et al., 2009)

$$\mathrm{SE}(\hat{\alpha}_H) = \frac{-\alpha - 1}{\sqrt{n}} + \mathcal{O}(\frac{1}{n}). \tag{6.34}$$

Because the number of samples is very large ($> 10^{11}$), the $\mathcal{O}$-term will be very small. Therefore, we will calculate the quantity without the term. As with the standard error for the slope of a straight line in Eq. (6.32), the standard error for the Hill estimator yields very small quantities. Again, this is because it assumes the underlying power law has a fixed exponential, while in our case the power law seems to vary. Therefore, this value is given only for completeness.

Because our distributions are huge, we decided not to do goodness-of-fit tests, because these require many similar distributions to be simulated in order to reach accurate decision. Instead, we will try to evaluate the distributions visually.

6.3 Data description

We have analysed the distributions of two data sets. Both data sets are 24 h LOFAR RFI surveys and are extensively analysed in Chapter 5. In one set, the low-band antennae (LBA) were used and the frequency range 30.1–77.5 MHz was recorded, while in the other the high-band antennae (HBA) were used to record the frequency range 115.0–163.3 MHz. More stations were used in the LBA set. The specifications of the two sets are listed in Table 6.1. The stations that have been used are geometrically spread over an area of about 80 km and 30 km in diameter at maximum for the LBA and HBA sets respectively.

Although we have used Hata's model to estimate the RFI log-log histogram slope, our frequency range falls partly outside the frequency range over which Hata's model has been verified. However, according to Hata's model the observing frequency does not influence the power-law exponent in the frequency range 150–1500 MHz, thus it can be assumed the exponent will at least not significantly differ over the HBA range.

To detect RFI, the AOFlagger (Offringa et al., 2010a,b) was used. However, since this flagger is partly amplitude-based, it is likely that low-level RFI will leak through the detector. Since it is also low-level RFI we are interested in, we will analyse unflagged data and the RFI classified data.

6.4 Results

In this section we present the histograms of the LBA and HBA sets and the results that were obtained by applying the methodology discussed in Sect. 6.2.

Table 6.1: *Data set specifications*

	LBA set	**HBA set**
Observation date	2011-10-09	2010-12-27
Start time	06:50 UTC	0:00 UTC
Length	24 h	24 h
Time resolution	1 s	1 s
Frequency range	30.1–77.5 MHz	115.0–163.3 MHz
Frequency resolution	0.76 kHz	0.76 kHz
Number of stations	33	13
Total size	96.3 TiB	18.6 TiB
Field	NCP	NCP
Amount RFI detected	1.77%	3.18%

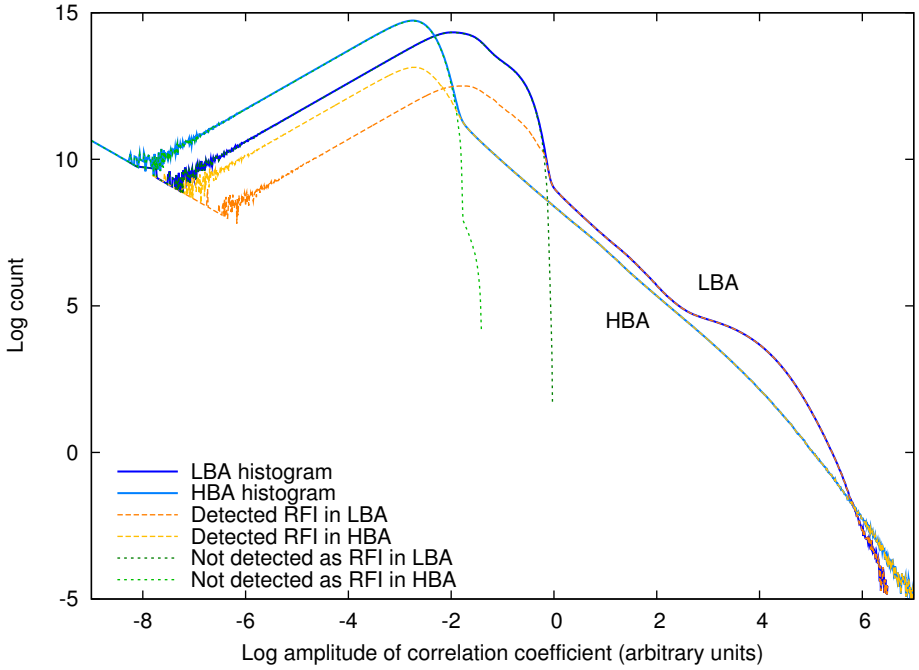


Figure 6.6: *The histograms of the two data sets before pass-band correction and flux calibration.*

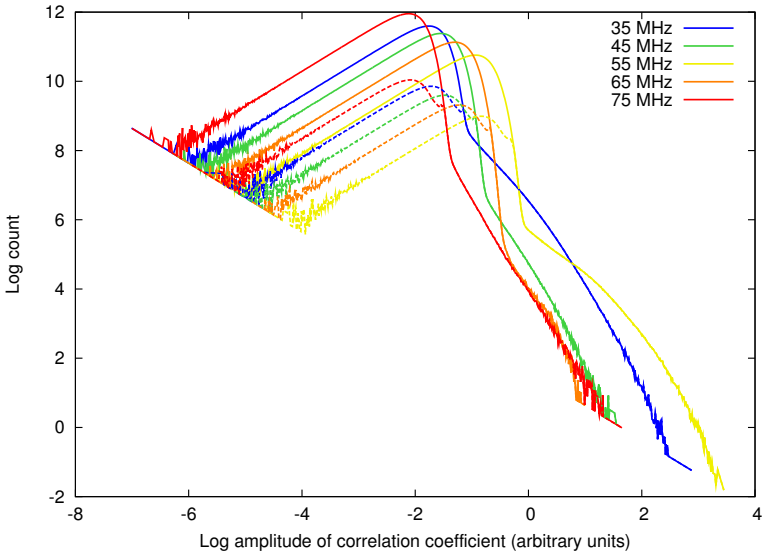


Figure 6.7: *Histograms for* 5 *different* 0.2 *MHz LBA sub-bands without pass-band correction and flux calibration. The continuous lines represent the data before RFI flagging. The dashes lines are the histograms of the samples that have been classified as RFI.*

6.4.1 Histogram analysis

Fig. 6.6 shows the histograms with logarithmic axes for the LBA and HBA sets. In both sets, it is clear that at least one variate with a Rayleigh and one with a power-law distribution have been observed. The left part of the histogram matches the Rayleigh distribution well up to the mode of the distribution. The bulge around the mode of the LBA histogram is wider due to the larger effect of the antenna response as discussed in §6.1.4. As can be seen in Fig. 6.7, the Rayleigh-bulges of individual sub-bands are not that wide, but they are not aligned because of the differing noise levels. This effect is not so strong in histograms of the HBA sub-bands in Fig. 6.8, because the HBA antenna response changes less over frequency.

It is to be expected that the RFI-dominated part of the distributions at different frequencies will reflect the underlying source populations. Both Figs. 6.7 and 6.8 show that the power-law part of the distributions are very different for different sub-bands. Nevertheless, combining the data of all the sub-bands results in reasonably-stable power-law distributions. This indicates that the individual sub-bands have an approximately similar underlying power-law distribution, that is not yet apparent because not enough samples are combined in the histograms of individual sub-bands.

To make sure that the antenna response does not influence the result of the slope, we have also analysed the curves after a simple band-pass calibration. This was performed by dividing each sub-band by its standard deviation after RFI excision. Because the standard deviation of the distribution might be affected by the RFI tail of the distribution, we compare the two histograms to make sure the power-law distribution is not significantly changed. The resulting histograms are shown in Fig. 6.9. This procedure makes the bulge of the LBA histogram similar to the

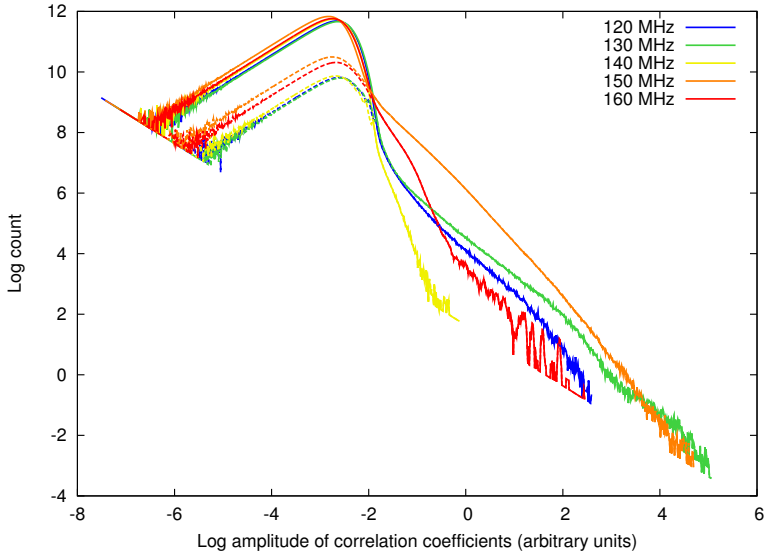


Figure 6.8: *Histograms for* 5 *different* 0.2 *MHz HBA sub-bands without pass-band correction and flux calibration. The continuous lines represent the data before RFI flagging. The dashes lines are the histograms of the samples that have been classified as RFI.*

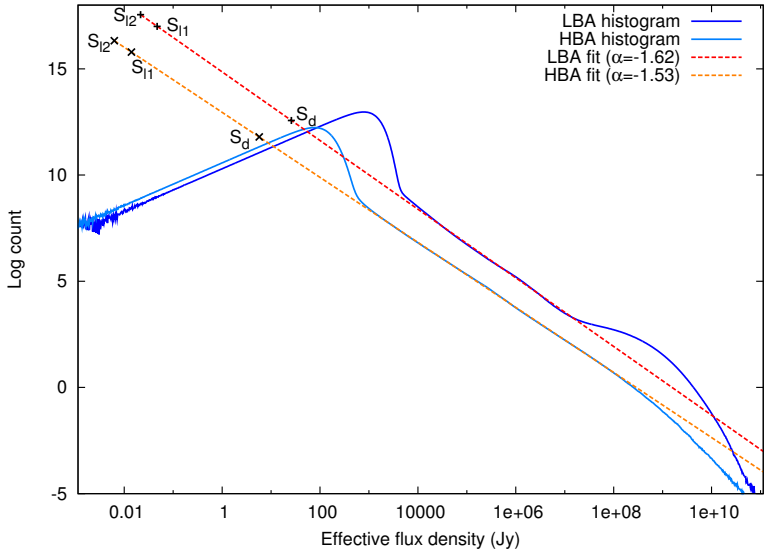


Figure 6.9: *LBA distribution after pass-band correction and flux calibration.* S_{l1} *and* S_{l2} *denote the limits of the distribution with a sharp lower cut-off (Eq.* (6.24)*) and uniform lower limit (Eq.* (6.26)*),* S_d *is the average lower limit of RFI that is detected.*

bulge of a Rayleigh curve and extends the power-law part. Nevertheless, it does not change the log-log slope of the power law in either histograms. This validates that the variable gain that is caused by the antenna response does not change the observed power law. Consequently, it can be expected that other stochastic effects, such as the intrinsic source strength and the beam gain due to a differing direction of arrival, will similarly not effect the power law. Because the pass-band corrected histograms provide a more representative analysis, we will use the corrected histograms for further analysis.

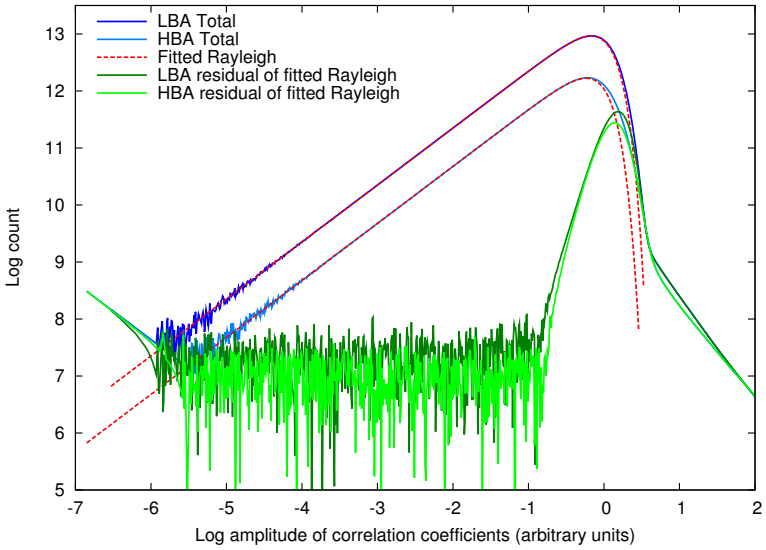


Figure 6.10: *Least-squares fits of Rayleigh distributions to the LBA and HBA histograms, after pass-band correction but without flux calibration.*

The Rayleigh parts of the distributions are plotted in Fig. 6.10, along with a least-squares fit and its residuals. Both histograms follow the Rayleigh distribution for about five orders of magnitude, which is validated by the residuals that show only noise. It breaks down about one order of magnitude before the mode of the distributions. This is because of the multi-variate nature of the distributions, as was described in §6.1.3.

If we go back to Fig. 6.9, we see that in the LBA the power law is stable for about three orders of magnitude, and one order more in the HBA. Fig. 6.11 shows the slope of the log-log plot as a function of amplitude, which was constructed by performing linear regression in a sliding window, with a window size of 1 decade. The HBA shows very little structure in the slope, but the HBA is less stable and shows some features in its power-law part. Linear regression on the power-law part of the log-log plot results in a slope of -1.62 for the LBA and -1.53 for the HBA. These and the other derived quantities have been summarized in Table 6.2. Although the HBA slope does not show any other significant features besides the Rayleigh and power-law curves, the LBA power law ends with a bulge around an amplitude of 10^6. This bulge is caused by a very strong source affecting lots of samples, and is an apparent single outlier in the spatial distribution. We found this is caused by RFI observed for about an hour in the late afternoon in the lower LBA frequency regime, around 30–40 MHz. Leaving this frequency range out flattens the bulge significantly,

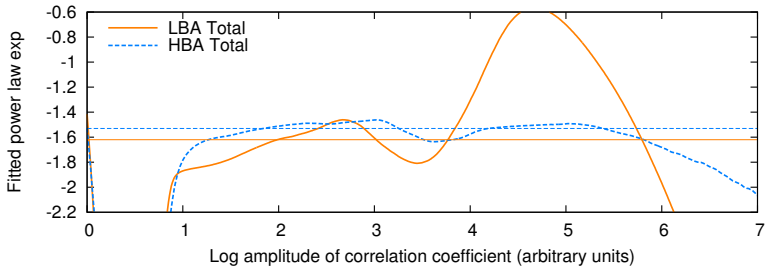


Figure 6.11: *The slope of the band-pass corrected log-log histogram as a function of the brightness. The horizontal lines indicate the fitted slope over the full (semi-) stable region. The horizontal axis is not calibrated.*

but does not completely eliminate it, because the source turned the receivers in non-linear state, causing leakage at lower intensity levels in the other sub-bands. Unlike linear regression, the fitting region of the Hill estimator is not limited at the high end. Consequently, because of the bulge, the Hill estimator evaluates for the LBA into a slope that is less steep, with a value of -1.53. For the HBA set, the Hill estimator is equal to the -1.53 value found by linear regression.

On the assumption that the histogram is zero below amplitude S_L, we find that $S_L = 21$ mJy for the LBA and $S_L = 6.2$ mJy for the HBA (see Table 6.2). If instead it is assumed that the histogram has a uniform distribution below some amplitude $\tilde{S}_L$, we find that the amplitude at which the power-law distribution breaks down is approximately a factor two higher. The two different assumptions on how the power-law distribution breaks down have a small effect on $E(S_{\text{leak}})$, the expected value of the leaked RFI. By using $\tilde{S}_L$ instead of S_L, it is a few percent lower. By assuming a 100% RFI occupancy, we find that the expected value of leaked RFI is 484–496 mJy for the LBA and 167–171 mJy. With 10% occupancy, the value for $E(S_{\text{leak}})$ is about 25% reduced. The RFI occupancy only starts to have a significant effect on $E(S_{\text{leak}})$ if it is well below 10%.

6.5 Conclusions and discussion

We have analysed the histogram of visibility amplitudes of LOFAR observations and found that, within a significant range of the histogram, the contribution of RFI sources follows a power-law distribution. The found power-law exponents of -1.62 and -1.53 for the 30–78 MHz LBA and 115–163 MHz HBA observations respectively, can be explained by a uniform spatial distribution of RFI sources, affected by propagation described surprisingly well by Hata's electromagnetic propagation model. Taken at face value these exponents imply in Hata's model that the average transmitting heights for sources affecting the LBA and HBA are 79 and 13 m respectively. Hata's model only goes down to 150 MHz, and it is possible that the electromagnetic fall-off due to propagation will be different for lower frequencies, e.g. because the effect of the ionosphere becomes stronger. Intervals for the exponents with representative 3σ confidence boundaries are $[-1.80; -1.44]$ for the LBA and $[-1.53; -1.49]$ for the HBA. The estimate of the HBA is thus more accurate, because its histogram deviates less from the power law.

On the assumption that the power-law distribution for RFI sources will continue down into the noise, we have constructed a full parametrization of the RFI apparent flux distribution. By

Table 6.2: *Estimated distribution quantities per data set.*

Symbol	Name	LBA set	HBA set
N_{total}	Total number of samples in histogram	8.0×10^{11}	5.4×10^{11}
σ	Rayleigh mode (assumed to be SEFD/$\sqrt{2\Delta t \Delta\nu}$)	770 Jy	77 Jy
Estimators for power-law distribution parameters			
α	Exponent of power law in RFI distribution	-1.62	-1.53
$SE(\alpha)$	Standard error of α	2.8×10^{-3}	6.9×10^{-4}
α_H	Hill estimator for power-law exponent	-1.53	-1.53
$SE(\alpha_H)$	Standard error of α_H	8.9×10^{-6}	1.0×10^{-5}
ϵ_α	Sampled estimate of standard deviation of α	6.1×10^{-2}	1.2×10^{-2}
β	Scaling factor of power law with exponent α	4.0×10^{17}	3.4×10^{15}
Limits			
S_L	Constraint on lower fall-off point of power law	21 mJy	6.2 mJy
$\tilde{S}_L$	As S_L, but assuming $Ig/r^2 \sim$ uniform	47 mJy	14 mJy
S_d	Expected lowest apparent level of RFI detected	26 Jy	5.7 Jy
$E(S_R)$	Apparent RFI flux density	2700 Jy	140 Jy
$E(S_{\text{leak}})$	Residual apparent RFI flux density after excision	484–496 mJy	167–171 mJy
	Same as above, but by assuming 10% occupancy	384 mJy	120 mJy
REFD	RFI equivalent flux density	18.9–19.3 Jy	6.5–6.7 Jy

assuming that all samples contain some contribution of RFI, we find that the average flux density of RFI that leaks through the detector is 484–496 mJy for the LBA and 167–171 mJy for the HBA (depending on the used method). These values should be compared to the noise in individual samples of 770 Jy (LBA) and 77 Jy (HBA) (see Table 6.2), and are upper limits for what can be expected. If in fact not all samples are affected by RFI, the leaked RFI flux will be smaller, and will of course be zero in the extreme case that the detector has found and removed all RFI.

In experiments such as the LOFAR EoR project, a simulation pipeline is used to create a realistic estimate of the signal that can be expected. Currently, these simulations do not include the effects of RFI. With the construction of empirical models for the RFI source distributions, we are one step closer to including these effects in the simulation. Using Eq. (6.16), one can sample the strength of a single RFI source, add the feature to the data and run the AOFlagger. What is still needed for accurate simulation, is to obtain a likely distribution for the duration that one such source affects the data. For example, it is neither realistic that all RFI sources are continuously transmitting nor that they affect only one sample. The RFI detector is highly depending on the morphology of the feature in the time-frequency domain. Finally, the coherency properties of the RFI might be even more important to simulate correctly, but these have been not been explored. However, these have large implications for observations with high sensitivity. This will be discussed in the next section.

The derived values for the average lower level of detected RFI, S_d, show that the AOFlagger has detected a large part of the RFI that is well below the sample noise. In both sets, S_d is more than one order of magnitude below the Rayleigh mode. This can be explained with two of the algorithms it implements. The first one is the `SumThreshold` method (Offringa et al., 2010a),

that thresholds on combinations of samples, and is thus able to detect RFI that is weaker than the sample noise. The second one is the scale-invariant rank (SIR) operator (Offringa et al., 2012b). This operator is not dependent on the sample amplitude, but flags based on morphology.

6.5.1 Implications for very long integrations

In theory, faint RFI could impose a fundamental limit on the attainable noise limit of long integrations. As an example, we will analyse the situation for the LOFAR EoR project. This project will use the LOFAR high-band antennae to collect on the order of 50–100 night-time observations of 6 h for a few target fields. The final resolution required for signal extraction will be about 1 MHz. The project will use about 60 stations, each of which provides two polarized feeds. This will bring the noise level in a single 6 h observation in 1 MHz bandwidth to

$$\sigma_{\text{eor-night}} = \text{SEFD}\left(2\Delta t \Delta\nu N_{\text{feed}} N_{\text{interferometers}}\right)^{-\frac{1}{2}} \approx 250\ \mu\text{Jy}. \tag{6.35}$$

Therefore, after 100 nights the thermal noise level will be 25 μJy.

Because some RFI sources might be stationary, the signals from these sources will add coherently over time. Therefore, the amount that time integration can lower the flux density of RFI might be limited. Additionally, some RFI sources will be received by multiple stations of the array, and by multiple feeds of the individual antennae. Therefore, integrating data from different interferometers and data from the two polarized feeds might also not bring the noise level that is caused by RFI down. In summary, unlike normal noise, the RFI might be coherent over time, interferometer and feed.

On the other hand, many RFI signals observed in the LOFAR bands have a limited bandwidth. Indeed, the majority of the detected RFI sources affect only one or a few LOFAR channels of 0.76 kHz. Therefore, frequency averaging will lower the flux density of the RFI signal. If the frequency range contains only one stationary RFI source, the strength of this RFI source will go down linearly with the total bandwidth. If we assume that all channels are affected by RFI sources and all transmit in approximately one channel, then the noise addition that is produced by RFI will go down with the square root of the number of averaged channels. This is a consequence of the random phase that different RFI sources have.

In summary, some class of stationary RFI sources are expected to be coherent over time, polarization and interferometer, but not over frequency. Therefore, in this case the noise level at which RFI leakage approximately becomes relevant is given by

$$\sigma_{\text{RFI}} = \frac{\text{REFD}}{\sqrt{2\Delta\nu}}, \tag{6.36}$$

where REFD is the RFI equivalent flux density at 1 Hz and 1 s resolution for a single station, in analogue to how the SEFD is defined. This only holds when the observational integrated bandwidth $\Delta\nu$ is substantially higher than the average bandwidth of a single RFI source. Otherwise, if the $\Delta\nu$ is small relative to the average bandwidth of RFI sources, some RFI might show up earlier. The empirically found upper limits in this work are $\text{REFD}_{\text{LBA}} = 18.9$–$19.3$ Jy and $\text{REFD}_{\text{HBA}} = 6.5$–$6.7$ Jy (see Table 6.2).

For the EoR project with 1 MHz resolution, Eq. 6.36 results in $\sigma_{\text{RFI}} \approx 4.7$ mJy. However, the first EoR results of observations of one day have approximately reached the thermal noise of about 1.7 mJy per subband (Yatawatta et al., *in preparation*), and the resulting images show no

signs of RFI. Clearly, Eq. 6.36 is therefore not applicable to most of the RFI that is observed with LOFAR. In the following section we will discuss effects that could cause a reduced contribution of RFI.

6.5.2 Interference-reducing effects

When integrating data, it is likely that the actual noise limit will be significantly lower than the given upper limit, which was determined on high resolution. There are several reasons for this, which we will summarize one by one.

- Many RFI sources have a variable geometric phase, because they move or because their path of propagation changes. This would cause them to sum (partly) incoherently over time, and thus go approximately down with the noise.

- Many RFI sources will be seen by only a few stations. This would make the histogram go down more quickly at lower amplitudes.

- For the shortest baselines at 150 MHz, the far fields starts around 1 km. Some RFI sources will therefore be in the near field, especially in the longer baselines. In that case, a source will not add up coherently over the interferometers that see the particular source, as the interferometers see them with different phases.

- Many stationary sources are not constant over time, thus will be attenuated somewhat by averaging.

- We have assumed 100% of the spectrum is occupied by RFI. If, say, only 10% of the spectrum is occupied, the expected value of the leaked RFI level decreases with about 25%, and if the detected 2.68% true-positives contain all RFI, there is no leaked RFI at all. With current data, one can only speculate how much the electromagnetic spectrum is truly occupied.

- Stationary RFI sources in a uniform spatial distribution will interfere both constructively and destructively with each other. Individually, the sources will have fixed geometric phases, but because the baselines are much longer than the wavelength, their phases will become uniformly spread. Therefore, they will add incoherently.

Two other effects can improve the RFI situation as well:

- Fringe stopping interferometers can partly average out RFI sources. Nevertheless, stationary RFI that is averaged out by fringe stopping will leave artefacts behind in the field centre (Offringa et al., 2012a). This is not relevant when observing the North Celestial Pole — which is one of the LOFAR EoR fields — because no fringe stopping is applied when observing the NCP.

- The Fourier transform that is involved in data imaging will localize the contribution from stationary RFI near the NCP. If RFI artefacts would show in the image of the NCP field, they can be easily detected and possibly be removed, or processing could ignore data near the pole.

Because of the above two arguments, when considering RFI it is a risk to use the NCP as one of the EoR target fields. At the same time, this field is useful for analysing the RFI coherency properties. It is also a simple field to observe with LOFAR, because it is always at high declination in the Netherlands and it does not contain bright foreground sources. Preliminary analysis of EoR NCP observations of a single night have reached the thermal noise (Yatawatta et al., *in preparation*), but do not show leaked RFI at the pole.

Finally, future RFI excision strategies will further enhance detection accuracy. Currently, RFI excision is applied only on the raw data from single interferometers at high resolution. Once data from a large number of nights are collected, it will be possible to detect and excise RFI more accurately, by looking at the averaged data from multiple nights and/or multiple interferometers. We have shown that the current detection algorithms can detect RFI well below the noise. Therefore, if data from different nights or interferometers are summed, and there is stationary RFI in the set, it will become detectable. All RFI that is below the noise but is not stationary, will act like normal noise and will therefore be harmless.

With the current strategy, it is likely that the LOFAR EoR project will encounter some RFI on some frequencies when averaging lots of observing nights, although this still has to be seen. To mitigate this leaked RFI, the detection can be executed at higher signal-to-noise levels. The current results indicate that a lot of RFI is not coherent, and the situation is promising. Considering the current RFI results, and the availability of further mitigation steps, it is very likely to assume that RFI will not be problematic for the detection Epoch of Reionisation.

Chapter 7

Conclusions & outlook

THE RED LINE throughout this thesis is that the current situation of radio-frequency interference for LOFAR is well manageable. All chapters in this thesis provide a part in this overall conclusion. In particular, we have seen that radio-astronomical observations at low frequencies can be performed in urban areas without the loss of a significant amount of sensitivity or bandwidth. This, however, is not possible without taking some measures during the hardware design and processing of the data. It is also imperative to have a stable frequency allocation, in order to avoid the strongest broadband transmitters. This chapter will briefly revisit the chapters one by one, and combine their conclusions in order to get to the overall conclusions of this work. During this discussion, the future developments and directions that relate to radio astronomy and interference mitigation will be considered as well.

Most of the conclusions are generic conclusions. If a conclusion is LOFAR specific, this will be stated.

7.1 Detection methods

Of all methods to deal with radio-frequency interference, post-correlation detection and mitigation is probably the most important method. It is relatively simple, does not require special hardware, it is used during almost all observations and it is often the only required RFI mitigating method to get to high-quality radio images. Although a lot of research described in the literature deals with generic detection methods, little research has been devoted to the specific purpose of *radio-frequency interference* detection. In the context of interference, most effort has been put into other methods, such as spatial filtering, adaptive cancellation and high-resolution blanking. These techniques are of crucial importance when the channel that is observed is fully contaminated. Nevertheless, these techniques are less demanding compared to the accurate detection methods that are used in almost every observation. Generic methods, such as simple thresholding, have been used regularly for the detection of interference. However, these methods are not very accurate and do not take all known information about the interference to their advantage. Now that modern interferometers start to provide much higher spectral and temporal resolution, accurate and fast detection methods have become even more important. In Chapter 2, this gap in our knowledge of

RFI detection methods is filled, and several novel detection methods are designed and compared, resulting in an much improved accuracy and performance.

In Chapter 2, the detection problem is split in three parts: (i) signal estimation or fitting algorithms; (ii) detection for statistical outliers; and (iii) morphological detection. In past work, the most advanced methods that were used for these steps are polynomial fitting or median filtering; amplitude thresholding; and a morphological dilation, respectively. In Chapter 2, the `SumThreshold` and scale-invariant rank (SIR) operator were introduced. The `SumThreshold` detects statistical outliers by performing combinatorial thresholding, while the SIR operator finds samples that are likely contaminated based on the morphology of the flag mask. These two methods have been shown to significantly enhance the accuracy of the detection, up to a level that it performs as good as manual selection "by eye". For signal estimation, a simple Gaussian low-pass filter is used. While this filter is as accurate as common methods, e.g. median filtering or polynomial fitting, it is also very fast. All these methods are effective in the LOFAR frequency range of 20–250 MHz, but have also shown to work well on higher frequencies.

A useful future direction is to make the signal estimation step robust for any LOFAR observation. Especially in observations with very strong off-axis sources (such as Cassiopeia A, Cygnus A or the Sun at low frequencies in the Northern hemisphere), we sometimes see that the default pipeline flags too much of the data. This is a problem that occurs only in a small range of observations, and can be solved by manually tweaking some parameters of the pipeline. Nevertheless, projects that survey a large part of the sky and thus target hundreds of directions simultaneously, require a completely automated reduction process. An example of such a project is the LOFAR MSSS project. Currently, the LOFAR flagger uses no prior knowledge about the target field. To automate the process successfully in all cases, the flagger could use knowledge about the field direction and the relative position of the strongest sources, and adopt the flagging strategy accordingly. Alternatively, one can flag everything with the default strategy, and detect observations with a relative high number of flags. Because the RFI occupancy is quite stable, a high number of flags indicates most likely a problem due to off-axis sources. If this case is detected, the set can be flagged again with a strategy that makes less assumptions about the temporal variation of the visibilities.

The SIR operator that was introduced in Chapter 2, is a new generic morphological operator. Other morphological operators, such as the morphological dilation, erosion, opening and closing operators, have shown to be applicable in a wide range of applications. Therefore, the use of the SIR operator might not be limited to an astronomical context. The SIR operator algorithm is closely related to the all maximal sub-sequence sum algorithm, which is used in protein and DNA sequence analysis (Karling and Brendel, 1992). The SIR operator might therefore also be relevant in that application. Another application where the algorithm might be useful is image processing.

7.2 Pipeline efficiency

The individual algorithms that were introduced in Chapter 2 are important building blocks for a fully automated pipeline. However, to let such a pipeline work robustly and successfully, the individual algorithms need to work together. Some of the input parameter need to be tweaked, while others need to be determined automatically. This is discussed in Chapter 3. The resulting pipeline is optimized for the LOFAR case specifically. However, optimizing the pipeline for a telescope with a different frequency range or different time and frequency resolution requires

adjustments of only two or three of the parameters.

The culminating pipeline, named the "AOFlagger[1]", that was formed from several previously-introduced algorithms, consists of an iterative procedure. During a single iteration, the algorithms starts by estimating the signal very roughly and subsequently finds the very strong outliers. This is followed by two iterations that successively increase the detection sensitivity. Finally, the morphological procedure is applied on the full flag mask. This procedure has been extensively tested, and is now the recommended way to flag any LOFAR imaging observation.

The AOFlagger is integrated in the default LOFAR pipeline (NDPPP), to minimize the number of times data needs to be reread from disk. By extensive profiling of the flagger code, the largest computational bottlenecks were identified. After many optimizations (e.g., see Appendix A), a point has been reached at which the pipeline can process observations faster than the data can be read from disks. In other words, the pipeline is input-output (IO) dominated. This is important to note, because it implies that further optimizations will not improve the speed of the flagger much further. Nevertheless, a few of the processing steps still rely on initial "experimental" (and thus unoptimized) code, such as the Gaussian smoothing algorithm. That step could be significantly increased with the use of optimized code and single-instruction-multiple-data (SIMD) instructions. However, because profiling showed that it is a few times faster than the heavily optimized `SumThreshold` method, improving it will only have a very slight repercussion — even when the pipeline would not be IO-dominated.

Nowadays, graphical processing units (GPUs) are often used in computational tasks. They are attractive, because they are relatively cheap for the amount of operations they can perform per second. On the other hand, it takes a lot of work to implement algorithms on GPUs, and certain algorithms are not efficiently convertible to GPU code. Currently, the AOFlagger can not make use of GPUs. It would be a lot of work to implement algorithms such as the `SumThreshold` to GPU code. At the same time, this will not improve the situation a lot, because the flagger is IO-limited as was discussed above. Moreover, the AOFlagger code is run on many platforms and different clusters, of which only a few provide GPUs that can be used for generic computations. The usage of GPUs for RFI detection might become more relevant in observatories where flagging has to be done on-line and in real-time.

The situation might of course change when faster hard disks become available in the future. A significant increase in IO-speed can be achieved by using solid-state drives (SSD). This uprising technology currently shows an increased speed of an order of magnitude over conventional hard disk drives (HDD), and the technique is continuously further improved. However, SSDs are also significantly more expensive per unit of storage size. Moreover, the AOFlagger step is only a small step of the full processing step that is applied on observations, and the calibration and imaging steps typically take an order of magnitude more time. All in all, instead of further optimizing the pipeline, at present it is more useful to focus on improving the calibration and imaging algorithms. Nevertheless, the importance of efficient IO should not be forgotten.

These conclusions are very important for future observatories that will deal with even higher data volumes compared to LOFAR. The Square-Kilometre Array (SKA) will correlate more than a thousand stations, while it will at the same time observe with more bandwidth and an order of magnitude higher time resolution. This implies that, to maintain real-time flagging speed, the IO should be accelerated by at least four orders of magnitude compared to LOFAR. Where LOFAR reads the data from approximately 100 HDDs during flagging, the SKA should distribute its data

[1] This name was introduced by Ger van Diepen when the first version of the flagger code was committed to the LOFAR software repository.

over millions of HDDs of similar speed to be able to perform the same strategy. Of course, the use of faster drives, such as SSDs, could lower this value. Providing an infrastructure that can distribute and transport the data from millions of HDDs efficiently will be a major challenge, and a lot of work on distributed computing for astronomical purposes is required before this can be realized.

Focal-plane arrays, such as the Apertif receivers that will soon be installed on the WSRT dishes, provide a different case where large volumes of data need to be flagged. Apertif will provide 37 beams on the sky at once, with 16,384 frequency channels of 18 kHz[2]. On the other hand, LOFAR provides 62,464 channels of 0.76 kHz, with an order of magnitude more baselines. Therefore, to first order, the LOFAR flagging strategy has also sufficient performance for Apertif. The AOFlagger will probably work as good on Apertif data as it does on LOFAR data, although it does not use all information that multi-beam systems provide to differentiate RFI from signal. An enhanced version of the `SumThreshold` method could take into account that a high signal in only one feed is less likely to be RFI. Potentially, this can improve the detection accuracy further. In Flöer et al. (2010), a RFI detection scheme for the multi-beam system of the Effelsberg telescope is shown, that makes use of the information from the various feeds. However, it is to be seen whether a detector can efficiently combine information from three dimensions (time, frequency and beam or feed). Another possibility of multi-beam systems are the spatial filtering techniques that have been successfully applied on the Parkes telescope (Kocz et al., 2012). This technique can be applied on the focal-plane arrays of individual telescopes.

7.3 Filters

Processing power is a major concern for modern telescopes. Low-frequency telescopes such as LOFAR have a wide field of view, which complicate the calibration process. An advanced calibration scheme is required to retrieve accurate parameters of celestial sources (such as flux density and polarization). Once a source has been calibrated, it can be subtracted from the data to show underlying emission. This is required to reach the full sensitivity of the instrument.

In Chapter 4, new techniques that help to improve the data quality of interferometric radio observations are considered. Filters are used to attenuate radio-frequency interference and off-axis sources without calibrating them. Several new filters are introduced and tested. A low-pass filter in time and frequency direction on a single baseline data is successfully used to lower the noise in the area of interest and to remove sidelobes coming from unmodeled off-axis sources and RFI. Chapter 4 analyses related side effects of data integration, averaging and gridding, and show that these can cause ghosts and an increase in noise, especially when using long baselines or interferometric elements that have a large field of view. Initial tests show that the filters can be several factors faster compared to common source separation techniques such as peeling and a variant of peeling that is currently being tested on LOFAR observations called “demixed peeling” (the LOFAR cookbook, Pizzo (2012, Chapter “Demixing”), Jeffs et al. (2006)).

The filter that is considered to be most attractive, is a low-pass filter applied on the visibilities after regular calibration. Such a filter can subtract the residuals of sources that are not perfectly modeled and subtracted. It is fast and uses information in the data that complements the information that is used during calibration, and thus the filter complements calibration and subtraction. Currently, this filter has only been used on WSRT data, where it shows good results. Applying the

[2]See `http://www.astron.nl/general/apertif/apertif`.

filter on LOFAR data requires some work, because the process needs to be efficiently distributed over the nodes. During most other tasks, each node processes a single subband. However, because the filter requires information from all sub-bands at once, the nodes need to communicate efficiently. We will discuss this in §7.6.

Chapter 4 also discussed stationary RFI that is not flaggable, because it is too weak or broadband. Such RFI sources are very hard to subtract with techniques such as peeling, because they can be very weak and time variable. When integrating these data, a stationary source might add coherently. In such a case, the RFI source will mimic a celestial source near the North Celestial Pole (NCP). This effect is not yet seen in LOFAR observations, but could appear when integrating longer observations. When such RFI appears, the presented filters could be useful to excise the RFI from the data.

7.4 LOFAR & RFI

In Chapter 5, the LOFAR radio environment is analyzed. These analyses are of high importance for future observing strategies of LOFAR. Moreover, since LOFAR is build in an urban area and operates at low frequencies, these results provide new insights in the general effects of an urban radio environment on radio-astronomical observations.

The radio environment is mapped with the help of two RFI-surveying observations of 24 h. One observation covers 30–78 MHz in the LBA frequency range, while the other observation covers 115–163 MHz in the HBA frequency range. Both surveys show a very small RFI occupancy of 1.8% for the LBA and 3.2% for the HBA. Indeed, the LBA range shows less RFI sources in observations, but even the HBA spectrum is largely unoccupied.

Several reasons are given for the small impact of RFI on LOFAR: (i) LOFAR's high time and frequency resolutions of 1 s and 0.76 kHz respectively, minimize the amount of data loss caused by interfering sources; (ii) the AOFlagger interference detection pipeline has an unprecedented accuracy; (iii) LOFAR's hardware is designed to deal with the strongest interfering sources that are found in its environment; (iv) in contrast to dishes with feeds in the focal point, the receiving elements of LOFAR are close to the ground; and (v) there is no evident self-contamination, i.e., RFI that is generated by the hardware itself.

Below 163 MHz, there are no large ($\geq$ 1 sub-band of 200 kHz) gaps in the VHF spectrum caused by RFI, with the obvious exception of the FM broadband frequency around 87–108 MHz. These frequencies are filtered in hardware. However, the frequency range of 216–230 MHz contains terrestrial digital audio broadband (T-DAB) transmitters. These kind of transmitters are highly problematic devices for radio astronomy, because they transmit continuously within their full bandwidth. The frequency range of 174–216 MHz is also allocated for T-DAB services, but the range is (still) mostly unused in the Netherlands. Consequently, for now this range can be used in LOFAR radio observations. An upgraded service called DAB+ might be rolled-out at some point. It is to be seen if and when DAB or DAB+ services will extend their frequency coverage.

Another change in the radio environment might be caused by new wind turbines. Currently, plans are formed to build wind turbines in the area of LOFAR. These turbines might themselves generate harmful radiation, but the propellers might also reflect other RFI sources that are normally well below the horizon. It is not yet clear what kind of effects nearby wind turbines will cause on radio astronomical observations.

During the construction phase of new telescopes such as the SKA, the radio environment has played a very important role in site selection. With the analysis of the LOFAR environment, we have proved that the amount of RFI below 170 MHz is — with some pre-cautions and after proper flagging — negligible, even in a country as populated as the Netherlands. Of course, this excludes the FM broadcasts in the frequency range 87–108 MHz, which is filtered in hardware. The "price" of building a telescope in a non-remote location, caused by radio-frequency interference, is therefore mostly limited to losing the 87–108 MHz FM bands and the DAB bands above 170 MHz (currently 216–230 MHz, but in the future possibly 174–230 MHz). This loss should be compared with the extra costs and logistics associated with remote locations, such as the necessity of building new infrastructure. Whether the given frequency ranges are permanently lost is as of yet unknown, as it might even be possible to excise FM and DAB signals from the data. Filters that exploit the spatial or cyclo-stationary properties of the RFI signals are considered, and might be able to recover the cosmic spectrum. Although the loss in frequency ranges that are not contaminated by broad-band broadcasts will also be somewhat reduced in remote locations, this loss is already negligible. Moreover, even the most remote locations will have some RFI that is caused by satellites and air-traffic. Therefore, it will remain necessary to deal with RFI during hardware design and by applying flagging strategies, even in the case of SKA. Consequently, for future observatories, the RFI environment should be checked for very strong and broadband sources, but other kinds of RFI sources should not have a lot of weight in site selection.

7.5 RFI implications for reionisation experiments

Several projects are currently underway to detect redshifted neutral hydrogen from the Epoch of Reionisation (EoR). One of those projects is the LOFAR EoR project. These experiments require very high point-source sensitivities of a few tens of μJy. In the LOFAR case, this is to be achieved by integrating 50–100 nights of LOFAR observations. The project is concerned about low-level stationary RFI, that could coherently add over many nights. Such RFI could in theory place a lower limit on the sensitivity, thereby making it impossible to detect the signal from the EoR.

In Chapter 6, the probability of stationary low-level RFI is explored. The RFI surveys from Chapter 5 are used for the analysis. It is found that the amplitude distribution of the RFI sources matches a uniform distribution of RFI sources on the surface of the Earth, affected by propagation described surprisingly well by the electromagnetic propagation model of Hata (1980). With worst-case assumptions, it is found that RFI that leaks through the detector has an average flux density of approximately 490 mJy for the LBA and 170 mJy for the HBA in a 1 kHz / 1 s sample. These values should be compared to the noise in individual visibility samples of 770 Jy (LBA) and 77 Jy (HBA). This confirms that the AOFlagger RFI detector is very accurate.

To which extent RFI can add up coherently is yet unknown. An important argument in this discussion, is that current observations are not showing any effects due to leaked RFI. This even holds for the North Celestial Pole (NCP) field, that is one of the LOFAR EoR target fields. If stationary RFI would leak through the detector and would add coherently, such RFI would generate a fake source near[3] the NCP. The first NCP observations have reached the current thermal noise limit of about 250 μJy in 1 MHz bandwidth after an integration time of 6 hours. No fake

[3]Radio observations are normally referenced in the sky frame of the year 2000 (J2000), but because of precession of the Earth the pole has slightly moved since then. A stationary source on Earth shifts by 20 arcsec/year in the J2000 reference frame.

sources caused by RFI are found in the resulting images, thus most of the leaked RFI must add incoherently. Currently, RFI flagging on the highest resolution is the only excision technique that is performed. If a contribution from RFI sources appears at the NCP after longer integration, it can be suppressed with more advanced RFI excision techniques. Consequently, it is unlikely that stationary RFI will limit the detection of signals from the EoR.

7.6 Data distribution & file formats

Measurement sets are normally stored in the order polarization, channel, baseline, time step, and finally these are distributed based on sub-band. The AOFlagger needs all data per baseline, thus it needs the data in a different order than in which it is stored. It was found that for observations with large data volumes, it is significantly more efficient (typically a factor 2–4, but depending on size of observation and platform) to read the observation from start to end and simultaneously rewrite the set to disk in a different order. This is caused by the physical properties of the hard disks and the installed file system. The performance of a hard disk degrades considerably when reading small chunks from many different locations.

Another related technical topic is the distribution of data. Currently, applications that need data from different sub-bands at once, need to implement some way to transport the data efficiently from several nodes to one (or multiple) locations. Efficient communication between cluster nodes is something which is often needed when processing high-volume radio observations, and it will be beneficial if more generic solutions become available. A solution that transparently implements distributed node access, such as clustered file systems, is easy to implement because they already exist. However, these generate an enormous amount of network traffic no matter how the data is accessed. Therefore, the current standard way of processing, in which each node processes a single sub-band, will be seriously slowed down by such a solution. Standardized communication interfaces such as the Message Passing Interface (MPI) make the process easier as well, but implement the communication on a rather low level. A program that performs a simple operation, such as removing the flags from a full observation, and which implements this with MPI, will still have a rather complex implementation (although, of course, MPI simplifies the implementation). A software library that allows access to the data at the same level as the Casacore Measurement Set library, but is aware that an observation can be distributed over several nodes, would make it much easier to implement tools that can deal efficiently with these observations.

The HDF5 file format has been suggested as a possibility for the LOFAR data format (Anderson et al., 2010), that could potentially improve the situation of data distribution. HDF5 adds a layer between the file system and the application, such that data can be easily stored in a hierarchical structure. Although efforts are made to standardize the way to store astronomical data inside the HDF5 format, it is far from being as complete as the CASA measurement set format for interferometric data, especially on the level of measurement units and their conversion and tools to process astronomical data. In fact, HDF5 abstracts only the lowest level of data access, but forces the data into a hierarchical structure. Consequently, by itself it does not yet provide a solution for fast distributed access, and to a large end the responsibility remains on the side of the applications to implement this.

During the course of this PhD project, I have been asked at least a dozen of times whether I could make the AOFlagger work on a different file format – in particular to let it work with UV FITS files and Miriad files. I started to work on some UV FITS file format support, but soon

realized that support for all stored meta data would be too much work (especially for a format that might become deprecated). Whatever the solution to fast distributed access to data will be, a common interferometric file format that is adopted by all observatories and software packages would be a major improvement. Each of the current file formats, i.e., the CASA MS, UV FITS and the Miriad file formats, provide different advantages (and liabilities), and it might be impossible to wrap every feature into one ultimate file format. Nevertheless, radio observatories should have the same needs for a file format, and a common denominator (or most popular format) should be just good enough.

A yet-unexplored area is the compression of radio data. Reducing the file size of interferometric data would not only decrease the required storage space, but also make the data much easier to handle. Although interferometric data is often dominated by the noise, and noise is incompressible, it might be possible to (partly) filter the irrelevant noise from the data. To this end, the low-pass filters that were presented in Chapter 4 could be useful. This can be combined with a different quantization or encoding scheme to optimize the required space. This approach would be similar to popular compression schemes for audio, such as MP3 and the Free Lossless Audio Codec (FLAC). If this would lead to a similar effective decrease of a few factors in size, it would be a huge benefit to the radio astronomical community.

7.7 Main thesis questions

During the introduction in Chapter 1, the aims of this thesis were summarized in the form of four questions. Now that all the conclusions of this work have been collected, we have reached the point at which those four questions can be answered.

- *What existing methods can one use to excise radio-frequency interference in LOFAR observations?*
  In §7.1, it is concluded that detection is the most important RFI excision technique for LOFAR. However, pre-existing techniques, in which median statistics or polynomial fits were used to estimate the signal, and were followed by normal thresholding, were found to be neither fast nor accurate. We have not yet seen stationary, broadband RFI in LOFAR data. Therefore, the methods that are aimed to mitigate this kind of RFI are not (yet) relevant. In particular, this concerns the singular value decomposition (SVD) mitigation method introduced by Pen et al. (2009) (see §2.2.7) and the fringe fitting method introduced by Athreya (2009) (see §4.2), both of which were useful for GMRT data.

  For experiments that have to make use of the frequency ranges 87–108 MHz or 216–230 MHz, which are permanently occupied by broadband transmissions from the FM and DAB broadcasts respectively, detection might not be sufficient. Both these ranges are outside of the optimized frequency range of LOFAR. This especially holds for the FM frequencies, which are attenuated by hardware filters. Nevertheless, they can be observed. For observing in permanently occupied ranges, spatial or cyclo-stationary filters might be useful. Applying such filters on real LOFAR data will require further research.

- *Can the accuracy and performance of currently available interference excision methods be improved?*
  Chapters 2, 3 and 4 test several new ways of excising RFI. Chapters 2 and 3 deal with detection, while Chapter 4 deals with filtering.

For detection, two important new methods were developed: the `SumThreshold` (§2.2.6) and the scale-invariant rank (SIR) operator (§2.4.1). In combination with a signal estimation algorithm, together these methods find all of the RFI that is apparent when data is visually inspected. Therefore, these methods are significantly more accurate than pre-existing methods (e.g., see Fig. 3.4 on page 67 for comparison with the MAD flagger). For signal estimation, we suggest to replace polynomial fitting and median filtering with a trivial Gaussian low-pass filter (§2.2.2). Such a filter is faster and shows the same accuracy. The full pipeline that combines all these techniques, has been thoroughly optimized and, in fact, is faster than hard disks can deliver the data. The implementation of this pipeline, named the "AOFlagger", is one of the results of this thesis. It is used by default on all LOFAR imaging observations, and is being tried by several other astronomers for other observatories.

For filtering, the SVD and fringe filtering techniques were extended with a filter that can remove from an observation any contribution that does not correspond with a source in the field of interest (§4.3.1 and §4.3.4). This is not only useful for RFI removal, but might also be useful for removing off-axis sources. So far, these filters have shown good result on real WSRT data, but need to be tested more extensively to see whether they work in the LOFAR case.

- *What are the observational consequences of building LOFAR in a populated area?*
  In Sect. 7.4, it is concluded that the effect of the LOFAR radio environment is very benign. Because LOFAR is in a populated area, some existing infrastructure could be used. This is probably a stronger weighting argument compared to the small increase of RFI. The most apparent down-side of building LOFAR in a populated area, is that observations in the frequency ranges 87–108 MHz and 216–230 MHz are not possible without further mitigation techniques, because of the FM and DAB broadcasts. However, even extremely remote areas are not free of satellites, air-traffic and possibly the FM and DAB broadcasts.

  Almost all observations that exclude the broadband broadcasting frequencies need no further post-processing after running the AOFlagger.

  The radio environment is not the only site condition that influences the data quality. Another important parameter for low-frequency astronomy, is the ionospheric condition. Its effect are much more pronounced in the calibration and imaging stages, and are often related to the total electron content (TEC) value. Currently, it appears that the TEC quantity has a significant impact on the quality of the data. In general, the ionospheric conditions are better away from the equator. Therefore, at 53° latitude, the Netherlands is also in this aspect a good place for low-frequency radio astronomy.

- *Will RFI cause a limit on the sensitivity with which LOFAR can observe?*
  With the results so far, it is unlikely that RFI is going to be a major issue for experiments involving long integrations, e.g., very deep extra-galactic surveys or the LOFAR EoR project. Thus far, it appears that for RFI to present a fundamental limit, it has to be stationary (i.e., fixed to the Earth). Such RFI will end up near the North Celestial Pole (NCP). Current observations are reaching the thermal noise limit, and are not showing any such effects. This holds even near the NCP, where fringe rotation does not quench the RFI. At the same time, it is possible to increase the RFI detection rate by applying the excision methods on integrated data, if such RFI were to show up. Consequently, the prospects for LOFAR to detect the epoch of reionisation are excellent.

7.8 Looking forward

For now, the problem of RFI with LOFAR *is largely solved.* However, several open and interesting questions have been posed during this thesis. Moreover, the RFI environment of LOFAR might change because of a change in frequency occupancy and the installation of wind turbines near the stations. We will briefly summarize the posed questions that are interesting future research directions.

- *Can the AOFlagger be generalized to make it work on data from any observatory?*
  There are many observatories, each having its own unique specifications. The flagging accuracy is depending on these specifications, such as the observational frequency range, frequency resolution, time resolution, antenna types and feed types. Often, simple methods such as thresholding are used — sometimes combined with polynomial fitting or median filtering — and are applied manually by the astronomer. For them, it would be significantly easier and faster if a single pipeline could accurately flag the RFI, such as is now the case for LOFAR. For certain observatories, such as the EVLA, the GMRT and the WSRT, it is easy enough to optimize the flagging parameters with the `rfigui` platform (see §5.2.1). However, for single dish, aperture arrays and for the SKA, that will produce significantly larger data volumes, more fundamental changes might be required.

- *Is it useful to apply compression schemes on observations?*
  Currently, compression of radio observations has not been addressed at all in the literature. Nevertheless, the data volumes and the associated storage costs grow with the increasing angular resolution of modern radio observatories. Consequently, if it would be possible to compress radio observations without losing the required time and frequency resolutions, it might significantly decrease the storage costs. Moreover, it makes it easier to handle the data. The methods that are discussed in §7.6 are an attractive direction for further research.

- *Will the visibility low-pass filters (§4.3.1) enhance LOFAR observations?*
  For now, the filters of §4.3.1 have only been applied to a WSRT observation. It is not yet known how much these filters can help to remove artefacts in LOFAR observations. The WSRT tests use only 2.5 MHz of bandwidth. LOFAR can already observe with 48 MHz continuous bandwidth, and soon this might be extended to 96 MHz. Because the effectiveness of these filters are largely depending on the bandwidth, a larger bandwidth will increase the effectiveness of the filters. The filters are very fast when compared to current methods, such as peeling and demixing. If they are effective, they are an attractive alternative (or complementary method) for off-axis source subtraction.

- *What are the coherency properties of RFI?*
  We have seen that RFI that leaks through the detector, does currently not add coherently. In §6.5.2, several possible arguments are given for this apparent incoherent behaviour of leaked RFI. Which of these arguments is the most dominating effect is not known. It is also unclear whether there will be a sensitivity level at which the RFI does show up. The NCP target field of the LOFAR EoR project will give an unprecedented opportunity to analyse the RFI coherency properties. Potentially, this would give better insight in the behaviour of RFI in astronomical observations. This might improve RFI excision and could enhance the observing strategy of LOFAR. It could even have impact on the design parameters of future telescopes.

- *What other applications could benefit from the SIR operator?*
  In this thesis, a new morphological operator was introduced, called the scale-invariant rank (SIR) operator. For this operator, a very fast algorithm with linear time complexity was developed. This operator is used for the specific case of RFI detection. However, because it is a very generic method that relates to other widely-employed operators, such as the morphological dilation and maximum sequential sum algorithm, it might be relevant in other fields.

- *What other signal processing techniques can enhance radio observations?*
  For this thesis, a lot of signal processing techniques were combined and improved. Signal processing techniques are still improving, and future techniques might provide (better) solutions for current problems. The filtering techniques for off-axis sources and compression techniques have already been mentioned, but are only two of many possibilities. The signal processing possibilities with the current state of technology are higher than ever, and improving receivers, correlators and calibration and imaging techniques is utterly important for the advance of radio astronomy.

Finally, in the closing words of this thesis, I would like to address the future prospects of LOFAR. LOFAR has only just started to explore the low-frequency Universe. As must be clear from this thesis, this fantastic instrument is working and in an excellent position, both geographically and symbolically speaking. Soon, LOFAR will engage into an unexplored parameter space of our Universe. Without doubt, this is going to be an exciting time!

Appendices

Appendix A

Technical details of the SumThreshold method

(The contents of this chapter are to be published as a technical report)

IN THIS appendix, the SumThreshold method is briefly described, and implementation details are given. The algorithm that is considered to be optimal will be given, and details of how to vectorize it with the SSE instruction set will be discussed.

A.1 Problem statement

Consider a data set consisting of a sequence of samples. The samples contain noise sampled from some distribution, and occasionally a feature of unknown intensity and length. The SumThreshold method is an algorithm for detecting such features, including its start and end position.

The method is introduced in Offringa et al. (2010a), where it is shown to be useful for detection of radio-frequency interference (RFI). For this case, it is applied separately in the time and frequency directions at high resolution. A pipeline using the SumThreshold method was described in Offringa et al. (2010b).

Paraphrasing Offringa et al. (2010a), the input of the SumThreshold method is a one dimensional sequence of values and its output is a binary mask of samples in which features are detected. If the input contains a consecutive sub-sequence with M samples, for which the average of the sub-sequence exceeds a threshold function $\chi(M)$, this sub-sequence will be selected in the mask. However, an added requirement is that a sample that exceeds some threshold $\chi(M)$, should not be used when testing larger sub-sequences. For example, if $\chi(1) = 1$ and $\chi(2) = 0.7$, then the sequence $[0, 3, 0]$ produces an output mask in which only the number 3 has been flagged, even though the size 2 sub-sequence $0 + 3 > 2\chi(2)$. Because the number 3 will be masked by the threshold limit of $\chi(1)$, it will be replaced by the sub-sequence averages (which is 0 here), and

the sub-sequences do not exceed the other thresholds. Table A.1 lists a few examples.

Table A.1: *Example outputs with* $\chi(1) = 1$, $\chi(2) = 0.7$ *and* $\chi(4) = 0.5$

Input					Output mask				
0,	0,	3,	0,	0	_	_	X	_	_
0,	0.9,	3,	0.9,	0	_	X	X	X	_
0,	0.9,	0.9,	0.9,	0	_	X	X	X	_
0.5,	0.9,	0.9,	0.9,	0.5	X	X	X	X	X

A.2 Algorithm

The method has some correspondence with the scale-invariant rank (SIR) operator. The SIR operator can also be said to detect features, though in a binary mask. It detects sub-sequences in which the ratio of masked values exceed the threshold. On the other hand, the `SumThreshold` method works on continuous values, but otherwise also detects sub-sequences in which the average value exceeds the threshold. If we would assign the values '0' and '1' to the possible binary values in the definition of the SIR operator, the `SumThreshold` seems to select the same sub-sequences.

However, the two methods differ on one important point: the SIR operator is meant to extend the mask beyond the original (binary) features. The purpose of the `SumThreshold` is to exactly select the feature. This difference is expressed by the iterative definition of the `SumThreshold` method: first, the least strict threshold is applied on single samples. Next, a slightly stricter (more sensitive) threshold is applied on the sum of size-two sub-sequences, but individual samples that were already detected in the first round, do not trigger the second iteration. In the third iteration, samples that were detected in the first or second will not trigger size three sub-sequences, etc.

This difference has a major consequence on the algorithm. While the SIR operator can be implemented by three passes over the data, we have not been able to find a similar fast algorithm for the `SumThreshold` method. The `SumThreshold`'s inherent iterative definition requires a pass over the data for each sub-sequence size. Therefore, if every sub-sequence size needs to be tested, one has to test $\mathcal{O}(N^2)$ sub-sequences, which requires the same time complexity. For some applications, a $\mathcal{O}(N^2)$ algorithm might be sufficient, but in the case of LOFAR, such an algorithm would be too slow. To overcome this problem, we allow a slight decrease in accuracy by constraining the number of sub-sequence lengths.

A.2.1 Constraining the tested sub-sequence lengths

We consider two relaxations in the the tested sub-sequence lengths, that increase the efficiency of the algorithm. Both relaxations are constraints on the tested sub-sequence lengths. The first constraint is to only consider exponentially increasing sub-sequence sizes, e.g. sizes of $[1, 2, 4, 8, 16, \ldots]$. This decreases the time complexity to $\mathcal{O}(N \log N)$, while having only a benign effect on the accuracy. This is because it is likely that features of non-tested sizes will be detected by one of the tested sizes, e.g., a feature of size 3 is likely detected as two features of size 2 if the

feature is strong enough. If it is not strong enough, it might still be detected as a feature of size 4. In this case, one falsely detected sample is included, but in the RFI detection case, it is more important to flag possible features, than minimizing false-positives. Features whose total average is larger than $\chi(3)$ and, together with their two neighbouring samples, are smaller than $\chi(4)$, are not detected.

The second constraint is to not consider sub-sequence sizes larger than a given size limit. In the LOFAR pipeline, we only consider features up to 1024 samples. Features larger than that are probably detected by one of the smaller sub-sequence tests, but one should note that a class of very extended and faint (but detectable) features are now ignored. This further optimizes the efficiency of the algorithm such that it has linear time complexity. A linear time complexity is particularly important in real-time environments.

All in all, by only considering exponentially increasing sub-sequence sizes up to 1024 samples in size, we have to perform 11 iterations of the algorithm, and each iteration has linear time complexity. We will now consider how to efficiently perform the individual iterations.

A.2.2 A single `SumThreshold` iteration

The following steps efficiently implement a single `SumThreshold` iteration:

- Slide a window over the data, with size equal to the sub-sequence size M to be tested in this iteration.
- Maintain the sum and the number of unflagged samples in the window. In particular, when moving the window one sample to the right:
  - If the sample to the right was not flagged in previous iterations, add it to the sum and increase the counter.
  - If the sample to the left was not flagged in previous iterations, subtracted it from the sum and decrease the counter.
- For each window position, the average can be calculated by dividing the sum with the counter. If this average exceeds the threshold χ, flag all samples in the window.

Listing 1 performs a single iteration.

Listing 1: Calculate one iteration of the `SumThreshold` mask

Require: χ is the average value threshold for sub-sequences of size M,
$x(0 \dots N-1)$ is the input sequence of N values,
$y(0 \dots N-1)$ is the mask generated by the previous iteration.
Ensure: $y(0 \dots N-1)$ contains the output mask
$z \leftarrow 0, i \leftarrow 0$, count $\leftarrow 0$
$t(0 \dots N-1) \leftarrow y(0 \dots N-1)$
while $i \neq M$ **do**
if NOT $y(i)$ **then**
$z \leftarrow z + x(i)$
count $\leftarrow$ count $+1$

```
        end if
        i ← i + 1
    end while
10: while i ≠ N do
        if z > χ × count OR z < −χ × count then
            t(i − M ... i − 1) ← set
        end if
        if NOT y(i) then
15:         z ← z + x(i)
            count ← count +1
        end if
        if NOT y(i − M) then
            z ← z − x(i − M)
20:         count ← count −1
        end if
        i ← i + 1
    end while
    y(0 ... N − 1) ← t(0 ... N − 1)
```

We note that the algorithm is not as fast for all cases. In the case that many large sub-sequences need to be flagged, the statement in line 12 (which will actually extend into a loop) will iterate over all samples in the window to flag those, for each window position. To make the algorithm fast even in such cases, an extra variable can be added to register the most recently flagged sample. If this value is larger than the start of the window to be flagged, only samples after the most recently flagged sample need to be set. Adding such a variable makes the speed of the algorithm less dependent on the number of samples to be flagged, thus is probably recommended in most cases. It proves to be more difficult to solve this in the vectorized algorithm however.

When optimizing and profiling the LOFAR RFI detection pipeline, it was found that this `SumThreshold` algorithm was dominating the computation time of the RFI pipeline. Therefore, the algorithm was vectorized, which will be discussed in the next section.

A.2.3 Using SSE instructions for vectorization

We have implemented a vectorized version of the algorithm that can compute the `SumThreshold` over multiple sequences at once. In the case of RFI detection, this is very useful, as the algorithm needs to be applied on all time steps and frequency channels. The sizes of both these dimensions are typically at least on the order of thousands. While using SSE instructions is a very specific and less portable solution, the algorithms that we use are commonly executed on Intel cluster machines that provide these instructions, and recent CPU's all implement the SSE instruction set. To implement the SSE algorithm, we have used `gcc` intrinsics. These intrinsics are functions that map back to assembly instructions, but one does not need to think about registry allocations, etc., as that is still performed by the compiler. Moreover, unlike literal assembly code, the compiler is able to perform certain optimizations such as instruction pairing and loop unrolling.

Because the algorithm contains multiple branched statements, some care need to be taken when vectorizing the algorithm, as these need to be replaced by conditional moves. Another

point of care is the use of booleans that are implemented with 1 byte and the use of floats of 4 bytes. This requires to combine both SSE instructions and 'regular' instructions, but because the SSE instructions have dedicated registers, efficiently sharing data between these requires some work. Because the SSE instruction set contains instructions that perform 4 computations at once, the vectorized algorithm can process 4 sequences at once. However, because of the overhead created by avoiding branching, we see about a 2–3 times speed-up over the normal algorithm. Newer processors also provide the Advanced Vector Extensions (AVX) instruction set, which can simultaneously process 8 floating point computations. The algorithm can easily be extended to AVX instructions, thereby further increasing its speed. However, this extension is only available in recent processors, and because of scarce availability this is not yet used in our implementation.

Listing 2: Vectorized algorithm of a `SumThreshold` iteration

Require: χ is the average value threshold for sub-sequences of size M,
$\mathbf{x}(0 \dots N-1)$ are 4 input sequences of $i < N$ values stored in a vector,
$\mathbf{y}(0 \dots N-1)$ are 4 masks generated by the previous iteration, stored in a vector.
Ensure: $\mathbf{y}(0 \dots N-1)$ contain the output masks
$i \leftarrow (0)$, $\mathbf{z} \leftarrow (0,0,0,0)$, $\mathbf{count} \leftarrow (0,0,0,0)$
$\mathbf{t}(0 \dots N-1) \leftarrow \mathbf{y}(0 \dots N-1)$

{*calculate first window sum and count*}
while $i \neq M$ **do**
  {*add sample to the right*}
  $\mathbf{isnflagged} \leftarrow$ (NOT $\mathbf{y}(i)$) ? 0xFFFFFFFF : 0x0
  $\mathbf{z} \leftarrow \mathbf{z} + ((\mathbf{x}(i)$ & $\mathbf{isnflagged}) \mid (\mathbf{0.0}$ & $\neg\mathbf{isnflagged}))$
  $\mathbf{count} \leftarrow \mathbf{count} + (1$ & $\mathbf{isnflagged})$
  $i \leftarrow i + 1$
end while

{*slide window over the data*}
while $i \neq N$ **do**
  {*if threshold exceeded, set mask*}
  $\mathbf{exceedsThreshold} \leftarrow$
    $((\mathbf{z} > \chi \times \mathbf{count})$ OR $(\mathbf{z} < -\chi \times \mathbf{count}))$? 0xFFFFFFFF : 0x0
  if $\mathbf{exceedsThreshold} \neq (0,0,0,0)$ **then**
    $\mathbf{byteFlags} \leftarrow \mathbf{exceedsThreshold}$? set : unset
    for $s \in t(i - M \dots i - 1)$ **do**
      $\mathbf{t}(s) \leftarrow (\mathbf{t}(s) \mid \mathbf{byteFlags})$
    end for
  end if

  {*add sample to the right*}
  $\mathbf{isnflagged} \leftarrow$ (NOT $\mathbf{y}(i)$) ? 0xFFFFFFFF : 0x0
  $\mathbf{z} \leftarrow \mathbf{z} + ((\mathbf{x}(i)$ & $\mathbf{isnflagged}) \mid (\mathbf{0.0}$ & $\neg\mathbf{isnflagged}))$
  $\mathbf{count} \leftarrow \mathbf{count} + (1$ & $\mathbf{isnflagged})$

30: $\{$*subtract sample to the left*$\}$
$\mathbf{isnflagged} \leftarrow (\text{NOT } \mathbf{y}(i - M))$? 0xFFFFFFFF : 0x0
$\mathbf{z} \leftarrow \mathbf{z} - ((\mathbf{x}(i)\ \&\ \mathbf{isnflagged}) \mid (\mathbf{0.0}\ \&\ \neg\mathbf{isnflagged}))$
$\mathbf{count} \leftarrow \mathbf{count} - (1\ \&\ \mathbf{isnflagged})$
$i \leftarrow i + 1$
35: **end while**
$\mathbf{y}(0 \ldots N-1) \leftarrow \mathbf{t}(0 \ldots N-1)$

The vectorized version follows the presented scalar version of the algorithm and is given in Listing 2. In the algorithm, some operators are applied on vectors. These operands and their corresponding symbols are add (+); subtract (-); bitwise or (|); bitwise and (&); and conditional move (x $\leftarrow$ test ? a : b). When applied on vectors, these symbols denote the element-wise operations. The following SSE intrinsics are used to implement the vectorized algorithm:

_mm_set_ps : sets a vector to constant float values. Returns (a, b, c, d).

_mm_set_epi32 : similar as _mm_set_ps, but for constant integer values.

_mm_set1_ps : sets all values in a vector to one constant value. Returns (a, a, a, a).

_mm_load_ps : loads a vector from (non-constant) values in memory. Returns (a, b, c, d).

_mm_cmpeq_epi32 : element-wise compare of two integer vectors and return all bits set if equal, or all bits unset otherwise. Returns:
$(\mathbf{y} = \mathbf{z})$? 0xFFFFFFFF : 0x0.

_mm_cmpgt_ps and **_mm_cmplt_ps** : element-wise compare of float vectors similar to _mm_cmpeq_epi32, but for "greater than" and "less than" comparisons.

_mm_and_ps and **_mm_or_ps** : bitwise `and` and bitwise `or` between two vectors. Return $(\mathbf{x}\ \&\ \mathbf{y})$ and $(\mathbf{x} \mid \mathbf{y})$ respectively.

_mm_andnot_ps : bitwise and of a vector with the bitwise negation of another vector. Returns $(\mathbf{x}\ \&\ \neg\mathbf{y})$.

_mm_add_ps, **_mm_sub_ps** and **_mm_div_ps** : element-wise add, subtract and divide two float vectors.

_mm_cvtepi32_ps : convert integer vector to float vector.

_mm_movemask_ps : creates a 4 bit mask from the most significant bits of a float vector. This allows to store the result of a vector comparison into a single word, that can be used in regular (non-SSE) instructions.

From the vectorized algorithm, we can extract three essential sub-operations: adding a sample at the right of the window to the window (lines 7–9 and 25–27), subtracting a sample at the left from the window (lines 30–32) and testing the current window and setting the corresponding mask if necessary (lines 16–22). The other statements provide the loops and the initialization, and are trivial to implement.

In listing 3, a SSE algorithm is given in the C++ language, that adds a sample to the window. The subtraction can be implemented similarly.

Listing 3: Adding samples to windows with SSE instructions

Requires:
`rowFlagPtr`: a `const bool*` pointer to an array of flags, such that `rowFlagPtr`[x] with $0 \leq x < 4$ is the flag for window x.
`rowValPtr`: a `const float*` pointer to an array of samples, similar to `rowFlagPtr`.
`zero4i`: a `__m128i` vector containing zeros.
`ones4`: a `__m128i` vector containing ones.
`count4` : the number of samples in the windows (integer `__m128i` vector).
`sum4` : the sum of the unflagged samples in the windows (float `__m128` vector).

```
// Assign each integer in the vector to one bool in the mask
// Convert true to 0xFFFFFFFF and false to 0
__m128 conditionMask = _mm_castsi128_ps(
    _mm_cmpeq_epi32(_mm_set_epi32(rowFlagPtr[3], rowFlagPtr[2],
                                  rowFlagPtr[1], rowFlagPtr[0]),
                    zero4i));

// Conditionally increment counters
count4 = _mm_add_epi32(count4,
    _mm_and_si128(_mm_castps_si128(conditionMask), ones4));

// Add values with conditional move
__m128 m = _mm_and_ps(_mm_load_ps(rowValPtr), conditionMask);
sum4 = _mm_add_ps(sum4, _mm_or_ps(m,
    _mm_andnot_ps(conditionMask, zero4)));
```

The remaining algorithm to threshold the window and output the flags is given in Listing 4. This part interchanges between using the C++ boolean type of 1 byte and SSE masks, and therefore applies some tricks to convert between the two, as well as to "or" 4 booleans at a time.

As discussed, due to the `for` loop that sets the flags and that might be executed for each window position, the algorithm is optimized for a low number of positives. Nevertheless, the loop is fast, as it consists of one statement that is not a floating point operations. Therefore, even in cases where a lot of windows need to be flagged, the loop will not excessively slow down the algorithm.

When calculating the average, the `count4` variable is not tested for zero. Therefore, calculating the average might perform a division by zero. However, this can be ignored, since the outcome is not important if all samples are already flagged.

Listing 4: Thresholding the windows with SSE instructions

Requires:
Function `outputMask->RowFlagPtr(x, y)`: returns a `bool*` pointer to an array of flags, such that `RowFlagPtr`(x, y) with $0 \leq x < 4$ is the y-th output flag for window x, ordered in x direction.
`threshold4Pos`, `threshold4Neg`: positive and negative thresholds, χ and $-\chi$.
`count4` : the number of samples in the windows (integer `__m128i` vector).

`sum4` : the sum of the unflagged samples in the windows (float `__m128` vector). `M` : tested subsequence size, hence the number of samples in the window.

```
// if sum/count > threshold || sum/count < -threshold
__m128 avg4 = _mm_div_ps(sum4, _mm_cvtepi32_ps(count4));
const unsigned flagConditions =
  _mm_movemask_ps(_mm_cmpgt_ps(avg, threshold4Pos)) |
  _mm_movemask_ps(_mm_cmplt_ps(avg, threshold4Neg));

// The assumption is that most of the values are actually not
// thresholded. If this is the case, we circumvent the whole loop
// at the cost of one extra comparison:
if(flagConditions != 0)
{
  union
  {
    bool theChars[4];
    unsigned theInt;
  } outputValues = { {
    (flagConditions&1)!=0,
    (flagConditions&2)!=0,
    (flagConditions&4)!=0,
    (flagConditions&8)!=0 } };

  for(size_t i=0;i<M;++i)
  {
    unsigned *outputPtr = reinterpret_cast<unsigned*>(
      outputMask->RowFlagPtr(x, yTop + i));

    *outputPtr |= outputValues.theInt;
  }
}
```

Our implementation is written in C++ and uses a template variable for M. By doing so, the compiler creates a specialized version of the algorithm for each M value, and this allows the compiler to fully unwrap the two loops with M limits. The gcc compiler will only do this if the optimization parameter `-funwrap-loops` is specified on the command line. Specifying this compiler option will also partially unwrap the main loop, which speeds up the implementation considerably, because the compiler is now free to optimize between loop iterations and perform an optimization technique called instruction pairing.

Different machine architectures give different speed ups, but we generally see a factor 2–3 increase. Apart from the `-funwrap-loops` option, we also specify `-march=native` to enable the compiler to use any instruction set which the host computer provides. Memory that is used in the `SumThreshold` SSE implementation need to be aligned on 16 byte boundaries. On certain architectures, notably Apple machines, memory return from `malloc()` is already aligned correctly, but other architectures require the use of the `posix_memalign()` function instead.

A.3 Discussion & conclusions

We have shown how the feature detection accuracy of the `SumThreshold` method can be slightly decreased to create a fast implementation. Using exponentially increasing tested subsequence sizes decreases the time complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, and limiting the subsequence size decreases the time complexity to $\mathcal{O}(N)$. Our final implementation performs about 10 passes over the data — one for each tested subsequence size — and each pass requires a few floating point calculations per sample.

Using SSE instructions gave another factor of 2–3 increase in speed, without compromising the accuracy. However, such optimizations come at the expense of limiting its portability and increasing its implementation complexity. It is therefore only useful for the most time-critical parts of the software. Some algorithms might get an additional improvement from using GPUs. The down-side of GPUs is, that they require stricter memory access patterns, are not good at branched code and programming them is in my opinion somewhat more complex than using SSE instrinsics. GPUs that are efficient at scientific calculations, such as the NVIDIA Tesla machines, are also generally less available. In the LOFAR case, the central processing cluster does not provide them, thus they were not an option. Also astronomers that run the software at home probably have less benefit from GPU code. This in contrast to the SSE implementation, because in the last few months this implementation has processed all LOFAR recorded imaging observations on the LOFAR cluster, and has caused no issues. The SSE implementation is also shipped for some time in the latest AOFlagger package, and no problems have been reported from its users. Using AVX instructions is an attractive future improvement that might give another factor of 2 increase. However, this instruction set is very new and not yet generally available.

Bibliography

F. S. Acton. *Analysis of Straight-Line Data*. New York: Dover, 1966.

C. E. R. Alves, E. N. Càceres, and S. W. Song. A BSP/CGM algorithm for finding all maximal contiguous subsequences of a sequence of numbers. Technical report, Universidade de São Paulo, 2005.

K. Anderson, A. Alexov, L. Baehren, J.-M. Griessmeier, M. Wise, and A. Renting. LOFAR and HDF5: Toward[s] a new radio data standard. In *Proc. of ISKAF2010*. Astron, PoS, June 2010.

R. Athreya. A new approach to mitigation of radio frequency interference in interferometric data. *AJ*, 696:885–890, Apr. 2009.

W. A. Baan, P. A. Fridman, and R. P. Millenaar. Radio frequency interference mitigation at the Westerbork Synthesis Radio Telescope: Algorithms, test observations, and system implementation. *AJ*, 128:933–949, Aug. 2004.

W. A. Baan, P. A. Fridman, S. Roy, and R. Millenaar. The RFI Mitigation System at WSRT. In *Proc. of RFI2010*. Astron, PoS, Mar. 2010.

C. Barnbaum and R. F. Bradley. A new approach to interference excision in radio astronomy: Real-time adaptive cancellation. *AJ*, 115:2598–2614, Nov. 1998.

M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes - Theory and Application*. Prentice-Hall, Inc., Apr. 1993.

J. Bentley. Programming pearls: algorithm design techniques. *Commun. ACM*, 27:865–873, September 1984. ISSN 0001-0782.

M. Bentum, A.-J. Boonstra, R. Millenaar, and A. Gunst. Implementation of LOFAR RFI mitigation strategy. In *URSI General Assembly 2008*, Chicago, August 2008. URSI.

M. J. Bentum, A. J. Boonstra, and R. P. Millenaar. Assessment of RFI measurements for LOFAR. *Proc. of RFI2010*, Mar. 2010.

G. Bernardi, A. G. de Bruyn, M. A. Brentjens, B. Ciardi, G. Harker, V. Jelić, L. V. E. Koopmans, P. Labropoulos, A. R. Offringa, V. N. Pandey, J. Schaye, R. M. Thomas, S. Yatawatta, and S. Zaroubi. Foregrounds for observations of the cosmological 21 cm line: I. First Westerbork measurements of Galactic emission at 150 MHz in a low latitude field. *A&A*, 500:965–979, Mar. 2009.

G. Bernardi, A. G. de Bruyn, M. A. Brentjens, B. Ciardi, G. Harker, V. Jelić, L. V. E. Koopmans, P. Labropoulos, A. R. Offringa, V. N. Pandey, J. Schaye, R. M. Thomas, S. Yatawatta, and S. Zaroubi. Foregrounds for observations of the cosmological 21 cm line: II. Westerbork observations of the fields around 3C196 and the North Celestial Pole. *A&A*, 522, 2010.

N. D. R. Bhat, J. M. Cordes, S. Chatterjee, and T. J. W. Lazio. RFI identification and mitigation using simultaneous dual station observations. *Radio Science*, 40, June 2005.

A. J. Boonstra. *Radio frequency interference mitigation in radio astronomy*. PhD thesis, June 2005.

A. J. Boonstra, S. J. Wijnholds, S. van der Tol, and B. Jeffs. Calibration, sensitivity and RFI mitigation requirements for LOFAR. *Proc. IEEE Int. Conf. on Acoustics, Speech & Signal Processing*, 5:869–872, Mar. 2005.

J. D. Bowman and A. E. E. Rogers. A lower limit of dz $>$ 0.06 for the duration of the reionization epoch. *Nature*, 468:796–798, Dec. 2010.

J. D. Bregman. Concept design for a low-frequency array. volume 4015, pages 19–32. SPIE, 2000.

A. H. Bridle and F. R. Schwab. Bandwidth and time-average smearing. In Taylor, G. B. and Carilli, C. L. and Perley, R. A., editor, *Synthesis Imaging in Radio Astronomy II*, volume 180, page 371. Astron. Soc. Pac., 1999. ISBN 1-58381-005-6.

F. Briggs, J. F. Bell, and M. J. Kesteven. Removing radio interference from contaminated astronomical spectra using an independent reference signal and closure relations. *AJ*, 120:3351–3361, Dec. 2000.

C. J. Chandler and R. A. Perley. *The Expanded Very Large Array observational status summary*. NRAO, Mar. 2010.

E. Chapman, F. B. Abdalla, G. Harker, V. Jelić, P. Labropoulos, S. Zaroubi, M. A. Brentjens, A. G. de Bruyn, and L. V. E. Koopmans. Foreground removal using FastICA: A showcase of LOFAR-EoR. 2012. submitted to MNRAS.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

A. G. de Bruyn and G. Bernardi. The first deep WSRT 150 MHz full polarization observations. In D. J. Saikia, D. A. Green, Y. Gupta, and T. Venturi, editors, *ASP Conf. Ser.*, volume 407, page 3, Pune, India, sep 2009. Astron. Soc. Pac.

A. G. de Bruyn, M. A. Brentjens, L. V. E. Koopmans, S. Zaroubi, P. Labropoulos, and S. B. Yatawatta. Detecting the EoR with LOFAR: Steps along the road. pages 1–4, aug 2011. ISBN 978-1-4244-5117-3.

R. Deriche. Recursively implementing the Gaussian and its derivatives. In *Proc. 2nd International Conference on Image Processing*, pages 263–267, 1992.

R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson. Cosmic Black-Body Radiation. *ApJ*, 142:414–419, July 1965.

S. W. Ellingson and G. A. Hampson. A subspace-tracking approach to interference nulling for phased array-based radio telescopes. *IEEE Trans. on Antennas & Propagation*, 50:25–30, Jan. 2002.

H. I. Ewen and E. M. Purcell. Observation of a line in the galactic radio spectrum: Radiation from galactic hydrogen at 1,420 mc./sec. *Nature*, 168:356, Sept. 1951.

J. R. Fisher. Techniques for coping with radio frequency interference. In R. J. Cohen and W. T. Sullivan, editors, *Preserving the Astronomical Sky, proc. of IAU Symposium 196*, page 279, Vienna, Austria., July 2001.

L. Flöer, B. Winkel, and J. Kerp. RFI mitigation for the Effelsberg Bonn HI Survey (EBHIS). In *Proc. of RFI2010*, Mar. 2010.

P. A. Fridman. Estimates of variance in radio-astronomy observations. *AJ*, 35:1810–1824, May 2008.

P. A. Fridman and W. A. Baan. RFI mitigation methods in radio astronomy. *A&A*, 378:327–344, Oct. 2001.

P. Friedman. A change point detection method for elimination of industrial interference in radio astronomy receivers. *Proc. 8th IEEE Signal Processing Workshop on Statistical Signal & Array Processing*, pages 264–266, June 1996.

D. E. Gary, Z. Liu, and G. M. Nita. Hardware implementation of an SK spectrometer. In *Proc. of RFI2010*, Mar. 2010.

R. N. Ghose. *Interference Mitigation: Theory and Application*. New York: IEEE Press, 1996.

J. Goutsias and H. J. A. M. Heijmans. Fundamenta morphologicae mathematicae. *Fundam. Inf.*, 41: 1–31, January 2000. ISSN 0169-2968.

D. Halen. Recursive Gaussian filters, 2006. CWP Report.

J. P. Hamaker, J. D. Bregman, and R. J. Sault. Understanding radio polarimetry. I. Mathematical foundations. *A&AS*, 117:137–147, May 1996.

G. Harker, S. Zaroubi, G. Bernardi, M. A. Brentjens, A. G. de Bruyn, B. Ciardi, V. Jelić, L. V. E. Koopmans, P. Labropoulos, G. Mellema, A. Offringa, V. N. Pandey, A. H. Pawlik, J. Schaye, R. M. Thomas, and S. Yatawatta. Power spectrum extraction for redshifted 21-cm Epoch of Reionization experiments: the LOFAR case. *MNRAS*, 405 (4):2492–2504, 2010.

G. J. A. Harker, S. Zaroubi, G. Bernardi, M. A. Brentjens, A. G. De Bruyn, B. Ciardi, V. Jelić, L. V. E. Koopmans, P. Labropoulos, G. Mellema, A. R. Offringa, V. N. Pandey, J. Schaye, R. M. Thomas, and S. Yatawatta. Non-parametric foreground subtraction for 21-cm epoch of reionization experiments. *MNRAS*, 397(2):1138–1152, Aug. 2009a.

G. J. A. Harker, S. Zaroubi, R. M. Thomas, V. Jelić, P. Labropoulos, G. Mellema, I. T. Iliev, G. Bernardi, M. A. Brentjens, A. G. De Bruyn, B. Ciardi, L. V. E. Koopmans, V. N. Pandey, A. H. Pawlik, J. Schaye, and S. Yatawatta. Detection and extraction of signals from the epoch of reionization using higher-order one-point statistics. *MNRAS*, 393(4):1449–1458, Mar. 2009b.

F. J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51 – 83, jan. 1978. ISSN 0018-9219.

M. Hata. Empirical formula for propagation loss in land mobile radio services. *"IEEE Trans. on Vehicular Technology"*, VT-29, Aug. 1980.

G. Heald et al. Progress with the LOFAR imaging pipeline. June 2010.

A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, and R. A. Collins. Observation of a rapidly pulsating radio source. *Nature*, 217:709–713, Feb. 1968.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3: 1163–1174, 1975.

H. T. Intema, S. van der Tol, W. D. Cotton, A. S. Cohen, I. M. van Bemmel, and H. J. A. Röttgering. Ionospheric calibration of low frequency radio interferometric observations using the peeling scheme. I. Method description and first results. *A&A*, 501:1185–1205, July 2009.

D. C. Jacobs, J. E. Aguirre, A. R. Parsons, J. C. Pober, R. F. Bradley, C. Carilli, N. E. Gugliucci, J. R. Manley, C. van der Merwe, D. F. Moore, and C. Parashare. New 145 MHz source measurements by PAPER in the southern sky. *The Astrophysical Journal Letters*, 734(2):L34, 2011.

K. G. Jansky. Electrical disturbances apparently of extraterrestrial origin. *Proc. IRE*, 1933.

B. D. Jeffs, S. van der Tol, and A.-J. van der Veen. Direction dependent self calibration of large distributed sensor arrays. volume 4, page IV1069, May 2006.

V. Jelić. *Cosmological 21cm experiments : searching for a needle in a haystack*. PhD thesis, May 2010.

V. Jelić, S. Zaroubi, P. Labropoulos, R. M. Thomas, G. Bernardi, M. A. Brentjens, A. G. de Bruyn, B. Ciardi, G. Harker, L. V. E. Koopmans, V. N. Pandey, J. Schaye, and S. Yatawatta. Foreground simulations for the LOFAR-epoch of reionization experiment. *MNRAS*, 389:1319–1335, Sept. 2008.

S. Karling and V. Brendel. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257:39–49, July 1992.

S. Kazemi, S. Yatawatta, S. Zaroubi, P. Labropoulos, A. G. de Bruyn, L. V. E. Koopmans, and J. Noordam. Radio interferometric calibration using the SAGE algorithm. *MNRAS*, 414(2):1656–1666, 2011. ISSN 1365-2966.

J. Kocz, F. H. Briggs, and J. Reynolds. Radio frequency interference removal through the application of spatial filtering techniques on the Parkes multibeam receiver. *AJ*, 140(6):2086, 2010.

J. Kocz, M. Bailes, D. Barnes, S. Burke-Spolaor, and L. Levin. Enhanced pulsar and single pulse detection via automated radio frequency interference detection in multipixel feeds. *Monthly Notices of the Royal Astronomical Society*, 420(1):271–278, 2012. ISSN 1365-2966.

L. Kogan and F. Owen. RFI Mitigation in AIPS. The New Task UVRFI. In *Proc. of RFI2010*. Astron, PoS, Mar. 2010.

P. Labropoulos. *The LOFAR Epoch of Reionization Data Model: Simulations, Calibration, Inversion*. PhD thesis, Sept. 2010.

W. M. Lane, A. S. Cohen, N. E. Kassim, T. J. W. Lazio, R. A. Perley, W. D. Cotton, and E. W. Greisen. Postcorrelation radio frequency interference excision at low frequencies. *RS*, 40, 2005.

J. J. Lemmon. Wideband model of man-made HF noise and interference. *Radio Science*, 32:525–539, Mar. 1997.

A. Leshem and A.-J. van der Veen. Introduction to interference mitigation techniques in radio astronomy. In A. B. Smolders and M. P. Haarlem, editors, *Proc. of Perspectives on Radio Astronomy: Technologies for Large Antenna Arrays*, page 201. ASTRON, 2000a.

A. Leshem and A.-J. van der Veen. Radio astronomical imaging in the presence of strong radio interference. *IEEE Trans. on Information Theory*, pages 1730–1747, 2000b.

A. Leshem, A.-J. van der Veen, and A.-J. Boonstra. Multichannel interference mitigation techniques in radio astronomy. *ApJS*, 131:355–373, Nov. 2000.

J. C. Mather, E. S. Cheng, R. E. Eplee, Jr., et al. A preliminary measurement of the cosmic microwave background spectrum by the Cosmic Background Explorer (COBE) satellite. *ApJ*, 354: L37–L40, May 1990.

L. L. McCready, J. L. Pawsey, and R. Payne-Scott. Solar radiation at radio frequencies and its relation to sunspots. In *Proc. Roy. Soc. Lond. A 12*, volume 190, pages 357–375, Aug. 1947.

M. McQuinn, O. Zahn, M. Zaldarriaga, L. Hernquist, and S. R. Furlanetto. Cosmological parameter estimation using 21 cm radiation from the epoch of reionization. *ApJ*, 653(2):815, 2006.

E. Middelberg. Automated editing of radio interferometer data with `pieflag`. *Publications of the Astronomical Society of Australia*, 23:64–68, 2006.

D. A. Mitchell and J. G. Robertson. Reference antenna techniques for canceling RFI due to moving sources. *Radio Science*, 40, 2005.

D. A. Mitchell, J. G. Robertson, and R. J. Sault. Alternative adaptive filter structures for improved radio frequency interference cancellation in radio astronomy. *AJ*, 130:2424–2433, Nov. 2005.

N. Niamsuwan, J. T. Johnson, and S. W. Ellingson. Examination of a simple pulse-blanking technique for radio frequency interference mitigation. *Radio Science*, 40, June 2005.

J. Noordam. *The Newstar cookbook, Internal NFRA report*. 1994.

J. E. Noordam. LOFAR calibration challenges. *Proc. of SPIE*, 5489(1):817–825, 2004. ISSN 0277786X.

A. R. Offringa. Proposal for adding statistics subtables to a measurement set. Technical report, University of Groningen, Kapteyn Astronomical Institute, Dec. 2011.

A. R. Offringa and A. G. de Bruyn. Interference detection results with LOFAR. In *General Assembly and Scientific Symposium, 2011 XXXth URSI*, pages 1–4, aug 2011.

A. R. Offringa, A. G. de Bruyn, M. Biehl, S. Zaroubi, G. Bernardi, and V. N. Pandey. Post-correlation radio frequency interference classification methods. *MNRAS*, 405:155–167, June 2010a.

A. R. Offringa, A. G. de Bruyn, M. Biehl, and S. Zaroubi. A LOFAR RFI detection pipeline and its first results. In *Proc. of RFI2010*. Astron, PoS, Mar. 2010b.

A. R. Offringa, A. G. de Bruyn, and S. Zaroubi. Post-correlation filtering techniques for off-axis source and RFI removal. *MNRAS*, 422:563–580, May 2012a.

A. R. Offringa, J. J. van de Gronde, and J. B. T. M. Roerdink. A morphological algorithm for improved radio-frequency interference detection. *A&A*, 539, Mar. 2012b.

Y. Okumura et al. Field strength and its variability in VHF and UHF land-mobile. *Radio Service Rev. Elec. Comm. Lab.*, pages 825–873, 1968.

S. M. Ord, D. A. Mitchell, R. B. Wayth, L. J. Greenhill, G. Bernardi, S. Gleadow, R. G. Edgar, M. A. Clark, G. Allen, W. Arcus, L. Benkevitch, J. D. Bowman, F. H. Briggs, J. D. Bunton, S. Burns, R. J. Cappallo, W. A. Coles, B. E. Corey, L. deSouza, S. S. Doeleman, M. Derome, A. Deshpande, D. Emrich, R. Goeke, M. R. Gopalakrishna, D. Herne, J. N. Hewitt, P. A. Kamini, D. L. Kaplan, J. C. Kasper, B. B. Kincaid, J. Kocz,

E. Kowald, E. Kratzenberg, D. Kumar, C. J. Lonsdale, M. J. Lynch, S. R. McWhirter, S. Madhavi, M. Matejek, M. F. Morales, E. Morgan, D. Oberoi, J. Pathikulangara, T. Prabu, A. E. E. Rogers, A. Roshi, J. E. Salah, A. Schinkel, N. U. Shankar, K. S. Srivani, J. Stevens, S. J. Tingay, A. Vaccarella, M. Waterson, R. L. Webster, A. R. Whitney, A. Williams, and C. Williams. Interferometric imaging with the 32 element Murchison Wide-Field Array. *Pub. of the Astr. Soc. of the Pac.*, 122(897):1353–1366, 2010.

G. Paciga, T.-C. Chang, Y. Gupta, R. Nityanada, J. Odegova, U.-L. Pen, J. B. Peterson, J. Roy, and K. Sigurdson. The GMRT Epoch of Reionization experiment: a new upper limit on the neutral hydrogen power spectrum at z approx 8.6. *Monthly Notices of the Royal Astronomical Society*, 413(2): 1174–1183, May 2011.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, June 1954.

V. Pankonin and R. M. Price. Radio Astronomy and Spectrum Management: The Impact of WARC-79. *IEEE Trans. on Electromagnetic Compatibility*, EMC-23:308–317, Aug. 1981. ISSN 0018-9375.

A. R. Parsons and D. C. Backer. Calibration of Low-Frequency, Wide-Field Radio Interferometers Using Delay/Delay-Rate Filtering. *AJ*, 138:219–226, July 2009.

U.-L. Pen, T.-C. Chang, C. M. Hirata, J. B. Peterson, J. Roy, Y. Gupta, J. Odegova, and K. Sigurdson. The GMRT EoR Experiment: Limits on Polarized Sky Brightness at 150 MHz. *MNRAS*, 399:181–194, Oct. 2009.

R. R. Pizzo, editor. *The LOFAR Imaging Cookbook: Manual data reduction with the imaging pipeline*. 2012.

A. J. Poulsen, B. D. Jeffs, K. F. Warnick, and J. R. Fisher. Programmable real-time cancellation of GLONASS interference with the Green Bank Telescope. *The Astronomical Journal*, 130 (6):2916, 2005.

J. Raza, A.-J. Boonstra, and A. van der Veen. Spatial filtering of RF interference in radio astronomy. *Signal Processing Letters, IEEE*, 9(2):64 –67, feb 2002. ISSN 1070-9908.

J. W. Romein. Bandpass correction in LOFAR. Technical report, ASTRON, Aug. 2008.

J. W. Romein, J. D. Mol, R. V. van Nieuwpoort, and P. C. Broekema. Processing LOFAR telescope data in real time on a Blue Gene/P supercomputer. In *URSI General Assembly, 2011*, pages 1–4. IEEE, Aug. 2011. ISBN 978-1-4244-5117-3.

M. G. Santos, A. Cooray, and L. Knox. Multifrequency analysis of 21 centimeter fluctuations from the era of reionization. *ApJ*, 625(2):575, 2005.

A. P. Schoenmakers, A. G. de Bruyn, H. J. A. Röttgering, H. Van Der Laan, and C. R. Kaiser. Radio galaxies with a 'double-double morphology' — I. Analysis of the radio properties and evidence for interrupted activity in active galactic nuclei. *MNRAS*, 315(2):371–380, 2000. ISSN 1365-2966.

O. M. Smirnov. Revisiting the radio interferometer measurement equation. *A&A*, 527:A106, Mar. 2011.

B. Smolders and G. Hampson. Deterministic RF nulling in phased arrays for the next generation of radio telescopes. *IEEE Antennas & Propagation magazine*, 44:13–22, Aug. 2002.

P. Soille. On morphological operators based on rank filters. *Pattern Recognition*, 35(2):527 – 535, 2002. ISSN 0031-3203.

P. Soille and H. Talbot. Directional morphological filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1313–1329, 2001. ISSN 0162-8828.

B. W. Stappers, J. W. T. Hessels, A. Alexov, et al. Observing pulsars and fast transients with LOFAR. *A&A*, 530:A80, June 2011.

R. M. Thomas, S. Zaroubi, B. Ciardi, A. H. Pawlik, P. Labropoulos, V. Jelić, G. Bernardi, M. A. Brentjens, A. G. de Bruyn, G. J. A. Harker, L. V. E. Koopmans, G. Mellema, V. N. Pandey, J. Schaye, and S. Yatawatta. Fast large-scale reionization simulations. *MNRAS*, 393:32–48, Feb. 2009.

A. R. Thompson. The response of a radio-astronomy synthesis array to interfering signals. *IEEE Trans. on Antennae & Propagation*, 30:450–456, 1982.

A. R. Thompson, T. E. Gergely, and P. A. Vanden Bout. Interference and Radioastronomy. *Physics today*, 44:41–49, Nov. 1991.

H. van der Marel, E. E. M. Woestenburg, and A. G. de Bruyn. Low frequency receivers for the WSRT. A window of opportunity. In *Proc. of URSI-GA 2005*, New Delhi, India, Oct. 2005.

L. van Vliet, I. Young, and P. Verbeek. Recursive Gaussian derivative filters. In *14th Int. Conf. on Pattern Recognition*, Brisbane, Qld., Australia, Aug. 1998.

S. von Hoerner. Radio source counts and cosmology. *ApJ*, 186:741–766, 1973.

R. Weber, C. Faye, F. Biraud, and J. Dansou. Spectral detector for interference time blanking. *A&AS*, 126:161–167, 1997.

B. Widrow and P. N. Stearns. *Adaptive Signal Processing*. Prentice Hall, Mar. 1985.

B. Winkel, J. Kerp, and S. Stanko. RFI detection by automated feature extraction and statistical analysis. *AN*, 88:789–801, 2006.

B. Winkel, J. Kerp, and P. Kalberla. Data reduction strategy of the Effelsberg-Bonn HI Survey (EBHIS). *Proc. Panoramic Radio Astronomy: Wide-field 1-2 GHz research on galaxy evolution, PoS(PRA2009)043*, June 2009.

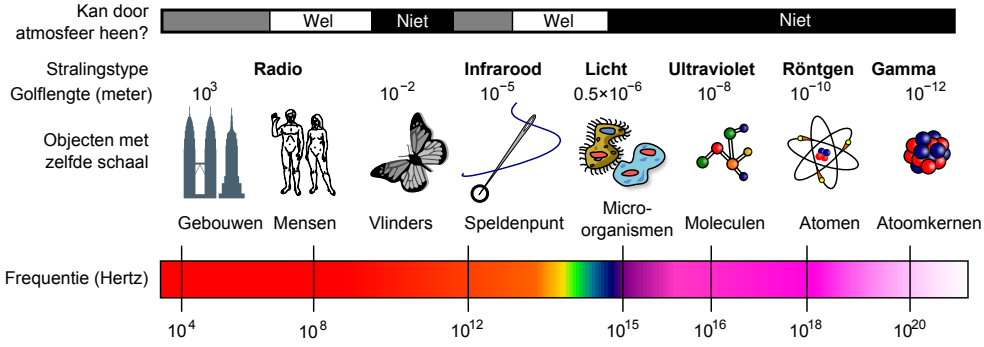
S. Yatawatta, S. Zaroubi, G. de Bruyn, L. Koopmans, and J. Noordam. Radio Interferometric Calibration Using The SAGE Algorithm. In *Proc. of 13th Dig. Sig. Proc. Workshop and 5th IEEE Sig. Proc. Education Workshop*, pages 150–155, 2009.

Index

Nederlandse samenvatting

RADIOASTRONOMIE IS een uitdagende tak van de sterrenkunde, waarin dankzij voortdurende technologische ontwikkelingen, vele verrassende en mysterieuze details van ons Universum zijn gevonden. Ondanks dat sterrenkunde één van de oudste wetenschappen in de geschiedenis is, werd pas in het jaar 1931 door Karl Jansky ontdekt dat de hemel niet alleen zichtbaar licht afgeeft. Deze invloedrijke ontdekking leidde tot nieuwe fantastische ontdekkingen, zoals in 1965 de ontdekking van radiostraling die is ontstaan vlak na de Oerknal, genaamd 'kosmologische achtergrondstraling', en de ontdekking van roterende neutronensterren ('pulsars') in 1967. Deze objecten gedragen zich als vuurtorens in de hemel.

Het verschil tussen radio- en optische sterrenkunde is, dat de hemel op een andere frequentie wordt bekeken. Zichtbaar licht bestaat uit elektromagnetische straling met een golflengte van ongeveer 400–800 nanometer, terwijl radiostraling een golflengte van enkele millimeters tot vele kilometers heeft. Een overzicht van het elektromagnetisch spectrum is gegeven in Fig. 1. Door ook op radiogolflengtes waar te nemen, kunnen we nieuwe fenomenen aanschouwen en ons Universum verder doorgronden. Echter, het radiospectrum wordt ook gebruikt door menselijke apparatuur, zoals satellieten en tv- en radiozenders. Dit zorgt voor storing ('interferentie') tijdens het waarnemen en deze signalen moet daarom gescheiden worden van het hemelsignaal. Dit proefschrift behandelt methodes voor het onschadelijk maken van deze interferentie, die ook wel *radio-frequency interference* (RFI) genoemd wordt.



Figuur 1: *Een overzicht van het elektromagnetische spectrum. De bovenste balk geeft aan of straling met de corresponderende golflengte de atmosfeer van onze aarde kan penetreren. (Bron: aangepaste afbeelding van Wikipedia en NASA.)*

Radioastronomie

Moderne radiosterrenwachten, zoals de sterrenwacht in Westerbork, kunnen de hemel met een enorme gevoeligheid en resolutie in kaart brengen. Radiotelescopen bestaan vaak uit één of meerdere grote schotelantennes. De gevoeligheid waarmee een telescoop kan observeren, is voornamelijk afhankelijk van de grootte van de schotel: hoe groter, hoe gevoeliger. De resolutie van een sterrenwacht kan worden verhoogd door het combineren van meerdere telescopen. Dit gebeurt door middel van een techniek die interferometrie wordt genoemd. Wanneer de waarneemresolutie door middel van het samenvoegen van meerdere telescopen wordt verhoogd, noemt men dit radiosynthese. Om steeds diepere details van ons Universum te bekijken, zijn er door de jaren heen grote sterrenwachten gebouwd. Naast de synthesetelescoop te Westerbork zijn een aantal bekende sterrenwachten die radiosynthese toepassen de Very Large Array (VLA) te Socorro (Nieuw-Mexico), de Australia Telescope Compact Array (ATCA) bij Narrabri (Australië), de Giant Metrewave Radio Telescope (GMRT) te Pune (India) en de Atacama Large Millimeter/sub-millimeter Array (ALMA) in de Atacamawoestijn (Chili).

Het gebruik van grote schotels is erg efficiënt voor golflengtes van centimeterschaal of kleiner. Voor golflengtes van een meter en groter heeft een schotel anderzijds relatief weinig toegevoegde waarde ten opzichte van een simpele antenne. Dankzij diverse technologische ontwikkelingen is het nu mogelijk om een groot aantal simpele en goedkope antennes te combineren tot een grote telescoop. Zulke antennes zijn dan esthetisch misschien niet zo aantrekkelijk als de traditionele schoteltelescoop, voor lange golflengtes zijn ze zeer goedkoop en efficiënt en bieden vele nieuwe mogelijkheden. De Low-Frequency Array, ofwel laagfrequente telescoop, is een nieuwe gedeeltelijk-Nederlandse radiosterrenwacht die als een van de eerste gebruik maakt van dit principe. Deze telescoop staat centraal in dit proefschrift, dus ik zal beginnen met een korte beschrijving van dit instrument.

De Low-Frequency Array

LOFAR, de Low-Frequency Array bestaat uit velden van antennes die elektromagnetische straling in het bereik 10–90 en 110–240 MHz kunnen ontvangen. De antennesignalen van een veld worden digitaal gecombineerd, en dit geheel wordt een station genoemd. Op het moment bestaat LOFAR uit 41 stations. Nog 7 stations worden gebouwd en mogelijk volgen er nog meer. Van de 41 voltooide stations staan er 33 in Nederland en 5 in Duitsland. Ook in Zweden, het Verenigd Koninkrijk en Frankrijk staat een LOFAR-station. Stations bevatten twee soorten antennes: de lage-bandantennes (low-band antennae, LBA) die 10–90 MHz kunnen ontvangen; en de hogebandantennes (high-band antennae, HBA) die gevoelig zijn voor 110–240 MHz. De HBA's worden samengevoegd in 'tegels' die 4x4 antennes bevatten. Foto's van de twee antennesoorten zijn te zien in Fig. 2. De Nederlandse stations bevatten 96 LBA's en één of twee velden van in totaal 48 HBA tegels.

Het centrum van LOFAR, waar de meeste stations zich bevinden, is gevestigd nabij Exloo. De hoogste dichtheid van stations bevindt zich op een kunstmatig schiereiland, waar zes stations zijn geplaatst. Dit eiland wordt de "Superterp" genoemd, en is opgehoogd land omgeven door water. Exloo bevindt zich in een landelijk gebied en heeft ten opzichte van de rest van Nederland een lage populatiedichtheid. Desalniettemin bevinden de stations zich, ten opzichte van andere sterrenwachten, erg dicht bij bewoond gebied, waardoor RFI van menselijke apparatuur grote

Figuur 2: *De twee soorten LOFAR-antennes. Afbeelding links: een lage-bandantenne met een kabinet in de achtergrond. Afbeelding rechts: deel van een station met hoge-bandantennes. Deze stations bestaan uit 24 tegels van 4 × 4 antennes.*

problemen kan geven. Deze RFI kan afkomstig zijn van opzettelijk transmissies, zoals bij het gebruik van walkietalkies en digitale video- en audiotransmissies, maar ook kunnen bijvoorbeeld auto's, schrikdraad, hoogspanningslijnen en windmolens onbedoeld radiostraling uitzenden.

In dit proefschrift wordt de schade geanalyseerd die RFI kan veroorzaken in LOFAR-observaties. Nieuwe technieken voor het meten en verwijderen van interferentie uit de data worden geïntroduceerd met als doel de schade te minimaliseren. Deze technieken kunnen vervolgens voor zowel LOFAR als voor andere radiotelescopen gebruikt worden. Omdat LOFAR een nieuw soort telescoop is, moeten deze technieken aan strenge eisen voldoen. Zo heeft LOFAR momenteel het grootste aantal stations van alle radiosynthese-telescopen, en omdat LOFAR met zeer hoge tijd- en frequentieresolutie zal waarnemen, zijn de gegenereerde datastromen van enorme omvang. De algoritmes die gebruikt worden tegen RFI moeten derhalve zeer snel zijn. In telescopen met slechts een klein aantal antennes of schotels kan een astronoom makkelijk één voor één de data van de verschillende antennes langs lopen. Echter, omdat het volume van de data nu zo veel groter is, zullen algoritmes volledig automatisch en robuust moeten zijn. Verder moet de hoeveelheid RFI die in de data achterblijft tot een minimum beperkt blijven. Dit houdt in dat de methodes zeer accuraat moeten zijn.

Detectiemethodes

Hoofdstuk 2 van dit proefschrift behandelt detectiemethodes. Bij de detectie van interferentie wordt gezocht naar monsters ('samples') in tijd-frequentieruimte die zijn beïnvloed door RFI. Wanneer dit slechts een klein gedeelte van de data betreft, kan deze data genegeerd worden en

kan de rest van de data gebruikt worden voor de verdere astronomische analyse. Deze techniek wordt aangeduid als het 'vlaggen' (*flagging*) van de data. In de literatuur was relatief weinig informatie te vinden over RFI-detectiemethodes: de meeste artikelen richtten zich op het herstellen van besmette data. Desalniettemin is detectie het belangrijkste wapen tegen RFI, aangezien voor iedere radiosterrenwacht bijna altijd een vorm van detectie wordt gebruikt.

Dit gat in de ontwikkeling van detectiemethodes wordt met dit proefschrift gevuld. In hoofdstuk 2 introduceren we diverse nieuwe detectietechnieken, die een significant hogere accuraatheid en snelheid hebben dan de methodes die voorheen gebruikt werden. Het probleem wordt opgesplitst in drie stappen:

I. Het signaal van de hemel wordt geschat en tijdelijk afgetrokken van de data. Dit wordt gedaan met een standaard Gaussisch filter. Het residu bevat voornamelijk RFI en normale ruis.

II. In de tijd-frequentieruimte worden samples gezocht die, alleen of gezamenlijk met andere samples, een buitensporige sterkte hebben. Hiervoor wordt het nieuw-ontworpen `Sum-Threshold` algoritme gebruikt.

III. Om ook samples te vinden die geen buitensporige sterkte hebben maar tóch beïnvloed zijn door RFI, wordt een morfologische operator toegepast. Deze operator, genaamd de schaalonafhankelijke rankoperator (*scale-invariant rank (SIR) operator*), kijkt naar de vorm van de gevlagde samples in tijd-frequentieruimte om zo te zoeken naar nieuwe RFI samples.

Deze drie stappen worden in hoofdstuk 3 gecombineerd tot een iteratieve *pipeline*. Met behulp van Westerbork- en LOFAR-data worden de parameters geoptimaliseerd. De resulterende pipeline vindt alle zichtbare interferentie in de data. Deze pipeline is vervolgens sterk geoptimaliseerd voor snelheid, onder andere door het gebruik van speciale SSE processorinstructies en het ontwikkelen van een SIR-operator-algoritme met lineaire tijdscomplexiteit. De uiteindelijke pipeline werkt sneller dan dat de harde schijven de data kunnen aanleveren. Dit betekent dat het LOFAR-rekencluster met ongeveer 100 computers een observatie (iets) sneller dan realtime kan vlaggen. De detectiemethodes komen weer terug in verdere hoofdstukken, waar ze verder worden getest en worden gebruikt voor het analyseren van de LOFAR-radio-omgeving.

Filtertechnieken

Hoofdstuk 4 behandelt het herstellen van de data wanneer langdurige en breedbandige RFI een observatie onbruikbaar heeft gemaakt. Ook behandelen we het filteren van hemelbronnen buiten het kaartcentrum, dat een gerelateerd probleem blijkt te zijn. Breedbandige RFI ontstaat meestal onbedoeld, bijvoorbeeld door schrikdraad dat regelmatig vonkjes genereert. Dit leidt er toe dat een observatie niet correct gekalibreerd kan worden of dat de RFI de hemelbronnen overschaduwt. Wanneer zulke RFI continu optreedt heeft het weinig zin om de samples die beïnvloed zijn te detecteren en te vlaggen, aangezien er dan geen samples overblijven.

Doordat een RFI-zender die vast staat op de aarde altijd dezelfde oriëntatie heeft ten opzichte van de antennes, gedraagt de zender zich als een bron aan de Noordelijke hemelpool. Dat is namelijk de enige plek aan de hemel waarvoor dit ook geldt. Doordat een interferometer bemonstert in het Fourierdomein, zien we een stationaire bron terug in de ruwe data als een bron met een specifieke golfbeweging (*fringe pattern*). Aangezien we de plek weten van de bron in het

reële domein — namelijk de pool — kunnen we ook de golfbeweging voorspellen. Deze kennis is gebruikt in een recent-ontwikkelde methode voor de GMRT radiosterrenwacht in India, waar erg veel RFI van deze soort zich voordoet. De methode schat de sterkte en fase van de bron en trekt vervolgens de golfbeweging van de data af. Dit proces wordt "fringe fitting" genoemd.

LOFAR ziet (gelukkig) geen sterke breedbandige bronnen. Zwakke bronnen zouden wel in de data aanwezig kunnen zijn en de kwaliteit van de hemelkaart kunnen verlagen. Na het analyseren van de fringe-fittingtechniek met simulaties, concluderen we dat eventuele zwakke bronnen niet correct kunnen worden verwijderd. We introduceren daarom diverse nieuwe technieken, waaronder een specifieke toepassing van een low-pass filter. Dit filter heeft in tegenstelling tot fringe fitting niet als doel om een bron op een specifieke locatie te filteren, maar filtert alle flux die afkomstig is van bronnen buiten het centrum van de observatie (*off-axis* bronnen).

De techniek is niet alleen een mogelijkheid voor het verwijderen van zwakke RFI, maar kan ook gebruikt worden voor het aftrekken van off-axis bronnen. Bij telescopen met een groot gezichtsveld — zoals LOFAR — zijn off-axis bronnen een groot probleem: ze maken het kalibratieproces zeer ingewikkeld en traag, en kunnen de kwaliteit van de uiteindelijke hemelkaart verlagen. We passen de techniek toe op een reële 150-MHz Westerborkobservatie met sterke off-axis bronnen en constateren dat de hemelkaart na toepassing van de techniek significant verbeterd is. De sterkte van het filter is afhankelijk van de hoeveelheid beschikbare aaneengesloten bandbreedte. In onze observatie is slechts 2,5 MHz bandbreedte gebruikt. Een volgende stap is het toepassen van de techniek op LOFAR-observaties, waar 48 MHz aaneengesloten bandbreedte beschikbaar is. De techniek is zeer snel, wat in de LOFAR-situatie ook een belangrijke eigenschap is.

De radio-omgeving van LOFAR

Door middel van de diverse besproken methodes, wordt in hoofdstuk 5 de radio-omgeving van LOFAR geanalyseerd. Dit gebeurt door analyse van een HBA- en LBA-LOFAR-observatie van beide 24 uur. We constateren dat, na het toepassing van de besproken automatische methodes, de observaties schoon zijn en dat direct verder gegaan kan worden met de kalibratie- en karteerprocedures.

De gevonden spectrale RFI bezetting in de 115–163-MHz HBA-meting is 3,2%, terwijl de 30–78 MHz LBA meting met 1,8% bezetting nog net iets minder storing bevat. Het verlies tengevolge van RFI is dus nagenoeg verwaarloosbaar in LOFAR's huidige situatie. Dit is een verrassende conclusie, aangezien LOFAR zich in bewoonde gebieden bevindt, in tegenstelling tot de meeste telescopen waar doorgaans meer RFI wordt gemeten. De redenen hiervoor zijn: (I) LOFAR heeft een zeer hoge tijd- en frequentieresolutie, waardoor korte of kleinbandige RFI slechts een miniem verlies van data veroorzaakt; (II) de AOFlagger detectiepipeline heeft een ongeëvenaarde accuraatheid; (III) LOFAR's hardware is ontworpen om met interferentie om te gaan; (IV) LOFAR-antennes bevinden zich laag bij de grond, waardoor zenders op een kleine afstand al geblokkeerd worden door de Aarde en haar bebouwing en begroeiing; en (V) er is geen noemenswaardige zelfbesmetting, dat wil zeggen, besmetting van RFI die gegenereerd wordt door LOFAR's eigen hardware.

Dit is een zeer belangrijke conclusie die laat zien dat radio-astronomie mogelijk is in een bewoonde omgeving. Niet alleen heeft LOFAR hiermee een hemels vooruitzicht, ook is dit belangrijk voor toekomstige radiosterrenwachten. Zo wordt momenteel de locatie bepaald van de

Square Kilometre Array (SKA). Dit is een toekomstige radiotelescoop met een nog grotere omvang dan LOFAR. RFI is bij deze keuze een belangrijk argument, en de overwogen opties zijn daarom zéér afgelegen plaatsen in Australië en Zuid-Afrika.

Eén belangrijke kanttekening is hierbij wel nodig. LOFAR kan op dit moment niet observeren op frequenties die bezet zijn door sterke, breedbandige zenders. Dit betekent dat het FM-zenderbereik tussen 87–108 MHz en het DAB (digital audio broadcasting) zenderbereik tussen 216–230 MHz niet gebruikt kunnen worden voor radioastronomie. Dit is een relatief klein bereik en met de verdere analyse van filtertechnieken zouden ook op deze frequenties nog observaties mogelijk kunnen zijn. Echter, het is belangrijk dat de frequentietoewijzing niet sterk verandert. Nieuwe, sterke en breedbandige zenders waar LOFAR niet op voorbereid is, kunnen zeer schadelijk zijn.

De spatiële- en helderheids-distributie van RFI bronnen

Als laatste onderwerp van dit proefschrift analyseren we diverse eigenschappen van RFI bronnen. In hoofdstuk 6 kijken we specifiek naar de invloed die gelekte RFI zou kunnen hebben op de meest gevoelige LOFAR-experimenten. Hoewel de AOFlagger zeer accuraat is, zou het kunnen zijn dat coherente patronen gevormd worden door minuscule gelekte RFI wanneer er geïntegreerd wordt over een zeer lange observatietijd. Of en in welke mate RFI inderdaad coherent is, is nog onduidelijk.

Een lange integratietijd is bijvoorbeeld nodig voor de detectie van signalen van de herionisatieperiode (*Epoch of Reionisation* of *EoR*). Hiervoor moet een enorme gevoeligheid van slechts tientallen μJy bereikt worden. In dit cruciale era in het ontstaan van ons Universum zijn de eerste objecten gevormd, zoals sterren, sterrenstelsels en quasars. Verscheidene experimenten zijn onderweg om roodverschoven signalen van neutrale waterstof uit deze periode te detecteren, en het LOFAR-EoR project is er een van. Om de benodigde gevoeligheid te bereiken, zal voor dit project over 100 nachten aan LOFAR-observaties worden geïntegreerd.

We analyseren de helderheidsdistributie van RFI op een manier die vergelijkbaar is met de $\log N - \log S$ analyses die gebruikt worden in de kosmologie. We constateren een aantal opvallende en verrassende kenmerken, die we kunnen verklaren met een uniforme spatiële distributie van RFI bronnen, waarbij de elektromagnetische propagatie van het signaal vrij exact beschreven wordt door bestaande propagatiemodellen. Door extrapolatie van de distributie en met een aantal aannames berekenen we de invloed die resterende RFI zou kunnen veroorzaken in zeer gevoelige experimenten. Waar de systeemruis op 1 kHz en 1 s niveau 770 Jy in de LBA en 77 Jy in de HBA is, wordt de maximale schade van gelekte RFI geschat op 490 mJy voor de LBA and 170 mJy voor de HBA. Dit laat wederom zien dat de AOFlagger zeer accuraat is, aangezien er bronnen worden gevonden die zich onder de ruis bevinden.

Desondanks zou dit RFI-niveau negatieve implicaties kunnen hebben. Wanneer namelijk RFI-bronnen met dit niveau stationair en coherent zouden zijn, zou al binnen één observatienacht duidelijke effecten te zien moeten zijn aan de hemelpool. Omdat de hemelpool één van de hemelvelden is waar het LOFAR-EoR zich op richt, kunnen we deze mogelijkheid analyseren. Door de eerste EoR observaties weten we nu dat zulke effecten nog niet zichtbaar zijn. We kunnen dus concluderen dat, als er inderdaad RFI resteert, deze momenteel niet coherent genoeg is om schade te veroorzaken.

Mochten er bij langere integraties tóch RFI effecten verschijnen, dan kan de AOFlagger op

de geïntegreerde data worden toegepast. Dit zou de gevoeligheid van de flagger aanzienlijk verhogen. Alles lijkt er dus op dat gelekte RFI zeer waarschijnlijk geen problemen gaat geven bij het detecteren van de herionisatieperiode!

Nu LOFAR zelfs ondanks RFI haar werk kan doen, kan zij beginnen aan haar fantastische wetenschappelijke doelen. Het staat in de sterren geschreven dat ons een opwindende tijd wacht!

Acknowledgements

During the four years that it took to write this thesis, many, many, MANY persons have helped me to finish it. Thanks to all of them, this journey has been a lot of fun too! I could not possibly thank everyone that contributed in some way, but I will try. Let me start by thanking the three kings that have been crucial for reaching this point...

Ger, I had the huge honour and sincere pleasure of having you as my supervisor. Your sharp and limitless knowledge about everything keeps amazing me, and you are definitely the king of radio observations! And math! In our countless enthusiastic discussions you always challenged me to find more answers, explore more things and, in the end, you have brought me here at the end of this thesis. Often I was sure that I had considered all possibilities, but you made a habit of coming with yet another useful inspiring idea or argument. Also, you do this in an ever friendly and constructive way. I greatly appreciate your unbiased scientific view. I always see you care passoniately about the truth, no matter from which direction it comes. Also, I really like the freedom that you give me in doing my science. You allowed me to walk my own path through this project, though with a constantly accelerating pace because of your help.

Saleem, I very much enjoyed working with you. You were a great supervisor. Many times you hopped by my office, and showed interest in my latest discoveries. Before you would leave the office, you always managed to change me into a more optimistic — if not blissful — state. As the king of theory & cosmology, you are very good in connecting the theory with the issues I tumbled upon, which were mostly technical. I also think you have a sixth sense for interpreting formulae. Finally, I must also mention you can be very funny!

Michael, I got to know you as the most inspiring teacher I knew, back in 2006 when you were teaching the course of "Modelling and Simulation". Although the Ph.D. project quickly diverged from your expertise, it was very nice to have you, the king of machine learning, as my supervisor. During this project, I learned that you are a great cook, and the mud walking was a lot of fun!

Ger, Saleem and Michael, thank you for supervising me!

Next, I would like to thank my reading committee, prof. dr. Frank Briggs, prof. dr. Wim Brouw and prof. dr. Thijs van der Hulst. I would also like to thank the LKBF for financial support.

I was very lucky to be a member of the *LOFAR epoch of reionisation group*. We have shared a lot of lively meetings, conferences and discussions, and I really enjoyed hanging out with our group of great, incredibly clever yet slightly crazy people. One of the first and very memorable meetings that I attended with the EoR group, was the "E-LOFAR" workshop in Hamburg. Around this time, I got to know *Vibor*, *Panos*, *Sarod* and *Rajat*, which was great. I continued to work together with Sarod, Panos and Vibor, and I realize now that four years is not enough to get used to these crazy guys (I can't help noticing that Panos and Vamsi show similarities with Howard Wolowitz and Rajesh Koothrappali from the tv series "the Big Bang theory"). By the way, I really am Dutch (stop calling me a German!). I had fun! Sarod, Panos and Vibor, thank you for all the help you have offered me during this project. Also, thank you for driving me to Astron many times!

Later, the EoR group acquisitted *Vamsi*, *Marty*(/Oscar), and *Sanaz*, which made the group even more interesting. I got to know you very well on the Foregrounds conference in Croatia, with a very nice combination of science and sea, sun and drinks. Thanks to your nice company (and Vibor's great organisation skills) this was one of the greatest conferences ever! Marty, thank you very much for helping with my many data tasks. I would also like to thank all the other LOFAR-EoR members for a nice collaboration from which I learned a lot, especially *Gianni*, *Nicoletta*, *Pandey*, *Geraint*, *Leon*, *Parisa*, *Michiel* and *Stefan*.

Ger (van Diepen), a special thanks for the numerous times you have helped me is in order! Although we saw each other only a few times at Astron, we had a lot of e-mail communication, and this was very helpful. I was very glad to be acquinted with the expert on Casacore, C++ programming and much more. *Roberto* and *Antonis*, you were my primary contacts to the "LOFAR observatory". Thank you for your help and feedback on the software. Roberto, thank you also for maintaining the LOFAR cookbook (you did a great job!) and for driving me to Astron quite a few times. *Arno* and *Marcel*, thank you for your help getting and keeping the AOFlagger in the LOFAR repository (sorry for the few times I broke the daily build).

Jasper and *Jos*, we had a brief but very productive collaboration. This resulted in a nice paper and a vastly optimized algorithm! Thank you for this invaluable contribution. It was very nice to work together with two persons whom I knew from well before my Ph.D.; Jasper from my study and from dancing, and Jos initially as professor of visualization courses. During my master, you were the supervisor of my thesis, which was the project that moved me towards doing research! Also on the side of Computer Science, I would like to thank my former officemates from "my other office", *Petra* (you are a great dancer!), *Kerstin* and *Aree*. Thanks also go to *Michael* (Wilkinson, also a great dancer :)) and other participants of the Intelligent Systems group.

The *Kapteyn Astronomical Institute* is a wonderful place to work! Besides hosting a lot of smart people, the atmosphere is great! An important contribution for this has been made by my officemates, *Yang-Shyang*, *Boris*, *Carlos* and *Harish*. I also very much appreciate the aid from other people at the Kapteyn, in particular *Eite* for performing all the hacks and tweaks to solve all my terrible (sorry!) software requests, and for keeping

the EoR cluster running (despite us). Of course, a lot of thanks to *Wim* for helping me with hardware troubles and I would like to thank *Hennie*, *Jacky*, *Lucia*, *Martin* and *Gineke* for help with various things. Furthermore, it appeared this mysterious institute is packed with friendly people, and I would like to thank all of them for their nice company during coffee, lunch and other inspiring events. This includes *Omar*, *Leon*, *Patrick*, *Marlies*, *Eva*, *Hans*, *Maarten*, *Stephan*, *Johan*, *Hugo*, *Koshy*, *Job*, *Tjitske*, *Katinka*, *Alicia*, and all the others!

Everyone that sent me feedback on the AOFlagger and other tools, although you kept me quite busy (there were so many of you!) thank you for all the questions, bug reports and other types of feedback.

During these four years, I also had (sort of) a life outside work. In my personal life there are many people very dear to me, and I hope I will continue to see you all once I moved to Australia!

Lieve papa & mama, dank jullie wel voor jullie onvoorwaardelijke steun en interesse gedurende de afgelopen ~~4~~ 30 jaar! Het is altijd fijn om jullie te zien of te spreken en samen dingen te doen. Vanaf mijn vijfde vertelde ik jullie dat ik 'uitvinder' wilde worden en het is heel fijn om jullie bij me te hebben terwijl ik dat doel nastreef... De hele wereld om me heen is veranderd, maar bij jullie staat mijn wereld altijd even stil. Jullie zijn de liefste ouders die er zijn! *Lieve broers Johan* en *Lars*, en natuurlijk ook *Maaike* en *Shanthi*, bedankt voor jullie altijd-warme interesse en "broederschap" over de jaren, inclusief de leuke vakantie samen. Hopelijk komen jullie een keer langs in Australië!

Beste Vincent & Eva, jullie betekenen heel veel voor mij en ik ben dan ook erg blij dat Vincent mijn paranimf is! Altijd zijn jullie er voor mij geweest en wanneer ik jullie zie, voelt het onmiddelijk vertrouwd... en leuk! Bedankt voor jullie steun. Ik hoop jullie te zien in Australië! Dat geldt natuurlijk ook voor *Susan* & *Anne Jan* en *Clement* & *Romy*.

Beste Jarno & Marije, mijn dank aan Jarno voor het zijn van mijn tweede paranimf! Niet alleen kende ik je als mede-geek bij The Blue Toes, maar later werd ik ook je collega, waar ik erg blij mee was. Dank jullie wel voor alle keren dat we samen hebben gegeten, gediscussiëerd over wetenschappelijke zaken (inclusief hulp met mijn artikelen), gedanst, film gekeken en alle andere dingen.

Beste Carolien, ik wil je bedanken voor je liefdevolle interesse tijdens de eerste helft van mijn PhD. Ik hoop dat je heel gelukkig wordt! Lieve mensen van *Studentenstijldansvereniging The Blue Toes*, jullie vulden mijn "tweede leven" naast mijn PhD en maakte het tot iets geweldigs! Sommige van jullie kende ik al ver vóór mijn PhD en het is super dat ik nog steeds de eer heb om met jullie les te geven of te dansen. Lieve *Joke*, *Mawije & Thomas*, *Christa*, *Sandra*, *Nikita & Liekele*, *Marten & Helena*, *Reeuwerd*, *Kelly*, *Maarten*, *Quirijn*, *Keimpe* en de rest van *Gevorderd II 2012*, mijn eigen allerbeste docent *Jan Willem Jansen*, en alle andere Blauwe Tenen, bedankt!

Liefste meisje van het Universum, liefste, dierbaarste, zorgzaamste, charmantste, schattigste *Ingeborg*, het leven is zoveel leuker met jou naast me! Ik hou van je.

Colofon

Algorithms for Radio Interference Detection and Removal
by Anne René (André) Offringa
First printed in May 2012

Printed by: Off Page, Amsterdam, The Netherlands

Print count: 250

Cover: Illustration of LOFAR low-band antennae. Photographed sky background from Wikipedia (license: CC). The stars in the sky on the back are from a LOFAR image of the 3C196 field that was made by Panos Labropoulos. The band of RFI is a recording of the FM band around 100 MHz, visualized with the 'rfigui' tool.

ISBN:
978-90-367-5545-0
978-90-367-5544-3 (electronic version)

Total pages: 212 (excluding cover)