# Robust Statistics Applied to Radio Astronomy: Radio Frequency Interference Mitigation and Automated Spectral Line Detection for Broadband Surveys

Christophe Belleval

PSL UNIVERSITÉ - OBSERVATOIRE DE PARIS

PhD in Astronomy and Astrophysics - Doctorat en Astronomie et Astrophysique

---

# Robust Statistics Applied to Radio Astronomy: Radio Frequency Interference Mitigation and Automated Spectral Line Detection for Broadband Surveys

---

*Author: Christophe BELLEVAL*

*Jury*

| | |
|---|---|
| *President:* | *Françoise COMBES* |
| *Supervisor:* | *Wim VAN DRIEL* |
| *Co-supervisor:* | *Jean-Michel MARTIN* |
| *Referee:* | *Albert-Jan BOONSTRA* |
| *Referee:* | *Lister STAVELEY-SMITH* |
| *Member:* | *Hélène COURTOIS* |
| *Invited:* | *Pierre COLOM* |

September 9, 2019

# *Abstract*

Cette thèse présente les algorithmes développés dans le but d'identifier et d'éliminer les interférences électromagnétiques (RFI) présentes dans les études à large bande HI effectuées avec des radiotélescopes à antenne unique dans la bande L. Ces procédés utilisent des matrices 2D (en colonne: les canaux fréquentiels; en ligne: les spectres moyennés de puissance) qui sont alimentées par des séries temporelles de valeurs spectrales moyennées de puissance accumulées durant une observation, celles-ci suivant une loi du $\chi 2$. En réglant le spectromètre afin que leur degré de liberté soit au moins égal à 50, de telles séries de valeurs spectrales moyennées de puissance sont distribuées suivant une loi normale, à la condition que le récepteur n'enregistre que des sources cosmiques et du bruit blanc gaussien. L'hypothèse qui structure ce travail est qu'un échantillon de telles valeurs qui est pollué par des RFI ne suit pas une distribution gaussienne. Pour identifier puis éliminer ces RFI avec une technique d'élimination des extrêmes (dite de "clipping"), on a étudié et sélectionné des estimateurs robustes (insensibles à la présence de valeurs extrêmes, dites "outliers") de position telle que la médiane et des estimateurs d'échelle (Median Absolute Deviation $MAD$, $Sn$ et $Qn$). De manière surprenante, $MAD$ est utilisé dans plusieurs paquets logiciels, alors que cet estimateur est biaisé pour des échantillons non symétriques. Pour cette raison, nous lui préférons $Sn$ qui est un estimateur non centré. Une fois que la matrice 2D a été réduite à un spectre 1D constitué des estimateurs de position et d'échelle des séries temporelles de valeurs de puissance spectrales moyennes de chaque canal fréquentiel, la détection automatique de raies spectrales est effectuée avec la régression robuste non linaire Least Trimmed Squares (LTS). L'ensemble de ces algorithmes est à présent mis en oeuvre avec le logiciel RObust Elusive Line detection (ROBEL). Les observations effectuées avec le radiotélescope décimétrique de Nançay et enregistrées avec le spectromètre large bande WIBAR sont présentées. Premièrement le quasar B0738+313 dont la ligne de visée a révélé deux absorbants Damped Lyman Alpha (DLA), l'un à $z{\sim}0.0912$ (nommé B1), l'autre à $z{\sim}0.2212$ (nommé B2) ce dernier étant occulté par un RFI. Ces derniers ont été éliminés par ROBEL, et par ailleurs, la précision des mesures a gagné un facteur 8 par rapport aux techniques usuelles d'évaluation effectuées sur le seul spectre 1D. Deuxièmement, ROBEL a éliminé les RFI puissants issus de la constellation de satellites Iridium qui occultent totalement les raies d'émission OH1665/1667 MHz du mégamaser III Zw 35. Trosièmement, les procédures d'élimination des RFI ont été testées sur les RFI causés par des radars ainsi que par des constellations Global Navigation Satellite System (GNSS) avec des résultats prometteurs: le *rms* obtenu à partir de ces observations courtes est à peu près deux fois supérieur à ce qu'il serait sans RFI, mais on n'a pas encore étudié sa diminution en fonction du temps d'observation: cette question motive une prochaine série extensive de tests complémentaires dans la bande $1160 - 1400$MHz. Finalement, les futurs développements sont discutés, en particulier l'adaptation des techniques spectrales exposées dans cette thèse à des interféromètres.

This Thesis presents algorithms which were developed with the aim of mitigating Radio Frequency Interference (RFI) and fostering automated spectral line detection in HI broadband surveys conducted with single-dish radio telescopes in the frequency band 1 - 2 GHz. These algorithms use 2D matrices (frequency channels vs. averaged power spectra) containing averaged time-line series of power spectral values accumulated during an observation, which in each frequency channel follows a $\chi 2$ distribution law. With enough degrees of freedom defined by the spectrometer setup (at least 50), such power spectral values samples are normally distributed if the receiver captures only cosmic source signals and white noise. To detect and remove RFI, the main assumption of this work is that when contaminated by artificial signals such as RFI the distribution of averaged power spectral values is no longer Gaussian. Robust estimators (i.e., those immune to sample outliers) of location (median) and scale (Median Absolute Deviation $MAD$, $Sn$ and $Qn$) are studied to assess these sample distributions and to excise RFI, the latter with a clipping technique. Though $MAD$ is used for most RFI mitigation applications in previously implemented post-processing software, it is biased in the presence of asymmetrical distributions, and for this reason $Sn$ is a better choice. Once the 2D matrix has been collapsed to a 1D spectrum containing estimators of location and scale of averaged power spectral values, a robust Least Trimmed Squares (LTS) non-linear regression is used for automated baseline fitting and spectral line detection. These algorithms were implemented in the RObust Elusive Line detection (ROBEL) software package which was used to post-process observations made with the 100m-class single-dish Nançay Radio Telescope (NRT) and captured by the WIBAR broadband spectrometer. Results for three case studies are presented: (1) QSO B0738+313 with two previously detected intergalactic Damped Lyman Alpha (DLA) HI absorption lines, with the one at $z{\sim}0.09$ (named B1) free of RFI and the other at $z{\sim}0.22$ (B2) swamped by RFI. RFI excision was successful, and measurement precision were notably improved with ROBEL algorithms even in the absence of RFI: error margins in measured line parameters were reduced by an 8 factor compared to usual calculations on 1D spectra. Moreover both B1 and B2 were automatically detected by ROBEL. (2) OH megamaser III Zw 35 whose 1665/1667 MHz emission lines, redshifted to $\sim$1621 MHz, have been hidden since two decades by strong RFI from the Iridium satellite constellation. The RFI was successfully removed and the lines made detectable again. (3) Mitigation of RFI from ground-based Radio Navigation (RN) radars and Global Navigation Satellite System (GNSS) constellations in the frequency range $1160 - 1382$ MHz, to prepare for reopening the band for astronomical observations. The first test results were promising: the *rms* of these short observations is about twice what it would be without RFI, but the reduction of the *rms* level as function of integration time has not yet been studied. Thus, extensive testing will be continued. Finally, foreseen developments are discussed, in particular the adaptation of the spectral techniques described in this Thesis to interferometers.

# *Acknowledgements*

My gratitude goes to people who supported me during the six years I took to complete my Thesis, in good and tough times, and believed in me and my work in spite of my unusual profile. We spent long hours arguing, brainstorming, sharing the best of ourselves, sometimes fighting to preserve scarce resources, or just simply to exist. Above all, we made a great team and we had fun. I admire the enthusiasm of these engineers and scientists who were able to create a great spectrometer with such a small budget and their astute job. It is the kind of magical science I have always dreamed of: nurtured with creativity and imagination by clever, demanding and benevolent gentlemen. Thus, my achievement is also theirs, and nothing would have possible without Jean-Michel Martin, Pierre Colom, Jean Borsenberger, and of course Wim van Driel who was a very committed and professional supervisor. The long discussions I had with Eric Gérard, the wise and enthusiast patriarch, also helped me to improve my thinking. When Wim and I discussed the composition of the jury, I asked for the best specialists in the field, even if I knew they would not make my life easy. And then again, I want to express my recognition for all their questions, remarks and contributions which allowed me to raise the quality of my Thesis.

Science can become an exclusive mistress. Nevertheless, my loving wife Marielle was an unwavering supporter, especially during tough times. Her sweet presence alongside helped me stand hardships and countless sleepless nights. I am also proud of my parents who not only passed on their values, enthusiasm, and optimism, but also supported me with love, patience, and endless discussions late at night.

My friends, colleagues, and beloved family, rest assured: this is only the beginning!

# Contents

# List of Figures

# List of Tables

*This Thesis is dedicated to my wife Marielle, my parents Geneviève and Pierre, as well as to my daughter Diane and my son Arnaud. May this adventure be inspirational to the following generation.*

# Chapter 1

# Introduction

## 1.1 Context

The advent of a new generation of radio telescopes coupled with digital processing hardware have provided tremendous new opportunities for extensive studies of the Galactic and extra-galactic environment. In parallel, the nowadays huge amounts of data produced have generated specific challenges. Among others: to secure high-speed data capture, to temporarily store huge quantities of raw data, to upgrade processing to this new data profile, and to provide researchers with practical tools of analysis.

In this Thesis I have focused on two issues. Firstly the mitigation of radio frequency interference (RFI)[1], and secondly practical methods for automated blind detection of spectral lines. The algorithms I present have been developed using observational data from the 100 m-class decimetric, single-dish Nançay Radio Telescope (NRT) located in the center of France (Monnier Ragaigne et al., 2003).

## 1.2 RFI mitigation: an overview

It is well recognized that RFI can cause serious problems for radio astronomical observations, with ever increasing pressure on use of the radio spectrum by actively emitting services and more and more sensitive radio telescopes. This is why regulatory bodies such as the International Telecommunication Union (ITU) issued Recommendations for the protection of radio astronomical observations, such as ITU-R RA.769[2] on "Protection criteria used for radio astronomical measurements" which specifies that RFI is detrimental if its level exceeds 10% of the background noise level (*rms*) (van Driel, 2011). Fig. 1.1 illustrates how extensive the RFI problem has become to radio astronomers outside of those frequency bands in which their observations are protected through regulatory measures.

On the regulatory protection of radio astronomical observations, it should be noted that (1) the provisions of Recommendation ITU-R RA.769 apply only to frequency bands in which the radio astronomy service (RAS) has a primary allocation according to the ITU-R Radio Regulations (RR). Within the spectral range of ∼1000-3400 MHz covered by the Nançay Radio Telescope, where the observations described in this Thesis was made, these are the bands 1400-1427, 1610.6-1613.8, 1660-1670 and

[1] In this document I consider RFI mitigation and excision as equivalent terms.
[2] https://www.itu.int/rec/R-REC-RA.769/en

2690-2700 MHz. (2) Footnote 5.340 of the RR states that all emissions are prohibited in certain bands, including 1400-1427 MHz. (3) Footnote 5.149 of the RR urges administrations to take all practicable steps to protect the RAS from harmful interference in making allocations to stations of other services to which certain bands are allocated, including 1330-1400, 1610.6-1613.8, 1660-1670, 1718.8-1722.2, 2655-2690, 3260-3267, and 3325-3339 MHz[3].

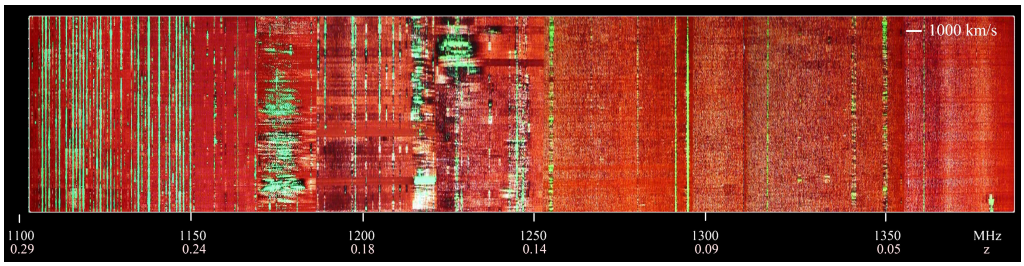


FIGURE 1.1: Broadband spectrum observed at the Nançay decimetric radio telescope (NRT) with the standard autocorrelator in 2011 (van Driel & Lehnert, private communication). Shown in this so-called waterfall diagram are one-hour long time series of $ON - OFF$ power spectra. Indicated along the horizontal axis are frequencies (from 1100 to 1380 MHz) and the corresponding redshifts $z$ of the 21 cm HI line (from 0.03 to 0.29); the horizontal white bar indicates a radial velocity range of 1000 km s$^{-1}$. RFI flagged by the standard NRT data reduction package (see Monnier Ragaigne et al. (2003)) is shown in green whereas the sky background is shown in red.

Unwanted emissions in radio telescope data are essentially generated by manmade sources of RFI, both ground-based or from satellites, as well as natural electric phenomena such as lightning and electronic noise (electromagnetic compatibility, or EMC) (Taylor et al., 2019). Moreover, single dish radio telescopes are especially vulnerable to RFI, as compared to interferometers (Finger et al., 2018).

The broad diversity of radio interferences implies the necessity of a complex and multi-layer strategy (Baan, 2019). Several strategies have been developed for data processing and RFI mitigation. Baan et al. (2019) and Kesteven (2010) have provided the following typology:

1. pro-active and pre-detection at station level;
2. mitigation at the system level aimed at dealing with strong RFI and protection against instrument saturation;
3. mitigation before correlation and processing which includes baseband as well as adaptive filtering;
4. mitigation during correlation or at the post-processing level.

Firstly, all possible proactive measures have to be taken to prevent any unwanted emissions from disrupting detectors. This is essentially done on one hand with the help of hardware shielding (e.g., Faraday cages) as well as isolation of any electronic equipment. On the other hand, the definition of a Radio Quiet Zone defined by

[3]Radio Regulations 2015, International Telecommunications Union (https://www.itu.int/pub/R-REG-RR/en)

regulatory authorities is of utmost importance (see e.g., the ITU-R Handbook on Radio Astronomy [4]).

Secondly, at the system level there should be filtering to avoid saturation by RFI, while maintaining gain level. However, filtering comes at a gain cost and is not easy to manipulate.

Thirdly, RFI may be excised by software at the baseband before correlation and processing. There are essentially four techniques for this:

1. signal blanking synchronized with regular interference such as radar pulses;
2. analysis of the signal kurtosis on a time-line basis (see, e.g. Gary et al. (2010));
3. the threshold technique notably used at the Westerbork Synthesis Radio Telescope (WSRT) (Baan et al., 2010) or the Giant Metrewave Radio Telescope (GMRT) (Buch et al., 2016);
4. the use of one or more reference antennas in order to substract RFI from the signal captured by the main telescope; this is called adaptive filtering (Kesteven, 2010) and targets the waveform. One of the first attempts was documented by Barnbaum et al. (1998). More recently Finger et al. (2018) have proposed a digital adaptive filter using a field programmable gate array (FPGA) to cancel out interference signals;
5. spatial filtering which was implemented at the LOFAR phased array test station (Boonstra et al., 2005) and coupled with adaptive filtering techniques.

Although working on the waveform theoretically opens the way to very interesting processes (see e.g., Hellbourg et al. (2012) and Weber et al. (1997)), it is limited in practice by the huge amount of data collected which constrains both the accessible bandwidth and the duration of acquisition.

Enter the fourth and last stage: the excision of the remaining RFI after the first three stages[5]. This is done when observational time-series of averaged power spectra in each frequency channel are reduced to their statistical estimators of location (such as the sample mean) and scale (such as the standard deviation) parameters (see section 2.1 on p.11).

RFI is discarded according to statistical criteria applied on time-series of averaged power spectra (any value exceeding a threshold given in terms of the sample standard deviation is discarded) according to a well-established method documented in the seminal paper of Peirce (1852).

There are two main schools in RFI mitigation software development: one that favors real-time mitigation and another proposing more sophisticated off-line algorithms.

The first strategy favors real-time data-processing in order to minimize data storage. Such an approach implies severe compromises to the complexity of the RFI mitigation algorithms, due to restrictions in both processing time and available memory.

[4] www.itu-ilibrary.org/science-and-technology/handbook-on-radio-astronomy_pub/809847c8-en

[5] Though in practice for certain radio telescope configurations the RFI is essentially mitigated at this post-processing level only; such is the case for the NRT.

A typical example of this family of software is AOFlagger which has been implemented on the LOFAR pipeline, though it does not in itself excise RFI. van Nieuwpoort (2016) developed a real-time RFI mitigation package called LOF (LOFAR Online Flagger) which mainly uses the SumThreshold algorithm: this performs thresholding "with an exponentially increasing window size, and an increasingly sharper threshold", allowing the detection with various time scales. It is applicable to voltages as well power spectral values, and has a linear computational complexity rise. Another example is the RFI Mitigation System at the WSRT interferometer (Baan et al., 2010). Winkel et al. (2007) has documented a real-time Digital Fast-Fourier-Transform (DFFT) based on the FPGA technology used at the Effelsberg 100m single-dish radio telescope. Buch et al. (2019) and Buch et al. (2016) describe a real-time FPGA excision system using the Median Absolute Deviation (MAD)[6].

The other strategy consists of post-processing data after acquisition – Fridman (2009) even argues that sophisticated algorithms can only be implemented off-line.

At single-dish telescopes, the data of the Parkes Galactic All-Sky Survey (GASS) has been processed using median filtering techniques (Kalberla, 2010), and the Effelsberg-Bonn HI Survey (EBHIS) also uses such sigma-clipping techniques (Flöer et al., 2010). In interferometry, sigma-clipping is used for Epoch of Reionization (EoR) HI line studies at the 21CM Array (21CMA) Huang et al. (2016). Constant RFI in interferometry is mitigated at the GMRT using fringe-stop patterns (Athreya, 2009), where it has mostly been tested off-line though it could also be implemented on-line with enough CPU resources.

Note that some other techniques have also been tested, including one applied to interferometers and taking advantage of cyclostationary RFI properties (Hellbourg et al., 2012), and deep convolutional neural networks, i.e., deep learning (Akeret et al., 2017).

## 1.3 The need for data quality assessment

A spectrometer converts digitized real signal voltages to complex spectral data by Discrete Fourier Transform (DFT). These complex values are then correlated, i.e., multiplied by their conjugates, giving power spectra. Averaged power spectra result from averaging $N$ of these power spectra. If the captured signal voltages are randomly distributed, the individual power spectra are $\chi^2$ distributed with 2 degrees of freedom. Average power spectra are also $\chi^2$ distributed but with $2N$ degrees of freedom. For $2N \geq 50$, the $\chi^2$ law tends to the normal law (by virtue of the Central Limit Theorem), which means that with sufficient degrees of freedom, averaged power spectra from a cosmic source mixed with blank Gaussian noise can be considered normally distributed.

With data quality assessment, I mean the evaluation of the deviations from a normal distribution of time-series of averaged power spectra for each frequency channel in a spectrum (see 3.2.3 on p.36). In the presence of RFI, the need to evaluate data quality is of utmost importance.

The main assumption of this work is the following: any artificial signal carries information which by definition is not randomly distributed. This means only natural

[6] See 2.4.2 on p. 20 for a discussion on this estimator of scale.

sources and sky background may exhibit a normal distribution of averaged power spectra as captured by a radio spectrometer. While this paradigm has already been expressed (Huang et al., 2016), I want to add the following to it:

Because the captured signal passes through a series of instruments, each with its own particular properties of non-neutrality and transparency (such as specular return, imperfect filtering, instrumental and even digital noise at the software level), *no signal* (even those of cosmic origin) recorded by an instrument *is ever* exactly randomly distributed on a time-line basis, even when observing a source such as a non-polarized HI spectral line.

This is the main reason why part of this project has been dedicated to defining a set of data quality estimators.

Data quality assessment must be applied before any kind of clipping on averaged power spectral time-series. For this work I have chosen a series of data quality estimators, some being classical and others based on robust estimators of scale.

In the first category, I have selected the comparison between the theoretical noise of the instrument and the actual *rms* of times-series of averaged power spectra of each frequency channel (see 3.2.3.1 on p.36). This indicates how far the observational data lie from the best possible dataset the instrument can produce.

The classical data quality estimators I have chosen are skewness (see 2.2.8.1 on p.16) and kurtosis (see 2.2.8.2 on p.17) of time-series. An analysis of both these moments give a good insight of some characteristics of the time-line series of averaged power spectra: skewness evaluates the symmetry of the dataset around its median whereas kurtosis measures the dataset spread. Strong RFI are typically identified with high absolute values of skewness and positive kurtosis, whereas a strictly normal distribution returns exactly 0 bor both.

Since classical statistical scale parameters such as the standard deviation are not robust to outliers (see 1.6 on p.7), I have also applied robust estimators of scale (see section 1.7 on p.7) for reasons I will explain in section 1.6 on p.7.

If applied to a strictly normal distribution, the Median Absolute Deviation robust estimator of scale ($MAD$ - see 2.4.2 on p.20), when corrected by a constant factor, exactly equals the non-robust standard deviation. Thus, the use of the $\sigma / MAD$ ratio is an excellent indicator of the dataset "Gaussianity".

## 1.4 Collapsing time-series spectral data to 1D spectra

Until recently, "classical" data analysis did not automatically include detailed analysis of time-series[7]. From what has been written in the previous sections, it is clear that any data quality assessment must be applied to time-series of spectral power data.

For a single dish radio telescope, collapsing power time-series in each frequency channel into 1D spectra and assessing data quality through through estimators of both location and scale is done according to the following steps (see 3.5.1 on p.48):

1. spectral data normalization, to ensure the maximum coherence of the dataset;

[7]For instance such a process is possible with CLAS but not proposed in standard features: it must be setup by the observer.

2. Doppler correction of the Earth's movement;
3. binning of frequency channels if required;
4. calculation of estimators of location and scale plus data quality estimators of each frequency channel;
5. excision of unwanted dataset outliers including RFI from each channel frequency;
6. step 4 reiterated on clipped dataset.

With the use of robust estimators of location and scale, it is no longer possible to apply binning by $n$ simply by averaging $n$ frequency channels. Indeed, because the median of $n$ medians is not the overall median (the same applies to all robust estimators) as opposed to the case of the average of averages, binning $n$ channels implies merging all the $n$ time-series of power spectral data into one (i.e., building a time series of power values with the $n$ time-series), and then calculating robust estimators of location and scale. I call such a process robust binning (see 3.2.2 on p.34).

Setting interval limits to data quality estimators allows flagging each frequency channel as either "good" or "bad" before and after excision of unwanted ouliers including RFI, according to predefined criteria: such information greatly reduce the number of false positive when applying blind automated detection. This step is performed on the resulting 1D spectra.

## 1.5 Automated detection of spectral lines

Until recently, correlators provided user data which could be handled manually. For instance, spectra have mostly been examined visually on graphs, and this has been enough for the observer to detect a spectral line. What has been done on a limited number of channels (no more than a few thousands per spectrum) has now proven to be impossible to handle manually. Indeed, with broadband spectrometers such as WIBAR see section 3.1 on p.29), it is not uncommon to deal with several million channels per spectrum.

This means the need to automate the detection of spectral line candidates as far as possible. Such was the case for the ALFALFA survey of galaxies in the 21cm HI line at the Arecibo radio telescope, which used a Fast Fourier Transform of cross correlation between the radio signal and templates (Giovanelli et al., 2007; Haynes et al., 2018; Saintonge, 2007).

However, automated line detection is far from a mature technology. For instance, the algorithm initially used for automated source detection in the HIPASS HI line survey at Parkes (Barnes et al., 2001), later applied to its HI Zone of Avoidance Survey (HIZOA), was then considered too unreliable, so its source catalog was created using only visual inspection (Staveley-Smith et al., 2016). I have therefore spent a great deal of time crafting a reliable process of spectral line detection, bearing in mind that this would be an essential prerequisite for so-called blind surveys in which areas on the sky are sampled without a priori knowledge of sources in the field, including their redshifts, as well as broadband pointed surveys.

The post-processing sequence I have implemented is the following. After reducing the time-series of power values to both estimators of location and scale as well as associating data quality indicators to every frequency channel, spectra are then ready for the detection of spectral lines. For 1D spectra, this is done by baseline fitting (i.e., non-linear regression), whereas for 2D channel maps, other techniques may be applied such as shape recognition. Spectral lines are the "good" channels characterized as points deviant from the fitted sky background.

## 1.6 The issue of statistical outliers

From a statistical point of view, any value deviating from the main cloud of measured data is qualified as an outlier (see section 2.2.5 on p.15). RFI as well as spectral lines are thus indifferently identified as outliers from the sky background, hence the absolute need for data quality assessment in order to discriminate between "good" and "bad" values.

Classical estimators of location and scale are respectively the mean and the standard-deviation. Fitting polynomials is usually done with the help of the Least Squares method (LS).

However, here we face a considerable obstacle. Neither mean nor standard deviation are robust: i.e., if only one point of a given sample tends to an arbitrary high value, the mean and standard deviation will behave the same way, thus giving pathological results.

In fact, in the presence of strong RFI, non-robust indicators of location and scale are inoperative.

Since we want to blindly assess samples (we do not know in advance where data will be of good quality or polluted for any reason), using mean and standard deviation is not appropriate. In general what we name an outlier will affect any non-robust indicator, because their influence function is not bounded (see section 2.2.2 on p.14).

Nor is the LS algorithm immune to outliers (see section 2.5.1 on p.23).

Moreover, strong outliers may hide weaker ones. For instance finding a spectral line within frequency channels polluted by strong RFI is an especially difficult task caused by what is called the masking effect (see section 2.2.7 on p.16). We therefore need robust estimators, which by definition are immune to outliers.

## 1.7 Robust statistics

The aim of robust statistics is to provide estimators of location and scale as well as regression algorithms which have the following properties (Rousseeuw et al., 2018):

1. be unbiased (see section 2.2.3 on p.14);
2. have a bounded influence function (limited effect induced by a marginal change in dataset) as explained in section 2.2.2 on p.14);
3. have maximum resistance to outliers (i.e., breakdown point) up to 50% (see section 2.2.6 on p.16) ;

4. be efficient (cf. Fisher Information), i.e., fast convergence (see section 2.2.4 on p.15).

Robust statistics are not new and most algorithms date back to the 1980's and 1990's. They were seldom used at the time since the proposed algorithms are very demanding in terms of computer resources. This has changed however during the last decade due to the rapid evolution of computer resources.

It should be noted that their use in radio astronomy has been relatively limited so far, essentially through the use of the median and MAD.

Nevertheless most authors tend to ignore the fundamental setbacks of MAD that I discuss in section 2.4.2 on p.20: this estimator of scale is relevant only for statistical samples which are symmetrically distributed around their median. Such is seldom the case in the presence of RFI when spectral data time-series are not normally distributed and are often highly skewed.

This led me to explore, evaluate and select other robust estimators of location and scale which I discuss in sections 2.3 on p.18 and 2.4 on p.19. In my studies, as estimators of location I have chosen the median (see 2.3.1 on p.18) and Tukey's biweight (see 2.3.2 on p.18), and as estimators of scale Tukey's biweight (see 2.4.1 on p.19), MAD (see 2.4.2 on p.20), Sn (see 2.4.3 on p.21) and Qn (see 2.4.4 on p.22).

Furthermore, I had to evaluate robust substitutes for the LS method. This led to me to consider the Minimum Volume Ellipsoid MVE (see section 2.5.3 on p.25) as well as the Least Trimmed Squares LTS (see section 2.5.2 on p.24).

## 1.8 The development of the ROBust Elusive Line post-processing software

The writing of post-processing robust software was not planned at the start of my Thesis project. The initial topic of my Thesis was the evaluation of the cosmological density of cold baryonic gas in the nearby universe ($z < 0.2$) using observations of HI absorption lines in the intergalactic medium along the lines of sight towards hundreds of quasars at the NRT.

Before the beginning of my work, more than 200 quasars had been observed between 2005 and 2007 using the current auto-correlator of the NRT for the Nançay Absorption Program (NAP), in a standard position-switching mode using pairs of *ON-OFF* position observations. After spending months analyzing the results it became clear that the existing hardware and software were not adapted to such an extensive study, for the following reasons:

1. the limited capabilities of the auto-correlator (maximum bandwidth of 50 MHz) implied that each QSO had to be observed many times to cover the required frequency band (between 1100 et 1425 MHz). This meant thousands of observations for a few hundred QSOs[8];

2. there were hundreds of false positives in ON spectra alone, making them unusable;

[8] At least 500 QSOs should be observed to make the results statistically significant.

3. worse, the use of OFF-position observations with the aim of normalizing the ON spectra (there is no way to get a true offset with the NRT) multiplied the false positives on the resulting $ON - OFF$ data;
4. the auto-correlator used a faulty process for adjusting output levels in the presence of strong RFI;
5. using the CLAS data reduction software in standard mode (no time series) prevented any data quality assessment, and setting it to include time-series analysis was far from trivial.

However, the new WIBAR broadband spectrometer for the NRT was in its development stage at the time of my abovementioned diagnosis, and its team welcomed me to start again with new campaigns of NAP observations using WIBAR.

I was conscious that WIBAR was not yet fully operational then, and that my observations made on a shared-risk basis would help the debugging process. This meant that observations had to be retried several times, while testing many different setups to eventually get satisfactory and homogeneous results.

Moreover, there was no post-processing package to handle these broadband observations. Its development was necessary specifically to complete extragalactic broadband line surveys.

The development of the ROBust Elusive Line (ROBEL) post-processing software was thus the outcome of empirical needs. It followed an abductive reasoning process which I characterized in my first Thesis on management sciences as "repetitive round trips between field and theory" (Belleval, 2001)[9].

This development has become the bulk of my Thesis work. Dozens of QSOs have been observed in search of narrow HI absorption lines in the intergalactic medium with WIBAR, but the data is still undergoing final processing and will not be ready for scientific purposes before the end of this Thesis.

I have used WIBAR results in order to develop and test the features of ROBEL: in this Thesis, some of them illustrate the functions and properties of WIBAR and ROBEL.

## 1.9 Contents

This manuscript is organized as follows:

In Chapter 2 I lay down the basics of robust statistics used for this project: after having recalled basics of classical statistics in the context of radio astronomy, I discuss the pros and cons of several robust estimators of location and scale, followed by an evaluation of options for robust regression pertaining to the aims of this project, i.e., RFI mitigation and automated spectral line detection.

In Chapter 3, I first describe the technical properties of the WIBAR broadband receiver, and then the fundamental assumptions and choices I have made to set up the architecture of the ROBEL post-processing software.

In Chapter 4, I present results of ROBEL post-processing of the observations of the quasar B0738+313 and the III Zw 35 megamaser which both require RFI excision.

[9] The two others being inductive and deductive reasoning.

I also assess the Global Navigation Satellite Systems (GNSS) and radio navigation (RN) RFI mitigation by ROBEL with case studies in the 1164 − 1400 MHz band.

And in the Conclusions, after having summarized the work completed and in progress, I present several possible paths for future developments.

# Chapter 2

# Theoretical background of robust statistics applied to this project

## 2.1 Introduction: Useful distribution laws and their estimators of location and scale

In radio astronomical spectroscopy we study natural phenomena which, when measured by a receiver, mostly show normal (Gaussian) distributed voltage input values, in particular when the samples are large enough. When converted to spectra by a Discrete Fourier Transform, these become complex values which are subsequently correlated to their conjugates to give power spectra (see 3.1 on p.29). Such squared values samples follow a $\chi^2$ law. Under some observational conditions as stated in 3.2.1, these power values may become normally distributed, by virtue of the Central Limit Theorem. In this chapter, we will consider that such conditions are fulfilled and thus, the following developments are essentially focused on assessing normal distributions through a set of estimators.

### 2.1.1 The normal law

#### 2.1.1.1 Definition and probability density function

A random variable $X$ expressed as an $n$-vector of coordinates $x_i$ , with $1 \leq i \leq n$ and following a normal distribution $\mathcal{N}$ with mean $\mu$ and variance $\sigma^2$ is referred to as:

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{2.1}$$

Its probability density function (PDF) is:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.2}$$

#### 2.1.1.2 Estimators of location and scale

$\mu$ represents the long term expectancy $\mathrm{E}[X]$ of repeated experiments; it is also called the location parameter:

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i \, p_i. \tag{2.3}$$

where $p_i$ are the respective probabilities of $x_i$.

The expectancy $\mathrm{E}[aX_1 + bX_2]$ of the linear sum of two random variables $X_1$ and $X_2$ with real coefficients $a$ and $b$ is the linear sum of their respective expectancies:

$$\mathrm{E}[aX_1 + bX_2] = a\,\mathrm{E}[X_1] + b\,\mathrm{E}[X_2] \tag{2.4}$$

We also need to characterize the dispersion of the sample by evaluating a "typical distance" between points and the mean: the scale parameter. $\widehat{\mathrm{Var}}(x) = \sigma^2$ is the expectation of the squared deviation between the random variable and $\mu$:

$$\widehat{\mathrm{Var}}(X) = \mathrm{E}\left[(X-\mu)^2\right] = \mathrm{E}\left[(X-\mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \tag{2.5}$$

Note that the variance $\mathrm{Var}\left(\sum_{i=1}^{N} a_i X_i\right)$ of a linear combination of $N$ random variables $X_i$ and real numbers $a_i$ ($1 \leq i \leq N$) is (Pelat, 2015):

$$\widehat{\mathrm{Var}}\left(\sum_{i=1}^{N} a_i X_i\right) = \sum_{i=1}^{N} a_i^2 \,\widehat{\mathrm{Var}}(X_i) + 2 \sum_{1 \leq i < j \leq N} a_i a_j \,\mathrm{Cov}(X_i, X_j) \tag{2.6}$$

Where $\mathrm{Cov}(X_i, X_j)$ is the covariance between random variables $X_i$ and $X_j$.

If the $X_i$ are uncorrelated, then $\mathrm{Cov}(X_i, X_j) = 0$. In this case, we get:

$$\widehat{\mathrm{Var}}\left(\sum_{i=1}^{N} a_i X_i\right) = \sum_{i=1}^{N} a_i^2 \,\widehat{\mathrm{Var}}(X_i) \tag{2.7}$$

Actual observations generate datasets which are not only limited in size, but which also include unwanted values due to system imperfections as well as the environment. This is why $\mu$ and $\sigma^2$ cannot be directly calculated: they need to be estimated.

For $\mu$ we need to choose what we call an estimator of location. When repeated experiments are associated with the same probability of occurrence (i.e., a normal distribution when $p_i = cte$), the usual estimator of location is the sample arithmetic mean:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.8}$$

With $\sigma^2$ we associate an estimator of scale which usually is the standard deviation. For the same probability $p_i = cte$ of occurrence, from the sample variance given by:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2.9}$$

we then deduct the standard deviation $\sigma$:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.10}$$

### 2.1.2 The chi-square distribution

A sample of a squared independent standard normal variable follows a $\chi^2$ law with 1 degree of freedom. The sum of the squares of $k$ independent standard normal variables follows a $\chi^2$ law with $k$ degrees of freedom. Its probability density function (PDF) is a special case of the $\Gamma$ function:

$$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1}e^{-x/2} \tag{2.11}$$

Therefore, the average of $k$ squared values of independent standard normal variables also follows a $\chi^2$ law with $k$ degrees of freedom.

When $k$ increases, the Central Limit Theorem (CLT) states that such a sample asymptotically reaches the normal distribution. In practice, the condition is considered fulfilled when $k$ exceeds 50, and in Chapter 3 this will be related to observational and spectrometer setup parameters, since the latter produces power spectral values. Therefore, all following developments refer to the normal law.

There are many different possible estimators of location and scale and one of the goals of this project is to choose those that are being appropriate both to broadband radio astronomical observations and RFI mitigation. More precisely, we need them to be convergent, efficient, unbiased and robust.

Moreover we want to implement a robust version (i.e., one insensitive to sample outliers) of the Least Squares polynomial adjustment to the baseline of spectra, aimed at automated blind spectral line detection.

In the following sections, I will first set some useful definitions and then enumerate robust estimators of location and scale, based on which I will explain my choices for this project. And finally I will discuss robust regression.

## 2.2 General definitions used in this work

Statistical outliers significantly deviate from the rest of the sample data. Outliers may result from intrinsic data variability, errors in recording, or pollution from external sources.

Robust statistical methods are aimed at providing estimators of location and scale as well as regressions which are not affected by outliers (Rousseeuw et al., 2018). Such methods allow the detection of both spectral lines ("good outliers") and RFI ("bad outliers").

In this section I present essential definitions useful for a discussion of the properties of such estimators, both robust or not.

### 2.2.1 M-estimators

M-estimators are a class of extremum estimators aimed at minimizing or maximizing objective functions of parametric models (Rousseeuw et al., 2003, p. 12). An M-estimator of a statistical model is the zero point of the derivative function which allows the estimation of its parameters (Huber, 2011). A set of estimating equations simultaneously embedding data and unknown parameters allow the solving of these parameters.

The parameters $\hat{\theta}$ are the solution of the minimizing of a measurable $\rho$ function:

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} \rho(x_i, \theta) \tag{2.12}$$

The choice of the $\rho$ function defines the type of M-estimator.

If it is possible to differentiate eq. 2.12, the M-estimator is of type-$\psi$, otherwise it is of type-$\rho$.

For type-$\psi$ M-estimators, finding the $\hat{\theta}$ parameters implies solving the equation:

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{n} \rho(x_i, \theta) = 0 \tag{2.13}$$

M-estimators are used to define location and scale estimators as well as regression (Maronna et al., 2018).

For instance, the mean is a type-$\rho$ M-estimator of location being the minimum of $\rho(x, \theta) = \frac{(x-\theta)^2}{2}$. It can be shown that the median (see 2.3.1 on p.18) is also a special case of M-estimator of location (Maronna et al., 2018).

Redescending M-estimators use a type-$\psi(x)$ function which tend to 0 when $x \to \infty$ (Maronna et al., 2018). Such is the case of Tukey's biweight estimators of location and scale (see 2.3.2 on p.18 and 2.4.1 on p.19). Tukey's biweight has been used in the HI Parkes All Sky Survey (HIPASS - see Barnes et al. (2001)) as well as in the Parkes HI Zone of Avoidance (HIZOA) survey (Staveley-Smith et al., 2016).

### 2.2.2 Bounded influence function

The influence function returns the outcome of a single marginal change in the sample on the estimator (Pelat, 2015).

The influence function of a robust estimator is by definition bounded since outliers (see 2.2.5 on p.15) have a limited impact.

### 2.2.3 Unbiased estimator

The bias of an estimator X is the difference between its expectancy $E(X)$ and its true value. It is unbiased if the bias equals 0.

The sample mean and median are unbiased estimators of $\mu$.

$$\widehat{\mathrm{Var}}_{unbiased}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}\,)^2 \tag{2.14}$$

is an unbiased estimator of the variance $\sigma^2$.

In the following sections we will study robust estimators of $\mu$ and $\sigma^2$ which need corrections to become unbiased.

### 2.2.4 Efficiency

The minimum possible variance $e(T)$ of the unbiased estimator $T$ is given by:

$$e(T) = \frac{1/\mathcal{I}(\theta)}{\mathrm{var}(T)} \tag{2.15}$$

(see eq. 2.2.6 on p.16) where $\mathcal{I}(\theta)$ is the Fisher information (Pelat, 2015, p. 284)[1]. If $e(T) = 1$ the estimator is considered fully efficient.

### 2.2.5 Outliers as leverage points

An outlier is defined as a point lying outside the bulk of observations. Such statistical anomalies are the outcome of either faulty experimental setup or true observed anomalies.

In dimension 2, an observation $(x_k, y_k)$ is called a leverage point if $x_k$ lies significantly far away from the majority of the $x_i$ values of the sample.

Generally in dimension $n$, a leverage point $(x_k 1, ..., x_k 1, y_k)$ is an outlier from the $(x_i 1, ..., x_i n)$ values of the dataset (Rousseeuw et al., 2003, p. 6).

In radio astronomy we consider the sky background as the reference for the detection of outliers. Unwanted or "bad" outliers can be due to various instrument malfunctions as well as radio frequency interference (RFI), whereas desired or "good" outliers come from the detection of spectral lines (or transients such as pulsars or fast radio bursts - FRBs).

Whatever their desirability, outliers often perturb statistical analysis, especially when there is no a priori visual assessment of the data cloud by the observer (such is the case for automated blind data processing). Because they have leverage properties (i.e., they exert a strong influence on regression estimators), they can either mask a true spectral line or create numerous false positives when applying statistical regression. Thus the aim of robust statistics is to deal specifically with outliers by applying algorithms which have high breakdown points.

[1] The Fisher information is a measurement of the amount of information that the observable random variable $T$ carries on an unknown parameter $\theta$ which constrains the probability of $T$.

### 2.2.6 Robustness - high breakdown point

The breakdown point of an estimator of a sample of $n$ data points is the smallest fraction of contamination $m/n, m < n$ that can increase the bias of this estimator to an arbitrarily high value (Donoho, 1983; Rousseeuw et al., 2003).

As an illustration, the mean of such a sample has a breakdown point of $1/n$ since a single outlier can make the mean arbitrarily high. This is also the case for variance and standard deviation.

### 2.2.7 Masking effect

After the removal of points with strong leverage properties, others may then appear as strongly influential on the regression whereas these were previously hidden by the strongest. This is called the masking effect which is notably illustrated by Davies et al. (1993) Becker et al. (1999) and Rousseeuw et al. (2003, p. 229).

Indeed the presence of strong RFI signals in a spectrum may mask a close or embedded spectral line in given frequency channels.

### 2.2.8 Third and fourth univariate moments

In the context of this work, data quality assessment consists of evaluating how close the spectral time-series of averaged power values for each frequency channel are to a normal distribution (for a full explanation, see 3.2.1 on p.33). Among the data quality estimators, I have chosen to use the $3^{rd}$ and $4^{th}$ univariate moments.

The $n^{th}$ moment $\mu_n$ of the mean of a variable $X$ is defined using the expectation $E(X)$ as

$$\mu_n = \mathrm{E}\left[(X - \mathrm{E}[X])^n\right] \tag{2.16}$$

while the $n^{th}$ power of standard deviation, $\sigma^n$ is

$$\sigma^n = \left(\sqrt{\mathrm{E}[(X-\mu)^2]}\right)^n \tag{2.17}$$

The $n^{th}$ standardized moment $\frac{\mu_n}{\sigma^n}$ is the ratio

$$\frac{\mu_n}{\sigma^n} = \frac{\mathrm{E}\left[(X-\mu)^n\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{n/2}} \tag{2.18}$$

#### 2.2.8.1 Skewness

The third standardized moment $\frac{\mu_3}{\sigma^3}$ is called skewness. It characterizes the asymmetry of a distribution about the mean:

$$\frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} \tag{2.19}$$

For an $n$-sample of values $x_i$ and mean $\overline{x}$ its skewness $M3$ is calculated as:

$$M3 = \frac{\sqrt{n\,(n-1)}}{n-2} \left[ \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}} \right] \tag{2.20}$$

Since we study univariate distributions (i.e., with only one random variable) in this project, we interpret skewness as follows:

1. $M3 = 0$ implies either that the distribution is symmetrical around its mean or asymmetrical but weighted around its mean;
2. $M3 < 0$ means in general that there is a longer tail on the values distribution below the mean;
3. $M3 > 0$ mostly indicates that there is a longer tail on the values distribution above the mean.

However, the interpretation of $M3$ is not straightforward since its value takes in account the respective weights of tails below and above the mean. Moreover, $M3$ as calculated above is biased.

Nevertheless, a normal distribution implies that $M3 = 0$. Any departure from this nil value may be either part of the instrument transfer function, or an indicator of some potential technical glitch or interference polluting the radio signal.

#### 2.2.8.2 Kurtosis

The fourth standardized moment is the kurtosis. It characterizes the distribution tails and in particular the likeliness of outliers (Westfall, 2014).

$$\frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{4/2}} \tag{2.21}$$

Throughout this manuscript, when I refer to "kurtosis" I actually will use what is called the excess kurtosis $M4$, i.e, the kurtosis $-3$.

So for an $n$-sample of values $x_i$ and mean $\overline{x}$ we calculate its excess kurtosis $M4$ as:

$$M4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3 \tag{2.22}$$

A normal distribution has a nil excess kurtosis. We will see in the following chapters that the kurtosis may become negative for various reasons (notably after data filtering). On the other hand, a large kurtosis may imply calibration or normalization problems between aggregated observations or/and the presence of RFI (see 3.2.4 on p.38).

Kurtosis has been used for RFI detection, as explained by Taylor et al. (2019).

## 2.3 Robust estimators of location

All this analysis demonstrates that non-robust estimators of location and scale should not be used on data samples whose quality is not guaranteed ex ante, such as radio spectra polluted by non-asymptotic Gaussian signals. In the next sections, I will discuss the usage of several robust estimators of location and scale.

### 2.3.1 The Median

The most obvious estimator of location is the median, especially in the context of this project, as the selected estimator must be applicable whatever the distribution of time-series of averaged power values for each frequency channel.

Considering an $n$-sample of numerically ordered values $x_i$, the median is $x_{(n+1)/2}$ if $n$ is odd, or $(x_{n/2} + x_{(n+1)/2})/2$ if $n$ is even.

The median has the maximum possible breakdown point of 50%, and asymptotically reaches its maximum efficiency (i.e., asymptotic efficiency - see 2.2.4 on p.15) of 64% (Rousseeuw et al., 1993).

In the case of a normal distribution $median = mean$, so here the median is an unbiased robust estimator of location.

Note that the median is also a type-$\rho$ M-estimator resulting from minimizing $\rho(x, \theta) = |x - \theta|$ (see 2.2.1 on p.14).

### 2.3.2 Tukey's biweight estimator of location

Tukey's biweight are type-$\psi$ redescending M-estimators (see 2.2.1 on p.14), which means that their $\psi$ functions redescend to 0 with a non-vertical slope. This implies that outliers are not completely ignored, only the extremes are. The general formula of a biweight $\Psi(x)$is (Rousseeuw et al., 2003, p. 129):

$$\Psi(x) = x(1 - (x/k)^2)^2 for\ |x| \leq k; and\ 0\ otherwise \tag{2.23}$$

Tukey's biweight have high breakdown points up to the possible maximum value of 50%.

Extensive experiments must be conducted to appropriately select $k$ in order to maximize efficiency (see 2.2.4 on p.15).

As proposed by the U.S. National Institute of Standard and Technology (NIST) in its Dataplot software[2], the biweight location estimate $y*$ is given by the following recursive process between $y*$ on the one hand and $w_i$ on the other:

$$y* = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2.24}$$

where

[2] https://itl.nist.gov/div898/software/dataplot/refman2/auxillar/biwloc.htm

$w_i = (1 - (\frac{y_i - y*}{cS})^2)^2 \qquad \text{for } (\frac{y_i - y*}{cS})^2 < 1$

$w_i = 0 \qquad \text{otherwise}$

$S = \text{median}\{|y_i - y*|\}$

$c$ is a coefficient defining the threshold of accepted residuals (for example $c = 6$ means up to $4\sigma$).

For reasons of complexity as well as for lack of extensive experiments to select $k$, I decided not to use this robust estimator of location so far. Lecacheux et al. (2013) wrote a prototype of RFI mitigation software for the Nançay decametric radio telescope. After evaluating Tukey's biweight estimator of location, they decided to select the median for the same reason.

Nevertheless, the fact that this estimator has the ability to take weak outliers into account when $c \neq 0$ makes it a candidate for future trials and development of ROBEL software.

## 2.4 Robust estimators of scale

During the past forty years, many robust estimators of scale have been tested, one of the most popular being the Median Absolute Deviation ($MAD$). But surprisingly, the usage of robust estimators of scale is still somehow scarce, and apart from $MAD$, I have basically not noted any application of the two other estimators I have selected for testing and implementing: $S_n$ and $Q_n$, with one exception on the latter.

Here I will describe the relevant properties of the Tukey's biweight, $MAD$, $S_n$ and $Q_n$ robust estimators of scale, and discuss the pros and cons of their application in a data processing package.

### 2.4.1 Tukey's biweight estimator of scale

Tukey's biweight (see 2.3.2 on p.18) can also be used as a robust estimator of scale. For instance, the U.S. National Institute of Standard and Technology (NIST) Dataplot software includes the following[3]:

$$ns_{bi}^2 = \frac{n \sum_{i=1}^{n} (y - y')^2 (1 - u^2)^4}{(\sum_{i=1}^{n} (1 - u^2)(1 - 5u^2))(-1 + \sum_{i=1}^{n} (1 - u^2)(1 - 5u^2))} \tag{2.25}$$

where the summation is restricted to $u_i^2 \leq 1$ ,

$y' = \text{median } y$

and

$u_i = \frac{y_i - y'}{9 * MAD} \qquad \text{for } (\frac{y_i - y*}{cS})^2 < 1$

For matters of complexity as mentioned in 2.3.2 p.18, I have not yet selected this M-estimator of scale.

[3] https://itl.nist.gov/div898/software/dataplot/refman2/auxillar/biwscale.htm

However, as I mentioned there, this particular estimator is worth exploring further. It may have interesting properties in the case of spectral lines being masked by strong RFI. The others, although they are quite efficient in the removal of powerful RFI, have abrupt thresholds that may not always be ideal for preserving weak natural signals which may be hidden by strong ones by the masking effect described in section 2.2.7 on p.16.

### 2.4.2 Median Absolute Deviation

The Median Absolute Deviation ($MAD$) is one of the only robust estimators of scale used of date in radio astronomical software.

For real-time RFI mitigation it is used at the GMRT in the system described by Buch et al. (2019) and Buch et al. (2016) as well as in the LOFAR Online Flagger (van Nieuwpoort, 2016).

The use of MAD for robust correlators has been evaluated through simulations by Fridman (2009) and Fridman (2008).

Among off-line post-processing packages using $MAD$ we find one at the 21CMA interferometer (Huang et al., 2016), and DUCHAMP which is available on the machines of the Australia Telescope National Facility (Whiting, 2012).

#### 2.4.2.1 Definition of $MAD$

The $MAD$ is defined as the median of the absolute deviations from the sample median multiplied by a correction factor $b$ for bias (Rousseeuw et al., 1993):

$$MAD \;=\; b\; med_i \mid x_i - med_j x_j \mid \tag{2.26}$$

The consecutive steps in the calculations are: (1) the median of the sample, (2) absolute values of the differences between each point and the sample median, (3) the median of these absolute differences, and (4) multiplication of the result by $b$.

Since $MAD$ is the median of the half normal distribution of standard deviation $\sigma$, with $b = 1$ we get:

$$MAD = \sigma\sqrt{2}\mathrm{erf}^{-1}(1/2) \approx 0.6744888\sigma \tag{2.27}$$

where $erf$ is the error function[4].

so with $b \approx 1.4826042$ we get $\sigma \approx MAD$.

Note that $MAD$ is also the $75th$ percentile of a symmetric distribution with 0 mean.

[4]For a normally distributed variable $Z$ with mean 0 and $\sigma = \frac{1}{\sqrt{2}}$, $erf(x)$ is the probability of $Z$ taking a value in the $[-x, x]$ interval.

#### 2.4.2.2 Properties of $MAD$

We will therefore use the $b$ value defined above throughout our work, thus making $MAD$ an unbiased estimator of $\sigma$ for normal distributions.

The advantages of $MAD$ are a maximum breakdown point of 50%, a bounded influence function with the sharpest possible bound. It is also fast to compute. However, it has a low efficiency of 37% (Rousseeuw et al., 1993).

$MAD$ should be used with due caution, however:

Because it computes the sample median and then the median of absolute differences with this median, the underlying assumption is that the distribution is symmetrically dispersed around its median. Therefore, if the sample skewness is anything than nil (see eq. 2.20 on p.17), $MAD$ will underestimate $\sigma$, as I will illustrate with examples later on.

Being thus forwarned, I will show that useful applications of $MAD$ do exist for assessing data quality. In particular, I use the $\sigma/MAD$ ratio to evaluate the signal "Gaussianity". If the averaged power spectral values distribution in any given frequency channel is normal, then $\sigma/MAD$ should exactly equal 1. However, we will see that in practice this is almost never the case.

Note that the main reason why $\sigma/MAD$ never equals 1 is often ignored: the simple fact that there is always an instrument transfer function shaped both by software and hardware guarantees that the collected data almost NEVER consists of perfect Gaussian samples.

To flag RFI in a given spectrum I will use $\sigma/MAD$ ratio, which can reach very high values in case RFI pollutes the sampling with outliers being much stronger than most cosmic radio sources present, when the averaged power spectral values distribution on a given frequency channel becomes highly asymmetrical.

Finally $MAD$ is used by the Least Trimmed Squares (LTS) robust regression algorithm (see section 2.5.2 on p.24).

The estimators of scale described in the next two sections, $Sn$ and $Qn$, as proposed by Rousseeuw et al. (1993), do not rely on any estimator of location, but rather on absolute pairwise differences.

### 2.4.3 $Sn$

#### 2.4.3.1 Definition of $Sn$

For each $x_i$, one computes the $n$ absolute differences $|x_i - x_j|$, from which the median $Sn$ is calculated:

$$Sn = cMedian_i \left[ Median_j |x_i - x_j| \right] \tag{2.28}$$

where $c$ is a constant defined as $Sn \approx \sigma$, obtained by asymptotic argument (Rousseeuw et al., 1993):
$c \approx 1.1926$ for a normal distribution.

#### 2.4.3.2 Properties of $Sn$

Because $Sn$ is insensitive to any asymmetry of the distribution, it returns a "typical distance between observations" (Rousseeuw et al., 1993).

$Sn$ is unbiased for a normal distribution and is more efficient $(58\%)$ than $MAD$. Its breakdown point is also the maximum 50%. Its influence function is bounded but has jumps (Rousseeuw et al., 1992), so sharp changes can be expected at given thresholds.

The algorithm to compute $Sn$ directly would take $O(n^2)$ computation time, but Rousseeuw et al. (1993) have provided a routine which only takes $O(n \log n)$ time and $O(n)$ space. I adapted and integrated this routine into my software package.

$Sn$ is one of my favorite statistics since it provides an almost unbiased estimator of scale at a reasonable cost in terms of computing time.

### 2.4.4 $Qn$

Rousseeuw et al. (1993) also proposed the $Qn$ estimator of scale as an interesting alternative to $MAD$.

#### 2.4.4.1 Definition of $Qn$

$Qn$ is the first quartile, i.e., the middle point between the sample median (the second interquartile $Q2$) and the lowest sample value of the absolute pairwise differences of the distribution[5]:

$$Qn = c_n \text{first quartile of } \left(\left|x_i - x_j\right| : i < j\right) \tag{2.29}$$

where $c_n$ is a constant depending on the number of observations $n$ to make $Qn$ unbiased for a normal distribution:

$$c_n = \frac{1}{\sqrt{2}\Phi^{-1}\left(\frac{5}{8}\right)} \approx 2.2219 \tag{2.30}$$

With $\Phi^{-1}$ being the quantile function (i.e., the inverse of the cumulative distribution function)[6].

[5]The first quartile $Q1$ is the middle value between the median ($Q2$) and the lowest value of the sample. The third quartile $Q3$ is the middle point between the sample median and the highest sample value. Note that the interquartile range $IQR = Q3 - Q1$ itself can also be used as a robust estimator of scale, since it can be made unbiased when corrected by a factor $\frac{1}{\sqrt{2}\Phi^{-1}\left(\frac{3}{4}\right)} \approx 1.349$) - in normal distributions, values below $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$ are considered outliers. However, the $IQR$ breakdown point is lower (25%). I therefore have not considered it as a possible choice for data processing in the context of this project.

[6]For a random variable $X$, the cumulative distribution (CDF) $\Phi(x)$ is the probability $p$ that $X$ will be less or equal to $x$. Symmetrically, for a given probability $p$, the quantile function $Q(p)$ returns the value $x$, i.e., the threshold value below which the probability of $X$ having the value $x$ is $p$.

#### 2.4.4.2 Properties of $Qn$

$Qn$ is unbiased for large samples like those I have processed derived from my observations. Because $Qn$ has a serious bias for small samples, Rousseeuw et al. (1993) have introduced a correction from numerical integration for small $n$ samples in order to reduce its bias. Its influence function is smooth (Rousseeuw et al., 1992).

$Qn$ has an efficiency of 82%, much better than $MAD$ or $Sn$.

Rousseeuw et al. (1993) have also provided a routine that reduces computing time and required memory, which takes only $O(n \log n)$ time and $O(n)$ space. However, the experience drawn from my NRT data shows that $Qn$ still takes up to 30 times longer to compute than $Sn$, which in practice makes it impossible to apply to samples exceeding 10000 points.

Note that Fridman (2009) when designing robust correlators investigated the efficiency of $Qn$ but only with the help of simulations and limited LOFAR observations. He also noticed that $Qn$ requires more operating capacity than other estimators of scale such as $MAD$.

This is the reason why $Sn$ is the most competitive indicator of scale among this selection.

## 2.5 The Least Squares and its robust alternatives

### 2.5.1 The Least Squares Regression

Least Squares Regression (LS) is an M-estimator where $\rho(x) = x^2$ (see 2.2.1 on p.14).

Considering an $n$ sample of $(x_i, y_i)$ where $x_i$ is an independent variable and $y_i$ the corresponding observed value.The function $f(x, \boldsymbol{\theta})$ where the $m$ parameters of the $\theta$ vector (also called regression coefficients) must be adjusted in order to minimize $S_{LS}$:

$$S_{LS} = \sum_{i=1}^{n} r_i(\theta)^2 \tag{2.31}$$

Where $r_i$ are the LS residuals defined as

$$r_i = y_i - f(x_i, \boldsymbol{\theta}) \tag{2.32}$$

It is clear that any outlier on the $y_i$ axis will affect the regression coefficients: the LS influence function is therefore not bounded. The LS regression has a breakdown point of $1/n$.

Furthermore, a leverage point strongly affects some parameters of the $\theta$ vector. However, this does not mean that the LS residuals attached to a leverage point are necessarily large: if such a point is coherent with the bulk of the data, it is called a good leverage point.

We should find a robust substitute to the LS regression method, since the presence of outliers that are leverage points somewhere among the millions of frequency channels will not only make baseline fitting unreliable, but it also prevents the detection of spectral lines which are "desirable" outliers against the sky background.

### 2.5.2 Least Trimmed Squares

Rousseeuw et al. (1984) introduced S-estimators that are robust measures of residuals scattering, including the the Least Trimmed Squares (LTS).

#### 2.5.2.1 Definition of LTS

While the Least Squares (LS) regression is aimed at fitting an objective function (see Equ. 2.31 2.2.6 p.16) by minimizing squared residuals on the overall $n$-sample (see Equ. 2.32 p.16), the LTS will minimize the same sum of squared residuals, but only over a subset of $k$ points ($1 < k < n$).

Hence the objective function $S_{LTS}$ becomes (Rousseeuw et al., 2003, p. 132):

$$S_{LTS} = \sum_{j=1}^{k} r_j(\theta)^2 \tag{2.33}$$

where $r_j(\theta)$ are the ordered values of the residuals of the $k$-subset.

This means that only the first $k$ of the smaller squared residuals are considered, while the largest $n - k$ others are ignored in order to make the function fitting free of outliers.

#### 2.5.2.2 Properties of LTS

The breakdown point of the LTS equals $\frac{n+1}{2n}$ and thus converges to $1/2$ when $n \to \infty$ (Rousseeuw et al., 1993).

When $k = n/2$ the breakdown point is 50%.

The number of $k$ subsets with no repetition among an $n$-sample grows quite rapidly as it is equal to:

$$\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!} \tag{2.34}$$

The LTS can therefore only be performed in a practical sense on small samples. In practice we shall rely on randomly generated subsets which should ideally be representative of the overall sample. The chances of getting a subset which contains no more than 50% outliers is proportional to the number $p$ ($p < \binom{n}{k}$) of subsets we built randomly. From the perspective of data processing this means finding a compromise between the cost in CPU time and acceptable results.

RFI or cosmic spectral lines may then be detected by setting a threshold level (e.g., $\pm 3\sigma$), where any $x_j$ with $|r_j| > threshold$ is considered an outlier.

### 2.5.3 Ellipsoid of Tolerance and the Minimum Volume Ellipsoid estimator - MVE

When considering a multivariate sample which follows a normal law of dimension $n$, a usual method to detect outliers is the ellipsoid of tolerance: any value exceeding a defined distance (called the Mahalanobis distance - $MD$) from the sample mean (which is the center of the ellipsoid) is considered an outlier.

However, since $MD$ is not robust, Rousseeuw et al. (2003, p. 258) have proposed the Minimum Volume Ellipsoid (MVE) which is inferred from robust Mahalanobis distances.

#### 2.5.3.1 $n$ dimension normal law

A random vector $X = (X_1, ..., X_n)$ follows a normal law of dimension $n$ $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V})$ if its probability density function (PDF) $f(x)$ equals (Pelat, 2015):

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{V}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right) \tag{2.35}$$

where $\mu$ is a column vector of $n$ components $\mu_i$, $V$ the covariance matrix (which is positive), $|V|$ its determinant (non nil since $V$ is positive), and $V^{-1}$ the inverse of the $V$ matrix (which necessarily exists since $|V| > 0$) also named the precision matrix.

#### 2.5.3.2 Mahalanobis distance

The Mahalanobis distance $MD$ (also called quadratic distance) measures distances from the mean of multivariate data, by taking in account variances as well as covariances between axes; when variables are uncorrelated, the $MD$ becomes a Euclidean distance. Thus the $MD$ is unitless and scale invariant.

From eq. 2.35 we extract the Mahalanobis distance$MD(x)$:

$$MD(x) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\mathbf{x}-\boldsymbol{\mu})} \tag{2.36}$$

If $V$ is the identity matrix then the $MD$ is the standard Euclidean distance between $x$ and $\mu$.

If $V$ is a diagonal matrix (when variables are independent), the distance between two vectors $\vec{x}$ and $\vec{y}$ the $MD$ is the standardized Euclidean distance:

$$MD = d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}} \tag{2.37}$$

When $n = 1$, the $MD$ is reduced to the central score, i.e., the normalized distance to $\mu$:

$$MD = z = \frac{x - \mu}{\sigma} \tag{2.38}$$

The $MD$ is used to identify outliers outside the ellipsoid of tolerance. However, the $MD$ is not robust since it relies on means and variances which have a 0 breakdown point.

#### 2.5.3.3 Ellipsoid of tolerance

The equation $MD(x) = k^2$ is that of an ellipsoid on which the pdf of the normal law is constant. When $k = 1$ it is named a correlation ellipse; the $MD$ is $\chi^2$ distributed (Pelat, 2015).

Considering the random variable $\chi^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})$

The point $X$ lies inside the ellipsoid if $\chi^2 \leq k^2$.

The probability that a point $X$ lies inside the ellipsoid $MD(x) = k^2$ is:

$P_{k^2} = Pr\{\chi^2 \leq k^2\}$. Hence:

$$P_{k^2} = F_{\chi^2{}_n}(k^2) \tag{2.39}$$

where $F_{\chi^2{}_n}$ is the cumulative distribution function (CDF) of the $\chi^2$ law with $n$ degrees of freedom (for $n$ dimensions).

We name $b$ the probability that the normal deviate of the variable lies inside the interval $[\mu - m\sigma; \mu + m\sigma]$, $m$ being a multiple of the standard deviation:

$$b = F(\mu + m\sigma) - F(\mu - m\sigma) = \Phi(m) - \Phi(-m) = \operatorname{erf}\left(\frac{m}{\sqrt{2}}\right) \tag{2.40}$$

where $F$ is the CDF and $\Phi$ is the standard normal distribution.

So any point with an $MD$ less than or equal to $\sqrt{\chi^2{}_{n,b}}$ lies within the ellipsoid of tolerance.

In practice, for the confidence interval of $3\sigma$ I have mostly used in this project,

$b \approx 0.9973$.

The fact that both $\mu$ and $\sigma$ have a breakdown point of 0 means that the ellipsoid of tolerance also has a 0 breakdown point. In the presence of outliers, the value of $\sqrt{\chi^2{}_{n,b}}$ may be excessively high. The ellipsoid of tolerance may therefore include some potential outliers which $MD$ puts more or less close inside the tolerance border, and these points will therefore not be detected as such. This is another illustration of the masking effect (see 2.2.7 on p.16).

#### 2.5.3.4 Minimum Volume Ellipsoid

The Minimum Volume Ellipsoid (MVE) is a robust estimator of both location and scale.

Considering a sample of $n$ points, the MVE provides the ellipsoid of minimum volume centered on its estimator of location and covering at least $h$ points of the sample,

with $h$ between $[n/2]+1$ and $n$ (Rousseeuw et al., 2003, p. 258), where $[n/2]$ signifies the largest integer less than or equal to $n/2$.

Van Aelst et al. (2009) have demonstrated that for a multivariate sample of $n$ points scattered over $p$ dimensions, the maximum breakdown value (50%) is reached when $h = (n+p+1)/2$.

Robust Mahalanobis distance are inferred from the MVE using the same principles explained in 2.5.3.2 on p.25.

The MVE converges weakly to a non-normal distribution at a rate of $n^{-1/3}$, which is rather inefficient (Davies et al., 1992).

Also, the MVE suffers from the same problem as the LTS: it needs random subsets when the data sample becomes too large, because processing the overall data sample may become too time-costly (see 2.5.2.2 on p.24).

Experiments on radio astronomical datasets have shown that the MVE is far too costly to compute compared to the LTS which has a convergence rate of $n^{-1/2}$. This is why the MVE, though being available in the ROBEL post-processing software (see 3 on p.29), was in practice discarded for further use in my project.

## 2.6 Conclusion: selected estimators

In this Chapter, I have demonstrated that using the mean and standard deviation as estimators of location and scale is not appropriate in the presence of strong outliers in the data, such as RFI.

As robust estimator of location, I have chosen the median: it is unbiased and has a robustness of 50% (which is the maximum).

As robust estimator of scale I have chosen $Sn$: not only it is unbiased and has a robustness of 50%, but it also requires less operating capacity than its equivalent $Qn$.

Though $MAD$ is the most commonly used robust estimator of scale in real-time and post-processing software, it has severe limitations which in my opinion are crippling: $MAD$ is only relevant for data samples that are symmetrically distributed around their means.

However, because $MAD$ is arithmetically and exactly a function of the standard deviation $\sigma$, I do use it for data quality assessment through the $\sigma/MAD$ ratio: $\sigma/MAD = 1$ indicates a sample which is normally distributed. Any deviation from 1 means that the distribution may be contaminated by outliers. We will see that using this quotient is an efficient way to detect RFI.

Assessing the quality of time-series of averaged power values of each frequency channel is necessary to discriminate between good outliers (spectral lines) and bad ones (RFI). For this purpose, I also use skewness and kurtosis besides the $\sigma/MAD$ ratio. When defining limits to these indicators, it is possible to flag frequency channels from clean to more or less contaminated by RFI.

Regression is then used for adjustment of the baseline of spectra to detect outliers (either RFI or true spectral lines). Because the Least Squares has a 0 robustness, it is

subject to masking effects and therefore unable to perform blind detection of spectral lines. Hence, It should be replaced by a robust equivalent.

From both alternatives to Least Squares regression (LS), the Minimum Volume Ellipsoid (MVE) and the Least Trimmed Squares (LTS), I have chosen the latter since it is much more efficient.

Tukey's biweight estimators of location and scale have very interesting properties linked to their smooth limits. They could be used to discriminate between strong and weak outliers. However, their application remains to be implemented and will be part of future developments.

# Chapter 3

# Overall presentation of WIBAR and ROBEL

Though the ROBust Elusive Line detector post-processing software (ROBEL) has its origins in supplementing the standard data reduction software of the existing auto-correlator of the Nançay radio telescope (NRT), its development has been closely linked to that of the WIBAR broadband spectrometer which has been installed in parallel on the NRT.

In order to understand the choices made for ROBEL, I will first describe in this chapter the overall architecture of WIBAR, followed by the main assumptions and choices I made for automated blind spectral line detection. After having given an overview of the main features of ROBEL, I will present the possible setups of WIBAR in respect to the ROBEL data processing capabilities. And finally, I will describe the steps of ROBEL data processing.

## 3.1 The WIBAR spectrometer

WIBAR is a broadband spectrometer installed on the NRT for use in parallel with the existing auto-correlator; its architecture consists of:

1. a Field-Programmable Gate-Array (FPGA) Berkeley processor plugged onto CASPER Roach 2 card on board connected to an Analog-to-Digital Converter (ADC) KatADC installed on the NRT focal carriage (550 MHz bandwidth in each of the two linear polarization channels in spectroscopic mode);
2. first line PCs installed in the laboratory and connected by fiber optics, equipped with GTX660 Graphic Processing Units (GPU) cards for data acquisition and long arrays Fast Fourier Transform (FFT);
3. second line PCs dedicated to post-processing (i.e., the production of FITS files upstream from ROBEL after optional calibration and elementary noise filtering[1]).

If required, WIBAR PCs can also produce cross-polarization spectra to calculate Stokes parameters.

[1] ROBEL is installed on other servers connected to the Paris Observatory network.

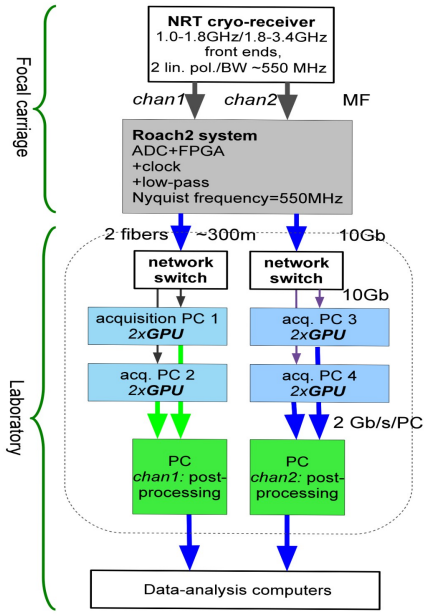


FIGURE 3.1: WIBAR functional diagram.

WIBAR captures real signal voltages from an analogue to digital converter (ADC), sampled at a (Nyquist) rate of $F_s = 1.1GHz$, i.e., 1.1 billion samples per second. Each of the two Nyquist zones has a bandwidth $BW$ which per definition is:

$$BW = \frac{F_s}{2} \tag{3.1}$$

The analogue filter selects the first Nyquist zone, i.e. the $0 - 550$ MHz band. The desired sky frequencies are mapped to the first Nyquist zone by means of mixing, which will not be detailed further in this overview. The ADC sample signal voltages are denoted by $x_n$, with $n$ representing the time index. The corrugated horns of the NRT can be rotated by 90°. In the case of NAP, the receiver detects both horizontal and vertical polarization signals, but only one of the two polarizations is described in this section as the mathematics for the other polarization is the same. Each of the two polarization signals is converted to spectra by applying a Discrete Fourier Transform (DFT), implemented by means of a Fast Fourier Transform (FFT) scheme in WIBAR. A spectrum with $N_{fc}$ frequency channels requires $2^{N_{fft}}$ samples, where $N_{fft} = 2N_{fc}$. The spectrometer applies the DFT to data blocks of duration $T_s$ seconds. The number of real samples $N_{fft}$ in each data block therefore is

$$N_{fft} = F_s T_s \tag{3.2}$$

The signals in a scan are stacked in a transpose vector $\mathbf{x}$ :

$$\mathbf{x} = \left[ x_1, \cdots, x_{N_{fft}} \right]^t \tag{3.3}$$

and $N_{scn}$ scan vectors can be stacked in a $(N_{fft} \times N_{scn})$ matrix $\mathbf{X}$:

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_{scn}}] \tag{3.4}$$

Defining $\mathbf{D}$ as the size $(N_{fft}, N_{scn})$ of the Discrete Fourier transform (DFT) matrix, blocks of $N_{fft}$ samples are converted by Fourier transform $\mathcal{F}$ to spectra $\mathbf{y}_i$, with $\mathbf{y}_i = \mathcal{F}\{\mathbf{x}_i\} = \mathbf{D}\mathbf{x}_i$ or in compact form:

$$\mathbf{Y} = \mathbf{D}\mathbf{X} \tag{3.5}$$

with the Fourier spectra $\mathbf{y}_i$ stacked in a size $(N_{fft}, N_{scn})$ matrix $\mathbf{Y}$ (a triangle windowing is applied to $\mathbf{x}_i$). As the input signals $\mathbf{x}_i$ are real, the spectra $\mathbf{y}_i$ are complex. This part of the length $N_f$ data vectors $\mathbf{z}_i$ of the transpose vector is:

$$\mathbf{z}_i = [y_{i,1}, \cdots, y_{i,N_f}]^t \tag{3.6}$$

where $N_f = \frac{1}{2}N_{fft}$ is the number of frequency bins The power spectrum $\mathbf{r}_i$ is obtained by correlation, that is by multiplying the complex samples in $\mathbf{z}_i$ with their complex conjugate values $\overline{\mathbf{z}_i}$:

$$\mathbf{r}_i = \mathbf{z}_i \odot \overline{\mathbf{z}_i} \tag{3.7}$$

Where $\odot$ is the element-wise matrix multiplication (Hadamard product).

The length $N_f$ vector $\mathbf{r}_i$ is real, and is averaged $N_{scn}$ times yielding an averaged spectral power vector $\mathbf{r}_{av}$:

$$\mathbf{r}_{av} = \frac{1}{N_{scn}} \sum_{i=1}^{N_{scn}} \mathbf{r}_i \tag{3.8}$$

Now, for the required spectral resolution $\Delta f$ of 262 Hz for NAP (see 4.1.2 on p.56), we need an elementary integration time of

$$T_s = \frac{2BW}{\Delta f} \frac{1}{F_s} = \frac{1}{\Delta f} \tag{3.9}$$

which is $T_s \approx 3.82$ms. For NAP, averaged spectra are required every $\tau = 200$ ms (see also 4.1.2), then $N_{scn} \approx 52$.

Denoting $OT$ as the observation time towards one sky position ($ON$ or $OFF$) such as $mod(OT, T_s) = 0$, $N_{rav}$ as the number of averaged spectra during this observation time, and $j$ as the temporal index of averaged power values $\mathbf{r}_{av_j}$ $(1 \leq j \leq N_{rav})$ gives:

1. $1 \leq i \leq (OT/T_s)$;
2. $N_{rav} = OT/(T_s \times N_{scn})$;

and Eq. 3.8 can be updated in order to take into account the $j$ chronological index as follows:

$$\mathbf{r}_{av_j} = \frac{1}{N_{scn}} \sum_{i=(j-1)N_{scn}+j}^{jN_{scn}} \mathbf{r}_i \tag{3.10}$$

Note that WIBAR power averaging is performed at the GPU level, which means that the $\mathbf{r}_i$ power spectra are not accessible to the user due to hardware and storage limitations. Therefore, as a result of a given observation, for each $ON$ and $OFF$ position on the sky (if the latter is required), WIBAR produces (in a series of FITS

files) a 2D $(N_{fc} \times N_{rav})$ matrix which is filled with the chronologically ordered $\mathbf{r}_{av_{j,q}}$ values, $q$ being the frequency channel index ($1 \leq q \leq N_{fc}$) and $j$ the temporal index mentioned above. This 2D matrix is the one that will be referred to in the description of the ROBEL processing in the following sections and chapters.

Compared to the existing auto-correlator which is limited to 50 MHz bandwidth as well as average power spectral values calculated every second at best, WIBAR can produce spectra with millions of frequency channels (up to 33 million) with average power spectral values calculated during intervals as short as $\approx 200\mu$s. Limitations on the combination of both parameters (channel width and sampling rate) are essentially set by the hardware, in particular network bandwidth and disk recording speed.

WIBAR is therefore especially suitable for e.g.:

1. RFI mitigation;
2. broadband HI line surveys covering an 0.2 instantaneous range in redshift;
3. observations of redshifted OH spectral lines, in particular those requiring RFI excision;
4. radio continuum sources;
5. the observation of fast radio bursts (FRBs).

Two apodisation windows have been implemented in the CPU/GPU codes to mitigate the impact of strong signals on nearby and weaker lines of interest. The triangle window (Harris, 1978) has been used all along the observations planned for this Thesis, with a 50% overlap to minimize the loss of observing time. The correlation between two successive spectra is then 25%, and will produce a time-wise statistical covariance on the 2D spectra. This aspect should be kept in mind since residual correlation between two successive windows contributes to create a small but non-nil statistical bias in the temporal series of power values for each frequency channel, which value depends on WIBAR settings. Its potential impact will be discussed in the case studies presented in Chapter 4.

## 3.2 Automated blind spectral line detection: main assumptions and choices

In this Section, I will address the following five topics:

1. the main assumption at the basis of this work: that time-lines series of averaged power spectral data collected from cosmic sources without any artificial interference have close-to-normal distributions, which occurs when the $\chi^2$ law they follow has enough degrees of freedom for the Central Limit Theorem to apply;
2. binning frequency channels, if necessary, requires a specific process which takes into account the properties of robust estimators of location and scale: robust binning, as I call it, consists of merging time-series of averaged power values in the channels to be binned;

3. flagging frequency channels contaminated by RFI, through applying selected non-robust and robust data quality estimators to assess how closely time-series resemble normal distribution;
4. the issue of data calibration, which is a possible factor of increasing *rms*, and spectra normalization, aimed at compensating the increase in *rms* induced by unwanted RF power level variations;
5. my suggestion that, under certain conditions, automated blind detection of spectral lines in pointed observations may be better performed with uncalibrated and normalized *ON* spectra

### 3.2.1 Statistical signatures of cosmic sources vs. RFI

It is supposed that the real samples $x_i$[2] captured by the spectrometer are composed of the sum of two normally distributed independent sub-samples:

- one from white noise $b_i$;
- one from cosmic source signals $s_i$.

From the additive property of the normal law, we deduct that the $x_i = b_i + s_i$ samples are also normally distributed. Therefore, in any given frequency channel $q$ of the resulting spectrum ($1 \leq q \leq N_{fc}$):

- $\mathbf{r}_{i,q}$ power value samples are $\chi^2$ distributed with 2 degrees of freedom (one for the real component, one for the imaginary component);
- $\mathbf{r}_{av_{j,q}}$ averaged power value samples are $\chi^2$ distributed with $2N_{scn}$ degrees of freedom.

By virtue of the Central Limit Theorem, a sample distribution which follows a $\chi^2$ law with $2N_{scn} \geq 50$ (i.e., $N_{scn} \geq 25$) degrees of freedom is well approximated by a normal law (see 2.1.2 on p.13).

It is worth mentioning that the number of averaged spectral power values $\mathbf{r}_{av_{j,q}}$ in a given frequency channel for a given observation must be sufficiently large (in practice at least 200) for the sample distribution to converge sufficiently towards a normal law. This sets a constraint on the observation time $OT$ such that $[OT/(T_s \times N_{scn})] \geq 200$.

Now suppose that the $x_i$ samples include a third sub-sample $a_i$ which originates from an artificial source such as RFI, and thus is not normally distributed. Then the resulting $x_i = b_i + s_i + a_i$ values are not normally distributed. Therefore, the resulting $\mathbf{r}_{i,q}$ and $\mathbf{r}_{av_{j,q}}$ values of the $qth$ frequency channel do not follow a $\chi^2$ law, and by extension the $\mathbf{r}_{av_{j,q}}$ distribution cannot follow a normal law.

Therefore, the main assumption I use to mitigate RFI is that time-line series of averaged power spectral values of cosmic radio sources will exhibit properties of statistical normal distributions for each frequency channel if the Central Limit Theorem applies, whereas any artificial signal carries coherent information which by definition cannot be normally distributed. However, since observational data go through a

[2] See 3.1 for the nomenclature of variables.

hardware setup, the output may be altered due to imperfections, partial failures, imperfect filtering, specular return or digital noise, among others. This implies that the data collected for any cosmic radio source is never perfectly normally distributed.

### 3.2.2 Robust binning of frequency channels

Binning spectra while taking advantage of robust estimators of location and scale requires a specific process: robust binning. Indeed, for a binned frequency channel, contrary to non robust estimators such as the mean and standard deviation, it is not possible to simply infer such estimators from robust estimators of unbinned channels. By principle, robust estimators rely on sorting time-series of averaged power values for a given frequency channel. Thus, when binning frequency channels, one has to merge the time-series of averaged power values and then calculate these estimators on the resulting time-series. Before I explain the details of robust binning, I recall the basics of classical binning.

We can define $\mathbf{R}_{av_q}$, the non-robust estimator of location (EoL) of the $\mathbf{r}_{av_{j,q}}$ averaged power values of the $q^{th}$ unbinned frequency channel ($1 \leq q \leq N_{fc}$), as follows:

$$\mathbf{R}_{av_q} = \frac{N_{scn}}{OT} \sum_{j=1}^{OT/N_{scn}} \mathbf{r}_{av_{j,q}} \tag{3.11}$$

Thus, $\mathbf{R}_{av_q}$ is the averaged power spectral value of the $q^{th}$ frequency channel.

Binning channels in blocks of $k$ (with $mod(N_{fc}/k) = 0$) results in reducing a spectrum to $N_{fc}/k$ frequency channels, each being $k$ times larger in frequency width than the original resolution.

From Eq. 2.4 on p.12, we deduct that the resulting averaged power value $\mathbf{R}_{avbin_p}$ of the $p^{th}$ binned channel ($1 \leq p \leq N_{fc}$, $p = k, 2k, ..., \frac{N_{fc}}{k}$) is given by:

$$\mathbf{R}_{avbin_p} = \frac{1}{k} \sum_{q=(p-1)k+1}^{pk} \mathbf{R}_{av_q} \tag{3.12}$$

In other words, the averaged power value of the $p^{th}$ frequency channel binned $k$ times is the average of the $k$ averaged power values from the $(p-1)k+1$th to the $pk$th channel.

With $k = 2$, the variance $\sigma bin_p^2$ of the time series of averaged power values of two consecutive frequency channels $FC$ of respective index $2(p-1)+1$ and variance $\sigma^2_{2(p-1)+1}$, $2(p-1)+2$ and $\sigma^2_{2(p-1)+2}$ into the $p$th binned frequency channel is given by:

$$\sigma bin_p^2 = \sigma^2_{2(p-1)+1} + \sigma^2_{2(p-1)+2} + 2cov[FC_{2(p-1)+1}, FC_{2(p-1)+2}] \tag{3.13}$$

where $cov[FC_{2(p-1)+1}, FC_{2(p-1)+2}]$ is the covariance of the averaged power values of the two unbinned frequency channels.

Because signals tend to overflow neighboring frequency channels (and this is specially true for strong signals such as RFI), Harris (1978) has demonstrated the need

of windowing the FFT to limit the risk of weak signals being masked by strong ones. In the case of the observations conducted with WIBAR for this thesis, the FFT is apodized with a triangle window in order to limit neighboring channel swamping (see 3.1 on p.29). Since its 6-dB bandwidth approximately equals 1.78 bin (Harris, 1978), neighboring frequency channels still show some level of correlation. However, when binning with $k > 2$ and $k$ being a power of 2 ($2, 4, 8, 16, 32, 64, ...$), such a correlation sharply decreases (Harris, 1978): this means that the covariance tends to zero for $k \geq 4$[3]. Therefore, from Eq. 2.6 on p.12, we deduct that the standard deviation $\sigma bin_p$ of the $p$th frequency channel calculated with the $k$ standard deviations $\sigma_i$ of unbinned channels approximately equals their quadratic sum divided by $k$ when $k \geq 4$ at least:

$$\sigma bin_p = \frac{1}{k}\sqrt{\sum_{i=(p-1)k+1}^{pk} \sigma_i^2},\ 1 \leq p \leq m,\ p = k, 2k, ..., \frac{m}{k} \tag{3.14}$$

In the absence of time-series of spectral data, no standard deviation per frequency channel is available, which is equivalent to assigning the same value $\sigma_i = \sigma$ to each of them. Eq. 3.14 becomes:

$$\sigma bin_p = \frac{\sigma}{\sqrt{k}} \tag{3.15}$$

Given an $rms$ of the overall spectrum, binning frequency channels per $k$ units will result in a reduced $rms_{bin} = \frac{rms}{\sqrt{k}}$.

Hence, binning $k$ times ($k \geq 4$) reduces the overall $rms$ by a factor of $\sqrt{k}$ (less with $k = 2$ because one must take into account the covariance between two consecutive unbinned frequency channels). Indeed, the loss of precision (i.e., broader frequency channels) is the price to pay for an improved signal to noise ratio (SNR) by a factor of $\sqrt{k}$.

When setting a frequency channel width for observations, one should keep in mind that:

1. with a triangle window, frequency channels tend to become uncorrelated starting at binning 4 at least;
2. the binning factor must be a power of 2;
3. if for practical reasons 2D spectra have to be sliced into smaller 2D parts, the number of frequency channels must be a multiple of the binning factor $k$;

Therefore, the observer must set the channel width at least at one-fourth of the maximum frequency width precision required, to ensure an optimum data quality.

Note that assessing the covariance value of two consecutive frequency channels is not straightforward. In order to get the exact standard deviation of a binned channel, one efficient method consists of concatenating the $k$ averaged power values time-series of each frequency channel to be binned into one data set of a frequency channel that is $k$ times larger, and calculating its standard deviation. Furthermore, from

[3]However, the only way to ensure that covariance is nil is to use only one frequency channel on four.

Eq.3.13, we deduct that it is possible to get the covariance as the difference between the variance of the binned channel and the variances of unbinned channels, all of them being calculated with observational data time-series.

To take advantage of frequency channel binning with robust estimators of location and scale requires a specific process. Since all robust estimators of location and scale rely on data sorting, for a binned frequency channel there is no way to infer these estimators directly from those of unbinned channels (for instance, the median of a $k$-binned frequency channel does not equal the median of all medians of unbinned channels). Thus processing time-series of averaged power values becomes compulsory. This is the reason why I call frequency channels binning with averaged power values time-series **robust binning**: such is the standard method implemented in ROBEL.

Consider $k$ consecutive frequency channels, each of them having $n$ time-series of averaged power spectral values. Robust binning of these $k$ channels consists of merging their time-series datasets into one. The resulting $k$-binned channel has a $nk$ time-series of averaged power spectral values on which all the same estimators of location and scales are then calculated. Robust binning may be part of the RFI mitigation strategy: it is much easier to get rid of unwanted outliers if frequency channels are uncorrelated, i.e., not swamped by strong neighboring signals. in Chapter 4, I will present examples of robust binning and analyse resulting estimators of location and scale in the presence or absence of RFI.

### 3.2.3 Data quality estimators

Because data quality assessment is performed on averaged power values time-line series of each frequency channel (see 1.3 on p.4), it must be performed before collapsing 2D arrays into 1D spectra. I have selected a set of non-robust and robust data quality estimators which are now part of ROBEL.

#### 3.2.3.1 Non robust data quality estimators

The first data quality estimator involves the comparison between the theoretical noise of the instrument and the actual *rms* of times-series of averaged power values for each frequency channel.

The theoretical noise *rms* is given by (Rohlfs et al., 2013):

$$rms^2 = \frac{T_{sys}^2}{B\tau} \tag{3.16}$$

where $T_{sys}$ is the system temperature in $K$, $B$ is the frequency channel width in $Hz$ and $\tau$ is the integration time of an averaged power spectrum in seconds: $\tau = T_s N_{scn}$ with $T_s$ being the elementary integration time and $N_{scn}$ the number of averaged spectra (see Eq. Eq. 3.8 and 3.9 on p.31).

In order to compare the theoretical *rms* to the observed standard deviation $\sigma_q$ of the $\mathbf{r}_{av_{j,q}}$ power spectral values time-series of the $qth$ channel, it is easier to work with normalized noise. When neutralizing the system temperature by setting $T_{sys} = 1$ we get the theoretical normalized noise $rms_t$:

$$rms_t = \frac{1}{\sqrt{B\tau}} \tag{3.17}$$

The $rms_t$ is to be compared with the observed normalized standard deviation of the power spectral values time-series of each of the $q$ frequency channels ($\sigma_{norm_q}$), which is given by:

$$\sigma_{norm_q} = \frac{\sigma_q}{\mathbf{R}_{av_q}} \tag{3.18}$$

where $\mathbf{R}_{av_q}$ is the non-robust estimator location (EoL) of the time-line series of averaged power values of the $qth$ frequency channel (see Eq. 3.11 on p.34).

Such a comparison between theoretical $rms$ and observed $\sigma_{norm_q}$ is not used to discriminate between "good" and "bad" frequency channels, as it has proven to be too unreliable essentially because it allows too many entries that may be affected by external factors. Hence there is no consensus on setting quality limits. However it provides extra information if a detailed analysis of a spectrum is required.

The second data quality estimator is skewness (see 2.2.8.1 on p.16), a good quality indicator of observational data time-series. A non-nil value along most of the frequency channels may indicate a temporal drift of averaged power values in frequency channels (see 3.2.5 on p.41). Moreover, the presence of RFI is almost invariably revealed by an exceedingly high skewness.

The third estimator I use is kurtosis (see 2.2.8.2 on p.17) since it provides at least two kinds of information on observational data time-series:

1. an exceedingly large value may reveal the presence of RFI, or a system failure by signaling a data spread which is abnormally high;
2. after time-series clipping, kurtosis is often negative since the wings of the distribution have been cut off. This value gives an idea of the remaining data distribution.

#### 3.2.3.2 The $\frac{\sigma}{MAD}$ robust data quality estimator

From the definition of the robust estimator of scale $MAD$ (see 2.4.2.1 on p.20), we know that $\frac{\sigma}{MAD} = 1$ for a normal distribution. Therefore, this ratio can be considered as a robust data quality estimator, which is ideal for RFI detection and assessment of RFI excision:

- its median on the overall spectrum indicates how close the averaged power time-series of all frequency channels are to a normal distribution;
- any RFI is almost always flagged by a value that is far from 1;
- after data clipping which hopefully removed most unwanted outliers including RFI:
  - any remaining RFI is again flagged by a $\frac{\sigma}{MAD}$ ratio far from 1,
  - its median is once again an assessment of the overall spectrum quality.

Another major advantage of using the $\frac{\sigma}{MAD}$ ratio resides in its ability to reveal subtle RFI which may be invisible when visually inspecting an *ON* spectrum, yet strong enough to compromise data if the unlucky observer is searching for a shallow spectral line at the location of such RFI.

Note that $\frac{\sigma}{MAD}$ never equals 1 in practice because of the intrinsic instrument characteristics which prevent obtaining perfectly normal distributions of spectral data time-series.

#### 3.2.3.3 The Kullback–Leibler divergence

Though not used in this thesis, it is worth mentioning the possible future application of the Kullback–Leibler (KL) divergence (or relative entropy), which measures differences between two probability distributions. For two discrete distributions $Q$ and $P$, their KL divergence $D$ is given by:

$$D(P \parallel Q) = \sum P(x) \log \left( \frac{P(x)}{Q(x)} \right) \tag{3.19}$$

Although for each frequency channel $l$, it would be possible to compare the actual distribution of the averaged power values $P(\mathbf{r}_{av_{j,l}})$ to a theoretical normal distribution $Q$, this does not provide much information on the nature of this divergence. As Eguchi et al. (2006) and Huillery et al. (2007) explain, the KL divergence is also a tool suitable to assess the average decrease of the probability of false alarms $\Delta P_{fa}$ (i.e., in our case, the probability to wrongly flag a detection of a spectral line in a given frequency channel, when the EoL of its averaged power values exceeds the detection threshold, but for any reason except a genuine spectral line) as well as the probability of detection $\Delta P_d$ compared to the ones given by the probability density function (PDF) of the $Q$ distribution law. When the $Q$ distribution is the theoretical modeling of the white noise, the KL divergence measures $\Delta P_{fa}$, whereas if $Q$ represents a non-Gaussian perturbation such as RFI, the KL difference measures $\Delta P_d$. The latter is difficult to implement as long as there is no extensive RFI statistical model, but this could change when sufficient RFI statistical databases and modeling allow efficient machine learning algorithms.

### 3.2.4 The issue of calibration and normalization of spectral data for automated line detection in broadband surveys of pointed observations

ROBEL has been primarily designed for blind automated spectral line detection on pointed observations. Such a task imposes constraints which are distinct from those related to assessing spectral lines characteristics once they are detected. One of them is to minimize the *rms*. While the standard observing mode for line detections is $ON - OFF$ in total power mode, where the $OFF$-source observations are made through frequency or position switching, a tempting strategy would consist of getting rid of the $OFF$ observations which intrinsically adds noise: I discuss its conditions of application[4]. Second, for blind detection in broadband spectra, data

[4]This discussion does not apply to blind HI surveys like HIPASS and ALFALFA: there, no OFF-position observations are made, the spectra devoid of spectral lines are averaged to form the "OFF" reference spectrum which is then substracted from each ON spectrum. In such a case, there is no increase of *rms*

calibration may not be necessarily productive, as it generates an additional source of uncertainty in the *rms*. Third, spectra normalization is a process which may compensate, at least partially, time-drift of power values during an observation (which mainly occurs because of temporal changes in $T_{sys}$ as well as environmental conditions), and thus reduce time-series standard deviation for each frequency channel. From now on, I will refer to *OFF* as position switching which is the only mode I have experimented with so far, though the conclusions I present may be relevant to frequency switching as well.

#### 3.2.4.1 About the use of *OFF*-position observations

Calculating $ON - OFF$ spectra is a classical way of subtracting the frequency response of the telescope (i.e., the *OFF* position) from the on-source observational data (i.e., the *ON* position). Consider an *ON* frequency channel with a $n$-sample of averaged power spectral data time-series. The corresponding *OFF* channel has the same number $n$ of averaged power spectral data for *ON* and *OFF* cycles have the same duration. From Eq. 2.6 we know that for a given frequency channel, the standard deviation $\sigma_{ON-OFF}$ is deducted from the respective standard deviations $\sigma_{ON}$ and $\sigma_{OFF}$:

$$\sigma_{ON-OFF} = \sqrt{\sigma_{ON}^2 + \sigma_{OFF}^2 + 2cov(ON, OFF)} \tag{3.20}$$

If *ON* and *OFF* channels are uncorrelated, we get:

$$\sigma_{ON-OFF} = \sqrt{\sigma_{ON}^2 + \sigma_{OFF}^2} \tag{3.21}$$

By simplification, if we assume that $\sigma_{ON}^2 \approx \sigma_{OFF}^2$ because data are collected in the same frequency channel, the *OFF* position being in the vicinity of the cosmic source and the data collection of both *ON* and *OFF* only separated by a sufficiently short time interval (i.e., one integration cycle duration), we get:

$$\sigma_{ON-OFF} \approx \sqrt{2}\sigma_{ON} \tag{3.22}$$

In fact, the *rms* is theoretically increased by a minimum factor of $\sqrt{2}$ if times series of the corresponding *ON* and *OFF* frequency channels are both uncorrelated and their standard deviation is comparable (the best case). In practice, it is often worse. $ON - OFF$ also multiplies false positives:

- the mechanical movement of the focal carriage, cable strectching and compressing between *ON* and *OFF* positions creates vibrations;
- there is always a time-lapse at the very beginning of the recording of each *ON* and *OFF* when data is less consistent;
- there is no guarantee that the *OFF* position provides a true offset: if unlucky the presence of another radio-source which has an emission spectral line may jeopardize the result (subtracting such an *OFF* spectrum from a line-free *ON* spectrum may give the false impression that an absorption line was detected)[5].

[5] And reciprocally if there is an absorption line on the *OFF* visibility field.

Furthermore, getting rid of $OFF$ during pointed observations multiplies by two the useful observation time on the targeted source for detection on any allocated time-slot.

For all these reasons, detecting spectral lines using only $ON$-source position for pointed observations of targeted sources is an interesting option providing the following conditions are fulfilled:

- the bandwidth around the spectral line should be free of strong RFI which may swamp the cosmic background;
- the spectral line width must be sufficiently differentiated from the specular reflection baseline ripple (Rohlfs et al., 2013) which occurs in some portions of the $ON$ spectra (wavelength around 115 km/s at the NRT at 21 cm);
- robust baseline fitting such as LTS is applied, as non-robust medthods are not reliable enough for automated blind detections.

These ripples can be locally removed in the FFT domain (Butcher et al., 2016). However specular reflection has proven to be by far too complex to be removed from broadband spectra.

Though using only $ON$-source position minimizes the $rms$, it does not provide exact profile parameters of spectral lines. Nor is it applicable for weak and wide lines such as some galaxy profiles. And so far, we do not have enough experience with LTS to propose a solution which could fit the $ON$ spectra baselines (including ripples) while ignoring such shallow and wide spectral lines which could then be detected as good outliers.

#### 3.2.4.2 Calibrating spectra with noise diodes

Calibration aims at deducting a source true visibility (in $Jy$) from its observed averaged power spectral value measured as antenna temperature in K. Primary calibration is performed using documented strong continuum sources with high SNR and no spectral features. This allows to take in account the radio telescope characteristics (i.e., the K/Jy scale). Resulting values are used for the secondary calibration of averaged power spectral values which is done with noise diodes. At the NRT, it is performed at the beginning of each cycle. The cycle is the smallest time interval during which a 2D sub-array is made. Cycle duration is defined in the observational setup: in general, it lasts between one and two minutes. For instance, an observation of 30 minutes with no $OFF$ may generate 15 2D $ON$ FITS files of 2 minute cycles. During one second, power spectral data which include both diode signals and the sky (either $ON$ or $OFF$) are averaged and recorded in arbitrary units into calibration files. A calibration table of diodes is available to calculate the flux density $F_{Jy}$ in $Jy$ :

$$F_{Jy} = \frac{F_{au} T_{cal}}{C_{au} - F_{au}} \tag{3.23}$$

where:

- $F_{au}$ is the recorded flux in arbitrary units when the noise diode is off,
- $C_{au}$ is the averaged recorded flux of both polarizations in arbitrary units when the noise diode is on,

- $T_{cal}$ is the corresponding calibration value extracted from the calibration table used to transform diode or signal deviation $F_{au} - C_{au}$ into a flux density in $Jy$.

$C_{au} - F_{au}$ is an approximation of the flux generated by the noise diode in arbitrary units.

The great advantage of the noise diode calibration is that it takes into account the variations of the electronic gain which is included in $C_{au}$.

Note that one $T_{cal}$ is relevant only on a limited bandwidth (6.25MHz filters for the auto-correlator): therefore the calibration table includes values which do not cover the entire bandwidth.

Also the calibration process has to be adjusted according to the source declination (Matthews et al., 2000).

However, the secondary calibration adds noise and intrinsic variations. Using the work from Kempen et al. (2000), we can approximate the variance of the ratio $\frac{F_{au}}{C_{au}-F_{au}}$[6]:

$$\begin{aligned}\mathrm{Var}\left(\frac{F_{au}}{C_{au}-F_{au}}\right) &= \frac{1}{E(C_{au}-F_{au})^2}\mathrm{Var}(F_{au}) + \frac{E(F_{au})^2}{E(C_{au}-F_{au})^4}\mathrm{Var}(C_{au}-F_{au}) \\ &\quad -2\frac{E(F_{au})}{E(C_{au}-F_{au})^3}\mathrm{Cov}(F_{au}, C_{au}-F_{au}) \\ = \frac{1}{E(C_{au}-F_{au})^2}\mathrm{Var}(F_{au}) &+ \frac{E(F_{au})^2}{E(C_{au}-F_{au})^4}[\mathrm{Var}(C_{au}) + \mathrm{Var}(F_{au}) + 2\mathrm{Cov}(F_{au}, C_{au})] \\ &\quad +2\frac{E(F_{au})}{E(C_{au}-F_{au})^3}\mathrm{Cov}(F_{au}, C_{au})\end{aligned} \tag{3.24}$$

where $E$ is the expectancy of the respective random variables being estimated by their mean (see 2.1 on p.11).

Eq. 3.24 contains only positive terms since $\mathrm{Cov}(F_{au}, C_{au}) > 0$ (i.e., both are positively correlated). Assessing conditions for which $\mathrm{Var}\left(\frac{F_{au}}{C_{au}-F_{au}}\right) \leq \mathrm{Var}(F_{au})$, meaning that calibration would not increase noise, appears rather complex. One of the reasons is the fact that the covariance matrix is rectangular: for any given cycle, the calibration exposure time lasts only one second vs. one to two minute exposure of $ON$ or $OFF$. Because of the variance asymptotic efficiency it is reasonable to assert that intrinsically $\mathrm{Var}(C_{au}) > \mathrm{Var}(F_{au})$. Moreover, if the calibration data includes RFI for any reason, the process becomes faulty. Indeed, it is reasonable to conclude that such a calibration process increases the uncertainty on the result[7].

### 3.2.5 Normalization of spectra

Because WIBAR records data from the instrument without any kind of calibration at first, I have noticed a recurrent problem with the NRT: the aging receiver being

[6] $T_{cal}$ is considered as a constant; therefore it is not included in this calculation.

[7] Also there are specific processes for observations which require high-level flux density precision, such as primary calibration at the beginning and at the end of the observation.

not very stable, it is not uncommon to notice temporal fluctuations of the RF power level from one observation to another, or even within one observation.

Temporal drift of power values time series is a matter of concern since it generates skewness in the data sample. Hence, non robust estimators of location and scale of averaged power values time series may be biased. Nevertheless, a low frequency *rms* is not incompatible with a significant temporal drift of power values time series providing such a temporal drift exhibits a similar pattern on each of the frequency channel. In such a case, the 1D spectrum resulting from the 2D1D process may consist of biased averaged power values albeit with a low frequency *rms*. Indeed, in a spectrum devoid of RFI, a low frequency *rms* only indicates that averaged power values time series of each frequency channel exhibit close statistical distribution properties. Therefore, the frequency *rms* is not proportional to the estimator of location (i.e., mean or median) of the normalized estimator of scale (i.e., standard deviation or Sn) of the temporal drift along the spectrum, but to the estimator of scale of the normalized estimator of scale of temporal drift. In the case studies presented in Chapter 4, I will present spectra with a median of temporal drift of averaged power values close to 13% and a normalized standard deviation of $\approx 0,06\%$, whereas the frequency *rms* is also $\approx 0,06\%$. Note that there is no obvious link between those two in the presence of RFI, though.

Discussions with observers dealing with observational data generated by different instruments on different sites convinced me that such a problem is recurrent at different degrees. Apart from partial hardware failures, such fluctuations can simply find their origins, among others, in fluctuations in air humidity or temperature, as well as in changes in system temperature ($T_{sys}$) and electronic gain during the observation. For instance, at NRT it has been noticed that $T_{sys}$ increases when the focal carriage moves closer to the rail tips, whereas it decreases when approaching the meridian. Also during several weeks in 2018, analysis of uncalibrated power spectral data showed a significant temporal drift during observations, due to a partial hardware failure.

However, for the latter data it appeared to be possible to recover the observations which were only moderately affected, by normalizing intensity of these uncalibrated spectra. The process is the following:

1. choose a protected bandwidth (i.e., RFI free) which will be the array of reference for normalization (e.g: 1421 to 1421.5$MHz$);
2. for each 2D FITS file[8]:
   2.1. calculate the median of observed intensity within the array of reference,
   2.2. divide all the observed data by this median.

More generally, it is not uncommon to notice slight temporal drifts of power values in frequency channels even under optimal conditions. Indeed, almost any observation exhibits a non-nil skewness before clipping and RFI excision. Therefore, I decided to include this normalization process into ROBEL.

It is assumed that data temporal drift is similar along the overall spectral bandwidth. This assumption has proven to be empirically satisfactory so far, but it will require

[8] Each 2D FITS file generated by WIBAR comprises observational data time-series recorded during one cycle.

future thorough investigations. At least observational datasets are rendered completely coherent within, and in the vicinity of, the selected protected bandwidth for normalization.

Should the fluctuations of RF power level prove to be non-linear, it would be possible to implement a more complex normalizing algorithm, for instance by defining several sub-arrays related to sub-bandwidths. So far this has not been needed.

While the ROBEL process of normalizing is quite efficient with linear polarization, it should be noted that it is irrelevant with cross correlation of the Stokes parameters (see 3.1 on p.29), because in this case there is no system temperature in the spectral data. One possibility would then consist of normalizing such spectra with a robust estimation of the *rms* over a protected bandwidth. This option will soon be tested.

### 3.2.6 Data collection setup for automated blind detection of spectral lines

For broadband surveys which require automated blind detection of spectral lines, some of them for long exposures, two essential success factors consist of:

1. minimizing uncertainty on data quality, in particular unwanted temporal fluctations which are uncorrelated to the cosmic sources (and may increase the *rms*);
2. minimizing noise by removing unnecessary inputs: only data which are strictly necessary for spectral lines detection should be kept.

When the standard NRT auto-correlator provides averaged power spectra which are normalized and calibrated, not only are such fluctations barely noticeable as long as the output is not incoherent, but it is impossible to correct them ex post. Worse, the process implemented on the auto-correlator is unreliable in the presence of strong RFI since each spectrum is normalized by its integral. In such a case, this integral deviates because its calculation is not immune to unwanted outliers. This means that the NRT auto-correlator, which has limited capabilities in spectral bandwidth, is moreover not suitable for surveys in strong RFI environments.

On the contrary, because WIBAR produces separate sets of data which allow a total control of normalization and calibration if requested, ROBEL has taken advantage of such a configuration to process normalization of power spectra in a way which minimizes risk: for each cycle, it is processed with a sub-array of power spectral data being extracted from a selected protected bandwidth.

Nevertheless, it should be noted that mixing spectra normalization and calibration should be applied with special care: such a combination of processes could create unexpected sources of errors which would be impossible to isolate, or worse even to detect. This led me to the conclusion that, for broadband surveys and automated blind detection of spectral lines on a single-dish radio telescope, using uncalibrated *ON* spectra allows minimizing noise while keeping the ability to assess the data quality[9]. With this, ROBEL has the ability to:

- precisely diagnose most instrumental problems through its data quality assessment process, which might be invisible with classical analysis;

[9]This conclusion cannot be extended to interferometers which, by their nature, experience specific calibration issues.

- reduce the *rms* of accumulated observations for a given source to its minimum, and so provide the best possible SNR.

Though $ON$ spectra are by definition non-linear and may show numerous irregularities, robust regression using the LTS algorithm (see 2.5.2 on p.24) allows effective baseline fitting with no visual adjustment, as I will show it later.

Exceptions may arise when looking for wide spectral line detection which may require uncalibrated $ON - OFF$ power spectra.

ROBEL is of course also able to apply its algorithms to calibrated $ON - OFF$ power spectra. This allows subsequent detailed analysis of spectral lines with dedicated software.

## 3.3 The ROBEL post-processing software

### 3.3.1 Technical choices

Because ROBEL is intended to process large datasets, it had to be written in a very stable and fast language. Furthermore, part of it had to be parallelized in order to take advantage of the capacities of multi-CPU calculators.

Also I wanted to provide the ability to install the package on almost any kind of Linux machine, the only limitation being both available threads and memory.

The libraries used had to be widely available and documented.

And finally I wanted it to be open-source and run in an open-source environment[10].

Hence ROBEL is written in gfortran. It uses the following libraries:

- cfitsio to read the WIBAR FITS files,
- OpenMP for the parallelized processes,
- pgplot to produce graphs.

It has been tested and validated on all Linux Debian releases from version 7 as well as every Ubuntu LTS from version 12.04 till now.

Although ROBEL has an input data module adapted to the WIBAR 2D FITS file format (axis 1: frequency channels; axis 2: time-line series of averaged power values), it can handle any kind of 2D time-series data file, assuming that a specific sub-program is added to handle the given input format.

The existing software architecture allows future developments for 3D time-series interferometric data reduction if requested.

The ROBEL package essentially consists of two main programs:

- the first reduces 2D time-series into 1D spectra while processing data quality assessment as well as RFI mitigation. I will henceforth refer to this step as the "2D1D process or program";

[10] Such is not the case yet, since a few modules are excerpts from proprietary libraries; they will be replaced in the next versions of ROBEL.

- the second one completes automated spectral line detection essentially with the help of robust polynomial baseline fitting on 1D spectra. I will refer to this step as the "1D process or program"[11].

### 3.3.2 Third-party software inputs

I have used the following pieces of software written by third-parties.

Proprietary software from Numerical Recipes (Press, 1996):

1. Least Squares (non robust regression - see section 2.5.1 on p.23);
2. "ran1" random number generator as a substitute for the original random generator of the LTS and MVE algorithms (see below).

Open Source software:

1. the solar system barycentric doppler shift correction using the SLALIB software (modified by Pierre Colom and myself);
2. the quicksort program which replaced all former versions of sort subroutines which were found to be less efficient[12];
3. The *Sn* (see 2.4.3 on p.21) and *Qn* (2.4.4 on p.22) programs written and documented by Rousseeuw et al. (1993);
4. the PROGRESS program developed and discussed in Rousseeuw et al. (2003) for the LTS (see 2.5.2 on p.24) and the MVE (see 2.5.3 on p.25).

All these modules have been updated firstly to handle 8 bytes real numbers and secondly as non-interactive subroutines.

Furthermore, PROGRESS has been deeply modified:

1. manual settings have been discarded and are now the results of parameters given by the main program;
2. only the modules being strictly necessary to the MVE and LTS have been kept;
3. because ROBEL handles huge amounts of data, MVE and LTS need random selection of samples to be processed (see 2.5.2.2 on p.24 and 2.5.3.4 on p.26). The original random generator of PROGRESS, developed to handle only a few hundred points, has proven to be unsatisfactory and had to be replaced[13].

### 3.3.3 Computer hardware requirements

The 2D1D program has been parallelized with OpenMP. There are several FITS files per observation (lasting from 10 to 90 minutes), and most of the time several observations per source. Because FITS arrays need to be transposed to produce x-axis

[11] This program can also process 1D spectra produced by third-party software; however in the absence of data quality assessment of each frequency channel, it will not be able to discriminate between RFI and actual spectral lines.

[12] Though quicksort can also be quite inefficient when it randomly selects the wrong index to start its sorting process.

[13] This issue has not completely been solved though.

frequency channels and y-axis time line averaged power values series, I have chosen to load all the files into memory in one go and then transpose and process them into one array (resulting outputs being recorded in 1D subsets which are sized to be easily manipulated).

While this is the fastest way to process such data samples (only one upload from the disks is required and no temporary files), in practice it imposes a serious demand on the hardware used:

1. the required memory is proportional to the amount of data (several hundred GB is not uncommon for multiple hours of observations);
2. robust statistics is essentially a matter of data sorting. There is no parallelized sort algorithm, which means that only one thread can process the time-series from a given frequency channel;
3. implementing multiple nodes has not been tested yet (e.g., OpenMP combined with MPI).

At the Paris Observatory, one computer (named Johannes) is quite well adapted to such requirements. Its architecture consists of 48 threads on one node and 1.5TB memory. Most sources are processed with this computer.

2D1D processing of one source with WIBAR, while being set up for a broadband survey, produces hundreds of sliced 1D spectra. The 1D program which handles a 1D file at a time is not parallelized. Moreover, it requires very little memory: the reduced quantity of data is not proportional to the original 2D dataset. While the number of frequency channels for a given 1D file is a parameter, we mostly choose 4096 as a good compromise for an easy use of files and graphs: a 2D spectra made of more than a million frequency channels is then collapsed into 256 1D spectra of 4096 channels each.

I have written a series of bash programs which are able to generate batches of "embarassingly parallel"[14] 1D jobs. They can be spread on several nodes and threads depending on availability[15]. Thus any calculator is able to handle 1D jobs, whether they be simple PCs or configurations with hundreds of threads. This is typically done on the Tycho computer of the Paris Observatory[16].

## 3.4 Setting up WIBAR for ROBEL data processing

As explained in 3.1, WIBAR Roach cards capture waveforms at a Nyquist rate of 1.1GHz. In practice, hardware limitations (e.g., data transfer rate, disk speed) prevent capturing data at such a rate during more than a few minutes. Moreover, even in the absence of such limits, the dataset would be so huge that post-processing would become problematic. Hence compromises had to be tested and implemented in view of the observational goals (broadband survey, RFI mitigation among others).

Apart from using, or not, *OFF*-position observations, the main parameters of the WIBAR setups are the following:

[14] "Embarassingly parallel" is a term used for batches of unparallelized and independent jobs simultaneously launched, each of them on one thread.

[15] Job queues are managed by the SLURM program on the mutualized Paris Observatory servers

[16] which, as opposed to Johannes, has a classical architecture consisting of hundreds of threads spread on dozens of nodes each with a much more limited memory.

1. center frequency of the 550MHz bandwidth;
2. frequency channel width $\Delta f$. For most broadband observations done so far, this has been set at $\Delta f = 262Hz$, i.e., $\sim 55m/s$ at $z = 0$;
3. the integration time $\tau$ of an averaged power spectrum of $N_{scn}$ power values (for instance $\tau = 50ms$). It should be noted that averaging goes against the accuracy of robust estimators of location and scale since the mean is not robust (see 2.2.6 on p.16). The more the power values are averaged the higher the chances are that outliers may be masked;
4. triangle window applied to the FFT in order to prevent sidelobe structures of strong signals such as RFI from swamping weaker signals in neighboring frequency channels. Because signals tend to overflow neighboring frequency channels (and this is particularly true for strong signals such as RFI), Harris (1978) has demonstrated the need of windowing the FFT to limit the risk of weak signals being masked by strong ones. The triangle window is generally considered as the sharpest FFT filter, though it would be interesting to tests others[17].

The WIBAR software suite also includes elementary routines aimed at data filtering, but they are not activated when ROBEL is used because the latter provides much more efficient algorithms. Furthermore, excising unwanted outlier averaged power values with non-robust algorithms could interfere with the ROBEL data quality assessment process as well as create unwanted masking effects (see 2.2.7 on p.16).

Typical WIBAR setups consist of:

1. high sampling rates (around one ms) coupled with kHz frequency channel width for mitigating targeted bandwidths which are heavily polluted by RFI;
2. low sampling rate (around a fraction of a second) together with narrow channel width (hundreds of Hz) for broadband surveys.

## 3.5 ROBEL data processing steps

An observation consists of time-series of averaged power spectral data. These time-series are split into cycles (depending on instrumental setup: in general, between one and two minutes long), each of them being recorded as FITS files by WIBAR. ROBEL first builds the 2D matrix of $N_{obs}$ observations (with respective index $1 \leq l \leq N_{obs}$, $N_{rav_l}$ being the number of averaged power spectra of the $l^{th}$ observation) of the same source with the same WIBAR setup, by reading and sorting the averaged power values stored in these FITS files, according to the process described in 3.1 (p.29). Doppler correction is then applied to each cycle (see 3.5.1.2 on p.48). Before robust binning, the resulting 2D averaged power matrix has a size of $(N_{fc} \times N_{ravtot})$, where $N_{ravtot} = \sum_{l=1}^{N_{obs}} N_{rav_l}$. If robust binning is required (see 3.2.2 on p.34), the resulting 2D matrix has a size of $(N_{fc}/k \times kN_{ravtot})$ size ($k$ being the binning factor).

The resulting 2D matrix is finally collapsed into a 1D spectrum.

The ROBEL data process for a radio source observed at the NRT with WIBAR is split into two steps:

[17] Harris (1978) mentions Tukeys's Biweight (see 2.2.1 on p.14) among others.

1. create a 2D matrix as explained above, to perform data quality assessment on each frequency channel, and to collapse it into a 1D power spectrum after excision of RFI or any other unwanted outlier;
2. perform baseline fitting with non-linear robust regression as well as blind automated detection of spectral line candidates.

### 3.5.1 Collapsing spectral time-series after data quality assessment

The overall 2D1D process is the following:

1. spectra normalization;
2. Doppler correction (solar system barycentric);
3. robust binning, if required;
4. the following calculations are processed on all values of the time-series of each resulting frequency channel (binned or not):
   4.1. non robust and robust estimators of location and scale,
   4.2. skewness and kurtosis;
5. undesired outliers (including RFI) are excised according to an iterated clipping process;
6. the same estimators of location and scale as well as skewness and kurtosis are calculated again, this time on clipped time-series data of each frequency channel.

The result is a 1D spectrum file reduced to estimators of quality, location and scale before and after clipping for each frequency channel, which is ready to be used in the second phase of processing (baseline fitting and automated spectral line detection).

#### 3.5.1.1 Normalizing spectral data time-series

Before any further processing, averaged power spectral data are normalized according to the principles described in 3.2.5 on 3.2.5.

#### 3.5.1.2 Doppler correction

Doppler correction (solar system barycentric) $dfssb(rf)$ is applied to each cycle of the observation. Hence, the time precision is around one to two minutes. During a one-hour observation, this implies a redshift of one channel of 262Hz at most.

The frequency correction applied at rest frequency $rf$ to each spectrum is given by:

$$dfssb(rf) = rf\left(\frac{olsr}{c}\right) \tag{3.25}$$

Where $c$ is the speed of light, $olsr$ the barycentric radial velocity correction due to the motion of the observer with respect to the Local Standard of Rest (LSR), which is the sum of the solar system barycentric velocity ($ervb$), the radial velocity correction

due to the Sun in the LSR ($vsun$), and the radial velocity correction due to Earth spin ($espin$):

$$olsr = ervb + vsun + espin \tag{3.26}$$

Note that the radial velocity correction due to Sun/LSR (km/s) in SLALIB is ∼8.32 km/s. This value changes over time thanks notably to data processed from the Hipparcos and Gaia satellites, and this may cause discrepancies between frequencies where spectral lines lie (see for instance 4.1.5.1 on p.68).

The rest frequencies $rf$ used are those of HI and OH line transitions:

1. HI: 1420.40575177$MHz$
2. OH:
   - 1612.231010$MHz$
   - 1665.401840$MHz$
   - 1667.359039$MHz$
   - 1720.529980$MHz$

These $rf$ values along with the cycle duration provide a precision of $4m/s$ ($\approx 2Hz$), well over the requirements for extra-galactic observations[18].

#### 3.5.1.3 Robust binning

Robust $k$-binning, if required, is applied on normalized spectra after Doppler correction. Only powers of 2 are allowed (in practice: 2, 4, 8, 16, 32, 64) to avoid frequency channels correlation (see 3.2.2 on p.34).

Because each spectrum consists of a very large number of frequency channels (typically over 1 million for a broadband survey), 2D files are cut into sub-arrays made of a number of frequency channels which is equal to a power of 2 (typically 4096) to allow proper $k$-binning as well as results readability[19].

#### 3.5.1.4 Calculations on unclipped 2D spectra

At this stage, the following calculations are performed on time-series of averaged power spectral data for each frequency channel:

1. non robust estimators of location (i.e., the mean) and scale (i.e., the standard deviation) which are essentially used for benchmarking as well as checking coherence of the
2. robust estimators of location (i.e., the median) as well as selected robust estimators of scale, being:
   2.1. $MAD$,
   2.2. $Sn$ ($Qn$ is an option);

[18] The original program was designed for comet observations.
[19] Graphs made of more than a few thousand points are notoriously difficult to interpret

3. as well as the data quality estimators:
   3.1. skewness,
   3.2. kurtosis.

#### 3.5.1.5 Clipping time-series of averaged power spectral data to remove unwanted outliers

Clipping time-series of averaged power spectral data is the process applied by ROBEL to remove unwanted outliers, including RFI. This method is based on the assumption that cosmic source power spectral data time-series follow a statistical distribution which is close to normal with an adequate WIBAR setup (see 3.2.1 on p.33). Data clipping excises data which lie far from the time-series $EoL$ (outside a typical Gaussian profile) and recursively closes the interval.

The clipping interval is centered on the median: $[median - pEoS, median + pEoS]$ where $p$ is a multiple of the selected estimator of scale ($EoS$). The clipping is performed recursively $q$ times: during each iteration, estimators of location and scale are updated and the clipping is repeated with these new estimators. Typical values used with ROBEL are $p = 3$ and $q = 5$ (these are parameters set by the user). Any averaged power spectral data which lies outside the clipping interval at each stage of the process is removed from the time-series. ROBEL simultaneously performs $sigma$, $MAD$, $Sn$, and optionnally $Qn$ clipping. For reasons explained in 2.4.3.2 on p.22,the most relevant clipping is done with $Sn$ and should be the reference for 1D spectra analysis. $sigma$ and $MAD$ clipping are essentially performed to compute the $\frac{\sigma}{MAD}$ data quality estimator (see 3.2.3.2 on p.37).

#### 3.5.1.6 Calculations on clipped 2D spectra

Calculations similar to those listed in 3.5.1.4 in p.49 are performed on the residual clipped time-series:

- robust estimators of location (i.e., the median) as well as selected robust estimators of scale ($MAD$, $Sn$, and optionnally $Qn$);
- data quality estimators (skewness and kurtosis).

Eventually, each 2D spectrum is collapsed into a 1D spectrum file in which data quality as well non robust estimators of location and scale are recorded.

### 3.5.2 Baseline fitting and automated spectral line detection

The 1D file spectrum, being the output of the 2D1D process, is then ready for the second phase of processing: baseline fitting and automated spectral line detection.

#### 3.5.2.1 Flagging frequency channels with data quality estimators

Each frequency channel is flagged as valid if its data quality estimators fit within the criteria defined by the user[20]:

- skewness (maximum: 1);
- kurtosis (minimum: $-1.5$; maximum: 3);
- $\frac{\sigma}{MAD}$ (maximum: 2).

The validity of each frequency channel is tested before and after clipping. Of course the highest quality is reached when testing is positive before clipping.

#### 3.5.2.2 Baseline fitting using robust regression

Baseline fitting consists of applying robust regression to 1D spectra. ROBEL performs polynomial fitting using LTS (see 2.5.2 on p.24) and optionnally MVE (though testing has shown its relative inefficiency - see 2.5.3.4 on p.26), as well as non-robust LS for comparison (see 2.5.1 on p.23). The process is the following:

- only valid channels are used: this means that bad outliers, especially RFI, are ignored;
- robust regression with LTS (and MVE) requires setting a maximum frequency width (LMW) for the targeted spectral lines to be detected. This is due to the fact that a spectral line is a subset of good outliers and can only be considered as such by LTS if their proportion is below its breakdown point (i.e., 50% - see 2.5.2.2 on p.24). This allows to define a working interval for the LTS which is more than twice the LMW;
- the 1D spectrum is then sliced in these successive working intervals;
- the user defines minimum and maximum degrees of polynomials to be fitted with the LTS (between degrees 3 and 10);
- because LTS cannot be performed on the entire dataset of the working interval, random samples must be selected (see 2.5.2.2 on p.24). When setting the size of the random sample, the user must arbitrate between robust fitting accuracy and processing time;
- first polynomial fitting: robust regression within each working interval is performed by LTS with polynomials between minimum and maximum degrees. The ultimately selected polynomial is the one which results in the lowest minimization of squared residuals as defined in Eq.2.33 on p.24;
- second polynomial fitting: to minimize possible fitting errors, the same process is repeated but with a starting gap of half a working interval;
- for each frequency channel, the selected value of the robust regression is the one which has the lowest squared residual in either the first and the second polynomial fitting.

[20] Values given here are from tests conducted during development: they were permissive to allow thorough analysis. With experience they will be narrowed.

This algorithm, while minimizing pathological polynomial fitting, is subject to possible mishaps: if the LTS is improperly setup (for instance with a random sample which is too small), minimizing residuals could mask a spectral line because the selected polynomial could stick to its shape. Indeed, if the LMW maximum allowed frequency width for spectral lines is set too large, the proportion of outliers may be too small and thus at least some or even all of them may not be included in the random subsample. On the contrary, if the actual spectral line gets close to the LMW, the random sample has a good chance to include too many points of the spectral line and pass the LTS breakdown point, preventing automated detection. Therefore, when searching for spectral lines of unknown maximum width, a good method consists of successively fitting the same spectra with different LMW and sizes of random samples.

Note that the same algorithm can be applied with the MVE, except that computation is performed on the robust Mahalanobis distance (see 2.5.3.4 on p.26).

#### 3.5.2.3 Automated blind detection of spectral lines candidates

Spectral lines candidates are flagged as subsamples being considered as good outliers (bad outliers including RFI, are excluded from baseline fitting). For each frequency channel $i$, LTS calculates the normalized residual $nr_f$:

$$nr_i = \frac{Fobs_i - Fadj_i}{FSCE} \tag{3.27}$$

where:

- $Fobs_i$ is the estimator of location ($EoL$) of the averaged power values recorded in the 1D spectrum (see 3.5.1 on 48);
- $Fadj_i$ is the value calculated by LTS as described in section 3.5.2.2;
- $FSCE$ is the final scale estimate performed by LTS after selection of the $k$-subset according to the algorithm described in 2.5.2.1 on 24. This $FSCE$ is the $MAD$ of the $k$-subset of $Fobs_i$.

In classical statistics, one would say that the $EoL$ of the averaged power values of the frequency channel $i$ is at $nr_i\sigma$ from the fitted polynomial.

The user has to select the $thr_{det}$ detection threshold as a multiple of $\sigma$: any frequency channel for which $nr_i \geq thr_{det}$ is then considered as a good outlier and thus as a spectral line candidate. In most cases, $thr_{det}$ is set at 3.

#### 3.5.2.4 Output and graph files from the 1D process

Results of the 1D process are stored in files which can be computed by specialized software, for instance for precise baseline fitting of a detected spectral line, or to draw targeted graphs[21]. To allow visual inspection, files are in ASCII format. They include for each frequency channel:

[21] Though ROBEL produces standardized graphs for each spectrum, they are intended to be practical and readable, but their purpose is not to generate clean outputs for articles or formal presentations. For example, a spectral line does not need to be centered on a graph, and the scales are calculated to fit the entire spectrum.

- data quality estimators before and after clipping (from the 2D1D process);
- estimators of location and scale of averaged power values (unclipped, $\sigma$-clipped, $MAD$-clipped, $Sn$-clipped and optionally $Qn$-clipped, all of them calculated during the 2D1D process);
- proportion of time-series of the averaged power spectral values clipped during the 2D1D process;
- flagging (good or bad channel) according to the user parameters set for 1D process before and after clipping;
- results of LTS (and optionally MVE) for each averaged power values median (unclipped, $\sigma$-clipped, $MAD$-clipped, $Sn$-clipped and optionnally $Qn$-clipped);
- normalized averaged power value $EoL$ $\frac{Fobs_i}{Fadj_i}$ (see 3.5.2.3 on p.52);
- flagging as an outlier (i.e., as a spectral line candidate) if the normalized residual from LTS (or MVE) exceeds the detection threshold set by the user.

Also, for each spectrum, a file including a list of spectral line candidates is produced, along with overall data quality assessment results before and after clipping.

A series of standard graph files is produced by ROBEL for each spectrum before and after clipping:

- normalized noise (see 3.5.2.4 on p.52);
- comparison between the theoretical noise and the normalized non robust $rms$ (see 3.2.3.1 on p.36);
- $\frac{\sigma}{MAD}$ (see 3.2.3.2 on p.37);
- percentage of clipped time-series;
- skewness;
- kurtosis;
- residuals of the LTS (and optionnally robust Mahalanobis distances for MVE).

## 3.6 Summary of WIBAR and ROBEL setup for RFI mitigation and automated line detection

This section summarizes the steps taken in WIBAR and ROBEL parametrization for RFI mitigation in 2D time-line series of averaged power spectra, and automated spectral line detection in collapsed 1D power spectra.

For WIBAR:

1. select the frequency band which includes both cosmic source(s) and the required normalization band (see 3.2.5 on p.41);
2. in addition to scientific goals, take into account the following requirements: define (1) the number of frequency channels $N_{fc}$, and their width $\Delta f$ (see 3.1 on p.29), (2) the integration time of an averaged power spectrum $\tau$, as well as (3) the minimum observation time $OT$ (see 3.2.1 on p.33) so that:

- the number of power values $N_{scn}$ used to calculate each of the averaged power spectral values $\mathbf{r}_{av_j}$ exceeds 25, to ensure that their distribution is sufficiently close to normal in the absence of any non-Gaussian signal captured by the receiver,
- the number of averaged spectral values $\mathbf{r}_{av_j}$ exceeds 200 in each frequency channel, to ensure they are sufficiently numerous for a qualitative assessment and that theirs can be compared to a normal distribution.

For ROBEL:

1. Data quality assessment and RFI mitigation for 2D time-line series of power spectra (see 3.2.3 on p.36):
   - 1.1. select the power spectra normalization frequency band (see 3.2.5 on p.41);
   - 1.2. select upper and lower limits to the following three parameters to validate time-line power spectral values samples in each frequency channel (see 3.5.2.1 on p.51):
     - skewness (see 2.2.8.1 on p.16),
     - kurtosis (see 2.2.8.2 on p.17),
     - $\sigma/MAD$ (see 2.4.2.2 on p.21);
   - 1.3. select $p$ and $q$ factors for outliers excision from power spectra time-line samples (i.e., RFI and anomalies) in each frequency channel by $q \times pEoS$ recursive clipping (see 3.5.1.5 on p.50), where $EoS$ (Estimator of scale) equals $\sigma$ (see 2.1.1.1 on p.11), and calculate $MAD$ (see 2.4.2 on p.20), $Sn$ (see 2.4.3 on p.21), and optionally $Qn$ (see 2.4.4 on p.22);
2. Automated spectral line detection from 1D spectra with non-robust (LS - see 2.5.1 on p.23) and robust (LTS - see 2.5.2 on p.24) baseline fitting (see 3.5.2.2 on p.51):
   - 2.1. set the detection threshold as a multiple of the frequency channels $EoS$ (non robust $rms$, robust $MAD$, $Sn$, and optionnally $Qn$);
   - 2.2. define the maximum line width of spectra ($LMW$) for which the $EoL$ of averaged power values exceeds the detection threshold;
   - 2.3. set the bandwidth to be processed by LTS for the overall spectrum, as a multiple of $LMW$, so that any spectral line will be flagged as a "good outlier" (spectral line candidate) when it has a width below $LMW$ with a corresponding frequency channels $EoL$ of averaged power spectral values above the detection threshold (for instance, $LMW = 250kHz$, LTS interval set at $501kHz$ and detection threshold at $5EoS$),
   - 2.4. select the random sample size for LTS processing, proceed by trial and error: find an acceptable compromise between on the one hand processing time and detection accuracy which are proportional to the random sample size, and on the other hand the probabilities of false or non-detections. Based on first experiences, the reliability of LTS automated spectral line detection grows asymptotically.

# Chapter 4

# Case studies

In this chapter, I present WIBAR observations and their analysis with ROBEL of two different objects in different RFI environments. The first is the QSO B0738+313 where two HI absorption lines had been previously detected in the intervening intergalactic medium. Since their profiles are well documented, I decided to use them as benchmarks for the ROBEL post-processing of WIBAR data. Moreover, in the same observations, I took the opportunity to test the RFI mitigation capabilities of ROBEL on Global Navigation Satellite Systems (GNSS) signals as well as ground-based radars. Here I present a few significant cases (not all of them can be included in this Thesis). The second object is the III Zw 35 OH megamaser whose 1667 and 1665 MHz spectral lines were also well documented in the 1980's and 1990's, before being swamped by RFI from the Iridium satellite constellation. I compare our new NRT results with those derived when the frequency band was devoid of RFI.

## 4.1 Observations of previously known intergalactic HI absorption lines towards the quasar B0738+313

### 4.1.1 Background

B0738+313 was a quasar with a high 21 cm continuum flux density of 2 Jy selected as one of the test cases for the Nançay Absorption Program (see 1.8 on p.8) with the standard autocorrelator on the NRT in 2005-2007, and it was later retained as a test target for WIBAR broadband observations and ROBEL data analysis. The reason for this is that there were two previously known, very narrow HI absorption lines in the intergalactic medium along the line of sight towards the quasar that had been detected with other telescopes and which were of sufficiently large optical depth to make them detectable at the NRT within a reasonable integration time.

The redshift of the quasar is 0.63, and the HI absorption lines were detected at $z = 0.0912$ (1301.62 MHz) for the B1 line system and at $z = 0.2212$ (1163.12 MHz) for the B2 line (Kanekar et al., 2001; Lane et al., 1998; Lane et al., 2000).

A total of 25 hours observing time was scheduled to observe the quasar at the NRT in 2018, but due to an exceptionally high rate of mainly mechanical telescope failures during that period, only a bit less than 2 hours of integration time could be used in practice for further analysis with ROBEL.

### 4.1.2 WIBAR setup

The observational setup was the following:

- frequency band: $1100 - 1425$ MHz. The maximum WIBAR bandwidth was not used because (1) the purpose of this observation was to test known spectral lines, and (2) using unnecessary spectral bandwidth is disk and CPU-time consuming;
- channel frequency width: $\Delta f = 262$ Hz (i.e., $\sim 55 m/s$ at $z = 0$);
- one averaged spectrum every $\tau = 199$ms;
- alternate 2 minute cycles of $ON$ and $OFF$-positions in total power mode;
- spectra in two linear polarizations.

With this setup, from Eq. 3.9 on p. 31, the elementary integration time is $T_s \approx 3.82$ms and power spectra are averaged $N_{scn} = 199/3.82 \approx 52$ times. The Central Limit Theorem applies: in the absence of non-Gaussian samples generated by RFI, the averaged power spectral values of each frequency channel follow a normal law (see 3.2.1 on p.33).

B1 and B2 were first detected at Arecibo. B1 was observed there for 45 minutes (Lane et al., 2000); B2 for 90 minutes by Lane et al. (1998) and 9 hours by Kanekar et al. (2001) who wished to study its complex triple gaussian fitted shape. In order to obtain NRT data with a sensitivity similar to that of the 45-minute Arecibo data, a total of 25 hours of observing time could be scheduled in 2018. However, mainly due to an exceptionally high rate of multiple NRT hardware failures, which are unrelated to WIBAR operations, 93% of the planned observation time could not be used or were lost. Indeed, long duration observations require that spectra profile stay as stable as possible, otherwise, the instability will generate its own noise simply because the normalization process can only compensate up to a certain limit. While such a situation may be acceptable with former generation sofware, this would make little sense to feed ROBEL with mediocre data sets (though, when applied to other observations, its robust algorithms showed their ability to handle such situations and return coherent results), especially because the purpose of the B0738+313 observation was to assess ROBEL potential in RFI excision. Hence, only about 1.75 hours of integration (31906 spectra) was homogeneous and used – equivalent to only about 6% of the integration time of Lane et al. Moreover, the data of one polarization could not be retrieved.

Nevertheless, this dataset was sufficiently large to detect B1 and B2: at 1163 MHz, the latter not only lies relatively close to the observable NRT frequency limit (around 900 MHz), but also right among RFI. Moreover, these WIBAR broadband spectra allowed tests of RFI mitigation, particularly those generated by GNSS satellite constellations (see Appendix A on p.137).

### 4.1.3 ROBEL setup

ROBEL data processing parameters used for this observation are the following:

- total band width 325 MHz;
- 4096 frequency channels per 1D spectrum;

- unbinned channel width 262 Hz;
- rest center frequency: 1420.40575 MHz;
- Solar System barycentric radial velocity correction to LSR (Local Standard of Rest - see 3.5.1.2 on p.48);
- normalization frequency band: $1421 - 1421.5$ MHz;
- clipping at $5 \times 3EoS$ (EoS: estimator of scale);
- line maximum width (LMW) for automated detection: 250 kHz for unbinned channel frequency width being 262 Hz;
- frequency interval for LS and LTS baseline fitting (fit_int.) and automated detection: 525 kHz;
- size of the randomly selected sample for LTS processing (frs_lts): 250 times the number of frequency channels within the interval set for baseline fitting;
- data quality estimators limits:
  - skewness $\leq 1$);
  - kurtosis (minimum: $-1.5$; maximum: 3);
  - $\frac{\sigma}{MAD} \leq 2$;
- detection threshold at $\pm 3EoS$ off the spectrum normalized median.

Before presenting spectra of B1 and B2, I will first give details on the spectrum normalization process which was applied in the frequency band $1421 - 1421.5$ MHz, selected for this purpose because it is part of the protected radio astronomy service band $1420 - 1427$ MHz and as no Galactic HI lines are expected within it over the observed sky region. Then I will present results of mitigation of Global Navigation Satellite System (GNSS) RFI between 1170 and 1380 MHz.

### 4.1.4 Results for the spectrum normalization band

The observation of B0738+313 consists of 7 scans, each of them made of alternate $ON$ and $OFF$-position cycles lasting 1 minute each. One cycle includes 301 spectra, each of them being integrated during $\tau = 0.199$s, the frequency channel width is set at $B = 262Hz$. The total number of $ON$ spectra (which is the same as $OFF$) recorded in the 7 scans is 31906.

The B0738+313 spectra, as well as those of all other targeted sources at $0 \leq z \leq 0.3$, are normalized within the $1421 - 1421.5$ MHz protected frequency band (see 3.2.5 on p.41). This band is split into 1908 frequency channels. Therefore, statistics are calculated for each cycle on $301 \times 1908 = 574308$ uncalibrated averaged power spectral values (in arbitrary units). In the absence of RFI nor Galactic lines, the analysis of this portion of spectrum returns the best possible results in terms of temporal and frequency noise. Here I present results for unbinned data (noted as BIN1 on graphs).

#### 4.1.4.1 Preliminary assessment of observational data

The preliminary assessment of observational data within the selected protected band before any subsequent processing by ROBEL is an important step to detect any serious anomaly (such as hardware malfunction) recorded in the scans. At this stage, no Doppler correction is applied; also, the information is available on a chronological basis, because data recorded in cycles have not yet been merged in a single array, and subsequently sorted for the purpose of building robust estimators of scale. All this preliminary statistical information is processed and stored by ROBEL for further analysis, such as that presented here.

As explained in 3.2.5 on p.41, for each cycle, spectra normalization consists of dividing all averaged power values by the median of the averaged power values within the protected band. Appendix C on p.153 illustrates the importance of such a process: it provides statistics on ON cycles only. Columns are:

1. Scan #: scan number recorded in the NRT database;
2. Cycle #: 60-second cycle number;
3. Doppler: Doppler frequency shift (Solar System Barycentric) calculated at rest frequency 1420.40575 MHz in the Local Standard of Rest (LSR) and expressed in:
   (a) number of frequency channels (of 262 Hz each),
   (b) *Hz*.
4. Intensity: statistics calculated on uncalibrated averaged power spectral values (i.e., in arbitrary units) in the spectral normalization frequency band for each cycle:
   (a) Median (1): median,
   (b) % of median variation compared to the previous cycle,
   (c) average,
   (d) % of average variation compared to the previous cycle,
   (e) standard deviation,
   (f) normalized standard deviation (i.e, standard deviation / average),
   (g) skewness,
   (h) kurtosis.

For each scan, the following quantities are calculated between the minimum and maximum, as well as the values between the first and the last cycle for:

- median;
- average;
- standard deviation;
- skewness;
- kurtosis.

The same quantities were also determined for the overall observation.

The analysis of the *ON* cycle observations reveals that, not only are there variations of the median of uncalibrated averaged power spectral values within a given scan (between 4.42% for scan 243923 and 7.53% for scan 244808) which are far from negligible, but the variation for the entire observation is as high as 22.2%. Without normalization, processing estimators of location and scale would simply be meaningless, as homogeneous samples need to be processed at comparable intensity levels..

In each cycle, the normalized standard deviation is relatively constant (around 13% with an overall variation of 0.86%). Skewness is slightly negative and close to 0, which indicates a small but persistent temporal drift during a scan. Kurtosis is also slightly negative and close to 0, which indicates that the sample distribution inside a 1-minute cycle is very close to a normal distribution. This is also confirmed by the proximity of the sample mean and median (a relative difference around 0.6 and 0.7%). The fact that each sample mean is slightly inferior to its median confirms the skewness values mentioned above.

It is important to underline that these temporal drifts of median and average are anything but linear. For instance, in scan 244045, the variations in the median between two consecutive cycles range from $-0.56$ to 2.58% with irregular trends.

#### 4.1.4.2 Data quality assessment of the normalization band before clipping

As expected in this protected band, the spectra are smooth and devoid of any RFI. Data quality assessment graphs for a twice a larger frequency band, $1421 - 1422$ MHz, show that before clipping:

- the theoretical normalized noise (see 3.18 on p.37) is close to the median of the observed normalized temporal noise before clipping. However the latter is slightly lower (Fig. 4.1 on p.60 compares the squared inverses of normalized theoretical and observed temporal noises). This is because (1), the averaged power spectral values samples in the frequency channels are not exactly normally distributed. According to Yuan et al. (2005), with a positive kurtosis, $\sigma$ is biased and underestimated. Though there are methods to estimate the bias of $\sigma$, there is no obvious solution for robust estimators of scale.(2) Overlapping and windowing in the Fourier domain induce another statistical bias (see 3.1 on p.29), and windowing cannot completely suppress a covariance factor between neighboring frequency channels which are thus not completely independent (see Eq. 3.13 on p.34). With the WIBAR setup used for this observation, less than 1% of data is overlapped, but it is difficult to assess their impact on estimators of location and scale. Since spectral data are homogeneous and devoid of any RFI, we can consider by approximation that these overlapped data would create a bias of underestimation on estimators of scale of about the same percentage, but this question remains to be deepened;
- a moderately positive skewness (median $\sim$0.29) highlights a globally homogeneous temporal drift of averaged power spectral values along the frequency channels (see Fig. 4.2 on p.61)
- kurtosis is close to that of a normal distribution of averaged power spectral values (median $\sim$0.13) and also globally homogeneous along the frequency channels (see Fig. 4.2);

- the overall median of $\frac{\sigma}{MAD}$ is ∼1.19 in a $3\sigma$ interval of 0.019 (see Fig. 4.3 on p.61): empirically and by comparison with other observations, such values are acceptable, but other sources were observed with a $\frac{\sigma}{MAD}$ as low as ∼1.07 (i.e., much closer to a normal distribution);
- the normalized median of temporal $\sigma$ and $Sn$ are ∼0.13 (to be compared with similar cycle values discussed in 4.1.4.1 on p. 58), whereas $MAD$ is ∼0.11 (see Fig. 4.4 on p.62; Fig. 4.5 on p.62 provides a zoom on temporal $\sigma$ in the 1421 − 1421.5 MHz band, which is the blue line of Fig. 4.4). The fact that $MAD < \sigma$ is another indication of skewness of the flux density temporal distribution (see 2.4.2.2 on p.21);

In order to compare the *rms* of $(ON - OFF)/OFF$ data to those of $ON$ or $OFF$ in themselves, it is necessary to divide their averaged power spectral values in each frequency channel by the value of the fitted poynomial. The unclipped mean of averaged power spectral values are LS-fitted, whereas $MAD$ and $Sn$-clipped averaged power spectral values are fitted with LTS.

By doing so, in the 1421 − 1421.5 MHz band, for the unclipped mean of averaged power spectral values, the frequency *rms* of the $ON$ spectrum is ∼0.00066 (∼17 mJy for $T_{sys} = 35$ K and a pointed source efficiency $PSE = 1.4$ K/Jy[1]), ∼0.00064 (∼16 mJy with the same $T_{sys}$ and $PSE$) for the $OFF$ spectrum, whereas the $(ON - OFF)/OFF$ *rms* is ∼0.00106 (∼27 mJy). The ratio between the latter and the quadratic sum of the $ON$ and $OFF$ *rms* values (∼0.0009) is ∼1.158, underlying again some negative covariance between frequency channels.

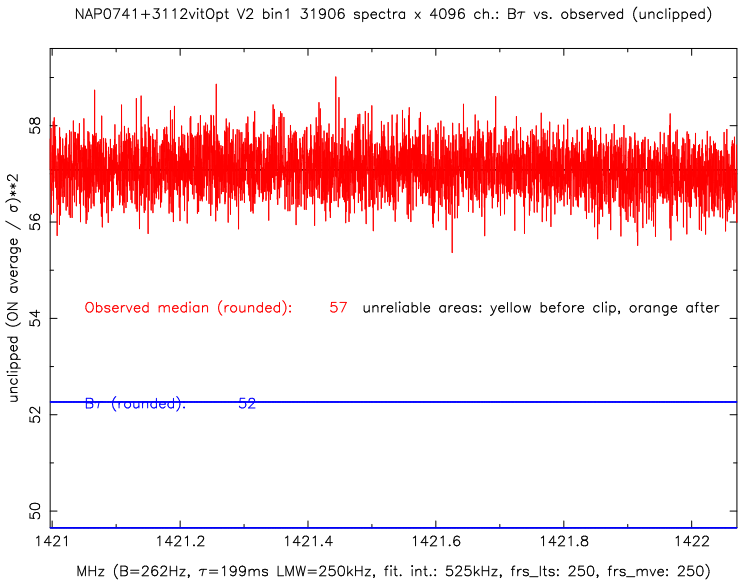


FIGURE 4.1: B0738+313: comparison between squared inverses of normalized theoretical noise ($B\tau$) and unclipped observed temporal noise, (unclipped $ON$ flux density average/$\sigma^2$), within the protected 1421 − 1422 MHz frequency band. This graph, and all others showing the 1421 − 1422.2 frequency range, were generated automatically as part of the standard ROBEL package.

[1] http://nrt.obspm.fr/nrt/obs/NRT_tech_info.html

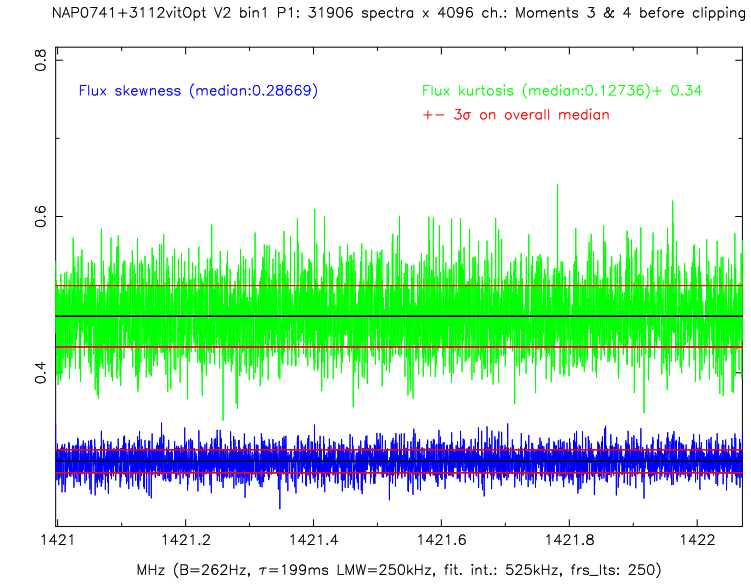


FIGURE 4.2: B0738+313: skewness and kurtosis before clipping within the 1421 - 1422 MHz band.

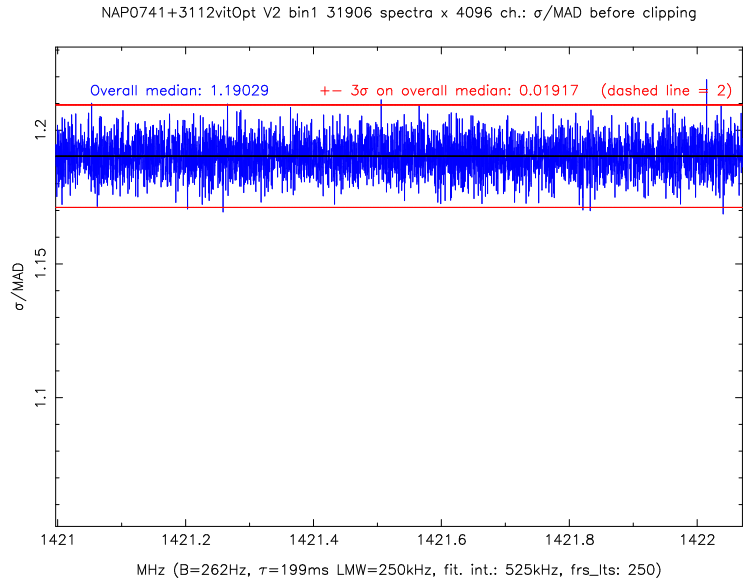


FIGURE 4.3: B0738+313: $\frac{\sigma}{MAD}$ before clipping within the 1421 - 1422 MHz frequency band.

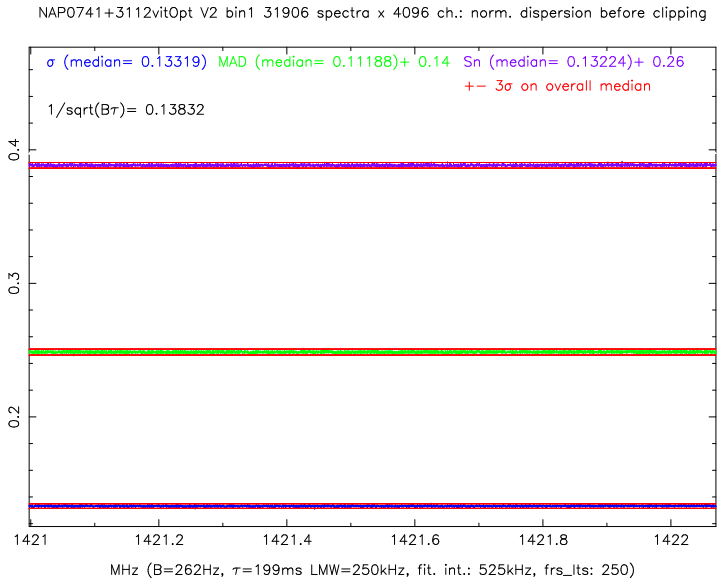


FIGURE 4.4: B0738+313: normalized estimators of scale before clipping within the 1421 - 1422 MHz frequency band.

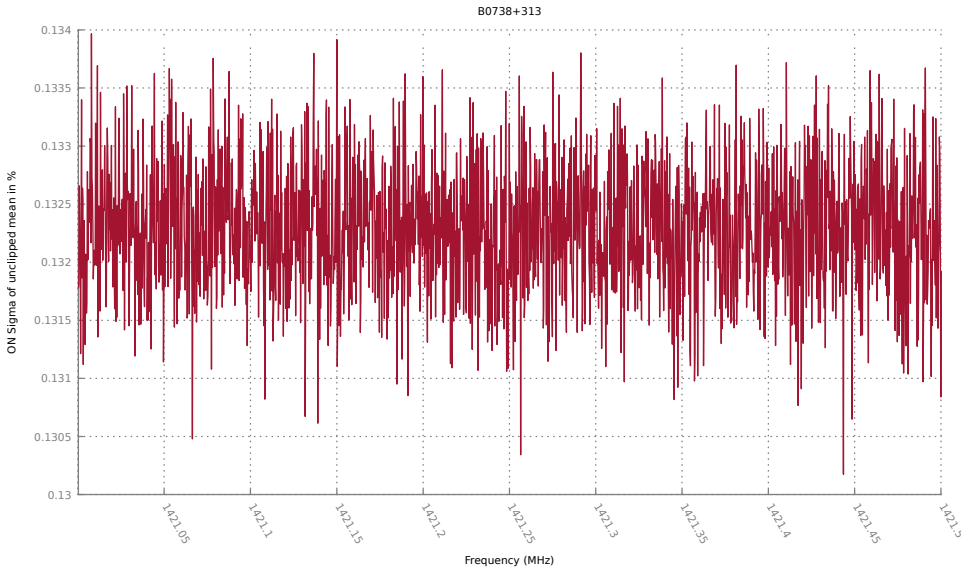


FIGURE 4.5: B0738+313: normalized standard deviation before clipping within the 1421 − 1421.5 MHz frequency band of the *ON* spectrum. This graph is a zoom of the blue line in Fig. 4.4.

#### 4.1.4.3 Data quality assessment of the normalization frequency band after clipping

Data quality assessment graphs for the protected band 1421 − 1422 MHz) show that after $5 \times 3EoS$ (EoS: estimator of scale) clipping:

- a median of 0.5% data has been excised by $\sigma$-clipping, 0.46% by $Sn$-clipping, and 1.5% by $MAD$-clipping (with a non-nil skewness, $MAD$ underestimates $\sigma$), see Fig. 4.6 on p.64;
- the skewness median has been reduced to $\sim$0.176 after $\sigma$-clipping, and $\sim$0.18 after $Sn$-clipping (see Fig. 4.7 on p.64);
- the kurtosis median has been severely reduced to $\sim -0.2$ after $\sigma$-clipping, and $\sim -0.19$ after $Sn$-clipping (see Fig. 4.8 on p.65);
- the overall median of $\frac{\sigma}{MAD}$ after clipping is $\sim$1.18 in a $3\sigma$ interval of 0.02, almost exactly the same as the unclipped values (see Fig. 4.9 on p.65)
- the normalized median of temporal $\sigma$ and $Sn$ have barely changed, and stay around 0.13% (see Fig. 4.10 on p.66); Fig. 4.11 on p.66 provides a zoom on temporal $Sn$ in the $1421 - 1421.5$ MHz band, which is the purple line of Fig. 4.10).
- the LTS fitting of the $ON$ $Sn$-clipping has performed well enough so that the sum of squared residuals $\chi^2$ for the $1421 - 1422$ MHz band is $\sim$0.18 for 4096 channels;
- there are 16 single frequency channels flagged by ROBEL for possible detections at $-3\sigma$ flagged, and 4 at $+3\sigma$ (none of them are consecutive). These are obviously false positive, since their total number of 20 is close enough to the 0.23% probability at $\pm 3\sigma$, i.e., $\sim$9 frequency channels[2].

Fig. 4.12 on p.67 shows the normalized residuals of the LTS from which detection candidates are extracted: for a detection at $\pm 3\sigma$, listed candidates are frequency channels which normalized LTS residuals exceed $\pm 3$.

In the $1421 - 1421.5$ MHz band, for $Sn$-clipped averaged power spectral values, the frequency $rms$ of the $ON$ spectrum is $\sim$0.00084 ($\sim$21 mJy for $T_{sys} = 35$K and a pointed source efficiency $PSE = 1.4$ K/Jy, $\sim$0.00086 ($\sim$21 mJy with the same $T_{sys}$ and $PSE$) for the $OFF$ spectrum , whereas the $(ON - OFF)/OFF$ $rms$ is $\sim$0.00133 ($\sim$33 mJy). The ratio between the latter and the quadratic sum of $ON$ and $OFF$ $rms$ ($\sim$0.0012) is $\sim$1.1.

These frequency standard deviation values are higher than those calculated using unclipped means of averaged power spectral values. Because this frequency band is devoid of any RFI, temporal samples of averaged power spectral values in each frequency channel are close to a normal distribution. In such a case, robust estimators of location and scale are less efficient, since they asymptotically tend to the mean and standard deviation: indeed, clipping a close-to-normal distribution may slightly deteriorate samples.

[2]ROBEL can be set to ignore single detected channels to decrease the number of false positive.

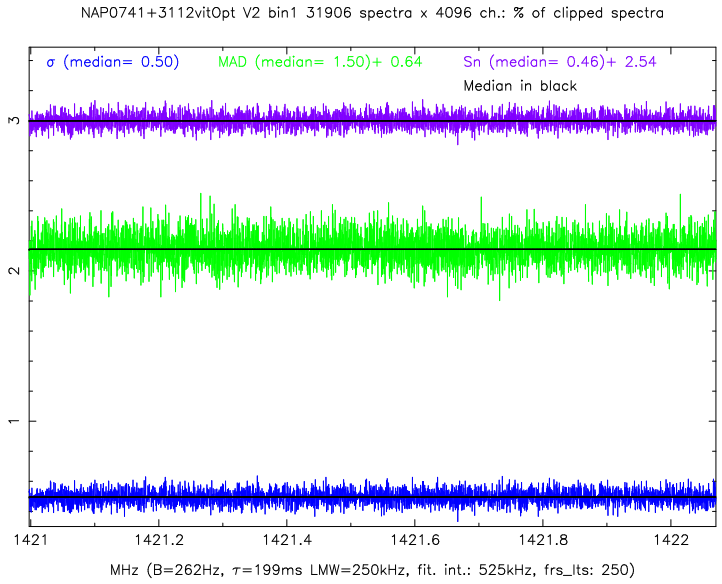


FIGURE 4.6: B0738+313: percentage of averaged power spectral values $\sigma$-clipped (blue line), *MAD*-clipped (green line), and *Sn*-clipped (purple line) within the 1421 - 1422 MHz frequency band.

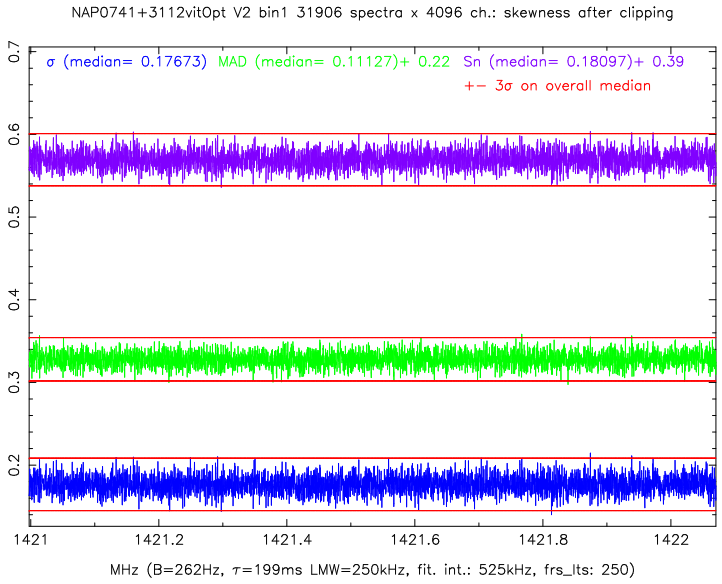


FIGURE 4.7: B0738+313: skewness after $\sigma$-clipping (blue line), *MAD*-clipping (green line), and *Sn*-clipping (purple line) within the 1421 - 1422 MHz frequency band.

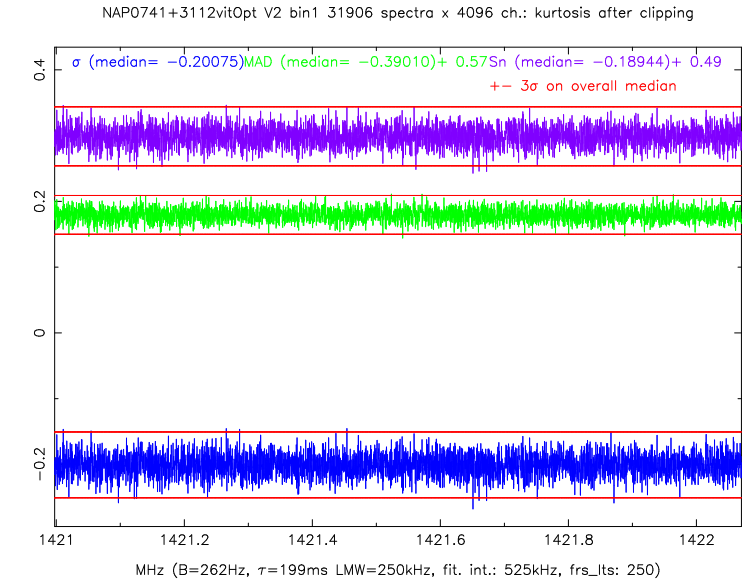


FIGURE 4.8: B0738+313: kurtosis after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1421 - 1422 MHz frequency band.

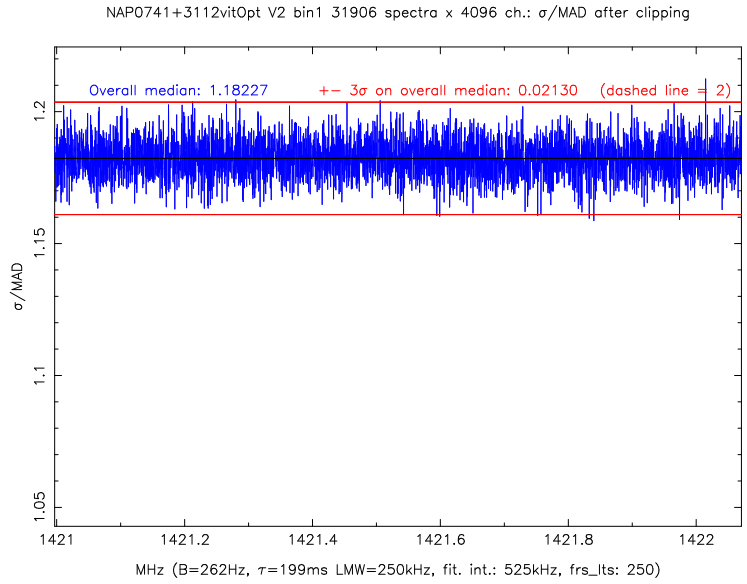


FIGURE 4.9: B0738+313: $\frac{\sigma}{MAD}$ after clipping within the 1421 - 1422 MHz frequency band.

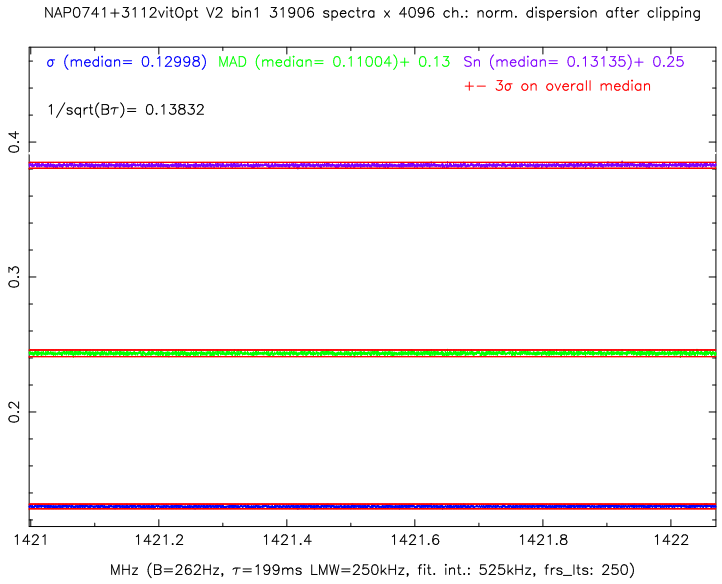


FIGURE 4.10: B0738+313: normalized estimators of scale after $\sigma$-clipping (blue line), *MAD*-clipping (green line), and *Sn*-clipping (purple line) within the 1421 - 1422 MHz frequency band.

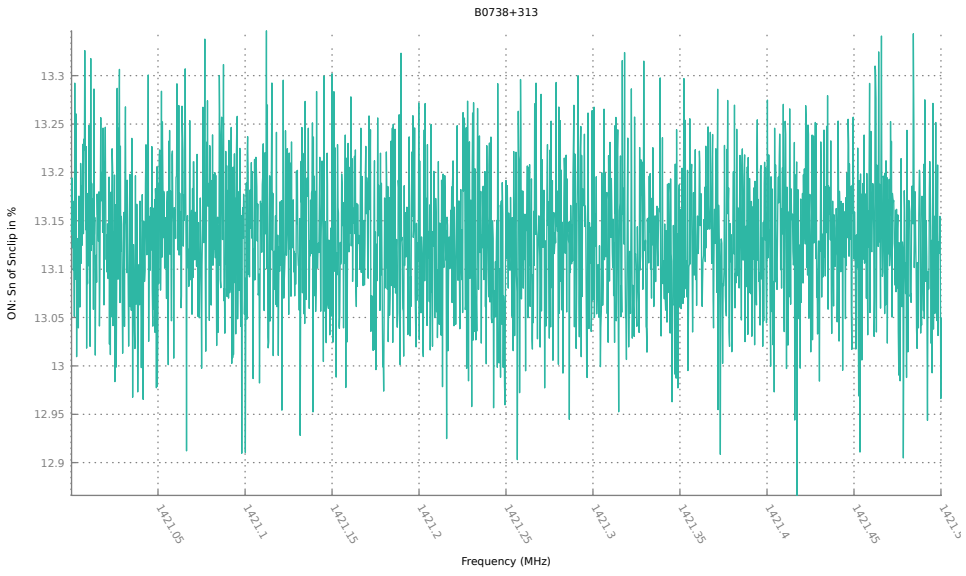


FIGURE 4.11: B0738+313: normalized *Sn* of the *Sn*-clipped averaged power spectral values within the $1421 - 1421.5$ MHz frequency band of the *ON* spectrum. This graph is a zoom focused on the purple line of Fig. 4.10.

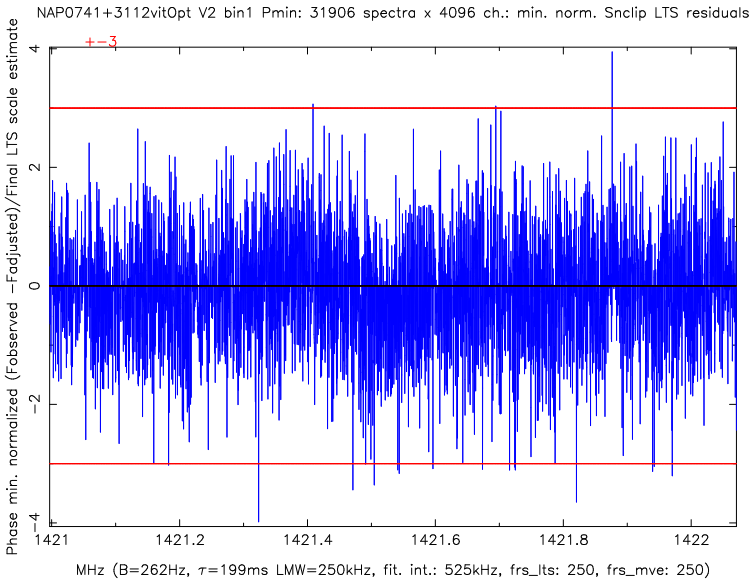


FIGURE 4.12: B0738+313: normalized residuals after LTS baseline fitting of the *Sn*-clipped spectrum within the 1421 - 1422 MHz frequency band. Values represented are the normalized difference between the *Sn*-clipped averaged power spectral values and the polynomial, divided by the final estimator of scale (i.e., the *EoS* calculated after outliers rejection by LTS). Frequency channels are flagged as possible detections if their value exceed $\pm 3$.

### 4.1.5 Results for the first absorber at $z = 0.0912$

According to Lane et al. (2000), the 21 cm line of the B1 absorber, which they observed at a velocity resolution of $0.35\,\mathrm{km\,s^{-1}}$ (or 1.5 kHz), consists of three components:

- the first is centered on 1301.6493 MHz heliocentric frequency, has an optical depth of $\tau = 0.2462 \pm 0.0010$ and an $FWHM$ velocity width of $\Delta v = 3.68 \pm 0.02$ $\mathrm{km\,s^{-1}}$;
- the second is separated from the first by $\Delta V_{offset} = 7.69 \pm 0.04$ $\mathrm{km\,s^{-1}}$. It has a $\tau = 0.0253 \pm 0.0010$, and an $FWHM$ velocity width of $\Delta v = 2.18 \pm 0.11$ $\mathrm{km\,s^{-1}}$;
- the third is separated from the first by $\Delta V_{offset} = -1.59 \pm 0.66$ $\mathrm{km\,s^{-1}}$. It has a $\tau = 0.0630 \pm 0.0008$, and an $FWHM$ velocity width of $\Delta v = 15.2 \pm 1.4$ $\mathrm{km\,s^{-1}}$.

Though it is not specified in the paper, when examining Fig. 4.13, the peak of the absorber can be estimated at $\sim -480$ mJy. This value will be compared with the WIBAR results in what follows.

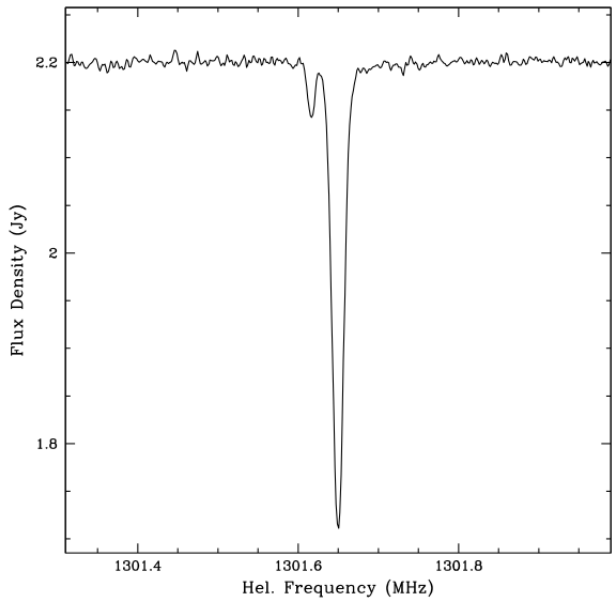


FIGURE 4.13: B0738+313: detection of the B1 first HI absorber at $z = 0.0912$ made at Arecibo during a 45 minutes observation by Lane et al. (2000).

#### 4.1.5.1 Results before clipping

In the $1300.68 - 1301.76$ MHz band analyzed by ROBEL in which lies the B1 absorber, data quality assessment tells us that the frequency band is smooth, and the unclipped temporal noise medians are similar to those in the protected band (see Fig. 4.17 on p.72 to be compared with Fig. 4.4 on p.62). This confirms that the spectrum normalization produced coherent data in this band. There are two RFI signals at frequencies which were flagged by ROBEL as unreliable before clipping (see the yellow lines in the $ON$ spectrum in Fig. 4.18 on p.73):

- the strongest peaks at 1301.29 MHz and its signature is clearly revealed by a $B\tau$ (see 3.2.3 and Eq. 3.17 on p.36) deviating from the spectrum median (Fig. 4.14 on p.70), high kurtosis (see Fig. 4.15 on p.71) and $\frac{\sigma}{MAD}$ (see Fig. 4.16 on p.71), as well as a surge in non robust temporal $\sigma$ (see Fig. 4.4) .
- the second, much weaker, peak centered on 1301.46 MHz is only revealed by excess kurtosis, and no particular RFI pattern appears on the $B\tau$, $\frac{\sigma}{MAD}$, or temporal noise graphs.

These are interesting and not uncommon cases of stealth RFI which are invisible when examining the $ON$ spectrum (Fig. 4.18, even when zooming), and only revealed by data quality estimators (but not by all of them simultaneously). Because B1 lies in a band which is devoid of close-by RFI and the integration was long enough, it is easily detected by ROBEL (even with LS, which is often unreliable) on the unclipped mean $ON$ spectrum which is smooth around B1 (see Fig. 4.18 on p.73, and the zoom on the absorber shown in Fig. 4.19 on p.74). The surge on the right side of the absorber is an RFI recorded in both $ON$ and $OFF$ (only visible on data quality assessment processed with at least 4 times robust binning, because it is weak and spread on several frequency channels: when binned, it is excised by

ROBEL), and is therefore removed in the $(ON - OFF)/OFF$ spectrum (see Fig. 4.20 on p.74. However with such a short exposure time, the graph does not reveal all three components described by Lane et al. (2000). The $ON$ spectrum frequency normalized $rms$ is $\sim$0.0016, (i.e., 41 mJy for $T_{sys} = 35$K and a pointed source efficiency $PSE = 1.4$ K/Jy) and the $OFF$ normalized $rms$ is $\sim$0.0024 ($\sim$62 mJy). By comparison, the $(ON - OFF)/OFF$ frequency $rms$ is $\sim$0.0049 ($\sim$123 mJy), almost 1.65 times higher than the quadratic sum of the two others, and can be explained by positive covariance between $ON$ and $OFF$ frequency channels.

The B1 peak non robust estimator of location is located at 1301.6983 MHz in the $(ON - OFF)/OFF$ spectrum, the line is 453 mJy deep and is detected at $-3.69\sigma$. The peak velocity error $\Delta V_{peak}$ can be assessed using the work of Schneider et al. (1986), Schneider et al. (1990), and Thuan et al. (1999):

$$\Delta V_{peak} = 1.5\,(W_{20} - W_{50})\left(\frac{S}{N}\right)^{-1} \tag{4.1}$$

Where $W_{20}$ is the spectral line width at 20%, $W_{50}$ the spectral line width at 50% (i.e., $FWHM$) and $\frac{S}{N}$ is the absolute value of the peak signal divided by the frequency $rms$.

The errors in $W_{50}$ ($FWHM$) and $W_{20}$ can be estimated as two and three times $\Delta V_{peak}$, respectively.

With $W_{50} = 3.5$ km s$^{-1}$, $W_{20} = 5.5$ km s$^{-1}$ and $rms = 123$ mJy, we estimate $\Delta V_{peak} = 0.8$ km s$^{-1}$. We can thus estimate the uncertainties: $V_{peak} = 27339.4 \pm 0.8$ km s$^{-1}$, $W_{20} = 5.5 \pm 2.4$ km s$^{-1}$, $W_{50} = 3.5 \pm 1.6$ km s$^{-1}$ and $rms = 123$ mJy This yields a $1301.696 - 1301.700$ MHz confidence interval for $V_{peak}$.

At first glance, in the absence of a better assessment method, the peak flux density error can be estimated as equal to the frequency $rms$, i.e.: 123 mJy. Thus, its value is $-453 \pm 62$ mJy.

Now ROBEL offers the opportunity to calculate intervals of confidence related to the location of the spectral line peak frequency, as well as for the calculation of its full width at half maximum ($FWHM$) using time-series of averaged power spectral values. For each channel frequency , I chose the $[EoL - 3EoS, EoL + 3EoS]$ 99.74% confidence interval ($EoL$ being the estimator of location of the temporal distribution of averaged power spectral values, $EoS$ its estimator of scale). The lower boundary frequency is that of the peak in the $EoL - 3EoS$ curve, whereas the upper lies at the peak of the $EoL + 3EoS$ curve. Since I did not find a fully applicable method of $EoS$ bias correction, I kept these values as they were calculated by ROBEL, even if they fall slightly below the minimum theoretical noise (see Eq. 3.17 on p.37). The $EoS$ being biased by underestimation, this means that the confidence intervals I present along this chapter should be enlarged of a few percent if reevaluated with a new process version.

The error on the peak depth is the median of temporal EoS plus 3 times the standard deviation of EoS within the 99.74% frequency confidence interval defined above. This yields a maximum value to which the error on the peak depth has a 99.74% probability to be inferior or equal.

In the $ON$ spectrum, within the 99.74% confidence interval, the peak of the B1 absorber is located at $1301.6967 \pm 0.00144$ MHz, i.e., $27339.8 \pm 0.3$ km s$^{-1}$. Its peak

level lies in the range $-490 \pm 20$ mJy and was detected by ROBEL at $-12\sigma$. We also determined the $FWHM$ of the B1 main component at $3.54 \pm 0.64$ km s$^{-1}$. The other two components detected by Lane et al. (2000) are too weak to be detected with such a short exposure time, although there is a hint of the second component around 1301.665 MHz in Fig. 4.29 on p. 85.

Still within the 99.74% confidence interval and this time on the $(ON - OFF)/OFF$ spectrum, the peak of the B1 absorber is located at $1301.6983 \pm 0.00104$ MHz, i.e., $27339.4 \pm 0.2$ km s$^{-1}$. Its peak level is evaluated at $-453 \pm 14$ mJy, and such a narrow interval is due to a significant frequency *rms* around the peak which is also thin and not rounded (see Fig. 4.20 on p.74). The B1 main component $FWHM$ lies in the range $3.54 \pm 0.3$ km s$^{-1}$.

All these frequencies are given in the local standard of rest (LSR) calculated by ROBEL (see Eq. 3.5.1.2 on p.48). The $\sim 10$ km s$^{-1}$ difference with the 1301.6493 MHz heliocentric frequency of the first component peak measured by Lane et al. (2000) is somewhat larger than the radial velocity correction we adopted for the Sun-LSR motion ($\approx 8.6$ km s$^{-1}$), but can be considered acceptable if one keeps in mind that this velocity has regularly been updated by several km s$^{-1}$ as new radial velocity and parallax measurements of nearby stars became available.

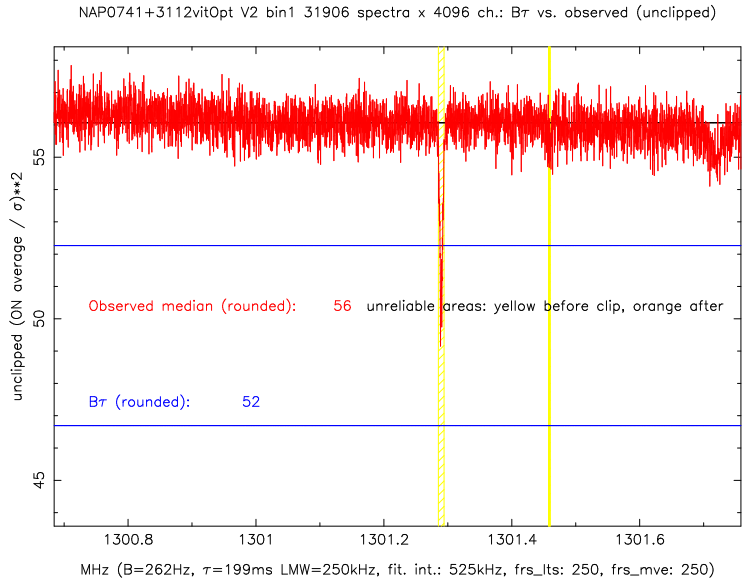


FIGURE 4.14: B0738+313: comparison between squared inverses of normalized theoretical noise ($B\tau$) and unclipped observed temporal noise, (unclipped $ON$ flux density average/$\sigma^2$), within the 1300.68 − 1301.76 MHz Local Standard of Rest (LSR) frequency band. The yellow lines indicate RFI regions flagged as unreliable before clipping, and the dip centered on 1301.7 MHz is related to the B1 absorption line. This graph, and all others showing the 1421 − 1422.2 frequency range, were generated automatically as part of the standard ROBEL package.

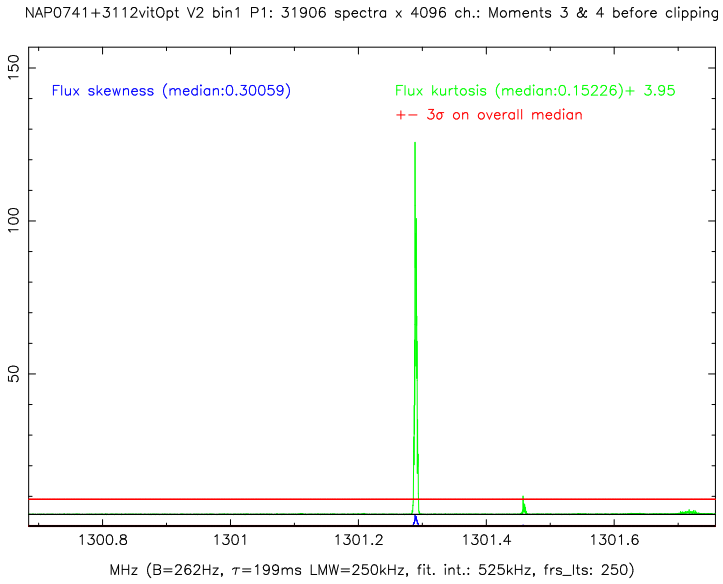


FIGURE 4.15: B0738+313: skewness and kurtosis before clipping within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band.

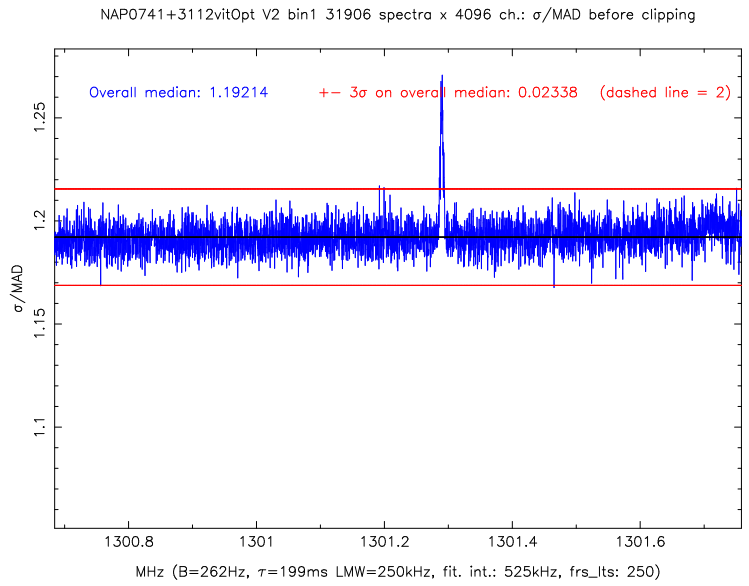


FIGURE 4.16: B0738+313: $\frac{\sigma}{MAD}$ before clipping within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band.

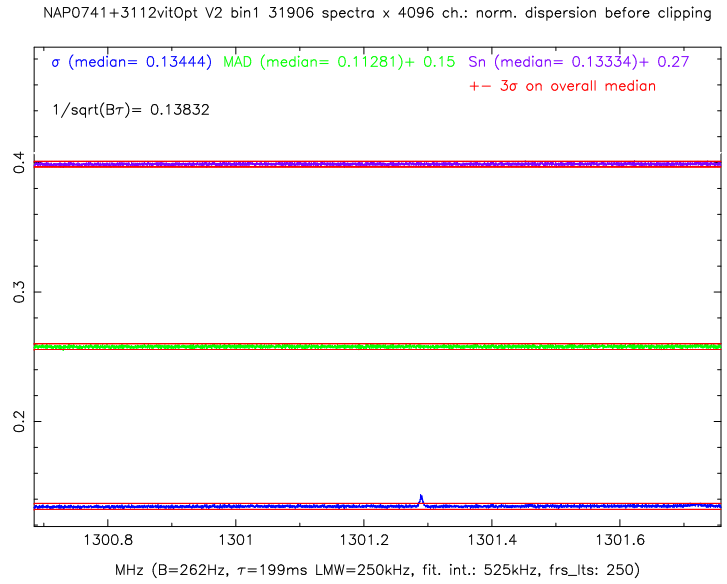


FIGURE 4.17: B0738+313: normalized estimators of scale before clipping within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band.

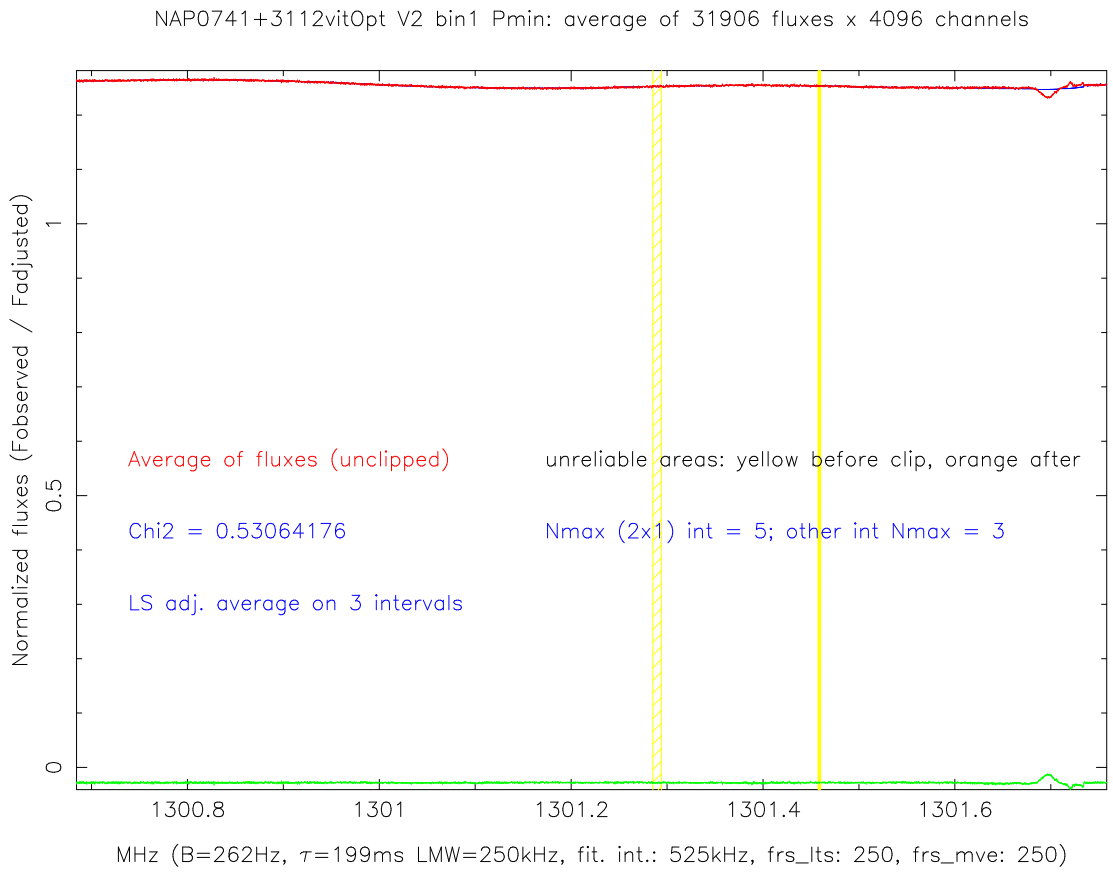


FIGURE 4.18: B0738+313: unclipped mean of uncalibrated averaged power spectral values (in red) within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band of the *ON* spectrum. The B1 absorber lies at 1301.62 MHz. The blue line is the LS fitted polynomial, the green line represents the LS residuals. Observation and processing parameters below the abscissa are the same than those detailed previously. Nmax are the parameters set for the maximum degree of polynomial to be used for LS baseline fitting by ROBEL: 5 for the first and the last two intervals to be fitted, and 3 for the other intervals. Because there are only 3 intervals, the maximum degree for all of them is 5. Chi2 is the sum of LS squared residuals. Spectral channels with white background are flagged by ROBEL as reliable before and after clipping; yellow background underlines channels valid only after clipping, orange background channels unreliable even after clipping.

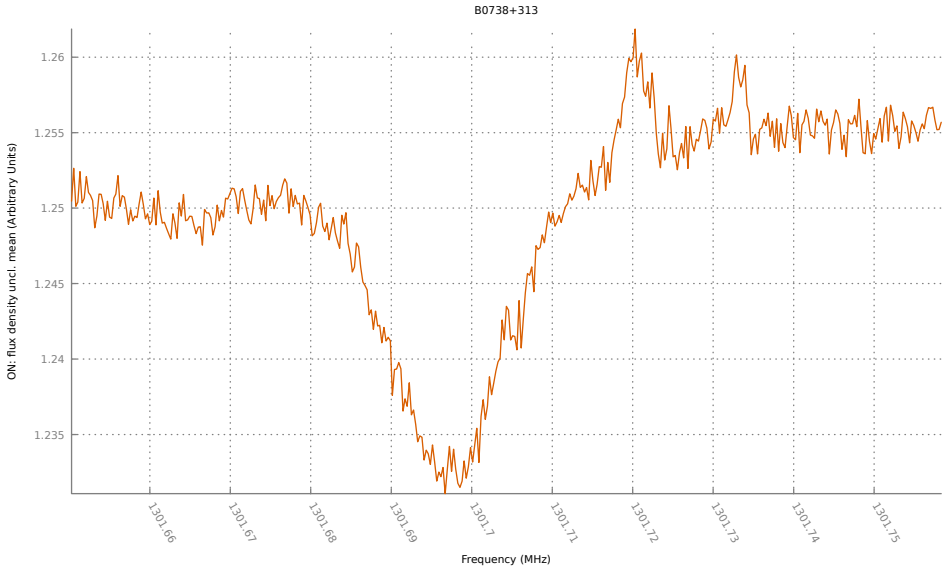


FIGURE 4.19: B0738+313: unclipped mean of uncalibrated averaged power spectral values of the $ON$ spectrum around the first absorber B1 (Local Standard of Rest -LSR- frequencies).This graph is a zoom of the red line in Fig. 4.18

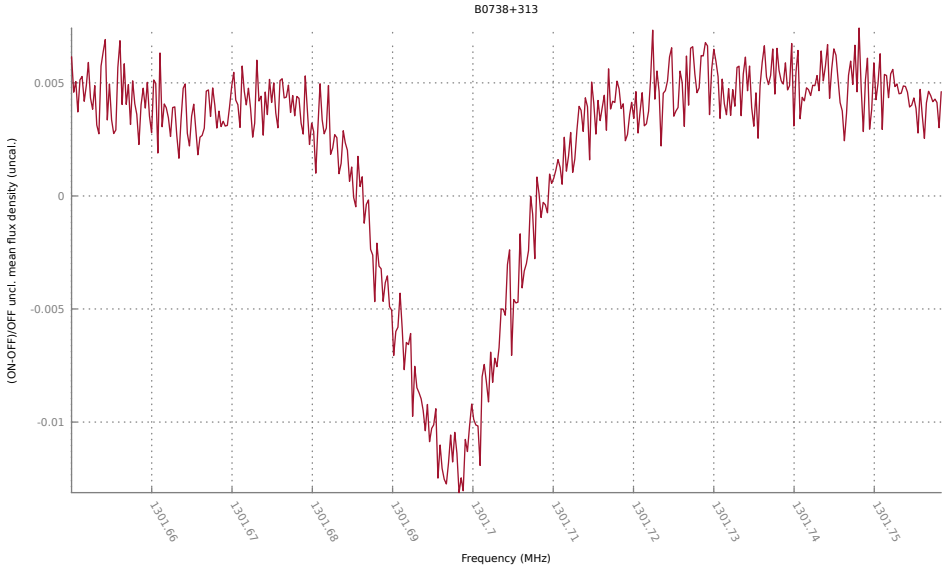


FIGURE 4.20: B0738+313: $(ON - OFF)/OFF$ spectrum (Local Standard of Rest -LSR- frequencies) around the first absorber B1 of unclipped mean of uncalibrated averaged power spectral values.

#### 4.1.5.2 Results after clipping

After having calculated the $Sn$-clipped median of uncalibrated averaged power spectral values, we find the following results:

- since the 1300.68 - 1301.75 is relatively quiet (except for the two RFI mentionned above), the percentage of excised averaged power spectral values in each frequency channel is low: Fig. 4.21 on p.79 shows that $Sn$-clipping has eliminated a median of 0.48% of outliers. The channels where the RFI lie have not been clipped more than the others. This indicates that these outliers are made of a few strong pulses;
- skewness (Fig. 4.22 on p.79), kurtosis (Fig. 4.23 on p.80) and $\frac{\sigma}{MAD}$ (Fig. 4.24 on p.80) are now smooth along the frequency band and prove that the two RFI have been eliminated;
- the median of skewness after $Sn$-clipping is still $\sim$0.19. This means that not all of the temporal drift has been eliminated, though it is homogeneous along the frequency band;
- the kurtosis median is around $\sim -0.19$. This indicates that the samples of averaged power spectral values for each frequency channel are narrower than a normal distribution. This is not a surprise, since clipping tends to trim the sample wings;
- the median of $\frac{\sigma}{MAD}$ is $\sim$1.18 and has not improved, so the profile does not change, except for the RFI mitigation;
- Fig. 4.25 on p.81 shows that temporal estimators of scales have barely changed after clipping.

The $Sn$-clipped $ON$ spectrum frequency normalized $rms$ is $\sim$0.0018, (i.e., 45 mJy for $T_{sys} = 35$K and a pointed source efficiency $PSE = 1.4$ K/Jy) and the $OFF$ normalized $rms$ is $\sim$0.0025 ($\sim$63 mJy). By comparison, the $Sn$-clipped $(ON - OFF)/OFF$ frequency $rms$ is $\sim$0.005 ($\sim$125 mJy), around 1.62 times higher than the quadratic sum of the two others. This also can be explained by positive covariance between $ON$ and $OFF$ frequency channels.

The B1 peak $Sn$-clipped median is located at 1301.6967 MHz in the $(ON - OFF)/OFF$ spectrum, its minimum flux density level at $-460$ mJy and is detected at the $-3.67\sigma$ significance level. Again using the work of Schneider et al. (1986), Schneider et al. (1990), and Thuan et al. (1999), from Eq. 4.1 on p.69, the peak velocity error is estimated at $\Delta V_{peak} = 0.8$ km s$^{-1}$, with $W_{20} = 5.53$ km s$^{-1}$, $W_{50} = 3.65$ km s$^{-1}$ and the frequency $rms = 125$ mJy. Eq. 4.1 on p.69 gives us the error $\Delta W_{50} = 1.54$ km s$^{-1}$ on $W_{50}$ ($FWHM$). Also, if not using temporal estimators of scale, I consider the error on the peak flux density as equal to the frequency $rms$, i.e., 125 mJy. Thus, it is included in the $366 - 491$ mJy interval.

Using the temporal distribution of averaged power spectral values as explained in 4.1.5.1 on p.68, we find for the B1 main component absorber:

- in the $ON$ spectrum (see Fig. 4.27 on p.83):
  - peak frequency $1301.6967 \pm 0.0008$ MHz, i.e., at $V_{LSR} = 27339.76 \pm 0.17$ km s$^{-1}$,
  - peak level $-496 \pm 20$ mJy, detected at $-11\sigma$, and an LTS residual of $\sim -12$ as shown in Fig. 4.26 on p.82 and Fig. 4.28 on p.84,
  - $FWHM$ $3.04 \pm 0.42$ km s$^{-1}$;
- in the $(ON - OFF)/OFF$ spectrum (see Fig. 4.29 on p.85):

- peak frequency $1301.6967 \pm 0.0004$ MHz, i.e., at $V_{LSR}$ = $27339.76 \pm 0.01$ km s$^{-1}$,
- peak level $-460 \pm 14$ mJy detected at $-3.67\sigma$ (see Fig. 4.29 on p.85),
- $FWHM$ $3.65 \pm 0.19$ km s$^{-1}$.

The optical depth $\tau_{21}(v)$ at radial velocity $v$ is given by:

$$e^{-\tau_{21}(v)} = \frac{I(v)}{I_0} \tag{4.2}$$

where $I_0$ is the average flux density of the continuum and $I(v)$ is the flux density at radial velocity $v$.

Thus, we derive the spectral line peak optical depth $\tau_{peak}$ from the following equation:

$$e^{-\tau_{peak}} = \frac{I_{peak}}{I_0} \tag{4.3}$$

where $I_{peak}$ is the spectral line peak flux density.

The equivalent width $EW_{21}$ is given by:

$$EW_{21} = \int \tau_{21}(v) \, \frac{dv}{\mathrm{km\,s^{-1}}} \quad . \tag{4.4}$$

It appears that Lane et al. (2000) used the simplified formula, which is only valid for $\tau_{21}(v) << 1$:

$$EW_{21} \;=\; 1.06 \; \tau_{peak} \left( \frac{\mathrm{FWHM}}{\mathrm{km\,s^{-1}}} \right) \quad , \tag{4.5}$$

And the column density $N(\mathrm{HI})$ is given by:

$$N_{\mathrm{HI}} \mathrm{atoms\,cm^{-2}}] \;=\; 1.823 \times 10^{20} \left( \frac{< T_s >}{100\,\mathrm{K}} \right) \, EW_{21} \quad . \tag{4.6}$$

Table 4.1 on p. 78 compares the results of Lane et al. (2000) for the first, dominant component of the B1 absorbers to those we derived from the NRT observations for the main component as a whole. While Lane et al. (2000) adopted a 297 K spin temperature, equal to the kinematic temperature they measured, I adopted the generally used 100 K for cold HI since the various optical depths calculated from my observation will be consistent with the Arecibo results. Note that a much longer exposure time would be required to compare the three line components in detail. I do not include B1 peak line frequency EoL in this table because of discrepancies explained in 4.1.5.1 on p.68. Comparing peak velocity errors is relevant, though. The main conclusions to draw are the following:

- since the $ON$ spectrum frequency $rms$ is lower than the one of the $(ON - OFF)/OFF$ spectrum (and in all cases studied here, the quadratic sum of the frequency $rms$ of $ON$ and $OFF$ is also lower than the $(ON - OFF)/OFF$ $rms$

for matters of covariance), automated blind detections of narrow absorbers are better performed using $ON$ spectra only; the B1 second component which is centered $\sim$33 kHz lower, while hidden in the noise of the $(ON-OFF)/OFF$ spectrum, its presence can be guessed in the $ON$ spectrum (though not above $3\sigma$, see Fig. 4.27 on p.83)

- baseline fitting is more approximate in $ON$ spectra than in $(ON-OFF)/OFF$ spectra, so optical depths derived from $ON$ spectra are more imprecise;
- using time-series of averaged power spectral values results in smaller error margins. Whether these are more accurate is a matter of debate since there is no obvious way of assessing them. But direct access to temporal noise for each frequency channel provides more accurate information on data profiles than inferring error margins from a few parameters such as $W_{20}$ and $W_{50}$;
- The NRT observed column density which is closest to that of Lane et al. (2000) is first, the one derived from the $(ON-OFF)/OFF$ $Sn$-clipped spectrum of median averaged power spectral values, and second, the $ON$ $Sn$-clipped median spectrum.

| **B0738+313 B1 absorber (z=0.0912)** | **Spin Temperature** | **Peak optical depth** | **EW21= 1.06tau** | **FWHM (km.s$^{-1}$)** | **N(HI) ($10^{20}$ cm$^{-2}$)** |
|---|---|---|---|---|---|
| **1. Lane et al.** | 297 | 0.2462± 0.0010 | 0.2610± 0.0010 | 3.69± 0.02 | 1.754097± 0.00003 |
| **2. NRT observation :** | | | | | |
| 2.1. Before clipping (non robust) | | | | | |
| 2.1.1.ON | | | | | |
| *ROBEL* | 100 | 0.2520± 0.0092 | 0.2671± 0.0098 | 3.54± 0.64 | 1.723838± 0.011373 |
| 2.1.2. (ON-OFF)/OFF | | | | | |
| *2.1.2.1. 1D spectrum only :* | 100 | 0.2306± 0.0288 | 0.2444± 0.0305 | 3.50± 1.57 | 1.559624± 0.087294 |
| *2.1.2.2. ROBEL :* | | 0.2306± 0.0288 | 0.2444± 0.0305 | 3.54± 0.30 | 1.577448± 0.016681 |
| 2.2. After Sn-clipping | | | | | |
| 2.2.1.ON | | | | | |
| *ROBEL* | 100 | 0.2555± 0.0092 | 0.2708± 0.0098 | 3.04± 0.42 | 1.500918± 0.007504 |
| 2.2.2. (ON-OFF)/OFF | | | | | |
| *2.2.2.1. 1D spectrum only :* | 100 | 0.2346± 0.0292 | 0.2487± 0.0310 | 3.65± 1.54 | 1.654678± 0.087324 |
| *2.2.2.2. ROBEL :* | 100 | 0.2346± 0.0064 | 0.2487± 0.0068 | 3.65± 0.19 | 1.654678± 0.002350 |

TABLE 4.1: B0738+313: comparison of results for the B1 absorber profile. The peak depth for Lane et al. (2000) is estimated from Fig. 4.13 on p.68. The peak depth as well as $FWHM$ values and errors are calculated either from 1D spectra only or from 2D spectra processed by ROBEL according to the methods described in 4.1.5.1 on p.68. The line peak optical depths are calculated using Eq. 4.3 on p.76, the column density $N$(HI) using Eq. 4.6 on p.76 with corresponding HI spin temperatures.

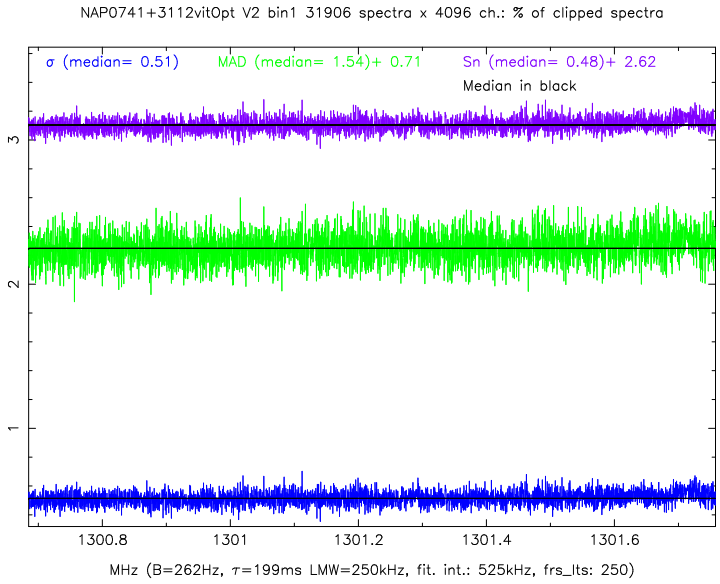


FIGURE 4.21: B0738+313: percentage of averaged power spectral values $\sigma$-clipped (blue line), $MAD$-clipped (green line), and $Sn$-clipped (purple line) within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band.

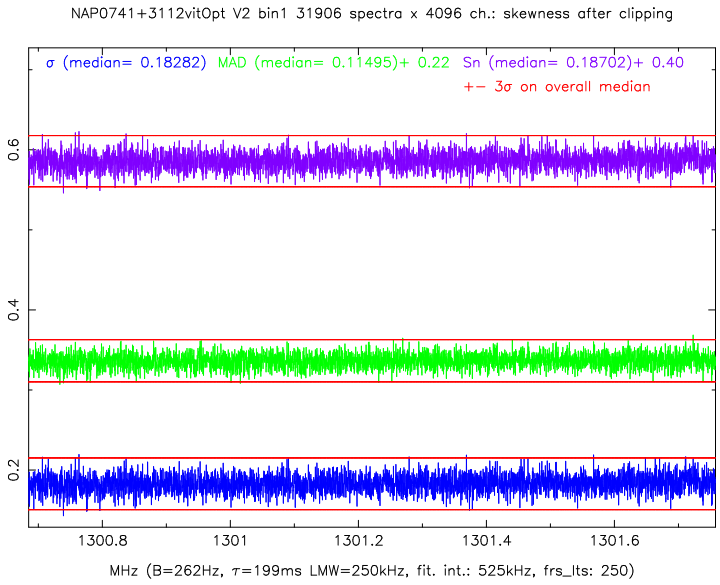


FIGURE 4.22: B0738+313: skewness after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band.

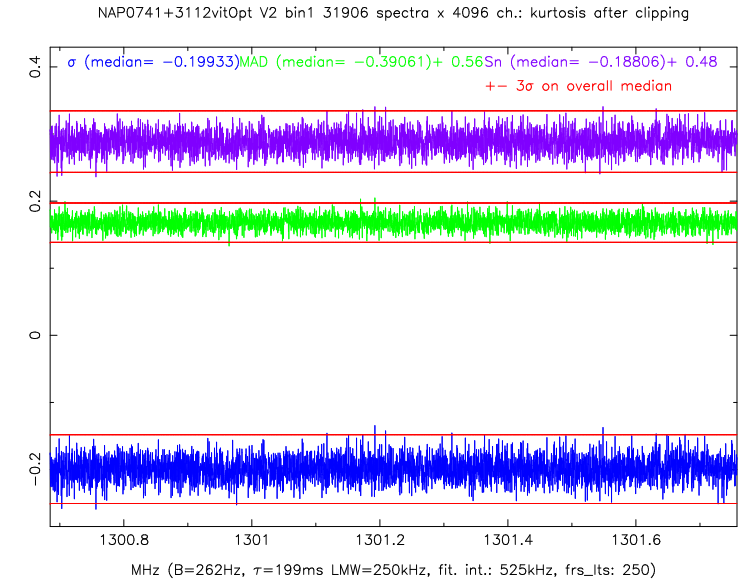


FIGURE 4.23: B0738+313: kurtosis after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1300.68 - 1301.75 MHz band.

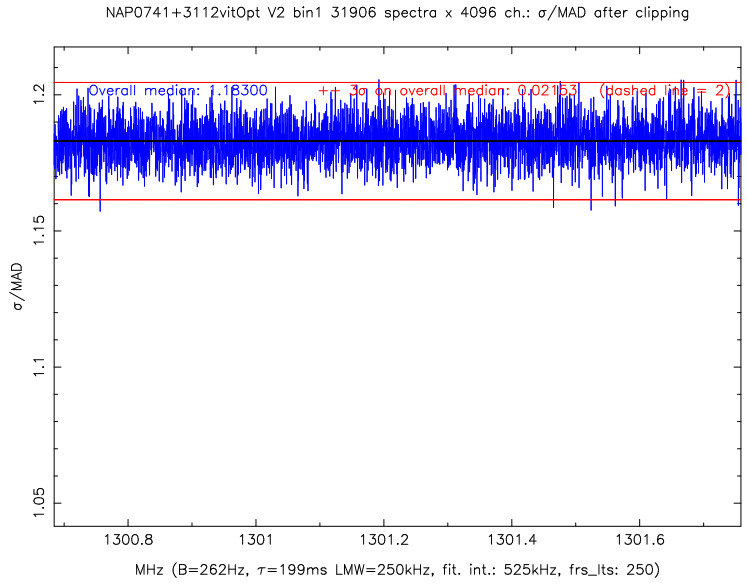


FIGURE 4.24: B0738+313: $\frac{\sigma}{MAD}$ after clipping within the 1300.68 - 1301.75 MHz band.

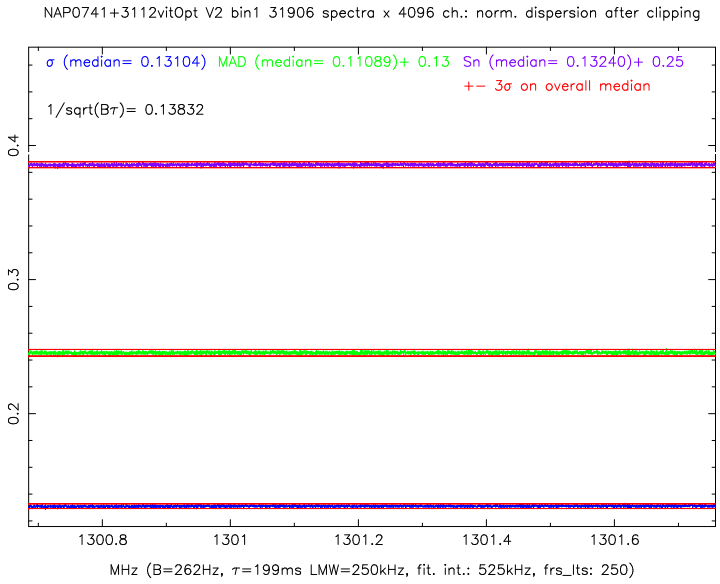


FIGURE 4.25: B0738+313: normalized estimators of scale after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1300.68 - 1301.75 MHz band.

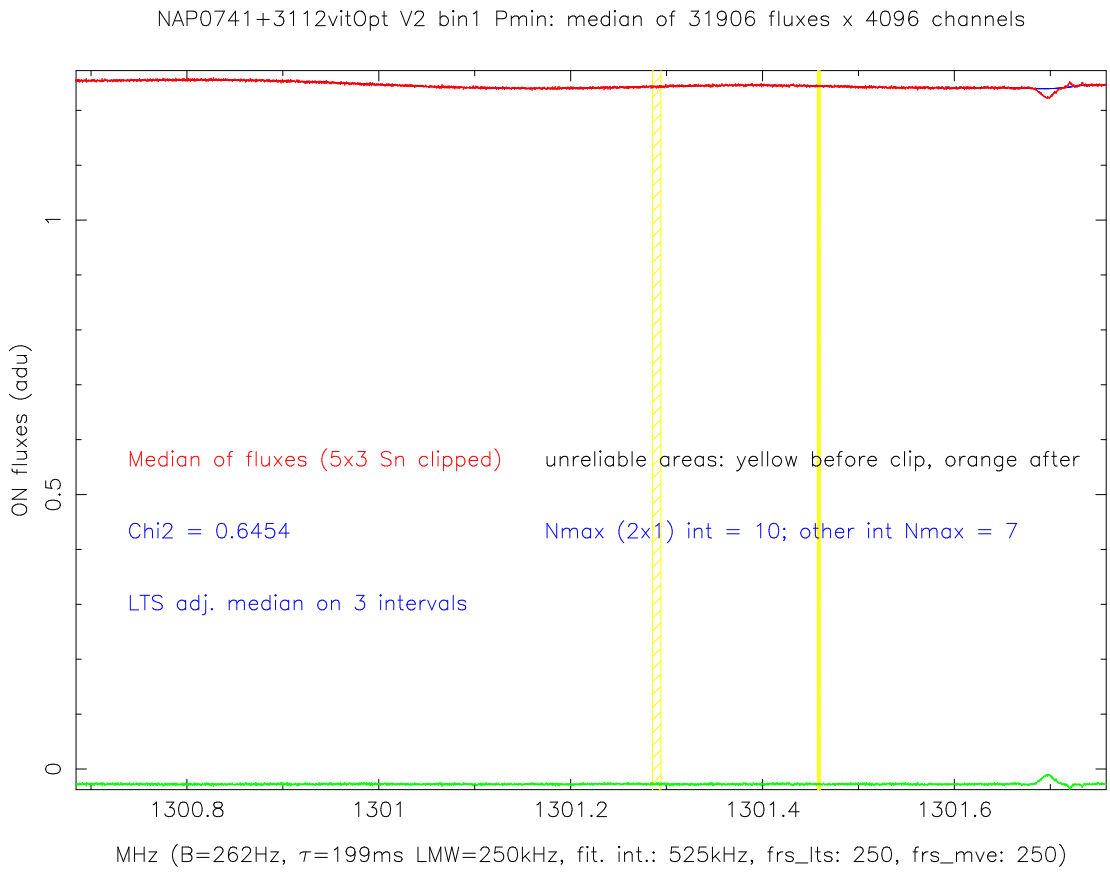


FIGURE 4.26: B0738+313: *Sn*-clipped median of uncalibrated averaged power spectral values (in red) within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band of the *ON* spectrum. The blue line is the LTS fitted polynomial, the green line represents the LTS residuals. Observation and processing parameters below the abscissa are the same than those detailed previously. Nmax are the parameters set for the maximum degree of polynomial to be used for LTS baseline fitting by ROBEL: 10 for the first and the last two intervals to be fitted, and 7 for the other intervals. Because there are only 3 intervals, the maximum degree for all of them is 10. Chi2 is the sum of LTS squared residuals. Spectral channels with white background are flagged by ROBEL as reliable before and after clipping; yellow background underlines channels valid only after clipping, orange background channels unreliable even after clipping.

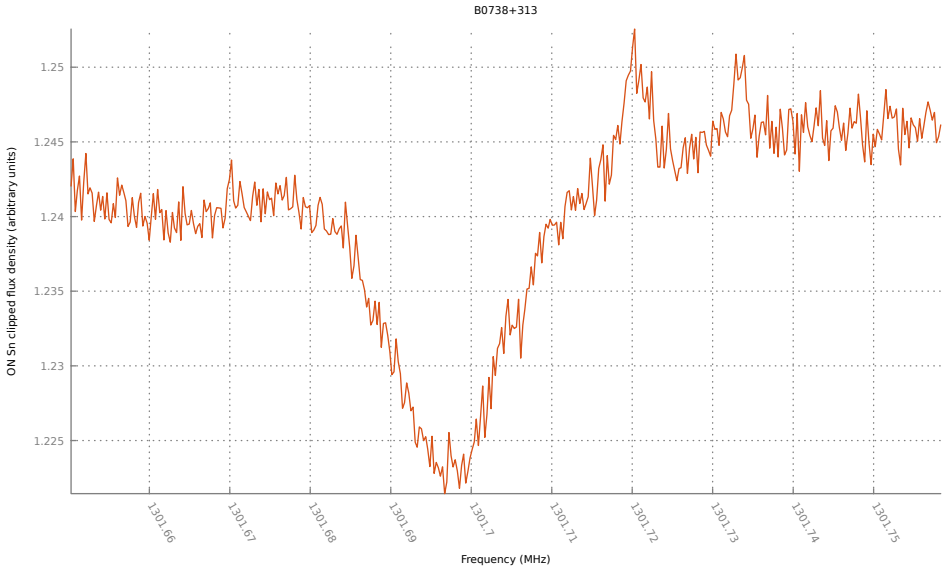


FIGURE 4.27: B0738+313: *Sn*-clipped median of uncalibrated averaged power spectral values of the *ON* spectrum around the first absorber B1.This graph is a zoom of the red line in Fig. 4.26.

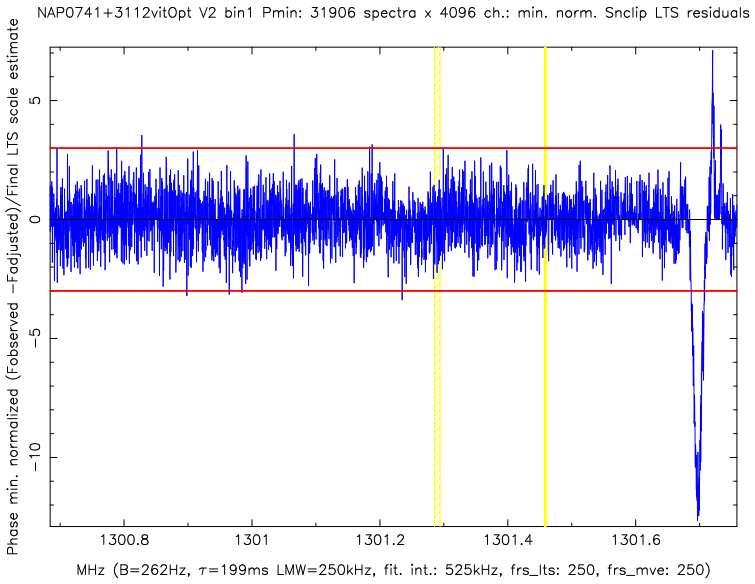


FIGURE 4.28: B0738+313: LTS residuals (blue line) of $Sn$-clipped median of uncalibrated averaged power spectral values of the $ON$ spectrum within the 1300.68 - 1301.75 MHz Local Standard of Rest (LSR) frequency band. The two red lines mark the $\pm 3$ limit for automated detection. Spectral channels with white background are flagged by ROBEL as reliable before and after clipping; yellow background underlines channels valid only after clipping, orange background channels unreliable even after clipping. The B1 absorber which lies around 1301.7 MHz has got a maximum LTS residual of $\sim -14$. Any valid channel is flagged by ROBEL as a spectral line candidate if its LTS residual exceeds $\pm 3$. The positive residuals superior to 3 nearby the B1 absorber are false positive due to artefacts equally present on both $ON$ and $OFF$ spectra. Hence they do not appear on the $(ON - OFF)/OFF$ spectrum in Fig. 4.29.

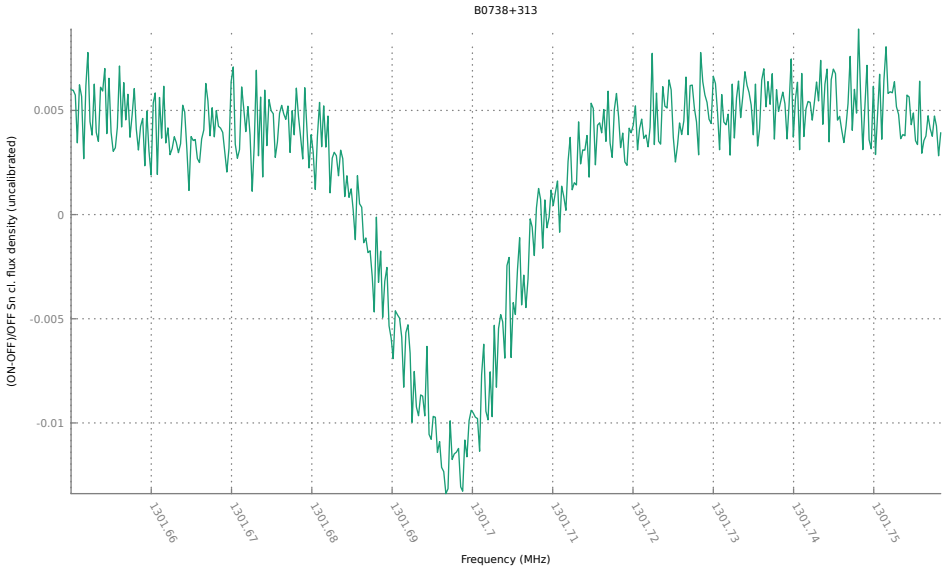


FIGURE 4.29: B0738+313: $(ON - OFF)/OFF$ spectrum around the first absorber B1 of $Sn$-clipped median of uncalibrated averaged power spectral values.

### 4.1.6 results for the second absorber at $z = 0.2212$

At Arecibo, Kanekar et al. (2001) obtained an extremely sensitive spectrum in 9 hours on-source integration at a velocity resolution of 0.4 $\mathrm{km\,s^{-1}}$ and an extremely high velocity resolution (0.025 $\mathrm{km\,s^{-1}}$) spectrum in 35 minutes on-source (see Fig. 4.30 on p.86). From these, they fitted parameters for three distinct line components:

- the first profile has a $0.016 \pm 0.007$ optical depth, a $3.85 \pm 0.30$ $\mathrm{km\,s^{-1}}$ $FWHM$ and a heliocentric redshift $z = 0.221250 \pm 10^{-6}$ (i.e., 1163.075 MHz heliocentric frequency);
- the second has a $0.080 \pm 0.002$ optical depth, a $1.75 \pm 0.69$ $\mathrm{km\,s^{-1}}$ $FWHM$ and a redshift $z = 0.221240 \pm 10^{-6}$ (i.e., 1163.0849 MHz LSR frequency);
- the third has a $0.0046 \pm 0.0005$ optical depth, a $18.55 \pm 1.11$ $\mathrm{km\,s^{-1}}$ $FWHM$ and a redshift $z = 0.221239 \pm 10^{-6}$ (i.e., 1163.0858 MHz LSR frequency).

It should be noted that the collecing area of the Arecibo mirror is almost ten times larger than the NRT one. Therefore, our 1.7 hours of NRT integration time represent only a tiny fraction of the 9 hours long observation by Kanekar et al. (2001).

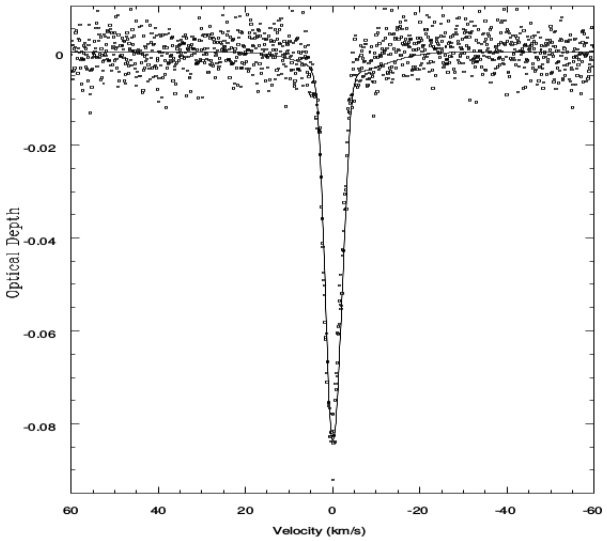


FIGURE 4.30: B0738+313 B2 absorber high-resolution ($\sim$0.1 km s$^{-1}$) profile observed at Arecibo by Kanekar et al. (2001).

At the NRT, the B2 center frequency of 1163 MHz lies right in the middle of RFI at 1162.8 − 1163.7 MHz (see the yellow areas in Fig. 4.31 on p.87) strong enough to hide the spectral line (see Fig. 4.37 on p.90). There is no clear consensus about the nature of this RFI. Debates with colleagues tend to favor a Radionavigation (RN) signal, though the presence of GNSS unwanted emissions cannot be excluded. The French spectrum regulatory agency (ANFR) stipulates that the 960 − 1215 MHz frequency band is allocated to the Aeronautical Radionavigation service (RN) for airborne operations on a shared primary basis[3], whereas the GNSS co-primary allocation starts at 1164 MHz (i.e., well outside the frequency range plotted in Figs. 4.31 - 4.34). This means that two different kinds of RFI can occur in the 1164 − 1215 MHz band, both from transmitters that move across the sky. Indeed, there is a good chance to find unwanted out-of-band emissions of either GPS L5I band (See Appendix A.3.1 on p.137 and Fig. A.3 p.139) or Galileo E5a band (Appendix A.3.2 on p.140 and Fig. A.5 p.141). The L5-E5 overlap is also illustrated by Fig. A.6 on p.141. Data quality assessment by ROBEL underlines the unmistakable signatures of these strong RFI.

#### 4.1.6.1 Results before clipping

Therefore, compared to B1, the RFI landscape is very different. The B2 center frequency of 1163 MHz lies at the edge of the GPS L5 and Galileo E5 frequency band. Their signal is clearly visible on Fig. 4.31 on p.87 as a noise surge, as strong increase of skewness, and above all kurtosis (Fig. 4.32 p.88), as well as a similar pseudo-periodic increase of $\frac{\sigma}{MAD}$ before clipping (Fig. 4.33 p.88). Note that the semi-periodic noise coming from secondary lobes of L5-E5 are clearly visible at a miminum frequency of $\sim$1162.6 MHz on Fig. 4.31, well below 1164 MHz.

B2 being swamped by this RFI, could not be detected with non-robust processing as Fig. 4.36 on p.90 of the $ON$ spectrum illustrates. The $(ON - OFF)/OFF$ spectrum Fig. 4.37 on p.90 is not even linear and a baseline fitting would be very difficult to

[3] https://www.anfr.fr/gestion-des-frequences-sites/tnrbf/

process. The $\sigma$ temporal *EoS* Fig. 4.34 on p.89 automatically generated by ROBEL in the 1162.11 − 1163.18 MHz band is dominated by the noise scale of GPS L5 and Galileo E5. A zoom centered on B2 reveals that the RFI has a semi-periodic noise profile which sticks to the one of the *ON* spectrum (see Fig. 4.35 p.89).

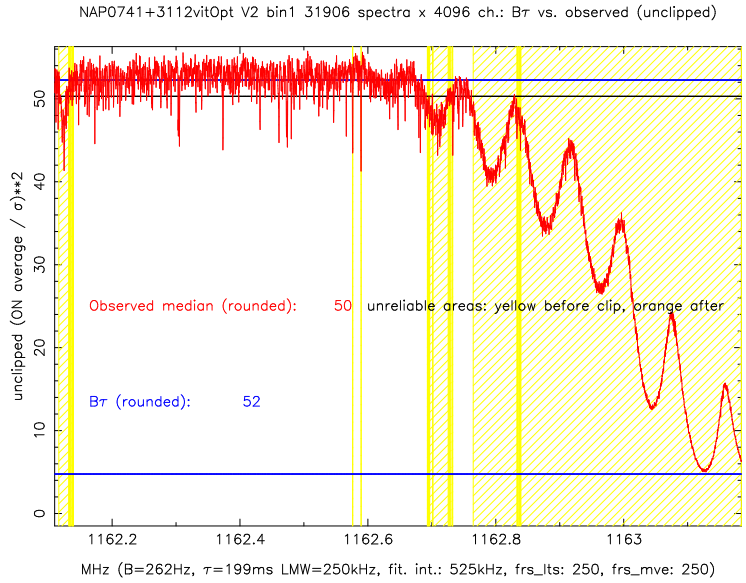


FIGURE 4.31: B0738+313: comparison between squared inverses of normalized theoretical noise ($B\tau$) and unclipped observed temporal noise, (unclipped *ON* flux density average/$\sigma^2$), within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band. The yellow lines indicate RFI regions flagged as unreliable before clipping. This graph, and all others showing the 1162.11 - 1163.18 MHz frequency range, were generated automatically as part of the standard ROBEL package.

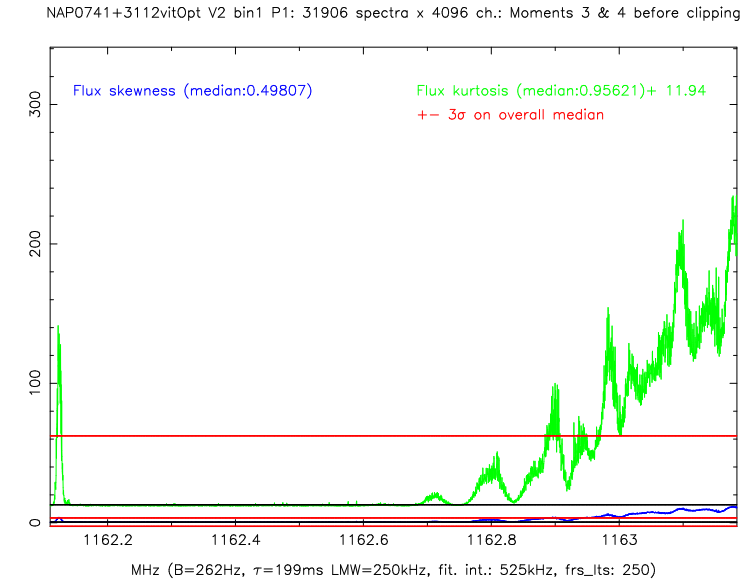


FIGURE 4.32: B0738+313: skewness and kurtosis before clipping within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

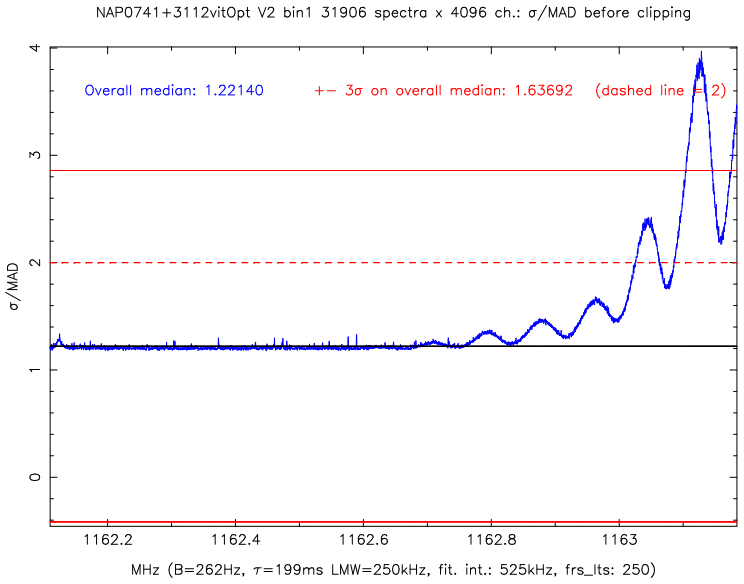


FIGURE 4.33: B0738+313: $\frac{\sigma}{MAD}$ before clipping within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

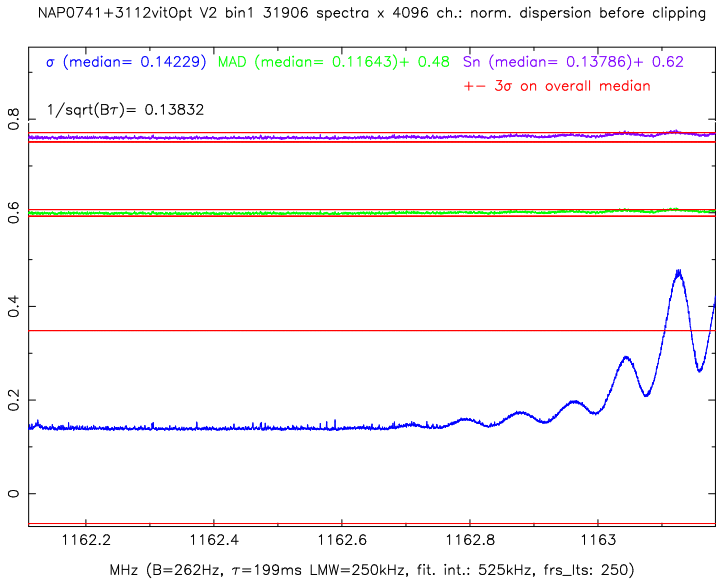


FIGURE 4.34: B0738+313: normalized estimators of scale before clipping within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

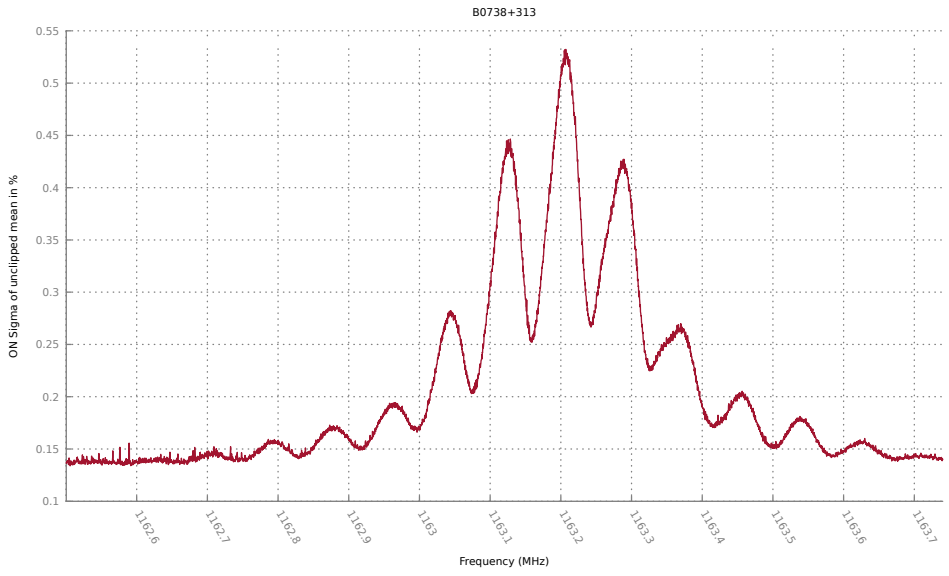


FIGURE 4.35: B0738+313: normalized temporal $\sigma$ of uncalibrated averaged power spectral values for the *ON* spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber. This portion is a zoom of the blue line of Fig. 4.34.

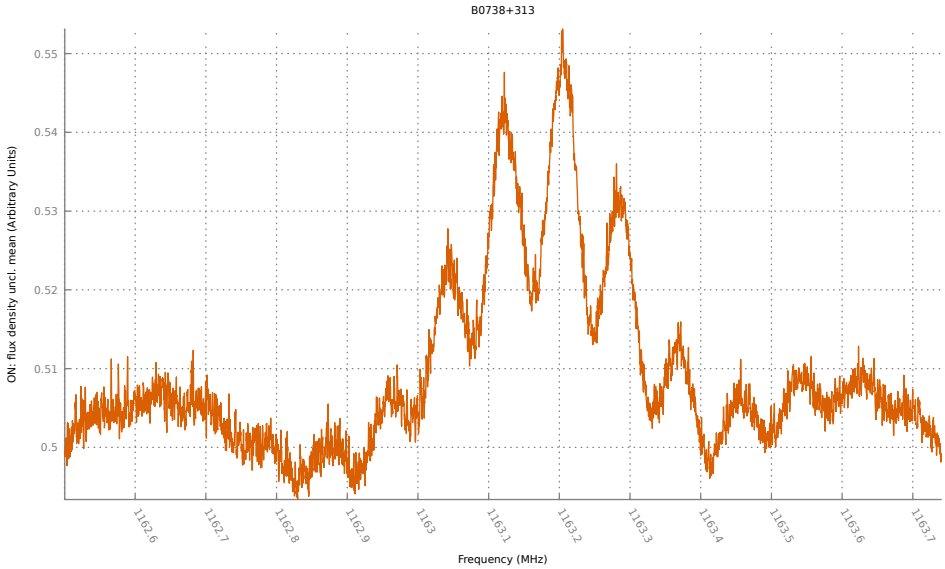


FIGURE 4.36: B0738+313: unclipped mean of uncalibrated averaged power spectral values for the $ON$ spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber.

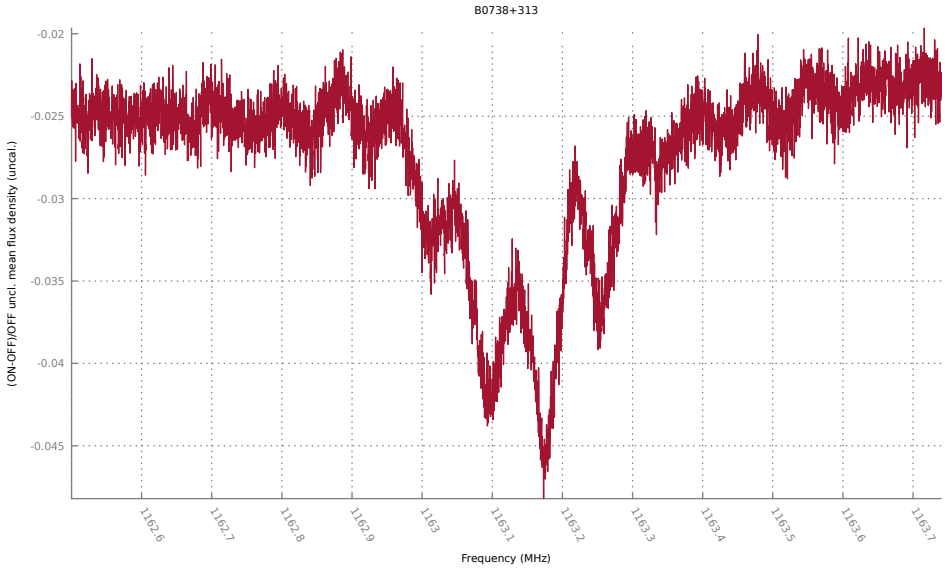


FIGURE 4.37: B0738+313: $(ON - OFF)/OFF$ spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber of unclipped mean of uncalibrated averaged power spectral values.

#### 4.1.6.2 Results after clipping

After having calculated $Sn$-clipped median of uncalibrated averaged power spectral values, we get the following results about the $1162.110 - 1163.184$ MHz band automatically processed by ROBEL:

- clipping is strong along the GNSS band (up to ∼4.5% averaged power spectral values excised compared to a 0.78% median), with a surge related to the B2 close-by RFI;
- a 0.26 skewness median which is still twice the median value of the normalization band but almost one-half of the unclipped spectrum skewness. Nevertheless the semi-periodic variations due to L5-E5 are notably smoothed;
- the kurtosis median is now $-0.17$ and reflects this strong clipping. Values are smoothed along the spectrum with residual slight variations in the L5-E5 band;
- the $\frac{\sigma}{MAD}$ ratio also follows the same trend;
- the normalized estimators of scale lie around their median (i.e., 0.136), slightly below the theoretical temporal noise (0.138) for reasons explained earlier in 4.1.4.2. A zoom on the $Sn$ temporal EoS of the $Sn$-clipped spectrum (Fig. 4.43 on p.97 highlights a local surge due to the RFI residuals.

The $Sn$-clipped $ON$ spectrum frequency normalized $rms$ is ∼0.0408, (i.e., 121 mJy for $T_{sys} = 35$K and a pointed source efficiency $PSE = 1.4$ K/Jy) and the $OFF$ normalized $rms$ is ∼0.0416 (∼104 mJy). By comparison, the $Sn$-clipped $(ON - OFF)/OFF$ frequency $rms$ is ∼0.016 (∼28.4 mJy), only 0.27 times the quadratic sum of the two others, a surprisingly low value. This also can be explained by negative covariance between $ON$ and $OFF$ frequency channels.

The first B2 component was easily detected in $Sn$-clipped spectra with a $-4$ LTS residual (see Fig. 4.47 on p.99. However, with such a low integration time compared to Kanekar et al. (2001) at Arecibo, frequency $rms$ is high enough for the following results to be considered with caution, as illustrated in Fig. 4.49 on p.100. Due to high frequency $rms$, the error in $Vpeak$ is irrelevant: variations of the estimators of location alone are too rapid; thus, in practice, $W_{50}$ ($FWHM$) and $W_{20}$ are quite difficult to assess. With this in mind, using temporal series of averaged power spectral values, we get:

- in the $ON$ spectrum (see Fig. 4.45 on p.98):
  - the peak located at $1163.1288 \pm 0.0014$ MHz, i.e., 245491 $\mathrm{km\,s^{-1}}$ with no relevant error margin,
  - the peak level in the range $-166 \pm 21$ mJy and detected by ROBEL at $1.63\sigma$ significance level, and a LTS residual of $\sim -4$ as shown in Fig. 4.47 on p.99;
  - the $FWHM$ around $0.83 \pm 0.69$ $\mathrm{km\,s^{-1}}$;
- on the $(ON - OFF)/OFF$ spectrum (see Fig. 4.48 on p.99):
  - the peak in the range of $1163.1345 \pm 0.0039$ MHz, i.e., $245494 \pm 1$ $\mathrm{km\,s^{-1}}$,
  - the peak level around $128 \pm 15$ mJy and detected by ROBEL at $3.22\sigma$ (see Fig. 4.48 on p.99),
  - the $FWHM$ around $3.04 \pm 0.19$ $\mathrm{km\,s^{-1}}$.

It is interesting to evaluate the impact of $Sn$-clipping by assessing the attenuation factors between the unclipped non-robust spectrum and the $Sn$-clipped one. These are only relevant either on $ON$ or $OFF$ spectra, since RFI must be excised separately from $ON$ and $OFF$ spectra before calculating the resulting $(ON - OFF)/OFF$.

The attenuation factor of the temporal estimator of location of averaged power spectral values $attn_{loc}(i)$ for each frequency channel $i$ is defined as:

$$attn_{loc}(i) = (Snclipped\ median(i)\ -\ unclipped\ mean(i))/(Snclipped\ median(i)) \tag{4.7}$$

The attenuation factor of the temporal estimator of scale of averaged power spectral values $attn_{scale}(i)$ for each frequency channel $i$ is given by:

$$attn_{scale}(i) = (Sn\ of\ Snclipped\ median(i)\ -\ \sigma\ of\ unclipped\ median(i))/(Sn\ of\ Snclipped\ median(i)) \tag{4.8}$$

Fig. 4.46 on p.98 illustrates the smoothing effect of $Sn$-clipping on the temporal EoL (according to Eq. 4.7) which then allows the detection of B2 first component on the $Sn$-clipped spectrum shown on Fig. 4.45. Eq. 4.8 reveals how efficient the $Sn$-clipping is to remove noise generated by RFI. Fig. 4.43 on p.97 shows that the attenuation factor of temporal noise exceeds 3 around 1163.2 MHz in the $ON$ spectrum. We will see that in the case of III Zw 35, these indicators go far beyond in the presence of strong RFI.

| B0738+313 B2 absorber (z=0.2212) | Error on Vpeak (km/s) | Peak depth (mJy) | Peak detection | Spin Temperature (K) | Peak optical depth | FWHM (km/s) |
|---|---|---|---|---|---|---|
| **1. Kanekar et al.** | ± 0.3 | | | 890 ± 160 | | |
| low resolution | | | | | 0.076 ± 0.002 | 3.76 ± 0.12 |
| high resolution | | | | | 0.080 ± 0.002 | 3.85 ± 0.30 |
| **2. NRT observation after Sn-clipping:** | | | | 244 | | |
| 2.1.ON | less than 0.01 | -166 ± 21 | 1.63σ | | 0.087 ± 0.010 | 0.83 ± 0.69 |
| 2.2. (ON-OFF)/OFF | ± 0.83 | -128 ± 15 | 3.22σ | | 0.066 ± 0.008 | 3.04 ± 0.33 |

TABLE 4.2: B0738+313: comparison of results for the profile of the first component of the B2 absorber. Vpeak is the absorber peak velocity, the peak detection is given in units of frequency *rms*, the peak depth as well as $FWHM$ values and errors are calculated from 2D spectra processed by ROBEL according to the methods described in 4.1.5.1 on p.68. NRT optical depths are calculated with the average 244 K spin temperature of the first two components as measured at Arecibo by Kanekar et al. (2001). From the NRT observation, only results for $Sn$-clipped spectra are presented, since the absorber is not detected with unclipped averaged power spectral values.

The third component signature on the spectrum is weak enough to be almost unnoticeable in Fig. 4.30 on p.86, as Kanekar et al. (2001) state. With such a low integration time at NRT, there was no way of getting a profile precise enough to observe the three B2 components and to discuss multiphase IGM with different spin temperatures. Therefore, it seemed logical to assess our results with the average of spin temperatures from the first and second components, i.e., 244 K. From Table 4.2 on p.93, we notice that the $Sn$-clipped $(ON - OFF)/OFF$ spectrum yields the closest $FWHM$ whereas the $ON$ spectrum provides the closest optical depth from Kanekar et al. (2001) values.

And finally the B2 NRT observation illustrated in Fig. 4.49 on p.100 confirms the asymmetry of the spectral line which is underlined by Kanekar et al. (2001).

In conclusion, even in difficult conditions (low integration time, RFI lying right in the middle of the band), and in a frequency band which is seldom explored at NRT since it is polluted by numerous unwanted signals, the $(ON - OFF)/OFF$ $Sn$-clipped spectrum exhibits a restored profile close to the one described by Kanekar et al. (2001), albeit less precise, whereas the non-robust unclipped spectrum in the same frequency band is completely swamped by RFI, making B2 undetectable.

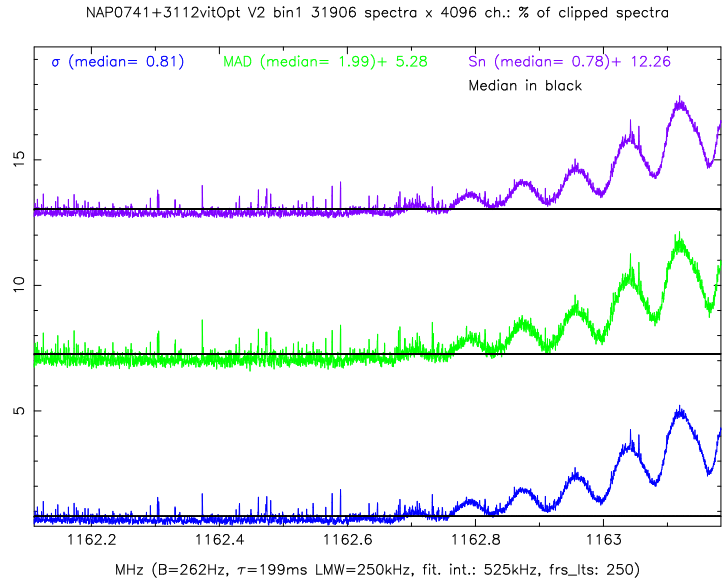


FIGURE 4.38: B0738+313: percentage of averaged power spectral values $\sigma$-clipped (blue line), $MAD$-clipped (green line), and $Sn$-clipped (purple line) within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

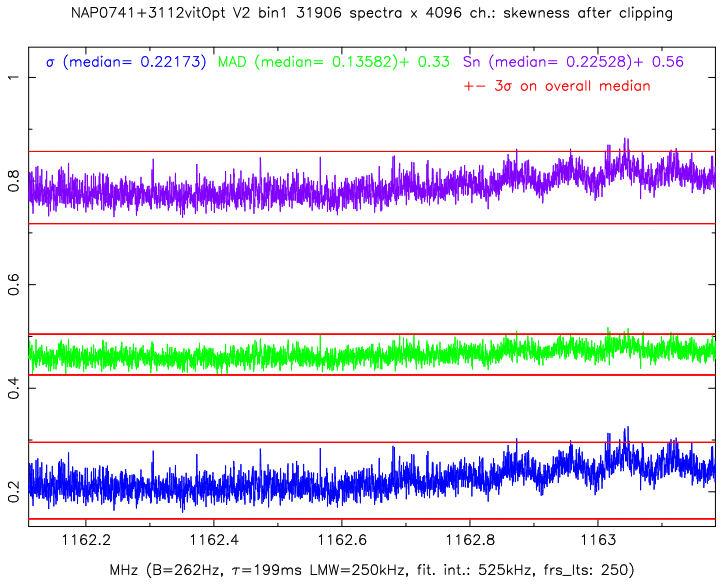


FIGURE 4.39: B0738+313: skewness after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

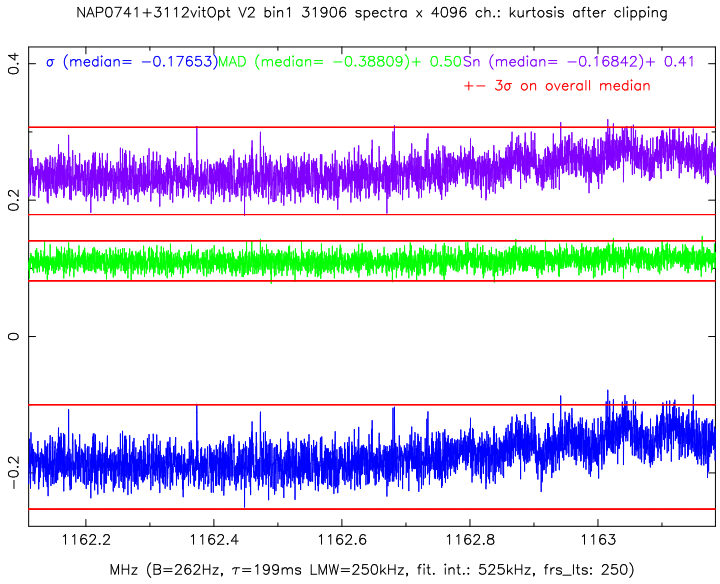


FIGURE 4.40: B0738+313: kurtosis after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

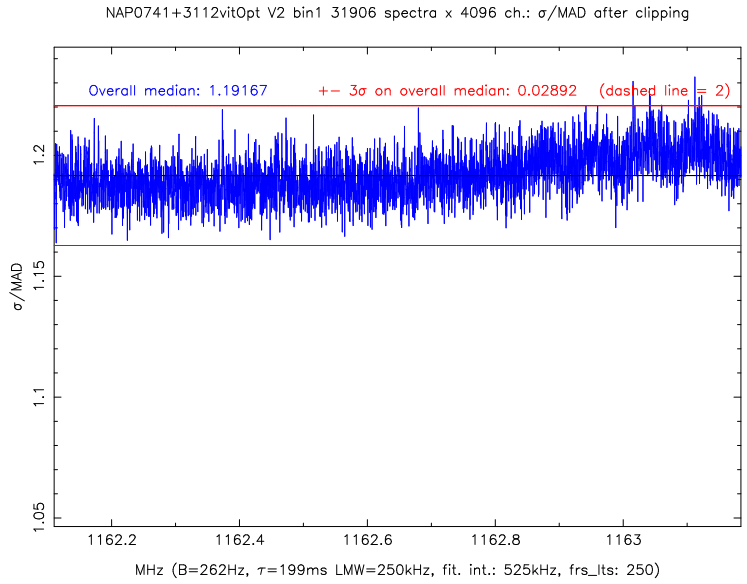


FIGURE 4.41: B0738+313: $\frac{\sigma}{MAD}$ after clipping within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

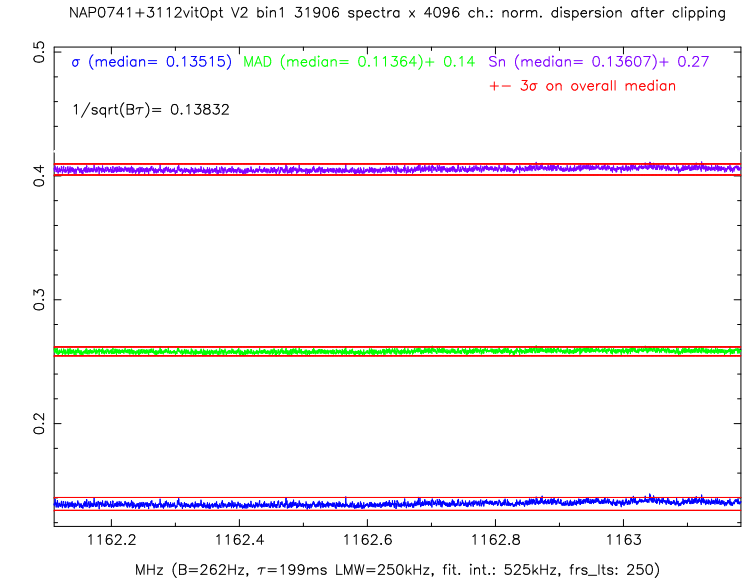


FIGURE 4.42: B0738+313: normalized estimators of scale after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band.

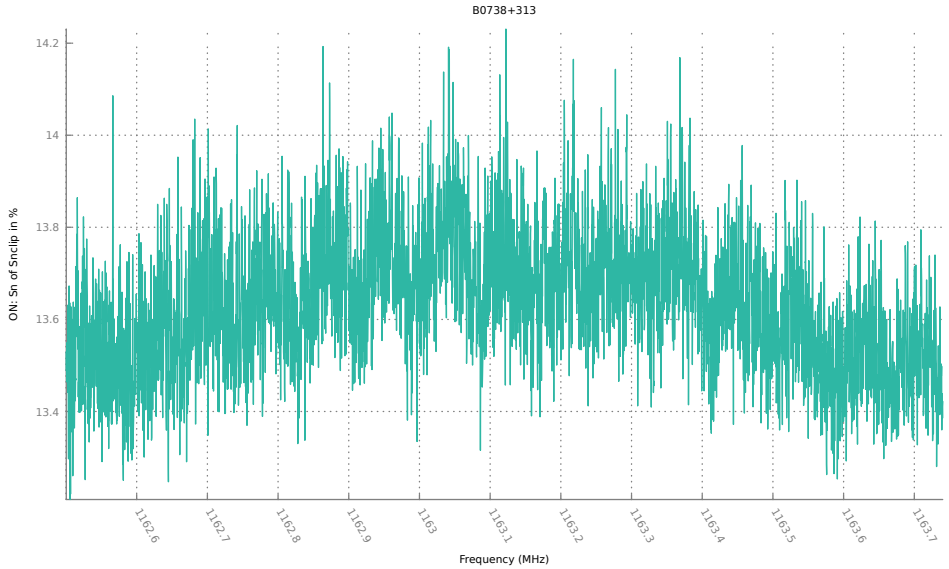


FIGURE 4.43: B0738+313: normalized temporal $Sn$ of uncalibrated $Sn$-clipped averaged power spectral values of the $ON$ spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber. This portion is a zoom of the purple line of Fig. 4.42.

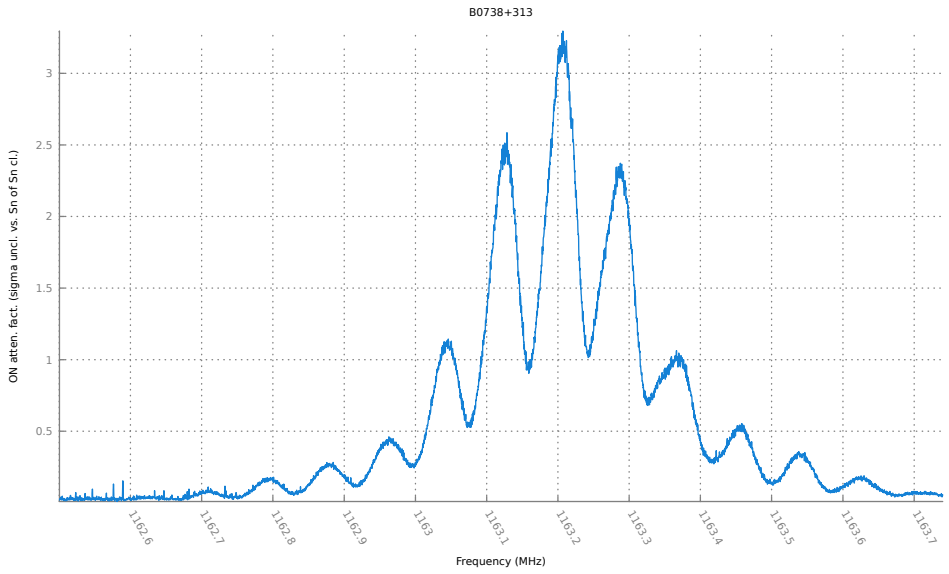


FIGURE 4.44: B0738+313: Temporal EoS attenuation factor ($Sn$ of $Sn$-clipped median $-$ $\sigma$ of unclipped median)/($Sn$ of $Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber.

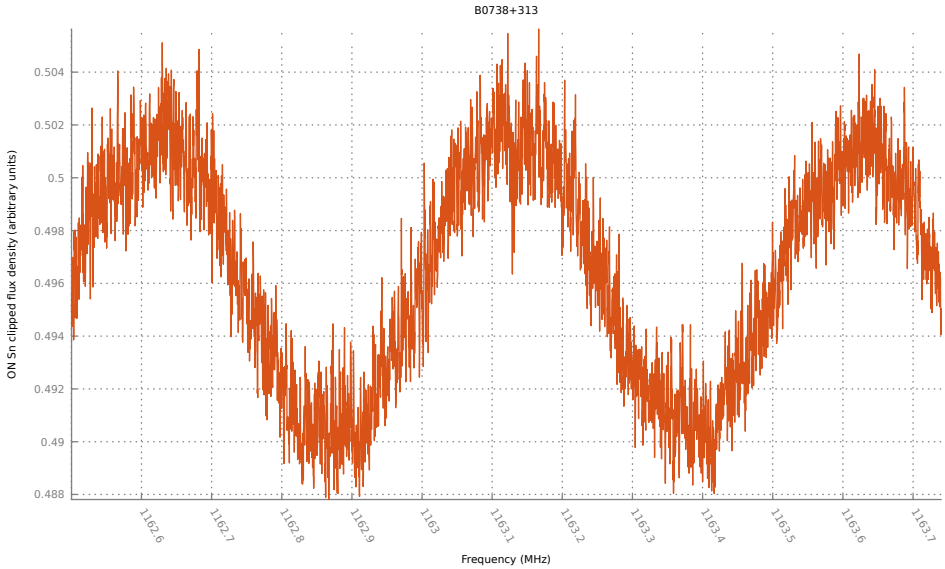


FIGURE 4.45: B0738+313: *Sn*-clipped median of uncalibrated averaged power spectral values of the *ON* spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber.

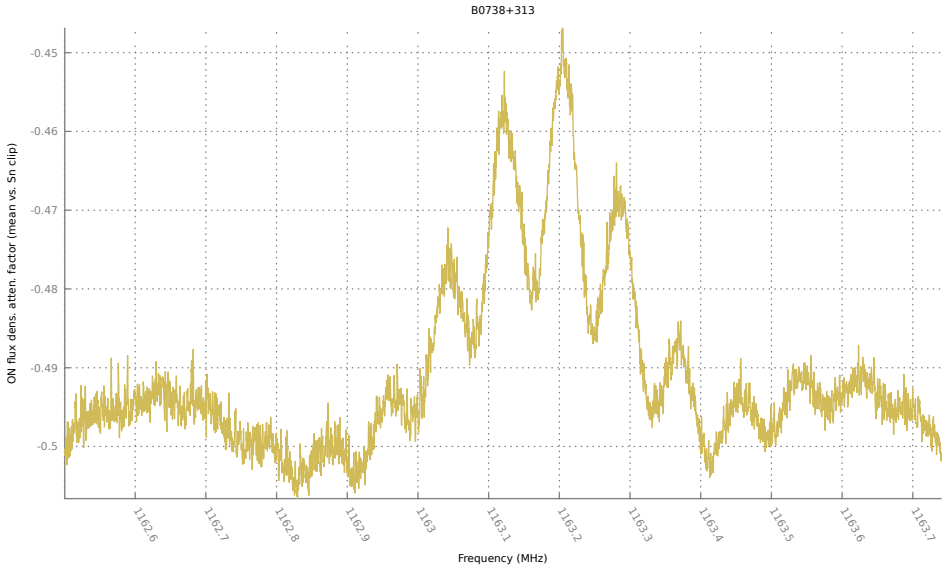


FIGURE 4.46: B0738+313: EoL attenuation factor (unclipped mean - *Sn*-clipped median)/(*Sn*-clipped median) of uncalibrated averaged power spectral values of the *ON* spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber.

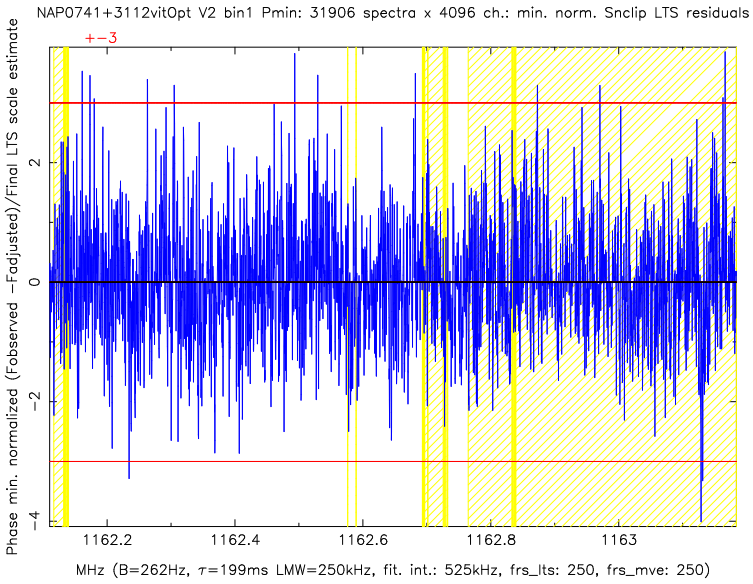


FIGURE 4.47: B0738+313: LTS residuals (blue line) of *Sn*-clipped median of uncalibrated averaged power spectral values of the *ON* spectrum within the 1162.11 - 1163.18 MHz Local Standard of Rest (LSR) frequency band. The two red lines mark the $\pm 3$ limit for automated detection. Spectral channels with white background are flagged by ROBEL as reliable before and after clipping; yellow background underlines channels valid only after clipping, orange background channels unreliable even after clipping. The B2 absorber which lies around 1163.12 MHz has got a maximum LTS residual of $\sim -4$. Any valid channel is flagged by ROBEL as a spectral line candidate if its LTS residual exceeds $\pm 3$.

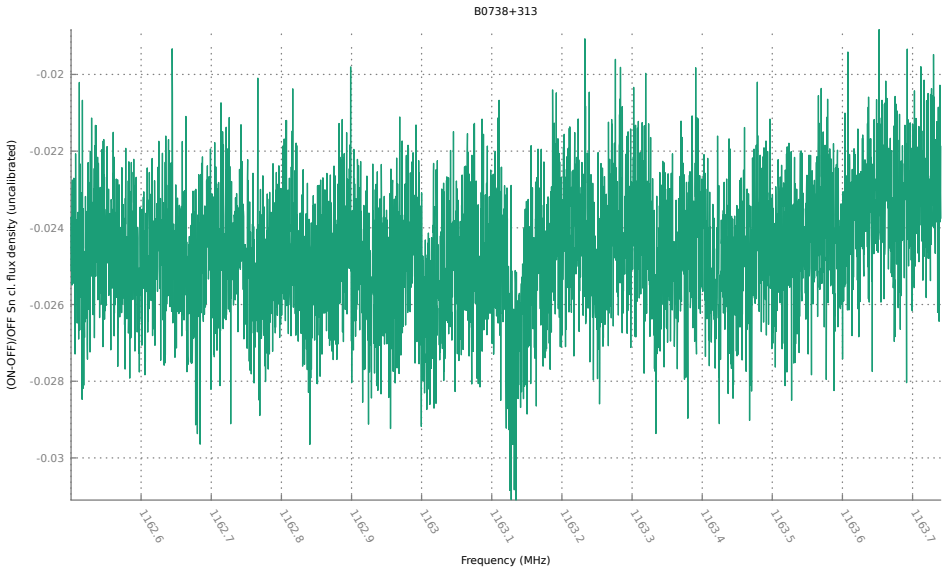


FIGURE 4.48: B0738+313: $(ON - OFF)/OFF$ spectrum (Local Standard of Rest -LSR- frequencies) around the B2 absorber of *Sn*-clipped median of uncalibrated averaged power spectral values.

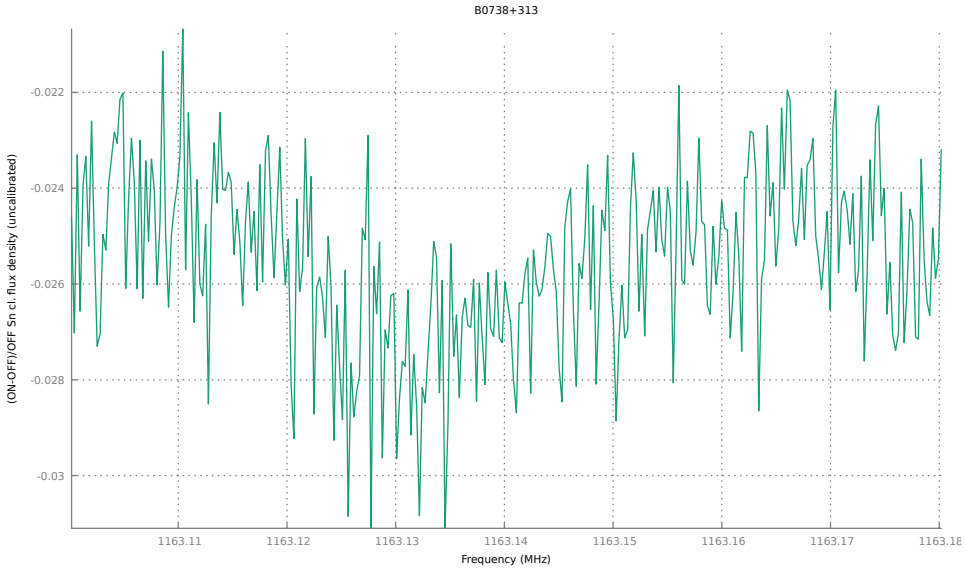


FIGURE 4.49: B0738+313: $(ON - OFF)/OFF$ spectrum (Local Standard of Rest -LSR- frequencies) of the B2 absorber only, with $Sn$-clipped median of uncalibrated averaged power spectral values. This graph is a zoom from Fig. 4.48

### 4.1.7 GNSS and radionavigation radar RFI mitigation

Between 1164 and 1400 MHz, there are numerous Global Navigation Satellite System (GNSS - see Appendix A on p. 137) and radionavigation (RN) radar signals which NRT radio astronomers have to deal with. Since it would be too long to present case studies for all of them, I have selected first one of the most powerful radars observed in the $1300 - 1400$ MHz band, and second on some GNSS signals in the $1164 - 1400$ MHz band, where I focus on the GPS L3 signal which occurs around 1381 MHz, and is particularly annoying for radio astronomers studying HI in galaxies at low redshifts. I conclude with a summary of RFI mitigation results for these cases.

All RFI discussed hereafter are from the B0738+313 broadband observations presented above. Therefore, all frequency *rms* are calculated with the WIBAR setup listed in 4.1.2 on p. 56, i.e., with channel width $B = 262$ Hz ($55ms^{-1}$) and $\tau = 0.199$ ms. From Eq. 3.15 on p. 35, I also give the equivalent frequency *rms* for a typical velocity resolution used for extragalactic line surveys, $B = 10$ km s$^{-1}$, i.e., attenuated by a factor $\sqrt{\frac{55}{10000}}$ (which is a theoretical estimate only, since we know that frequency channels are somehow correlated).

#### 4.1.7.1 Radionavigation radar at $1324 - 1332$ MHz

Radio astronomers at the NRT face a number of strong RFI signals generated by radars. One of the most powerful RN signals lies in the $1324 - 1332$ MHz frequency band. I chose to test the ROBEL RFI detection and excision algorithm on this source of strong unwanted emissions, which is clearly visible in the $ON$ spectrum in Fig. 4.50 on p. 102, as well as in the $(ON - OFF)/OFF$ spectrum on Fig. 4.51 on p. 102.

As always, data quality assessment and RFI excision must be processed on $ON$ and $OFF$ spectra separately before the robust $Sn$-clipped $(ON - OFF)/OFF$ spectrum is derived. These RFI have a specific profile. Contrary to GNSS signals which spread out over broad bands (as a result, with the narrow frequency channel set for this observation - 262 Hz, we don't see spikes, except at the vicinity of their central frequency), radar signals clearly reveal strong peaks along the frequency band. In particular:

- before clipping:
  - skewness and above all kurtosis (see Fig. 4.52 on p. 103) exhibit very high levels;
  - in parallel, the $\frac{\sigma}{MAD}$ ratio reaches very high levels in the same frequency channels (see Fig. 4.53 on p. 103);
  - the frequency $rms$ is extremely high (606 mJy) making the band unusable without RFI excision;
- after clipping:
  - a 6% median of averaged power spectral values have been excised from each frequency channel (see Fig. 4.54 on p. 104);
  - skewness has been reduced by 99.3% (see Fig. 4.55 on p. 104) and kurtosis by 99.9% (see Fig. 4.56 on p. 105), both to levels comparable to the baseline, except for a mild excess;
  - though we still notice a mild peak of $\frac{\sigma}{MAD}$, its 1.18 median marks the spectrum as entirely valid according to ROBEL set criteria (see Fig. 4.57 on p. 105);
  - the $Sn$-clipping process has smoothed the radar peaks (see Fig. 4.58 on p. 106) and severely reduced the temporal noise (see Fig. 4.59 on p. 106). The result is a smoothed $(ON - OFF)/OFF$ spectrum (see Fig. 4.60 on p. 107).
  - the frequency $rms$ of the $(ON - OFF)/OFF$ spectrum has been reduced by 87% to 79 mJy. For a B=10 $\mathrm{km\,s^{-1}}$ velocity resolution, the estimated equivalent $rms$ would be 6 mJy: this implies that any spectral line which peaks at 30 mJy could in principle be detected with a $5\sigma$ confidence.

FIGURE 4.50: RN radar: $ON$ unclipped spectrum of averaged power values in the $1324 - 1332$ MHz frequency band.

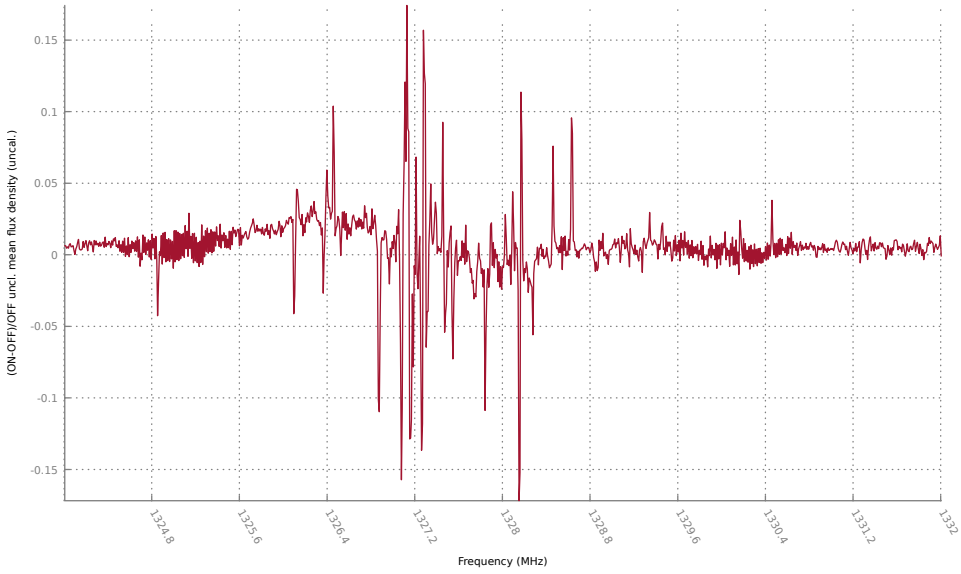


FIGURE 4.51: RN radar: $(ON - OFF)/OFF$ unclipped spectrum of averaged power values in the $1324 - 1332$ MHz frequency band.

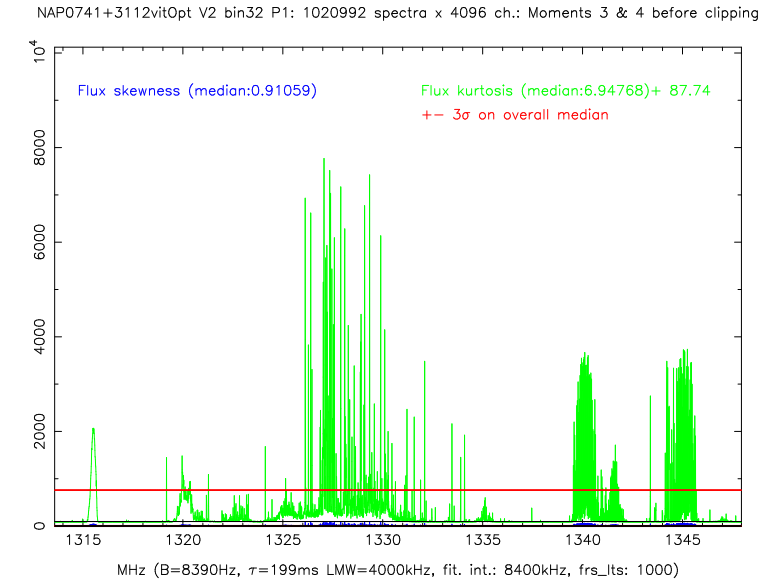


FIGURE 4.52: RN radar: skewness and kurtosis before clipping for the *ON* unclipped spectrum in the $1324 - 1332$ MHz frequency band (robust binning 32 times).

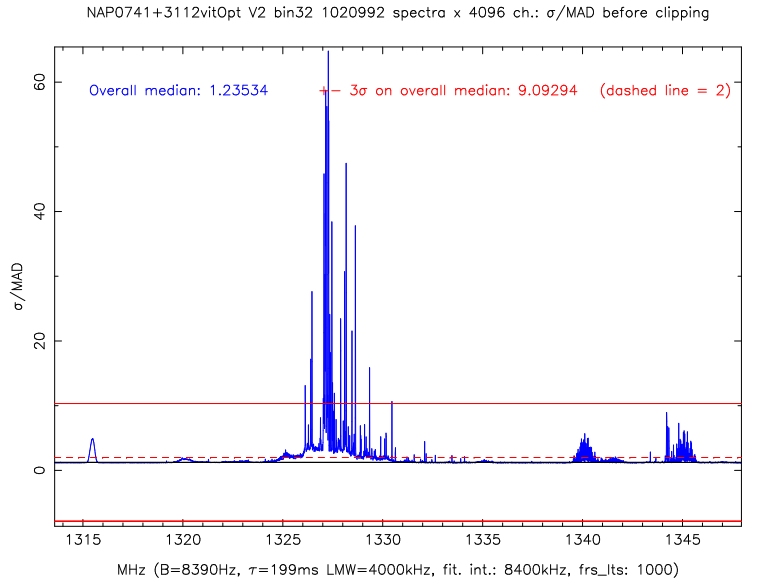


FIGURE 4.53: RN radar: $\frac{\sigma}{MAD}$ for the *ON* unclipped spectrum in the $1324 - 1332$ MHz frequency band (robust binning 32 times).

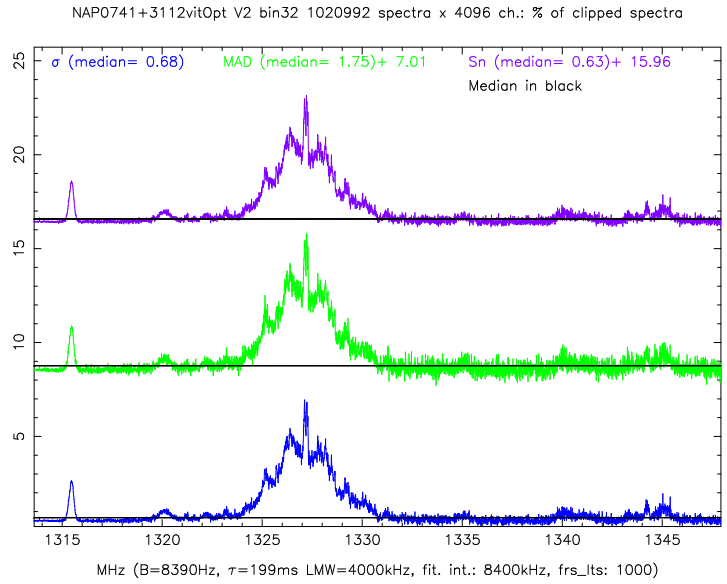


FIGURE 4.54: RN radar: percentage of averaged power spectral values $\sigma$-clipped (blue line), $MAD$-clipped (green line), and $Sn$-clipped (purple line) for the $ON$ spectrum in the $1324 - 1332$ MHz frequency band (robust binning 32 times).

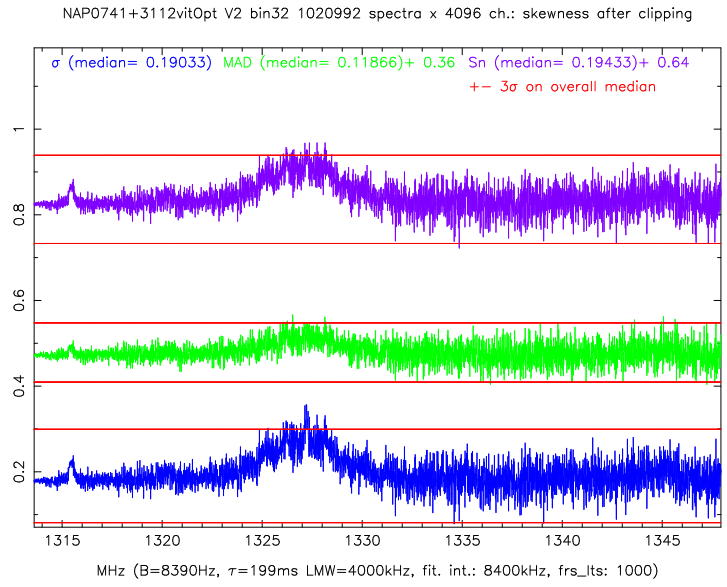


FIGURE 4.55: RN radar: skewness after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the $1324 - 1332$ MHz frequency band (robust binning 32 times).

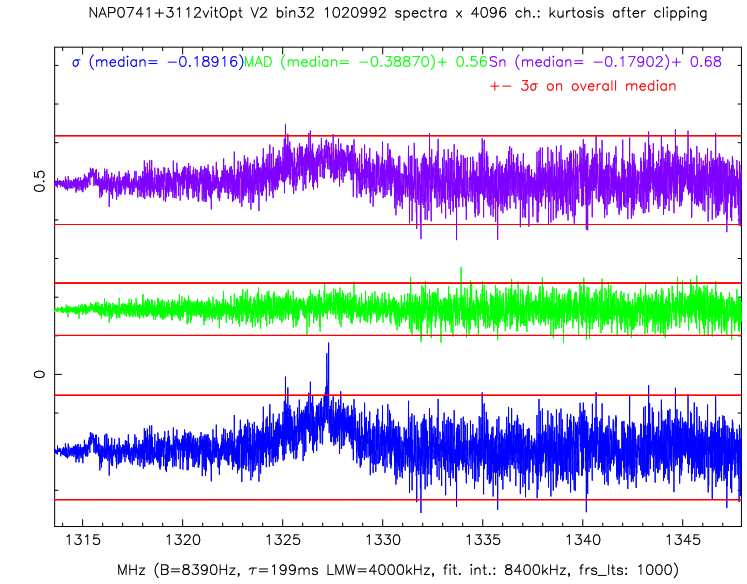


FIGURE 4.56: RN radar: kurtosis after $\sigma$-clipping (blue line), $MAD$-clipping (green line), and $Sn$-clipping (purple line) within the $1324-1332$ MHz frequency band (robust binning 32 times).

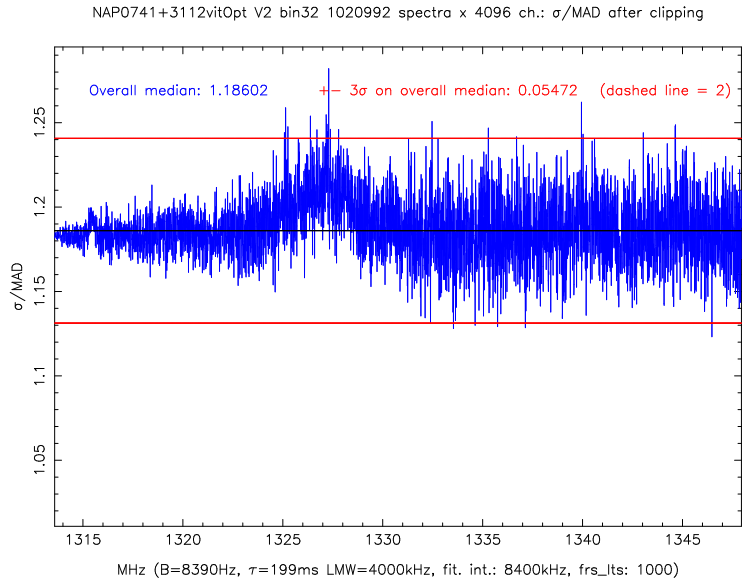


FIGURE 4.57: RN radar: $\frac{\sigma}{MAD}$ for the $ON$ clipped spectrum in the $1324-1332$ MHz frequency band (robust binning 32 times).

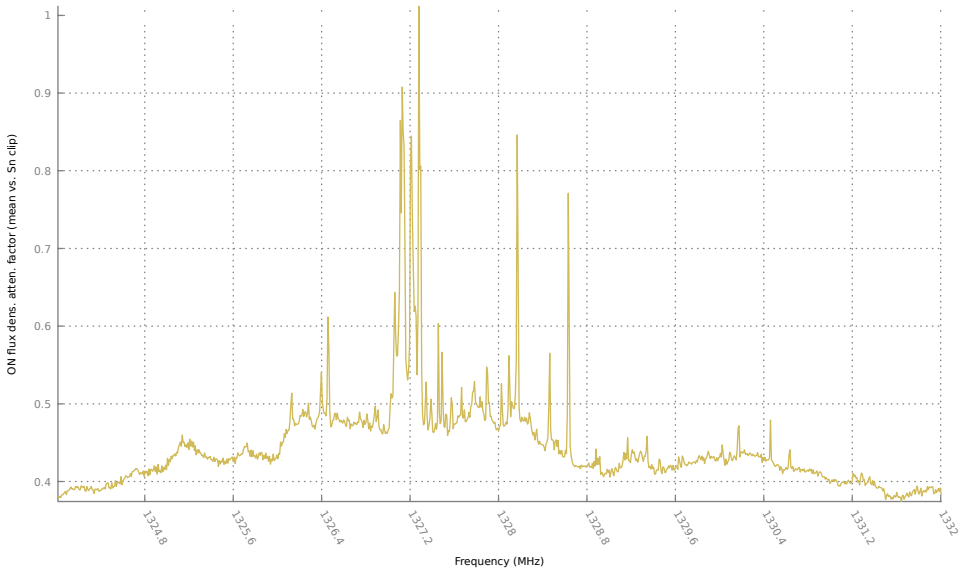


FIGURE 4.58: RN radar: EoL attenuation factor (unclipped mean - *Sn*-clipped median)/(*Sn*-clipped median) of uncalibrated averaged power spectral values of the *ON* spectrum around the 1324 – 1332 MHz frequency band.

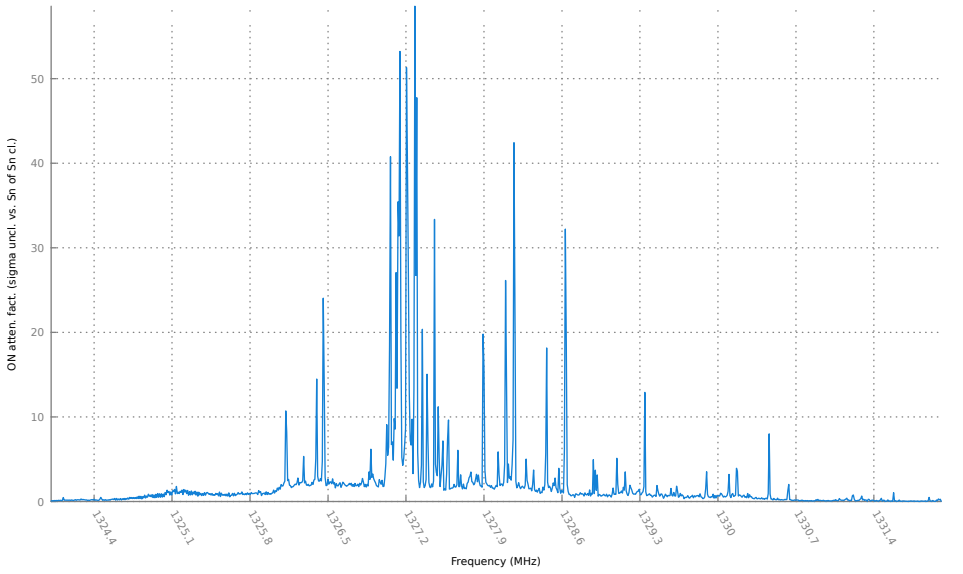


FIGURE 4.59: RN radar: temporal EoS attenuation factor (*Sn* of *Sn*-clipped median – $\sigma$ of unclipped median)/(*Sn* of *Sn*-clipped median) of uncalibrated averaged power spectral values for the *ON* spectrum around the 1324 – 1332 MHz frequency band.

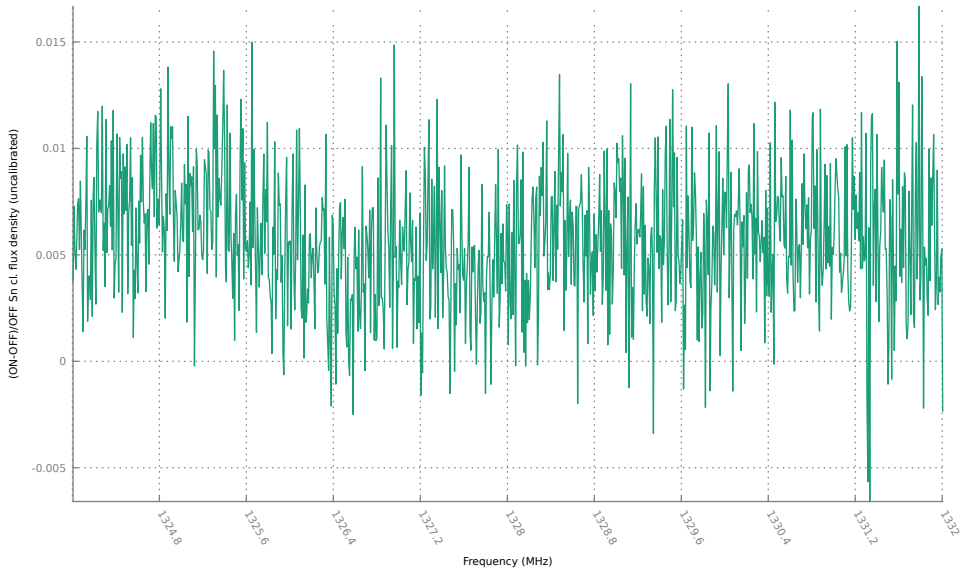


FIGURE 4.60: RN radar: $(ON - OFF)/OFF$ $Sn$-clipped spectrum in the $1324 - 1332$ MHz frequency band.

#### 4.1.7.2 GNSS at $1164 - 1400$ MHz

There are numerous Global Navigation Satellite System (GNSS) signals within the L band ($1 - 2$ GHz) and Fig. 4.61 on p.108 shows those in the $1164 - 1290$ MHz frequency band:

- GPS L5 centered on 1176.45 MHz (see Fig. A.3 on p.139);
- Galileo E5 centered on 1191.795 MHz (see Fig. A.5 on p. 141, and Fig. A.6 on p. 141 for GPS and Galileo overlapping);
- GLONASS L3 centered on 1201 MHz (see Fig. A.10 on p. 144 for option 1, and Fig. A.11 on p. 144 for option 2);
- Beidou B2 centered on 1207.14 MHz (see Fig. A.13 on p.146);
- GLONASS L2 centered on 1246 MHz (see Fig. A.9 on p.143);
- Beidou B3 centered on 1268.52 MHz (see Fig. A.14 on p.146);
- Galileo E6 centered on 1278.75 MHz (see Fig. A.7 on p.142);
- GPS L3 centered on 1381.05 MHz.

In spite of numerous RFI, ROBEL data quality assessment with thresholds set as enumerated in 4.1.3 on p.56 tells us that only the frequency channels from the $1265 - 1272$ MHz band are not valid after clipping: this matches the Beidou B3 signal which appears to be much stronger than other GNSS. As for the others, around 20% are valid before clipping and 80% after. In spite of this restrictive diagnosis, I wanted to assess the conditions of spectral line detections within or close to GNSS signals, including Beidou B3. Before presenting selected case studies of RFI mitigation within the $1164 - 1290$ MHz frequency band, I will now focus on the GPS L3.

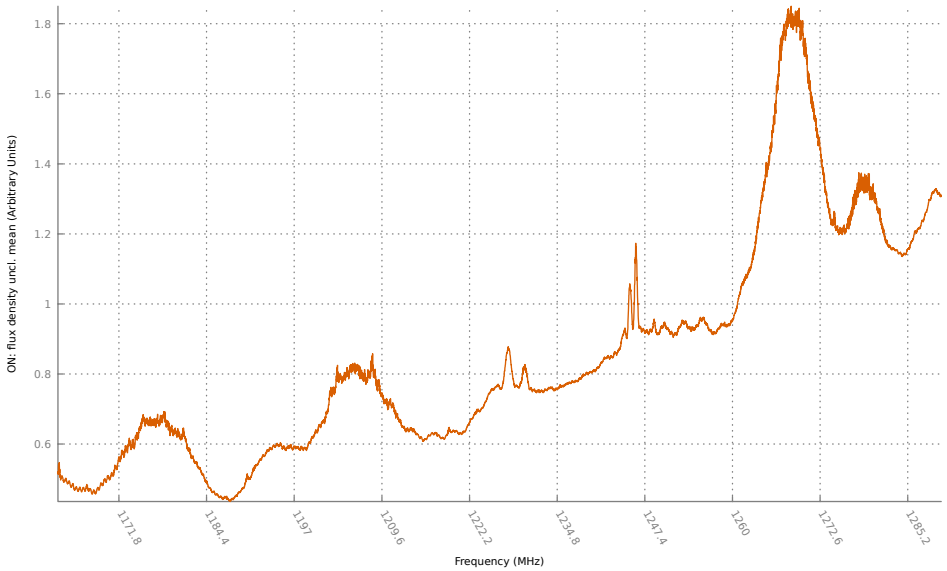


FIGURE 4.61: *ON* unclipped spectrum of averaged power values of the $1163 - 1290$ MHz frequency band after robust binning 64 times (i.e., $31906 \times 64 = 2041984$ averaged power spectral values in each of the $262 \times 64 = 16780$ Hz wide frequency channels).

The intermittent GPS L3 signal is centered on 1381.05 MHz and normally appears to occupy the $1380 - 1382$ MHz band (i.e., the $8300 - 8800$ km s$^{-1}$ HI LSR radial velocity range), but when exceptionally strong it can cause RFI over a much larger range. It is of particular concern for radio astronomers because it covers a redshift range where numerous galaxies could in principle be detected in HI. Note that it is difficult to assess the lower bound of L3 since several radio-navigation RFI lie in the vicinity (see Fig. 4.62 on p.110)). One of them overlaps the L3 signal (see Fig. 4.63 on p.110)), but it has disappeared on the unclipped $(ON - OFF)/OFF$ spectrum (see Fig. 4.64 on p.111) which has a 167 mJy frequency *rms*. The EoL attenuation is not as spectacular as for the radar (see Fig. 4.65 on p.111), but the EoS attenuation is significant (see Fig. 4.66 on p.112). The *Sn*-clipped spectrum (see Fig. 4.67 on p.112) has been reduced by 41% to 98 mJy frequency *rms*. For a B=10 km s$^{-1}$ velocity resolution, the estimated equivalent *rms* would be 7 mJy.

The *Sn*-clipping of other GNSS signals brings the following results, as Table 4.3 on p.113 shows that:

- though the processing of the GPS L5 signal in the $1176.07 - 1177.15$ MHz band brings significantly reduced temporal noise, skewness and kurtosis, such improvements are not translated in frequency *rms* which stays high and stable after *Sn*-clipping, probably because it includes the 1176.45 MHz central frequency where the signal peaks. For a velocity resolution $B = 55$ m s$^{-1}$, the frequency *rms* of the $(ON - OFF)/OFF$ spectrum is 160 mJy, and its equivalent at B=10 km s$^{-1}$ resolution would be a 12 mJy *rms*;
- the same GPS L5 signal processed this time in the $1182.52 - 1183.59$ MHz band which is relatively close to the central frequency but on a negative slope of the *ON* spectrum (see Fig. 4.61 on p.108) has a 56 to 57 mJy *rms* frequency before and after clipping, equivalent to 4 mJy at B=10 km s$^{-1}$;

- the processing of the $1269.25 - 1269.75$ MHz band close to the Beidou B3 1268.52 MHz central frequency brings better results after clipping in reducing temporal noise, skewness and kurtosis. The *rms* frequency is reduced by 12% to 120 mJy, equivalent to 9 mJy at B=10 $\mathrm{km\,s^{-1}}$;
- there is a significant improvement with the $1272.75 - 1273.83$ MHz band which lies a bit further to the B3 central frequency and on a negative slope of the *ON* spectrum. Reductions of temporal noise, skewness and kurtosis are still very good, but this time the *rms* frequency is reduced by 39% to 66 mJy (i.e., 5 mJy and thus a 25 mJy $5\sigma$ detection confidence at B=10 $\mathrm{km\,s^{-1}}$), a better value than for GPS L3 and radar;
- the processing of the Galileo E6 RFI in the $1278.12 - 1279.20$ MHz band which includes its 1278.75 MHz central frequency brings surprisingly good results on a signal at its peak. The reductions of temporal noise, skewness and kurtosis happen to be among the best of all these case studies. The *rms* frequency is reduced by 51% to 69 mJy, which is equivalent to 5 mJy for B=10 $\mathrm{km\,s^{-1}}$.

In conclusion, these case studies illustrate the ability of ROBEL to deliver $5\sigma$ peak signal-to-noise detections of 60 mJy line peaks right across strong radar signals at 10 $\mathrm{km\,s^{-1}}$ velocity resolution and $\tau \approx 0.2$ s almost everywhere in the $1164 - 1400$ MHz band, except in windows of a few MHz wide.

While these studies have provided a glimpse of RFI mitigation with ROBEL, it is difficult to infer conclusions on the overall frequency band: there are sub bands where RFI are weaker, but since GNSS and RN RFI are often superimposed, the band landscape is rather complex and few frequency channels are completely devoid of RFI. A systematic study of *rms* levels in the entire frequency band would provide detection thresholds for sub bands: the implementation of the required extra software development is planned in the near future.

For the following observation of the III Zw 35 megamaser, WIBAR was setup to a much higher sampling rate on a targeted and relatively narrow frequency band.

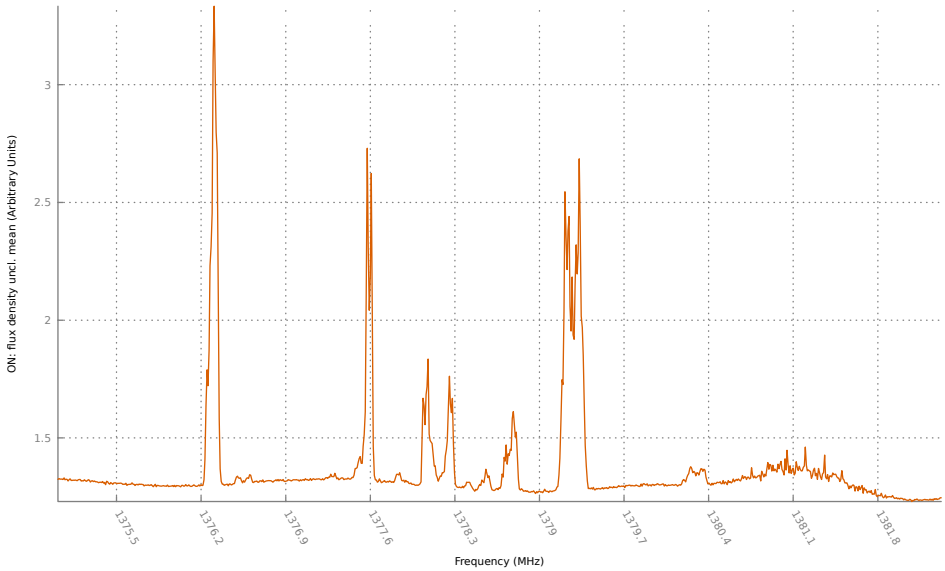


FIGURE 4.62: *ON* unclipped spectrum of averaged power values in the 1375 − 1382 MHz frequency band. The GPS L3 signal is centered on 1381.05 MHz. The other RFI are radio-navigation signals.

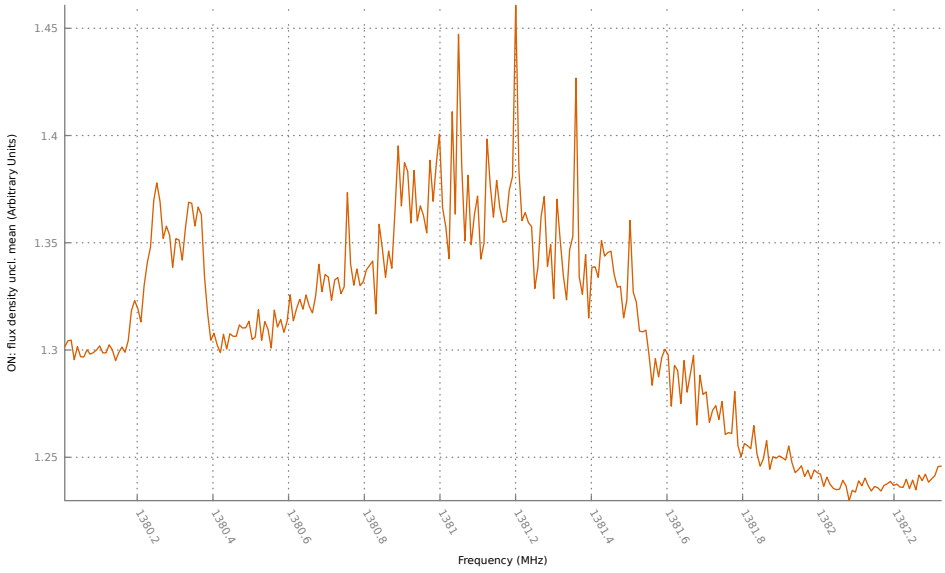


FIGURE 4.63: *ON* unclipped spectrum of averaged power values in the 1380 − 1382 MHz GPS L3 frequency band. Another RFI appears to overlap the L3 signal between 1380.2 and 1380.4 MHz.

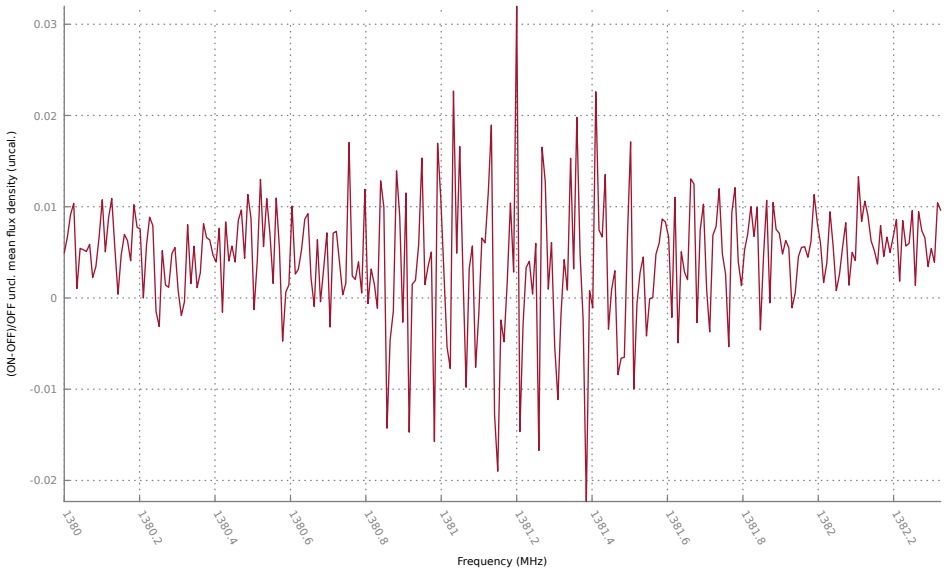


FIGURE 4.64: $(ON - OFF)/OFF$ unclipped spectrum of averaged power values in the $1380 - 1382$ MHz GPS L3 frequency band.

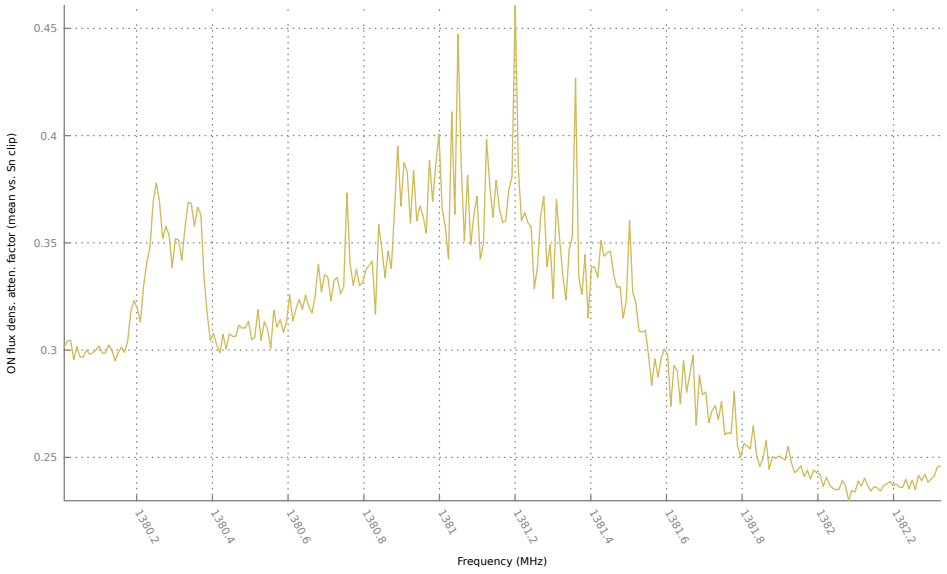


FIGURE 4.65: EoL attenuation factor (unclipped mean - $Sn$-clipped median)/($Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum around the $1380 - 1382$ MHz frequency band.

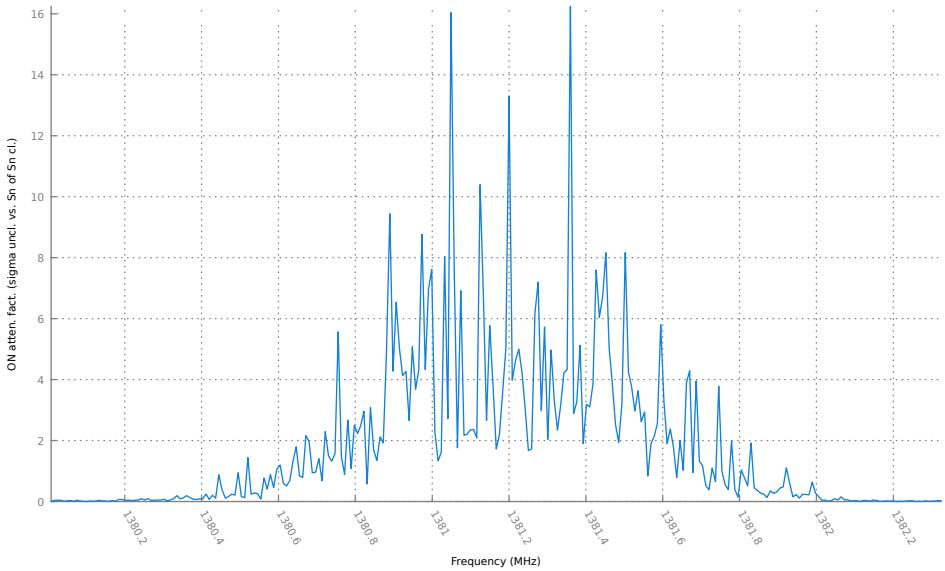


FIGURE 4.66: Temporal EoS attenuation factor ($Sn$ of $Sn$-clipped median $-$ $\sigma$ of unclipped median)/($Sn$ of $Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum around the 1380 $-$ 1382 MHz GPS L3 frequency band.

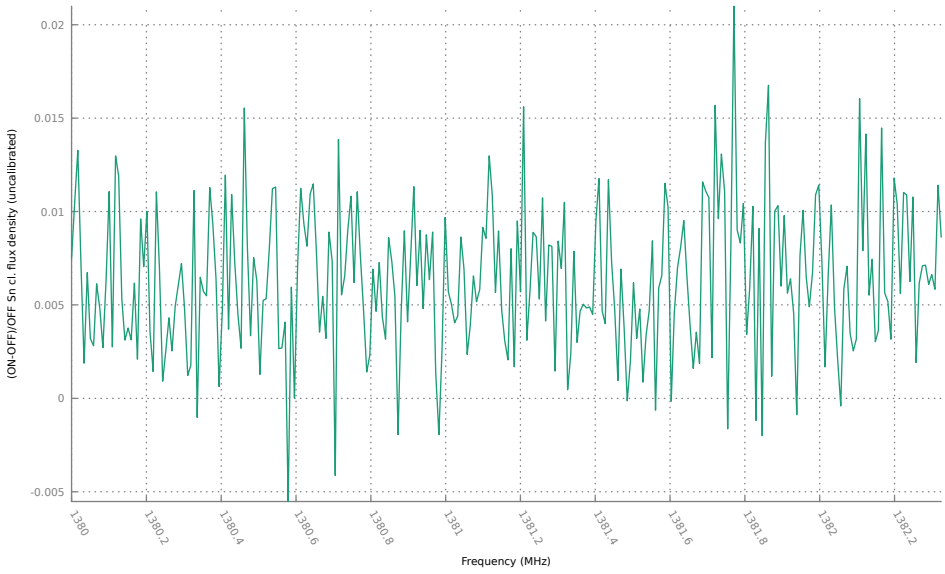


FIGURE 4.67: $(ON - OFF)/OFF$ $Sn$-clipped spectrum in the 1380 $-$ 1382 MHz GPS L3 frequency band.

| B = 55 m.s$^{-1}$<br>τ = 199 ms | Processed frequency band (MHz) | Before clipping | | | | After Sn-clipping | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ON spectrum : time-line median | | | Frequency rms (ON-OFF)/OFF (mJy) | ON spectrum : time-line median | | | | | | Frequency rms (ON-OFF)/OFF (mJy) | Frequency rms Δ % | Equivalent frequency rms for B= 10 km.s$^{-1}$ (mJy) | 5 σ confidence for B= 10 km.s$^{-1}$ (mJy) |
| | | σ | Skewness | Kurtosis | | % clipping | Sn | Δ % Sn σ | Skewness | Δ % skewness | Kurtosis | Δ % kurtosis | | | |
| **GNSS signal** | | | | | | | | | | | | | | | |
| GPS L5 | 1176.07 - 1177.15 | 0.38 | 2.47 | 8.11 | 154 | 11% | 0.18 | -53% | 0.57 | -67% | -0.11 | -98% | 160 | 4% | 12 | 60 |
| | 1182.52 – 1183.59 | 0.17 | 1.13 | 2.90 | 56 | 3% | 0.15 | -12% | 0.35 | -69% | -0.11 | -96% | 57 | 2% | 4 | 20 |
| Beidou B3 | 1269.25 – 1269.75 | 0.8 | 2.45 | 7.00 | 137 | 25% | 0.20 | -75% | 0.68 | -72% | -0.15 | -96% | 120 | -12% | 9 | 45 |
| | 1272.75 – 1273.83 | 0.44 | 2.57 | 8.17 | 108 | 13% | 0.17 | -61% | 0.57 | -78% | -0.10 | -99% | 66 | -39% | 5 | 25 |
| Galileo E6 | 1278.12 – 1279.20 | 0.49 | 4.18 | 21.61 | 140 | 9% | 0.15 | -69% | 0.41 | -91% | -0.08 | -99.7% | 69 | -51% | 5 | 25 |
| GPS L3 | 1380.00 – 1382.32 | 0.20 | 5.84 | 79.92 | 167 | 14% | 0.14 | -30% | 0.25 | -96% | -0.15 | -99.8% | 98 | -41% | 7 | 35 |
| **Radar** | 1324.00 – 1332.00 | 5.20 | 43.82 | 800 | 606 | 6% | 0.19 | -96% | 0.31 | -99.3% | -0.12 | -99.99% | 79 | -87% | 6 | 30 |

1. Before clipping :
1.1. for the ON spectrum, median of :
temporal standard deviation (σ) (col. 3)
temporal skewness (col.4)
temporal kurtosis (col.5)
1.2. frequency rms of the (ON-OFF)/OFF spectrum (col. 6)

2. After Sn-clipping :
2.1. For the ON spectrum :
median of the percentage of clipped flux densities (col. 7)
temporal Sn (col. 8)
variation between the unclipped σ and Sn (col. 9)
temporal skewness (col. 10)
variation between the unclipped and Sn-clipped skewness (col. 11)
temporal kurtosis (col.12)
variation between the unclipped and Sn-clipped kurtosis (col. 13)
2.2. frequency rms of the (ON-OFF)/OFF spectrum (col. 14)
Variation between unclipped and Sn-clipped frequency rms of the (ON-OFF)/OFF spectrum (col. 15)
Equivalent frequency rms at B=10km.s$^{-1}$ (col. 16) = frequency rms (col 14) x √(55/10000)
5 σ confidence with B=10km.s$^{-1}$ (col. 17) = Equivalent frequency rms at B=10km.s$^{-1}$ (col. 16) x 5

TABLE 4.3: Results on selected GNSS and radar RFI mitigation.

## 4.2 Observation of OH 1665 and 1667 MHz spectral lines of the III Zw 35 megamaser

### 4.2.1 Background

Luminous Infrared Galaxies (LIRGs) contain very dusty star formation regions which emit infrared (IR) radiation. III Zw 35 is a LIRG which has two nuclei as the result of a probable galaxy merger. One of these hosts a hydroxyl (OH) megamaser which was first observed at Jodrell Bank by Staveley-Smith et al. (1987), and thereafter by Martin (1989) at the NRT and Trotter et al. (1997) at the Very Long Baseline Array, among others. The radio continuum background source, which in the case of III Zw 35 is believed to be a high dust concentration heated by SNRs (Pihlström et al., 2001), pumps OH molecules with IR photons (which have higher energy levels than the OH 1667 and 1665 Mhz hyperfine transitions). The 1667 MHz line is always the strongest of the two, and megamaser 1667/1665 line flux ratio ranges from 2 to 20 (Randell et al., 1995); more specifically it is 9 for III Zw 35 (Staveley-Smith et al., 1987).

For III Zw 35, at Jodrell-Bank, Staveley-Smith et al. (1987) found an OH line LSR systemic velocity of $8262 \pm 5$ km s$^{-1}$ (within the 1667 MHz line rest frame), i.e., at $1622.641 \pm 0.026$ MHz (LSR), and a 1667 peak flux density of 177 mJy at low velocity resolution (40 km s$^{-1}$). At high resolution (6 km s$^{-1}$) three components were identified: the first lies at 8240 km s$^{-1}$, or 1622.7565 MHz (LSR), with a 245 mJy peak; the second lies at 8310 km s$^{-1}$, or 1622.3878 MHz,(LSR), with a $\sim$140mJy peak (i.e., 1.75 times lower than the first component); the third lies at $8160 \pm 5$ km s$^{-1}$, or $1623.178 \pm 0.026$ MHz (LSR), with a $\sim$30 mJy peak level.

Both Staveley-Smith et al. (1987) and Martin (1989) showed in their respective spectra an OH 1665 MHz spectral line covering the $8186 - 8338$ km s$^{-1}$ range (within the 1665 MHz rest frame, i.e., the $1620.3361 - 1621.1358$ MHz LSR frequency band) with a peak at $\sim$30 mJy, i.e., $\sim$8 times lower than the OH 1667 first component peak (except for the Staveley-Smith et al. (1987) low resolution spectrum, values are extracted from Fig. 4.68 on p.115 and 4.69 on p.116, no exact figure is available). Given these line strengths, without RFI, detections of both OH lines were well feasible with radio telescopes.

Both OH transitions could be readily observed before the onset of Iridium satellite constellation RFI in 1998. This voice and data telecommunications constellation transmits a right-hand circularly polarized (RHCP) signal in the frequency band 1616 - 1625 MHz (see Appendix B on p.151) which is up to 15 times more powerful than the cosmic background and increases the temporal noise up to 8000 times in frequency channels. Therefore, getting a spectrum from the OH1667 MHz line with no significant distortion compared to previous observations is a real challenge, whereas detecting the 1665 MHz line with more than $3\sigma$ confidence interval has not been performed until now.

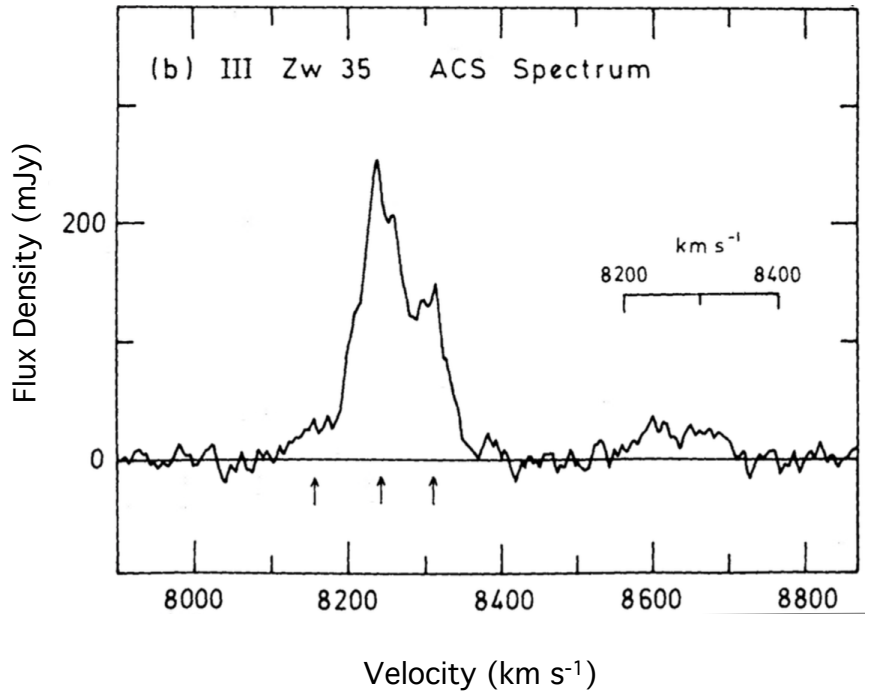


FIGURE 4.68: III Zw 35: high-resolution (6 $\mathrm{km\,s^{-1}}$) OH 1667 and OH 1665 MHz line spectra observed at Jodrell Bank by Staveley-Smith et al. (1987). Radial velocities along the horizontal axis are given in the 1667 MHz line rest frame, whereas the $8200 - 8400$ MHz inset refers to the 1665 MHz transition. The three arrows mark, from left to right, the center velocities of the third, first and second OH 1667 MHz components respectively. The OH 1665 MHz spectral lines cover the $8186 - 8338$ $\mathrm{km\,s^{-1}}$ band within its own rest frame (or $8550 - 8700$ $\mathrm{km\,s^{-1}}$ band in the 1667 MHz line rest frame)

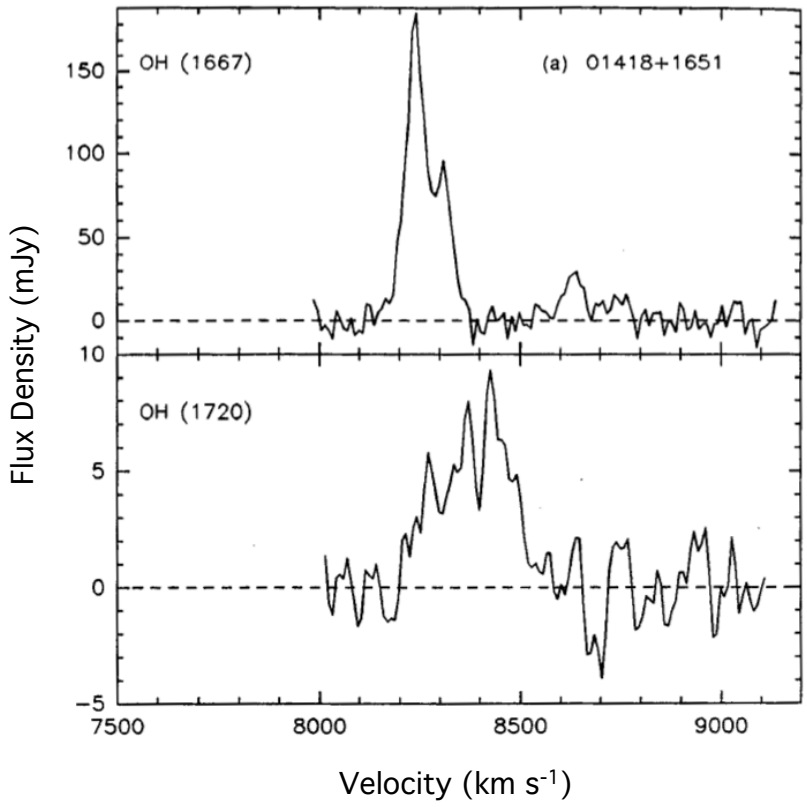


FIGURE 4.69: III Zw 35: OH 1667 1665 and 1720 MHz line spectra observed at the NRT with a 10.5 $km\,s^{-1}$ resolution by Martin (1989). Heliocentric radial velocities are given in the 1667 and 1720 MHz line rest frame(s), respectively. Please note that the OH 1720 MHz transition was not observed with WIBAR/ROBEL.

### 4.2.2 History of Iridium RFI mitigation attempts

One of the most recent attempts to mitigate Iridium RFI from single-dish radio telescope (Arecibo) data is from Deshpande et al. (2019) who worked on signal characteristics such as RHCP polarization and periodicity. They extracted data from the full Stokes parameters and excised RFI with an algorithm which is close to the ROBEL $3\sigma$ clipping. Their spectra demonstrate a high RFI mitigation efficiency. However, the paper gives no indication of their ability to discriminate between RFI and any underlying spectral lines. Moreover, their method cannot be called "robust" in the sense used in this Thesis, since neither their mean nor standard deviation are robust. In respect to such strong Iridium RFI, it is doubtful, based on various results presented in this Thesis, that non robust estimators of location and scale would excise as much RFI as $Sn$-clipping: indeed, in the presence of strong RFI, the residual temporal noise (sigma) after clipping is often higher than the one derived from robust clipping (Sn).

At the NRT, Dumez-Viou (2007) used an algorithm which blanked the frequency channels swamped by RFI as well as their neighboring channels, by defining a dynamic blanking threshold. With this, the OH 1667 MHz spectral line from III Zw 35 could be detected. In Fig. 4.70 on p.117 in the blanked spectrum on the right,

the first 1667 MHz line component detected by Staveley-Smith et al. (1987) peaks at $\sim 205 mJy$, the second at ∼125 mJy (i.e., ∼1.8 times lower) and the third component is not visible. The OH 1665 MHz peak value is uncertain, somewhere between ∼50 and 100 mJy (i.e., ∼1.9 times lower than the OH 1667 MHz line first component peak), due to the presence of two strong peaks which appear to be residual RFI. These values are significantly different from those determined by Staveley-Smith et al. (1987) (see Table 4.3). One of the reasons is the non-linearity of the baseline around the OH 1665 MHz spectral line, probably caused by a relative inefficiency of RFI excision in this frequency band. To assess spectral line fluxes, it is then necessary to adjust the baseline around for the OH 1667 1665 MHz lines. However, as we do not have the numerical data used for making the right-hand graph in Fig. 4.70, we cannot perform a higher-order polynomial baseline fit on the spectrum. For the 1667 Mhz line, choosing a $-10$ mJy linear baseline adjustment brings the main component peak to 215 mJy and the second to 135 mJy, the 1667/1665 MHz line peak ratio ratio becoming 1.6 to 1, which is closer to the 1.75 to 1 ratio from Staveley-Smith et al. (1987). Then, with a linear baseline at ∼20 mJy, the OH 1665 MHZ MHz line peak is now at ∼90 mJy. However, this spectral line is not so clearly detected, and certainly not with a $3\sigma$ confidence.

At the NRT, Dumez-Viou (2007)'s observations were made in 2005 whereas ours were made in 2018, with a more than tripled number of subscribers to Iridium's services.

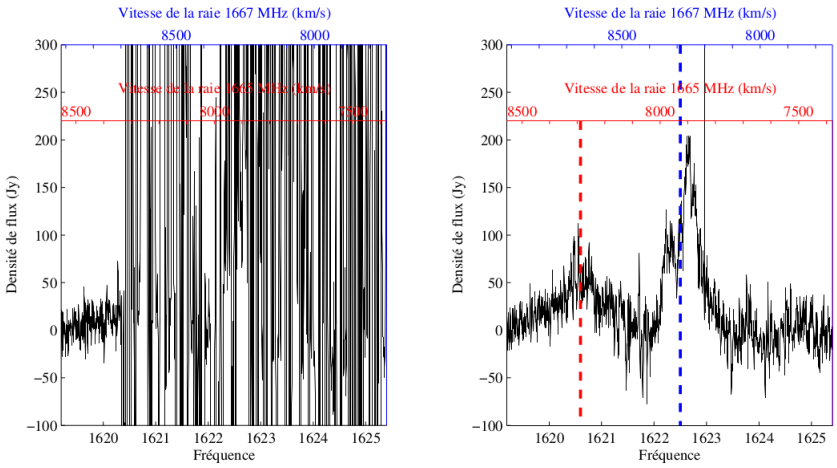


FIGURE 4.70: III Zw 35: OH 1667 and OH 1665 MHz line observation made at the NRT by Dumez-Viou (2007). The unclipped spectrum is on the left, and the RFI-blanked version is on the right. The horizontal axes show frequency (MHz), the vertical axes flux density (mJy). At the top of each spectrum, the blue and red scales show radial velocities in the 1667 and 1665 MHz frameworks respectively. The vertical dashed line mark the center velocities of both lines. No error margins are documented.

### 4.2.3 WIBAR and ROBEL setup

The WIBAR observation of III Zw 35 consisted of 3 scans recorded over a total bandwidth limited to 34 MHz to secure high speed data transfer due to the sampling rate $\tau = 4.67$ ms, and to limit post-processing. The frequency channel width was set

to $B = 4195Hz$ (or $\sim$0.9 $\mathrm{km\,s^{-1}}$). Each scan consisted of alternate $ON$ and $OFF$-position cycle pairs lasting 1 minute each. One cycle included 12587 spectra, The total number of $ON$ spectra (which is the same as $OFF$) recorded in the 3 scans was 780394, for a total $ON + OFF$ integration time of about 2 hours.

ROBEL data processing parameters used for this observation are the following:

- 4096 frequency channels per 1D spectrum;
- rest center frequency: 1612.231010 MHz chosen because it is the closest to the redshifted OH spectral lines;
- Solar System barycentric radial velocity correction to LSR (Local Standard of Rest - see 3.5.1.2 on p.48);
- normalization frequency band: $1611 - 1613$ MHz;
- clipping at $5 \times 3EoS$ (EoS: estimator of scale);
- line maximum width (LMW) for automated detection: 2000 kHz;
- frequency interval for LS and LTS baseline fitting and automated detection: 4200 kHz;
- size of the randomly selected sample for LTS processing: 500 times the number of frequency channels within the interval set for baseline fitting;
- data quality estimators limits:
  - skewness $\leq 1$);
  - kurtosis (minimum: $-1.5$; maximum: 3);
  - $\frac{\sigma}{MAD} \leq 2$;
- detection threshold at $\pm 3EoS$ off the spectrum normalized median.

With this setup, from Eq. 3.9 on p.31, we get the elemetary integration time $T_s \approx 0.24$ ms and the number of power spectra averaged during $\tau = 4.67$ ms: $N_{scn} \approx 4.67/0.24 \approx 20$. This is slightly less than the minimum of 25 required to state that the averaged power spectral values in each frequency channel follow a normal law in the absence of non-Gaussian samples generated by RFI (see 3.2.1 on p.33). This induces a mild positive skewness and kurtosis, but the $\chi^2$ probability density function with $2N_{scn} = 40$ degrees of freedom is close to the Gaussian distribution PDF, so it is possible to apply the assumptions for RFI mitigation described in 3.2 on p.32, albeit with caution.

### 4.2.4 Results for the spectrum normalization band

The III Zw 35 spectra are normalized in the $1611 - 1613$ MHz frequency band, within the protected radio astronomy service band $1610 - 1613.8$ MHz (see 3.2.5 on p.41). This band was split into 478 frequency channels. Therefore, statistics are calculated for each cycle for a total of $478 \times 12587 = 6016586$ uncalibrated averaged power spectral values (in arbitrary units). This band is not completely devoid of weak RFI though, and I will now present results on unbinned data.

The actual normalized non-robust temporal *rms* is not constant: it is negatively correlated with frequency as shown in Fig. 4.71 on p.119. No frequency channel is

considered valid before clipping, since the $\frac{\sigma}{MAD}$ median is ∼4.68, skewness median is above 23, and kurtosis above 850. This leads to a median of the non-robust unclipped $(ON - OFF)/OFF$ spectrum which is $-0.027$. Though the actual temporal $\sigma$ median on the $ON$ spectrum is 0.45, the theoretical temporal $\sigma$ median, which equals 0.22 is well above the actual temporal $Sn$ median (0.17) which highlights a serious bias generated by a very high kurtosis (see 4.1.4.2 on p. 59). The $(ON - OFF)/OFF$ frequency $rms$ calculated with the standard value $\frac{T_{sys}}{PSE} = 25$ chosen for all the case studies in this Thesis is 8 mJy . Fortunately, $Sn$-clipping corrects all these data quality estimators to acceptable levels: all frequency channels become valid, with median values being $\frac{\sigma}{MAD} = 1.19$, skewness = 0.23, and kurtosis = $-0.17$. The median of the $Sn$-clipped $(ON - OFF)/OFF$ spectrum is close to 0, and its frequency $rms$ is 8 mJy. The $ON$ $Sn$ temporal median is 0.17, still above the theoretical temporal $\sigma$ median, though. Fig. 4.72 on p.120 and Fig. 4.73 on p.120 show that the attenuation factors on averaged power spectral values $EoL$ and $EoS$ are far from negligible.

Thus, spectra normalization in this frequency band is coherent only with $Sn$-clipped spectra: there is RFI on each of its frequency channels which are strong enough to deteriorate temporal noise and generate skewness and kurtosis.

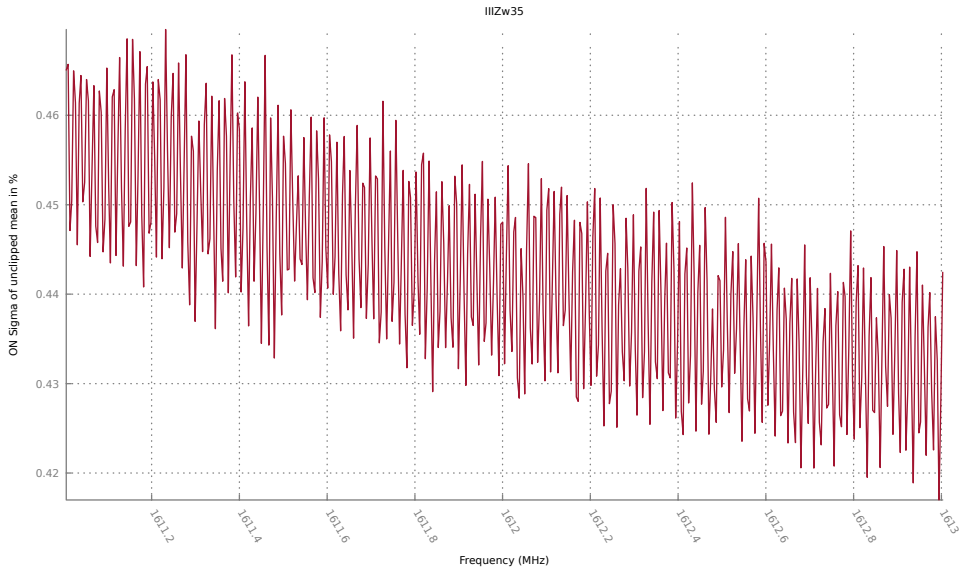


FIGURE 4.71: Normalized temporal $\sigma$ in the $1611 - 1613$ MHz protected radio astronomy service frequency band.

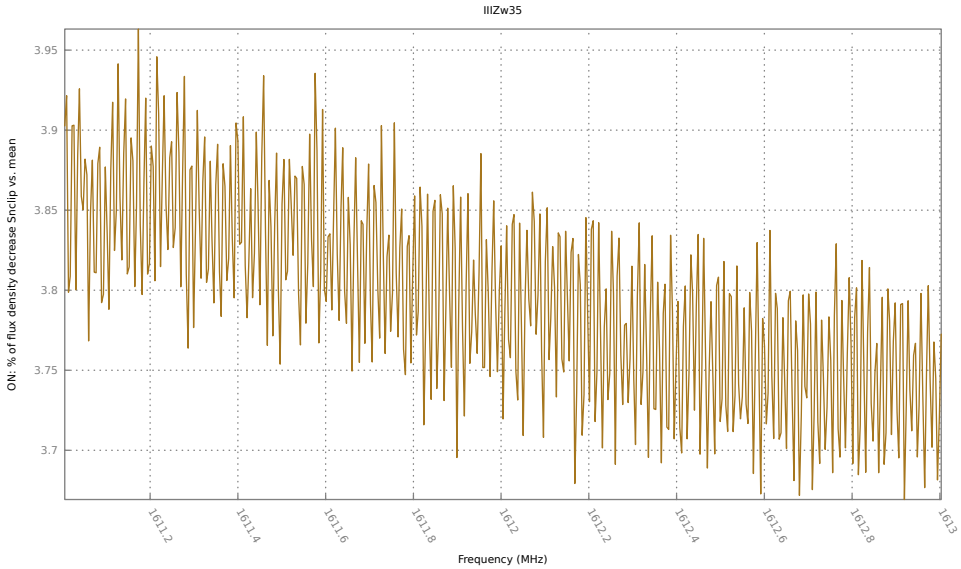


FIGURE 4.72: EoL attenuation factor (unclipped mean - $Sn$-clipped median)/($Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum in the 1611 − 1613 MHz protected frequency band.

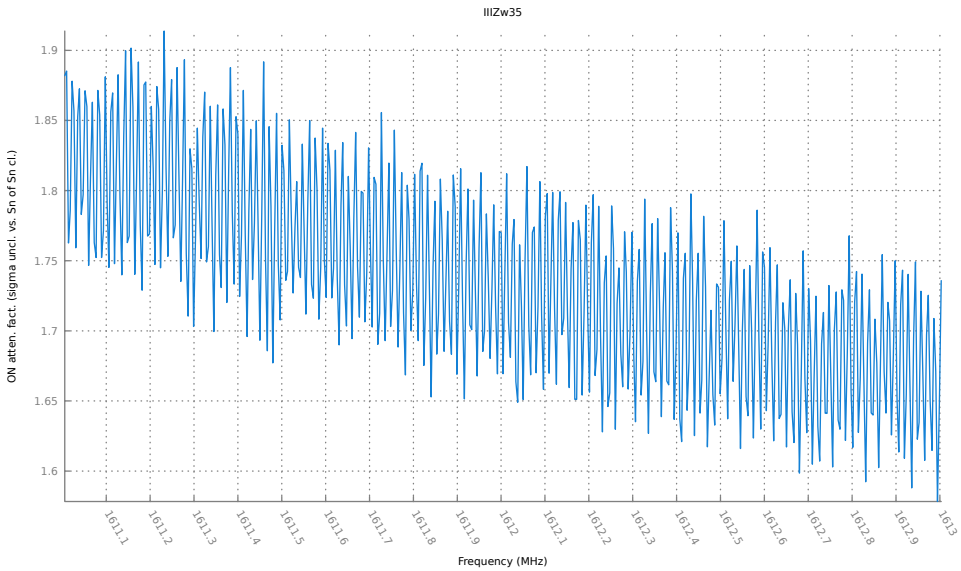


FIGURE 4.73: Temporal EoS (estimator of scale) attenuation factor ($Sn$ of $Sn$-clipped median − $\sigma$ of unclipped median)/($Sn$ of $Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum in the 1611 − 1613 MHz protected frequency band.

### 4.2.5 Results for the OH megamaser

The unclipped non-robust $(ON - OFF)/OFF$ spectrum of the 1617 − 1628.5 MHz band shows that the megamaser components are completely swamped by these RFI (see Fig. 4.74 on p. 124), and that the temporal $\sigma$ values are high and irregular (see

Fig. 4.75 on p. 124). All this makes RFI excision compulsory. The $Sn$-clipping process yields strong attenuation of both flux density $EoL$ (up to 95% around 1622.3 MHz-see Fig. 4.76 on p.125) and temporal noise (up to 3000 times in the same band - see Fig. 4.77 on p.125). We then get a much lower temporal $Sn$ along the $ON$ spectrum (see Fig. 4.76 on p.125), but still not completely smoothed. As a consequence, the $Sn$-clipped $(ON - OFF)/OFF$ spectrum reveals both the OH 1667 and 1665 MHz spectral lines: the first between 1622.2 and 1623.8 MHz, the second between 1620.1 and 1621.3 MHz (see Fig. 4.78 on p.126). Though this spectrum is flatter than Dumez-Viou (2007)'s, it still requires customized baselines for both lines: Fig. 4.79 on p.126 shows that robust temporal noise after clipping (i.e., $Sn$ of the $Sn$-clipped median of averaged power spectral values) still exhibits small RFI residuals able to alter the spectrum baseline.

Zooming in the OH1667 MHz spectral line in Fig. 4.80 on p.127, the first two components identified by Staveley-Smith et al. (1987) are clearly visible, the profile of the third one which is a negative slope is visible but not detected with a $3\sigma$ confidence. Choosing a $-0.001$ baseline and a $[EoL - 3EoS, EoL + 3EoS]$ 99.74% confidence interval ($EoL$ being the estimator of location of the temporal distribution of averaged power spectral values, $EoS$ its estimator of scale), and without gaussian fitting, we get with an overall frequency $rms$ equal to 27 mJy:

- the first component covering the 24 MHz range between 1622.66 and 1622.90 MHz at baseline level, i.e., $8235 \pm 23$ km s$^{-1}$ LSR. Its peak is $182 \pm 18$ mJy, is detected with $6.7\sigma$ confidence. This component has a FWHM of $78 \pm 16$ km s$^{-1}$ (it is the only one among the three whose width can be assessed without gaussian fitting);
- the second component peak in the range of $1622.41 - 1622.49$ MHz (i.e., $8290 - 8294$ km s$^{-1}$) at $108 \pm 18$ mJy, being detected with $4.0\sigma$ confidence;
- the third component peak in the range of $1623.32 - 1623.41$ MHz (i.e., $8115 - 8133$ km s$^{-1}$) at $64 \pm 18$ mJy, being detected with $2.3\sigma$ confidence.

From the OH 1665 spectral line profile (see Fig. 4.82 on p.128), we find a peak in the range of $1620.38 - 1621.17$ MHz (i.e., $8616 \pm 74$ km s$^{-1}$), and choosing a $-0.0008$ baseline yields to $76 \pm 18$ mJy, being detected with $2.8\sigma$ confidence; its $FWHM$ equals $195 \pm 5$ km s$^{-1}$. Note that setting the baseline value is not easy. If, instead of $-0.001$, we choose $-0.0005$, the peak is then in the range of $68 \pm 18$ mJy. Yet getting a 31 mJy similar to Staveley-Smith et al. (1987) and Martin (1989) would require a baseline set at 0.001 which is much too high given the overall spectrum shown in Fig. 4.78 on p.126. The temporal $Sn$ along the $ON$ spectrum is also irregular, albeit on a limited scale. Robust temporal noise ($Sn$) peaks at 1620.4 and 1620.7 MHz (see Fig. 4.83 on p.128) seem to match the irregularities of Dumez-Viou (2007)'s correponding spectrum (see Fig. 4.70 on p.117). In our case, these peaks do not appear to markedly influence the averaged power spectral values $Sn$-clipped median in Fig. 4.82. Table 4.4 on p.123 compares peak lines flux densities from Dumez-Viou (2007), Martin (1989), and Staveley-Smith et al. (1987) to ROBEL results.

In conclusion, the WIBAR/ROBEL observation of the III Zw 35 megamaser has restored the ability to obtain relevant OH 1667 and 1665 MHz spectral line profile parameters, within a frequency band swamped by strong RFI. There are several options to improve the results. One is to halve the sampling time $\tau$ from 4 to 2 ms, to decrease the RFI residuals after $Sn$-clipping. With the current WIBAR hardware configuration, this is the minimum applicable value. The other one is to test Tukey's

biweight estimators of location and scale (see 2.3.2 on p.18 and 2.4.1 on p.19): both have non vertical slopes which can be configured to maximize efficiency. It would also be interesting to discriminate between RHCP and LHCP spectra, since the Iridium signal is RHCP only. However, the problem resides in the secondary lobes where the signal polarization may be altered, hence it is not sure that the outcome would be much better than with linear polarization. WIBAR can also record the input waveform, in order to build various datasets with different FFT and stacking parameters. It would then be possible to test new excision algorithms.

| | OH1667 | | | OH1665 |
|---|---|---|---|---|
| (mJy) | first component peak | second component peak | third component peak | peak |
| S. Smith hi-res | 245 | 140 | 30 | 30 |
| Martin | 190 | 90 | | 30 |
| Dumez-Viou | 215 | 135 | | 90 |
| ROBEL | 182 ± 18 | 108 ± 18 | 64 ± 18 | 76 ± 18 |

TABLE 4.4: III Zw 35: comparison of peak flux densities for the three OH 1667 MHz line components and the OH 1665 MHz line derived from Dumez-Viou (2007), Martin (1989), and Staveley-Smith et al. (1987) and from ROBEL. Except for ROBEL, values are estimated from graphs and without error bars. Values from Staveley-Smith et al. (1987) are taken from the high-resolution spectrum, whereas for Dumez-Viou (2007), they are extracted from the blanked spectrum. For the latter and ROBEL, the 1667 and 1665 MHz line baselines were corrected.

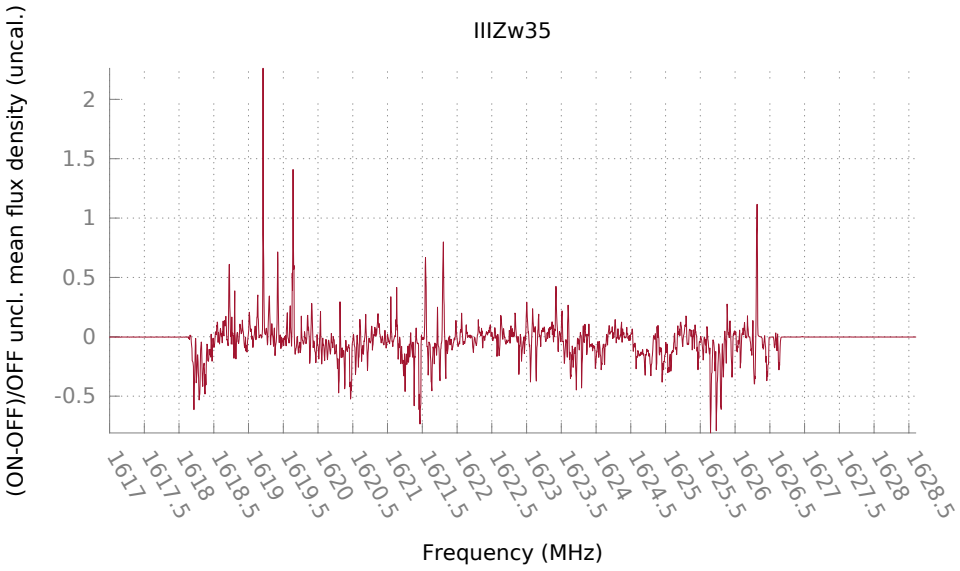


FIGURE 4.74: Non-robust unclipped $(ON - OFF)/OFF$ spectrum of the $1617 - 1628.5$ MHz band.

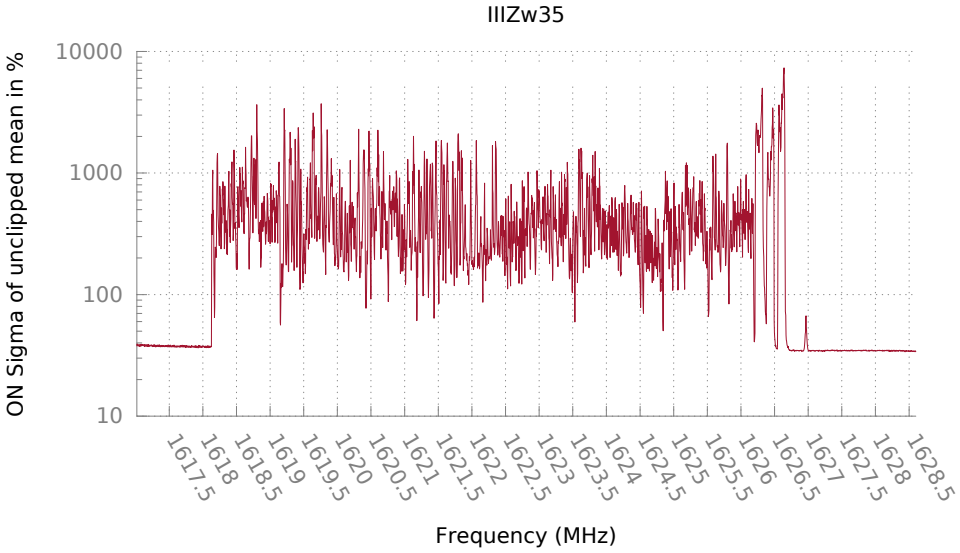


FIGURE 4.75: Temporal $\sigma$ of the unclipped $ON$ spectrum of the $1617 - 1628.5$ MHz band.

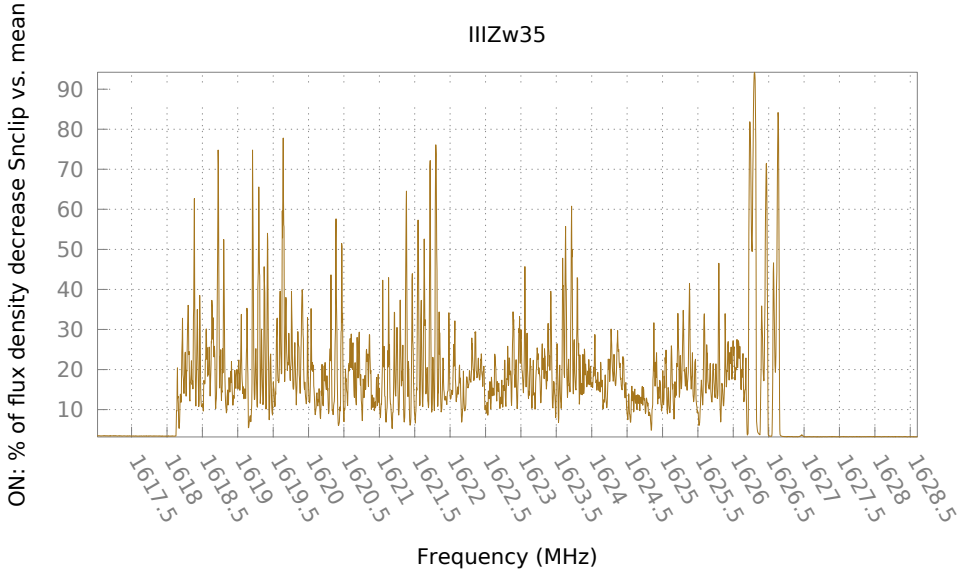


FIGURE 4.76: EoL attenuation factor (unclipped mean - $Sn$-clipped median)/($Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum in the 1617 − 1628.5 MHz frequency band.

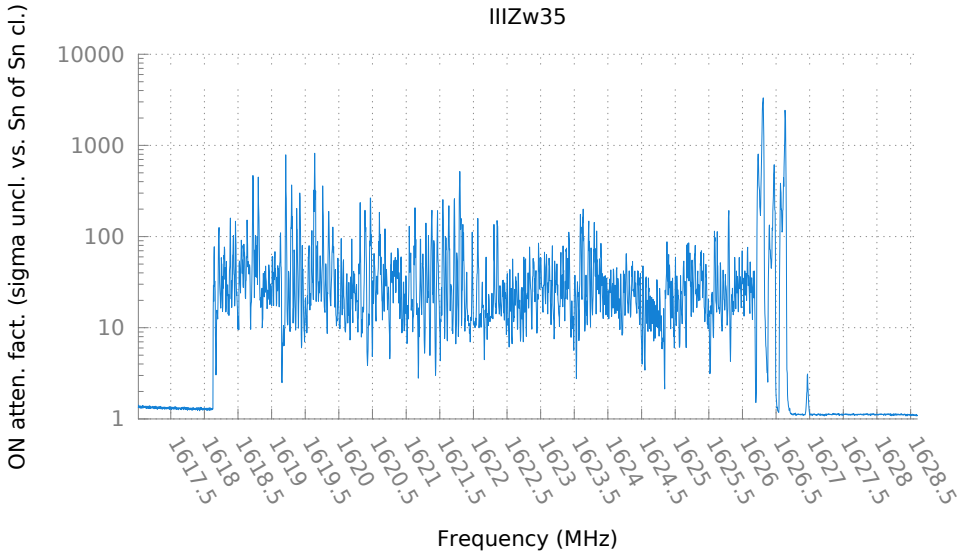


FIGURE 4.77: Temporal EoS attenuation factor ($Sn$ of $Sn$-clipped median − $\sigma$ of unclipped median)/($Sn$ of $Sn$-clipped median) of uncalibrated averaged power spectral values of the $ON$ spectrum in the 1617 − 1628.5 MHz frequency band.

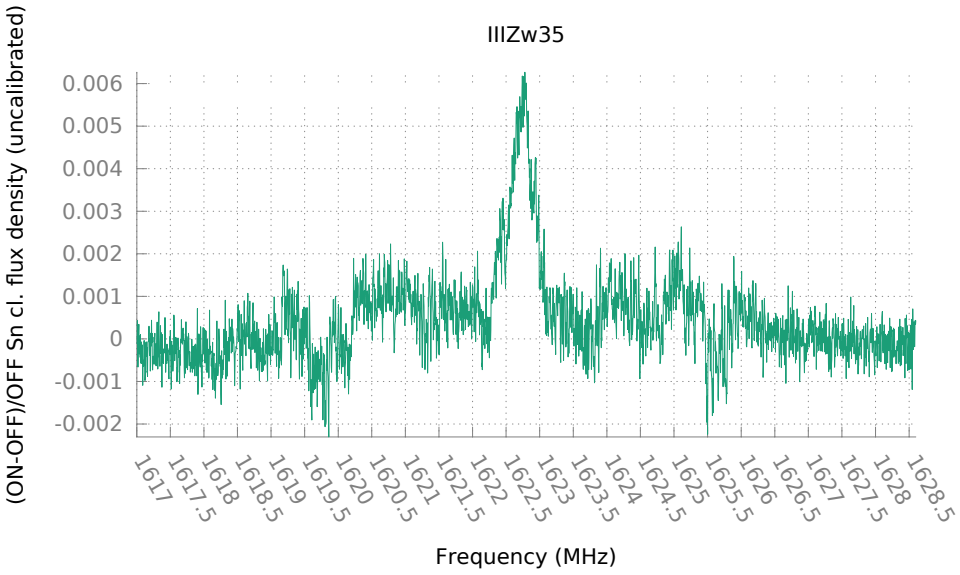


FIGURE 4.78: *Sn*-clipped $(ON - OFF)/OFF$ spectrum of the $1617 - 1628.5$ MHz band.

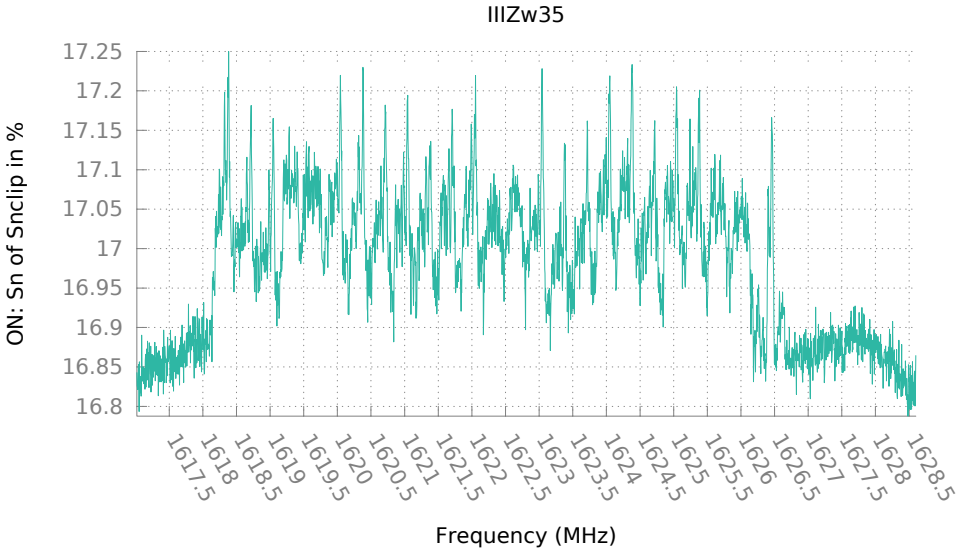


FIGURE 4.79: Temporal *Sn* of the *Sn*-clipped *ON* spectrum in the $1617 - 1628.5$ MHz band.

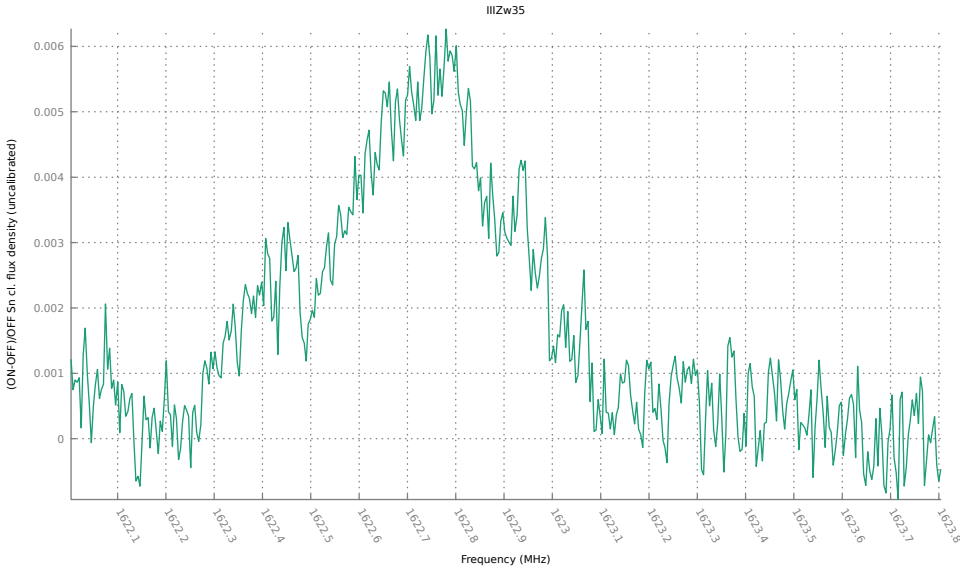


FIGURE 4.80: *Sn*-clipped $(ON - OFF)/OFF$ spectrum of the OH 1667 spectral line.

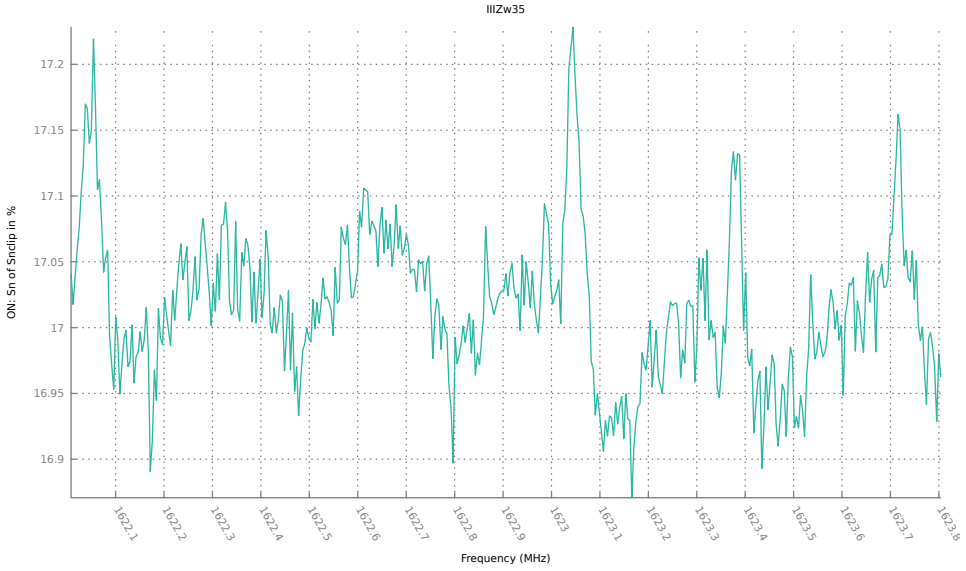


FIGURE 4.81: Temporal *Sn* of the *Sn*-clipped *ON* spectrum of the OH 1667 spectral line.

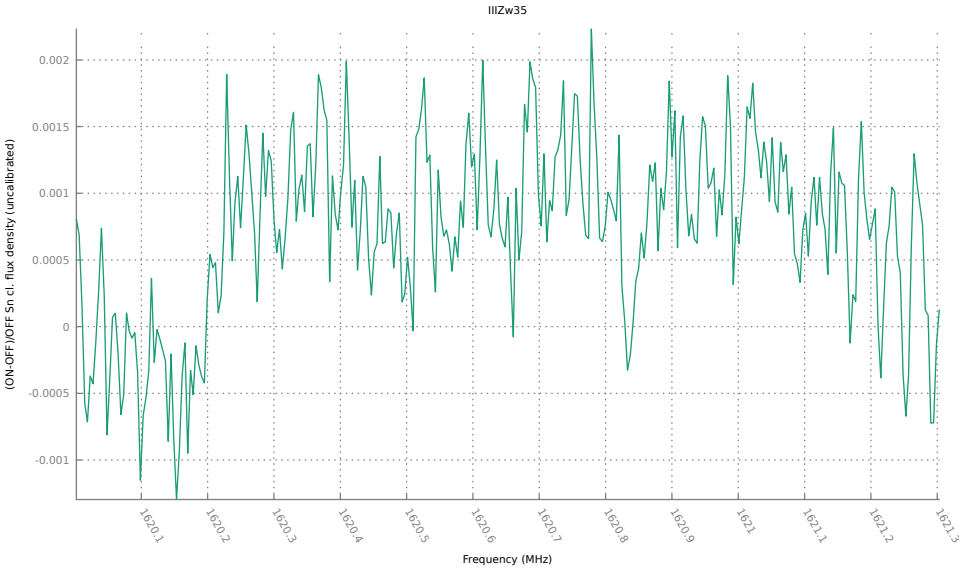


FIGURE 4.82: $Sn$-clipped $(ON - OFF)/OFF$ spectrum of the OH 1665 spectral line.

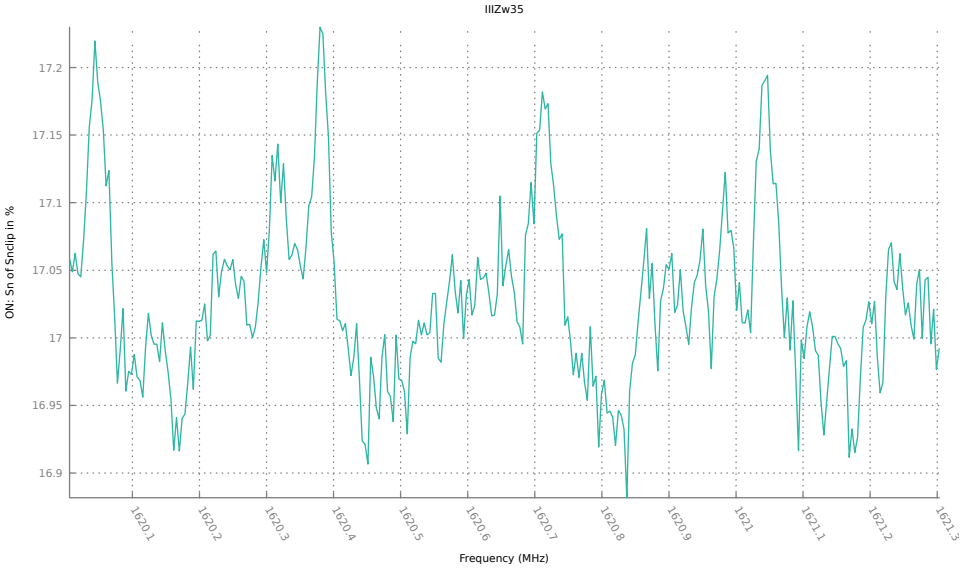


FIGURE 4.83: Temporal $Sn$ of the $Sn$-clipped $ON$ spectrum of the OH 1665 spectral line.

# Chapter 5

# Conclusion

## 5.1 Summary of developments

This thesis presents algorithms which were developed with the aim of mitigating RFI and fostering automated spectral line detection in broadband surveys. These algorithms take advantage of the cosmic source signal properties which, when captured by the radio telescope receiver and translated to real voltage samples by an analogue to digital converter (ADC), exhibit Gaussian properties, contrary to any man-made artificial signal.

These real voltage samples are converted into spectra made of complex values by Discrete Fourier Transform (DFT) processed by a spectrometer. Complex values are then multiplied by their conjugates, giving a real power spectrum (see 3.1 on p.29). A normal distribution of voltage samples does not imply that corresponding power spectral values are normally distributed, however. Indeed, in such a case, these are $\chi^2$ distributed with 2 degrees of freedom (one for the real component, one for the imaginary component). When power values are averaged $N_{scn}$ times, the resulting averaged power values samples in any frequency channel follow a $\chi^2$ law with $2N_{scn}$ degrees of freedom. By virtue of the Central Limit Theorem, a $\chi^2$ law with at least 50 degrees of freedom is considered well approximated by a normal law (see 3.2.1 on p.33).

Because of the Gaussian distribution additive properties, adding white (i.e., Gaussian) noise to the cosmic source signal gives voltage values which are still normally distributed. We deduct that any cosmic source signal observed through a white noise (but devoid of RFI) by a spectrometer whose setup allows averaged power spectral samples such as $N_{scn} \geq 25$, will produce a spectrum divided in frequency channels in which averaged power spectral values are normally distributed. Therefore, detecting and excising RFI consists of first assessing how close the sample of averaged power values is to a normal distribution in each frequency channel. Such a process is called data quality assessment.

A given frequency channel may be saturated by RFI, which means that the majority of its power values sample is made of RFI, and thus, it is not usable and should be flagged as such (though it would be possible in certain cases to retrieve good data which in this case are outliers). In real life, frequency channels are seldom completely devoid of any RFI, nor are the averaged power values exactly normally distributed, because of instrument characteristics or partial failures: indeed there always are values which lie outside the bulk of the samples. From a statistical point of view, these are called outliers (see 2.2 on p.13).

RFI signals are statistical outliers in a frequency channel if the sample of averaged spectral values in which they lie mostly consists white noise and/or cosmic source signals. One classical way of excising these unwanted values consists of recursive clipping to remove those outside a given distance to the mean of the normal distribution. For instance, recursive $3\sigma$ (standard deviation) clipping will remove $\sim$0.23% of the values which lie at more than $3\sigma$ from the center (mean) of a normal distribution. However, such process can become inefficient in the presence of strong outliers, because the mean (which is an estimator of the sample location) and the standard deviation (which is an estimator of scale of the sample) have 0 breakdown properties (see 2.2.6 on p.16): a single value which tends to infinity will lead both estimators to also tend to infinity. Moreover, because such outliers have a strong leverage effect on such estimators of location ($EoL$) and scale ($EoS$), they may mask the presence of subsamples of interest such as spectral lines (see 2.2.7 on p.16).

In order to circumvent this difficulty, the algorithms developed in this Thesis use robust estimators, which are immune to outliers as long as their contribution to the sample is less than 50%. Part of this work consisted of choosing appropriate robust estimators of location (studied in 2.3 on p.18) and scale (see 2.4 on p.19), i.e., unbiased (see 2.2.3 on p.14), efficient (see 2.2.4 on p.15) and convergent. The selected $EoL$ is the median. As for the robust $EoS$, several solutions were studied and notably:

- the Median Absolute Deviation ($MAD$ - see 2.4.2 on p.20);
- $Sn$ (see 2.4.3 on p.21);
- $Qn$ (see 2.4.4 on p.22)

While $MAD$ is used in several post-processing software packages (see 5.4 on p.134), it has severe limitations which are surprisingly ignored by post-processing software designers: it is unbiased only if the sample has a nil skewness, i.e., if symmetrically dispersed around its median (see 2.4.2.2 on p.21). However, since $MAD$ is algebrically proportional to the non-robust $\sigma$ (see Eq. 2.27 on p.20), the $\sigma/MAD$ ratio is a good estimator of the sample normal distribution properties as well as the presence of outliers. Although $Qn$ is unbiased only for large samples, the most efficient of all $EoS$ and insensitive to sample asymmetry, it is very CPU time-consuming. Therefore it can only be calculated with less than 10 000 values. And finally, $Sn$ is unbiased, more efficient than $MAD$, also insensitive to sample asymmetry (see 2.4.3 on p.21), and less CPU-time consuming than $Qn$. All these robust estimators are implemented in the ROBEL post-processing software (see below), but the only RFI excision results presented in the case studies of Chapter 4 are $Sn$-clipped based.

In case only relatively few RFI signals (even if extremely strong) are polluting a normal distribution in a given frequency channel, robust $EoL$ and $EoS$ will still be relevant indicators of a normal distribution since they will stay unaltered. In these conditions, not only divergent non-robust vs. robust $EoL$ and $EoS$ of the same sample will reveal the presence of outliers such as RFI, but recursive clipping using these robust estimators will be an efficient method to excise these unwanted signals.

Data quality assessment as well as RFI mitigation is performed on a 2D matrix which consists of frequency channel columns and time-line series of averaged power spectral values (see 3.2.3 on p.36). Each frequency channel is processed individually. For data quality assessment, skewness (see 2.2.8.1 on p.16) and kurtosis (see 2.2.8.2 on p.17) of averaged power values are calculated (for a normal distribution, skewness is nil and kurtosis equals 1). Then the standard deviation is compared to its robust

equivalent, the Median Absolute Deviation (MAD). For a normal distribution, they are algebrically equal. Any outlier will make these three data quality estimators diverge from their reference value. Thus, by setting limits to these estimators, it is possible to flag each frequency channel as reliable or not. After recursive clipping aimed at getting rid of unwanted outliers (see 3.5.1.5 on p.50), a new data quality assessment provides an evaluation of the cleaned averaged power value samples in each frequency channel. Therefore, for each frequency channel, we get non-robust and robust $EoL$ and $EoS$ as well as data quality estimators, both before and after clipping. The output consists of 1D spectra calculated with different estimators, and this allows their comparison.

Baseline fitting on resulting 1D spectra can be reliably processed to perform automated line detection on broadband surveys, assuming that unreliable frequency channels are ignored (see 3.5.2.2 on p.51). Spectral lines are statistical outliers if the majority of frequency channels $EoL$ are representative of the cosmic background, because they deviate from the bulk of spectral values. By using a robust version of the Least Squares polynomial regression for baseline fitting (i.e., the Least Trimmed Squares - LTS, see 2.5.2 on p.24), it becomes possible to perform automated blind detection of spectral lines inside a considerable amount of 1D data. Therefore, such an algorithm is a time-saver for large and/or broadband surveys.

## 5.2 The ROBEL post-processing software

The ROBEL post-processing software was written to develop, test and implement the aforementioned algorithms using the WIBAR broadband spectrometer installed on the Nançay Radio Telescope (NRT) in France. Its generic architecture allows the processing of any 2D time-lines series of power spectra from a single-dish radio telescope and would only require writing an adapted input module for other spectrometer data formats (so far FITS files were used). ROBEL is easily portable to any Linux calculator with gfortran, OpenMP, cfitsio and pgplot libraries installed. After years of development, ROBEL has become a dashboard which provides extensive information and allows us to:

- monitor the entire instrumental chain, from the receiver to the spectrometer;
- identify, analyze and excise RFI either ground-based (e.g., radars) or mobile (notably GNSS and Iridium satellite constellations);
- assess observation data quality before and after RFI mitigation;
- detect elusive spectral lines hidden by RFI;
- process line profiles with the full dataset of time-wise averaged power spectra;
- detect automatically spectral line candidates in broadband spectra.

This software package is stable and essentially needs the development of a friendly graphic user interface (GUI).

## 5.3 Case studies results

The 1164 − 1385 MHz band is highly polluted by radionavigation (RN) radars as well as Global Navigation Satellite System (GNSS) RFI. Reopening this band would offer the opportunity to observe many galaxies which are readily detectable with a 100m-class telescope such as the NRT.

The observations presented in this thesis were essentially a test bed for RFI mitigation. The first object targeted was QSO B0738+313 with two previously documented Damped Lyman Alpha (DLA) HI absorption lines in the intergalactic medium along the line of sight towards this quasar: one at $z = 0.0912$ (named B1) and the other at $z = 0.2212$ (named B2). Since both lines were observed previously at Arecibo without the presence of RFI by Kanekar et al. (2001), Lane et al. (1998), and Lane et al. (2000)), it was possible to use these observations as a benchmark for data obtained with WIBAR at the NRT and post-processed by ROBEL. Although the total 1.7 hours of integration time at the NRT was only a fraction of that at Arecibo, both B1 and B2 were detected.

At the NRT there were RFI signals near B1, but the line was automatically detected with blind LTS baseline fitting before and after clipping on the $ON$-position 1D spectrum. Table 4.1 on p.78 shows that the $FWHM$ closest to the Lane et al. (2000) Arecibo value is given by the $Sn-$clipped $(ON - OFF)/OFF$ spectrum. This table also compares error margins calculated on the one hand using 1D spectra according to methods described by Schneider et al. (1986), Schneider et al. (1990), and Thuan et al. (1999), and on the other hand by ROBEL with the on $[EoL - 3EoS, EoL + 3EoS]$ 99.74% confidence interval applied to time-series of averaged power spectra in each frequency channel (see 4.1.5.1 on p.68). Indeed, while Lane et al. (2000) at Arecibo found a B1 $FWHM$ of $3.69 \pm 0.02$ $\mathrm{km\,s^{-1}}$, the calculation from the 1D spectrum gave a rather uncertain result of $3.65 \pm 1.54$ $\mathrm{km\,s^{-1}}$, whose precision was greatly improved with ROBEL algorithms to $3.65 \pm 0.19$ $\mathrm{km\,s^{-1}}$. Thus, the B1 case study demonstrated the ability to improve the precision of measurements with the help of robust clipping of time-series of averaged spectral values, even in the absence of strong RFI: error margins greatly decreased (divided by a 8 factor on the $FWHM$ observed on the $Sn-$clipped $(ON - OFF)/OFF$ spectrum).

The B2 line presented an ideal case to study RFI mitigation, since the spectral line occurs right in the middle of a strong artificial signal. The NRT/WIBAR observation post-processed by ROBEL did not show any spectral line before robust clipping (see Fig. 4.36 on p.90). Though swamped by RFI, their excision made B2 clearly visible (see Fig. 4.49 on p.100) and it was automatically detected by robust LTS baseline fitting of the $Sn-$clipped $ON$ 1D spectrum. However, the short observation time and the correspondingly high *rms* level did not allow the determination of relevant results for all line parameters (for instance the $V_{peak}$ error of the $ON$ spectrum seems inconsistent). Nevertheless, the $FWHM$ is still comparable to the one measured by Kanekar et al. (2001) at Arecibo (see 4.2 on p.93).

The observation of the III Zw 35 Luminous Infrared Galaxy (LIRG) which contains an OH megamaser offered the opportunity to test the RFI mitigation algorithms with ROBEL under extreme conditions (i.e., strong intensity and high occurence). When first observed by Staveley-Smith et al. (1987) and then by Martin (1989), its OH 1665/1667 MHz spectral lines (redshifted to around 1621 MHz) were devoid of RFI. Starting in 1998, the Iridium satellite constellation created such strong RFI that

it swamped these lines which were easily observable before. To get rid of artificial signals which were up to 5 000 times more powerful than the cosmic background (see Fig. 4.77 on p.125), WIBAR was set up for a much shorter integration time to minimize possible RFI residuals in averaged power spectral values. This resulted in only 40 degrees of freedom for the $\chi^2$ law they followed, but that was still considered acceptable to apply the Central Limit Theorem (the lowest acceptable limit being 30, while most of the specialized literature set it to 50). Nevertheless, the results were rewarding, since the OH 1665/1667 MHz spectral line profiles calculated from resulting 1D $Sn-$clipped $(ON-OFF)/OFF$ spectra by ROBEL were similar to those originally observed without RFI (see Fig. 4.68 on p.115, 4.69 on p.116, 4.80 on p.127, 4.82 on p.128, and Table 4.4 on p.123). This case study demonstrates the ability of the algorithms applied by ROBEL to excise RFI in extreme conditions.

The broadband observation of the quasar B0738+313 also gave the opportunity to study the effects of robust clipping on various kinds of RFI signals in the range $1160-1385$ MHz for a typical velocity resolution used for extragalactic line surveys, $B = 10$ km s$^{-1}$. RFI mitigation was applied to a radar signal in the $1324-1332$ MHz band (see Fig. 4.51 on p.102, 4.59 on p.106, and 4.60 on p.107) the intermittent GPS L3 signal centered on 1381.05 MHz, GPS L5 centered on 1176.45 MHz (see Fig. A.3 on p.139), Galileo E6 centered on 1278.75 MHz (see Fig. A.7 on p.142), and Beidou B3 centered on 1268.52 MHz (see Fig. A.14 on p.146) which is the most powerful GNSS signal in this band (see Fig. 4.61 on p.108). The results summarized in Table 4.3 on p.113 are promising. After robust $Sn-$clipping, the kurtosis of averaged power values in frequency channels was reduced by 96 to 99.99% for all these RN/GNSS RFI. As For the 1D spectra, the frequency *rms* around the studied radar was reduced by 87% to an equivalent of 6 mJy for a frequency channel width $B = 10$ km s$^{-1}$. Whereas in the vicinity of the GPS L5 and Beidou B3 centers, the frequency *rms* was barely or not reduced, robust clipping of the latter just 3 MHz above the center resulted in a significant improvement ($-39\%$ on the *rms*) to an equivalent of 5 mJy at the same $B = 10$ km s$^{-1}$. The GPS L3 signal *rms* was decreased by 41% to an equivalent of 7 mJy with $B = 10$ km s$^{-1}$. The Galileo E6 signal was also clipped, including its most difficult part to process, i.e., its band center, and returned an *rms* reduced by 51% also to an equivalent of 5 mJy. Overall, these preliminary tests on RN/GNSS RFI mitigation demonstrated the ability of the robust clipping algorithms to significantly reduce *rms* over most of the studied band. The *rms* of these short observations is about twice what it would be without RFI, but the reduction of the *rms* level as function of integration time has not yet been studied. Therefore, extensive test observations will be conducted, this time with actual $B\sim10$ km s$^{-1}$ and with spectral lines in the line of sight.

These non-exhaustive case studies illustrated the ability of the algorithms developed in this Thesis and embedded into the ROBEL software to reveal spectral lines swamped by strong RFI and also to improve the precision of measured spectral line parameters. Also most of the selected GNSS/RN RFI signals were excised, resulting in an *rms* noise level reduced to a few mJy. Such values are still higher than those in frequency bands devoid of RFI, but future tests with longer integration times will explore the limits of such an exercise.

## 5.4 Comparison with other post-processing software

There are many attempts and solutions proposed for RFI mitigation[1]. Few have taken advantage of time-series though (see for instance Nita et al. (2007)). A significant part of post-processing software process RFI mitigation in the frequency domain, either on 1D spectra or in 2D interferometer maps. While it is not possible at this stage to assess relative performance (a benchmark would require the same observation on the same instrument to be post-processed with different packages, and the current ROBEL version is written for single-dish radio telescopes only), it is interesting to compare ROBEL with a few existing solutions, either on single-dish telescopes or interferometers.

For example, to search for HI emission lines in gravitationally lensed galaxies at $z\sim0.4$ using the Green Bank Telescope (GBT) in spectral bands polluted by RFI, Hunt et al. (2016) developed two RFI excision algorithms. Wide-band RFI signals were removed in the Fourier domain, whereas narrow-band RFI were mitigated with the help of a non-robust $4.3\sigma$-clipping applied in the frequency domain to 1D spectra. The total observation time was quite long (more than 77 hours in cumulated *ON* and *OFF* source positions) and the integration time $\tau$ was limited to 2s by the GBT Spectral Processor which was from an old generation. While these algorithms produced satisfactory spectral outputs in general, they were unable to remove RFI caused by a Distance Measurement Radar (DME), which emits between 24 to 30 pulses a second between 962 and 1024 MHz. It is difficult to compare these results with those of WIBAR/ROBEL since:

1. the WIBAR integration time $\tau$ can be set much lower (e.g. 4.67 ms for the III Zw 35 observation - see 4.2.3 on p.117) in order to minimize averaged power values contaminated by RFI, the main limit being the degrees of freedom of the $\chi^2$ law that the averaged power spectral values must follow for the ROBEL algorithms to apply;
2. there is no DME RFI at NRT, but a TACAN (TACtical Air Navigation) radar system around 1 GHz, which provides more extensive information to aircraft than the DME. Observations including TACAN RFI were performed with WIBAR, but at the time of this writing, not enough time could be devoted to an exhaustive analysis. Nevertheless, a first inspection of 1D spectra shows that ROBEL excised at least 90% of related RFI, but whether this makes corresponding frequency channels usable for spectral line detection has yet to be proven. Furthermore, several WIBAR setups will be tried and results will be compared.

Because ROBEL has a modular architecture, it would be easy to write a data input interface compatible with the FITS format used at the GBT and this would allow the processing of the DME signal from an observation lasting at least 10 to 15 minutes but with the new GBT spectrometer.

DUCHAMP (Whiting, 2012) is a 3D (as well as 2D and 1D) source-finder designed for interferometers aimed at HI sources. Its detection algorithm, which notably uses the *MAD EoS*, searches for signals above a signal-to-noise ratio. Though its use of *MAD* was seminal for this Thesis as it introduced the notion of robust estimators in radio astronomy, its results could not yet be compared with those of ROBEL since

[1] The RFI 2019 conference "Coexisting with Radio Frequency Interference" provides an overview of recent contributions

that would require the development of a 3D version of the latter, which remains to be implemented.

SOFIA (Serra et al., 2015) is a modular 3D source-finder aimed at supporting large HI surveys with the ASKAP interferometer (Serra et al., 2015). It produces two cubes and two masks:

- a data cube which includes instrument noise and cosmic source signals;
- a weights cube which the observer assigns to voxels;
- a binary mask for detection and non-detection flagging;
- an object mask to assign object indexes to detected voxels.

Among the *EoS* it uses are non-robust standard deviation, $MAD$ and a Gaussian fit. Methods of denoising are applied, including a 3D kernel and 2D-1D wavelet filtering. Several methods of spectral lines detection are implemented, including thresholding techniques, which will not be detailed here, though one is similar in principle to ROBEL automated detection algorithm (i.e., flagging voxels which exceed a multiple of the *rms*). SOFIA and ROBEL cannot be directly benchmarked since the latter can not yet process data cubes. Moreover, ROBEL algorithms rely on time-series of averaged power spectra, which SOFIA does not handle. As for the solutions described above, detection and RFI mitigation techniques apply in the absence of any chronological data series. It should also be noted that interferometers are less susceptible to RFI signals than single-dish radio telescopes. In any case, as stated above, the results of this Thesis advocate for a substitution of $MAD$ by the $Sn$ estimator of scale.

## 5.5 Foreseen developments

The case studies presented in this Thesis will be completed with extra observations for:

- B0738+313: extra time is needed to decrease *rms*, especially for the B2 spectral line;
- III Zw 35: some upcoming observation tests will be conducted to reach a better performance, this time with $\chi^2$ law of more than 50 degrees of freedom;
- RN radars/GNSS: Extensive test observations will be conducted, this time with actual $B{\sim}10$ km s$^{-1}$ and with spectral lines in the line of sight.

At the Nançay Radio Telescope time has been allocated for pilot projects involving WIBAR and ROBEL in the first semester 2020, with a view towards starting larger surveys, such as:

1. For HI absorption line searches, ROBEL will be used to (re)analyze NRT data obtained for HI absorption line searches for gas in samples of galaxies and clusters out to $z \sim 0.23$, i.e., down to the practical frequency limit of the telescope (1100 MHz). Most spectra are heavily contaminated by different types of RFI (see Figure 1.1) which ROBEL has been able to mitigate in the data shown in this Thesis;

2. For HI emission line surveys, ROBEL will be used to reopen the radial velocity range beyond 7500 $\mathrm{km\,s^{-1}}$ (1385 MHz) at the NRT, which is covered by different types of RFI. This is of particular interest for galaxy searches in the Zone of Avoidance, where this velocity range is crucial for mapping the overall galaxy and mass density distribution in the local Universe;
3. Systematic measurements and analysis of *rms* noise levels and RFI throughout the L band before and after $Sn$-clipping, in order to assess the feasibility of various kinds of surveys in the band;
4. ROBEL can also be used to remove RFI from broadband continuum observations, for example for monitoring of Gamma Ray Bursts and variable AGN sources.

The algorithms developed in this thesis and embedded in ROBEL are generic spectral techniques and should in principle also work for phased arrays interferometers (tied-array as well as phased array beams).

Since SOFIA is foreseen as a future post-processing package for SKA Precursor instruments, one of the possible ROBEL developments is writing a module complementary to SOFIA which would deliver data cubes with RFI mitigated as well as weights and binary mask cubes with reliability and detection flagging.

# Appendix A

# Global Navigation Satellite Systems (GNSS)

## A.1 Definition of GNSS

Global Navigation Satellite Systems (GNSS) consist of constellations of about 25 satellites which provide autonomous geo-spatial positioning with a global coverage. The satellites are spread on medium Earth Orbits (MEO) with altitudes of about 20000 km, with an orbital period of approximately 12 hours. They broadcast signals which carry information on their orbital position and precise time of emission.

## A.2 GNSS in Operation

The main GNSS in operation so far are:

- GPS, for Global Positioning System (GPS), United States;
- GLONASS, for GLObal NAvigation Satellite System, Russia;
- Galileo, European Union;
- BDS, for BeiDou Navigation Satellite System/BEIDOU, China.

## A.3 GNSS Frequency Bands

### A.3.1 Global Positioning System (GPS)

The GPS transmitted navigation signal is essentially a bipolar phase-shift key (BPSK) waveform (i.e., the phase is modulated between sine and cosine)[1] carried with Code Division Multiple access (CDMA). It is Right Hand Circulary Polarized (RHCP).

Of the GPS frequency bands three are available for civilian use. They are centered on the carrier frequencies listed below, with an ITU authorized bandwidth of $\pm$12MHz:

- L1 : 1575.42 MHz (see Fig. A.1);
- L2 : 1227.6 MHz (see Fig. A.2);

[1] Though some are Binary Offset Carrier - BOC, or Time Multiplex BOC - TMBOC. For more details, refer to: https://gssc.esa.int/navipedia/index.php/Time-Multiplexed_BOC_(TMBOC)

- L3 : 1381.05 MHz
- L4 : 1379.9133 MHz
- L5 : 1176.45 MHz (see Fig. A.3);

L3 (used by the Nuclear Detonation - NUDET - detection system payload - NDS) and L4 (being studied for ionospheric correction) are of special concern for radio-astronomers because many galaxies emit HI lines around these frequencies. RFI from L3 is regularly seen in NRT spectra[2]

The GPS M-CODE is a split spectrum signal with a 30 MHz bandwidth centered on L1 and L2 bands.

The following figures are scaled in Power Spectral Density (PSD) which is the Fourier transform of the auto-correlation function of the signal. For explanations on In-Phase and Quadra-Phase PSD, refer to the ESA Navipedia webpage about Coherent Adaptive Sub-Carrier Modulation (CASM) and Interplex on : `http:https://gssc.esa.int/navipedia/index.php/Coherent_Adaptive_Sub-Carrier_Modulation_(CASM)_and_Interplex`

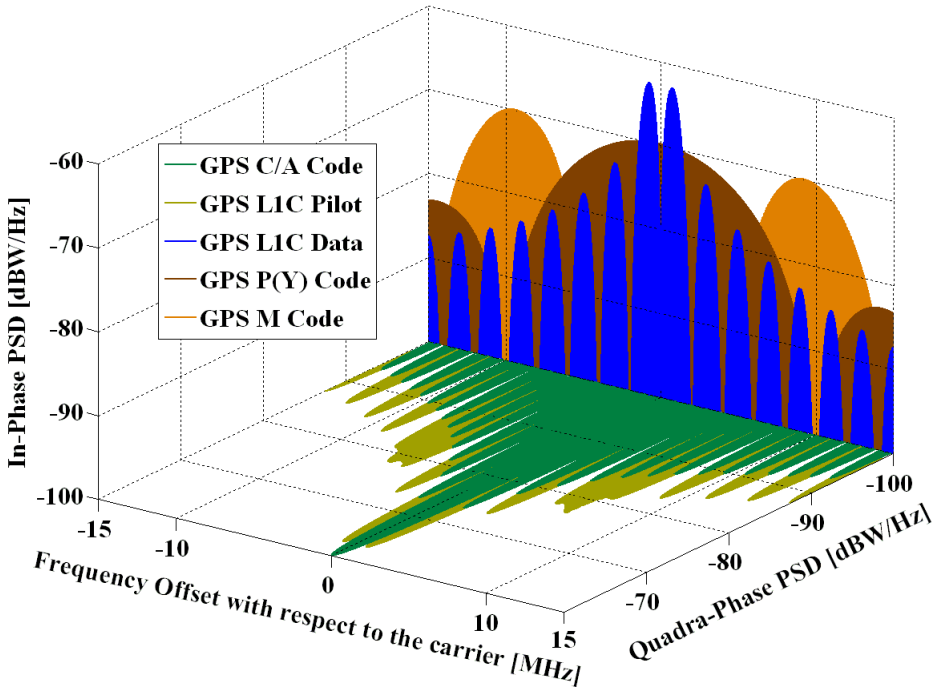


FIGURE A.1: L1 GPS frequency plan (center carrier frequency: 1575.42 MHz). Source: ESA Navipedia.

[2]van Driel, W. 2009, Keeping our windows on the radio Universe clean, Proceedings of the Second Marie Curie MCCT-SKADS Training School `http://pos.sissa.it/archive/conferences/065/017/2nd%20MCCT-SKADS_017.pdf`

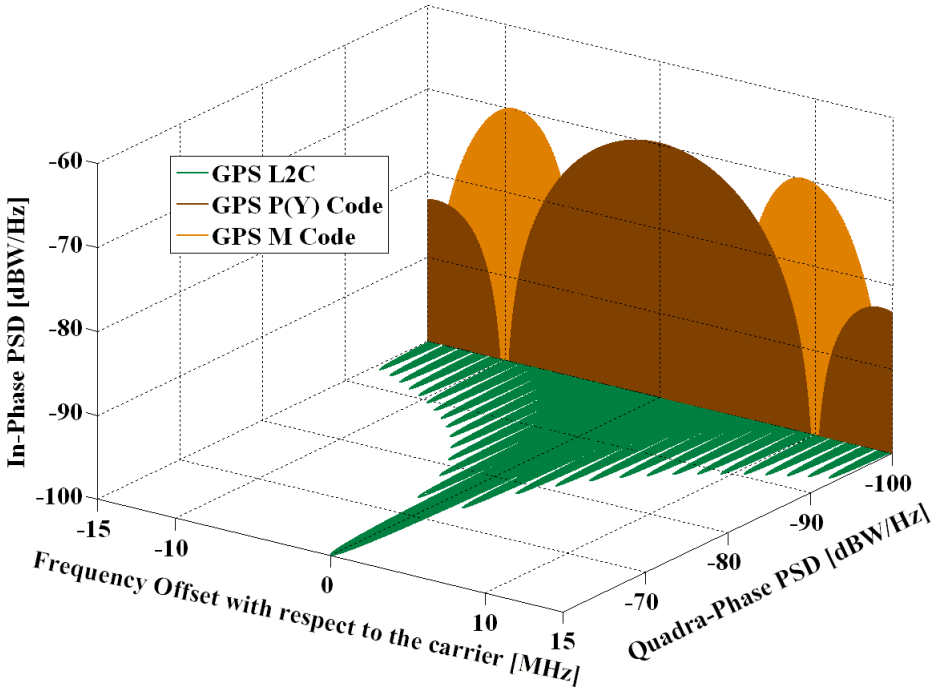


FIGURE A.2: L2 GPS frequency plan (center carrier frequency: 1227.6 MHz). Source: ESA Navipedia.

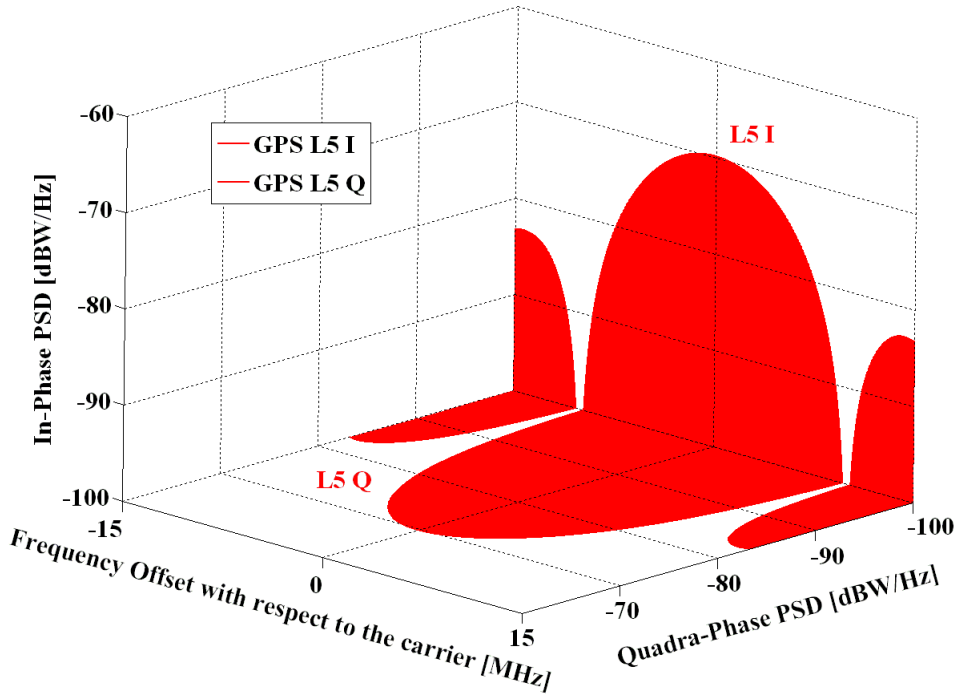


FIGURE A.3: L5 GPS frequency plan (center carrier frequency: 1176.45 MHz). Source: ESA Navipedia.

### A.3.2 Galileo

The Galileo transmitted navigation signal consists of different versions of Binary Offset Carrier (BOC) with Code Division Multiple access (CDMA). It is Right Hand Circulary Polarized (RHCP).

Galileo uses four frequency bands centered on:

- E1: 1575.42 MHz (see Fig. A.4);
- E2: 1561.098 MHz (no figure available);
- E5: 1191.795 MHz (see Fig. A.5, and Fig.A.6 for GPS and Galileo overlapping);
- E6 1278.75 MHz (see Fig. A.7).

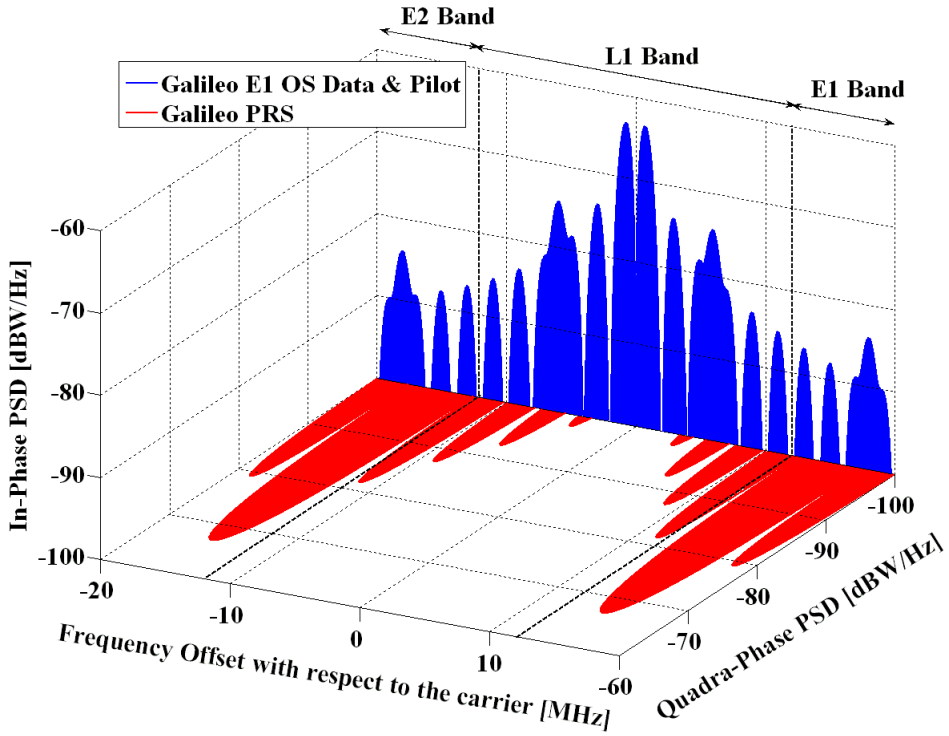


FIGURE A.4: E1 Galileo frequency plan (center carrier frequency: 1575.42 MHz). Source: ESA Navipedia.

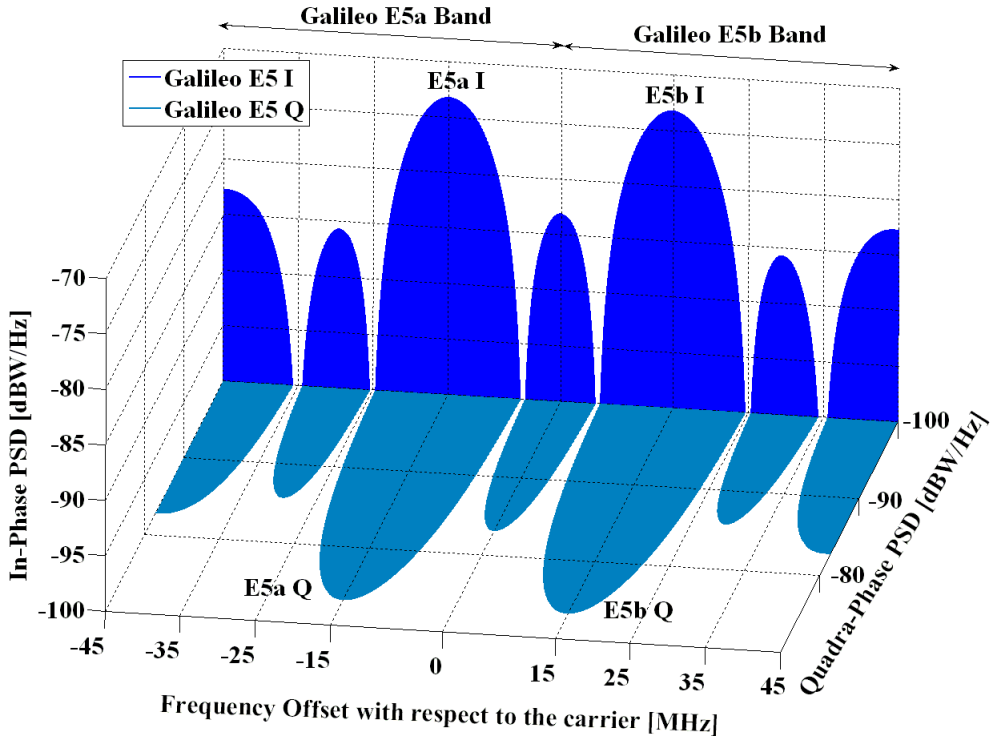


FIGURE A.5: E5 Galileo frequency plan (center carrier frequency: 1191.795 MHz). Source: ESA Navipedia.

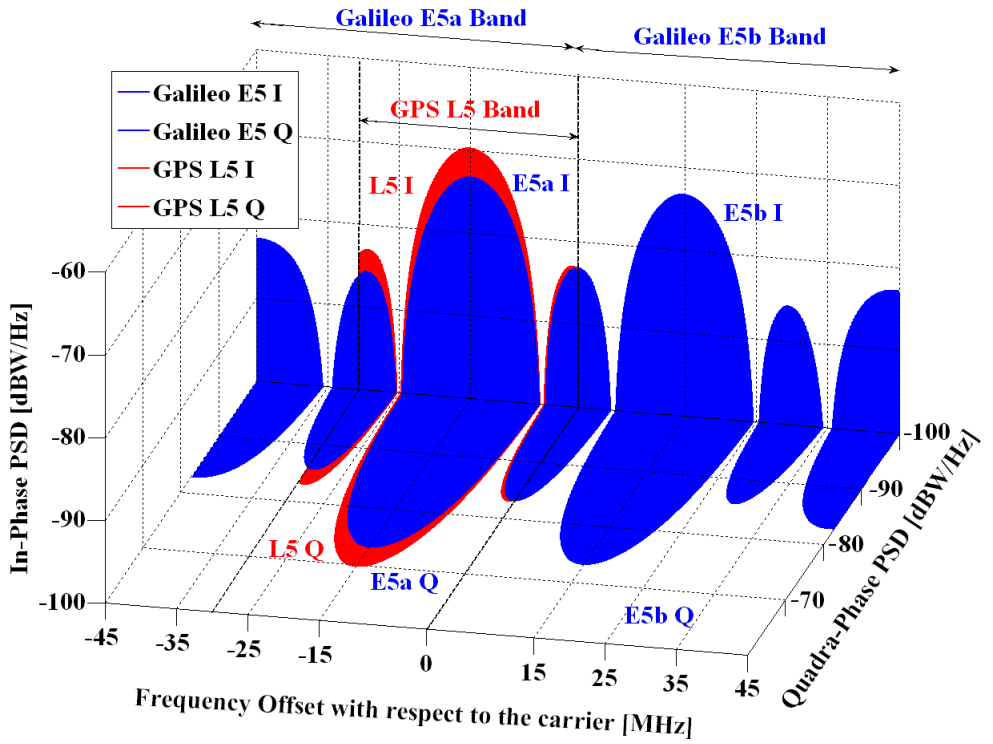


FIGURE A.6: E5 GPS Galileo frequency plan overlapping (center carrier frequency: 1191.795 MHz). Source: ESA Navipedia.

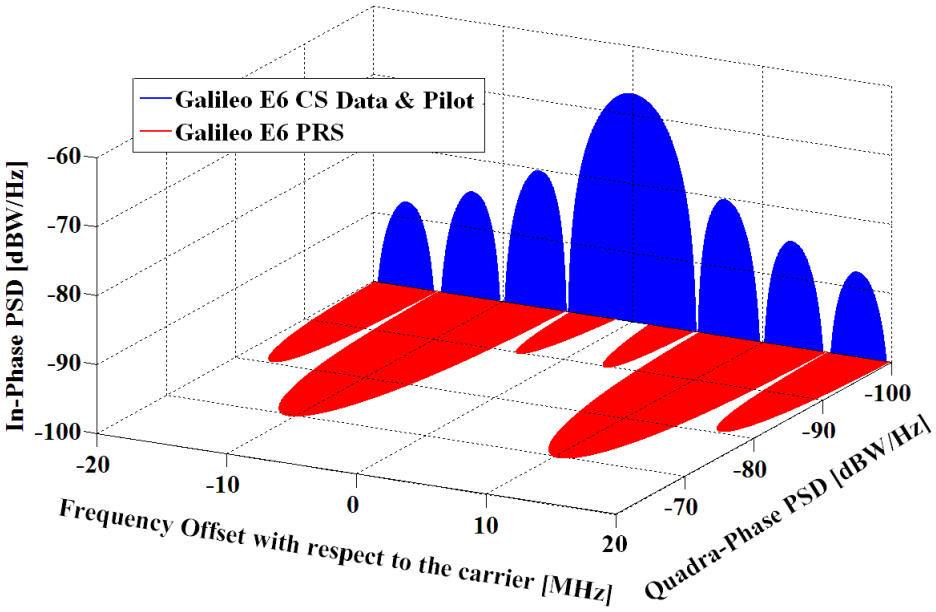


FIGURE A.7: E6 Galileo frequency plan (center carrier frequency: 1278.75 MHz). Source: ESA Navipedia.

### A.3.3 GLONASS

The GLONASS transmitted navigation signal is a bipolar phase-shift key (BPSK) waveform (i.e., the phase is modulated between sine and cosine) carried with Frequency Division Multiple Access (FDMA).

GLONASS uses two bands (L1 and L2) and there are two options for L3:

- L1: 1602 MHz (see Fig. A.8);
- L2: 1246 MHz (see Fig. A.9);
- L3: 1201 MHz option 1 (see Fig. A.10) and option 2 (see Fig. A.11).

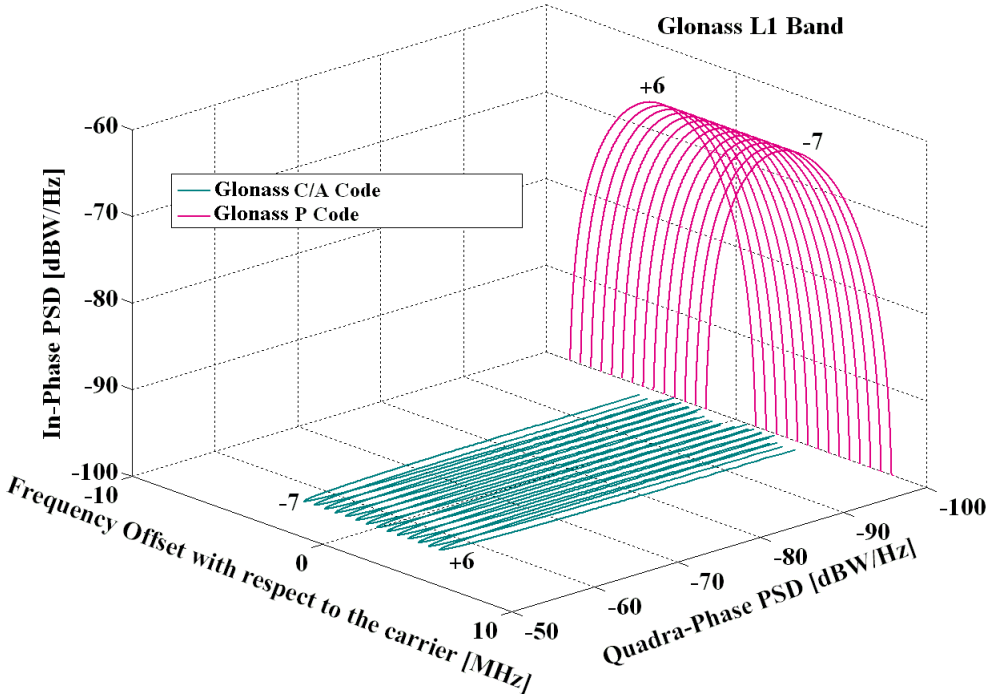


FIGURE A.8: L1 GLONASS frequency plan (center carrier frequency: 1602 MHz). Source: ESA Navipedia.

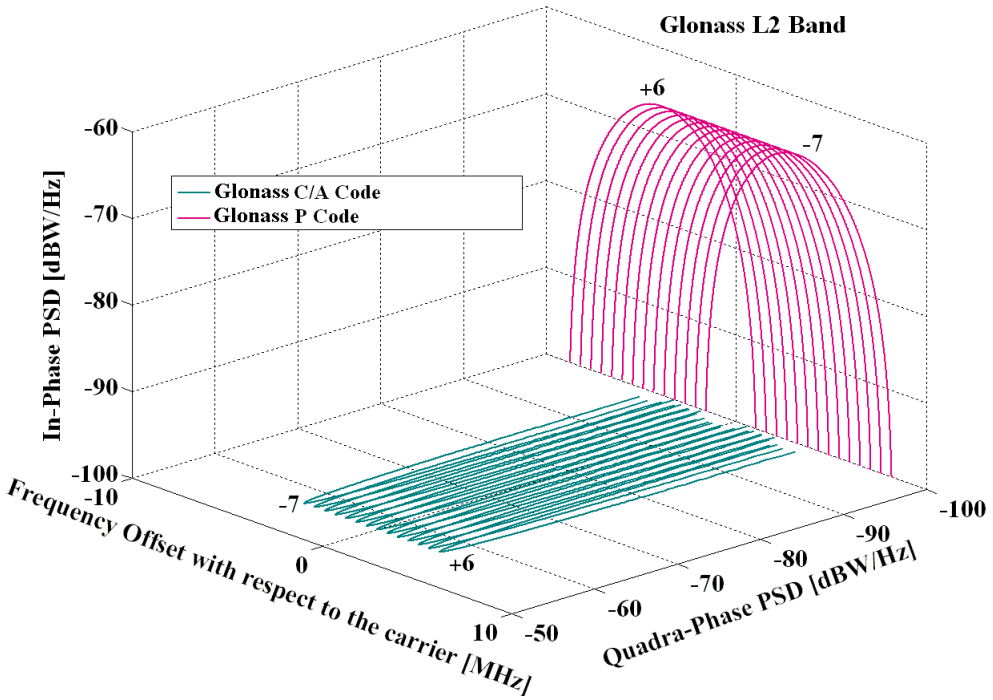


FIGURE A.9: L2 GLONASS frequency plan (center carrier frequency: 1246 MHz). Source: ESA Navipedia.

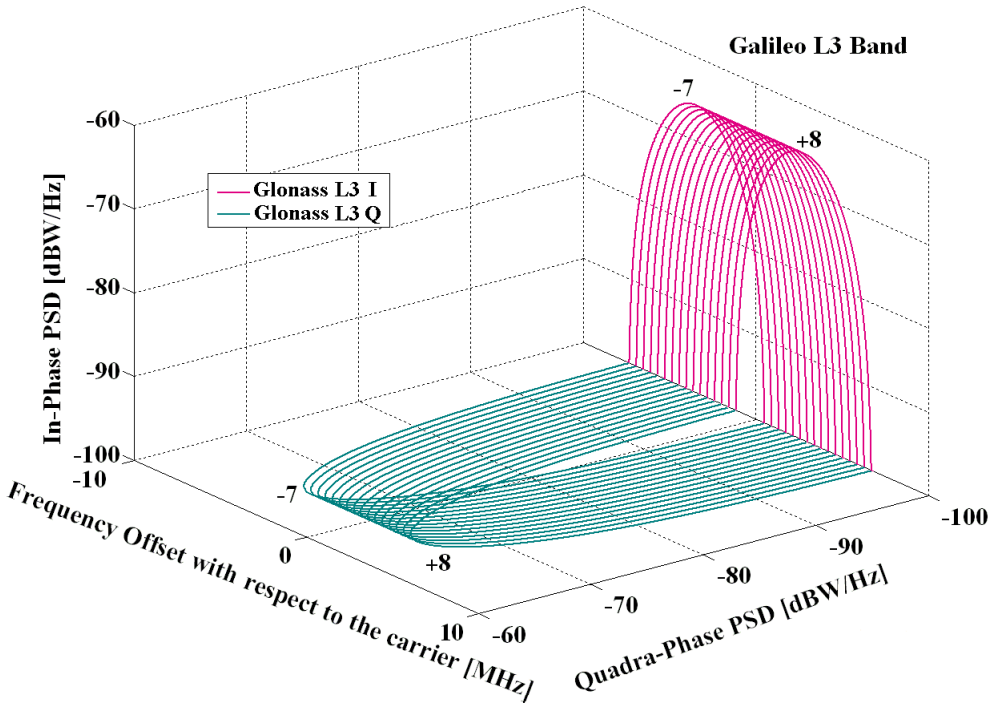


FIGURE A.10: L3 GLONASS frequency plan option 1 (center carrier frequency: 1201 MHz). Source: ESA Navipedia.

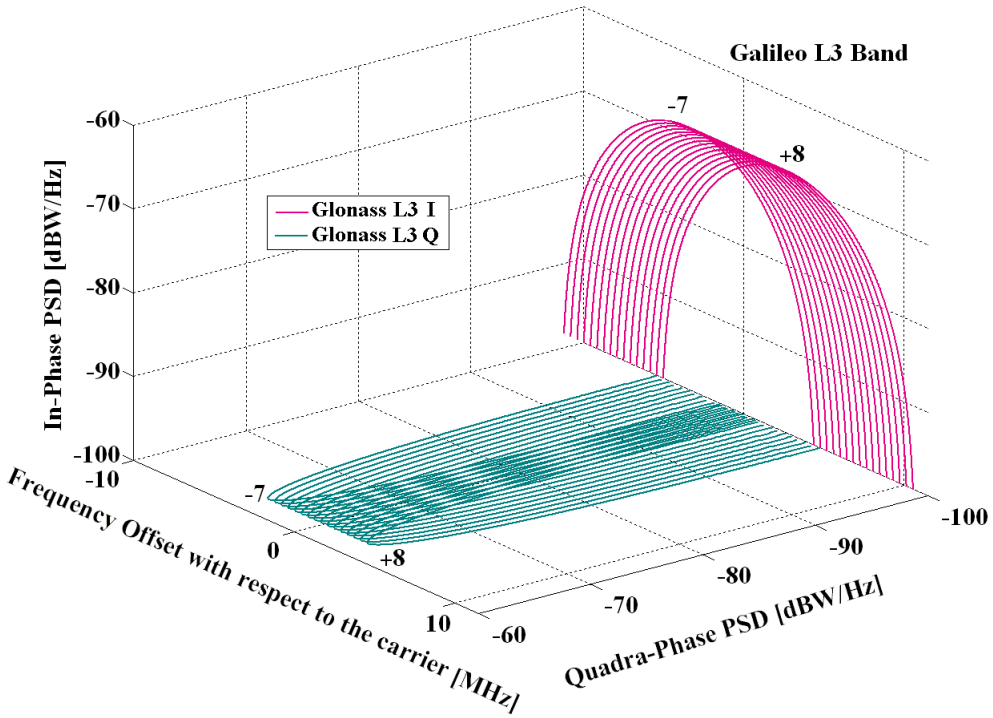


FIGURE A.11: L3 GLONASS frequency plan option 2 (center carrier frequency: 1201 MHz). Source: ESA Navipedia.

### A.3.4 BEIDOU

The COMPASS/BeiDou signal is a quadriphase PSK (Quadrature phase-shift keying - QPSK) waveform carried with Code Division Multiple access (CDMA).

COMPASS/BeiDou uses the following bands:

- COMPASS CPII/Beidou-B1 1561.098 MHz with a bandwidth of $\pm 4.092 MHz$ (see Fig. A.12;
- COMPASS CPII/Beidou-B1-2 1589.74 MHz with a bandwidth of $\pm 4.092 MHz$;
- COMPASS CPII/Beidou,B1-BOC 1575.42 MHzwith a bandwidth of $\pm$16.368MHz;
- COMPASS CPII/Beidou-B2 1207.14 MHz with a bandwidth of $\pm$24MHz (see Fig. A.13;
- COMPASS CPII/Beidou,B2-BOC 1207.14 MHz with a bandwidth of $\pm$5.115MHz;
- COMPASS CPII/Beidou,B3-BOC 1268.52 MHz with a bandwidth of $\pm$35.805MHz (see Fig. A.14;
- COMPASS CPII/Beidou,B3 1268.52 MHz with a bandwidth of $\pm$24MHz;
- COMPASS CPII/Beidou: L5 1176.45 MHz with a bandwidth of $\pm$24MHz (no figure).

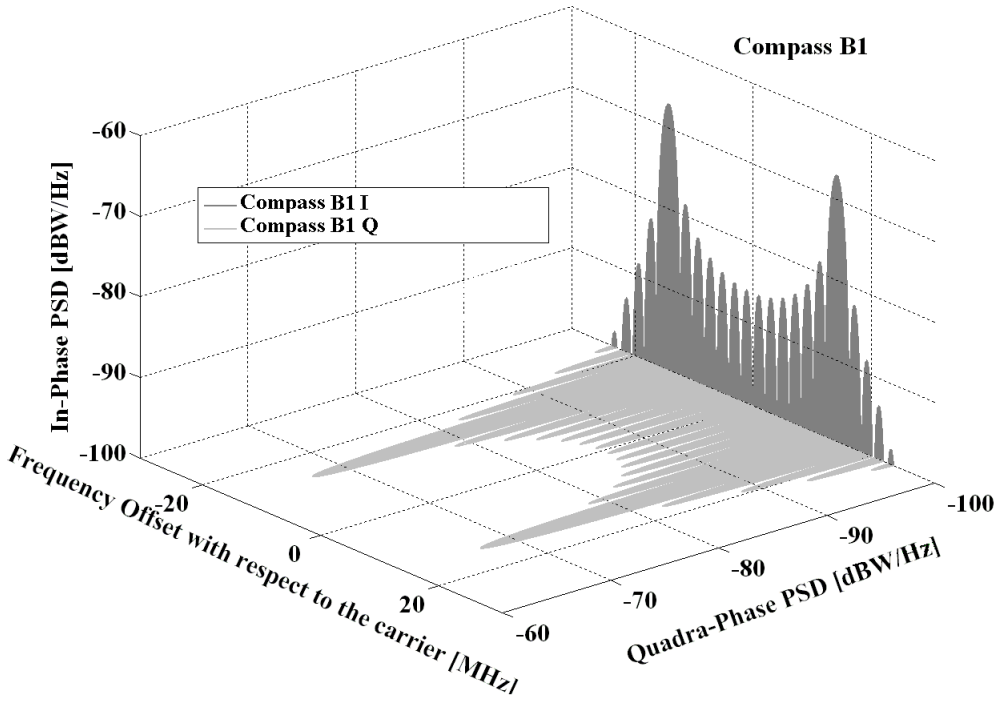


FIGURE A.12: B1 Compass/Beidou frequency plan (center carrier frequency: 1561.098 MHz). Source: ESA Navipedia.

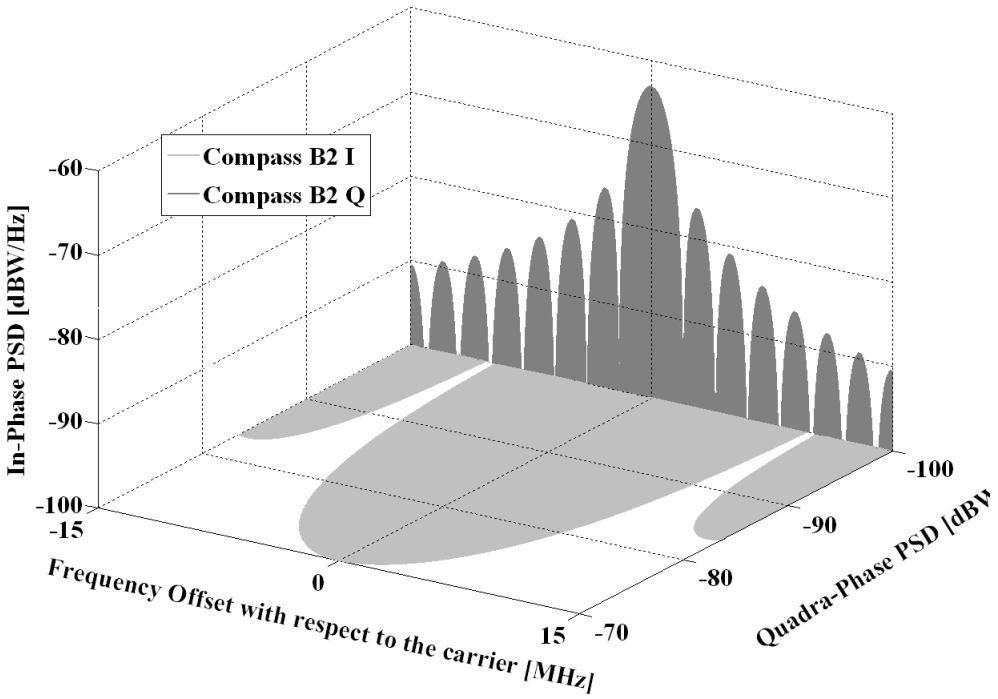


FIGURE A.13: B2 Compass/Beidou frequency plan (center carrier frequency: 1207.14 MHz). Source: ESA Navipedia.

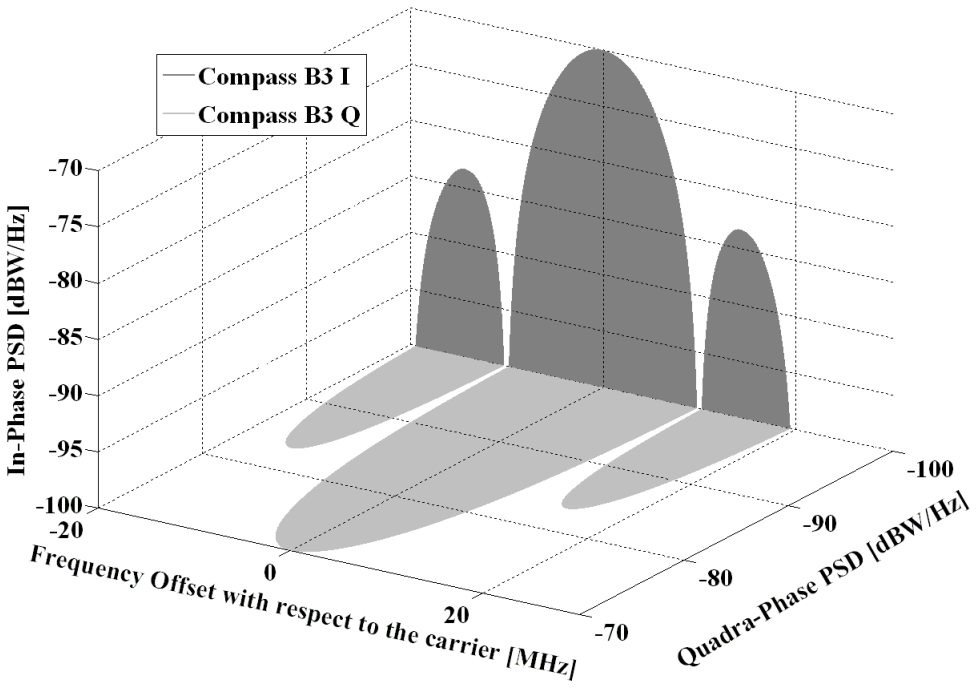


FIGURE A.14: B3 Compass/Beidou frequency plan (center carrier frequency: 1268.52 MHz). Source: ESA Navipedia.

## A.4 GNSS overall frequency bandwidths

Fig. A.15 shows the frequency bands in the range 1145 and 1615 MHz in which GNSS emit signals.

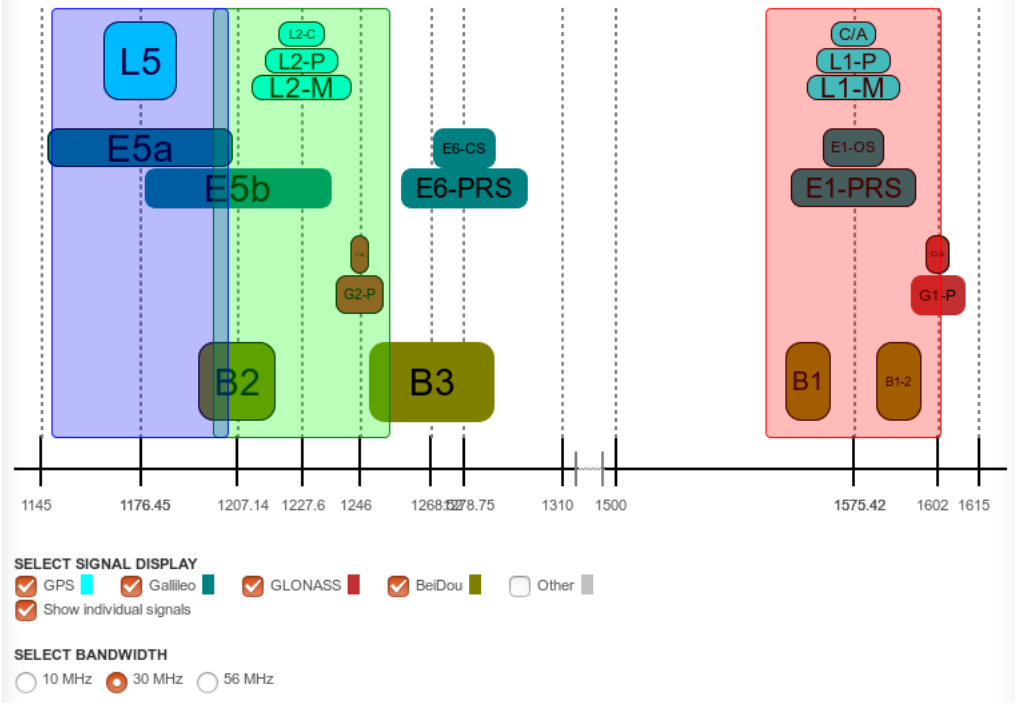


FIGURE A.15: Overall view of GNSS frequency plans. The bandwidth selected for L1/E1/B1, L2/E2/B2 and L5/E5/B5 is 30 MHz, in order to take into account signals below 10dB. Source: LabSat 3 Wideband.

The following figures provide a visual description of GNSS frequency plan overlaps for:

- L1 (cf. Fig.A.16)
- L5 - B2 (cf. Fig.A.17)
- E6 - B3 (cf. Fig.A.18)

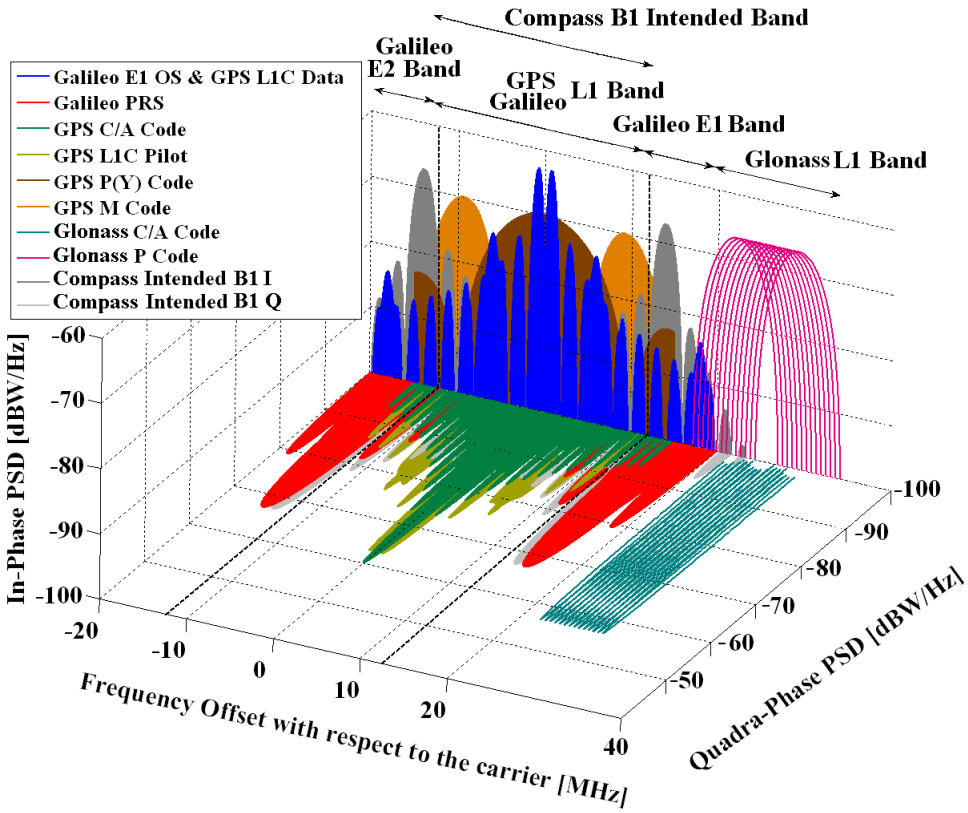


FIGURE A.16: L1 band GNSS frequency plans (GPS, Galileo, GLONASS Option 2 and BeiDou) (center carrier frequency: 1575.42 MHz). Source: ESA Navipedia.

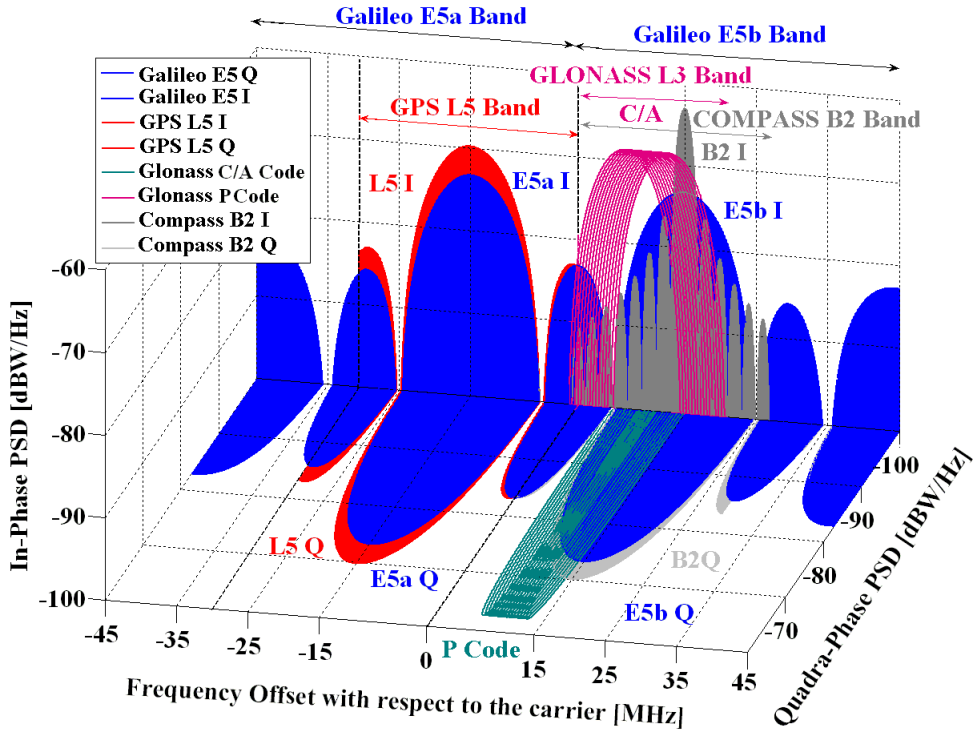


FIGURE A.17: L5 - B2 band GNSS frequency plans (GPS, Galileo, GLONASS and BeiDou) (center carrier frequency: 1191.795 MHz). Source: ESA Navipedia.

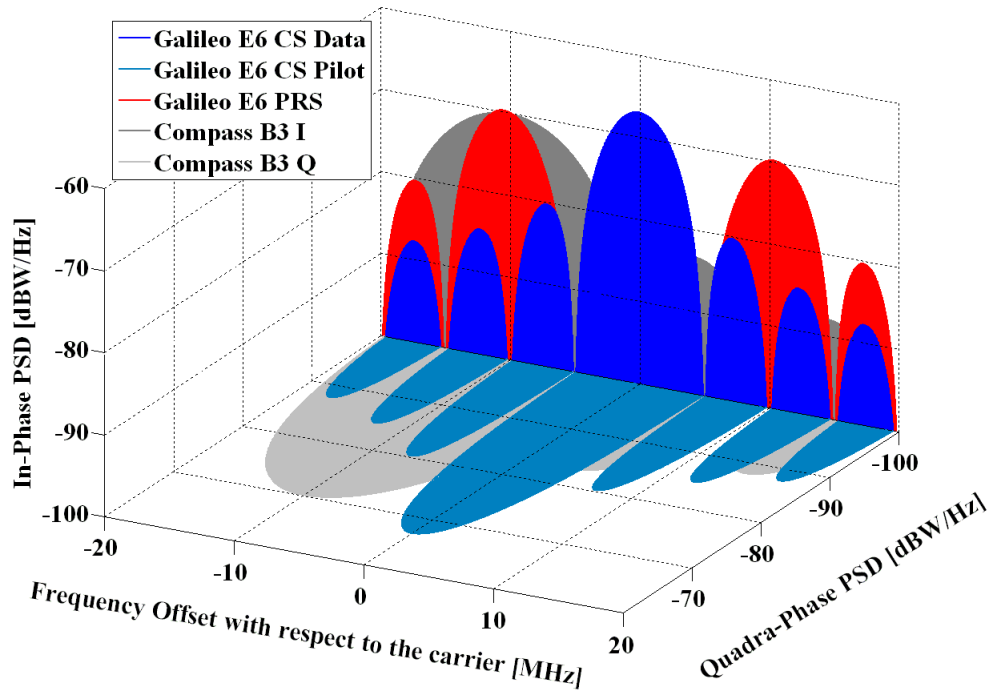


FIGURE A.18: E6 - B3 band GNSS frequency plans (Galileo and BeiDou) (center carrier frequency: 1278.75 MHz). Source: ESA Navipedia.

# Appendix B

# The Iridium satellite constellation

The Iridium satellite constellation consists of 66 active satellites (plus spares in storage orbits) which provide a global coverage of voice and data telecommunications. Of primary interest to radio astronomy is the use by Iridium of the 1616 - 1625 MHz bandwidth. The signal is right-hand circulary polarized (RHCP). Traffic channels are implemented with Time Division Duplex (TDD), an hybrid technology using Time Division Multiple Access/Frequency Division Multiple Access (TDMA/FDMA). A single TDMA time-slot lasts 90 ms (cf. Fig. B.1). The FDMA elementary frequency access occupies a 41.667 kHz bandwidth (cf. Fig. B.2). Frequency accesses used for duplex transmission are organized in sub-bands each consisting of 8 frequency accesses in a bandwidth of 333.333kHz (i.e., 8 times 41.667kHz). There are 30 sub-bands (i.e., 240 frequency accesses) in the 1616 - 1625 MHz bandwidth. Simplex channel band used for ring alert and messaging occupy a globally allocated 500 kHz within the 1616 - 1625 MHz bandwidth[1].

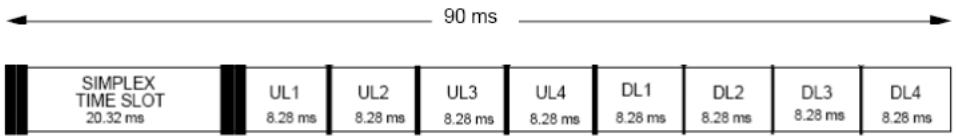


FIGURE B.1: The Iridium TDMA (Time Division Multiple Access) structure. The 90 ms are divided into (1) a 20.32 ms downlink simplex time-slot, (2) uplink (UL) and downlink (DL) time-slots of 8.28 ms each. Source: ICAO Manual mentioned before.

[1]Please note that all information mentioned about telecommunications was extracted from the Manual for ICAO Aeronautical Mobile Satellite (Route) Service Part 2-IRIDIUM Draft v4.0 published on 21 March 2007.

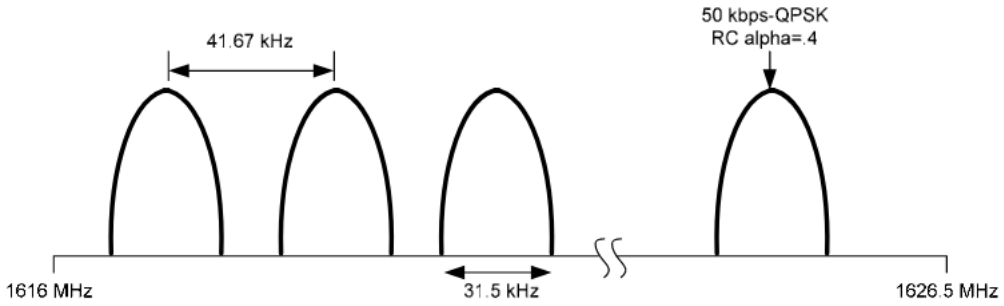


FIGURE B.2: The Iridium FDMA (Frequency Division Multiple Access) frequency plan. The signal format consists of 25 kilosymbols-per-second (ksps), i.e., 50 kilobits-per-second (kbps), quadrature phase shift keying (QPSK) modulation and are implemented with 40% square root raised cosine pulse shaping. Source: ICAO Manual mentioned before.

# Appendix C

# Observation of B0738+313: statistics on scans and cycles in the normalization bandwidth $1421 - 1421.5$ MHz.

The observation of BO738+313 consists of 7 scans, each of them made of alternate *ON* and *OFF*-position cycles lasting 1 minute each. One cycle includes 301 averaged power spectral values, each of them being integrated during $\tau = 0.199$s, the frequency channel width is set at $B = 262Hz$.

The spectra normalization bandwidth ($1421 - 1421.5$ MHz) is split into 1908 frequency channels. Therefore, statistics are calculated for each cycle on $301 \times 1908 = 574308$ uncalibrated averaged power spectra (in arbitrary units).

The total number of *ON* spectra (which is the same as *OFF*) recorded in the 7 scans is 31906.

The following table provides statistics on ON cycles only. Columns are:

1. scan number recorded in the NRT database;
2. Doppler frequency shift (Solar System Barycentric) calculated at rest frequency: 1420.40575 MHz and expressed in:
   (a) number of frequency channels,
   (b) frequency shift in *Hz*;
3. statistics calculated on uncalibrated averaged power spectral values (i.e., in arbitrary units) in the spectral normalization bandwidth for each cycle:
   (a) median,
   (b) % of median variation compared to the previous cycle,
   (c) average,
   (d) % of average variation compared to the previous cycle,
   (e) standard deviation,
   (f) normalized standard deviation (i..e, standard deviation / average),
   (g) skewness,

(h) kurtosis.

For each scan, variations are calculated between the minimum and maximum, as well as the values between the first and the last cycle for:

- median;
- average;
- standard deviation;
- skewness;
- kurtosis.

The same variations are eventually processed for the overall observation.

ON

| Scan # | Cycle # | Doppler nb of frequency channels shift | Hz (B=262Hz) | Averaged power spectral values (uncalibrated arbitrary units) in the 1421 – 1421.5 bandwidth Median (1) | Δ% (1) with previous cy. | Average (2) | Δ% (2) with previous cy. | Δ% (2)-(1)/(1) | std deviation σ (3) | normalized σ (3)/(2) | skewness | kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 243923 | 1 | -355 | -93010 | 12486.5098 | | 12562.9844 | | -0.61% | 1662.06665 | 0.132298712 | -7.62E-04 | -3.59E-04 |
| | 2 | -355 | | 12506.8877 | 0.16% | 12428.585 | -1.08% | 0.63% | 1654.12402 | 0.133090293 | -7.68E-04 | -3.53E-04 |
| | 3 | -355 | | 12567.7324 | 0.48% | 12487.373 | 0.47% | 0.64% | 1662.47156 | 0.13313221 | -7.69E-04 | -3.61E-04 |
| | 4 | -355 | | 12594.6855 | 0.21% | 12516.5625 | 0.23% | 0.62% | 1664.87756 | 0.133013961 | -7.44E-04 | -2.88E-04 |
| | 5 | -355 | | 12555.2725 | -0.31% | 12476.8008 | -0.32% | 0.63% | 1661.09644 | 0.133134805 | -7.62E-04 | -3.41E-04 |
| | 6 | -355 | | 12516.5195 | -0.31% | 12434.7383 | -0.34% | 0.65% | 1653.20984 | 0.132950915 | -7.77E-04 | -3.49E-04 |
| | 7 | -355 | | 12373.751 | -1.15% | 12294.5879 | -1.14% | 0.64% | 1634.69141 | 0.132960244 | -7.64E-04 | -3.50E-04 |
| | 8 | -355 | | 12392.9443 | 0.15% | 12310.6797 | 0.13% | 0.66% | 1638.32727 | 0.133081788 | -7.82E-04 | -3.64E-04 |
| | 9 | -355 | | 12665.3418 | 2.15% | 12586.0098 | 2.19% | 0.63% | 1674.57056 | 0.133050155 | -7.59E-04 | -3.30E-04 |
| | 10 | -355 | | 12920.2021 | 1.97% | 12841.9297 | 1.99% | 0.61% | 1710.01587 | 0.133158794 | -7.67E-04 | -3.88E-04 |
| Δ% between : | min. and max. | | | 4.42% | | 4.45% | | | 4.61% | 0.65% | -4.87% | -25.76% |
| | first and last cycle | | | 3.47% | | 2.22% | | | 2.88% | 0.65% | 0.67% | 8.13% |
| 244045 | 1 | -389 | -101918 | 11441.4141 | -12.92% | 11512.7061 | -11.55% | -0.62% | 1524.38867 | 0.13240924 | -7.75E-04 | -3.47E-04 |
| | 2 | -389 | | 11422.6328 | -0.16% | 11492.6699 | -0.17% | -0.61% | 1518.18323 | 0.132100134 | -7.60E-04 | -3.67E-04 |
| | 3 | -389 | | 11401.9102 | -0.18% | 11473.0938 | -0.17% | -0.62% | 1517.37891 | 0.132255426 | -7.64E-04 | -3.13E-04 |
| | 4 | -389 | | 11447.1221 | 0.39% | 11517.7754 | 0.39% | -0.62% | 1523.8634 | 0.132305358 | -7.52E-04 | -3.64E-04 |
| | 5 | -389 | | 11432.917 | -0.12% | 11505.665 | -0.11% | -0.64% | 1522.84534 | 0.132356134 | -7.61E-04 | -3.46E-04 |
| | 6 | -389 | | 11386.708 | -0.41% | 11460.6162 | -0.39% | -0.65% | 1514.41602 | 0.13214089 | -7.73E-04 | -3.42E-04 |
| | 7 | -389 | | 11322.998 | -0.56% | 11396.7842 | -0.56% | -0.65% | 1507.45154 | 0.132269903 | -7.77E-04 | -3.59E-04 |
| | 8 | -389 | | 11273.6035 | -0.44% | 11346.2871 | -0.45% | -0.64% | 1502.06458 | 0.132383798 | -7.70E-04 | -3.62E-04 |
| | 9 | -389 | | 11341.0059 | 0.59% | 11413.6133 | 0.59% | -0.64% | 1508.34802 | 0.132153419 | -7.69E-04 | -3.56E-04 |
| | 10 | -389 | | 11640.9922 | 2.58% | 11715.3652 | 2.58% | -0.64% | 1550.2915 | 0.132329763 | -7.70E-04 | -3.11E-04 |
| | 11 | -389 | | 11910.2129 | 2.26% | 11988.4326 | 2.28% | -0.66% | 1588.42834 | 0.132496749 | -7.69E-04 | -3.36E-04 |
| Δ% between : | min. and max. | | | 5.65% | | 5.66% | | | 5.75% | 0.30% | -3.23% | -15.35% |
| | first and last cycle | | | 4.27% | | 4.31% | | | 4.63% | 0.30% | 1.17% | -8.52% |

ON

| Scan # | Cycle # | Doppler nb of frequency channels shift | Hz (B=262Hz) | Averaged power spectral values (uncalibrated arbitrary units) in the 1421 – 1421.5 bandwidth Median (1) | Δ% (1) with previous cy. | Average (2) | Δ% (2) with previous cy. | Δ% (2)-(1)/(1) | std deviation σ (3) | normalized σ (3)/(2) | skewness | kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 244370 | 1 | -468 | -122616 | 11261.4756 | -5.76% | 11332.0938 | -5.79% | -0.63% | 1497.96387 | 0.13218774 | -7.65E-04 | -3.35E-04 |
| | 2 | -468 | | 11328.9102 | 0.60% | 11401.5039 | 0.61% | -0.64% | 1508.61401 | 0.132317107 | -7.71E-04 | -3.49E-04 |
| | 3 | -468 | | 11324.9766 | -0.03% | 11397.0137 | -0.04% | -0.64% | 1507.81445 | 0.132299082 | -7.66E-04 | -3.61E-04 |
| | 4 | -468 | | 11298.8262 | -0.23% | 11371.2764 | -0.23% | -0.64% | 1502.86475 | 0.132163242 | -7.64E-04 | -3.31E-04 |
| | 5 | -468 | | 11266.6279 | -0.29% | 11336.5303 | -0.31% | -0.62% | 1502.755 | 0.132558637 | -7.72E-04 | -4.09E-04 |
| | 6 | -468 | | 11293.5049 | 0.24% | 11366.5938 | 0.26% | -0.65% | 1501.92859 | 0.132135327 | -7.74E-04 | -3.31E-04 |
| | 7 | -468 | | 11294.8105 | 0.01% | 11366.6719 | 0.00% | -0.64% | 1503.21716 | 0.132247783 | -7.68E-04 | -3.48E-04 |
| | 8 | -468 | | 11250.3086 | -0.40% | 11320.123 | -0.41% | -0.62% | 1497.44214 | 0.132281437 | -7.59E-04 | -3.45E-04 |
| | 9 | -468 | | 11217.9287 | -0.29% | 11288.7188 | -0.28% | -0.63% | 1492.68042 | 0.13222762 | -7.60E-04 | -3.39E-04 |
| | 10 | -468 | | 11465.9551 | 2.16% | 11536.0137 | 2.14% | -0.61% | 1527.59631 | 0.132419773 | -7.55E-04 | -2.93E-04 |
| | 11 | -468 | | 11702.5273 | 2.02% | 11773.832 | 2.02% | -0.61% | 1559.4187 | 0.132447847 | -7.59E-04 | -3.23E-04 |
| Δ% between : | min. and max. | | | 4.32% | | 4.30% | | | 4.47% | 0.32% | -2.44% | -28.30% |
| | first and last cycle | | | 3.30% | | 3.27% | | | 3.37% | 0.10% | -1.59% | -7.61% |
| 244510 | 1 | -517 | -135454 | 12076.5273 | 3.10% | 12153.2383 | 3.12% | -0.64% | 1607.52649 | 0.132271453 | -7.65E-04 | -3.19E-04 |
| | 2 | -517 | | 12082.4668 | 0.05% | 12160.2822 | 0.06% | -0.64% | 1609.54102 | 0.132360499 | 7.73E-04 | -3.60E-04 |
| | 3 | -517 | | 12112.3564 | 0.25% | 12194.1611 | 0.28% | -0.68% | 1614.60742 | 0.132408241 | -8.00E-04 | -3.96E-04 |
| | 4 | -517 | | 12177.1895 | 0.53% | 12255.0859 | 0.50% | -0.64% | 1622.98389 | 0.132433498 | -7.74E-04 | -3.60E-04 |
| | 5 | -517 | | 12213.9453 | 0.30% | 12290.6699 | 0.29% | -0.63% | 1625.3562 | 0.132243093 | -7.58E-04 | -3.12E-04 |
| | 6 | -517 | | 12184.957 | -0.24% | 12263.6865 | -0.22% | -0.65% | 1623.55823 | 0.132387454 | -7.60E-04 | -3.23E-04 |
| | 7 | -517 | | 12117.3213 | -0.56% | 12195.6572 | -0.56% | -0.65% | 1612.67847 | 0.132233831 | -7.54E-04 | 3.13E-04 |
| | 8 | -517 | | 12079.6211 | -0.31% | 12159.8428 | -0.29% | -0.66% | 1607.35461 | 0.132185476 | -7.66E-04 | -3.55E-04 |
| | 9 | -517 | | 12072.957 | -0.06% | 12151.3574 | -0.07% | -0.65% | 1607.6731 | 0.132303993 | -7.83E-04 | -4.01E-04 |
| | 10 | -518 | -135716 | 12382.3242 | 2.50% | 12458.9346 | 2.47% | -0.62% | 1648.74573 | 0.132334408 | -7.64E-04 | -3.29E-04 |
| | 11 | -518 | | 12719.375 | 2.65% | 12800.5781 | 2.67% | -0.64% | 1695.08679 | 0.132422675 | -7.65E-04 | -3.52E-04 |
| Δ% between : | min. and max. | | | 5.35% | | 5.34% | | | 5.46% | 0.19% | -196.52% | -177.96% |
| | first and last cycle | | | 5.27% | | 5.27% | | | 5.31% | 0.05% | -199.06% | -2.29% |

ON

| Scan # | Cycle # | Doppler nb of frequency channels shift | Hz (B=262Hz) | Averaged power spectral values (uncalibrated arbitrary units) in the 1421 – 1421.5 bandwidth Median (1) | Δ% (1) with previous cy. | Average (2) | Δ% (2) with previous cy. | Δ% (2)-(1)/(1) | std deviation σ (3) | normalized σ (3)/(2) | skewness | kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 244690 | 1 | -572 | -149864 | 11819.1055 | -7.62% | 11895.7412 | -7.61% | -0.65% | 1577.53552 | 0.13261347 | 7.74E-04 | -3.68E-04 |
| | 2 | -572 | | 11779.9414 | -0.33% | 11852.4043 | -0.37% | -0.62% | 1568.17004 | 0.13230818 | -7.64E-04 | 3.28E-04 |
| | 3 | -572 | | 11825.5566 | 0.39% | 11901.5273 | 0.41% | -0.64% | 1575.39526 | 0.132369167 | -7.70E-04 | -3.47E-04 |
| | 4 | -572 | | 11771.4102 | -0.46% | 11847.71 | -0.45% | -0.65% | 1568.30347 | 0.132371865 | -7.74E-04 | -3.61E-04 |
| | 5 | -572 | | 11844.1191 | 0.61% | 11918.4971 | 0.59% | -0.63% | 1578.99109 | 0.132482399 | -7.60E-04 | -3.55E-04 |
| | 6 | -572 | | 11826.4609 | -0.15% | 11903.5195 | -0.13% | -0.65% | 1574.82666 | 0.132299246 | -7.74E-04 | -3.22E-04 |
| | 7 | -573 | -150126 | 11855.791 | 0.25% | 11932.7021 | 0.24% | -0.65% | 1577.97473 | 0.132239514 | -7.84E-04 | -3.63E-04 |
| | 8 | -573 | | 11888.0293 | 0.27% | 11965.3174 | 0.27% | -0.65% | 1583.73718 | 0.132360649 | -7.69E-04 | 3.44E-04 |
| | 9 | -573 | | 11946.3232 | 0.49% | 12024.4053 | 0.49% | -0.65% | 1589.51062 | 0.132190373 | -7.67E-04 | -3.34E-04 |
| | 10 | -573 | | 11898.583 | -0.40% | 11973.0078 | -0.43% | -0.63% | 1586.66467 | 0.13252014 | -7.74E-04 | -3.69E-04 |
| | 11 | -573 | | 11898.8125 | 0.00% | 11973.2012 | 0.00% | -0.63% | 1582.17969 | 0.132143415 | -7.69E-04 | -3.81E-04 |
| | 12 | -573 | | 11901.8086 | 0.03% | 11975.5234 | 0.02% | -0.62% | 1584.52283 | 0.132313451 | -7.57E-04 | -3.27E-04 |
| | 13 | -573 | | 11886.2803 | -0.13% | 11960.9414 | -0.12% | -0.63% | 1584.51648 | 0.132474228 | -7.59E-04 | -3.11E-04 |
| | 14 | -573 | | 11812.1455 | -0.63% | 11885.6338 | -0.63% | -0.62% | 1571.69983 | 0.132235256 | -7.66E-04 | -3.66E-04 |
| | 15 | -573 | | 11664.1484 | -1.27% | 11737.5098 | -1.26% | -0.63% | 1552.45959 | 0.132264817 | -7.68E-04 | -3.54E-04 |
| | 16 | -573 | | 11837.0459 | 1.46% | 11911.3955 | 1.46% | -0.63% | 1576.16101 | 0.132323791 | -7.68E-04 | -3.39E-04 |
| | 17 | -573 | | 11967.5664 | 1.09% | 12043.4824 | 1.10% | -0.63% | 1593.04541 | 0.132274483 | -7.67E-04 | -3.41E-04 |
| | 18 | -573 | | 12336.9346 | 2.99% | 12412.8477 | 2.98% | -0.62% | 1641.21716 | 0.13221923 | -7.56E-04 | -3.58E-04 |
| Δ% between : | min. and max. | | | 5.77% | | 5.75% | | | 5.72% | 0.36% | -198.72% | -190.39% |
| | first and last cycle | | | 4.38% | | 4.35% | | | 4.04% | -0.30% | -197.65% | -2.76% |

ON

| Scan # | Cycle # | Doppler nb of frequency channels shift | Hz (B=262Hz) | Averaged power spectral values (uncalibrated arbitrary units) in the 1421 – 1421.5 bandwidth Median (1) | Δ% (1) with previous cy. | Average (2) | Δ% (2) with previous cy. | Δ% (2)-(1)/(1) | std deviation σ (3) | normalized σ (3)/(2) | skewness | kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 244808 | 1 | -593 | -155366 | 11369.3008 | -8.51% | 11441.3516 | -8.49% | -0.63% | 1516.21375 | 0.13252051 | -7.61E-04 | -3.14E-04 |
| | 2 | -593 | | 11045.9033 | -2.93% | 11117.8516 | -2.91% | -0.65% | 1471.89746 | 0.132390458 | -7.58E-04 | -3.22E-04 |
| | 3 | -593 | | 10824.9609 | -2.04% | 10891.5713 | -2.08% | -0.62% | 1440.66589 | 0.132273466 | -7.63E-04 | -3.54E-04 |
| | 4 | -593 | | 10726.1025 | -0.92% | 10795.7725 | -0.89% | -0.65% | 1431.16321 | 0.132567003 | -7.74E-04 | -3.21E-04 |
| | 5 | -593 | | 10761.1113 | 0.33% | 10830.3613 | 0.32% | -0.64% | 1434.79541 | 0.132478998 | -7.58E-04 | -3.44E-04 |
| | 6 | -593 | | 10692.1992 | -0.64% | 10759.5898 | -0.66% | -0.63% | 1423.21338 | 0.132273944 | -7.69E-04 | -3.68E-04 |
| | 7 | -593 | | 10697.9609 | 0.05% | 10766.7754 | 0.07% | -0.64% | 1425.72827 | 0.132419245 | -7.73E-04 | -3.78E-04 |
| | 8 | -593 | | 10659.9756 | -0.36% | 10727.0342 | -0.37% | -0.63% | 1420.25745 | 0.132399825 | -7.70E-04 | -3.71E-04 |
| | 9 | -593 | | 10653.0312 | -0.07% | 10720.2783 | -0.06% | -0.63% | 1417.14148 | 0.132192602 | -7.77E-04 | -3.56E-04 |
| | 10 | -593 | | 10627.4473 | -0.24% | 10696.1279 | -0.23% | -0.65% | 1417.68823 | 0.132542191 | -7.72E-04 | -3.30E-04 |
| | 11 | -594 | -155628 | 10605.3291 | -0.21% | 10672.8457 | -0.22% | -0.64% | 1413.53528 | 0.132442211 | -7.61E-04 | -3.49E-04 |
| | 12 | -594 | | 10634.1953 | 0.27% | 10701.4971 | 0.27% | -0.63% | 1415.26208 | 0.13224898 | -7.72E-04 | -3.85E-04 |
| | 13 | -594 | | 10642.6816 | 0.08% | 10708.2363 | 0.06% | -0.62% | 1415.04211 | 0.132145208 | -7.74E-04 | -3.73E-04 |
| | 14 | -594 | | 10623.1367 | -0.18% | 10692.0996 | -0.15% | -0.65% | 1414.24255 | 0.132269863 | -7.86E-04 | -3.83E-04 |
| | 15 | -594 | | 10599.9824 | -0.22% | 10667.0762 | -0.23% | -0.63% | 1413.06824 | 0.132470061 | -7.70E-04 | -3.38E-04 |
| | 16 | -594 | | 10599.5391 | 0.00% | 10664.3662 | -0.03% | -0.61% | 1409.76379 | 0.132193865 | -7.61E-04 | -3.72E-04 |
| | 17 | -594 | | 10606.6504 | 0.07% | 10675.5146 | 0.10% | -0.65% | 1413.66016 | 0.132420798 | -7.60E-04 | -3.16E-04 |
| | 18 | -594 | | 10673.8477 | 0.63% | 10743.8848 | 0.64% | -0.66% | 1421.25049 | 0.132284599 | -7.73E-04 | -3.59E-04 |
| | 19 | -594 | | 10667.9131 | -0.06% | 10735.7559 | -0.08% | -0.64% | 1420.4115 | 0.132306613 | -7.76E-04 | -3.57E-04 |
| | 20 | -594 | | 10695.3887 | 0.26% | 10760.0176 | 0.23% | -0.60% | 1423.60327 | 0.13230492 | -7.65E-04 | -3.50E-04 |
| | 21 | -594 | | 10697.6348 | 0.02% | 10764.3359 | 0.04% | -0.62% | 1425.5282 | 0.132430669 | -7.69E-04 | -3.52E-04 |
| | 22 | -594 | | 10644.5586 | -0.50% | 10712.1982 | -0.49% | -0.64% | 1418.13013 | 0.132384605 | -7.68E-04 | -3.23E-04 |
| | 23 | -594 | | 10636.7383 | -0.07% | 10704.459 | -0.07% | -0.64% | 1417.05823 | 0.132380182 | -7.81E-04 | -3.54E-04 |
| | 24 | -594 | | 10572.8379 | -0.60% | 10638.4209 | -0.62% | -0.62% | 1406.05359 | 0.132167509 | -7.46E-04 | -3.21E-04 |
| | 25 | -594 | | 10873.9688 | 2.77% | 10942.6787 | 2.78% | -0.63% | 1448.74536 | 0.132394033 | -7.74E-04 | 3.72E-04 |
| | 26 | -595 | -155890 | 11101.8623 | 2.05% | 11170.9922 | 2.04% | -0.62% | 1474.89001 | 0.13202856 | -7.63E-04 | -3.77E-04 |
| Δ% between : | min. and max. | | | 7.53% | | 7.55% | | | 7.83% | 0.41% | -5.08% | -196.56% |
| | first and last cycle | | | -2.35% | | -2.36% | | | -2.73% | -0.37% | 0.38% | 20.19% |

ON

| Scan # | Cycle # | Doppler nb of frequency channels shift | Hz (B=262Hz) | Averaged power spectral values (uncalibrated arbitrary units) in the 1421 – 1421.5 bandwidth Median (1) | Δ% (1) with previous cy. | Average (2) | Δ% (2) with previous cy. | Δ% (2)-(1)/(1) | std deviation σ (3) | normalized σ (3)/(2) | skewness | kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 244990 | 1 | -621 | | 11897.584 | 6.69% | 11973.165 | 6.70% | -0.64% | 1584.33447 | 0.132323782 | -7.62E-04 | -3.51E-04 |
| | 2 | -621 | | 11901.1621 | 0.03% | 11974.2852 | 0.01% | -0.61% | 1586.90186 | 0.132525811 | -7.50E-04 | -2.86E-04 |
| | 3 | -621 | | 11868.6279 | -0.27% | 11942.6768 | -0.26% | -0.62% | 1582.1012 | 0.132474589 | -7.75E-04 | -3.59E-04 |
| | 4 | -621 | | 11856.0586 | -0.11% | 11930.5361 | -0.10% | -0.63% | 1576.28271 | 0.1321217 | -7.63E-04 | -3.34E-04 |
| | 5 | -621 | | 11894.7969 | 0.33% | 11972.3828 | 0.35% | -0.65% | 1585.73096 | 0.132449069 | -7.72E-04 | -3.29E-04 |
| | 6 | -621 | | 11837.1094 | -0.49% | 11911.5098 | -0.51% | -0.63% | 1574.89563 | 0.13221629 | -7.68E-04 | -3.54E-04 |
| | 7 | -621 | | 11835.9971 | -0.01% | 11914.8037 | 0.03% | -0.67% | 1576.93896 | 0.132351233 | -7.74E-04 | -3.45E-04 |
| | 8 | -621 | | 11832.1133 | -0.03% | 11910.1523 | -0.04% | -0.66% | 1573.78564 | 0.132138162 | -7.71E-04 | -3.15E-04 |
| | 9 | -621 | | 11827.7188 | -0.04% | 11900.7646 | -0.08% | -0.62% | 1575.17212 | 0.132358901 | -7.73E-04 | -3.47E-04 |
| | 10 | -621 | | 11836.1465 | 0.07% | 11913.3203 | 0.11% | -0.65% | 1575.81982 | 0.132273773 | -7.79E-04 | -3.24E-04 |
| | 11 | -622 | -162964 | 11849.1045 | 0.11% | 11921.416 | 0.07% | -0.61% | 1573.84607 | 0.132018384 | -7.59E-04 | -3.63E-04 |
| | 12 | -622 | | 11869.6816 | 0.17% | 11944.3193 | 0.19% | -0.63% | 1579.08093 | 0.132203509 | -7.68E-04 | -3.31E-04 |
| | 13 | -622 | | 11802.3184 | -0.57% | 11877.6064 | -0.56% | -0.64% | 1570.93591 | 0.13226031 | -7.78E-04 | -3.57E-04 |
| | 14 | -622 | | 11795.707 | -0.06% | 11869.3125 | -0.07% | -0.62% | 1571.50122 | 0.132400358 | -7.65E-04 | -3.49E-04 |
| | 15 | -622 | | 11770.8965 | -0.21% | 11846.5439 | -0.19% | -0.64% | 1565.70825 | 0.132165825 | -7.53E-04 | -2.79E-04 |
| | 16 | -622 | | 11723.1523 | -0.41% | 11798.6094 | -0.41% | -0.64% | 1560.74463 | 0.132282083 | -7.69E-04 | -3.40E-04 |
| | 17 | -622 | | 11889.0645 | 1.40% | 11963.6094 | 1.38% | -0.63% | 1581.61841 | 0.132202445 | -7.61E-04 | -3.27E-04 |
| | 18 | -622 | | 12077.0957 | 1.56% | 12153.751 | 1.56% | -0.63% | 1608.94739 | 0.132382784 | -7.67E-04 | -3.31E-04 |
| | 19 | -622 | | 12462.0381 | 3.09% | 12539.8799 | 3.08% | -0.62% | 1658.0658 | 0.132223419 | -7.59E-04 | -3.52E-04 |
| Δ% between : | min. and max. | | | 6.30% | | 6.28% | | | 6.24% | 0.38% | -3.70% | -23.25% |
| | first and last cycle | | | 4.74% | | 4.73% | | | 4.65% | -0.08% | -0.39% | 0.19% |
| Observation : | | | | | | | | | | | | |
| Total : | 106 | | | | | | | | | | | |
| Δ% between : | min. and max. | | | 22.20% | | 20.71% | | | 21.62% | 0.86% | -196.73% | -190.98% |
| | first and last cycle | | | -0.20% | | -0.18% | | | -0.24% | -0.06% | -0.35% | -1.98% |

# Bibliography


Akeret, J., C. Chang, A. Lucchi, and A. Refregier (2017). "Radio frequency interference mitigation using deep convolutional neural networks". In: *Astronomy and Computing* 18, pp. 35–39. DOI: 10.1016/j.ascom.2017.01.002. arXiv: 1609.09077 [astro-ph.IM].

Athreya, R. (2009). "A New Approach to Mitigation of Radio Frequency Interference in Interferometric Data". In: *Astrophysical Journal* 696, pp. 885–890. DOI: 10.1088/0004-637X/696/1/885. arXiv: 0902.3332 [astro-ph.IM].

Baan, W., P. A. Fridman, S. Roy, and R. Millenaar (2010). "The RFI Mitigations System at WSRT". In: *RFI Mitigation Workshop*, p. 24.

Baan, Willem A. (2019). "Implementing RFI Mitigation in Radio Science". In: *Journal of Astronomical Instrumentation* 08.01, p. 1940010. DOI: 10.1142/S2251171719400105. eprint: https://doi.org/10.1142/S2251171719400105.

Baan, Willem A., Axel Jessner, and Jaap Steenge (2019). "Measuring Data Loss Resulting from Interference". In: *Journal of Astronomical Instrumentation* 08.01, p. 1940007. DOI: 10.1142/S2251171719400075. eprint: https://doi.org/10.1142/S2251171719400075.

Barnbaum, Cecilia and Richard F. Bradley (1998). "A New Approach to Interference Excision in Radio Astronomy: Real-Time Adaptive Cancellation". In: *The Astronomical Journal* 116.5, pp. 2598–2614. DOI: 10.1086/300604.

Barnes, D. G., L. Staveley-Smith, W. J. G. de Blok, T. Oosterloo, I. M. Stewart, A. E. Wright, G. D. Banks, R. Bhathal, P. J. Boyce, M. R. Calabretta, M. J. Disney, M. J. Drinkwater, R. D. Ekers, K. C. Freeman, B. K. Gibson, A. J. Green, R. F. Haynes, P. te Lintel Hekkert, P. A. Henning, H. Jerjen, S. Juraszek, M. J. Kesteven, V. A. Kilborn, P. M. Knezek, B. Koribalski, R. C. Kraan-Korteweg, D. F. Malin, M. Marquarding, R. F. Minchin, J. R. Mould, R. M. Price, M. E. Putman, S. D. Ryder, E. M. Sadler, A. Schröder, F. Stootman, R. L. Webster, W. E. Wilson, and T. Ye (2001). "The HI Parkes All Sky Survey: southern observations, calibration and robust imaging". In: *Monthly Notices of the Royal Astronomical Society* 322, pp. 486–498. DOI: 10.1046/j.1365-8711.2001.04102.x.

Becker, Claudia and Ursula Gather (1999). "The masking breakdown point of multivariate outlier identification rules". In: *Journal of the American Statistical Association* 94.447, pp. 947–955.

Belleval, Christophe (2001). "Le Pilotage des grands projets de haute technologie dans une organisation centrée sur ses compétences de base: le cas des programmes spatiaux, application à la ligne de produits microsatellites" Myriade" du CNES". PhD thesis. Université Louis Pasteur (Strasbourg).

Boonstra, A. J. and S. van der Tol (2005). "Spatial filtering of interfering signals at the initial Low Frequency Array (LOFAR) phased array test station". In: *Radio Science* 40, RS5S09, RS5S09. DOI: 10.1029/2004RS003135.

Buch, K. D., K. Naik, S. Nalawade, S. Bhatporia, Y. Gupta, and B. Ajithkumar (2019). "Real-Time Implementation of MAD-Based RFI Excision on FPGA". In: *Journal of Astronomical Instrumentation* 8, 1940006, p. 1940006. DOI: 10.1142/S2251171719400063.

Buch, Kaushal D., Shruti Bhatporia, Yashwant Gupta, Swapnil Nalawade, Aditya Chowdhury, Kishor Naik, Kshitij Aggarwal, and B. Ajithkumar (2016). "Towards Real-Time Impulsive RFI Mitigation for Radio Telescopes". In: *Journal of Astronomical Instrumentation* 05.04, p. 1641018. DOI: 10.1142/S225117171641018X. eprint: https://doi.org/10.1142/S225117171641018X.

Butcher, Z., S. Schneider, W. van Driel, M. D. Lehnert, and R. Minchin (2016). "NIBLES - an HI census of stellar mass selected SDSS galaxies. II. Arecibo follow-up HI observations". In: *Astronomy & Astrophysics* 596, A60, A60. DOI: 10.1051/0004-6361/201628189. arXiv: 1609.06242.

Davies, Laurie et al. (1992). "The asymptotics of Rousseeuw's minimum volume ellipsoid estimator". In: *The Annals of Statistics* 20.4, pp. 1828–1843.

Davies, Laurie and Ursula Gather (1993). "The identification of multiple outliers". In: *Journal of the American Statistical Association* 88.423, pp. 782–792.

Deshpande, Avinash A. and B. M. Lewis (2019). "Iridium Satellite Signals: A Case Study in Interference Characterization and Mitigation for Radio Astronomy Observations". In: *Journal of Astronomical Instrumentation* 08.01, p. 1940009. DOI: 10.1142/S2251171719400099. eprint: https://doi.org/10.1142/S2251171719400099.

Donoho David; Huber, Peter J. (1983). "The notion of breakdown point". In: *A Festschrift for Erich L. Lehmann*. Ed. by K. Doksum P. J. Bickel and Jr. J. L. Hodges. Belmont, CA: Wadsworth, 157–184.

Dumez-Viou, Cedric (2007). "Restauration de sources radioastronomiques en milieu radioélectrique hostile: Implantation de détecteurs temps réel sur des spectres dynamiques." PhD thesis.

Eguchi, Shinto and John Copas (2006). "Interpreting kullback–leibler divergence with the neyman–pearson lemma". In: *Journal of Multivariate Analysis* 97.9, pp. 2034–2040.

Finger, R., F. Curotto, R. Fuentes, R. Duan, L. Bronfman, and D. Li (2018). "A FPGA-based Fast Converging Digital Adaptive Filter for Real-time RFI Mitigation on Ground Based Radio Telescopes". In: *Publications of the Astronomical Society of the Pacific* 130.2, p. 025002. DOI: 10.1088/1538-3873/aa972f. arXiv: 1805.06376 [astro-ph.IM].

Flöer, L., B. Winkel, and J. Kerp (2010). "RFI mitigation for the Effelsberg Bonn HI Survey (EBHIS)". In: *RFI Mitigation Workshop*, p. 42. arXiv: 1007.2328 [astro-ph.IM].

Fridman, P. (2009). "Robust correlators". In: *Astronomy and Astrophysics* 502, pp. 401–408. DOI: 10.1051/0004-6361/200912006. arXiv: 1009.5654 [astro-ph.IM].

Fridman, P. A. (2008). "Statistically Stable Estimates of Variance in Radio-Astronomy Observations as Tools for Radio-Frequency Interference Mitigation". In: *Astronomical Journal* 135, pp. 1810–1824. DOI: 10.1088/0004-6256/135/5/1810. arXiv: 1009.5655 [astro-ph.IM].

Gary, D. E., Z. Liu, and G. M. Nita (2010). "A Wideband Spectrometer with RFI Detection". In: *Publications of the Astronomical Society of the Pacific* 122, p. 560. DOI: 10.1086/652410.

Giovanelli, R., M. P. Haynes, B. R. Kent, A. Saintonge, S. Stierwalt, A. Altaf, T. Balonek, N. Brosch, S. Brown, B. Catinella, A. Furniss, J. Goldstein, G. L. Hoffman, R. A. Koopmann, D. A. Kornreich, B. Mahmood, A. M. Martin, K. L. Masters, A. Mitschang, E. Momjian, P. H. Nair, J. L. Rosenberg, and B. Walsh (2007). "The Arecibo Legacy Fast ALFA Survey. III. H I Source Catalog of the Northern Virgo Cluster Region". In: *Astronomical Journal* 133, pp. 2569–2583. DOI: 10.1086/516635. eprint: astro-ph/0702316.

Harris, F. J. (1978). "On the use of windows for harmonic analysis with the discrete Fourier transform". In: *Proceedings of the IEEE* 66.1, pp. 51–83. DOI: 10.1109/PROC.1978.10837.

Haynes, M. P., R. Giovanelli, B. R. Kent, E. A. K. Adams, T. J. Balonek, D. W. Craig, D. Fertig, R. Finn, C. Giovanardi, G. Hallenbeck, K. M. Hess, G. L. Hoffman, S. Huang, M. G. Jones, R. A. Koopmann, D. A. Kornreich, L. Leisman, J. Miller, C. Moorman, J. O'Connor, A. O'Donoghue, E. Papastergis, P. Troischt, D. Stark, and L. Xiao (2018). "The Arecibo Legacy Fast ALFA Survey: The ALFALFA Extragalactic H I Source Catalog". In: *Astrophysical Journal* 861, 49, p. 49. DOI: 10.3847/1538-4357/aac956. arXiv: 1805.11499.

Hellbourg, G., R. Weber, C. Capdessus, and A.-J. Boonstra (2012). "Cyclostationary approaches for spatial RFI mitigation in radio astronomy". In: *Comptes Rendus Physique* 13, pp. 71–79. DOI: 10.1016/j.crhy.2011.10.010.

Huang, Y., X.-P. Wu, Q. Zheng, J.-H. Gu, and H. Xu (2016). "The radio environment of the 21 Centimeter Array: RFI detection and mitigation". In: *Research in Astronomy and Astrophysics* 16, 36, p. 36. DOI: 10.1088/1674-4527/16/2/036. arXiv: 1602.06623 [astro-ph.IM].

Huber, Peter J (2011). *Robust statistics*. Springer.

Huillery, Julien and Nadine Martin (2007). "Approximation de la loi de probabilité du spectrogramme, KL Divergence et détection dans le plan temps-fréquence". In: *XXIème colloque GRETSI (GRETSI 2007)*. Troyes, France, pp. 457–460.

Hunt, Lucas R, DJ Pisano, and S Edel (2016). "The Search for HI Emission at z 0.4 in Gravitationally Lensed Galaxies with the Green Bank Telescope". In: *The Astronomical Journal* 152.2, p. 30.

Kalberla, P. (2010). "RFI Mitigation for the Parkes Galactic All-Sky Survey (GASS)". In: *RFI Mitigation Workshop*, p. 38.

Kanekar, N., T. Ghosh, and J. N. Chengalur (2001). "Detection of a multi-phase ISM at {vec z=} 0.2212". In: *Astronomy & Astrophysics* 373, pp. 394–401. DOI: 10.1051/0004-6361:20010545. eprint: astro-ph/0104321.

Kempen, Geert and Lucas Van Vliet (2000). "Mean and Variance of Ratio Estimators Used in Fluorescence Ratio Imaging". In: *Cytometry* 39, pp. 300–5. DOI: 10.1002/(SICI)1097-0320(20000401)39:43.0.CO;2-0.

Kesteven, M. (2010). "Overview of RFI mitigation methods in existing and new systems (invited)". In: *RFI Mitigation Workshop*, p. 7.

Lane, W., A. Smette, F. Briggs, S. Rao, D. Turnshek, and G. Meylan (1998). "H i 21 Centimeter Absorption in Two Low-Redshift Damped Lyalpha Systems". In: *The Astronomical Journal* 116, pp. 26–30. DOI: 10.1086/300422. eprint: astro-ph/9803243.

Lane, W. M., F. H. Briggs, and A. Smette (2000). "Detection of Warm and Cold Phases of the Neutral ISM in a Damped Ly$\alpha$ Absorber". In: *The Astrophysical Journal* 532, pp. 146–151. DOI: 10.1086/308578. eprint: arXiv:astro-ph/9911142.

Lecacheux, Alain, Cedric Dumez-Viou, and Karl-Ludwig Klein (2013). *Un spectrographe pour la radioastronomie aux ondes courtes, au voisinage de la coupure ionosphérique*.

Maronna, Ricardo A, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera (2018). *Robust statistics: theory and methods (with R)*. Wiley.

Martin, Jean-Michel (1989). "Physique du gaz interstellaire dans les galaxies fortement émettrice dans l'infrarouge lointain et étude particuli ère des mégamasers OH". PhD thesis. Observatoire de Paris.

Matthews, L. D. and W. van Driel (2000). "An H I survey of highly flattened, edge-on, pure disk galaxies". In: *Astronomy & Astrophysics Supplement* 143, pp. 421–456. DOI: 10.1051/aas:2000307.

Monnier Ragaigne, D., W. van Driel, S. E. Schneider, C. Balkowski, and T. H. Jarrett (2003). "A search for Low Surface Brightness galaxies in the near-infrared. III. Nançay H I line observations". In: *Astronomy and Astrophysics* 408, pp. 465–477. DOI: 10.1051/0004-6361:20030714. eprint: astro-ph/0305319.

Nita, Gelu M, Dale E Gary, Zhiwei Liu, Gordon J Hurford, and Stephen M White (2007). "Radio Frequency Interference Excision Using Spectral-Domain Statistics". In: *Publications of the Astronomical Society of the Pacific* 119.857, p. 805.

Peirce, B. (1852). "Criterion for the rejection of doubtful observations". In: *Astronomical Journal* 2, pp. 161–163. DOI: 10.1086/100259.

Pelat, Didier (2015). *Bases et Méthodes pour le Traitement des Données*. Observatoire de Paris.

Pihlström, Y. M., J. E. Conway, R. S. Booth, P. J. Diamond, and A. G. Polatidis (2001). "EVN and MERLIN observations of <ASTROBJ>III Zw 35 </ASTROBJ>. A starburst continuum and an OH maser ring". In: *Astronomy and Astrophysics* 377, pp. 413–424. DOI: 10.1051/0004-6361:20011107.

Press, William H (1996). *FORTRAN Numerical Recipes: Numerical recipes in FORTRAN 90: the art of parallel scientific computing*. Vol. 2. Cambridge University Press.

Randell, J., D. Field, K. N. Jones, J. A. Yates, and M. D. Gray (1995). "The OH zone in OH megamaser galaxies." In: *Astronomy and Astrophysics* 300, p. 659.

Rohlfs, Kristen and Thomas L Wilson (2013). *Tools of radio astronomy*. Springer Science & Business Media.

Rousseeuw, Peter and Victor Yohai (1984). "Robust regression by means of S-estimators". In: *Robust and nonlinear time series analysis*. Springer, pp. 256–272.

Rousseeuw, Peter and Christophe Croux (1992). "Explicit scale estimators with high breakdown point". eng. In: Dodge, Y. North-Holland; Amsterdam, pp. 77–92.

Rousseeuw, Peter and Annick M. Leroy (2003). *Robust Regression and Outlier Detection*. Wiley-Interscience.

Rousseeuw, Peter and Mia Hubert (2018). "Anomaly Detection by Robust Statistics". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8. DOI: 10.1002/widm.1236.

Rousseeuw, Peter J. and Christophe Croux (1993). "Alternatives to the Median Absolute Deviation". In: *Journal of the American Statistical Association* 88.424, pp. 1273–1283.

Saintonge, A. (2007). "The Arecibo Legacy Fast ALFA Survey. IV. Strategies for Signal Identification and Survey Catalog Reliability". In: *Astronomical Journal* 133, pp. 2087–2096. DOI: 10.1086/513515. eprint: astro-ph/0702178.

Schneider, SE, G Helou, EE Salpeter, and Y Terzian (1986). "Neutral hydrogen in small groups of galaxies". In: *The Astronomical Journal* 92, pp. 742–765.

Schneider, Stephen E, Trinh X Thuan, Christopher Magri, and James E Wadiak (1990). "Northern dwarf and low surface brightness galaxies. I-The Arecibo neutral hydrogen survey". In: *The Astrophysical Journal Supplement Series* 72, pp. 245–289.

Serra, Paolo, Tobias Westmeier, Nadine Giese, Russell Jurek, Lars Flöer, Attila Popping, Benjamin Winkel, Thijs van der Hulst, Martin Meyer, Bärbel S. Koribalski, Lister Staveley-Smith, and Hélène Courtois (2015). "SOFIA: a flexible source finder for 3D spectral line data". In: *Monthly Notices of the Royal Astronomical Society* 448.2, pp. 1922–1929. DOI: 10.1093/mnras/stv079. arXiv: 1501.03906 [astro-ph.IM].

Staveley-Smith, L., R. J. Cohen, J. M. Chapman, L. Pointon, and S. W. Unger (1987). "A systematic search for OH megamasers". In: *Monthly Notices of the Royal Astronomical Society* 226, pp. 689–701. DOI: 10.1093/mnras/226.3.689.

Staveley-Smith, L., R. C. Kraan-Korteweg, A. C. Schröder, P. A. Henning, B. S. Koribalski, I. M. Stewart, and G. Heald (2016). "The Parkes H I Zone of Avoidance Survey". In: *Astronomical Journal* 151, 52, p. 52. DOI: 10.3847/0004-6256/151/3/52. arXiv: 1602.02922.

Taylor, Jacob, Nolan Denman, Kevin Bandura, Philippe Berger, Kiyoshi Masui, Andre Renard, Ian Tretyakov, and Keith Vanderlinde (2019). "Spectral Kurtosis-Based RFI Mitigation for CHIME". In: *Journal of Astronomical Instrumentation* 08.01, p. 1940004. DOI: 10.1142/S225117171940004X. eprint: https://doi.org/10.1142/S225117171940004X.

Thuan, Trinh X, VA Lipovetsky, J-M Martin, and SA Pustilnik (1999). "HI observations of blue compact galaxies from the first and second Byurakan surveys". In: *Astronomy and Astrophysics Supplement Series* 139.1, pp. 1–24.

Trotter, A. S., J. M. Moran, L. J. Greenhill, X.-W. Zheng, and C. R. Gwinn (1997). "VLBA Imaging of the OH Maser in III Zw 35". In: *The Astrophysical Journal Letters* 485, pp. L79–L82. DOI: 10.1086/310818. eprint: astro-ph/9707007.

Van Aelst, Stefan and Peter Rousseeuw (2009). "Minimum Volume Ellipsoid". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1, pp. 71 –82. DOI: 10.1002/wics.19.

van Driel, W. (2011). "Radio quiet, please! - protecting radio astronomy from interference". In: *The Role of Astronomy in Society and Culture*. Ed. by D. Valls-Gabaud and A. Boksenberg. Vol. 260. IAU Symposium, pp. 457–464. DOI: 10.1017/S1743921311002675.

van Nieuwpoort, R. V. (2016). "Towards exascale real-time RFI mitigation". In: Radio Frequency Interference (RFI2016) Coexisting with Radio Frequency Interference. Socorro, New Mexico, USA, pp. 69–74. DOI: doi:10.1109/RFINT.2016.7833534.

Weber, R., C. Faye, F. Biraud, and J. Dansou (1997). "Spectral detector for interference time blanking using quantized correlator". In: *Astron. Astrophys. Suppl. Ser.* 126, pp. 161–167. DOI: 10.1051/aas:1997257.

Westfall, Peter H (2014). "Kurtosis as peakedness, 1905–2014. RIP". In: *The American Statistician* 68.3, pp. 191–195.

Whiting, Matthew T. (2012). "duchamp: a 3D source finder for spectral-line data". In: *Monthly Notices of the Royal Astronomical Society* 421.4, pp. 3242–3256. DOI: 10.1111/j.1365-2966.2012.20548.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2966.2012.20548.x.

Winkel, B., J. Kerp, and S. Stanko (2007). "RFI detection by automated feature extraction and statistical analysis". In: *Astronomische Nachrichten* 328.1, pp. 68–79. DOI: 10.1002/asna.200610661. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asna.200610661.

Yuan, Ke-Hai, Peter M Bentler, and Wei Zhang (2005). "The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication". In: *Sociological Methods & Research* 34.2, pp. 240–258.