Development of a Powerwall-based solution for the manual flagging of radio astronomy data from eMerlin

Felipe P. Mosquera

School of Computing

University of Leeds

A thesis submitted for the degree of

MSc by Research

30th September 2019

This thesis is dedicated to Mona and Tati.

Acknowledgements

First I would like to thank of all to my family and friends (s) for being there with me during this process. My supervisors Roy Ruddle and Melvin Hoare for their invaluable collaboration. German Chaparro for its unconditional support. Stanley Kurtz for believe in me. The RADA team Anna Scaiffe and Sally Cooper and Bios team Eduardo Gomez and Jorge Arboleda. Also Javier Moldon and Katharine Johnston.

This research was funded by the Science and Technology Facilities Council (ST/R001944/1).

Abstract

This project was created with the intention of establishing an optimisation method for the manual flagging of interferometric data of the eMerlin radio astronomy array, using a Powerwall as a visualisation tool. The complexity of this process which is due to the amount of variables and parameters demands a deep understanding of the data treatment. Once the data is achieved by the antennas the signals are correlated. This process generates undesired signals which mostly coming from radio frequency interference. Also when the calibration is performed some values can mislead the expected outcome. Although the flagging is supported with algorithms this method is not one hundred percent accurate. That is why visual inspection is still required. The possibility to use a Powerwall as a visualisation system allows different and new dynamics in terms of the interaction of the analyst with the information required to make the flagging.

Abbreviations

RFI	Radio Frequency Interference
Spw	Spectral window
ms	Measurement set
MHz	Mega Hertz
GHz	Giga Hertz
microJy	micro Jansky
B_i	Complex bandpass
uvfix	Complex visibility
UTC	Coordinated Universal Time
ITRF	International Terrestrial Reference Frame
px	pixel
Mpx	Mega Pixel
Gpx	Giga Pixel
LOD	Level of detail
DPI	Dots per inch MB
Mega byte	
GB	Giga byte

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

This research investigates how data visualisation on a Powerwall can substantially speed up the manual flagging of radio astronomy data by users, to reduce RFI (Radio Frequency Interference) and calibrators noise. Then this research brings together the areas of visualisation and radio astronomy.

Visualisation is a key component of the software that astronomers use to process data (e.g. CASA (Common Astronomy Software APPLICATION (Ott J., Kern J., 2013)). However, that software is designed for ordinary desktop displays. Powerwall are large scale visualisation systems which allows not only to obtain high resolution enhanced images but the possibility to display simultaneously multi dimensional parameters for a given application.

Radio astronomy is an observational technique of astronomy that uses cosmic waves for the study of the universe. This technique uses radio telescopes, generally parabolic antennas, in order to gather electromagnetic radiation from different astronomical sources. eMerlin is a radio astronomical interferometric system designed to be used initially in the study of star formation and black holes energetic processes. Technically, it consists in a 7 antennas array, that works as one, in order to obtain a higher angular resolution. In terms of the data processing, first the rough data from each antenna is centralised and correlated. Then, it is calibrated and processed in a dedicated pipeline (Argo M. K., 2015).

Figure 1.1: University of Leeds Powerwall.

This pipeline is a series of automatic processes that correct and adjust the data. Some of the data is not useful so it has to be eliminated.

For the purpose of this work, the undesired data is assumed to be RFI (Radio Frequency Interferences) and as calibration noise. The RFI is all the electromagnetic signals produced by human made electronic systems (satellites, air crafts, broadcast transmissions). The calibration noise is all the data obtained from misbehaviour of one or several elements or parameters analysed in the pipeline. Flagging is one of the calibration processes that takes place in the pipeline data treatment.

Specifically, the flagging is the identification and removal of undesired data. Usually, this flagging is performed automatically by specialised algorithms (Peck L. W. and Fenech D. M., 2013). However, this automated technique does not remove completely the noise in the information. So a visual inspection is required in order to complete the debug of the data. The visual inspection of undesired data for the manual flagging process is a time consuming key process in the data-treatment calibration.

There have been proposed many methods for the RFI mitigation (Offringa A. R., de Bruyn A. G., Biehl M., ZaroubiS S., Bernardi G. and Pandey V. N.,

2010) and (Ford J. M., Kaushal D. B., 2014). These methods mainly consisting algorithms to recognise certain patterns in the data. This patterns can be characterised in order to be automatically removed. Although many authors recognise the importance of the manual flagging process, it is not a recurrent research topic.

1.2 Research Question

Flagging is an important process for removing noise from radio astronomy data. Automatic flagging is effective for removing some types of noise, but manual flagging is also needed. However, manual flagging is very time-consuming because of the number of parameters that are involved and the number of visualisations that need to be checked. Powerwalls have the potential to provide a solution, because of their high-resolution. This research investigates how such a solution should be designed.

1.3 Approach

This research establishes several stages to address the stated problem: understanding of the data treatment and calibration (pipeline), identification of the parameters involve in the flagging process, design and implementation of visualisation environments, analysis of results and future work. The understanding of the data treatment involves the recognition of the architecture of the pipeline which is the tool in which the flagging process takes place. This conceptualisation allows to have an overview of the data processing process and therefore an understanding of the variables immerse in that process. The identification of the parameters used in the flagging process permits to recognise the main variables that have to be taken into account in order to define the guidelines for the design and implementation of the solution. This solution is based on a visualisation environment using pre-generated plots that helps the analyst to perform the manual flagging in a more effective way; this effectiveness is expressed in terms of time, interaction and accuracy.

1.4 Contributions

This project seeks to design visualisations which optimise the process of manual flagging for interferometric data. Specific contributions are:

1. Design of visualisation canvases, which exploit the large display real estate of Powerwalls for manual flagging (Chapter 3).
2. Implementation of software to generate canvases (Chapter 4).
3. Evaluation of the solution (Chapter 5).

1.5 Thesis outline

In order to address the project stages were developed as they are presented in each of the following chapters. This document is divided in 6 chapters:

Chapter 1 provides an overview of the project, mentioning the research question, approach and contributions. Then in chapter 2 are mentioned aspects related to radioastronomy and the characteristics of the data format in order to understand the complexity and challenges of the manual flagging process. Also is mentioned the eMerlin pipeline structure, what is flagging and how this process is actually done. Then a description of Powerwall configurations and visualization principles in big scale displays are explained. In Chapter 3 is explained the design process mentioning the parameters, criteria to select the variables and the conceptual design of the solution. Chapter 4 explains the process to implement the solution considering the plotting, merging and integration of the visualization solution to an users interface. Then in Chapter 5 an evaluation of the solution is presented. This evaluation was supported by two evaluation sessions performed by experts. Finally in Chapter 6 conclusions are provided.

Chapter 2

Background

2.1 Introduction

This thesis describes a Powerwall-based solution to optimise the manual flagging process for radio astronomy data. The solution is based on the development of a visualisation environment to reduce the amount of time required in the process of identification and removal of undesired data from the eMerlin radio telescopes array. Although the data passes through a dedicated pipeline in which the data is automatically processed, calibrated and flagged, this procedure is not enough to eliminate all the noise in the data. So human inspection is a process that still needs to be done and requires a lot of effort and time from the analyst.

The use of a Powerwall allows a new approach to the manual flagging process. Due to its characteristics, the Powerwall is able to visualise a great amount of data simultaneously. Also, the possibility of obtaining enhanced images of the required plots allows not only a benefit in time consuming but in the effectiveness and accuracy of the flagging process. This research integrates the understanding of the data handling architecture used in radio astronomy, the calibration processes, and the development of a solution to optimise the manual flagging process.

The following section describes the data format used in radio astronomy and more specifically in interferometric arrays data. After this the general architecture for the processing and the calibration of the data and the flagging process are explained. Then a manual flagging example is provided and its analysis is performed. Finally a description is provided of configurations and visualisation characteristics in Powerwalls.

2.2 Radio astronomy principles

Radio astronomy is an observational technique used in Astronomy involving the study of radio waves coming from space. These waves come from different types of radio sources such as stars, nebulae, galaxies, pulsars, the sun, the planets and molecules between the stars. The study of those radio waves makes it possible to know where they come from, how they are produced and the type of astronomical object involved. In order to understand the principles of radio astronomy some key terms will be mentioned:

2.2.1 Key concepts

In radio astronomy some key concepts are fundamental in order to understand the signals behavior and characteristics of the instrumentation involved. Amplitude indicates the voltage of a signal; it can also corresponds to the current level, field intensity or power level. In radio astronomy the unit that represents the intensity of a signal is the Jansky (Jy). It is the measure of the amount or radio energy per area in a specific frequency. The Frequency expresses the number of cycles of a wave in one second being the Hertz (Hz) its unit; one Hz corresponds to one cycle per second. The polarization, from an instrumental point of view, is related to the direction of of the waves perceived. Polarizations can be linear, horizontal or circular. For the purpose of this work it should be noted that right and/or left handed circular polarizations (RR or LL) are mainly used.

2.2.2 Atmospheric window

Due to the atmospheric window which explains how the atmosphere absorbs certain radiation, not all the cosmic waves can be perceived from earth. These waves are framed in the electromagnetic (em.) spectrum which categorizes the em. waves in function of the frequency. In Figure 2.1 is shown that the higher frequencies correspond to the Gamma Ray and the lowest frequencies to Radio.

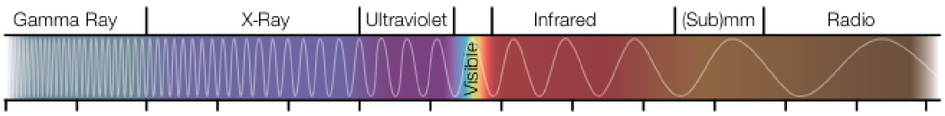


Figure 2.1: Electromagnetic spectrum chart.

As seen in Figure 2.2, the atmosphere absorbs most of the electromagnetic radiation. This is also called opacity. The bands or sections of that specter that can be perceived on the ground are the visible band and the radio band. In terms of frequency this radio band corresponds to a range between 10 Mega Hertz (MHz) to 1 Tera Hertz (THz). Those frequencies correspond to length waves between 10 millimeters to 10 meters. It is worth mentioning that some regions of the near ultraviolet, near infrared and some far infrared bands are perceptible on ground too.

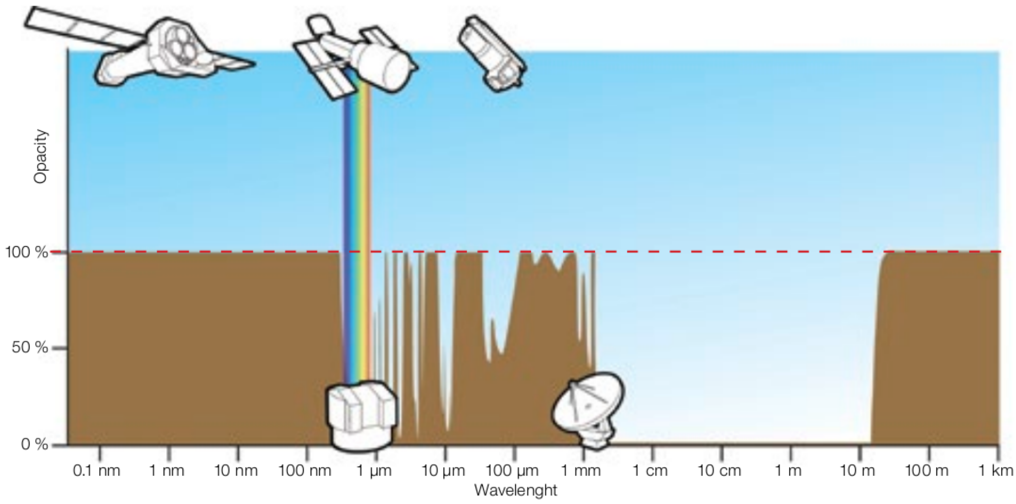


Figure 2.2: Atmospheric window (opacity generated by the atmosphere (percentage) v/s signal wavelength (m)).

Radio waves pass through the atmospheric window so radio telescopes can perceive radio cosmic waves. In bands such as IR (Infrared), UV (Ultraviolet), X ray or Gamma ray, telescopes have to be placed into space so that the atmospheric window does not block such emissions.

2.2.3 Cosmic Radio Sources

Many objects in the universe emit radio waves. This emission depends on physical conditions of the emitting objects involving electrons movement and loss of energy which generates radio emissions. Basically there are two main processes: thermal and non thermal. Thermal processes include slow moving electrons within hot clouds of gas that surround very hot stars. Non thermal processes corresponds to photon emissions due to electrons that have been accelerated through stellar or large scale explosions as supernovas. By studying those emissions we can infer, among other things, the chemical composition, evolution and distances of those objects.

2.2.4 Radio telescopes

Traditionally telescopes had been the most common tool used to study the universe. Nevertheless radio telescopes, which are the radio reception systems used in radio astronomy, are now used increasingly. Telescopes use the visible bandwidth of the electromagnetic spectrum (light waves). Radio telescopes reflect electromagnetic radio waves from objects in the universe and also from the interstellar medium.

Technically speaking, radiotelescopes are radio reception systems which can be used as a single system or as an array. These arrays generate a new technique called interferometry (Subsection 2.2.5). The process of how a radio telescope works is explained as follows. Radio emissions from space are gathered by an antenna. Those waves come in as an analog signal that has to be converted into a digital signal. Given the many signals obtained in an interferometric system, a synchronization pattern is needed to unify them. For that purpose atomic clocks add marks in the signals to synchronize them. After that, through correlation those signals are integrated into a unified data set. Finally the raw data is sent to a facility in which the information is processed, calibrated and analyzed. Figure 2.3 shows a conceptual diagram of ALMA (Atacama Large Milimeter Array).

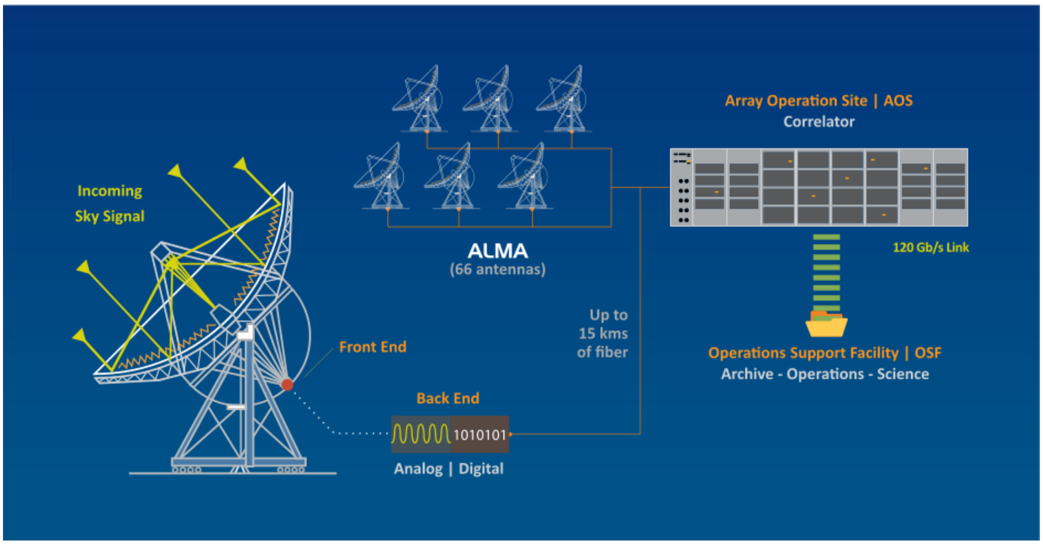


Figure 2.3: Signal processing using several radio telescopes).

2.2.5 Interferometry

A radio telescope is then a radio reception system that is designed to obtain cosmic waves in the range of radio frequencies. In radio astronomy one of the most used type of antennas is the paraboloidal antenna. An analogy used to explain the use of these antennas is that if in a rainy day you want to collect a large amount of water you use a big cup instead of a little one. In the same way, in radio astronomy the purpose is to obtain as much radiation as possible. So if you have a larger paraboloidal antenna you will obtain more radiation. Due to structural constraints it is impossible to build increasingly large antennas. The answer to this was the implementation of interferometry, a technique which uses several antennas that act as one big virtual antenna. This technique allows the use of as many antennas as you may have. One of the greatest benefits of this technique is to obtain higher angular resolution which basically is the smallest separation that you can measure or resolve. This resolution can be extrapolated to the magnification obtained in a telescope. If you use certain mirrors and lenses in a telescope you can enhance a certain area. In the case of radio telescopes, if you have a better angular resolution you can analyze deeper into space or examine a region in a more detailed way. Related to the angular resolution is the concept of baseline. In radio astronomy and more specifically in interferometry a baseline is the physical separation or distance between two antennas. The angular resolution is then the angle obtained from the relation of the wavelength of the signal over the diameter (D) of a single antenna that for the case of interferometry is translated in the maximum distance between the antennas of the array (longest baseline)(Figure2.4).

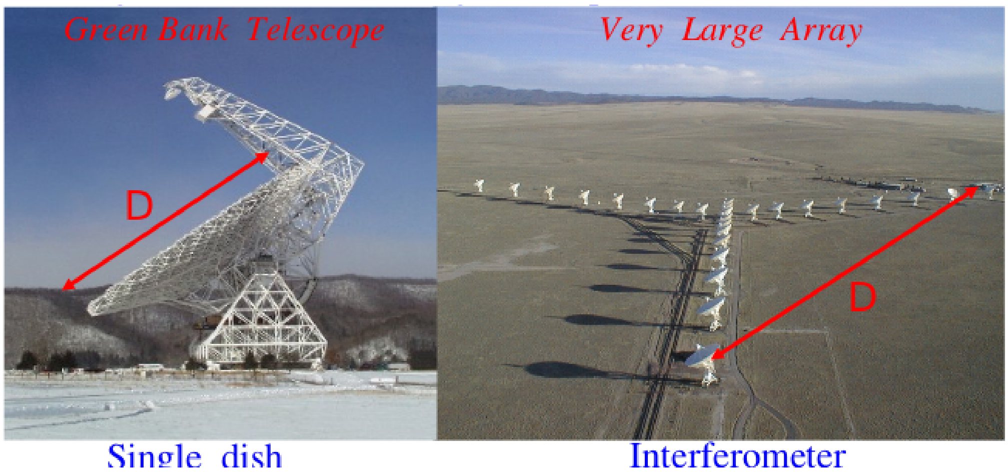


Figure 2.4: Diameter for a single dish and distance for an interferometric system (baseline).

2.2.6 Flagging

As explained, the cosmic signals go through several stages until the information is ready to be calibrated and analyzed. The method generally used for that purpose is a pipeline which is a software-based protocol that perform steps systematically to calibrate the data. It is in this process where the analysis for the removal of interferences and mis-calibrations has to be performed. This raises the concept of flagging which is the identification and elimination of bad data. This bad data can come either from radio frequency interferences generated by human made objects, such as satellites or radio communication systems, or by mis-calibration in the pipeline. Traditionally, flagging is performed by specialized algorithms or techniques. Nevertheless, this software and these processes do not remove all the undesired signals, so a human inspection has to be performed. This is what is called manual flagging. This manual flagging demands a lot of effort and is done in an intuitive way, depending on the preferences and personal methods of the analyst.

2.3 eMerlin radio astronomical array

The eMerlin (Multi Element Remotely Lined Interferometer Network) is an array of 7 radiotelescopes (Jodrell Bank Lovell, Jodrell Bank Mk2, Cambridge, Defford, Knockin, Darnhall, and Pickmere) distributed around UK. The technique that uses this configuration of antennas is called interferometry, which mainly is a technique that allows the use of several antennas that operate as one. This technique is used in order to enhance the angular resolution of the system. The incoming signals gather in the antennas are centralised and correlated in the Jodrell Bank Radio telescope by optical fibre. The eMerlin array operates at frequencies between 1.3 Giga Hertz (GHz) to 24 GHz and allows a resolution of 150 milliarcseconds. Another important factor is the sensitivity which is expressed in micro Jules (Jy). In order to establish the sensitivity of an interferometric system like eMerlin several aspects must be taken into account, such as integration time i.e. time aiming to the source (hr), number of antennas used in the array, frequency bands, bandwidths and apperture ratio (defined by the angular resolution of the system). The sensitivities obtain for the eMerlin according to the frequency bands are: 24 Jy for the L band (1.5 Giga Hertz (GHz)), 16 Jy for the C Band (5GHz) and 120 Jy for K Band (22GHz).

The eMerlin array is designed to be used for several astronomy techniques mainly in the fields of astrometry, polarimetry and spectroscopy. Some of the scientific questions that are intended to be address with this system are the history of star-formation and black hole growth as galaxies evolve, the physical processes which govern the formation of stars, the modes of activity in nearby galaxies and the energetic processes in relativistic outflows from jets generated by black holes and compact objects.

2.4 .ms Data Format

The data obtained from the eMerlin array is managed in a format called MS (Measurement Set) (Kemball A. J. and Wieringa M. H., 2000). This format was built according to the concepts proposed in the Measurement Equation 2.17 by Hamaker, Bregman and Sault, for radio astronomical data calibration (Hamaker J. P., Bregman J. D., Sault R. J., 1996). This equation describes the use of matrices for the polarimetric response of a radio interferometer, allowing to obtain general algorithms to the calibration of such system. Once the data is arranged in a .ms file, the parameters are segmented in specific columns. Each column is expressed in a a determinate data type. The Table 2.1 shows the columns in which the information is segmented according to determinate parameters, the data format of each parameter and an example value.

PARAMETER	DATA TYPE	EXAMPLE VALUE
UVW	Double array	[6431.55, 13472.4, -5539.68]
FLAG	Boolean array	[4, 128] Boolean
FLAG_CATEGORY	3 variable (boolean array)	[0, 0, 0] Boolean
WEIGHT	4 variables (float array)	[14.0318, 1.48005, 0.478257, 14.6269]
SIGMA	4 variables (float array)	[0.000353553, 0.000353553, 0.000353553, 0.000353553]
ANTENNA1	Integer	3
ANTENNA2	Integer	4
ARRAY_ID	Integer	0
DATA_DESC_ID	Integer	1
EXPOSURE	Integer	4
FEED1	Integer	0
FEED2	Integer	0
FIELD_ID	Integer	4
FLAG_ROW	Boolean	0
INTERVAL	Integer	4
OBSERVATION_ID	Integer	0
PROCESSOR_ID	Integer	-1
SCAN NUMBER	Integer	36
STATE_ID	Integer	-1
TIME	String	05/05/15 10:40 PM
TIME_CENTROID	String	05/05/15 10:40 PM
DATA	Complex array	[4, 128] Complex
CORRECTED DATA	Complex array	[4, 128] Complex
MODEL DATA	Complex array	[4, 128] Complex

Table 2.1: List of the parameters within a .ms file specifying the data type and providing an example of each parameter.

In order to visualise and manipulate the data, dedicated software is used. In this case the software is CASA (Common Astronomical Software Applications). This software provides an environment to make the calibration of interferometric data (Ott J., Kern J., 2013). It is possible to develop the complete processing of the data from the rough data base to the generation of an image. CASA has specific application as plotms which allows to produce plots with all the variables immerse in the data treatment.

2.4.1 Data handling architecture

In order to develop a pertinent solution, this project uses a real eMerlin data base for analysis purposes. The file size is 10 GB and is obtained in a compressed .tar file. Initially an inspection of the data is performed in order to identify the technical characteristics of the astronomical observations. For this purpose the commands listobs and tablebrowse are used. Figures 2.5 and 2.6 are obtained by using the listobs command. To obtain the visualisation as shown in Figures 2.7, 2.8, and 2.9 browsetable command was used. These commands provides a different kind of visualisation of the data architecture or structure within the database.

2.4.2 Ms file data segmentation

Once the rough data file is available, the first step is to unpack the file. A first approach to the data is obtained by calling the .ms file from a determinate observation using the listobs command. A first view is displayed in order to identify the rows and columns structure of the database. Figure 2.5 shows the columns structure divided in Date, Timerange (UTC), Scan, Field, Field Name, nRows, SpwIds (Spectral window identification), Average Interval(s) and ScanIntent.

Figure 2.6 provides the following information: field's ID, Code, Name (of the source), RA (right ascension), Declination, Epoch, SrcId (Source identification) and nRows, spectral window, spwID, name, the number of channels, frame, Ch0 (Channel frequency/MHz), CahnWid (Channel width/kHz), TotBW (Total Band width/kHz), CtrFreq(Central frequency/MHz) and Corrs (Polarisations/RR-RL-LR-LL), antenna's ID (of antennas of the array), Name, Station (code), Diam (Diameter), Long (Geographic Longitude), Lat (Geographic Latitude), Offset from array center (East, North and Elevation), and the ITRF geocentric coordinates (m / x,y and z axis).

```
Date          Timerange (UTC)            Scan  FldId FieldName             nRows     SpwIds    Average Interval(s)     ScanIntent
05-May-2015/18:32:09.5 - 18:35:31.3        1       0 1302+5748              3000  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            18:42:31.0 - 18:55:35.0        2       0 1302+5748             11700  [0,1,2,3]  [4, 4, 4, 4]
            18:55:39.0 - 19:02:38.5        3       3 1252+5634              6300  [0,1,2,3]  [3.99, 3.99, 3.99, 3.99]
            19:02:41.0 - 19:05:41.5        4       0 1302+5748              2700  [0,1,2,3]  [3.98, 3.98, 3.98, 3.98]
            19:05:44.0 - 19:12:38.0        5       3 1252+5634              6180  [0,1,2,3]  [4.01, 4.01, 4.01, 4.01]
            19:12:41.0 - 19:25:37.5        6       0 1302+5748             11580  [0,1,2,3]  [4, 4, 4, 4]
            19:25:40.0 - 19:32:35.5        7       3 1252+5634              6240  [0,1,2,3]  [3.99, 3.99, 3.99, 3.99]
            19:32:38.0 - 19:45:28.7        8       0 1302+5748             11460  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            19:45:33.0 - 19:52:34.3        9       3 1252+5634              6300  [0,1,2,3]  [4, 4, 4, 4]
            19:52:37.0 - 19:55:36.5       10       0 1302+5748              2700  [0,1,2,3]  [3.98, 3.98, 3.98, 3.98]
            19:55:39.0 - 20:02:05.0       11       3 1252+5634              5760  [0,1,2,3]  [4.01, 4.01, 4.01, 4.01]
            20:02:08.0 - 20:05:09.3       12       0 1302+5748              2700  [0,1,2,3]  [4, 4, 4, 4]
            20:05:11.0 - 20:12:33.7       13       3 1252+5634              6600  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            20:12:36.0 - 20:15:34.8       14       0 1302+5748              2640  [0,1,2,3]  [4.05, 4.05, 4.05, 4.05]
            20:15:37.0 - 20:22:30.3       15       3 1252+5634              6180  [0,1,2,3]  [4, 4, 4, 4]
            20:22:34.0 - 20:25:08.7       16       0 1302+5748              2280  [0,1,2,3]  [4.05, 4.05, 4.05, 4.05]
            20:25:11.0 - 20:32:05.7       17       3 1252+5634              6180  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            20:32:09.0 - 20:35:03.7       18       0 1302+5748              2580  [0,1,2,3]  [4.05, 4.05, 4.05, 4.05]
            20:35:06.0 - 20:42:01.5       19       3 1252+5634              6240  [0,1,2,3]  [3.99, 3.99, 3.99, 3.99]
            20:42:04.0 - 20:45:01.2       20       0 1302+5748              2640  [0,1,2,3]  [4, 4, 4, 4]
            20:45:04.0 - 20:52:10.7       21       3 1252+5634              6360  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            20:52:14.0 - 20:55:03.3       22       0 1302+5748              2520  [0,1,2,3]  [4, 4, 4, 4]
            20:55:06.0 - 21:02:00.8       23       3 1252+5634              6180  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            21:02:04.0 - 21:05:01.2       24       0 1302+5748              2640  [0,1,2,3]  [4, 4, 4, 4]
            21:05:04.0 - 21:12:02.8       25       3 1252+5634              6240  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            21:12:05.0 - 21:15:03.0       26       0 1302+5748              2640  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            21:15:05.0 - 21:22:00.5       27       3 1252+5634              6240  [0,1,2,3]  [3.99, 3.99, 3.99, 3.99]
            21:22:03.0 - 21:25:01.0       28       0 1302+5748              2640  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
            21:25:03.0 - 21:32:01.0       29       3 1252+5634              6240  [0,1,2,3]  [4.01, 4.01, 4.01, 4.01]
            21:32:04.0 - 21:35:02.0       30       0 1302+5748              2640  [0,1,2,3]  [4.02, 4.02, 4.02, 4.02]
```

Figure 2.5: Output from the listobs command (timeranges, number of scan, number of rows, averages and intervals).

```
Fields: 5
  ID   Code Name               RA               Decl           Epoch        nRows
  0         1302+5748          13:02:52.465277 +57.48.37.60932 J2000       201060
  1         0319+415           03:19:48.160110 +41.30.42.10330 J2000        53940
  2         1407+284           14:07:00.394410 +28.27.14.68990 J2000        38640
  3         1252+5634          12:52:26.285900 +56.34.19.48800 J2000       405600
  4    ACAL 1331+305           13:31:08.287300 +30.30.32.95900 J2000        40440
Spectral Windows:  (4 unique spectral windows and 1 unique polarization setups)
  SpwID  Name   #Chans   Frame   Ch0(MHz)  ChanWid(kHz)  TotBW(kHz) CtrFreq(MHz)  Corrs
  0      none     128    GEO     4816.500     1000.000    128000.0   4880.0000   RR  RL  LR  LL
  1      none     128    GEO     4944.500     1000.000    128000.0   5008.0000   RR  RL  LR  LL
  2      none     128    GEO     5072.500     1000.000    128000.0   5136.0000   RR  RL  LR  LL
  3      none     128    GEO     5200.500     1000.000    128000.0   5264.0000   RR  RL  LR  LL
The SOURCE table is empty: see the FIELD table
Antennas: 6:
  ID   Name  Station   Diam.    Long.         Lat.                Offset from array center (m)                ITRF Geocentric coordinates (m)
                                                                     East         North     Elevation               x               y               z
  0    Mk2   Mk2       24.0 m   -002.18.08.9  +53.02.58.7      19713.9103    20897.1596     6334.4681  3822473.365000  -153692.318000  5085851.303000
  1    Kn    Kn        25.0 m   -002.59.44.9  +52.36.18.4     -26733.5549   -28428.6831     6480.6814  3859711.503000  -201995.077000  5056134.251000
  2    De    De        25.0 m   -002.08.35.0  +51.54.50.9      30394.6148  -105100.8391     6688.6339  3923069.171000  -146804.368000  5009320.528000
  3    Pi    Pi        25.0 m   -002.26.38.3  +53.06.16.2      10235.7831    26985.6080     6271.3637  3817176.561000  -162921.179000  5089462.057000
  4    Da    Da        25.0 m   -002.32.03.3  +52.58.18.5       4186.5058    12262.8833     6330.1699  3828714.513000  -169458.995000  5080647.749000
  5    Cm    Cm        32.0 m   +000.02.19.5  +51.58.50.2     176561.6720   -97724.9405     6660.8614  3919982.752000     2651.982000  5013849.826000
```

Figure 2.6: Output from ths listobs command (Fields of observation, spectral windows and antennas).

The previous screenshots are taken from the CASA external terminal. Another possibility of interaction in terms of displaying of the data is by using the command browsetable

Figure 2.7 shows the uvw, flagging process (array), the flag category, weight, sigma, antenna1, antenna2, array identification, data description identification, exposure, feed1, feed2 and field identification.

	UVW	FLAG	FLAG_CATEGORY	WEIGHT	SIGMA	ANTENNA1	ANTENNA2	ARRAY_ID	DATA_DESC_ID	EXPOSURE	FEED1	FEED2	FIELD_ID
15	[1038.23, -10882.3, 26...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	0	3	0	0	4	0	0	0
16	[58364.6, 853.771, 34...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	0	1	0	0	4	0	0	0
17	[-57326.3, -11736.1, -3...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	1	3	0	0	4	0	0	0
18	[-44132.3, -5875.74, -2...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	1	4	0	0	4	0	0	0
19	[77970.7, 94933.2, 30...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	0	2	0	0	4	0	0	0
20	[19606.1, 94079.5, -40...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	1	2	0	0	4	0	0	0
21	[-76932.4, -105816, -27...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	2	3	0	0	4	0	0	0
22	[-63738.4, -99955.2, -2...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	2	4	0	0	4	0	0	0
23	[-89220.2, 99018.7, -67...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	2	5	0	0	4	0	0	0
24	[14232.3, -5021.97, 93...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	0	4	0	0	4	0	0	0
25	[13194.1, 5860.32, 67...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	3	4	0	0	4	0	0	0
26	[-11249.5, 193952, -37...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	0	5	0	0	4	0	0	0
27	[-69614.1, 193098, -71...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	1	5	0	0	4	0	0	0
28	[-12287.7, 204834, -39...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	3	5	0	0	4	0	0	0
29	[-25481.8, 198974, -46...	[4, 128] Boolean	[0, 0, 0] Boolean	[0, 0, 0, 0]	[0.000408247, 0.00040824...	4	5	0	0	4	0	0	0

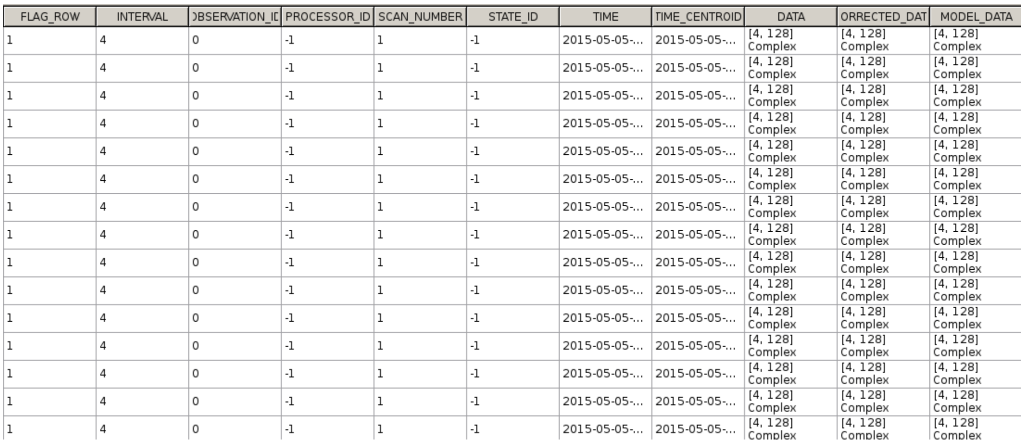
Figure 2.7: Visualisation of information using browsetable command.

Figure 2.8 shows the flag row, interval, observation identification. processor identification, scan number, state identification, time, time center identification, data(array/complex numbers), corrected data (array/complex numbers) and model data (array).

In Figure 2.9 is shown the complex array structure of where the data is located. It is an array of 4x128. Each of the four rows corresponds to each of the polarisations (RR, LL, RL AND LR) and the 128 columns corresponds to each specific frequency or channel. In conclusion each of this complex data array corresponds to all the polarisation for a specific spectral window; each spectral window has 128 frequencies.

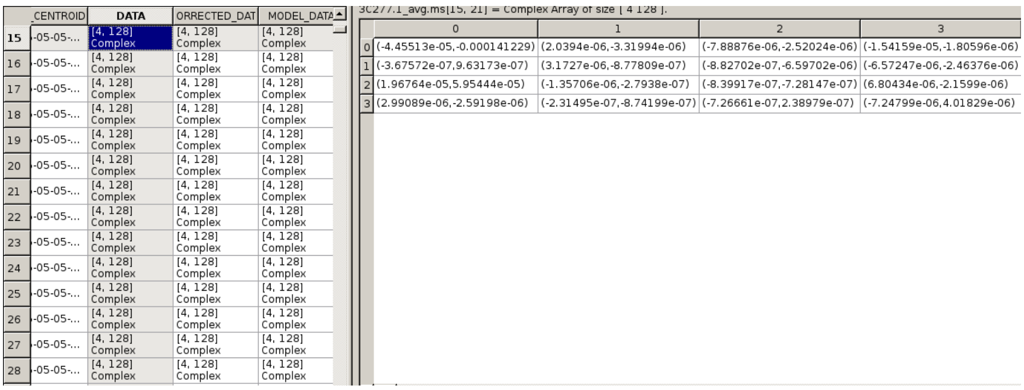
2.4.3 Main parameters

Certain parameters are important to be known and identified for the purpose of the manual flagging process. The concepts of field, spectral window, channel,



FLAG_ROW	INTERVAL	)BSERVATION_ID	PROCESSOR_ID	SCAN_NUMBER	STATE_ID	TIME	TIME_CENTROID	DATA	ORRECTED_DAT	MODEL_DATA
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
1	4	0	-1	1	-1	2015-05-05-...	2015-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex

Figure 2.8: Visualisation of information using browsetable command.



	_CENTROID	DATA	ORRECTED_DAT	MODEL_DATA
15	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
16	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
17	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
18	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
19	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
20	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
21	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
22	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
23	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
24	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
25	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
26	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
27	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex
28	-05-05-...	[4, 128] Complex	[4, 128] Complex	[4, 128] Complex

3C277.1_avg.ms[15, 21] = Complex Array of size [4 128].

	0	1	2	3
0	(-4.45513e-05,-0.000141229)	(2.0394e-06,-3.31994e-06)	(-7.88876e-06,-2.52024e-06)	(-1.54159e-05,-1.80596e-06)
1	(-3.67572e-07,9.63173e-07)	(3.1727e-06,-8.77809e-07)	(-8.82702e-07,-6.59702e-06)	(-6.57247e-06,-2.46376e-06)
2	(1.96764e-05,5.95444e-05)	(-1.35706e-06,-2.7938e-07)	(-8.39917e-07,-7.28147e-07)	(6.80434e-06,-2.1599e-06)
3	(2.99089e-06,-2.59198e-06)	(-2.31495e-07,-8.74199e-07)	(-7.26661e-07,2.38979e-07)	(-7.24799e-06,4.01829e-06)

Figure 2.9: Polarisation and data allocation within the data base.

scan, baseline and polarisation will be recurrent in the further processes. Following there is a description of the main parameters and the location within the data file.

Field of observation: the field of observation is the astronomical source of observation. It includes the calibration sources, required for the verification of the correct performance of the system. This variable is defined in the FIELD-ID parameter column of Table 2.1.

Spectral window/channel/frequency: the spectral window, channel and frequency parameters are related. The spectral window is the range of frequencies in which the observations are segmented. The channels are determinate respect to frequencies and also in relation to each spectral window (Figure 2.10). The frequency information corresponds to the data parameter. This is a complex array of 4 rows x 128 columns (Figure 2.11). Each of the 128 columns corresponds to one frequency. The channels or specific frequencies are defined in the DATA-DESC-ID parameter column of Table 2.1.

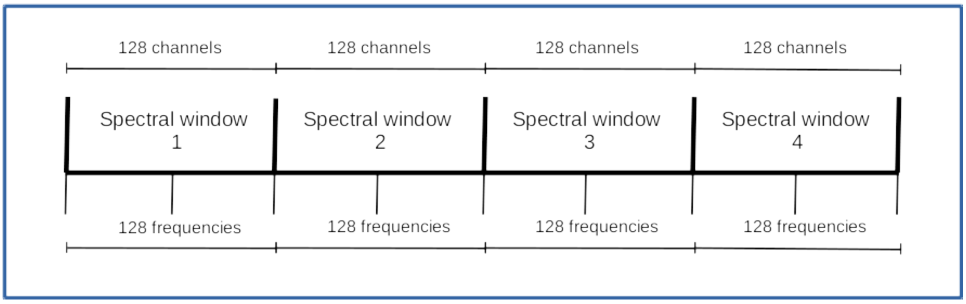


Figure 2.10: Relation between spectral window, channel and frequency.

Scan: the scan is determinate by a given number of rows of the database, according to the number of observations of a specific astronomical target. There are certain numbers of scans for each field of observation. The scan consecutive corresponds to the SCAN-NUMBER parameter column of Table2.1.

Baseline: the baseline is established in relation to the uv distance or physical baseline length between two antennas. The baseline is defined by the relation between two antennas specified in the columns ANTENNA1 and ANTENNA2. The values for the baseline are specified in the UVW parameter column of Table2.1.

Polarisation: the polarisation is expressed in four different types RR, LL, RL and LR. The polarisations are defined in the DATA parameter; they are related in each of the 4 rows of the 4x128 complex array (Figure 2.11).

	0	1	2	3	4	5
0	(0.00212635,0.00179103)	(0.00439044,-0.00661301)	(0.0073744,0.00379163)	(1.1586,-0.0791089)	(0.572468,-0.231151)	(-0.354202,0.550643)
1	(-0.00149082,0.00439923)	(-0.0020426,-0.000829371)	(-0.00508706,-0.000862557)	(0.567501,-0.755674)	(0.603071,0.521907)	(0.385666,-0.115025)
2	(-0.00212461,0.000662333)	(-0.00909242,-0.00115153)	(0.000769782,-0.000169141)	(0.214018,0.00315091)	(0.143947,-0.549328)	(0.172485,-0.250419)
3	(0.00693933,0.00215649)	(0.0068203,-0.0050204)	(-0.0036293,-0.00662244)	(-0.855519,-1.22357)	(-0.164335,0.00190679)	(-0.0383436,-0.224093)

Figure 2.11: .ms file polarisations and channels parameters.

2.4.4 Specifications of the 3C321.ms file (eMerlin database)

In order to understand the complexity of the manual flagging mainly due to the number of parameters immerse in that process the characteristics of an eMerlin database used to perform the analysis will be mention. This understanding will guide the design of the method for the optimisation of the manual flagging process. The used database is called "3C321". This is a measurement set file (.ms). This database was obtained from the eMerlin interferometric array. The most relevant variables will be determinate in accordance to the information available in the 3C321.ms file. First the specifications for fields, spectral windows, frequencies, antennas and baselines configurations will be identify. This identification will be supported by the visualisation of the real database by screenshots taken directly from the .ms file. Then a conceptual analysis of how the variables are processed will be presented.

2.4.4.1 eMerlin 3C321.ms data file characteristics

After the inspection of the data base the most relevant parameters to be taken into account are fields, spectral windows, antennas and baselines. This parameters are selected mainly because they are directly involve in the manual flagging and their use is recurrent. As seen in Table 2.1 there are many parameters that are part of an interferometric database. The criteria to select field of observation, spectral window, antennas and baselines as main parameters for this solution was mainly oriented to the use of few variables that integrates the rest of parameters. This integration intends to provide a visualisation perspective that summarises the signal behaviour in order to identify anomalous patterns. This decision was also oriented by the opinion of an eMerlin analyst. In Figure 2.12 are the specifications of those parameters.

FIELDS

code	reference	Purpose of the field of observation	# rows (scans)
0	1302+5748	phase reference source	201,060
1	0319+415	bandpase, polarization leakage calibrator	53,940
2	1407+284	bandpass calibrator	38,640
3	1252+5634	target	405,600
4	1331+305	flux scale/polarization angle calibrator	40,400

SPECTRAL WINDOWS

Id	# channels	central frequency (MHz)	frequencies range
0	128	4,880	4816.5-4944.5
1	128	5,008	4944.5-5072.5
2	128	5,136	5072.5-5200.5
3	128	5,264	5200.5-5392.5

ANTENNAS

Id	Reference
0	Mk2
1	Kn
2	De
3	Pi
4	Da
5	Cm

Baselines antennas configurations

0-3	2-5
0-1	0-4
1-3	3-4
1-4	0-5
0-2	1-5
1-2	3-5
2-3	4-5
2-4	

Figure 2.12: eMerlin 3C321.ms data base technical characteristics.

1. Fields: 5 fields of observation were identified. Each field has a code that is related to the reference of the source of observation. The number of rows in the data base that corresponds to the amount of scans are the number of observations performed for a specific field of observation. Every field of observation has a purpose. It could be for calibrating purpose or for the observation of an astronomical source.

2. Spectral Window (Spw): There are 4 spectral windows, each one with 128

channels assigned. Each channel corresponds to a frequency depending on the spectral window where is located.

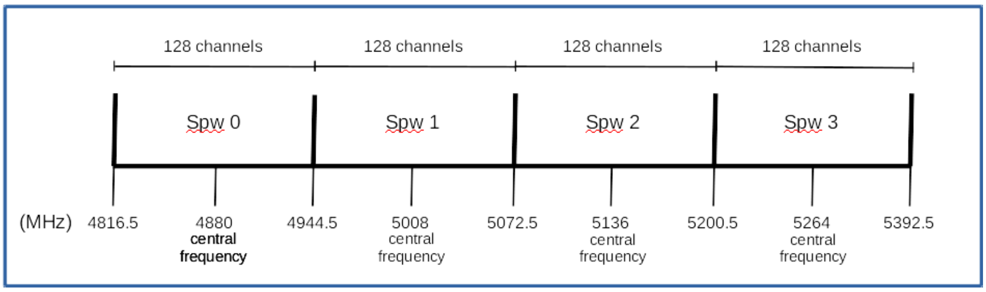


Figure 2.13: Conceptual diagram of the spectral windows, frequencies and channels assignation for the 3C321.ms eMerlin database.

3. Antennas: This data base uses 6 antennas, each one of them with its identification nomenclature.

4. Baselines: The information is ordered by segments of 15 antennas combinations. This means that one scan is considered as the information of that combinations to a same field n-times depending on the number of arrows.

2.4.4.2 Data structure

The data structure could be conceptualised as a 4 dimensions system. This dimensions consider baseline, time, frequency and polarisation. If all the variables except the polarisation are considered a 3 dimensional plot could be inferred.

In the Figure 2.14 is represented how the data can be conceptualised by considering all the frequencies for all the baselines and all the times. This could be extrapolated for each polarisation. Then if a specific frequency wants to be study, conceptually the representation will result as is showed in Figure 2.15.

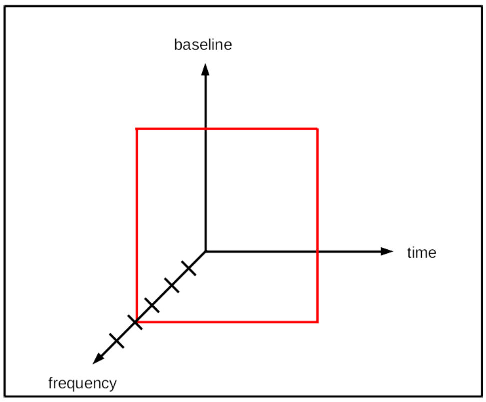


Figure 2.14: 3D data structure conceptual diagram (all baselines, all frequencies and all times).

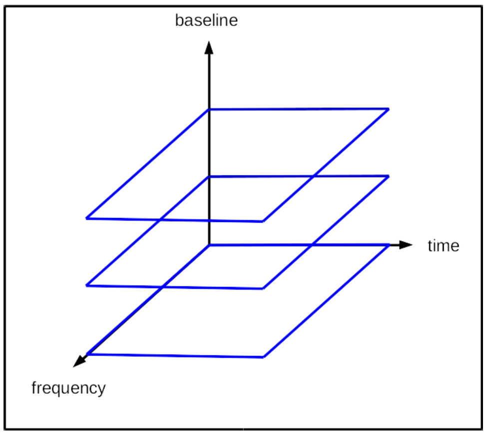


Figure 2.15: 3D data structure conceptual diagram (all baselines, all frequencies and all times).

2.5 Pipeline (processing and calibrating architecture)

After the rough data is obtained, the information is pre-processed and calibrated. A pipeline is used in order to have a unified protocol for the treatments of the data (Argo M. K., 2015) and (Moldon J., 2018). The pipeline integrates the processes immerse in the data treatment. This pipeline regulates the flow in which the data is processed. The main processes that integrate the data pre-processing and calibration procedure are shown in the Figure 2.18.

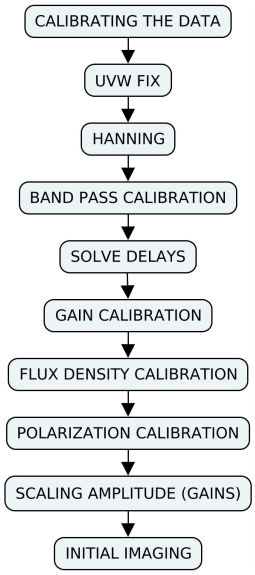


Figure 2.16: Main phases of the eMerlin pipeline.

The description of each phase is described as follows:

CALIBRATING THE DATA: Once the data is acquired and available in a .ms file, the first step is to examine and edit it. This procedure is completed with the synthesis calibration which is designed specifically implemented for that. It uses the synthesis calibration formula to relate the observed visibility between antennas to the true visibility. The visibility is expressed as a function of the frequency and spatial wave numbers.

$$V'_{i,j}(u,v,f)=b_{i,j}(t)[B_i(f,t)B_j(f,t)]g_i(t)g_j V_{i,j}(u,v,f)e^{i(\theta_i(t)-\theta_j(t))}$$

Figure 2.17: Synthesis calibration formula.

u and v: spatial wave numbers (as function of frequency f) g1 and omega1: amplitude and phase (complex gain) (expressed as function of time due to occasionally variations with temperature, atmospheric conditions and related variable)

Bi: complex band pass (the instrumental response as a function of frequency f; it also may vary as a function of time.

b(t): Baseline term. It is related to the antenna positions.

UVFIX: The uv fix corresponds to the solution of complex visibility parameters in order to define the coordinates parameters of the observations.

HANNING: A hanning filter is a smoothing function that is use in order to reduce the ringing, which is an impulsive oscillatory signal. Nevertheless this technique has effects in lowering the spectral resolution of the system.

BAND PASS CALIBRATION: As the data is taken in a spectral line mode a band pass solution is required. Spectral line mode is a correlation configuration (integration method) that uses small bandwidths to observe independent segments of frequency at a time. This process considers a band pass calibrator as reference to calculate the gain variations in relation to the frequency.

SOLVE DELAYS: As part of the band pass calibration antenna-based delays has to be solved. The delays of each antenna are solved by taking a reference antenna as pattern and then by the analysis of the phase and frequency signals behaviour for each spectral window.

GAIN CALIBRATION: To derive corrections for the complex antenna gains, absolute magnitude of the gain amplitudes are determinate by reference to a standard flux calibrator that is close to the target source; this is done in order to minimise differences generated by the atmosphere. If relative gain amplitudes and phases for different antennas are established, an absolute flux density scale can be achieved by comparing the gain amplitudes derived for the observation source with the ones derived from the source calibrator.

FLUX DENSITY CALIBRATION: Provides a flux density value for the amplitude calibrators. eMerlin uses 3C286, 3C86 and OQ208 as calibrators.

POLARISATION CALIBRATION: First, the instrumental polarisation has to be solved using an non-polarised source. Then the polarisation position angle is solved using a source with a known polarisation angle. Then LL and RR delays has to be solved as well as cross-hand delays RL and LR.

SCALING AMPLITUDE (GAINS): For this, a second calibrator has to be used. Using the primary flux calibrator to determine the system response to a source of known flux density, allows to know the flux density for the secondary calibrator.

INITIAL IMAGING: Once applied all the calibrations, and splitting off the target data into a separate measurement set, an initial image is possible to be obtained.

2.6 Flagging

The process of identification and removal of undesired data is called flagging and is one of the most critical data treatment phases in order to obtain a coherent result of the acquired data. It is not possible to establish a standard method or use only an automatic tool in order to perform the required inspections of the data due to the specific characteristics of each interferometric system.

For the purpose of this research the undesired data consists of two main sources: RFI (Radio Frequency Interferences) and calibration noise.

RFI has a enormous impact in the astronomical observations. This RFI is generated by man-made sources that emit electromagnetic radiation (computers, satellites, automobiles, telecommunications broadcasts). Although some bands are legally protected for the use of radio astronomy by the ITU International Telecommunication Union, RFI still compromises data from the weak cosmic signals in radio astronomy assigned bands (Wilson T. L., Rohlfs K., Huttemeister S., 2009). Calibration noise is either the resulting data from the misbehaviour of one of the systems in the radio telescopes systems e.g. a damaged antenna; or produced as a miscalibration of one of the parameters. In order to eliminate this data a flagging process has to be performed.

The flagging process usually is done by automatic algorithms. The eMerlin uses Serpent which is an automated RFI mitigation software (Peck L. W. and Fenech D. M., 2013). Many techniques for the RFI mitigation have been proposed (Offringa A. R., de Bruyn A. G., Biehl M., ZaroubiS S., Bernardi G. and Pandey V. N., 2010) (Ford J. M., Kaushal D. B., 2014) (Fridman P. A., Baan W. A., 2001) (Baan W. A., 2011). Specifically post-correlation thresholding, surface fitting and smoothing, cumulative sum, combinatorial thresholding and singular value decomposition are some of the most recurrent used techniques. Post correlation thresholding is mostly used to remove strong radio frequency interferences using mean or media values to establish ranges to determine flagging limits. Surface fitting and smoothing contrast the characteristic sharp edges shapes of RFI obtained in the time-frequency domain against smooth surfaces with small changes in frequency and time of astronomical sources; it is worth mentioning that this method is not used for strong lines sources such as pulsars. The cumulative sum is a method used to identify changes in distribution parameters which for the purpose of RFI detection estimates the variance of the signal taking a certain frequency obtained by one antenna as reference. Combinatorial thresholding considers the relation of frequency and time in RFI by establishing limits for sample combinations of those parameters. Singular value decomposition is a mathematical approach to find specific values of a matrix(dimensions U-V) to obtain properties of the elements used (flux, baseline-frequency index and time index).These methods mainly consist of algorithms to recognise certain patterns in the data. These patterns can be characterised in order to be automatically removed. Nevertheless these automated techniques do not eliminate all the RFI so a manual inspection is still required.

Manual flagging process demands human interaction. In this process the analyst produces and examine a series of plots to the identification of the source of any anomaly in the data patterns. Although many authors acknowledge the importance of the manual flagging process, this area of research has not been enough pursued, mainly arguing that data sizes are too big.

In order to understand the manual flagging process two Levels Of Detail (LOD) will be explain. LOD1 defines the stages in which the flagging takes place within the pipeline. In LOD2 it is explained a manual flagging procedure by establishing a conceptual diagram of the required steps and the performance of a practical example. Finally an efficiency analysis will be provided in order to identify and evaluate the manual flagging procedure.

2.6.1 LOD1 (Phases of flagging)

Flagging process takes place in several phases of the data treatment in the pipeline. There are 4 stages in which the flagging is developed. These stages are: between uvwfix and hanning; hanning and band pass calibration; band pass calibration and delays solving; and after all calibrations are done before the initial imaging.The stages in which the flagging is performed through the pipeline are shown in the Figure 2.18.

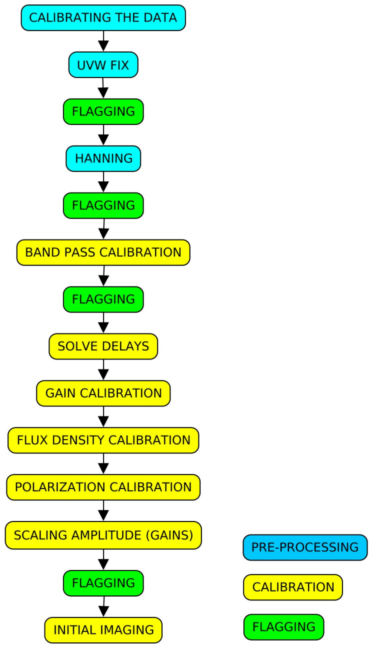


Figure 2.18: Stages of the eMerlin pipeline.

Even after the rough data is passed through automatic flaggers, human interactive inspection has to be developed by a specialist in order to remove residuals of noise in the information. Usually, the manual flagging procedure is performed in the last flagging stage. In this procedure a specialist has to watch initially a frequency vs time plot in which he can identify some kind of anomaly that the auto-flagger did not correct or eliminate. Then the analyst has to evaluate several plots combinations using different variables such as baselines polarisations, spectral window, channels and fields. When the pipeline has perform all the required calibrations the pipeline generates a serious of plots Amplitude vs Time, Phase vs Time, Amplitude vs Frequency, Phase vs Frequency, Amplitude vs UV wave, Frequency vs UV wave, among others. This analysis is done in order to establish the source of the undesired data.

2.6.2 LOD2 (Manual flagging procedure)

This section explains manual flagging using a simple example. As mentioned in the Table 2.1 , there are many variables that the analyst has to take into account. To provide clarity in the flagging process an example will be developed.

The Figure 2.19 the conceptual diagram of the example provided.

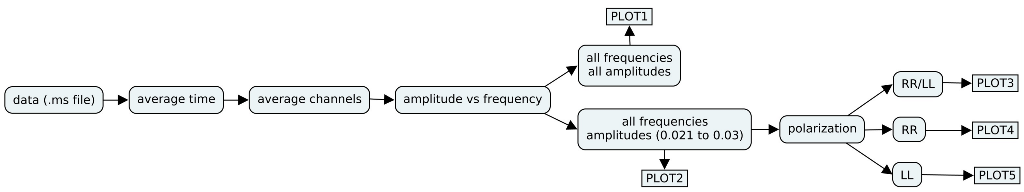


Figure 2.19: Steps of a manual flagging process.

Due to the fact that there are many many observations and many frequencies in the data base an average for the time and frequency is established to obtain an approximation of the behaviour of the signals for all the observation times and all the frequencies. Then the first two plots are done. The first one is frequency vs amplitude. The second one corresponds to the information delimited by amplitude. After this the behaviour of the selected data respect to polarisations is analysed . So three plots are done for the polarisation analysis. The first one provides a relation between RR and LL polarisations. The second one only for RR polarisation and the third one for LL polarisation. This is done to identify the source of the corrupted data. After this plots are done, an evaluation of the results provides the required information to flag certain specific data.

To provide more clarity in the manual flagging process an example will be developed in the CASA terminal using plotms. The following example considers field=1, and averaging of time and channels.

In first instance a plot of frequency vs amplitude is done in order to establish the response of the intensity of the signals respect to the frequencies (Figure 2.20). The averaging is established in order to have a general perspective and understanding of the complete observation. If a more detailed analysis wants to be done, a delimitation of time segments would have to be performed.

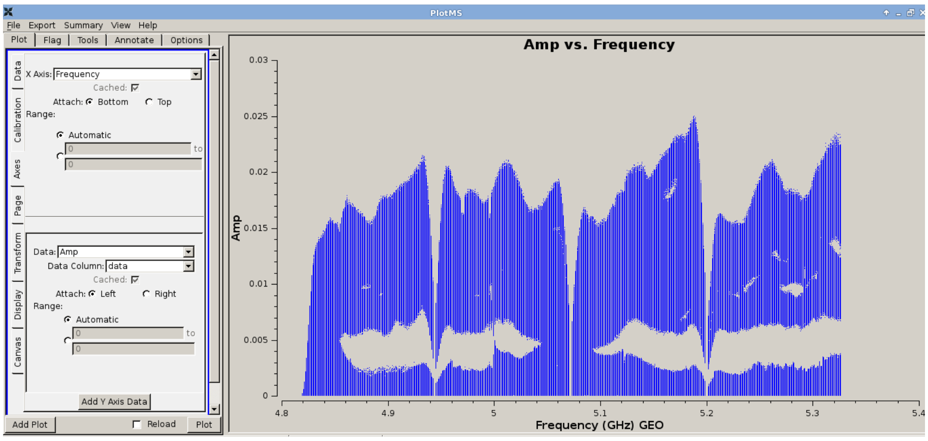


Figure 2.20: Amplitude vs Frequency plot (PLOT1).

After analysing the first plot an anomaly is found. The signals with an amplitude over 0.022 Janskys (Jy) require to be analysed. So the amplitude is delimited between 0.022 Jy and 0.03 Jy. Therefore we obtain the plot illustrated in Figure 2.21. The default colors provided by CASA plotms delivers a green-blue plot. It is worth mentioning that according to the rules of color-blindness these parameters should be changed for future applications and/or versions of this solution.

Then, the polarisation behaviour of such signals needs to be analysed. So a distinction in colours between polarisations RR and LL is done. From the analysis of this plot, it is inferred that the signals respond differently to each polarisation (Figure 2.22).

The next step is to provide separated plots for RR and LL polarisations. In the RR plot there are 2 groups of signals; between 5.15 GHz and 5.193 GHz, and another between 5.309 GHz and 5.325 GHz. This patterns show that the undesired signals have a RR polarisation in the mentioned frequencies (Figure 2.23).

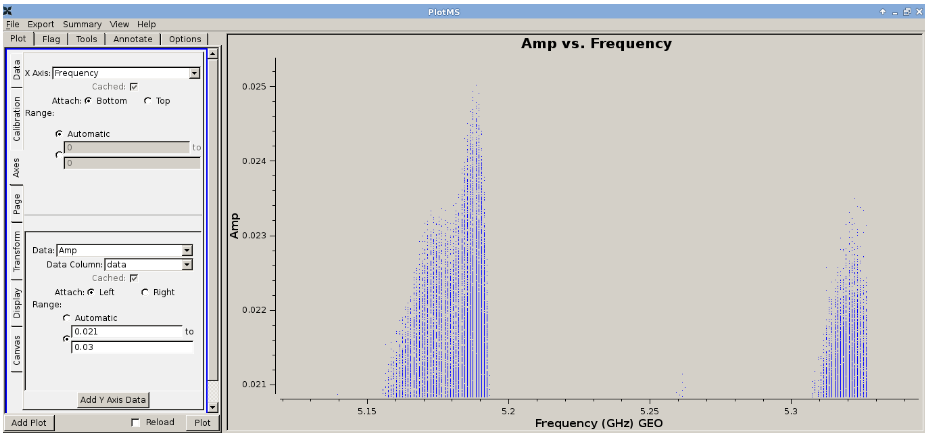


Figure 2.21: Amplitude (from 0.021 to 0.03 (Jy)) vs Frequency plot (PLOT2)

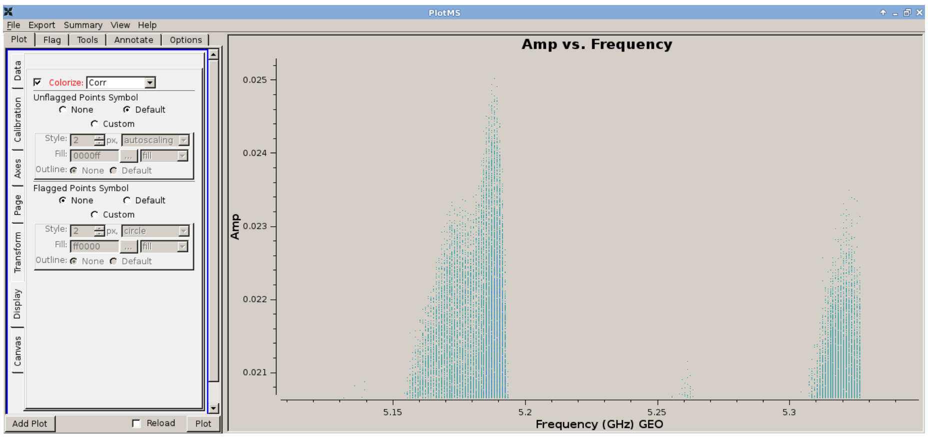


Figure 2.22: Amplitude (from 0.022 to 0.03 (Jy)) vs Frequency with colourised RR and LL polarisations plot (PLOT3).

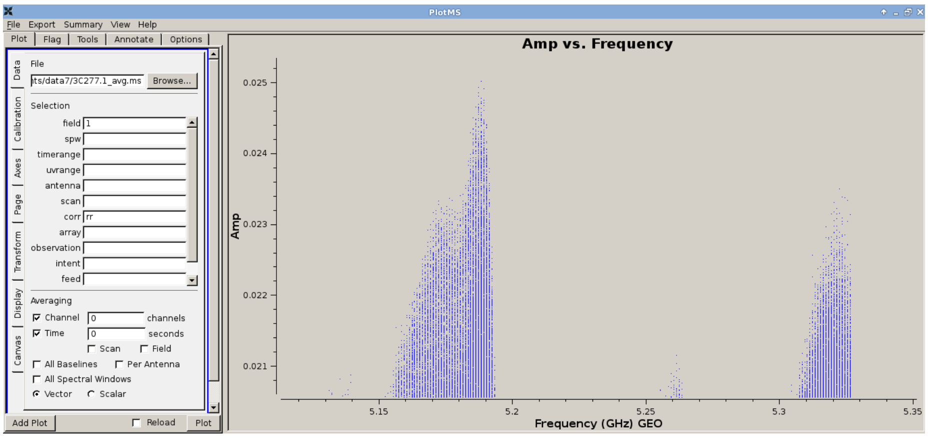


Figure 2.23: Amplitude (from 0.022 to 0.03 (Jy)) vs Frequency for RR polarisation plot (PLOT4).

For the LL polarisation plot the signals are between 5.182 GHz and 5.1912 GHz and between 5.31 GHz to 5.325 GHz (Figure 2.24). That Figure has a different axial scale for the x axis (frequency) than the one used for Figures 2.21-2.23. The reason to do that was to make noticeable the signal response to that specific frequency range.

Figures 2.22, 2.23 and 2.24 corroborate that for polarisations RR and LL the response in frequency is different. Although for both polarisations the undesired signals are present in two frequency ranges, the patterns of the signals are not the same. From here the analyst is able to flag that specific part of data for amplitudes between 0.022 Jy and 0.003 Jy and field of observation 1, considering frequency ranges between 5.18 GHz to 5.19 GHz and 5.309 GHz to 5.325 GHz for RR polarisation; and between 5.182 GHz to 5.1912 GHz and 5.31 GHz to 5.324 GHz for LL polarisation.

2.6.3 Efficiency analysis of the state of art manual flagging process

Flagging is an important process for removing noise from radio astronomy data. Automatic flagging is effective for removing some types of noise, but manual

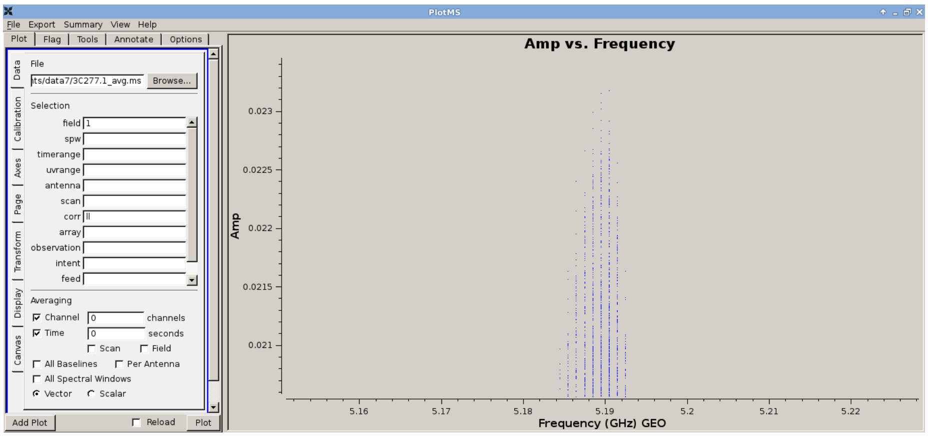


Figure 2.24: Amplitude (from 0.022 to 0.03 (Jy)) vs Frequency for LL polarisation plot (PLOT5).

flagging is also needed. However, manual flagging is very time-consuming because of the number of parameters that are involved and the number of visualisations that need to be checked. As mentioned, the actual process of manual flagging demands that the analyst performs a series of plots in order to understand the relation and origin of the undesired data. For each plot several clicks has to be performed and in some cases the definition of the variables by typing has to be added. Therefore, according to the amount of data and level of detail applied a proportional time is required. The information provided in Table 2.2 in terms of the options (clicks, number of clicks, typing and number of characters typed) and the required time of each action was obtained by performing the plotting of each of the plots in CASA.

As seen in the Table 2.2 each plot requires several actions to be done (clicks and characters typing). Before the plots are performed, a pre-configuration of the application is required. This step defines the general parameters according to the selected variables. According to the analysis each plot requires from 3 to 6 actions; in overall 30 actions were identified to produce the 5 plots. It can be inferred that for a more detailed analysis required actions will considerably increase. For the same example but analysing baselines and fields the amount of plots would be 15x4x5 [(number of baselines) x (number of polarisations) x

DESCRIPTION	OPTION	CLICK	NUMBER OF CLICKS	CLICKS CONSECUTIVE	TYPE	NUMBER OF CHARACTERS	TIME (seconds)
MAIN CONFIGURATION	data window (by default)						1
	file browse	X	1	1			1
	choose file	X	1	2			1
	field	X	1	3			1
	field				X	1	1
	averaging channel	X	1	4			1
	averaging time	X	1	5			11
PLOT 1	axes window	X	1	6			1
	x axis variables visualization	X	1	7			1
	x axis variable selection	X	2	8-9			4
	y axis variables visualization	X	1	10			1
	y axis variable selection	X	2	11-12			4
	plot	X	1	13			21
PLOT 2	amplitude manual range	X	1	14			1
	amplitude maximum value option	X	1	15			1
	amplitude maximum value selection				X	5	4
	amplitude minimum value option	X	1	16			1
	amplitude minimum value selection				X	4	4
	plot	X	1	17			19.5
PLOT 3	display window	X	1	18			1
	colorize option	X	1	19			1
	colorize option selection	X	2	20-21			4
	plot	X	1	22			15
PLOT 4	data window	X	1	23			1
	correlation option	X	1	24			1
	correlation selection				X	2	4
	plot	X	1				15.3
PLOT 5	correlation option	X	1	25			1
	correlation selection				X	2	4
	plot	X	1	26			12
		total	**27**	**26**	**5**	**14**	**138.8**

Table 2.2: Description of the actions required related to the number and time required of clicks and characters typing for the manual flagging example.

(number of fields)] obtaining 300 plots.

From this, the main parameters immerse in the manual flagging process were identified: spectral windows, baselines, fields and polarisations. In the Figure 2.25 is mentioned the number of possibilities for each variable that has to be taken into account in order to performed the manual flagging.

variable	options
spectral windows	4
baselines	15
fields	5
polarizations	4

Figure 2.25: Main variables immerse in the manual flagging.

It is worth to be mentioned that the 4 spectral windows corresponds to 502 frequencies and the 15 baselines are determinate in relation with the 6 antennas.

Finally, some aspects has been identified as source of the inefficiency of the state of art manual flagging process. The most relevant are: required time to perform one plot, excessive amount of possible parameters to be determinate to make a plot and lack of possibilities to visualise several plots at the same time.

2.7 Powerwall (large scale high resolution visualisation)

A Powerwall is a matrix of high resolution displays intended to be used for large-scale data visualisation. This system can be used to display a vast amount of information simultaneously, as a tool for data analysis, for the visualisation of different levels of detail and also as a collaborative work space (Rooney C. and Ruddle R. A., 2015).

Powerwalls have been used in many applications and domains such as security, medicine, earth sciences and physics. In security there have been studies of a dynamic visualisation of Bitcoins transaction patterns (McGinn D., Birch D., Akroyd D., Molina-Solana M., Guo Y. and Knottenbelt J., 2016), observation of intelligence analysts (Andrews C., Endert A. and North C., 2010), and in the support of cyber analytic processes by historic visualisation (Singh A., Endert A., Andrews C., Bradel L., Kincaid R. and North C., 2011). In medicine applications include clinical conferencing settings from the prototyping of a Radiology scenario (Olsen B. I., Dhakal S. B., Eldevik O. P., Hasvold P. and Hartvigsen G., 2008) and in bioinformatics for the display of genome visualisation (Aurisano J., Reda K., Johnson A., Marai E. G. and Leigh J., 2015). In earth sciences Powerwall have been used for the multi-variate analysis of seismic and satellite-based observational data (Yuan X., He X., Guo H., Guo P., Kendall W., Huang J. and Zhang Y., 2010) and in a research project to integrate geographic education in order to evaluate the potential of a GeoWall (Slocum T. A., Dunbar M. D. and Egbert S. L., 2007). In physics applications include the visualisation of large scale atomistic simulations in Ultra-resolutions (Reda K., Knoll A., Nomura K., Papka M. E., Johnson A. E. and Leigh J., 2013).

There is an interesting case in the cultural area. They proposed a new approach called Cultural Analytics. Taking visual analytics techniques as reference, a research group in the University of San Diego (Yamaoka S., Manovich L., Douglass J. and Kuester F., 2011) proposed a an approach for the challenges of how to access and visualise large contents of cultural media content. In order to do that they proposed techniques to be used in large displays systems. Among those techniques are image collage, image plot, multiple variable and multiple datasets, and visual extension to scatterplots. This kind of works corroborates the importance given in research groups for multiple applications taken the potential of large scale visualisation as reference.

2.7.1 Powerwall displays configurations

The first Powerwall was installed at the University of Minnesota in 1994 by using 4 projectors obtaining a resolution of 7.8 Mega pixel (Mpx). In 2006 to drive a 60 Mpx. Powerwall a 7 machine cluster was required. By 2012 to achieve the same resolution a Powerwall could be controlled by a single computer with three graphic cards and by 2015 the same Powerwall only required one graphic card (Rooney C. and Ruddle R. A., 2015). One example of this is the University of Leeds Powerwall. This high resolution wall-sized display has six 8 Mpx high resolution screens driven by a single PC. Actually, The Reality Deck of 1.5 Giga pixel (Gpx) is the largest high resolution visualisation system (Papadopoulos C., Petkov K., Kaufman A. E. and Mueller K., 2015).

The use of cluster driven Powerwalls allows to implement bigger visualisation systems in terms of overall resolution. The Reality Deck which is driven by a 18 PC cluster provide a 360 degrees horizontal field of view with a resolution of 1.5 Gpx (Papadopoulos C., Petkov K., Kaufman A. E. and Mueller K., 2015). From another perspective, clusters can be used as spreadsheet-based visualisation system. This means that each screen provides specif data within a common group of information. An example of this is The hyperwall (Sandstrom T. A., Henze C. and Levit C., 2003) which is a 18" monitors array. This cluster has been applications like molecular quantum mechanics, computational aerodynamics, weather modeling, planetary geology and remote sensing.

If considering the work of (Yamaoka S., Doerr K. and Kuester F., 2011) another approach is to classify the configurations is in terms of the management and distribution of data. From this the visualisation is generated and then distributed to visualisation environments or handle the visualisation by the cluster that manages the high resolution displays. Some examples of those techniques are The Scalabale Adaptive Graphics Environment (SAGE) (Jeong B. , Renambot L., Jagodic R., Singh R., Aguilera J., Johnson A., Leigh J., 2006), Magic Carpet, Juxta View (Krishnaprasad N.K., Vishwanath V., Venkataraman S., Rao A.G., Renambot L., Leigh J., Johnson A.E., 2004) and GigaStack (Ponto K., Doerr K., Kuester F., 2010).

The criteria to select a Powerwall configuration responds to several factors. The application determines the amount of information that is required to obtain benefits in terms of the deployment of data. Two main approaches can be identified. If the intention is just to deploy images without interaction of the user the configuration should responds to the type of graphics and therefore establish the resolution requirements and overall size of the visualisation system. One example of this is the Reality Deck which provides a 360 degrees visualisation experience. On the other hand, if the intention is to have interoperability the criteria are different. In this point is crucial to consider cognitive loads to limit the amount of information intended to be shown simultaneously. This is to avoid an overload or saturation of mental processes at any given moment. Also, the interaction tools as mouses, sensor's gloves or touch screens have a direct relation to the configuration characteristics.

2.7.2 Visualisation on Powerwalls

Powerwalls show orders of magnitude more display real estate (i.e., pixels) and this produces a corresponding increase in the amount of data, abstractions, and stages of analysis that users may display at any moment in time (Ruddle R. A., Thomas R. G., Randell R. S., Quirke P. and Treanor D., 2013).

Over a decade, there has been a substantial amount of research into the use of Powerwalls for visualisation in a number of application domains. Also some positive results have been obtained in terms of the optimisation of certain processes. Some example of benefits of the use of a Powerwall are: users found targets significantly faster when a display showed the full dataset, instead of users having to pan to see the full dataset (Ruddle R. A., Thomas R. G., Randell R. S., Quirke P. and Treanor D., 2013), users noticed unexpected patterns when using a Powerwall (Fateen W., Ruddle R. A., Treanor D., Sondergeld P. and Ouirke P., October, 2013), and Powerwalls can be used as a large scale collaboration environment (Westing B., Urick B., Esteva M., Rojas F. and Xu W., 2011).

There have been many studies related to large-scale information visualisation (Andrews C., Endbert A., Yost B. and North C., 2011), (Ni T., Schmidt G. S., Staadt O. G., Livingston M. A., Ball R. and May R., 2006), (Knudsen S., Jakobsen M. R., Hornaek K., 2012) and (Beaudouin-Lafon M., Huot S., Nancel M., Mackay W., Pietriga E., Primet R., Wagner J., Chapuis O., Pillias C., Eagan J., Gjerlufsen T. and Klokmose C., 2012). Those studies provide guides and common terminology in order to define the attributes that have to be considered at the moment of design, implementation and use of a visualisation system like a Powerwall. From a technological standpoint some key attributes are: size, pixel density, resolution, brightness, contrast, viewing angle, bezels, display technology and form factor. Size corresponds to the diagonal viewing area and is expressed in inches. Pixel density corresponds to the amount of pixels per inch and is expressed in DPI (Dots per inch). Resolution defines the overall number of pixels by multiplying horizontal and vertical pixel quantities. Brightness establishes the amount of light emitted by the display and is expressed in candelas per square meter. Contrast measures the luminance ratio between the brightest and darkest color. Viewing angle is the angle from which a display can be viewed with acceptable performance. Bezels corresponds to the frames that surround the monitors. Display technology refers to the type of technology used to display the visualization resources, being rear or front projected and generally using tiled LCD monitors or projectors. Form factor specifies the physical arrangement and shape of the display.

In terms of the graphics encodings they can be divided into scalable and not scalable as proposed in the visual scalability work by (Eick S. G. and Karr A. F., 2002). Some graphical encoding challenges that need to be considered are spatial position, glyph size, colour encoding and orientation. In the visualisation design is critical to consider visual acuity to provide a coherent solution. With respect to the user interaction, there are several approaches that can be used as reference. Some of them are navigation techniques, brushing and linking, selecting and marking, control panels (movable controls, pop-up controls, hand-held controls and gestures) and spatial interaction. Other studies consider a user evaluation approach. For example the User Evaluation of Polymetric Views (Anslow C., Marshall S., Noble J., Tempero E. and Biddle R., 2010) describes an experiment to visualise software metrics data by using a System Hotspots View technique.

Although is not possible to design a generic solution that works well every single application, it is reasonable to establish similarities that allows some kind of initial protocol that guides the definition of the design of the visualisations for a Powerwall. Where interoperability is required, a starting point is to define the process or processes that are intended to be addressed. This establishes the type and amount of data that is required. As will be seen in the following chapters, the pre-generation of graphics contributes to the optimisation of time. Also, by having more size capabilities compared to the traditional visualisation devices as desk monitors, it is possible to have a multidimensional approach towards to what can be simultaneously presented.

2.7.3 Visualization principles as design methodology

Following is the description of the project taking as reference the methodology proposed by (Sedlmar M., Meyer M., Munzner T., 2012) in terms of the visualization principles that are involved in the design of a problem-driven visualization research. Also key concepts in visualisation principles are mentioned. This project can be categorized in three main components: precondition, core and analysis (as seen in Figure 2.26. Precondition is divided in two stages:learn and discover. Core is divided in four stages: discover, design, implement and deploy. Analysis is divided in two stages: reflect and write.

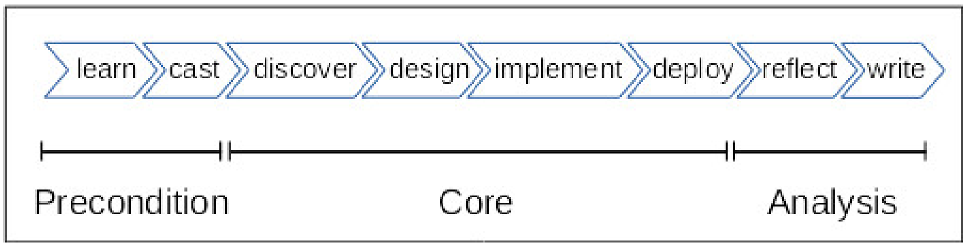


Figure 2.26: Components and stages of the project.

PRECONDITION: The learning stage was based in reading literature in visualization techniques and methodologies and also in the computing procedures for the data treatment of radioastronomical data. The casting stage involves the participation of two professional analysts. The roles of those analysts were different. One of them provided information related to the parameters involved in the manual flagging procedure and the other analyst performed the evaluation of the solution.

CORE: By having identified the valuable parameters involved in the manual flagging it was possible to define the problem. Then the designing process started taking as reference the information obtained in the previous stages. This information provided the criteria necessary for the visualization structures that defined the solution. Initially the powerwall was used as separate screens but after some tests it was decided to use it as a complete visualization system. After the design was ready the implementation stage begun. In this stage the software architecture was built in order to provide a scalable solution in terms of the algorithms structure. Finally the deployment of the solution was achieved and was ready to be tested.

ANALYSIS: Once the solution was deployed, an evaluation was performed by an analyst. This stage provides preliminary results in terms of the effectiveness of the solution. Finally, having all the information, conclusions were registered. The writing process starts from the beginning of the project, nevertheless at the end of the mentioned stages it was possible to integrate all the information in a coherent way.

2.7.3.1 Key concepts of visualization principles

There are many theories in terms of visualisation principles. For the purpose of our work it will be mentioned the ones that are more related such as balance, movement, pattern, proportion, rhythm, variety and unity. Balance includes aspects such as colour, shape and texture. In our case it was used the default parameters provided by CASA (Common Astronomical Software Application) in terms of the colours and plot types. A symmetrical type of balance was used. Movement provides the users attention in a certain direction. In our project this point was achieved by the allowance of having vertical and horizontal scans of all the baselines. A pattern is created by showing repeated objects, in our case it was possible to establish data combinations that generates systematic arrangements of the plots in order to provide a novel radioastronomical data analysis. Proportion is related to the sizes of the plots; they were calculated according to the maximum sizes of the Leeds Powerwall in terms of resolution dimensions. The design of the canvases achieved a coherent rhythm in terms of the flow of how the information was arranged. This allows to navigate through the data in a systematic way. In terms of variety, the intention of having data merged in different configurations responds to the principle of having several possibilities of visualisation of the information. This is a crucial aspect due to the personal method used by each analyst to perform the manual flagging process. Canvases can be perceived as a unit in terms of an homogeneous solution.

2.8 Using a Powerwall to make manual flagging more efficient

The manual flagging procedure (Section 2.5.2) has similarities with the trial and error approach that users typically take to optimise parameters for data processing. However, if outputs are pre-computed and displayed together then users perform a more thorough analysis of a parameters space, which significantly increases the quality of an image processing pipeline (Pretorius A. J., Bray M. A., Carpenter A. E. and Ruddle R. A., 2011).

The approaches to optimise the manual flagging process by the use of a Powerwall are: low-level plotting, simultaneously multi-plotting, higher resolution images, different levels of detail visualisation and dedicated flagging application. Following there is a brief description of the approaches.

The low-level plotting intends to directly show low-level plots (Figures 2.23 and 2.24) instead of requiring users to first create higher-level plots (Figures 2.20 and 2.21). Simultaneously multi-plotting shows multiple plots at once, allowing a general perspective of the data of specific variables. Higher resolution images allow a more detailed inspection avoiding the need for subsequent plots (as seen in Figure 2.22). A different L.O.D. visualisation permits users to view sets of plots at any level of detail (e.g., a plot like Figure 2.20 for each baseline) to ascertain which parts of the variable space do (not) need to be analysed in detail. A dedicated flagging application minimise the required time to make decisions in order to produce a specific plot.

2.9 Conclusions

This chapter has introduced the processes and general architecture of the data treatment for the eMerlin observations. An example of a manual flagging procedure was established to distinguish the difficulties involve in the process.

The main parameters and options that has to be consider in the identification of undesired data were established in order to understand the complexity of the manual flagging process. This complexity is assumed due to the multiple parameters involve and also in the particularities that every analyst has to perform the manual flagging. From that understanding raises the challenge to design and implement a solution that is highly adaptable and easily configurable.

The main problem immerse in the process is time consuming; in order to generate one plot the analyst needs to specify many parameters and also due to the number of variables immerse in the manual flagging the quantity of required plots is enormous. Considering those aspects, the pre-generation of plots is a feasible solution to optimise this process.

Chapter 3

Design of the solution

3.1 Introduction

The following design is developed with the purpose of creating a visualisation method to optimise in terms of time and effectiveness the manual flagging process of radio astronomical data. A novel visualisation approach is proposed taking as reference the observation of the existing techniques used in the manual flagging. Actually the analyst who performs the manual flagging produces one plot at a time in order to explore undesired data. From this initial plot the analyst gradually generates more detailed plots in order to study a specific segment of parameters. This process is very intuitive. That means that time becomes significant due to the amount of data involved in the astronomical observations and depends directly on the expertise of the analyst. As mentioned before, the cognitive load that the analyst require demands a great cost in terms of the memory and mental work. The capacity of memorise patterns observed in a previous plot determines the effectiveness of the process itself. The actual process relies in the memory capacity to make comparisons. Another important aspect is the number of interactions required to make a single plot. In Chapter 2 a description of the actions required to generate a plot were established. The amount of required time is significant when considering the number of variables immerse in the process and the size of data used.

This design considers the visualisation of combinations of plots. There are advantages and disadvantages of having multiple plots simultaneously. On one side this type of visualisation in terms of different LODs allows the user to have access to detailed data from the beginning. This means that a considerable amount of time is saved if both general and detailed data is shown simultaneously. If for instance the analyst identifies a certain region of data that he wants to study he can immediately have access to it by just moving his eyes. Also the pre-generation of plots reduces the numbers of steps or actions that the analyst require to obtain the information he requires in order to make the evaluation of the information. Another benefit is to have the information arranged according to similar paramters allowing to generate comparisons of hole dimensions of data. Also the possibility of having different configurations of the visualisation of the information allows the analyst to rule in and out specific segments of data. From here the identification of what data is valuable and what is worthless can be performed in a more efficient way in terms of time and effectiveness.

Although there are a number of positive aspects, the visualisation of simultaneous information demands more effort from the analyst and increases the cognitive load. In order to minimise that impact and/or reduce that load in the user the canvas configurations and the structures for the organization of the information seeks to provide clear patterns in the information. Those patterns are intended to provide an organised perspective in which the brain interprets the information in a coherent way. This coherence is related to the clear segmentation of parameters and in how the merges of the matrices of plots are arranged.

3.2 Conceptual design

The solution is based on the pre-generation of plots arranged in specific configurations. Arrangements are termed as canvas and have a nomenclature from P1 to P12. Each canvas responds to the Leeds Powerwall technical characteristics in terms of resolution capabilities (11520 x 4320 pixel (px)). Regarding the structure of the canvas, the selected configurations seek to allows an easy understanding of the deployed information. Due to amount of information there is no a single solution that can cover all the parameters involved in the process. That is why several configurations were designed to work as a unified method or solution.

From what was found in related works, studies has been conducted oriented to cognitive processes, specific applications and interaction. From them, strategies can be used in order to guide the designing process. For example in order to compare the accertivness of different designs attribute-centric and space-centric strategies were used by (Yost B., Haciahmetoglu Y. and North C., 2007). Another perspective is to study the input configuration process considering collaborative work as done by (Birnholtz J. P., Grossman T., Mak C. and Balakrishnan R., 2007). The work by (Bi X., Bae S. H. and Balakrishnan R., 2010) shows that bezels in the Powerwalls have to be considered as a variable in the definition of the design. The conclusion of this study is that the interior bezels of the Powerwall are detrimental for search accuracy but not so to target selection. So every design must take into account the specific technical configuration and the user requirements and objectives.

The criteria to select the parameters and configurations were based on the recommendations of a professional analyst who performs the manual flagging process for the eMerlin interferometric array. Taking those recommendations as a starting point there were produced several possibilities that finally converge in the 12 canvas configurations. Another important aspect is that this solution was designed from the beginning to be scalable. The canvases are based on matrices that can be merged in any way that a radio astronomy system requires. eMerlin uses 6 antennas. There are systems that use up to 32 antennas. This design responds to this principle of scalability in the means of having a base or initial structure for other systems with different characteristics. This scalability also could be embraced in two ways. On one hand this solution provides different configurations of parameters that can be used in any manual flagging process for any .ms files users of radio astronomy data. On the other hand by having this canvas-based structure of several configurations is also possible to easily adapt the matrices sizes, merging and parameters according to the interferometric system characteristics, analyst preferences, Powerwall specifications, or available visualisation tool. In order to easily modify those configurations, python scripts were produced to generate tailor-made configurations.

Another aspect is adpatability in terms of the different procedures that each analyst has to perform the manual flagging. By considering and comparing the interaction efficiency of different users, as concluded by (Liu C., Chapuis O., Beaudouin-Lafon M. and Lecolinet E., 2016), a better understanding of the processes immerse and a more coherent design in terms of develop a pertinent solution will be achieved.

It is worth saying that the design presented is the result of several attempts. The process to obtain this result was by making the design of many configurations and by performing visualisation tests on the Powerwall. The visualisation experience can only be really measured when this kind of canvas were putting together in the screens. Although the desk design is vital to produce a coherent work, the visualisation testing is an essential part of the designing process. Those tests resulted in feedback that improves the solution.

3.3 Visualization design

The processes involved in this project are explained taking as reference the Nested Model for Visualization Design by (Munzner T., 2009). This model indicates the flow of the project and explains it in a systematic way. Figure 3.1 shows the Model and its integration to the manual flagging powerwall-based solution.

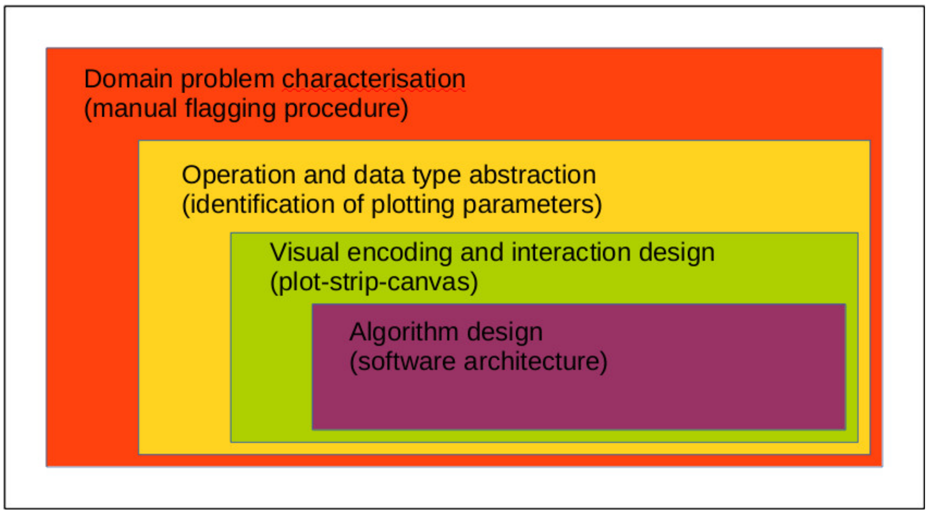


Figure 3.1: Nested model for visualization design for the manual flagging power-based solution.

Each step is explained as follows.

3.3.1 Domain problem characterization (manual flagging procedure)

In order to define the design process the first step was to determine how the manual flagging was performed. To that end, the actions and times of each of those actions were identified. As seen in Figure 3.2 the identification of a flagging section demands several actions to obtain the required plots that support such identification. From this it is noticeable that the manual flagging process implies different configurations to obtain the different plots.

3.3.2 Operation and data type abstraction (identification of plotting parameters)

Once the process was understood the next step was to identify the specific parameters that were required to obtain the plots that the analyst needs to perform the analysis and identification of flagging regions. In order to do that and taking as reference the concepts of a professional analyst the parameters were defined as seen in Figure 3.3.

3.3.3 Visual encoding and interaction design (plot-strip-canvas)

Once the parameters of the plots were identified the next step was to create a coherent visualization structure that would provide an understandable environment to display not only one plot but a series of them arranged by baselines, in order to contrast in each baseline different specific parameters. In this point the concept of strip of plots and canvas was conceived. A strip takes as basis its 15 single plots to show all the baselines one next to the other. From there it is possible to compare in one strip all the baselines for a specific configuration of parameters. In order to obtain the necessary number of strips to display all the parameters, 112 strips were required. Once the 112 strips were produced the next step was to merge the strips in coherent structures. These structures are called

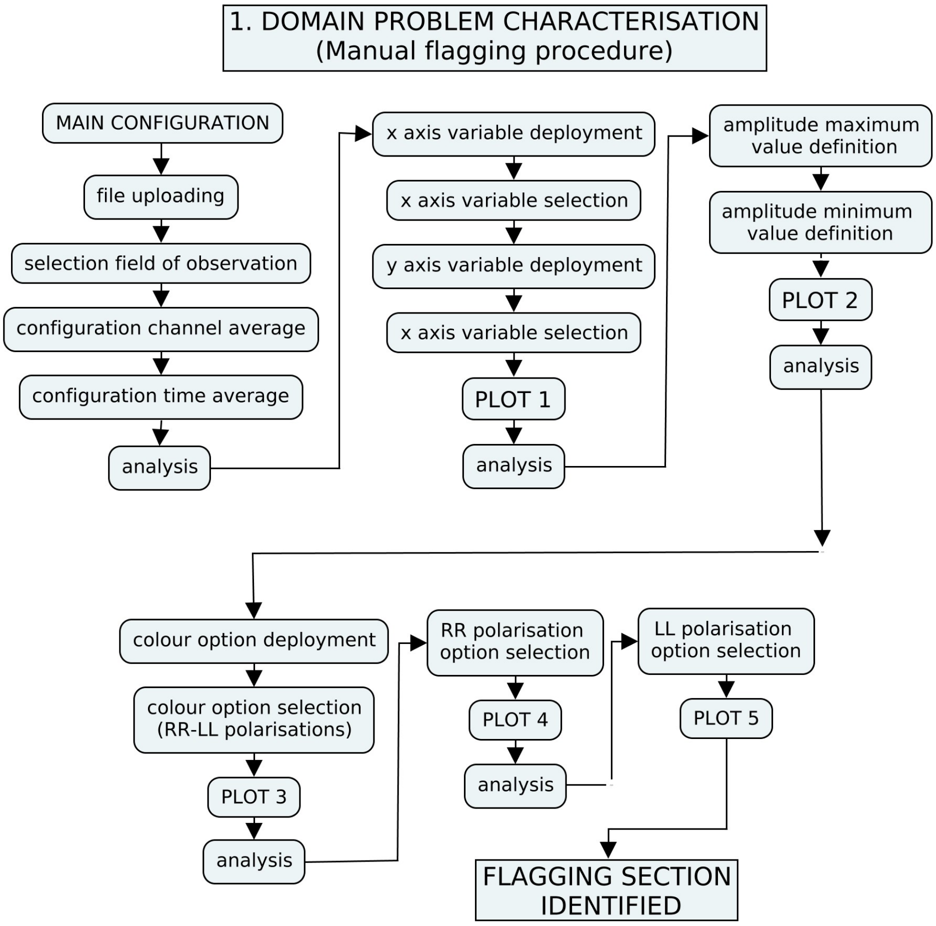


Figure 3.2: Manual flagging procedure

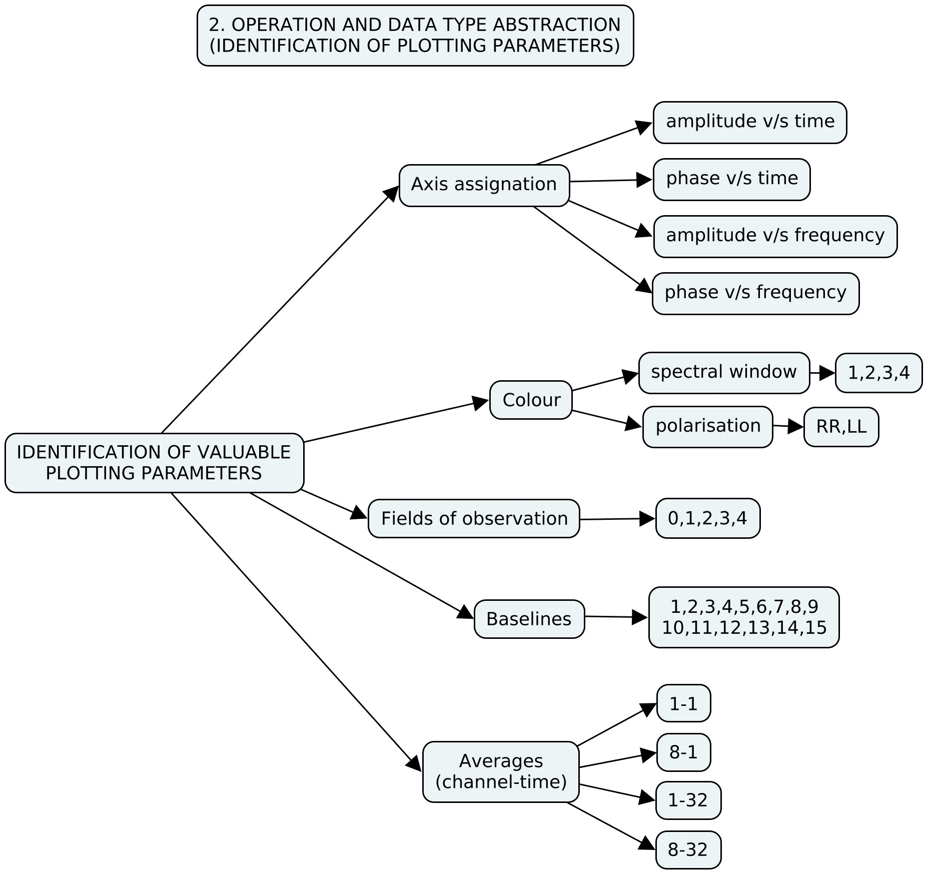


Figure 3.3: Identification of plotting parameters

canvases. In order to group all the strips, 12 canvases were produced. Once the 12 canvases were produced the final step was to integrate them in a single file using a visualization software. This software allows the user to control and have access to every canvas by just clicking the desired canvas.Figure 3.4 shows the process to generate the canvases taking as reference the strips of plots.

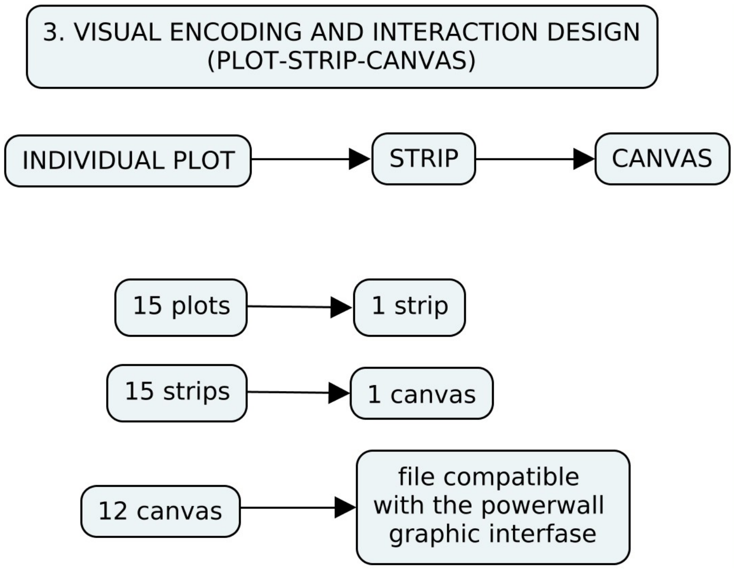


Figure 3.4: Plot, strip and canvas structures

3.3.4 Algorithm design (software architecture)

The software structure to produce and merge the strips and then integrate them in a user interface demanded the implementation of 2 algorithms and the use of a visualization software. The first algorithm generates the 112 strips according to the parameters identified. The second algorithm was designed to merge the strips according to the structure of each canvas. Figure 3.5 shows the structure of the software architecture used for the solution.

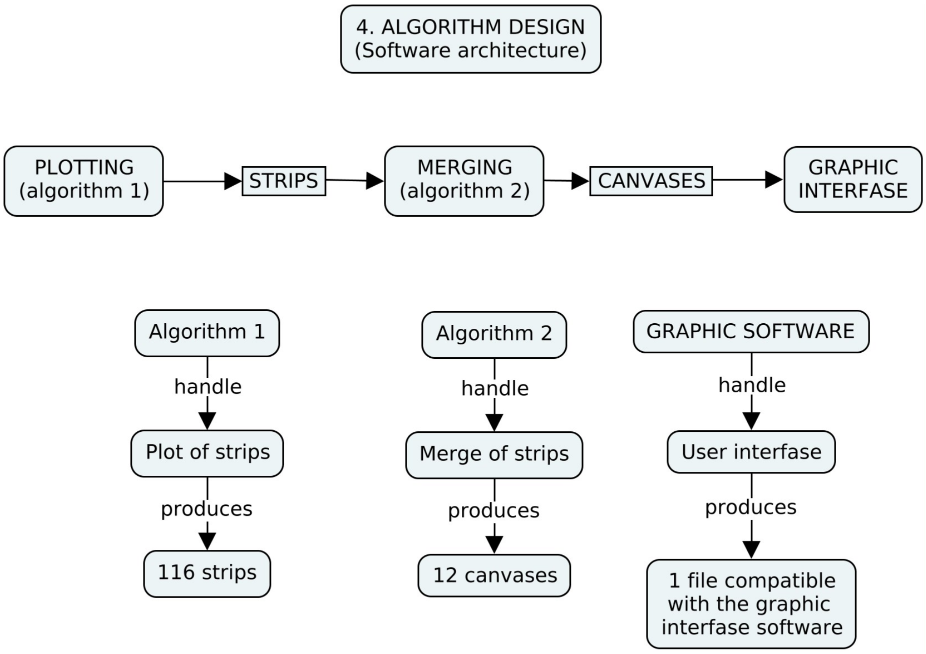


Figure 3.5: Software architecture

3.4 Design structure

The proposed design responds to several constraints identified in the initial discussions with the analyst. First of all the analyst relies on certain combinations in terms of axis assignation to build the plots. Those combinations were the basis to establish the design in terms of axis assignation for each plot and therefore for each strip configuration. Also, it was clear that in order to perform the comparisons between different plots the possibility of having different baselines was a crucial aspect to contrast the behavior of the signals in order to identify undesired patterns. As a result, it was decided that to show baselines side-by-side was a key factor in the designing process. Therefore, the design responds to the necessity of placing those baselines in a coherent way. That coherence was confronted by arranging the plots according to different axis assignations, averages and fields of observation maintaining the baselines as a visualization pattern. Once those decisions where taken in terms of defining the required combinations that allows the analyst to have all the valuable information 12 canvases were defined as a way to group the plots in an understandable way to visualize the different plots. The canvas responds as the design choice because it allows the information to be arranged by segmenting the plots according to the axis assignations, averages and fields of observation, taking the baselines one after another as reference for the visualization configurations.

In order to explain the design structure three terms will be used. Those terms are plot, strip of plots and canvas. The plot is defined by the variables are in each graphic. The strip defines the characteristics according to the parameters of the groups of plots and the orientation in which the individual plots are arranged. The canvas is the final image composition that integrates the plots and the strips in a specific configuration. The explanation of the design structure is supported in the Table 3.1 which integrates and defines the specifications of each plot, strip and canvas.

3.4.1 Plot

A plot is considered as the individual image that shows the signal response according to certain parameters and attributes. A plot considers two parameters: axis configurations and colour assignation.

3.4.1.1 Axis configuration

Considering the axis configurations there are 4 four combinations of parameters. Those combinations are:

- Amplitude vs Time.
- Phase vs Time.
- Amplitude vs Frequency.
- Phase vs Frequency.

Axis configurations define the variables to be used in every plot in terms of the x and y axis. Each of these configurations are used as source of information to observe and analyse different signal patterns. The possibility of show different axis configurations of the same baseline one near to each other allows to compare and find anomalies in the signal's response. Then these anomalies can be identified and be flagged.

3.4.1.2 Colour throughout

The colour assignation defines which parameter will be shown in different colours. By doing this we obtain another dimension that shows information of a determinate parameter. Two colour assignations were selected. The first one is spectral window and the second polarisation. For the plots in which time is one of the axis parameters the spectral windows will be colourised, and when a plot considers frequency as one of their axes parameters polarisation will be colourised.

3.4.2 Strip of plots

The individual plots will be automatically produced from CASA by using a python script "plot.py". The script will produce a total of 112 vertical and horizontal strips using 15 individual plots. The orientation depends on how the baselines are shown. In Figure 3.6 is shown how the baselines are arranged. For example if row one is considered (bottom figure) then different plots parameters will be shown horizontally only for baseline 1. Now if column 1 (top figure) is considered the different parameters will be shown vertically only for the baseline 1. The purpose of having two different orientations is to increase the analysis options that the analyst has with the same type of plots.

There are additional attributes that define each strip. Those attributes are time and channel averages, fields of observation and polarisation and spectral windows. Each attribute is explain as follows.

3.4.2.1 Time and channel averages

Another parameter is the average of the channels and the time. Average in channels is defined by an n factor, which specifies the number of channels that will be averaged together to obtain a total number of channels decreased by a factor of n. The value is expressed in the number of channels to be averaged. Time average corresponds to the amount of time that the data will be averaged

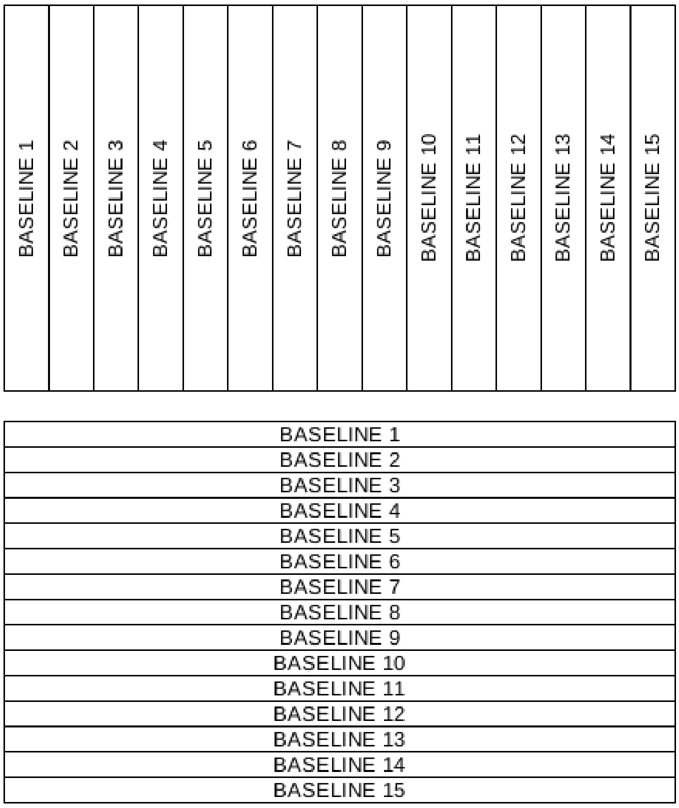


Figure 3.6: Strip of plots orientations: horizontal (top), vertical(bottom).

in terms of seconds. The quantity of seconds will define the averaging of the signals for a given interval.

The main purpose of using different averages is to obtain different levels of detail in each configuration. There are four possibilities for the definition of the averages 1-1, 8-1, 1-32 and 8-32 being the first term the channel average and the second the time average. According to the value of the averages, the software produces a sum and media of determinate number of channels or frames of time. The greater is the value assigned to the average, the more detailed the plot is produced.

3.4.2.2 Fields of observation

As seen in Table 3.1, two approaches were defined with respect to the fields of observation. First for the strips 1 to 16 (as referenced in column 1) and strips 57 to 72 (as referenced in column 3) all the fields were integrated in order to produce a general perspective of the signal response for each baseline. Then from the strips 17 to the 56 (as referenced in column 1) and strips 73 to 112 (as referenced in column 3) the fields of observation are shown individually. The purpose of showing each field of observation separately is to obtain a discriminated visualisation of the fields with respect to each baseline.

3.4.2.3 Polarisations and spectral windows

Is worth mentioning that all the plots consider all of the spectral windows and the polarisations RR and LL. The intention of having all the spectral windows is to provide the perspective of all the range of frequencies that the radio telescopes can provide. On the other hand the decision to select RR and LL polarisations where related to the polarisations that are mostly used in the flagging process.

Table 3.1 presents the information for every strip of plot. The first four columns corresponds to the consecutive strip numbers and the reference assigned to each strip. Columns five and six are the parameters assigned for each axis of the plots. Column seven provides information related to the colour assignation for each strip. Columns seven and eight defines the channel and time averages. Finally, in the ninth column is defined how the fields of observation are shown.

As an example, plot configuration number 1 will produce 15 plots horizontally oriented with the following characteristics: amplitude vs time axis configuration, spectral window as colour assignation, channel average 1 and time average 1, and visualisation of all the fields. As it was mentioned before each of these strips configurations were produced both vertically and horizontally, so the transpose strip for strip number 1 is strip number 57.

3.4.3 Canvas

A canvas is created by merging in a particular configuration strips of plots. The 12 canvas are obtained by merging in different orientations 112 strips as seen in Table 3.2 This allows the analyst to compare different sets of parameters. Each of the configurations may be arranged vertically or horizontally showing a different perspective of the same information. This means that P2, P4, P6, P8, P10, and P12 are the transpose of P1, P3, P5, P7, P9, and P11. All the canvas configurations considers all the baselines and all the averages. Regarding the fields of observation, the difference is that in canvas P1, P2, P3,and P4 the baselines are not shown separately but they are in P5, P6, P7, P8, P9, P10, P11, and P12.

Table 3.2 defines the general characteristics of each canvas. Those characteristics are the strip orientation, axis configuration and colour assignation. The reason why there are the same configurations for some canvas references is that although the strips parameters are the same the way that they are merged are different. Following there is a detailed description of each canvas configuration. First the purpose of the design of the canvas will be mentioned. Then a table with the specifications of each canvas will be provided. The table shows the parameters of the plots and strips (shown in yellow), and the orientation of the strips in the canvas (blue). This blue section represents what is visualised in the Powerwall. In terms of resolution the green area corresponds to 11520 x 4820 pixels. The parameters related to the axis configuration give the details of the x and y axis for the plots. The averages (Avg) are expressed with two numbers. The first one defines the average in channel and the second number defines the average in time.

strip #	1x15 strip reference (horizontal)	strip #	15x1 strip reference (vertical)	X axis	y axis	coloraxis	average channel	average time	field
1	101	57	201	time	amplitude	spectral window	1	1	all
2	102	58	202	time	amplitude	spectral window	8	1	all
3	103	59	203	time	amplitude	spectral window	1	32	all
4	104	60	204	time	amplitude	spectral window	8	32	all
5	105	61	205	time	phase	spectral window	1	1	all
6	106	62	206	time	phase	spectral window	8	1	all
7	107	63	207	time	phase	spectral window	1	32	all
8	108	64	208	time	phase	spectral window	8	32	all
9	109	65	209	frequency	amplitude	polarization	1	1	all
10	110	66	210	frequency	amplitude	polarization	8	1	all
11	111	67	211	frequency	amplitude	polarization	1	32	all
12	112	68	212	frequency	amplitude	polarization	8	32	all
13	113	69	213	frequency	phase	polarization	1	1	all
14	114	70	214	frequency	phase	polarization	8	1	all
15	115	71	215	frequency	phase	polarization	1	32	all
16	116	72	216	frequency	phase	polarization	8	32	all
17	301	73	401	frequency	amplitude	polarization	1	1	0
18	302	74	402	frequency	amplitude	polarization	8	1	0
19	303	75	403	frequency	amplitude	polarization	1	32	0
20	304	76	404	frequency	amplitude	polarization	8	32	0
21	305	77	405	frequency	phase	polarization	1	1	0
22	306	78	406	frequency	phase	polarization	8	1	0
23	307	79	407	frequency	phase	polarization	1	32	0
24	308	80	408	frequency	phase	polarization	8	32	0
25	309	81	409	frequency	amplitude	polarization	1	1	1
26	310	82	410	frequency	amplitude	polarization	8	1	1
27	311	83	411	frequency	amplitude	polarization	1	32	1
28	312	84	412	frequency	amplitude	polarization	8	32	1
29	313	85	413	frequency	phase	polarization	1	1	1
30	314	86	414	frequency	phase	polarization	8	1	1
31	315	87	415	frequency	phase	polarization	1	32	1
32	316	88	416	frequency	phase	polarization	8	32	1
33	317	89	417	frequency	amplitude	polarization	1	1	2
34	318	90	418	frequency	amplitude	polarization	8	1	2
35	319	91	419	frequency	amplitude	polarization	1	32	2
36	320	92	420	frequency	amplitude	polarization	8	32	2
37	321	93	421	frequency	phase	polarization	1	1	2
38	322	94	422	frequency	phase	polarization	8	1	2
39	323	95	423	frequency	phase	polarization	1	32	2
40	324	96	424	frequency	phase	polarization	8	32	2
41	325	97	425	frequency	amplitude	polarization	1	1	3
42	326	98	426	frequency	amplitude	polarization	8	1	3
43	327	99	427	frequency	amplitude	polarization	1	32	3
44	328	100	428	frequency	amplitude	polarization	8	32	3
45	329	101	429	frequency	phase	polarization	1	1	3
46	330	102	430	frequency	phase	polarization	8	1	3
47	331	103	431	frequency	phase	polarization	1	32	3
48	332	104	432	frequency	phase	polarization	8	32	3
49	333	105	433	frequency	amplitude	polarization	1	1	4
50	334	106	434	frequency	amplitude	polarization	8	1	4
51	335	107	435	frequency	amplitude	polarization	1	32	4
52	336	108	436	frequency	amplitude	polarization	8	32	4
53	337	109	437	frequency	phase	polarization	1	1	4
54	338	110	438	frequency	phase	polarization	8	1	4
55	339	111	439	frequency	phase	polarization	1	32	4
56	340	112	440	frequency	phase	polarization	8	32	4

Table 3.1: List of parameters for each of the 112 strip of plots to produce P1 to P12 canvases(strip number for horizontal and vertical configurations, x and y axis parameters, colour assignation, time and channel averages, and fields of observation)

canvas reference	strip orientation	axis configurations				color assignation
		amplitude vs time	phase vs time	amplitude vs frequecy	phase vs frequency	
1	horizontal	x	x	x	x	spw-polarization
2	vertical	x	x	x	x	spw-polarization
3	horizontal	x	x	x	x	spw-polarization
4	vertical	x	x	x	x	spw-polarization
5	horizontal			x		polarization
6	vertical			x		polarization
7	horizontal				x	polarization
8	vertical				x	polarization
9	horizontal			x		polarization
10	vertical			x		polarization
11	horizontal				x	polarization
12	vertical				x	polarization

Table 3.2: Description of the configuration for each canvas (strip orientation, axis parameters and colour assignation)

3.4.3.1 P1 canvas

This configuration allows the analyst to make vertical comparisons of the signals response in time in the first eight vertical strips and in frequency from vertical strip 9 to 16. Also by having each average one next to another is possible to achieve different levels of detail of each axis configurations in order to compare separately each baseline signal response. By having all the averages one next to another is possible to have a detailed perspective of each parameter combination specially if the behaviour of each individual baseline wants to be analysed (see Table 3.3).

3.4.3.2 P2 canvas

This configuration permits the user to have a detailed perspective of each of the parameter combination by having the averages of each of those parameters one next to the other. Also by having the baselines located horizontally is possible to observe and analyse all baselines and establish comparisons with respect to amplitude and phase. In this scenario, while the amplitude analysis provides evidence of undesired peaks in the signal, the analysis of the phase patterns becomes easily recognisable due to the visualisation arrange of information. Another approach is to perform an inspection of each baseline vertically. This approach allows a user to observe simultaneously the signal response with respect to time

P1		AMPLITUDE vs TIME				PHASE Vs TIME				AMPLITUDE vs FREQUENCY				PHASE vs FREQUENCY			
		Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32
BASELINES	1	Strip 101	Strip 102	Strip 103	Strip 104	Strip 105	Strip 106	Strip 107	Strip 108	Strip 109	Strip 110	Strip 111	Strip 112	Strip 113	Strip 114	Strip 115	Strip 116
	2																
	3																
	4																
	5																
	6																
	7																
	8																
	9																
	10																
	11																
	12																
	13																
	14																
	15																

Table 3.3: P1 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

and frequency. By doing this the user is able to find discrepancies in the signals of different averages for the same periods of time and same frequencies for each baseline (see Table 3.4).

3.4.3.3 P3 canvas

The P3 canvas enable the analyst to compare vertically time and frequency per baseline. This configuration permits to analyse all the parameter for specific averages for all the baselines. By having the parameters aligned horizontally it is possible to have a general approach of the signal response of all the baselines for specific averages. This also helps provide to have a perspective of the signal behaviour for different levels of detail based on the average configurations. Another benefit of this configuration is to have a sequential visualisation of every baseline and its response in time and frequency (Table 3.5).

P2		BASELINES														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
AMPLITUDE vs TIME	Av 1-1	Strip 201														
	Av 8-1	Strip 202														
	Av 1-32	Strip 203														
	Av 8-32	Strip 204														
PHASE vs TIME	Av 1-1	Strip 205														
	Av 8-1	Strip 206														
	Av 1-32	Strip 207														
	Av 8-32	Strip 208														
AMPLITUDE vs FREQUENCY	Av 1-1	Strip 209														
	Av 8-1	Strip 210														
	Av 1-32	Strip 211														
	Av 8-32	Strip 212														
PHASE vs FREQUENCY	Av 1-1	Strip 213														
	Av 8-1	Strip 214														
	Av 1-32	Strip 215														
	Av 8-32	Strip 216														

Table 3.4: P2 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

P3		Av1-1				Av 8-1				Av 1-32				Av 8-32			
		Amp vs Time	Pha vs Time	Amp vs Freq	Pha vs Freq	Amp vs Time	Pha vs Time	Amp vs Freq	Pha vs Freq	Amp vs Time	Pha vs Time	Amp vs Freq	Pha vs Freq	Amp vs Time	Pha vs Time	Amp vs Freq	Pha vs Freq
BASELINES	1	Strip 101	Strip 105	Strip 109	Strip 113	Strip 102	Strip 106	Strip 110	Strip 114	Strip 103	Strip 107	Strip 111	Strip 115	Strip 104	Strip 108	Strip 111	Strip 116
	2																
	3																
	4																
	5																
	6																
	7																
	8																
	9																
	10																
	11																
	12																
	13																
	14																
	15																

Table 3.5: P3 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.4.3.4 P4 canvas

The P4 configuration permits the user to visualise all the parameters combinations one next to the other for specific averages. Also permits to compare horizontally amplitude and phase per baseline. This allows to make comparisons of the signals amplitude and of the phase patterns for each baseline according to the desired level of detail that the analyst wants to see. This level of detail is expressed in the average configuration of the different sets of plots. This is a very important factor because the analyst can identify a certain region of information that wants to inspect and instantly obtain the plots in a more detailed way (Table 3.6).

P4		BASELINES														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Av 1-1	Amp vs Time	Strip 201														
	Pha vs Time	Strip 205														
	Amp vs Freq	Strip 209														
	Pha vs Freq	Strip 213														
Av 8-1	Amp vs Time	Strip 202														
	Pha vs Time	Strip 206														
	Amp vs Freq	Strip 210														
	Pha vs Freq	Strip 214														
Av 1-32	Amp vs Time	Strip 203														
	Pha vs Time	Strip 207														
	Amp vs Freq	Strip 211														
	Pha vs Freq	Strip 215														
Av 8-32	Amp vs Time	Strip 204														
	Pha vs Time	Strip 208														
	Amp vs Freq	Strip 212														
	Pha vs Freq	Strip 216														

Table 3.6: P4 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.4.3.5 P5 canvas

P5 enables the user to analyse vertically frequency for each average and field of observation. This configuration was designed with the purpose of visualise the signals response by analysing the amplitude with respect to the frequency for each field of observation. By having the field information segmented the analyst will be able to contrast the signal response in frequency with different levels of detail for a specific field of observation (Table 3.7).

P5		AXIS CONFIGURATION: AMPLITUDE vs FREQUENCY																			
		FIELD 0				FIELD 1				FIELD 2				FIELD 3				FIELD 4			
		Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32	Avg 1-1	Avg 8-1	Avg 1-32	Avg 8-32
BASELINES	1	Strip 301	Strip 302	Strip 303	Strip 304	Strip 309	Strip 310	Strip 311	Strip 312	Strip 317	Strip 318	Strip 319	Strip 320	Strip 325	Strip 326	Strip 327	Strip 328	Strip 333	Strip 334	Strip 335	Strip 336
	2																				
	3																				
	4																				
	5																				
	6																				
	7																				
	8																				
	9																				
	10																				
	11																				
	12																				
	13																				
	14																				
	15																				

Table 3.7: P5 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.4.3.6 P6 canvas

The P6 configuration allows the user to analyse horizontally the amplitude response for specific fields of observation, per baseline and average. Also by having the averages configurations one next to another, it is possible to have both a general and a specific perspective of the amplitude response for every field. In general, by considering all the baselines so the analyst will have an overview of all the antennas configurations' behaviour for all the fields. By having the different averages one next to each other a detailed inspection of the amplitude with respect to time may be performed (Table 3.8).

3.4.3.7 P7 canvas

This configuration was designed mainly with the purpose of obtaining a visualisation configuration to perform the inspection of the phase patterns by contrasting its response in frequency. By having the baselines in a horizontal direction, the

P6		AXIS CONFIGURATION: AMPLITUDE vs FREQUENCY														
		BASELINES														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FIELD 0	Av1-1	Strip 401														
	Av8-1	Strip 402														
	Av1-32	Strip 403														
	Av8-32	Strip 404														
FIELD 1	Av1-1	Strip 409														
	Av8-1	Strip 410														
	Av1-32	Strip 411														
	Av8-32	Strip 412														
FIELD 2	Av1-1	Strip 417														
	Av8-1	Strip 418														
	Av1-32	Strip 419														
	Av8-32	Strip 420														
FIELD 3	Av1-1	Strip 425														
	Av8-1	Strip 426														
	Av1-32	Strip 427														
	Av8-32	Strip 428														
FIELD 4	Av1-1	Strip 433														
	Av8-1	Strip 434														
	Av1-32	Strip 435														
	Av8-32	Strip 436														

Table 3.8: P6 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

phase pattern for each baseline and each field can be analysed. Also this configuration enables the analyst to specifically see how each of those patterns responds in function of frequency by a vertical inspection (Table 3.9).

<table>
<tr><th colspan="2" rowspan="3">P7</th><th colspan="20">AXIS CONFIGURATION: PHASE vs FREQUENCY</th></tr>
<tr><th colspan="4">FIELD 0</th><th colspan="4">FIELD 1</th><th colspan="4">FIELD 2</th><th colspan="4">FIELD 3</th><th colspan="4">FIELD 4</th></tr>
<tr><th>Avg 1-1</th><th>Avg 8-1</th><th>Avg 1-32</th><th>Avg 8-32</th><th>Avg 1-1</th><th>Avg 8-1</th><th>Avg 1-32</th><th>Avg 8-32</th><th>Avg 1-1</th><th>Avg 8-1</th><th>Avg 1-32</th><th>Avg 8-32</th><th>Avg 1-1</th><th>Avg 8-1</th><th>Avg 1-32</th><th>Avg 8-32</th><th>Avg 1-1</th><th>Avg 8-1</th><th>Avg 1-32</th><th>Avg 8-32</th></tr>
<tr><td rowspan="15">BASELINES</td><td>1</td><td rowspan="15">Strip 305</td><td rowspan="15">Strip 306</td><td rowspan="15">Strip 307</td><td rowspan="15">Strip 308</td><td rowspan="15">Strip 313</td><td rowspan="15">Strip 314</td><td rowspan="15">Strip 315</td><td rowspan="15">Strip 316</td><td rowspan="15">Strip 321</td><td rowspan="15">Strip 322</td><td rowspan="15">Strip 323</td><td rowspan="15">Strip 324</td><td rowspan="15">Strip 329</td><td rowspan="15">Strip 330</td><td rowspan="15">Strip 331</td><td rowspan="15">Strip 332</td><td rowspan="15">Strip 337</td><td rowspan="15">Strip 338</td><td rowspan="15">Strip 339</td><td rowspan="15">Strip 340</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
<tr><td>5</td></tr>
<tr><td>6</td></tr>
<tr><td>7</td></tr>
<tr><td>8</td></tr>
<tr><td>9</td></tr>
<tr><td>10</td></tr>
<tr><td>11</td></tr>
<tr><td>12</td></tr>
<tr><td>13</td></tr>
<tr><td>14</td></tr>
<tr><td>15</td></tr>
</table>

Table 3.9: P7 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.4.3.8 P8 canvas

This configuration allows the user to contrast the different phase patterns for every baseline and field by having an horizontal visualisation of the baselines and vertical from the fields. Also the accessibility of having the averages one next to the other allows the user to obtain different levels of detail for each field of observation (Table 3.10).

3.4.3.9 P9 canvas

This configuration allows the analyst to obtain a general perspective of the same configurations in average and parameters by showing the plots of each field one next to the other. By having the baselines shown horizontally it is possible to analyse the frequency response for each field of observation and average. Also

P8	AXIS CONFIGURATION: PHASE vs FREQUENCY															
		BASELINES														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
FIELD 0	Av1-1	Strip 405														
	Av8-1	Strip 406														
	Av1-32	Strip 407														
	Av8-32	Strip 408														
FIELD 1	Av1-1	Strip 413														
	Av8-1	Strip 414														
	Av1-32	Strip 415														
	Av8-32	Strip 416														
FIELD 2	Av1-1	Strip 421														
	Av8-1	Strip 422														
	Av1-32	Strip 423														
	Av8-32	Strip 424														
FIELD 3	Av1-1	Strip 429														
	Av8-1	Strip 430														
	Av1-32	Strip 431														
	Av8-32	Strip 432														
FIELD 4	Av1-1	Strip 437														
	Av8-1	Strip 438														
	Av1-32	Strip 439														
	Av8-32	Strip 440														

Table 3.10: P8 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

the plots of the different averages are easily recognisable if some specific region wants to be inspected in more detail (Table 3.11).

P9		AXIS CONFIGURATION: AMPLITUDE vs FREQUENCY																			
		Av1-1					Av 8-1					Av 1-32					Av 8-32				
		FIELD 0	FIELD 1	FIELD 2	FIELD 3	FIELD 4	FIELD 0	FIELD 1	FIELD 2	FIELD 3	FIELD 4	FIELD 0	FIELD 1	FIELD 2	FIELD 3	FIELD 4	FIELD 0	FIELD 1	FIELD 2	FIELD 3	FIELD 4
BASELINES	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	Strip 301	Strip 309	Strip 317	Strip 325	Strip 333	Strip 302	Strip 310	Strip 318	Strip 326	Strip 334	Strip 303	Strip 311	Strip 319	Strip 327	Strip 335	Strip 304	Strip 312	Strip 320	Strip 328	Strip 336

Table 3.11: P9 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.4.3.10 P10 canvas

The P10 configuration permits the user to compare the amplitude levels for all the baselines horizontally. Is worth mentioning that by having the fields discriminated in contiguous rows it is possible to analyse and contrast the signals' response in frequency for each baseline and field. Also this configuration allows the analyst to have a detailed analysis of the variables mentioned before by having the other sets of plots arranged by averages (Table 3.12).

3.4.3.11 P11 canvas

With this configuration it is possible to compare vertically the phase response of each baseline with respect to the frequency and horizontally frequencies for each average and time. By having each field one next to another it also is possible to compare the obtained phase patterns for each baseline. Another important factor is that due to the colour assignation a perspective of the polarisation response in each baseline and field is achieved (Table 3.13).

<table>
<tr><th colspan="2" rowspan="3">P10</th><th colspan="15">AXIS CONFIGURATION: AMPLITUDE vs FREQUENCY</th></tr>
<tr><th colspan="15">BASELINES</th></tr>
<tr><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th></tr>
<tr><td rowspan="5">Av 1-1</td><td>FIELD 0</td><td colspan="15">Strip 401</td></tr>
<tr><td>FIELD 1</td><td colspan="15">Strip 409</td></tr>
<tr><td>FIELD 2</td><td colspan="15">Strip 417</td></tr>
<tr><td>FIELD 3</td><td colspan="15">Strip 425</td></tr>
<tr><td>FIELD 4</td><td colspan="15">Strip 433</td></tr>
<tr><td rowspan="5">Av 8-1</td><td>FIELD 0</td><td colspan="15">Strip 402</td></tr>
<tr><td>FIELD 1</td><td colspan="15">Strip 410</td></tr>
<tr><td>FIELD 2</td><td colspan="15">Strip 418</td></tr>
<tr><td>FIELD 3</td><td colspan="15">Strip 426</td></tr>
<tr><td>FIELD 4</td><td colspan="15">Strip 434</td></tr>
<tr><td rowspan="5">Av 1-32</td><td>FIELD 0</td><td colspan="15">Strip 403</td></tr>
<tr><td>FIELD 1</td><td colspan="15">Strip 411</td></tr>
<tr><td>FIELD 2</td><td colspan="15">Strip 419</td></tr>
<tr><td>FIELD 3</td><td colspan="15">Strip 427</td></tr>
<tr><td>FIELD 4</td><td colspan="15">Strip 435</td></tr>
<tr><td rowspan="5">Av 8-32</td><td>FIELD 0</td><td colspan="15">Strip 404</td></tr>
<tr><td>FIELD 1</td><td colspan="15">Strip 412</td></tr>
<tr><td>FIELD 2</td><td colspan="15">Strip 420</td></tr>
<tr><td>FIELD 3</td><td colspan="15">Strip 428</td></tr>
<tr><td>FIELD 4</td><td colspan="15">Strip 436</td></tr>
</table>

Table 3.12: P10 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

<table>
<tr><th colspan="2" rowspan="3">P11</th><th colspan="20">AXIS CONFIGURATION: PHASE vs FREQUENCY</th></tr>
<tr><th colspan="5">1-1</th><th colspan="5">8-1</th><th colspan="5">1-32</th><th colspan="5">8-32</th></tr>
<tr><th>FIELD 0</th><th>FIELD 1</th><th>FIELD 2</th><th>FIELD 3</th><th>FIELD 4</th><th>FIELD 0</th><th>FIELD 1</th><th>FIELD 2</th><th>FIELD 3</th><th>FIELD 4</th><th>FIELD 0</th><th>FIELD 1</th><th>FIELD 2</th><th>FIELD 3</th><th>FIELD 4</th><th>FIELD 0</th><th>FIELD 1</th><th>FIELD 2</th><th>FIELD 3</th><th>FIELD 4</th></tr>
<tr><td rowspan="15">BASELINES</td><td>1</td><td rowspan="15">Strip 305</td><td rowspan="15">Strip 313</td><td rowspan="15">Strip 321</td><td rowspan="15">Strip 329</td><td rowspan="15">Strip 337</td><td rowspan="15">Strip 306</td><td rowspan="15">Strip 314</td><td rowspan="15">Strip 321</td><td rowspan="15">Strip 329</td><td rowspan="15">Strip 337</td><td rowspan="15">Strip 307</td><td rowspan="15">Strip 315</td><td rowspan="15">Strip 323</td><td rowspan="15">Strip 331</td><td rowspan="15">Strip 339</td><td rowspan="15">Strip 308</td><td rowspan="15">Strip 316</td><td rowspan="15">Strip 324</td><td rowspan="15">Strip 332</td><td rowspan="15">Strip 340</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
<tr><td>5</td></tr>
<tr><td>6</td></tr>
<tr><td>7</td></tr>
<tr><td>8</td></tr>
<tr><td>9</td></tr>
<tr><td>10</td></tr>
<tr><td>11</td></tr>
<tr><td>12</td></tr>
<tr><td>13</td></tr>
<tr><td>14</td></tr>
<tr><td>15</td></tr>
</table>

Table 3.13: P11 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.4.3.12 P12 canvas

There are two main advantages of the P12 configuration. The first one is that is that the analyst is able to compare horizontally the phase patterns for each baseline, field and average and also to analyse the polarisation response. The second one is to be able to compare the response of the mentioned phase patterns in frequency for the same baseline but in the different fields of observation (Table 3.14).

P12		AXIS CONFIGURATION: PHASE vs FREQUENCY														
		BASELINES														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Av 1-1	FIELD 0	Strip 405														
	FIELD 1	Strip 413														
	FIELD 2	Strip 421														
	FIELD 3	Strip 429														
	FIELD 4	Strip 437														
Av 8-1	FIELD 0	Strip 406														
	FIELD 1	Strip 414														
	FIELD 2	Strip 422														
	FIELD 3	Strip 430														
	FIELD 4	Strip 438														
Av 1-32	FIELD 0	Strip 407														
	FIELD 1	Strip 415														
	FIELD 2	Strip 423														
	FIELD 3	Strip 431														
	FIELD 4	Strip 439														
Av 8-32	FIELD 0	Strip 408														
	FIELD 1	Strip 416														
	FIELD 2	Strip 424														
	FIELD 3	Strip 432														
	FIELD 4	Strip 440														

Table 3.14: P12 canvas structure showing the numbers of each strip and horizontal and vertical configurations (axis parameters, averages and baselines)

3.5 Conclusions

The process to define the design of this solution went through several stages. First it was recognised that the pre-generation of plots was the most assertive perspective to this solution. From there, the idea was to produce groups of plots per screen. Once this was tested in the Powerwall it was found that the amount of actions that the user would require were excessive, mainly by not having a dedicated visualisation interactive application that allows the calling of the groups of plots in an effective way. Also it was found that by having the data segmented per screen the way that the plots could be arranged did not allow effective comparative configurations. By identifying those factors it was decided to exploit the Powerwall as a single overall visualisation tool, not considering it divided or segmented by screens. This was the beginning of the strips and canvas concepts. Once that criteria was defined the process of establishing the parameters began. Based on the perspective of an eMerlin analyst, and also considering the variables used to produce the plots delivered by the eMerlin pipeline, the parameters for the plots of this solution were defined. Once the parameters were established, the next step was to define the orientations and order in which the data will be organised. To do this 12 canvases were produced, integrating all the variables considered arranged in different configurations. It is worth mentioning that 6 canvases corresponds to the transpose of the other 6. This was done in order to maximise the utility of a plot. If the same group of plots are organised vertically is possible to compare easily the parameter of the x axis. On the other hand, if the plots are shown horizontally the y axis parameter will become more relevant to analysis purposes.

To perform difficult tasks, the Powerwall is significantly more efficient than a desktop monitor. As commented by (Liu C., Chapuis O., Beaudouin-Lafon M., Lecolinet E. and Mackay W. E., 2014) the use of classification and the partition of information according to classes depending on their properties or attributes is a reasonable direction to validate the design of the manual flagging solution.

Figure 3.7 shows a session intended to provide feedback from the supervisor's perspective in terms of the periodic advance of the designing process.

The working sessions using the Powerwall led to a realistic approach to the design phase. By knowing that this solution must respond and be pertinent to the Powerwall characteristics, the experimental sessions provide a coherent perspective of the assertiveness of the design (Figure 3.8).

Figure 3.7: Inspection session of the designing process.

Figure 3.8: Working session of the designing process.

Chapter 4

Implementation of the solution

4.1 Introduction

Once the parameters for plots and strips were defined, the next step was to create an implementation that aimed to reduce the time and actions required by the analyst or user. The implementation of the solution was done in three main phases: plotting, merging the plots to canvas and integration of the canvas with the visualisation interface. To achieve those phases five python scripts were created. One for the plotting and the other four for the merging of the strips. These are external scripts that were designed to minimise the required interaction of the final user. Just by calling the scripts 112, strips are produced and automatically merged in order to obtain the 12 canvases.

To manage interoperability there can be many approaches as tangible remote controllers as seen in (Jansen Y., Dragicevic P. and Fekete J. D., 2012) or by multi-touch applications as proposed in (Westing B., Urick B., Esteva M., Rojas F. and Xu W., 2011). Although this implementation uses an existing visualisation software controlled by mouse and key commands, the methodology used permits the resulting visualisation configurations to be adapted in order to be controlled by different interaction devices.

In order to implement the solution, the design was developed for the technical specifications of the University of Leeds Powerwall. The Powerwall is built with 6 high resolution screens. Each screen has a resolution of 3,840 x 2,160 pixels for a total of 8,294,440 pixels per screen. The complete Powerwall system provides a 49,776,400 pixel resolution. This configuration allows the user to obtain a single wide screen which generates a visualisation experience with multiple applications. The size of the synchronised screens, permits the users not only to obtain enhanced images in high resolution, but also the opportunity to visualise a great amount of data at the same time.

Is important to take into account the scalability limits of the solution. As seen in the work of (Papadopoulos C., Mirhosseini S., Gutenko I., Petkov K., Kaufman A. E. and Laha B., 2015). The definition of the strips resolution sizes in height and width were defined taking the Powerwall capabilities as a reference.

After developing the implementation and by making tests on the Powerwall it was clear that although the configurations were adequate, the sizes of the plots have to be different for some specific canvas configurations. That is the reason for having the same configurations of parameters but with different references. Each of those corresponds to a size of plot oriented to maximise the resolution characteristics of the Powerwall. The pixel sizes for each strip is shown in Table 4.1.

canvas reference	strips references	matrix (col x row)	strip size (px)	
			width	height
P1 P3	101-116	1x15	720	4320
P2 P4	201-216	15x1	11520	270
P5 P7 P9 P11	301-340	1x15	576	4320
P6 P8 P10 P12	401-440	15x1	11520	216

Table 4.1: Description of the resolution used for each strip (pixel).

As mentioned before, the intention is to maximise the capabilities of the Powerwall. Taking as reference the design structure of the solution the sizes for each strip were established in function of the overall area of the 6 screens (11520 x 4320 px). For example, for canvas P1 and P3 sixteen strips are merged horizontally. So that is the reason why each of this strips have a width of 720 px (720x16=11520). Now if P2 and P4 are considered, sixteen strips are merged vertically. So the height for the strips is 270 px considering that (270x16=4320).

For P5, P7, P9, and P11, twenty strips are merged horizontally. The width for the strips is then 576 px (576x20=11520). Finally for P6, P8, P10, and P12 twenty strips are merged vertically so the height is 216 px (216x20=4320).

The criteria to select the type of plots responds to the preferences presented by the analyst. In terms of the type of graphics delivered by CASA it was decided to use the default parameters. Although there are other graphic types such as scatter and bars the same type of graphics were maintained to preserve continuity in terms of how the process of manual flagging is performed, allowing a familiar environment for the users. In terms of the colours used to generate the plots, default parameters were used. For the building of the 12 canvases, two main parameters were selected to show different components of the data. Those parameters were spectral window and polarization. Spectral windows were used for the plots where frequency was not an axis parameter and polarization for the plots where frequency was considered as an axis parameter. The intention of doing that was to provide clarity of the frequency respond, for the first case, and to add another important parameter as polarization for the second case. The colour assignation also responds to the suggestions provided by the analyst. The other main criteria was the definition of the strips of plots taking the baselines as reference (section 3.3.2).

It should also be mentioned that is not enough just to consider the overall dimensions of the visualisation system and from there calculate the maximum size of the images. For the Leeds Powerwall, implementation tests show that those sizes were appropriate. In general, cognitive factors as acuity need to be considered. According to the work of (Yost B., Haciahmetoglu Y. and North C., 2007), visual acuity determines the scalability limit in a large high resolution display. This term refers to to the relation between the number of dots per inch (DPI) and the limit for a single pixel to be seen. In terms of the resolution of a system which is the number of pixels available to be used if the DPI is higher than a certain value an individual pixel will not be seen no matter how close the user is to the visualization display.

4.2 Software architecture

To develop a low cost solution in terms of time and interaction, the following software architecture were implemented. The conceptual diagram of this architecture is shown in Figure 4.1. Basically the user interacts with one folder “powerwall”. This folder contains one sub folder “canvas” and four files. To produce the strips of plots the user calls plot.py from CASA version 5.4. This file is linked with 3C27701.avg.ms which is the file that contains a radio telescope dataset. Once plot.py is executed 112 strips of plots are generated in the canvas folder. The next step is to open a Python3.5 environment and execute merge1.py and then merge2.py. Once these files are executed the 12 canvases are produced and automatically saved in the canvas folder. The final step is to open the 12 canvases as layers in an image visualisation software.

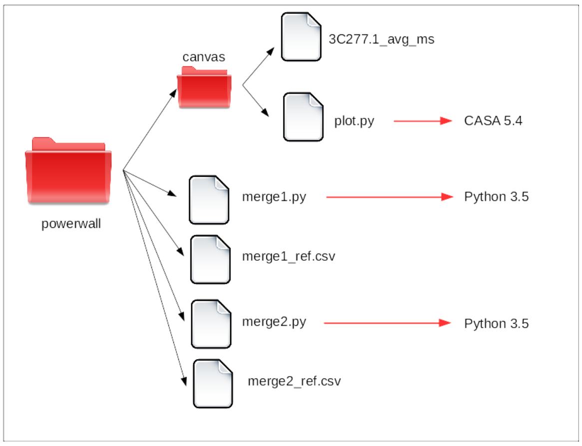


Figure 4.1: Conceptual diagram of the software architecture.

The files merge1_ref.csv and merge2_ref.csv are linked with merge1.py and merge2.py. In those files are specified the order for the merging of the strips of plots. The following sections explain in detail the internal structure of the plotting and merging scripts and the visualisation interface.

4.2.1 Plotting the strips

The implementation of the plots is based on the configuration structure as explained in subsection 2.3, taking as reference the parameters of plots (axis configuration and coloraxis), strips (channel and time averages, fields of observation, spectral windows and polarisations) and canvas (vertical and horizontal orientation). Then a reference was assigned to each strip of plots. This reference intends to provide an understandable nomenclature in order to built each of the canvas configurations. Finally, the plots were generated in a tailor-made python script, which automatically produces the required strips of plots to produce each canvas. The result are the 112 strips of plots required to produce the 12 canvases. As mentioned before, a python script was implemented in order to automatically generate all the strips of plots. This script is executed from CASA as an external file. After the script is executed the 112 strips are produced and saved in the folder in which the file is storage. By having the plots referenced for every canvas configuration the script automatically generates each plot according to its reference. Figure B1 in Appendix B shows the script to produce the first strip of plots. This command produces a time vs amplitude plot, using baseline as iteration parameter, spectral window as colour assignation, average channel of 1, average time of 1, shows all the spectral windows, RR and LL polarisations and all fields of observation. The strip is defined as a 1 column x 15 rows matrix, 720 px in width and 4320 px in height.

4.2.2 Merging the strips to canvas

Once the 112 strips are produced the next step is to merge them in a specific order according to each canvas configuration. Each strip is configured as a 15x1 or 1x15 matrix. In order to optimise the process of producing each canvas according to the configurations established in Chapter 2, two merging scripts were implemented. The merging python scripts allows to merge automatically all the strips that are used for every canvas. The script merge1.py is used for the horizontal orientation merging and merge2.py for the vertical merging. The input data are 112 strips, each one with 15 individual plots and the output are the 12 canvases already ordered according to the design parameters.

In the Figure B2 Appendix B is shown the merging script for the horizontal orientation merging.

Both merging scripts merge1.py and merge2.py are linked with a reference script which defines the order of the strips for each canvas. This scripts are merge1_ref.csv and merge2_ref.csv. The uneven canvases are merged horizontally and the even vertical. For example, to produce Canvas1 the strips 101 to 116 are merged horizontally and to produce Canvas2 the strips 201 to 216 are merged vertically.

The following images corresponds to the result of the merging of the strips by using the plotting and merging scripts for the first four canvases. Those images are provided taking as reference the configuration arranges for each canvas in terms of the strips nomenclature assigned. Figure 4.2 corresponds to canvas P1, Figure 4.3 to canvas P2, Figure 4.4 to canvas P3 and Figure 4.5 to canvas P4.

4.2.3 Integration of the canvases to the visualisation interface

The final step is to integrate the resulting 12 canvases to the interface which allows to have the control to the visualisation of each of them. The selected interface is image software called GNU GIMP. As each canvas is a PNG file, each of them has to be imported as a layer from the software. After this, is possibe to control the calling of each canvas by a layer control window as it can be seen in the right of the Figure 4.6 and the final result as seen on the Powerwall in Figure 4.7.

With all of the above complete, the solution is ready to be used. In order to make the vertical and horizontal scans and zooms of the images in the Powerwall seven commands are used:

1. Scroll up : up horizontal scan.
2. Scroll down: down horizontal scan.
3. Shift / scroll up : left vertical scan.

Figure 4.2: Canvas P1 final result (as referenced in Table 3.3).

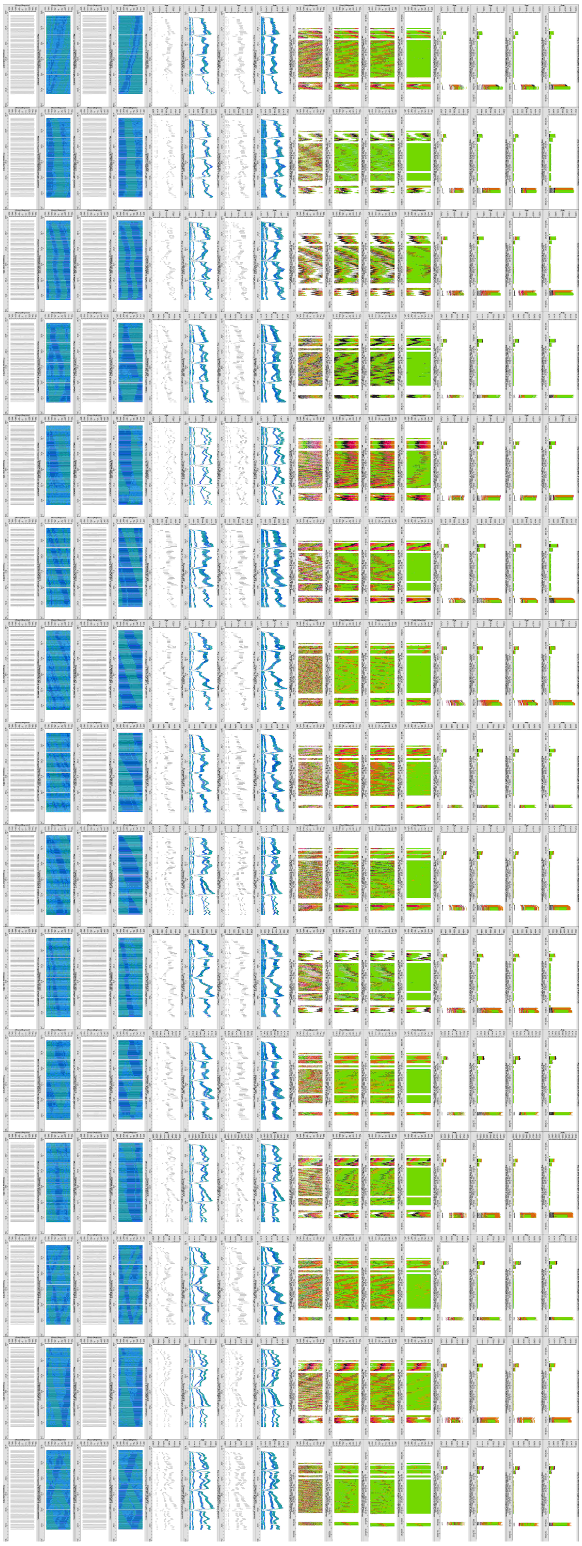


Figure 4.3: Canvas P2 final result (as referenced in Table 3.4).

Figure 4.4: Canvas P3 final result (as referenced in Table 3.5).

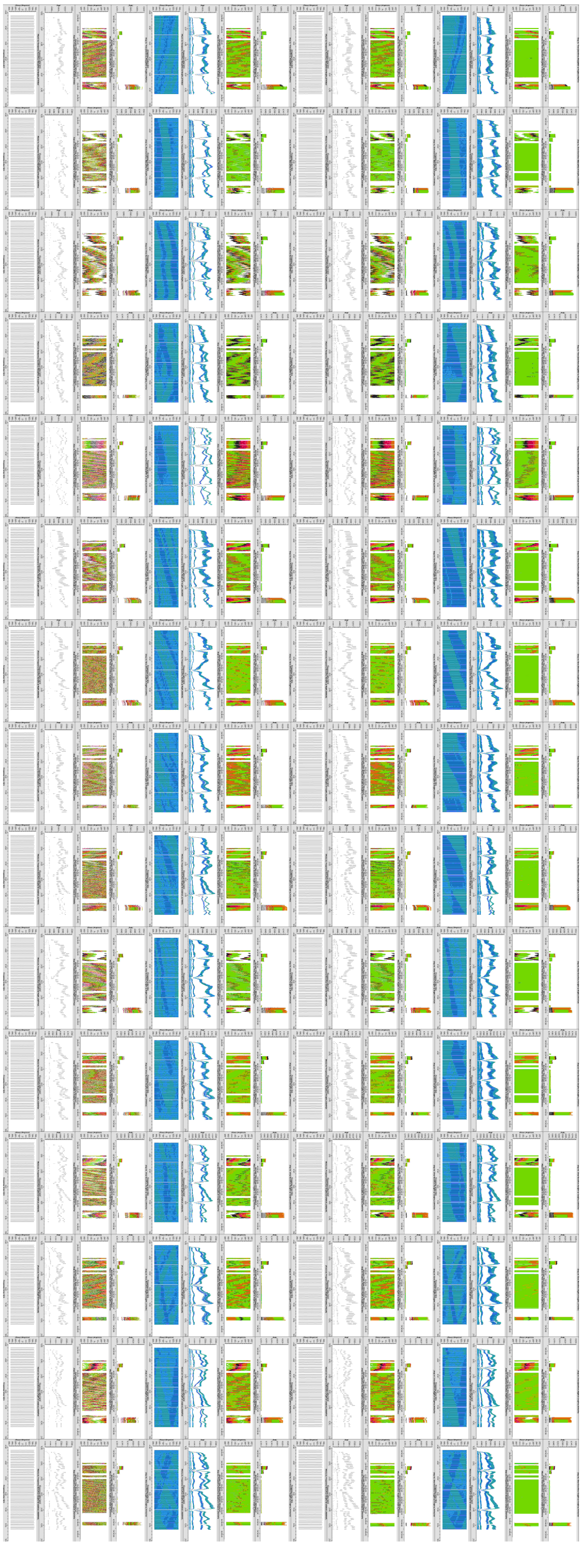


Figure 4.5: Canvas P4 final result (as referenced in Table 3.6).

4. Shift / scroll down: right vertical scan.

5. Control / scroll up: zoom in.

6. Control / scroll down: zoom out.

7. Ctrl / Shift / J: Zoom out to a complete canvas visualisation.

By the use of those commands, the user can inspect of any canvas and also control the sections of data that need to be analysed. The analysis can be either by scanning vertically or horizontally the strips of plots and/or by obtaining zooms of any desired data region.

A visualisation interface can be adapted to the specific requirements of the application. A example of this is the work of (Forlines C., Esenther A., Shen C., Wigdor D. and Ryall K., 2006) where a single display was adapted for multi-device, multi-user geospatial exploration. This was done by adding new tools to an existing visualisation software. This modifications allow synchronised and coordinated viewing, visually separating layers, touching to navigate and touching to reference as interaction resources. This is mentioned because has similarities to the interface used in this solution. Existing visualisation software can become a valuable interface solution by implementing comprehensive arranges of data with an understandable segmentation of parameters grouped by layers or canvases in our case.

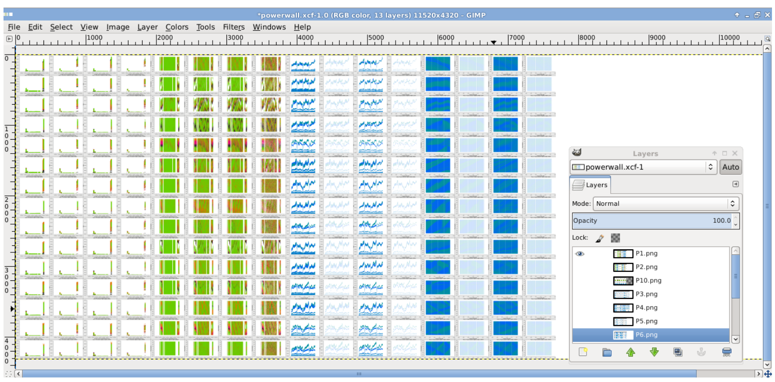


Figure 4.6: Canvas integrated with the visualisation interface.

Figure 4.7 shows the final result of the canvas P1 as visualised in the Powerwall. The layer window that controls the calling of all the canvases can also be seen. Appendix 1 User Manual provides a detailed description of this process.

Figure 4.7: Canvas P1 as seen in the Powerwall.

4.3 Identification of analysis possibilities

Once the canvases were produced, certain aspects related to signal response were identified. Due to the canvas configurations there are specific comparisons that can be achieved by the use of each o them. Some of the comparisons that are useful in the manual flagging are shown. P1 corresponds to canvas 1, P2 to canvas 3 and P3 as to canvas 5. The sub indexes define the horizontal and vertical lines in the graphics.

P1-1: Variation of amplitude levels in time contrasting the average configurations (Figure 4.8).

P1-2-1, P1-2-2, P1-2-3: Presence of signals of different spectral windows depending on the average configuration (Figure 4.8).

P1-3: Amplitude synchronisation in time between baselines (Figure 4.8).

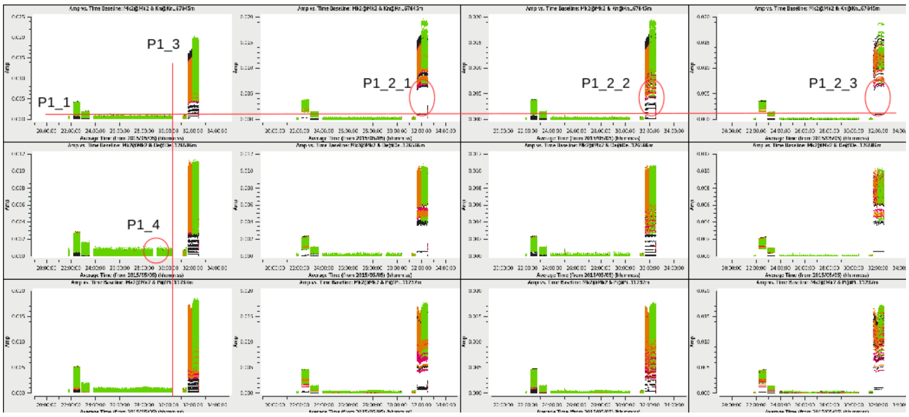


Figure 4.8: Vertical baseline sub-configuration / Canvas 1: columns 1-4 rows 1-3.

P1-4: Absence of signals in specific baselines discriminated in time (Figure 4.8).

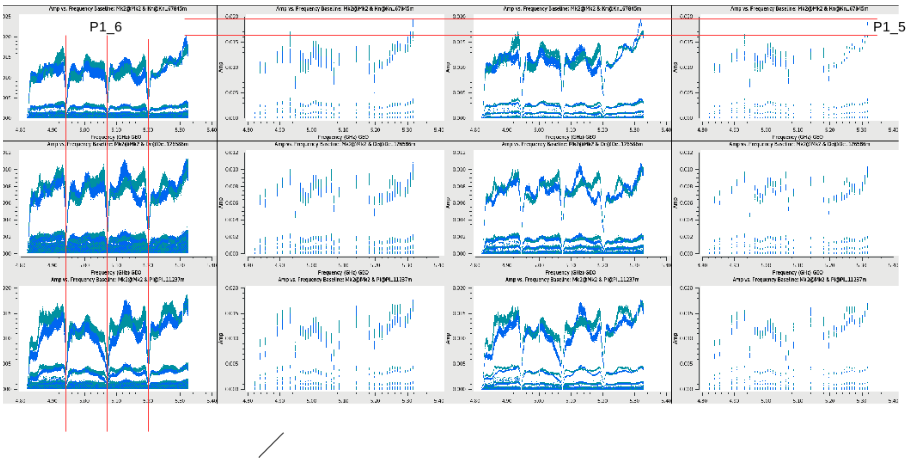


Figure 4.9: Vertical baseline sub-configuration / Canvas 1: columns 9-12 rows 1-3.

P1-5: Identification of maximum amplitude peaks levels for each average configuration (Figure 4.9).

P1-6: Corroboration of synchronisation in the initial and ending times for each spectral window (Figure 4.9).

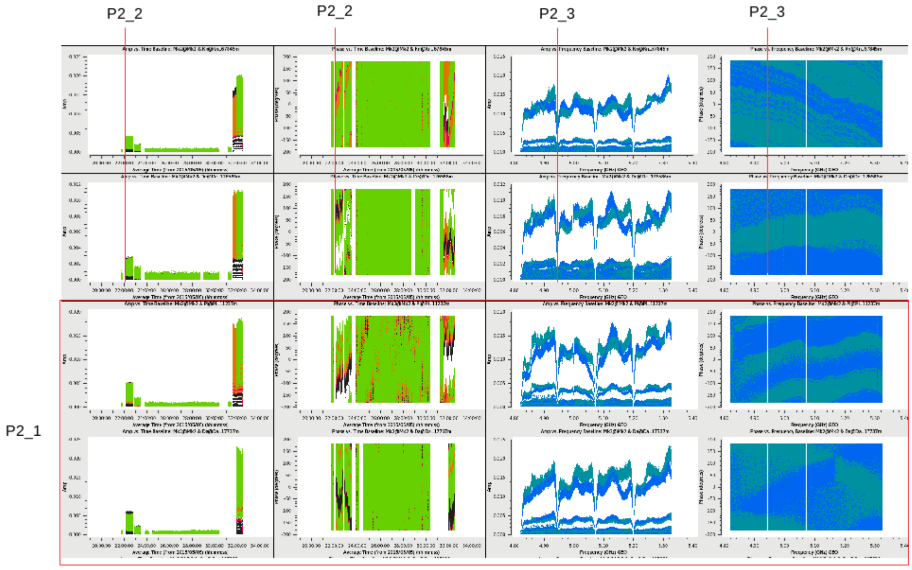


Figure 4.10: Vertical baseline sub-configuration / Canvas 3: columns 1-4 rows 1-3.

P2-1: Behaviour of each baseline in time comparing time v/s amplitude, time vs phase; and in frequency comparing frequency vs amplitude and frequency vs phase (Figure 4.10).

P2-2: Synchronisation in time by the inspection of starting points for each spectral window comparing amplitude and phase for each baseline (Figure 4.10).

P2-3: Visualisation of frequency patterns for each spectral window comparing amplitude and phase for each baseline (Figure 4.10).

P3-1: Amplitude response in time for each field, baseline and average configuration (Figure 4.11).

P3-2: Visualisation of the maximum peaks in amplitude for every average configuration and baseline (Figure 4.11).

4.3.1 Conclusions

The implementation of the pre-generation of plots as part of this solution were segmented in two main areas: plotting and merging. To aboard the plotting, it was decided to use the CASA plotms application. To use this, it was necessary

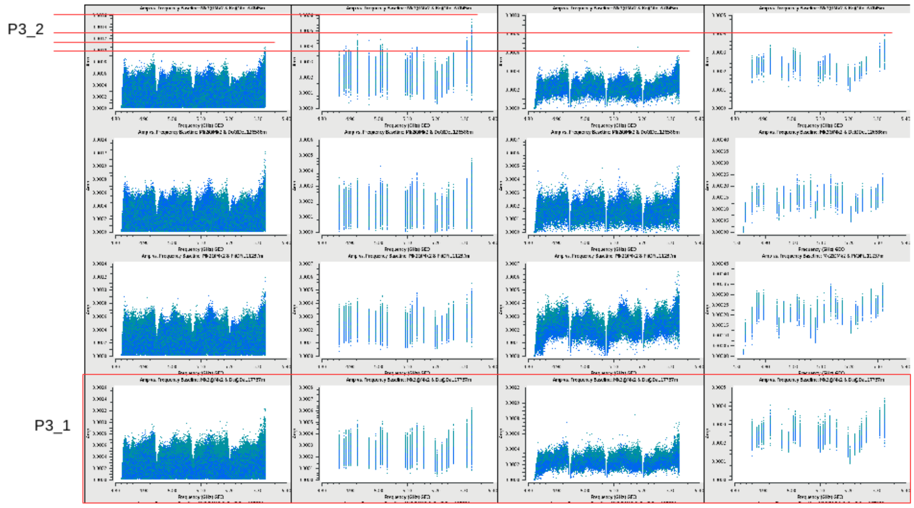


Figure 4.11: Vertical baseline sub-configuration / Canvas 5: columns 1-4 rows 1-3.

to understand how the CASA internal code structure worked; this in function of having an automatic generation of plots. Once this structure were understood the next step were to build a python script (plot.py) which will be externally called. The results of this were found appropriate in terms of an accessible solution for the users to change the parameters for further tailor-made configurations. This script was segmented by strips of plots in order to be understandable for analysts without to much experience in programming. This structure only requires to change the consecutive numbers of the strips and the desired parameters and values. The other aspect which was merging of the strips also responds to the intention of scalability. In order to produce the merge of the strips generated by the plot.py script the analyst only requires to define the order and orientations of the desired canvas by modifying the merge.py scripts. From a general perspective it can be said that by the use of those scripts, that were written to be easily modify and comprehend, any user can rapidly merge the strips to produce their own canvas configurations according to their preferences.

Chapter 5

Evaluation of the solution

5.1 Introduction

The evaluation of large scale visualisation systems demands a particular approaches in terms of the required variables that allows to establish effectiveness of a given system. If considering the work of (Papadopoulos C., Mirhosseini S., Gutenko I., Petkov K., Kaufman A. E. and Laha B., 2015) in terms of the evaluating the scalability limits of an immersive high resolution visualisation environment some criteria can be extracted. It has to be noted that this work used a different visualization configuration compared to the one used in our project. In our case, the Leeds Powerwall uses six high resolution screens, the Reality Deck, used for the mentioned work, uses 416 LCD panels obtaining a 360 degree visualisation system. This means that the users are surrounded by the screens experiencing an immerse visualisation experience. Nevertheless the tasks used can be applied to any visualisation system. Those tasks were visual search, attribute search, comparisons and pattern finding. The reason to highlight this example is to contextualise the kind of metrics that are used in order to verify the effectiveness of a visualisation system.

Another example is the work of (Shupp L., Andrews C., Dickey-Kurdziolek M., Yost B. and North C., 2009) which mainly intend to analyse the performance and behavioural impact of curving high resolution displays. To accomplish this evaluation tasks were also proposed by changing the configurations and by proposing each time harder tasks.

The evaluation to measure the effectiveness of the manual flagging solution extracts some criteria used in those evaluation projects. In order to establish the validity of our solution the analyst was asked to perform a visual inspection by comparing patterns in order to identify specific regions, as proposed in the work of (Papadopoulos C., Mirhosseini S., Gutenko I., Petkov K., Kaufman A. E. and Laha B., 2015). Also by contrasting the evaluator's session 1 and 2, two configurations were compared (single monitor in the first session and the Powerwall in the second) and that has a similar approach to the work of (Shupp L., Andrews C., Dickey-Kurdziolek M., Yost B. and North C., 2009).

Other examples of evaluation are the multiscale interaction technique by (Peck S. M., North C. and Bowman D., 2009) and the networks and tasks analysis by (Ebert A., Thelen S., Olech P. S., Meyer J. and Hagen H., 2010).

The evaluation of the solution was focused on the analysis of three main variables: time, interaction and accuracy. In terms of time the solution was analysed using different hardware and comparing the different amount of time required to complete the computation of the 12 canvases. This information was contrasted to the state-of-art method that is used in manual flagging (as seen in Chapter2). Also the time that the analyst required to perform a flagging process was analyzed by the use of the solution compared with the time required without using it. The other parameter was interaction. This focused on the actions that the user needs to make to achieve the same quantity of plots that the solution provides. Finally, in terms of accuracy this evaluation intends to have a perspective in terms of finding if the solution helps to produce more accurate flagging in terms of the recognition of signal patterns that usually are not identified. This evaluation was supported by testing the solution with the collaboration of a radio astronomer who performs manual flagging using eMerlin data.

5.2 Technical evaluation

The solution presented in this project relies in the pre-generation of images for a subsequent analysis in the Powerwall. This part of the evaluation provides data about the time required to produce those images. In order to establish the time required to make the plots for the 12 canvases two tests were developed. The plot.py script was run on two computers with different hardware characteristics. PC1 had 4 GB of RAM and an Intel(R) Core(TM) i3-2310M CPU @ 2.10GHz processor. The PC2 had 8 GB of RAM and an Intel(R) Xeon(R) CPU E3-1225 V2 @ 3.20GHz processor.

To plot the 112 strips the PC1 took 5 hours 5 minutes 25 seconds and the PC2 2 hours 34 minutes 1 second. This showed that by using the PC2 the same process took 2 hours 31 minutes less.

As it can be seen, there is a big difference in the required time according to the hardware characteristics of the computers in which the plotting script is run.

Figure 5.1 shows the time for each strip series and the total amount of time using PC1 and PC2. Figure 5.2 shows the comparative for each individual strip for both machines.

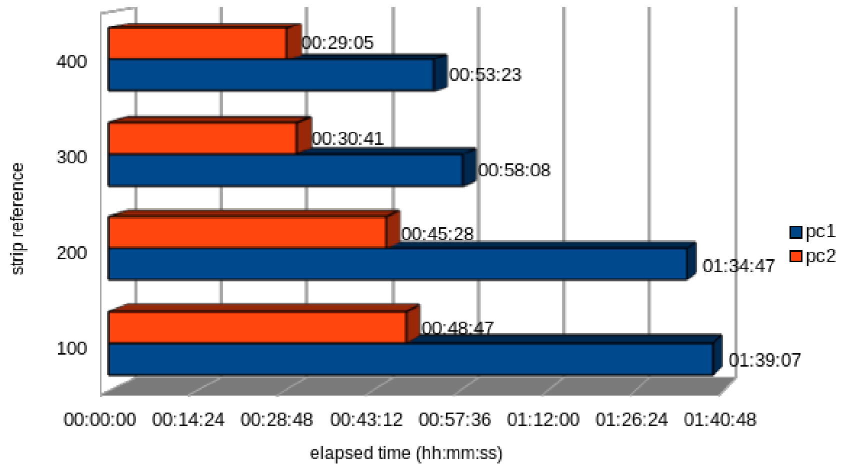


Figure 5.1: Graphic of strip series v/s elapsed time (for PC1 and PC2).

5.3 User evaluation

To be able to analyse the effectiveness of the proposed solutions, an user evaluation process was performed. This analyst was a Astrophisycs PhD student who

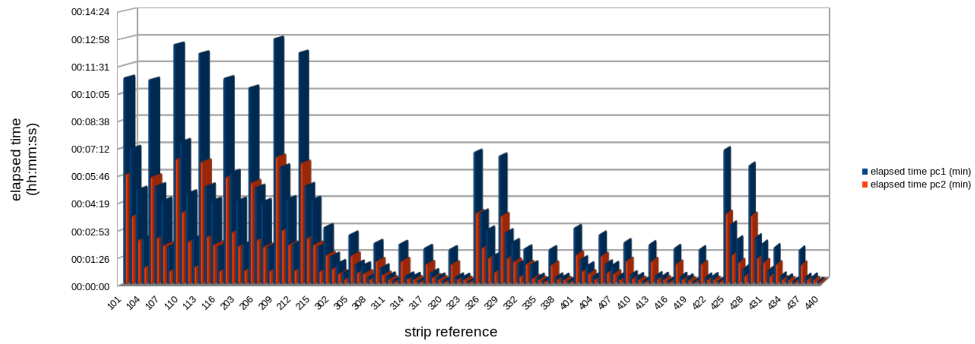


Figure 5.2: Graphic of individual strips v/s elapsed time (for PC1 and PC2).

works in the manual flagging using eMerlin data. Although the same radioastronomical observation was used for both evaluation sessions they differ in the stage of flagging as seen in Fig. 2.18 (Stages of the eMerlin pipeline). The data set for the first evaluation session had no flagging (before the 1st flagging stage). The second one had passed the first step of flagging (after the 1st flagging stage).The data sets used came from the eMerlin data archive and it was gathered using the complete configuration of the array by using the 6 antennas. This process was divided into two sessions with different purposes. The first session characterises how the expert develops the manual flagging. In the second session the Powerwall solution was tested using tailor-made canvases, as proposed by the analyst. This involved using 4 of the original canvases and producing 6 new canvases with the specifications of the analyst.

This evaluation was segmented in two main parts that consolidate the obtained data. These parts are:

1. identification of flagging points
2. scripts and verification

The identification of flagging points identifies specific data segments that require to be flagged. The scripts and verification process handle the writing of the script and the validation of the results.

5.3.1 Session 1 Analyst's manual flagging standard procedure

The objective of this session was to document the procedure followed by the analyst in terms of the required actions to do manual flagging using an eMerlin data file. To do this the analyst was asked to identify 5 flagging points and perform the flagging.

In this session, the analyst performed manual flagging as she usually does. This procedure began with the call of the .ms file to the find the characteristics of the data by using browsetable from the listobs commands (as seen in sub-section 2.3.2). After this the identification of the flagging points starts to write and execute the flagging scripts and finally confirm the results. During this process the actions to perform the tasks were identified in order to be compared with the ones performed when using the Powerwall solution. The session one is described in Tables 5.2, 5.3, 5.4 and 5.5.

As it can be seen in the flagging1 process (actions 1 to 14 in Table 5.2) the initial action is to call a listobs command from CASA (action 1) to retrieve sources numbers, identification nomenclatures and have the understand the characteristics of the data. In this step (action 2) the analyst seeks specifically the number of scans, names and id of the fields of observation. From this the analyst recognised the purposes of each field of observation in terms of the flux, bandpss and phase calibrator and target, "usually you start seeing the calibrators". This initial analysis is performed in order "to know how are they gonna look when plotted".

After this inspection the first plot was done (actions 3 to 4). This was an amplitude vs time plot to get a general perspective of how the data looks. The calling of this plot was done by writing a script with the intention of changing some values later interactively in plotms. After some inspections, the analyst changed the values of colour attributes from spectral window to field ("to see different sources in different colours"; actions 5 and 6). After zooming in and out this plot, certain actions were repeatedly performed (actions 7 to 13). Basically the analyst started zooming the plot by segments or chunks in a progressive way. Each time that some region is to be inspected in a more detailed way, from two to five zoom-in actions were performed. In this part some observations were produced, e.g. "Is strange this part... I would like to see better this tiny region... I look for consistent patterns of set of points, here is a discrepancy so in time I would probably look deeper that segment".

Once some specific regions were recognised as noise, the analyst began to take notes ("I start taking notes of time points and recognition of the source"; action 14). After this the initial notes were finished and the first flagged point was identified.

Once the first flagging point was described and noted, she started the analysis of the next source. Then a progressive inspection in time of the data starts, to obtain the second flagged point (actions 15 to 58, Table 5.2), the third (actions 59 to 80, Table 5.3), fourth (actions 81 to 107, Table 5.3) and fifth (actions 108 to 151, Table 5.4). After this "a final checking is done to confirm the times are right before making the flagging scripts". The final part is to write and execute the flagging scripts according to the notes taken (action 152, Table 5.5). The notes are verified by a plotms option that is called autologger (action 156). This function allows the user to know the specific characteristics of a segment of data localised in the plot. By drawing a rectangle in the selected region, CASA automatically delivers the information corresponding to that specific segment "now I go to the locating button to go to the logger and see the characteristics of the selected data". This examination also required some specific regions to be plotted and the analyst to make closer inspections by zooming into specific chunks of data (actions 153 to 181).

Although a complete flagging process could involve tens to a hundred flagged points, for the purpose of this evaluation five flagging points were sought and therefore 5 flagging scripts were produced. Also, in a full flagging process the analyst performs those steps first by plotting amplitude vs time, then amplitude vs phase, then amplitude vs frequency and finally phase vs frequency.

5.3.2 Session 2

The second evaluation session involves producing a new series of canvases. This was done in order to provide the analyst with tailor-made configurations. The following sections describe the new canvases and then the evaluation session number two.

5.3.2.1 Analyst tailor-made canvases

Six new canvases were produced for the second session of the evaluation. The Table 5.1 shows the parameters for canvases P101, P102, P103, P104, P105, and P106.

Next, the tables with the strips' parameters and references for each canvas and the resulting canvas images after the strips are merged are presented. Figure 5.3 corresponds to canvas P101 and P102, Figure 5.4 to canvas P103 and P104, and Figure 5.5 to canvas P105 and P106.

5.3.2.2 Manual flagging using the Powerwall solution

This session sought to observe and characterise the response of the analyst in function of the interaction with the Powerwall solution. Initially, the analyst began to inspect the canvas to identify 5 flagging points. Then, one script was written and executed in order to confirm the effectiveness of the flagged area. As in Session 1 the actions were registered in order to be compared with the results obtained that session. The session two is described in Tables 5.6 and 5.7.

Once the new configurations were presented to the analyst using the Powerwall visualisation interface, it was decided to achieve 5 regions or segments of data to be flagged and then produce the detailed inspection in CASA to finally write a flagging scripts for one flagging point. The first step was to select the P101 canvas in order to analyse amplitude vs frequency (action 1). From an initial inspection the analyst identified spikes in the spectral window three. From there she concluded that some channels at the end of that spectral window required flagging. From an horizontal scan through all the baselines she identified that the situation repeats in all the baselines ("for the different baselines there is high points here in all of them"; actions 2 and 3 in Table 5.6). After this, notes were taken and the first flagging point identified (action 4). Then "this particular strip shows us the hole fields together so I am going to do P103 cause then you get colourised by fields"; (actions 5 and 6).

canvas	1x15 strip reference (horizontal)	canvas	15x1 strip reference (vertical)	X axis	y axis	coloraxis	average channel	average time	field	spw
	101		201	time	amplitude	spectral window	1	1	all	all
	102		202	time	amplitude	spectral window	8	1	all	all
	103		203	time	amplitude	spectral window	1	32	all	all
	104		204	time	amplitude	spectral window	8	32	all	all
	109		209	frequency	amplitude	polarization	1	1	all	all
	110		210	frequency	amplitude	polarization	8	1	all	all
	111		211	frequency	amplitude	polarization	1	32	all	all
P101	112	P102	212	frequency	amplitude	polarization	8	32	all	all
	101		201	time	amplitude	spectral window	1	1	all	all
	109		209	frequency	amplitude	polarization	1	1	all	all
	102		202	time	amplitude	spectral window	1	1	all	all
	110		210	frequency	amplitude	polarization	8	1	all	all
	103		203	time	amplitude	spectral window	1	32	all	all
	111		211	frequency	amplitude	polarization	1	32	all	all
	104		204	time	amplitude	spectral window	8	32	all	all
	112		212	frequency	amplitude	polarization	8	32	all	all
	1201		11201	time	amplitude	spectral window	1	1	all	all
	1202		11202	time	amplitude	spectral window	all	1	all	all
	1203		11203	frequency	amplitude	spectral window	1	1	all	all
	1204		11204	frequency	amplitude	spectral window	1	all	all	all
	1205		11205	time	amplitude	field	1	1	all	all
	1206		11206	time	amplitude	field	all	1	all	all
	1207		11207	frequency	amplitude	field	1	1	all	all
P103	1208	P104	11208	frequency	amplitude	field	1	all	all	all
	1209		11209	time	amplitude	polarization	1	1	all	all
	1210		11210	time	amplitude	polarization	all	1	all	all
	1211		11211	frequency	amplitude	polarization	1	1	all	all
	1212		11212	frequency	amplitude	polarization	1	all	all	all
	1213		11213	time	amplitude	antenna1	1	1	all	all
	1214		11214	time	amplitude	antenna1	all	1	all	all
	1215		11215	frequency	amplitude	antenna1	1	1	all	all
	1216		11216	frequency	amplitude	antenna1	1	all	all	all
	2001		20001	time	phase	polarization	all	1	0	0
	2002		20002	time	phase	polarization	all	1	0	1
	2003		20003	time	phase	polarization	all	1	0	2
	2004		20004	time	phase	polarization	all	1	0	3
	2005		20005	time	phase	polarization	all	1	1	0
	2006		20006	time	phase	polarization	all	1	1	1
	2007		20007	time	phase	polarization	all	1	1	2
P105	2008	P106	20008	time	phase	polarization	all	1	1	3
	2009		20009	time	phase	polarization	all	1	2	0
	2010		20010	time	phase	polarization	all	1	2	1
	2011		20011	time	phase	polarization	all	1	2	2
	2012		20012	time	phase	polarization	all	1	2	3
	2013		20013	time	phase	polarization	all	1	4	0
	2014		20014	time	phase	polarization	all	1	4	1
	2015		20015	time	phase	polarization	all	1	4	2
	2016		20016	time	phase	polarization	all	1	4	3

Table 5.1: List of parameters for each strip of plots to produce P101 to P106 canvases (strip number for horizontal and vertical configurations, x and y axis parameters, colour assignation, time and channel averages, fields of observation and spectral window)

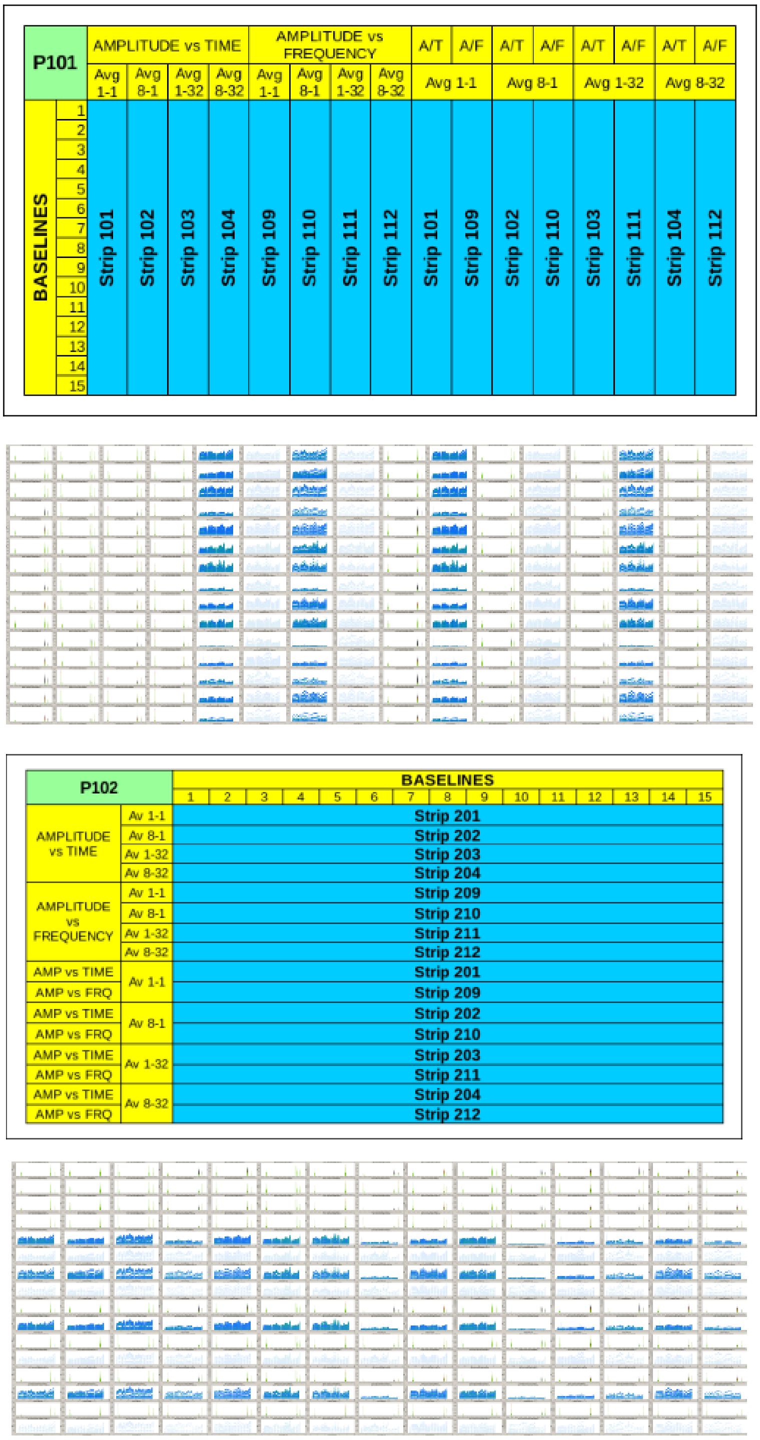


Figure 5.3: P101 and P102 canvases.

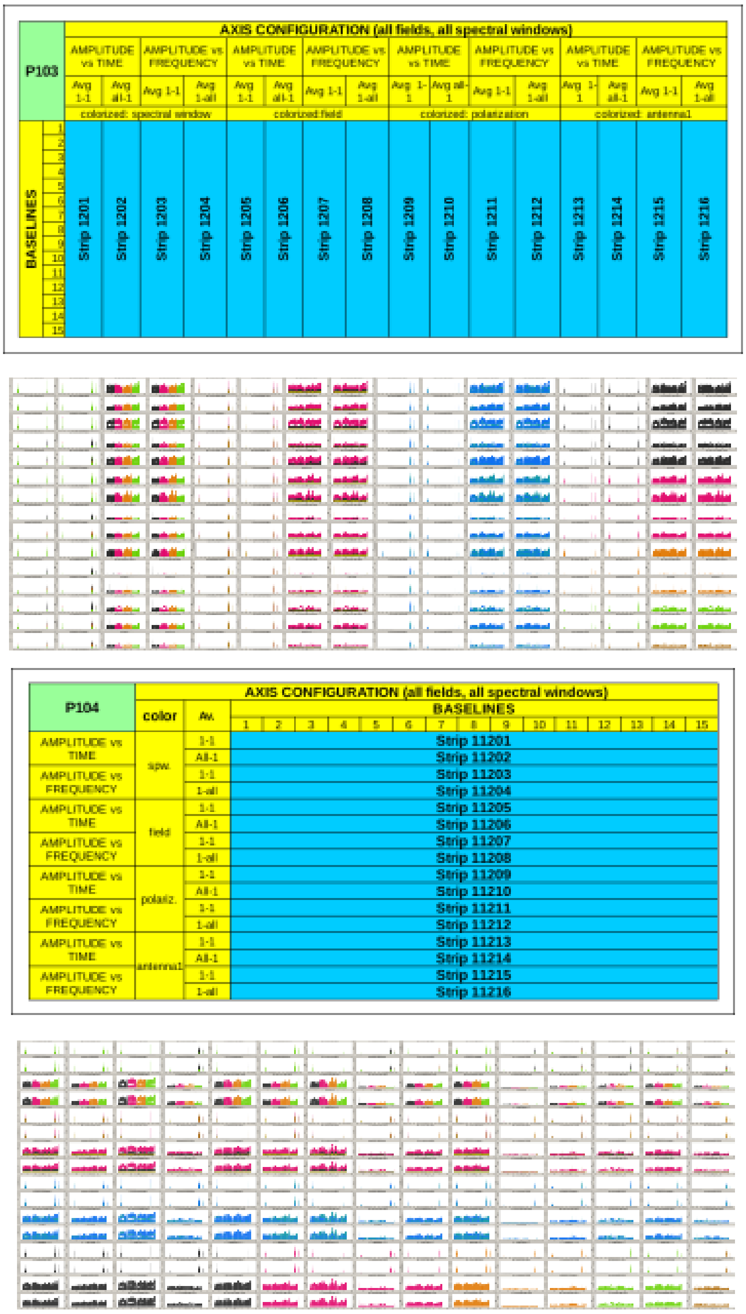


Figure 5.4: P103 and P104 canvases.

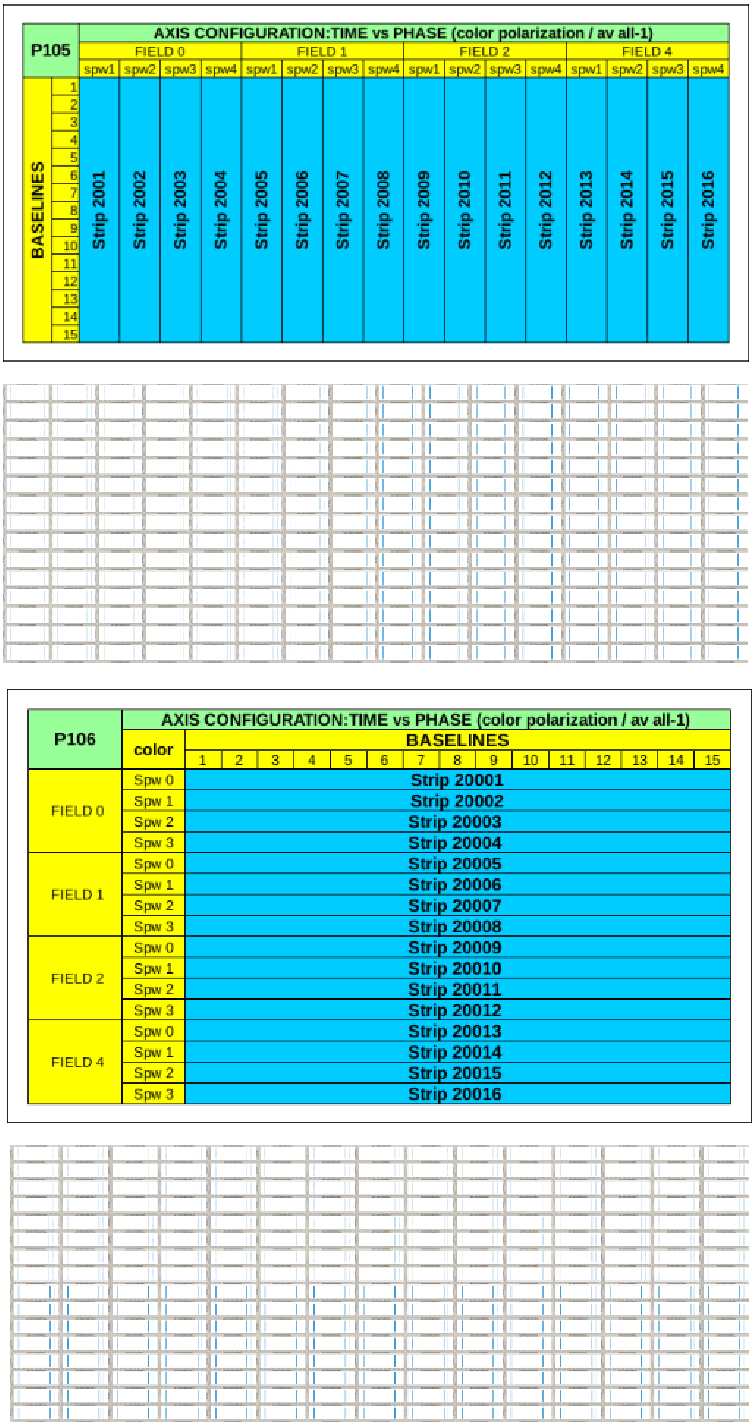


Figure 5.5: P105 and P106 canvases.

Once P103 was deployed and by scrolling horizontally to get to the desired strip, which was amplitude vs frequency colourised by field, something interesting occurred (actions 7 and 8). The intention to make different orientations, according to the canvas designs, were to provide different perspectives to analyse the information. Nevertheless the decision of the analyst in this point was to make an inspection by scanning vertically (lines 9 and 10) and then horizontally the strips provided in P103 ("in the first spectral window there is a spike so I can think is RFI seeing in the calibrators and not in the fields"). From here on, notes were taken and the second flagged point was established (action 11).

After this, a baseline scanning took place ("I am going to scroll down to see if that exists in other baselines"; actions 12 to 24). While analysing this, another parameter was recognised as a source of undesired data ("from here I can see that in the Pi antenna appears the same issue for all its baselines so I can think that this antenna is causing the RFI... also I can identify that there is not affected in all the fields I think is only in in spw 1 and 4 I am going write down"; action 25). This analysis shows that the canvas design provides a solution in terms not only of showing the region of undesired data but also helps to identify specific parameters that guide the process of manual flagging.

After this, and also using P103 the analyst decides to scan vertically the canvas to see the different contrasts according to the different colourizing parameters (actions 26 to 39). Specifically in the polarisation colourised plots she commented "this is very interesting, because it will indicates me if some of the things that need to be flagged are happening only in one polarisation... it also tells me when I go on CASA which polarisation i have to check, so that's good". Notes were taken and the fourth flagged point was established (action 40).

After that some analysis was performed and the decision was to look for the next canvas (actions 41 to 44, Table 5.7). P104 was deployed (action 45; "this is amplitude vs frequency in all the ways that you can image, this is good". From scanning canvas P104 vertically, she corroborates what was deduced before (actions 46 to 50, Figure 5.12). Then canvas P105 was deployed (action 51) in order to perform a deeper inspection (actions 52 to 61; "this is by field which is a good thing to look in"). This inspection was done in a specific field of observation

("I want to see what is happening in this field"). After this notes were taken and the fifth flagged point was identified (action 62).

The final step was to establish the specific regions in CASA to verify the selected flagging regions. After selecting one of the flagging regions, the analyst created a plot using the same parameters used in an specific canvas plot (actions 63 to 65). Then she zoomed in to specific regions in order to corroborate her notes ("this is interesting") while contrasting CASA plots with Powerwall canvases ("that looks exactly the same... it shows how the different averages provided in the Powerwall really can show you things that you are missing... because to do it manually you probably won't do that"). Also, in order to corroborate what was found, she analysed the response of the identified pattern in all the baselines for the same field by horizontal scans. Then she said "I will go vertically to compare this situation in all the fields... now I can see that this pattern is happening in all the fields so that is definitely something you want to get rid off". These corroborations were performed using both CASA and the Powerwall canvases (actions 66 to 70) "I am seeing evidence in CASA of things seen in the Powerwall which is good". After this and once the specific regions were established, she called the autologger in order to deploy the specifications of the region and identify the characteristics of the segment of data (actions 71). Then the script was written (action 72) and the flagging done.

Finally, she called an amplitude vs frequency plot to verify the flagging script results (actions 73 to 81). This corroborates that the flagging was adequately identified and performed (action 82).

5.3.3 Results

The results obtained are not conclusive taking into account that the files used in the evaluation sessions differ in the stage of flagging. By not having a like-for-like comparison it was not possible to conclude the effectiveness of the manual flagging solution. Nevertheless the results presented serve as a starting point to measure the effectiveness of our project. In order to be able to interpret the results of the evaluation sessions two main processes were used to segment the information. One involved segmentation of data segmentation according to the required actions to establish a flagging point. The other process consisted in the writing and running of the flagging scripts and the verification of the results when at the time of execution.

The results of the evaluation sessions are presented in Tables 5.2, 5.3, 5.4 and 5.5 for the session number one and in Tables 5.6 and 5.7 for the session number two. The information is organized according to actions performed by the analyst and the time required for each action. At the end of the table the overall time is calculated.

Considering the flagging identification process, it is noticeable that the process as it is actually done demands a lot of actions. These actions are done with the purpose of performing zoom of specific chunks of data. Also it is notable that this identification of flagging points takes between 5 minutes 54 seconds to 13 minutes 51 seconds. Also, it was observed that the second flagging (6 minutes 44 seconds) took the greatest number of actions. On the other hand, using the Powerwall considerably requires less actions and therefore less time. The flagging took between 1 minute 18 seconds to 6 minutes 23 seconds. Also, fewer actions were required to identify each flagging region. By analysing Figure 5.6 it can be seen that the progression in time is more linear in the Powerwall solution. It is clear that the manual flagging process can not be completely characterised in terms of defining an absolute cycle of steps mainly due to the different approaches, methods and preferences of each analyst.

At this point, the initial results validate the effectiveness of the solution. This validation is supported in finding that the actions and time required to identify the first 5 flagged points were considerably lower than with the traditional method.

Now by analysing the script writing and verification process, the results shows that more time was required by using the Powerwall solution than without it. This verification requires producing specific plots in CASA, in order to zoom in to particular regions. One explanation for that increase in time using the Powerwall solution is that the analyst requires to make from scratch plots which were already present in the Powerwall.

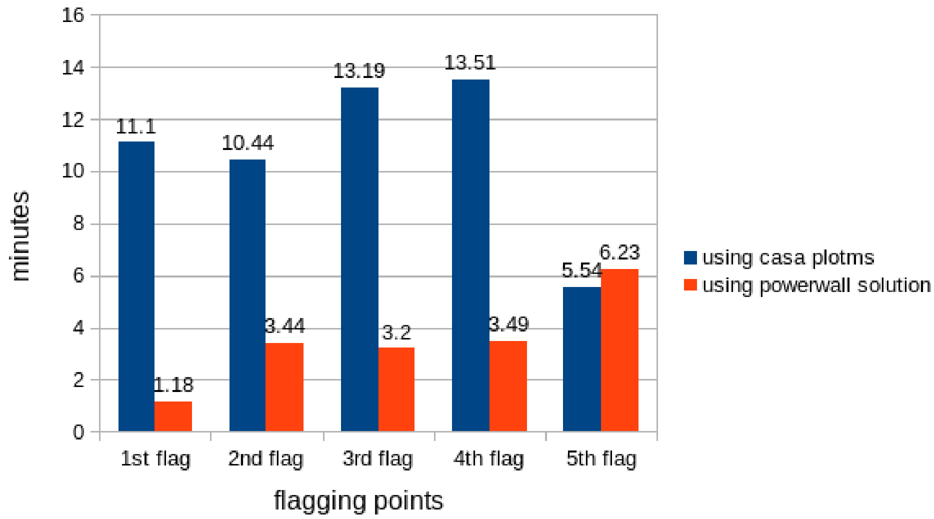


Figure 5.6: Required time to perform each flagging in sessions 1 and 2.

5.4 Discussion

In terms of the benefits of the solution, it was clear that having pre-generated plots is a real advantage in order to reduce time with respect to the time required in CASA. However, when using CASA, it is possible to have more interactivity in terms of producing zooms of a specific region. This factor was highlighted and recommended for future versions of this solution. One possibility is to be able to call individual high resolution plots allowing a more detailed inspection. Although with our solution it is possible to make zooms of specific regions, the resolution is not enough compared to what can be achieved in CASA, in order to achieved an advanced degree of LOD. There are two ways to achieve the same degree of resolution obtained using CASA. The first one is to integrate an interactive terminal linked to CASA. By doing this the analyst can generate specific plots of interest obtaining a level of detail and resolution as high as CASA can provide. The second option is to implement plots for the canvases with higher resolution. The constraints for the first scenario imply a decrease in the amount of space available to display the strips of plots. The second scenario would require re-design of the canvases. The resolutions used to implement the strips of plots were calculated to maximize the powerwall characteristics in terms of the overall

	#	DESCRIPTION	ACTIONS: DRAG	CLICK	TYPE	ANALYSIS	NOTES	NUMBER OF CLICKS	NUMBER OF CHAR.	TIME (sec)	TOTAL TIME (sec)	TOTAL TIME (min)
1ST FLAG	1	Listobs			X				26	10	661	11'1"
	2	Analysis of the listobs in casa terminal				X				300		
	3	Plot amp vs time			X				28	42		
	4	Plot ms process								93		
	5	Plot amp vs time(col, fld, avg in time and channel)			X				77	66		
	6	Plot ms process								27		
	7	Zoom in	X	X				1		6		
	8	Analysis of the plot				X				19		
	9	Zoom out original	X	X				1		3		
	10	Zoom in	X	X				1		6		
	11	Zoom in	X	X				1		6		
	12	Analysis of the plot				X				40		
	13	Zoom in	X	X				1		6		
	14	Notes					X			37		
2ND FLAG	15	Identification of characteristics of the data				X				53	644	10'44"
	16	Call a file to copy attributes	X	X				3		10		
	17	Analysis of the plot				X				80		
	18	Zoom out (clear region)	X	X				1		6		
	19	Analysis of the plot				X				7		
	20	Zoom in	X	X				1		6		
	21	Zoom out (clear region)	X	X				1		6		
	22	Analysis of the plot				X				8		
	23	Zoom in (to the next source)	X	X				1		6		
	24	Zoom in specific antenna	X	X				1		6		
	25	Zoom in chunk	X	X				1		6		
	26	Zoom out	X	X				1		6		
	27	Zoom in (to the next chunk)	X	X				1		6		
	28	Analysis of the plot				X				31		
	29	Zoom out	X	X				1		6		
	30	Zoom in specific antenna	X	X				1		6		
	31	Zoom out	X	X				1		6		
	32	Zoom in	X	X				1		6		
	33	Analysis of the plot				X				15		
	34	Zoom out	X	X				1		6		
	35	Zoom out initial plot	X	X				1		6		
	36	Zoom in chunks by time	X	X				1		6		
	37	Analysis of the plot				X				36		
	38	Zoom in chunk 5 sec for quak	X	X				1		6		
	39	Zoom out	X	X				1		6		
	40	Zoom in chunk	X	X				1		6		
	41	Zoom in cjunk	X	X				1		6		
	42	Zoom out	X	X				1		6		
	43	Zoom out initial plot	X	X				1		6		
	44	Zoom in chunk	X	X				1		6		
	45	Analysis of the plot				X				34		
	46	Zoom out initial plot	X	X				1		6		
	47	Zoom in chunk (stablished time)	X	X				1		6		
	48	Zoom out initial plot	X	X				1		6		
	49	Zoom in chunk	X	X				1		6		
	50	Zoom in next chunk	X	X				1		6		
	51	Analysis of the plot				X				56		
	52	Zoom in specific antenna	X	X				1		6		
	53	Analysis of the plot				X				52		
	54	Zoom out	X	X				1		6		
	55	Zoom out initial plot	X	X				1		6		
	56	Next chunk	X	X				1		6		
	57	Zoom in	X	X				1		6		
	58	Notes					X			70		

Table 5.2: Evaluation session 1 results in terms of actions and times (actions 1 to 58).

	#	DESCRIPTION	ACTIONS: DRAG	CLICK	TYPE	ANALYSIS	NOTES	NUMBER OF CLICKS	NUMBER OF CHAR.	TIME (sec)	TOTAL TIME (sec)	TOTAL TIME (min)
3 RD FLAG	59	Analysis of the plot				X				52	789	13'19"
	60	Listobs (check fields of observation)	X	X				2		5		
	61	Analysis of the plot				X				48		
	62	Zoom out	X	X				1		6		
	63	Zoom in specific chunk	X	X				1		6		
	64	Zoom out	X	X				1		6		
	65	Zoom out initial plot	X	X				1		6		
	66	Zoom in	X	X				1		6		
	67	Analysis of the plot				X				56		
	68	Select locate it log	X	X				2		13		
	69	Zoom out initial plot	X	X				1		6		
	70	Zoom in next source	X	X				1		6		
	71	Notes					X			138		
	72	Plot colorized by antenna			X				63	90		
	73	Zoom in	X	X				1		6		
	74	Plot antenna1 field 1 colorized spw	X	X				4		29		
	75	Analysis of the plot				X				78		
	76	Plot antenna2 field 1 colorized spw	X	X				4		27		
	77	Zoom in	X	X				1		6		
	78	Analzye				X				83		
	79	Zoom in	X	X				1		6		
	80	Notes					X			110		
4 TH FLAG	81	Plot rr ll polarizations	X	X				4		27	831	13'51"
	82	Zoom in initial chunk	X	X				1		6		
	83	Zoom out initial plot	X	X				1		6		
	84	Zoom in	X	X				1		6		
	85	Analysis of the plot				X				61		
	86	Zoom in source 3	X	X				1		6		
	87	Zoom in chunk	X	X				1		6		
	88	Analysis of the plot				X				63		
	89	Zoom in chunk	X	X				1		6		
	90	Zoom in chunk	X	X				1		6		
	91	Zoom out initial plot	X	X				1		6		
	92	Zoom in another source	X	X				1		6		
	93	Analysis of the plot				X				82		
	94	Notes					X			150		
	95	Casa terminal (check commands)			X				7	60		
	96	Notes					X			180		
	97	Plot initial no averaging			X				2	38		
	98	Analysis of the plot				X				26		
	99	Zoom in chunk	X	X				1		6		
	100	Zoom in chunk	X	X				1		6		
	101	Zoom out initial plot	X	X				1		6		
	102	Zoom in chunk	X	X				1		6		
	103	Zoom in chunk	X	X				1		6		
	104	Zoom in chunk	X	X				1		6		
	105	Zoom in chunk	X	X				1		6		
	106	Zoom out	X	X				1		6		
	107	Notes (checkin times for quacking)					X			42		

Table 5.3: Evaluation session 1 results in terms of actions and times (actions 59 to 107).

	#	DESCRIPTION	ACTIONS DRAG	CLICK	TYPE	ANALYSIS	NOTES	NUMBER OF CLICKS	NUMBER OF CHAR.	TIME (sec)	TOTAL TIME (sec)	TOTAL TIME (min)
5TH FLAG	108	Zoom out	X	X				1		6	354	5'54"
	109	Zoom out initial plot	X	X				1		6		
	110	Zoom in	X	X				1		6		
	111	Zoom in	X	X				1		6		
	112	Zoom in	X	X				1		6		
	113	Zoom in	X	X				1		6		
	114	Zoom out initial plot	X	X				1		6		
	115	Zoom in	X	X				1		6		
	116	Zoom in	X	X				1		6		
	117	Zoom in	X	X				1		6		
	118	Zoom out	X	X				1		6		
	119	Zoom out	X	X				1		6		
	120	Zoom out initial plot	X	X				1		6		
	121	Zoom in	X	X				1		6		
	122	Zoom in	X	X				1		6		
	123	Analysis of the plot				X				33		
	124	Zoom out	X	X				1		6		
	125	Zoom out	X	X				1		6		
	126	Zoom in chunk	X	X				1		6		
	127	Zoom out	X	X				1		6		
	128	Zoom out	X	X				1		6		
	129	Zoom out initial plot (checking no time averaging)	X	X				1		6		
	130	Zoom out	X	X				1		6		
	131	Zoom in	X	X				1		6		
	132	Zoom in	X	X				1		6		
	133	Zoom in	X	X				1		6		
	134	Zoom in	X	X				1		6		
	135	Notes					X			32		
	136	Zoom out initial plot	X	X				1		6		
	137	Zoom in final source	X	X				1		6		
	138	Plot colorized spw	X	X				4		21		
	139	Zoom in chunk	X	X				1		6		
	140	Zoom out	X	X				1		6		
	141	Zoom in chunk	X	X				1		6		
	142	Zoom in chunk	X	X				1		6		
	143	Zoom in chunk	X	X				1		6		
	144	Zoom in chunk	X	X				1		6		
	145	Zoom out	X	X				1		6		
	146	Zoom in chunk	X	X				1		6		
	147	Zoom in chunk	X	X				1		6		
	148	Zoom out	X	X				1		6		
	149	Zoom in chunk	X	X				1		6		
	150	Zoom out	X	X				1		6		
	151	Notes					X			28		

Table 5.4: Evaluation session 1 results in terms of actions and times (actions 108 to 151).

	#	DESCRIPTION	ACTIONS									
			DRAG	CLICK	TYPE	ANALYSIS	NOTES	NUMBER OF CLICKS	NUMBER OF CHAR.	TIME (sec)	TOTAL TIME (sec)	TOTAL TIME (min)
SCRIPTS AND VERIFICATION	152	Building and running flagging scripts			X				344	435	1365	22'45"
	153	Zoom out initial plot plot (check flaffing)	X	X				2		95		
	154	Zoom in chunk	X	X				1		6		
	155	Zoom in chunk	X	X				1		6		
	156	Zoom in region and call logger	X	X				3		21		
	157	Zoom out initial plot	X	X				1		6		
	158	Zoom in chunk	X	X				1		6		
	159	Analysis of the plot				X				34		
	160	Zoom in chunk	X	X				1		6		
	161	Modify quaking script			X				93	150		
	162	Run script								15		
	163	Plot data again	X	X				1		30		
	164	Zoom in chunk	X	X				1		6		
	165	Zoom in	X	X				1		6		
	166	Analysis of the plot				X				63		
	167	Zoom out	X	X				1		6		
	168	Zoom in	X	X				1		6		
	169	Analysis of the plot				X				48		
	170	Zoom out	X	X				1		6		
	171	Zoom in chunk next source	X	X				1		6		
	172	Zoom in chunk	X	X				1		6		
	173	Analysis of the plot				X				82		
	174	Zoom out initial plot	X	X				1		6		
	175	Zoom in chunk	X	X				1		6		
	176	Analysis of the plot				X				63		
	177	Zoom out	X	X				1		6		
	178	Zoom in chunk	X	X				1		6		
	179	Analysis of the plot				X				87		
	180	Logger (check segment)	X	X				2		26		
	181	Listobs (check scan)			X				26	120		
		TOTAL ACTIONS		**20**	**3**	**6**	**0**	**24**	**463**			
									TOTAL TIME (seconds)	**1365**		
									TOTAL TIME (minutes)	**77'24"**		
									TOTAL TIME (hours)	**1hr 17min 24 sec**		

Table 5.5: Evaluation session 1 results in terms of actions and times (actions 152 to 181).

	# Act.	PW PC	An. PC	DESCRIPTION	ACTIONS: DRAG	CLICK	SCROLL	TYPE	ANALYSIS	NOTES	NUMBER OF CLICKS	NUMBER OF CHAR.	TIME (sec)	TOTAL TIME (sec)	TOTAL TIME (min)
1ST FLAG	1	X		Canvas 101		X					1		5	78	1'18"
	2	X		Zoom in			X	X				1	3		
	3	X		Scan horizontally (baselines)			X						50		
	4	X		Notes (spw3						X			20		
2ND FLAG	5	X		Analysis					X				30	224	3'44"
	6	X		Canvas 103		X							5		
	7	X		Scan horizontally to find the desired			X	X				1	18		
	8	X		Analysis									75		
	9	X		Scan vertically to the top			X	X				1	6		
	10	X		Analysis					X				50		
	11	X		Notes						X			40		
3RD FLAG	12	X		Scan vertically (verify other baselines)			X	X				1	5	200	3'20"
	13	X		Analysis									20		
	14	X		Scan vertically (verify other baselines)			X	X				1	5		
	15	X		Analysis					X				15		
	16	X		Scan vertically (verify other baselines)			X	X				1	5		
	17	X		Analysis					X				23		
	18		X	Notes						X			14		
	19	X		Scan vertically (verify specific antenna behaviour)			X	X				1	7		
	20	X		Analysis					X				20		
	21	X		Scan vertically (verify specific antenna behaviour)			X	X				1	7		
	22	X		Analysis					X				25		
	23	X		Scan vertically (verify specific antenna behaviour)			X	X				1	7		
	24	X		Analysis					X				23		
	25		X	Notes						X			24		
4TH FLAG	26	X		Analysis					X				25	229	3'49"
	27	X		Scan vertically (chackin)			X	X				1	5		
	28	X		Analysis					X				35		
	29	X		Scan vertically (chackin)			X	X				1	5		
	30		X	Notes						X			9		
	31	X		Scan vertically (chackin)			X	X				1	10		
	32	X		Analyze					X				20		
	33	X		Reset image				X				3	2		
	34	X		Zoom in		X	X				1		4		
	35	X		Analyze									10		
	36	X		Scan horizontally			X						35		
	37	X		Scan vertically (chackin)			X	X				1	20		
	38	X		Zoom in		X	X				1		5		
	39	X		Analyze					X				29		
	40		X	Notes						X			15		

Table 5.6: Evaluation session 2 results in terms of actions and times (actions 1 to 40).

					ACTIONS										
	# Act.	PW PC	An. PC	DESCRIPTION	DRAG	CLICK	SCROLL	TYPE	ANALYSIS	NOTES	NUMBER OF CLICKS	NUMBER OF CHAR.	TIME (sec)	TOTAL TIME (sec)	TOTAL TIME (min)
5TH FLAG	41	X		Reset image				X				3	4	383	6’23”
	42	X		Analysis					X				40		
	43	X		Zoom in		X	X				1		4		
	44	X		Analysis					X				20		
	45	X		Canvas 104		X					1		3		
	46	X		Analyze									35		
	47	X		Zoom in		X	X				1		6		
	48	X		Analysis					X				9		
	49	X		Scan vertically			X	X				1	8		
	50	X		Analysis					X				25		
	51	X		Canvas p5		X					1		4		
	52	X		Zoom in		X	X				1		3		
	53	X		Analysis					X				16		
	54	X		Scan horizontally			X						14		
	55	X		Analysis					X				15		
	56	X		Scan vertically (analyze this particular field)			X	X				1	25		
	57	X		Analysis									10		
	58	X		Scan vertically (analyze this particular field)			X	X				1	5		
	59	X		Analysis					X				17		
	60	X		Scan horizontally (see what is happening in other fields)			X						51		
	61	X		Analysis					X				37		
	62	X		Final notes						X			32		
SCRIPTS AND VERIFICATION	63		X	Open casa				X				4	15	741	12’21”
	64		X	Plotms frrquency last spectral window color field		X					5		70		
	65		X	Plotting		X					1		20		
	66	X		Call p5		X					1		5		
	67	X		Zoom in		X	X				1		3		
	68	X		Scan vertically			X	X				1	8		
	69	X	X	Analysis					X				65		
	70		X	Zoom in	X	X	X				1		13		
	71		X	Zoom in region and call logger		X					1		8		
	72		X	Create flaggin command				X				83	60		
	73		X	Zoom out			X	X					3		
	74	X	X	Analysis	X	X							60		
	75		X	Plotms									10		
	76		X	Analysis change between baslines to corroborate		X							157		
	77		X	Plot ms (different field)		X					1		7		
	78		X	Analysis (different fields)					X				133		
	79		X	Run command in casa		X							35		
	80		X	Plotms amp freq		X					2		47		
	81	X	X	Final analysis					X				22		
				TOTAL ACTIONS	2	16	12	8	11	1	18	94			
				TOTAL (seconds)									1124		
				TOTAL (minutes)									30’55”		

Table 5.7: Evaluation session 2 results in terms of actions and times (actions 41 to 81).

size in pixels in function of showing all the baselines one next to another.

Also the analyst found that the different canvas structures in terms of how is the information arranged were adequate to reduce the time to have a general perspective of the database. These configurations also allows the user to establish the origin of the data in terms of parameters such as antenna ID, channels, polarisations, baselines or fields of observation. This can be assumed as a benefit in terms of having a visualisation aid that can be used both to have a big picture of the data in terms of recognition of undesired patterns, and also in terms of helping to guide the process of the identification of specific parameters that produce a signal response.

The canvas configurations allows a proper accessibility in terms of allowing to see the plots very quickly. Also just by scrolling through the canvas quickly can be identified patterns where if using CASA the only way to do that is to do the iteration for each baseline, field or spectral window. Although it is possible to do panels in CASA ("they are very clunky and difficult to produce... also they can not be looked at the same time as it can be done by using the canvas in the Powerwall"). Another factor is that is easier to compare different parameters settings and to see patterns by using a Powerwall, otherwise the analyst must remember those patterns. Considering cognitive loads, the use of canvases decreases the amount of information that has to be memorised, although it may demands more effort in terms of analysing more information. In terms of the analyst perspective of the 12 canvases some comments were concluded. For the P1, P2, P3, and P4 canvases, the amplitude vs time and frequency strips were found useful, however phase vs time and frequency not that much because the plots overlap all the fields making hard to see any patterns on them.

From P5 to P12, overlapping fields were no longer an issue because in those canvases the fields were presented separately. P5 and P6 were found very useful in terms of seeing easily particular RFI problems or bad data. Between them P6 was found more adequate arranged in terms of the orientation of the strips because it was easier to scan across side to side; having similar plots going horizontally makes easier to compare them. By using P7 and P8 the analyst identified patterns in frequency (phase vs frequency). In those patterns the analyst found deviations which helped in the process of establishing regions of interest in which to perform flagging. Between P7 and P8 the analyst prefers P7.

At this point some interesting issues arise. The orientation of the strips may serve better depending on the type of parameters of the plots. For amplitude vs frequency it was preferred a horizontal strip orientation but for the phase vs frequency a vertical one. It was commented that it will be adequate to still produce both orientations allowing the analyst being able to select one or another according to its preferences and the parameters evaluated in that particular moment.

Although P9, P10, P11, and P12 were found valuable in order to have the possibility to have the data arranged in different configurations to confirm or see new patterns, they were considered not so useful because of the existence of the previous canvas. This argument can be interpreted in two ways. On one hand the information presented in the first 8 canvases were enough in order to accomplish the purposes of the solution. On the other hand the perspective of the analyst before performing the implementation is crucial in order to obtain a tailor-made environment. This will increase the effectiveness of the solution both in time and interaction.

Regarding the new canvases made in accordance to the analyst preferences, the response was positive. P101 and P102 use the same structure of P1 and P2, but removing the phase vs time and frequency strips. This was done because for those specific configurations the overlapping fields did not allow to see any RFI patterns. P103 and P104 shows amplitude vs time and frequency colourised by spectral window, field, polarisation and antenna1. By having different colourised parameters it is possible to evaluate specific origins or which dimensions of the data need to be flagged. For instance in the polarisation colourisation it is possible to identify in which specific polarisation is the undesired data.

Regarding how the analyst responds to the handling and manipulation of the visualisation interface, some aspects were identified. From the analyst perspective, at the beginning it was hard to know the commands and the key strokes that are needed to go through the data smoothly. Although one recommendation was the use of a touch screen in order to make the interaction easier and allowing the analyst just to worry about the analysis, during the session she made significant progress, and at the end she knew how to manipulate the visualisation interface.

Also it was mentioned that the Powerwall allows collaborative work in terms of two or more astronomers interacting and discussing specific regions of data. The solution was found valuable for pattern recognition in terms of a realistic experience by having both a complete and detailed perspective of datasets of interferometric data.

Chapter 6

Conclusions

Regarding the initial objective of this project, which was investigating how to design a solution that optimises the process of manual flagging for radio astronomy data, the result was satisfactory. This statement is supported by the initial results that were found, which show that the solution is useful to reduce time in identifying points to flag. The solution allows the analyst to see patterns and make comparisons that otherwise will be hard to see or too complex to achieve if plots have to be deployed one by one.

Once multiple parameters were identified it was clear that this project required the solution to be based on an adaptable environment in function of the analyst requirements. Also by the interaction of the analysts it was perceived from the beginning that manual flagging is a process defined by the analyst who performs it. Two astronomers participated during this process. Just by having these perspectives it was recognised that the procedures and methods that are needed to perform the manual flagging process were considerably different. Even so, this perspective allows this project to be stronger in terms of adaptability, because the design stage was based on these criteria. The solution must respond not only to be effective in showing specific visualisation structures and configurations, but also to the different requirements of end users. This is how the software architecture was designed and implemented.

In terms of the initial evaluation criteria, the result of the evaluation is satisfactory. Nevertheless, there is no evidence that two data sets of different stages of flagging are or are not equivalent in complexity for manual flagging purposes.

As it was mentioned in Section 5.3.3 the interpretation of the evaluation can not generate conclusive results. One hypothesis is that the further that the data sets go through flagging stages, the flagging areas will be each time more difficult to find. But this is only a hypothesis. Comparissons of the response of analysts using data sets in the same stage of flagging are required to obtain final conclusions of our solution against the traditional manual flagging procedure. In any case, and taking as reference what was obtained, some aspects can be inferred. Regarding the time required to produce the plots, the conclusion is that is possible to considerably reduce the time by using modern hardware. To produce the 1680 plots for the 12 canvases, the results showed that with a computer using a core i3 processor the amount of time was 5 hours 5 minutes, and with a computer with a processor 5 generations newer the time was 2 hours 34 minutes. This is a reduction of time of 2 hours 31 minutes. Typically, an analyst spends from 30 seconds to 2 minutes to produce just one plot. From here it can be assumed that this pre-plotting process is adequate to reduce the time required to obtain the plots.

Another important point to take into account is that with the proposed solution, the interaction to produce the plots is minimal. In order to produce the plots the analyst just has to execute a file and all the plots will automatically be produced. This is interpreted as another benefit in terms of minimizing the interaction of the user in order to produce the desired plots. The learning curve in order to manipulate the interface was satisfactory. The analyst was able to understand, learn and manipulate the commands and key strokes to go along the different canvases in order to analyse the plots. She was also able to perform horizontal and vertical scans rapidly. Regarding this point, it is worth mentioning that, although the vertical and horizontal orientations were designed to be used as different possibilities to see the information, the analyst found it very useful to do the scans in specific canvases.

In terms of the accuracy of the process, it was observed that the canvas design is not only a way to deploy a large amount of information, but also allows analysts to identify previous unseen undesired patterns. While comparing CASA plots with a canvas, the analyst commented that it was interesting to see new patterns that would be difficult to observe just by plotting using CASA in a standard environment.These patterns were recognizable due to the possibility of having a complete perspective of the behavior of the signals by showing all the baselines one next to another. That aspect allows to perform scans of specific signal tendencies and also to make comparisons in time, frequency and amplitude that by using single plots would not be feasible. To recognise such data it would take a considerable amount of time when going plot by plot in CASA. Also, the analyst mentioned that before she had to rely on her memory in order to remember patterns but by having the information as it is in the canvas this is not required anymore. The result of visualizing different averages and different baselines and to have accessibility to this information immediately results in the identification of previously unobserved patterns when using the traditional method, and shows to be a useful tool in corroborating important regions to be flagged.

In the process of developing the evaluation sessions it was decided to test also the scalability of the solution. In order to perform the second session using the Powerwall, 6 new canvases were produced in response to the analyst preferences. The structure of the solution's scripts allows the analyst to obtain the new canvases just by adjusting the plot.py script and the merging.py scripts. This shows that the proposed design is pertinent in terms of being rapidly configured and adapted to the users preferences and requirements.

6.1 Future work

In terms of how the visualisation is accomplished, it is possible to establish algorithms that calculate and resize the plots in order for the plots not get cut by the screens edges of the Powerwall. From the experience of this project, it is valuable to think of applications made specifically for the Powerwall. This is in terms of the interfaces that control the deployment of images and in terms also of facilitating interoperability of any given application.

Regarding producing new plots, the design of an application which allows the user to select the variables and parameters for the canvases would speed the process. That application can be linked to the plotting and merging script in order to automatically produce the new canvases.

Having high resolution images would reduce the number of plots required to make the manual flagging. As it was mentioned in the evaluation chapter, to obtain a high level of detail of a particular chunk of data requires an analyst to perform several zooms in specific regions. An option is to have an external file with pre-generated high resolution images. Also by having different pre-generated plots with different levels of zoom and by chunks of data, the amount of individual plots produced in CASA will considerably decrease in number.

Deploying this solution in a mid-range multi-processor workstation would also significantly reduce computing time, as the plotting commands are readily parallelisable. Thus, pairing a Powerwall with a workstation or an HPC machine node would greatly increase the efficiency of the plot pre-generation process.

Another important improvement to the solution is to produce plots with different time divisions. This aspect will benefit the identification of the source of an undesired signal pattern.

6.2 Final comments

The development of this solution corroborates the enormous potential of Powerwalls in terms of the many applications that can be designed and implemented. As was seen, the possibility to use a high resolution visualisation environment allows analysts to re think how certain processes can be performed. Although the details of how the visualisation configurations need to be accomplished depends on the specific applications, there are common aspects that can be addressed as Powerwall protocols. For example, managing how images are parsed to the visualisation tool requires similar dedicated end user interfaces. In the context of research activities requiring large-scale visualisation systems, having a multi-dimensional approach to data visualisation enhances the possibility of reducing time and increasing the accuracy for specific processes. This solution will provide an initial approach for how manual flagging can be managed by using a Powerwall in order to optimise the process. As was seen, there are some benefits like the reduction of time in the identification of flagging segments, and aspects to be improved like the user interface. Nevertheless the solution was satisfactory in terms of the initial intent.

Appendix A

User's manual

The following manual provides a detailed explanation of the process to generate new canvas by the use of the python scripts and the procedure with the interface visualisation software.

POWERWALL-BASED SOLUTION FOR THE MANUAL FLAGGING OF RADIO ASTRONOMY DATA USER MANUAL

This manual provides the steps to produce new canvas configurations according to the user's preferences using the python scripts developed for the plotting and merging of strips.

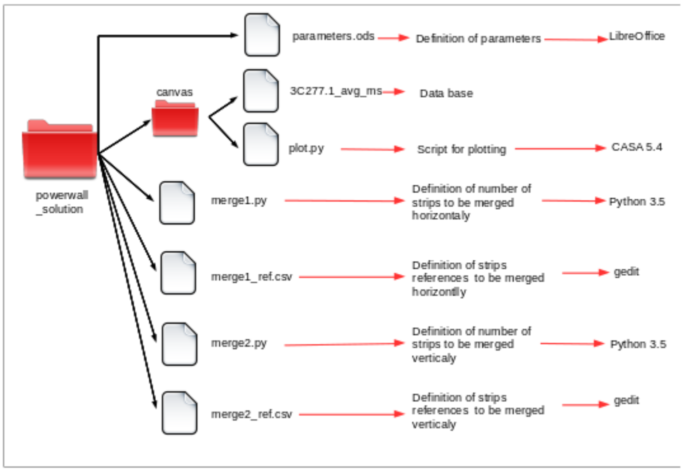


Image 1 Software architecture of the powerwall solution

From a general perspective the steps are:

1) Define the parameters for the canvas using the template located in parameters.ods. as a guide.

2) Load the .ms data set file.

2) Modify the plot.py script in function of the parameters assigned in parameters.ods.

3) Modify the merging scripts in order to specify the amount of plots and references per canvas for the horizontal and vertical orientations; merge1.py and merge1_csv for horizontal merging, and merge2.py and merge2_csv for vertical merging.

4) Execute plot.py from casa 5.4.1.

5) Execute merge1.py and merge2.py. using python 3.5.

6) Load the resulted canvas to the visualization interface.

A description of the parameters template, plotting and merging scripts and the integration to the visualization interface is provided as follows in order to update the information according to the the user preferences:

Figure A.1: User's manual.

Parameters template

plot ref.	size	matrix	coloraxis		# plots	X axis	y axis	time vs amplitude	time vs phase	frequency vs amplitude	frequency vs phase		iteraxis	coloraxis	avgchannel	avgtime	spw	corr	field

Table 1 Parameter configuration template

The parameters template intends to provide a guide in order to establish tailor made canvas.
In the Table 1 are shown the parameter template to guide the process of definition of variables and canvas configurations (plot reference, size of the matrix, coloraxis, number of plots per strip, x an y axis assignation, iteraxis, averages, spectral window, correlation and field)

Plotting script

Once the parameters were established the next step is to update the plot.py script. This script will automatically produce the strips for the canvas, both horizontally and vertically oriented.

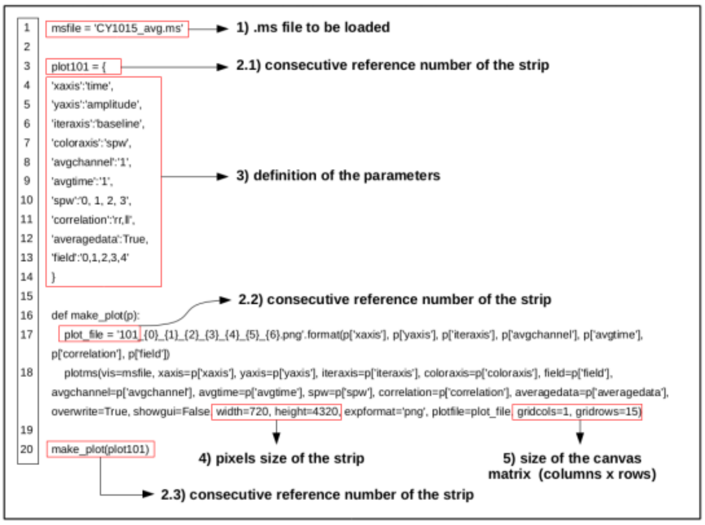


Image 2 plot.py script description

Figure A.2: User's manual.

1) .ms file to be loaded: in the line #1 is defined the .ms file which will be used.

2.1), 2.2) and 2.3) consecutive reference number of the strip: in lines #3, #17 and #20 has to be specified the consecutive number of the strip.

3) definition of parameters: in lines #4 to #13 are defined the values for the parameters to be plotted. Those parameters are x axis, yaxis, iteraxis, coloraxis, avgchannel, avgtime, spectral window, correlation and field of observation.

4) pixels size for the strips: in this part of line #18 is specified the size of the strip in terms of weight and height.

5) size of the canvas matrix (columns x rows): here is defined the matrix arrange in terms of columns and rows to produce each strip.

The 20 lines showed before defines one strip. When produced the required strips the next step is to execute the file using casa 5.4.1. From here the strips will be automatically stored in the folder which contains the scripts.

Image3 Strip 101 as produced from casa by the use of the plot.py script

Merging script

```
#!/usr/bin/env python
# create_destroy_toplevel.py

import numpy as np
import glob
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import pandas as pd
path = 'mock_test/'
files = np.array([f for f in glob.glob(path + "*.png", recursive=True)])
df=pd.read_csv("custom_screen_guide1.csv")
CSlist=np.unique(df.CustomScr)
for CS in CSlist:
    imlist=df.PlotRef[df.CustomScr==CS]
    plfiles=[]
    for i in imlist:
        plfiles+=[files[np.array([str(i) in file for file in files])][0]]
    if len(imlist)==16:
        im1=np.hstack((mpimg.imread(plfiles[0]), mpimg.imread(plfiles[1]), mpimg.imread(plfiles[2]), mpimg.imread(plfiles[3]),
mpimg.imread(plfiles[4]), mpimg.imread(plfiles[5]), mpimg.imread(plfiles[6]), mpimg.imread(plfiles[7])))
        im2=np.hstack((mpimg.imread(plfiles[8]), mpimg.imread(plfiles[9]),mpimg.imread(plfiles[10]),mpimg. imread(plfiles[11]),
mpimg. imread(plfiles[12]), mpimg.imread(plfiles[13]), mpimg.imread(plfiles[14]), mpimg.imread(plfiles[15])))
        imag=np.hstack((im1,im2))
        plt.imsave(path+CS+'.png',imag,format='png')
```

1) folder that contains the .py scripts

2) reference merging .csv file

3) number of strips per canvas

4) assignation of parameters according to the number of strips

Image 4 merge1_ref.py script description

Figure A.3: User's manual.

```
CustomScr,PlotRef

P1,101
P1,102
P1,103
P1,104
P1,105
P1,106
P1,107
P1,108
P1,109
P1,110
P1,111
P1,112
P1,113
P1,114
P1,115
```

Image 5 merge1_ref.csv script

1) folder that contains the .py script: in line #9 is defined the folder in which is the plotting script.
2) reference merging.csv files: in line #11 is specified the .csv which defines the strips to be merged for each canvas. Image 5 provides an example of the structure of how has to be defined the strips order for each canvas. In this example is stipulated that canvas P1 will be arranged vertically using the sixteen strips from the references 101 to115.
3) number of strips per canvas: in line # 18 is stipulated the number of strips to be merged for each canvas according to the .csv file.
4) assignation of parameters according to the number of strips: in function of the number of strips of canvas the same amount of strips require to be assigned.
Note: this procedure requires to be done in merge1.py to define the horizontal merging and in merge2.py for the vertical merging.

Once the merging scripts are ready the next step is to execute the merging scripts1.py and merging script2.py in order to obtain the final .jpg images of each canvas. The result of merging the 16 strips for the P1 canvas is shown in Image 6.

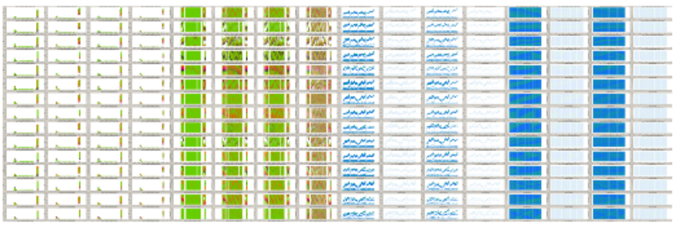


Image 6 P1 canvas as merging the 16 strips

Figure A.4: User's manual.

Integration with the powerwall visualization interfase

Finally by having the each canvas as an image, they can be called from a visualization software. If using GIMP GNU the .png files requires to be invoke as layers. By doing that all canvas will be able to be called from the layer terminal as required by the analyst.

Invoke the png. files as layers. To do this open GIMP GNU and click /file/Open as Layers/ and select the png. files in the canvas folder.

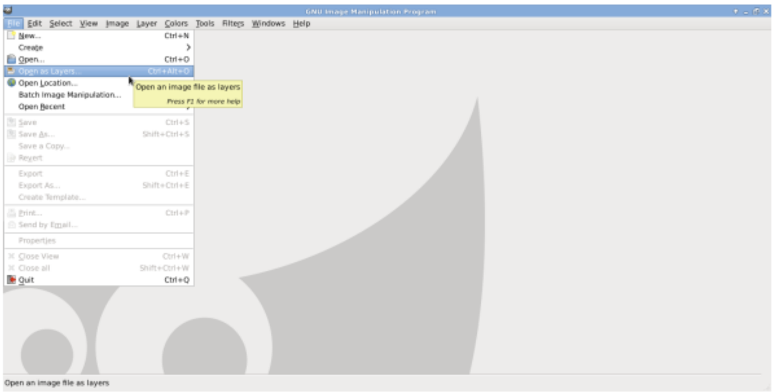


Image 7 Invoking the png. files in order to be managed as layers

The images will appear as shown.

Image 8 P1 canvas .png file invoked as a layer

Figure A.5: User's manual.

Save the file and then call it from the powerwall using also GIMP GNU.

Once the file is open click in /Windows/Dockable Dialogs/Layers.

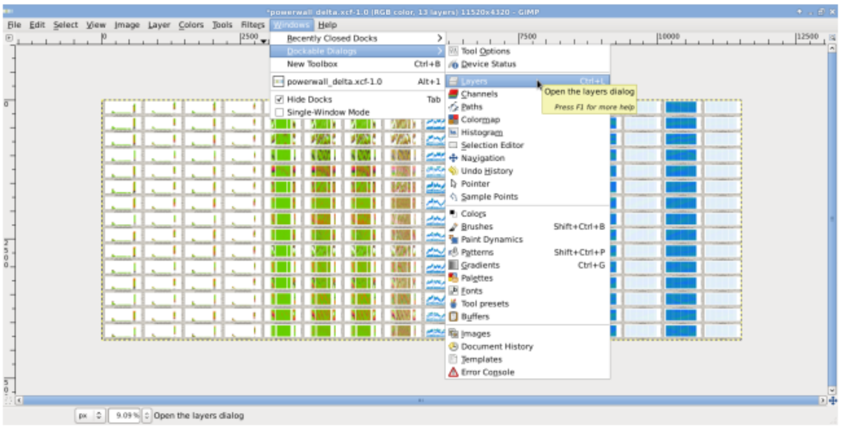


Image 9 Calling of the visualization layer window

By doing this a layer window will be deployed. This window allows to invoke each of the canvas by clicking in the desired one.

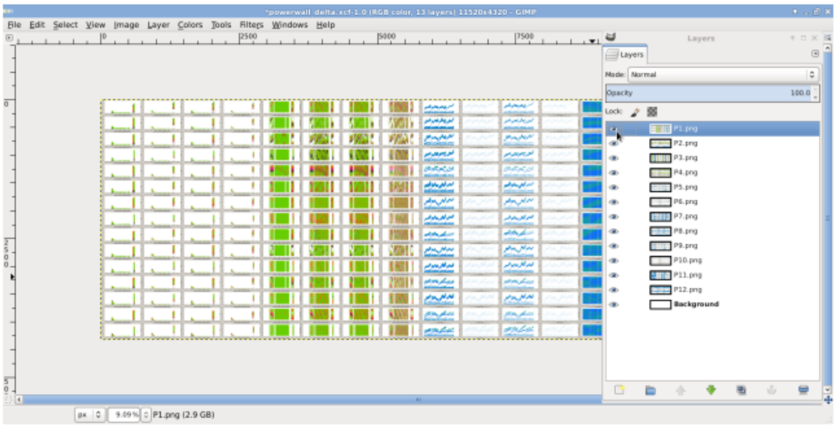


Image 10 Visualization layer window

Figure A.6: User's manual.

This is the final result of the process as seen in the powerwall.

Image 11 Powerwall visualization of the P1 canvas and visualization layer window to invoke each canvas

The control commands are:

- Scroll up : up horizontal scan.
- Scroll down: down horizontal scan.
- Shift / scroll up : left vertical scan.
- Shift / scroll down: right vertical scan.
- Control / scroll up: zoom in.
- Control / scroll down: zoom out.
- Ctrl / Shift / J: Zoom out to a complete canvas visualisation.

Figure A.7: User's manual.

Appendix B

Scripts

This appendix shows the plotting and merging scripts required to generate the canvas (plot.py, merge1.py, merge1 ref.csv).

Plt.py script (lines 1 to 21 and 2204 to 2221)

```
msfile = 'CY1015_avg.ms'

plot101 = {
'xaxis':'time',
'yaxis':'amplitude',
'iteraxis':'baseline',
'coloraxis':'spw',
'avgchannel':'1',
'avgtime':'1',
'spw':'0, 1, 2, 3',
'correlation':'rr,ll',
'averagedata':True,
'field':'0,1,2,3,4'
}

def make_plot(p):
    plot_file = '101_{0}_{1}_{2}_{3}_{4}_{5}_{6}.png'.format(p['xaxis'], p['yaxis'], p['iteraxis'],
p['avgchannel'], p['avgtime'], p['correlation'], p['field'])
    plotms(vis=msfile, xaxis=p['xaxis'], yaxis=p['yaxis'], iteraxis=p['iteraxis'], coloraxis=p['coloraxis'],
field=p['field'], avgchannel=p['avgchannel'], avgtime=p['avgtime'], spw=p['spw'],
correlation=p['correlation'], averagedata=p['averagedata'], overwrite=True, showgui=False, width=720,
height=4320, expformat='png', plotfile=plot_file, gridcols=1, gridrows=15)

make_plot(plot101)


plot440 = {
'xaxis':'frequency',
'yaxis':'phase',
'iteraxis':'baseline',
'coloraxis':'corr',
'avgchannel':'8',
'avgtime':'32',
'spw':'0, 1, 2, 3',
'correlation':'rr,ll',
'averagedata':True,
'field':'4'
}

def make_plot(p):
    plot_file = '440_{0}_{1}_{2}_{3}_{4}_{5}_{6}.png'.format(p['xaxis'], p['yaxis'], p['iteraxis'],
p['avgchannel'], p['avgtime'], p['correlation'], p['field'])
    plotms(vis=msfile, xaxis=p['xaxis'], yaxis=p['yaxis'], iteraxis=p['iteraxis'], coloraxis=p['coloraxis'],
field=p['field'], avgchannel=p['avgchannel'], avgtime=p['avgtime'], spw=p['spw'],
correlation=p['correlation'], averagedata=p['averagedata'], overwrite=True, showgui=False,
width=11520, height=216, expformat='png', plotfile=plot_file, gridcols=15, gridrows=1)

make_plot(plot440)
```

Figure B.1: Plotting script plot.py.

Plt.py merge1.py

```
#!/usr/bin/env python
# create_destroy_toplevel.py

import numpy as np
import glob
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import pandas as pd
path = 'canvas/'
files = np.array([f for f in glob.glob(path + "*.png", recursive=True)])
df=pd.read_csv("merge1_ref.csv")
CSlist=np.unique(df.CustomScr)
for CS in CSlist:
    imlist=df.PlotRef[df.CustomScr==CS]
    plfiles=[]
    for i in imlist:
        plfiles+=[files[np.array([str(i) in file for file in files])][0]]
    if len(imlist)==16:
        im1=np.hstack((mpimg.imread(plfiles[0]),mpimg.imread(plfiles
[1]),mpimg.imread(plfiles[2]),mpimg.imread(plfiles[3]),mpimg.imread(plfiles[4]),mpimg.imread(plfile
s[5]),mpimg.imread(plfiles[6]),mpimg.imread(plfiles[7])))

im2=np.hstack((mpimg.imread(plfiles[8]),mpimg.imread(plfiles[9]),mpimg.imread(plfiles[10]),mpimg.
imread(plfiles[11]),mpimg.imread(plfiles[12]),mpimg.imread(plfiles[13]),mpimg.imread(plfiles[14]),m
pimg.imread(plfiles[15])))
        imag=np.hstack((im1,im2))
        plt.imsave(path+CS+'.png',imag,format='png')

    if len(imlist)==20:

im1=np.hstack((mpimg.imread(plfiles[0]),mpimg.imread(plfiles[1]),mpimg.imread(plfiles[2]),mpimg.i
mread(plfiles[3]),mpimg.imread(plfiles[4]),mpimg.imread(plfiles[5]),mpimg.imread(plfiles[6]),mpimg.
imread(plfiles[7]),mpimg.imread(plfiles[8]),mpimg.imread(plfiles[9])))

im2=np.hstack((mpimg.imread(plfiles[10]),mpimg.imread(plfiles[11]),mpimg.imread(plfiles[12]),mpi
mg.imread(plfiles[13]),mpimg.imread(plfiles[14]),mpimg.imread(plfiles[15]),mpimg.imread(plfiles[16
]),mpimg.imread(plfiles[17]),mpimg.imread(plfiles[18]),mpimg.imread(plfiles[19])))
       imag=np.hstack((im1,im2))
       plt.imsave(path+CS+'.png',imag,format='png')
```

Figure B.2: Merging script merge1.py.

Plt.py merge2.py

```
#!/usr/bin/env python
# create_destroy_toplevel.py

import numpy as np
import glob
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import pandas as pd
path = 'canvas/'
files = np.array([f for f in glob.glob(path + "*.png", recursive=True)])
df=pd.read_csv("merge2_ref.csv")
CSlist=np.unique(df.CustomScr)
for CS in CSlist:
   imlist=df.PlotRef[df.CustomScr==CS]
   plfiles=[]
   for i in imlist:
      plfiles+=[files[np.array([str(i) in file for file in files])][0]]
   if len(imlist)==16:

im1=np.vstack((mpimg.imread(plfiles[0]),mpimg.imread(plfiles[1]),mpimg.imread(plfiles[2]),mpimg.i
mread(plfiles[3]),mpimg.imread(plfiles[4]),mpimg.imread(plfiles[5]),mpimg.imread(plfiles[6]),mpimg.
imread(plfiles[7])))

im2=np.vstack((mpimg.imread(plfiles[8]),mpimg.imread(plfiles[9]),mpimg.imread(plfiles[10]),mpimg.
imread(plfiles[11]),mpimg.imread(plfiles[12]),mpimg.imread(plfiles[13]),mpimg.imread(plfiles[14]),m
pimg.imread(plfiles[15])))
      imag=np.vstack((im1,im2))
      plt.imsave(path+CS+'.png',imag,format='png')

   if len(imlist)==20:

im1=np.vstack((mpimg.imread(plfiles[0]),mpimg.imread(plfiles[1]),mpimg.imread(plfiles[2]),mpimg.i
mread(plfiles[3]),mpimg.imread(plfiles[4]),mpimg.imread(plfiles[5]),mpimg.imread(plfiles[6]),mpimg.
imread(plfiles[7]),mpimg.imread(plfiles[8]),mpimg.imread(plfiles[9])))

im2=np.vstack((mpimg.imread(plfiles[10]),mpimg.imread(plfiles[11]),mpimg.imread(plfiles[12]),mpi
mg.imread(plfiles[13]),mpimg.imread(plfiles[14]),mpimg.imread(plfiles[15]),mpimg.imread(plfiles[16
]),mpimg.imread(plfiles[17]),mpimg.imread(plfiles[18]),mpimg.imread(plfiles[19])))
      imag=np.vstack((im1,im2))
      plt.imsave(path+CS+'.png',imag,format='png')
```

Figure B.3: Merging script merge2.py.

Plt.py merge2.csv

```
CustomScr,PlotRef

P1,101
P1,102
P1,103
P1,104
P1,105
P1,106
P1,107
P1,108
P1,109
P1,110
P1,111
P1,112
P1,113
P1,114
P1,115
P1,116
P3,101
P3,105
P3,109
P3,113
P3,102
P3,106
P3,110
P3,114
P3,103
P3,107
P3,111
P3,115
P3,104
P3,108
P3,112
P3,116
P5,301
P5,302
P5,303
P5,304
P5,309
P5,310
P5,311
P5,312
P5,317
P5,318
P5,319
P5,320
P5,325
P5,326
P5,327
P5,328
P5,333
P5,334
P5,335
P5,336
P7,305
P7,306
P7,307
P7,308
P7,313
P7,314
P7,315
P7,316
P7,321
P7,322
P7,323
P7,324
P7,329
P7,330
P7,331
P7,332
P7,337
P7,338
P7,339
P7,340
P9,301
P9,309
P9,317
P9,325
P9,333
P9,302
P9,310
P9,318
P9,326
P9,334
P9,303
P9,311
P9,319
P9,327
P9,335
P9,304
P9,312
P9,320
P9,328
P9,336
P11,305
P11,313
P11,321
P11,329
P11,337
P11,306
P11,314
P11,322
P11,330
P11,338
P11,307
P11,315
P11,323
P11,331
P11,339
P11,308
P11,316
P11,324
P11,332
P11,340
```

Figure B.4: Merging script merge1.csv.

Plt.py merge2.csv

```
CustomScr,PlotRef
P2,201
P2,202
P2,203
P2,204
P2,205
P2,206
P2,207
P2,208
P2,209
P2,210
P2,211
P2,212
P2,213
P2,214
P2,215
P2,216
P4,201
P4,205
P4,209
P4,213
P4,202
P4,206
P4,210
P4,214
P4,203
P4,207
P4,211
P4,215
P4,204
P4,208
P4,212
P4,216
P6,401
P6,402
P6,403
P6,404
P6,409
P6,410
P6,411
P6,412
P6,417
P6,418
P6,419
P6,420
P6,425
P6,426
P6,427
P6,428
P6,433
P6,434
P6,435
P6,436
P8,405
P8,406
P8,407
P8,408
P8,413
P8,414
P8,415
P8,416
P8,421
P8,422
P8,423
P8,424
P8,429
P8,430
P8,431
P8,432
P8,437
P8,438
P8,439
P8,440
P10,401
P10,409
P10,417
P10,425
P10,433
P10,402
P10,410
P10,418
P10,426
P10,434
P10,403
P10,411
P10,419
P10,427
P10,435
P10,404
P10,412
P10,420
P10,428
P10,436
P12,405
P12,413
P12,421
P12,429
P12,437
P12,406
P12,414
P12,422
P12,430
P12,438
P12,407
P12,415
P12,423
P12,431
P12,439
P12,408
P12,416
P12,424
P12,432
P12,440
```

Figure B.5: Merging script merge2.csv.

References

Andrews C., Endbert A., Yost B. and North C. (2011). Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, **10**, 341–355. 36

Andrews C., Endert A. and North C. (2010). Space to Think: Large, High-Resolution Displays for Sensemaking. *Proc. ACM CHI*. 33

Anslow C., Marshall S., Noble J., Tempero E. and Biddle R. (2010). User Evaluation of Polymetric Views Using a Large Visualization Wall. *Proc. IEEE SoftVis*. 37

Argo M. K. (2015). The e-MERLIN Data Reduction Pipeline. *Journal of Open Research Software*, **3**, e2. 1, 21

Aurisano J., Reda K., Johnson A., Marai E. G. and Leigh J. (2015). BactoGeNIE: a large-scale comparative genome visualization for big displays. *BMC Bioinformatics*. 34

Baan W. A. (2011). RFI mitigation in radio astronomy. *Conference Paper DOI: 10.1109/URSIGASS.2011.6051248*. 24

Beaudouin-Lafon M., Huot S., Nancel M., Mackay W., Pietriga E., Primet R., Wagner J., Chapuis O., Pillias C., Eagan J., Gjerlufsen T. and Klokmose C. (2012). Multisurface Interaction in the WILD Room. *IEEE Spectrum*. 36

Bi X., Bae S. H. and Balakrishnan R. (2010). Effects of Interior Bezels of Tiled-Monitor Large Displays on Visual Search, Tunnel Steering, and Target Selection. *Proc. ACM CHI*. 44

Birnholtz J. P., Grossman T., Mak C. and Balakrishnan R. (2007). An exploratory study of input configuration and group process in a negotiation task using a large display. *Proc. ACM CHI*. 44

Ebert A., Thelen S., Olech P. S., Meyer J. and Hagen H. (2010). Tiled++: An Enhanced Tiled Hi-Res Display Wall. *IEEE Transactions on Visualization and Computer Graphics*, **16**. 86

Eick S. G. and Karr A. F. (2002). Visual scalability. *Computational and Graphical Statistics*, **11**, 22–43. 37

Fateen W., Ruddle R. A., Treanor D., Sondergeld P. and Ouirke P. (October, 2013). Leveraging wall-sized high-resolution displays for comparative genomics analyses of copy number variation. *In 2013 IEEE Symposium on Biological Data Visualization (BioVis)*, pp. 89–96. 36

Ford J. M., Kaushal D. B. (2014). RFI mitigation techniques in radio astronomy. *Conference paper DOI 10.1109 IGARSS.2014.6946399*. 3, 24

Forlines C., Esenther A., Shen C., Wigdor D. and Ryall K. (2006). Multi-User, Multi-Display Interaction with a Single-User, Single-Display Geospatial Application. *Proc. ACM UIST*. 80

Fridman P. A., Baan W. A. (2001). RFI mitigation methods in radio astronomy. *Astronomy and Atrophysics*, **378**, 327–344. 24

Hamaker J. P., Bregman J. D., Sault R. J. (1996). Understanding radio polarimetry. *Astronomy Astrophysics*, **117**, 137–147. 12

Jansen Y., Dragicevic P. and Fekete J. D. (2012). Tangible Remote Controllers for Wall-Size Displays. *Proc. ACM CHI*. 70

Jeong B. , Renambot L., Jagodic R., Singh R., Aguilera J., Johnson A., Leigh J. (2006). High-performance dynamic graphics streaming for scalable adaptive graphics environment. *ACM/IEEE SC 2006 Conference*. 35

Kemball A. J. and Wieringa M. H. (2000). *MeasurementSet definition version 2.0*. National Radio Astronomy Observatory. 12

Knudsen S., Jakobsen M. R., Hornaek K. (2012). An Exploratory Study of How Abundant Display Space May Support Data Analysis. *Proc. ACM NordiCHI*. 36

Krishnaprasad N.K., Vishwanath V., Venkataraman S., Rao A.G., Renambot L., Leigh J., Johnson A.E. (2004). JuxtaView a tool for interactive visualization of large imagery on scalable tiled displays. *IEEE International Conference on Cluster Computing*, **4**, 411–420. 35

Liu C., Chapuis O., Beaudouin-Lafon M. and Lecolinet E. (2016). Shared Interaction on a Wall-Sized Display in a Data Manipulation Task. *Proc. ACM CHI*. 45

Liu C., Chapuis O., Beaudouin-Lafon M., Lecolinet E. and Mackay W. E. (2014). Effects of Display Size and Navigation Type on a Classification Task. *Proc. ACM CHI*. 68

McGinn D., Birch D., Akroyd D., Molina-Solana M., Guo Y. and Knottenbelt J. (2016). Visualizing Dynamic Bitcoin Transaction Patterns. *Big Data*, **4**. 33

Moldon J. (2018). The e-MERLIN CASA Pipeline. *University of Manchester*. 21

Munzner T. (2009). A Nessted Model for Visualization Design and Validation. *Transactions on visualization and computer graphics*, **15**, 921–928. 45

Ni T., Schmidt G. S., Staadt O. G., Livingston M. A., Ball R. and May R. (2006). A Survey of Large High-Resolution Display Technologies, Techniques, and Applications. *Proc. IEEE Virtual Reality*. 36

Offringa A. R., de Bruyn A. G., Biehl M., ZaroubiS S., Bernardi G. and Pandey V. N. (2010). Post-correlation radio frequency interference classification methods. *Monthly Notices of the Royal Astronomical Society*, **405**, 155–167. 2, 24

Olsen B. I., Dhakal S. B., Eldevik O. P., Hasvold P. and Hartvigsen G. (2008). A Large, High Resolution Tiled Display for Medical Use: Experiences from Prototyping of a Radiology Scenario. *eHealth Beyond the Horizon - Get IT There*, 535–540. 34

Ott J., Kern J. (2013). *CASA Synthesis Single Dish Reduction Reference Manual Cookbook*. National Radio Astronomy Observatory. 1, 13

Papadopoulos C., Mirhosseini S., Gutenko I., Petkov K., Kaufman A. E. and Laha B. (2015). Scalability Limits of Large Immersive High-Resolution Displays. *Proc. IEEE Virtual Reality*. 71, 85, 86

Papadopoulos C., Petkov K., Kaufman A. E. and Mueller K. (2015). The Reality Deck–an Immersive Gigapixel Display. *IEEE computer graphics and applications*, **35(1)**, 33–45. 35

Peck L. W. and Fenech D. M. (2013). Automated reduction and RFI-mitigation software for e-Merlin. *Astronomy and Computing*, **2**, 54–56. 2, 24

Peck S. M., North C. and Bowman D. (2009). A Multiscale Interaction Technique for Large, High-Resolution Displays. *Proc. IEEE 3DUI*. 86

Ponto K., Doerr K., Kuester F. (2010). Giga-stack: a method for visualizing giga-pixel layered imagery on massively tiled displays. *Future Generation Computer Systems*, **5**, 693–700. 35

Pretorius A. J., Bray M. A., Carpenter A. E. and Ruddle R. A. (2011). Visualization of parameter space for image analysis. *IEEE Transactions on Visualization and Computer Graphics.*, **17(12)**, 2402–2411. 40

Reda K., Knoll A., Nomura K., Papka M. E., Johnson A. E. and Leigh J. (2013). Visualizing Large-Scale Atomistic Simulations in Ultra-Resolution Immersive Environments. *Proc. IEEE LDAV*. 34

Rooney C. and Ruddle R. A. (2015). HiReD: a high-resolution multi-window visualisation environment for cluster-driven displays. *In Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pp. 2–11. 33, 34

Ruddle R. A., Thomas R. G., Randell R. S., Quirke P. and Treanor D. (2013). Performance and interaction behaviour during visual search on large, high-resolution displays. Information Visualization. *In 2013 IEEE Symposium on Biological Data Visualization (BioVis)*, **14(2)**, 137–147. 36

Sandstrom T. A., Henze C. and Levit C. (2003). The hyperwall. *Proc. CMV*. 35

Sedlmar M., Meyer M., Munzner T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *Transactions on visualization and computer graphics*, **18**, 2431–2440. 38

Shupp L., Andrews C., Dickey-Kurdziolek M., Yost B. and North C. (2009). Shaping the Display of the Future: The Effects of Display Size and Curvature on User Performance and Insights. *Human-Computer Interaction*, **24**, 230–272. 85, 86

Singh A., Endert A., Andrews C., Bradel L., Kincaid R. and North C. (2011). Supporting the cyber analytic process using visual history on large displays. *Proc. VIZSEC*. 34

Slocum T. A., Dunbar M. D. and Egbert S. L. (2007). Evaluating the Potential of the GeoWall for Geographic Education. *Geography*, **106**. 34

Westing B., Urick B., Esteva M., Rojas F. and Xu W. (2011). Integrating Multi-touch in High-Resolution Display Environments. *Proc. Scientific Computing*. 36, 70

Wilson T. L., Rohlfs K., Huttemeister S. (2009). *Tools of Radioastronomy*. Springer. 24

Yamaoka S., Doerr K. and Kuester F. (2011). Visualization of high-resolution image collections on large tiled display walls. *Future Generation Computer Systems*, **27**, 498–505. 35

Yamaoka S., Manovich L., Douglass J. and Kuester F. (2011). Cultural analytics in large-scale visualization environments. *IEEE Computer*. 34

Yost B., Haciahmetoglu Y. and North C. (2007). Beyond physical acuity: The perceptual scalability of information visualizations for large displays. *Proceedings ACM CHI*. 44, 72

Yuan X., He X., Guo H., Guo P., Kendall W., Huang J. and Zhang Y. (2010). Scalable multi-variate analytics of seismic and satellite-based observational data. *IEEE Transactions on Visualization and Computer Graphics*, **16**, 1413–1420. 34