# Chapter 3

# Automatic evaluation of data quality

---

*"The totality of features and characteristics*
*which decide the ability to satisfy needs."*
- a phrase for quality assurance

*"A problem cannot be solved,*
*at the same level of consciousness as it was created."*
- Albert Einstein

In this chapter, we discuss the concept of data quality, its requirement for our astronomy work and present an algorithm[1] for its automatic evaluation for our visibility data. The main aim of the algorithm is to automatically classify the visibility data (1 TB) depending on its quality and usefulness for imaging. The framework is based on finding the key parameters (criteria) which affect the data quality. To a visibility file, for each key parameter, a numerical index from 1 to 9 is assigned which is called the Quality Factor (QF). The QF based on each parameter assesses the extent up to which the observed visibilities satisfy the expected requirements for that criterion. The overall quality of a visibility file is calculated using all the QFs assigned based on individual key criterion, namely the completeness, interference statistics, rms noise and the effect of sun on the visibility data. We also present the results of application of such a framework to our database. The results are very encouraging and application of this technique has made the data classification objective, efficient and fast compared to the traditional manual approach. Its critical role in our surveying cannot be overemphasized. With suitable modifications, an astronomical tool based on such an algorithm can be potentially employed on data sets from other interferometric arrays and to monitor the performance of an observatory itself.

## 3.1 Introduction

The advancement in technology and the quest for higher angular resolution and sensitivity is fueling rapid growth of data in astronomy. Such an increase in volume of astronomical data and a variety of factors affecting its suitability for imaging, has forced us to spend considerable effort on the analysis of data and its management, often at the expense of addressing the real underlying physics of the problem. For example, while handling data from radio interferometers such as GMRT (Giant Metrewave Radio Telescope), VLA (Very Large Array), ATCA (Australia Telescope Compact Array), the usual approach lies in identifying discrepant and severely corrupted data, using imaging softwares like AIPS[2] and MIRIAD[3] etc. with handy display tools which aid in discovering faulty data using several forms of editing. It is very important to edit data of bad quality as it is usually worse than no data at all. The errors are either gross which are easily identifiable or subtle in which there is uncertainty in identifying them. Gross errors can be generally handled easily by using automatic methods or data visualization tools or by a combination of both.

[1] In the $9^{th}$ century Al-Khawarizmi introduced the concept of algorithm, which provided a universal method for solving a problem by repeatedly using a simpler computational method. This formed the basis of formalization of methods in engineering. The word “algorithm” is derived from his name.

[2] AIPS is a software package for calibration, data analysis, image display, plotting and a variety of ancillary tasks on astronomical data. It comes from the National Radio Astronomy Observatory.

[3] Miriad is a radio interferometry data reduction package, alternative to AIPS, for reduction of continuum and spectral line astronomy data. It comes from the Australia Telescope National Facility.

Data visualization makes it possible for an astronomer to gain a deeper and more intuitive understanding and is generally the main form of editing for subtle data. It allows user to focus on certain patterns and trends, form a hypothesis regarding the possible reasons for the occurrence of such problems and carry out manual editing of data depending on the reasons. The problem gets further compounded when dealing with data from synthesis telescopes, as the number of baselines and frequency channels make the task painstaking and time consuming. Since it is difficult to display all the data from the modern synthesis telescopes, as a compromise one carefully chooses a subset of representative data and it is usually possible to discover faulty data in nearly all the data set except in pathological cases.

The process to analyse astronomical data is complicated and at times subjective due to which human intervention is still far from being inevitable. There are many archives which have been constructed for the astronomical data gathered by the Hubble Space Telescope (HST), Sloan Digital Sky Survey (SDSS), the Chandra X-ray Observatory, the Two Micron All Sky Survey (2MASS) etc.. In the near future, several large array telescopes like ALMA (Atacama Large Millimeter Array), LOFAR, SKA etc. are being planned which will churn out huge amounts of data. In addition, the planned Global Virtual Observatory will give access to a large quantity of data, where one will have to spend considerable time in choosing the data one would like to use. The amount of unprecedented data will be difficult to handle by traditional methods as in order to organize and make its good use, it would require enormous time and high quality human resource. In the years to come the problem of handling astronomical data may soon become unsurmountable with the traditional approach. This encouraged us to look for alternative approaches to confront this challenge and arrive at possible solutions. Here, we discuss our technique devised to satisfy the need for automatic classification of visibility data for the MRT survey. Our main emphasis is on the approach to data classification and its implementation in software with a view to automate it.

## 3.2 Automatic classification in astronomy

There have been numerous efforts to use a variety of techniques both deterministic and non-deterministic for automatic classification, a few of which include using include Knowledge Based Systems (KBS) (Gesu, 1989), Artificial Neural Networks (ANN), Principal Component Analysis (PCA), Minimum Distance Methods (MNM) and Gaussian Probabilistic Models (GPM) etc. (see Bailer-Jones (2002b) for an overview). The non-deterministic algorithms have been mainly used for catalogue extraction (Andreon et. al.,

2000), star/galaxy classification (Odewahn et. al., 1992; Nail et. al., 1995; Miller & Coe, 1996; Mähönen et. al., 1995; Bertin & Arnout, 1996; Bazell & Peng, 1998), galaxy morphology (Storrie-Lombardi et. al., 1992; Lahav et. al., 1995) and classification of stellar spectra (Bailer-Jones et. al., 1998; Allende et. al., 2000). There have also been attempts using deterministic algorithms for morphological classification of galaxies into ellipticals, spirals, lenticulars and irregulars (Thonnat, 1989; Okamura et. al., 1989; Doi et. al., 1992; Spiekermann, 1992). To our knowledge most of the previously described techniques have been mainly used for object classification in astronomy, where one deals primarily with pattern recognition in images and these have been mostly dominated by non-deterministic algorithms.

## 3.3 Need for data classification at MRT

As mentioned earlier in Chapter 2, imaging for the MRT survey is carried out on a sidereal hour basis. Fig. 2.3 shows the main steps of data processing to synthesize the full resolution images. To synthesize a full resolution image[4] covering one sidereal hour range and entire declination range of MRT, the first step is to choose good visibility data files for each of the four delay zones and all the allocations of the sidereal hour range to be imaged. Optimized usage of the survey data is important to reap the efforts gone in for carrying out the observations.

Due to paucity of strong point sources in the southern sky we use only three sources for calibration of visibilities namely, MRC0915-118, MRC1932-464 and MRC2211-172, having flux densities 252 Jy, 96 Jy and 84 Jy respectively at 151.5 MHz (Large et. al., 1981; Slee, 1995). This leaves us with a situation where we can calibrate the array only three times during a sidereal day (Sec. 5.2). So along with choosing a good data file we also need to look at the availability and quality of its calibration file in order to decide whether a given file can be satisfactorily imaged.

In any one sidereal hour range there are on an average ≈3,500 data files which include observations of 63 allocations with appropriate delay zones. In the traditional human approach all the data files need to be examined to estimate their quality. For each data file we need to check the availability and estimate the quality of all the calibration files, corresponding to the observations of the three calibrators and subsequently choose the most suitable calibration file. To accomplish this using the traditional approach, assuming ≈4 min (based on manual experience with the data) of time taken to visually inspect

[4]We recall that images for a survey with partial resolution (17′×23′) with MRT were made earlier by Golap (1998) while data was being collected with longer NS baselines. Due to this the term full resolution image is used and refers to an image with a resolution of $4' \times 4\overset{s}{.}6 \sec(\delta + 20\overset{\circ}{.}14)$ made using data from all the allocations.

a data file, its all available calibration files for various parameters, comprehend them and take a decision, a rough calculation reveals that it would require ≈30 man days assuming 8 working hours a day ($\because$ 3,500×4 min=14,000 min≈240 hours). For the entire database (all the sidereal hours) this translates to about 2 man years just for selecting the good visibility data for imaging. In case an error occurs due to some parameter inadvertently skipped during the decision making process, redoing the entire process later becomes a very strenuous exercise. In view of this there was a compelling need to develop a framework to classify the visibility data which can be handled by automation and can replace the insight of an astronomer to the best possible extent.

## 3.4 A new classification scheme

The problem of defining the data quality in radio astronomy does not have an exact prescription. Human experience plays a vital role whenever we are dealing with cloudy properties which are inexact, vague and are dependent on many parameters in a complicated manner which is not fully understood. Since the problem falls in the purview of fuzzy logic and it may be useful to utilize concepts involved to develop tools to solve similar fuzzy logic problems. Nevertheless we feel it may still be beneficial to formulate the problem based on parameters on which the behavior of the problem is well understood and use approximations for those wherever one does not have a precise knowledge of its effect[5]. Application of such a framework to the data may yield useful results. We worked on possible automatic methods for classification of data as an alternative to the traditional approach. Our initial attempts for visibility data classification are briefly discussed in Pandey & Udaya Shankar (2002b). One of the important advantages of such automation algorithms is that if one wants to improve, the entire exercise can be repeated after incorporating the appropriate corrections. The repeatability may not be practically possible with the usual manual traditional approach due to large amount of time required. An interesting and somewhat similar approach has been developed for automated identification of emission lines from the spectra (Sharpee et. al., 2003). In their approach automatic identification of spectral lines based on assigning numerical identification index to identification criteria is attempted.

To start with, we listed various factors which an astronomer generally looks for to decide the quality of a visibility file based upon our experience in handling the MRT survey

[5]In our view, mature solutions to such problems which do not have an exact conceptual prescription like quality of data may evolve on similar lines to an engineering discipline. Shaw (1990) in her model for evolution of an engineering discipline states *"Historically, engineering has emerged from ad hoc practice in two stages: First, management and production techniques enable routine production. Later, the problems of routine production stimulate the development of a supporting science; the mature science eventually merges with established practice to yield professional engineering practice."*

data. The overall suitability of a data file to be considered for imaging depends upon,

a. The quality of the data file.

b. The availability and quality of the calibration file.

In order to express the overall suitability of the data file for imaging on a quantitative basis, we define Total Quality Factor (TQF) as,

$$TQF = \frac{QF_{data} + QF_{calib}}{2} \tag{3.1}$$

where $QF_{data}$ and $QF_{calib}$ are the QFs of the data and its calibration file.

Quantification of quality of data is difficult since there is no unique way to define it. However astrophysical consistency, a priori information about the field of view, expected instrument response, working health of the telescope, RFI, ionospheric conditions and simple logic etc. are important inputs which help us to distinguish good measurements against possible incorrect measurements.

The observations for the MRT survey were carried out around the solar minimum (May ,1996) when the Sun is relatively quiet due to which we can expect the ionospheric conditions to be less harmful (See Sec. 3.4.1 and Fig. 3.1 for details). In addition MRT is located in a zone where the effect of ionospheric variations are minimum and less harmful. In view of this the effects of ionospheric variation for the maximum baseline of 2 km of MRT are not significant. We thus evolved with the following list of parameters to decide the quality of the data and its calibration file.

1. Inputs from the observers log report.
2. The completeness of the visibility file.
3. RFI statistics of the visibility file.
4. The rms noise in the visibility file.
5. The strength and the position of the Sun during observation of the visibility file.

In order to take into account the adverse remarks in the observers log report, all such visibility files were excluded from the database. After this, we were left with the last four parameters. This problem can be addressed by an automated procedure which keeps track of these key parameters to at least some desired acceptable satisfaction level. With this approach in mind, we arrived at a QF assignment process, which is based on assigning an integer numerical index from 1 to 9, called the Quality Factor (QF), to the observed visibility file based on each key criterion. This numerical QF for each of the criterion assesses the extent up to which the observed visibilities satisfy the expected requirements for that criterion. The individual QFs based on each key criterion are denoted by $QF_1$, $QF_2$, $QF_3$

and $QF_4$ which stand for QF based on completeness, RFI statistics, rms noise and effect of the Sun on the visibilities respectively. A lower value of QF indicates a better quality data. QF of 1 indicates data of best quality while QF of 9 indicates data of a very poor quality. We generally consider a QF less than 5 as acceptable. The overall quality of a data ($QF_{data}$), or a calibration ($QF_{calib}$) file is defined as the simple average of the individual QFs.

$$QF_{data} \; or \; QF_{calib} = \frac{QF_1 + QF_2 + QF_3 + QF_4}{4} \tag{3.2}$$

For a data file the individual QFs are assigned taking into account the visibilities for one complete sidereal hour range. In case of the calibration file, only a short duration of the visibility data relevant for calibration is taken into account for assigning the individual QFs (See Sec. 5.2). Now, we describe each of the key criteria and the assignment of QFs based on them.

### 3.4.1 Assignment of the Quality Factor

**QF based on completeness** :

The first step while assigning the $QF_1$ to a visibility file is to check for the completeness of the data. Completeness refers to uninterrupted observation for the duration of one sidereal hour range for a data file. For calibration file the duration considered is $8 \times \sec(\delta)$ minutes around the calibrator source transit (to transit the FWHM of the EW group along RA, a source at $\delta$=0° takes ≈8 min while a source at $\delta$=-70° takes ≈23 min). The interruption during the observations is most often due to mains power failures and occasionally due to instrumental failures. A visibility file which is complete is assigned a $QF_1$ of 1. A visibility file which is complete for less than half of the duration considered is assigned a $QF_1$ of 9. The $QF_1$ for visibilities is decided as specified in the Table 3.1. Although for imaging for the survey, we have used files which are complete for one sidereal hour (See Sec. 3.4.4), we have still assigned $QF_1$ for completeness in a number of bins as in general one may also consider imaging subparts of one sidereal hour range. In such a case only the threshold limit (QF cut-off within which files are considered acceptable for imaging; See Sec. 3.4.4) for completeness has to be changed in the present scheme for classification.

**QF based on RFI statistics** :

At MRT we use 480 complex visibilities (32 EW×15 NS) each second for imaging (See Sec. 5.3). The interference detection is carried out in the domain of *sum of magnitudes* of visibilities on all these baselines based on the assumption that interference generally

| S.No. | $C_r$ | $QF_1$ |
|---|---|---|
| 1 | $0.99 \leq C_r \leq 1.00$ | 1 |
| 2 | $0.95 \leq C_r < 0.99$ | 2 |
| 3 | $0.90 \leq C_r < 0.95$ | 3 |
| 4 | $0.85 \leq C_r < 0.90$ | 4 |
| 5 | $0.80 \leq C_r < 0.85$ | 5 |
| 6 | $0.70 \leq C_r < 0.80$ | 6 |
| 7 | $0.60 \leq C_r < 0.70$ | 7 |
| 8 | $0.50 \leq C_r < 0.60$ | 8 |
| 9 | $0.00 \leq C_r < 0.50$ | 9 |

Table 3.1: Guidelines for assignment of $QF_1$ based on completeness for data and calibration files. In case of data file $C_r$ is fraction of the time for which the observations are available in one sidereal hour, while for a calibration file it is the fraction of time for which the observations are available in the time range relevant for calibration.

| **S.No.** | **Data File** (3300*) | | | **MRC0915-118 (450*)** | | | **MRC1932-464 (640*)** | | | **MRC2211-172 (460*)** | | | **QF**$_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $N_{int}$ | $T_{max}$ | $N_{HPBW}$ | $N_{int}$ | $T_{max}$ | $N_{HPBW}$ | $N_{int}$ | $T_{max}$ | $N_{HPBW}$ | $N_{int}$ | $T_{max}$ | $N_{HPBW}$ | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | $\leq$ 50 | $\leq$ 16 | 0 | $\leq$ 10 | $\leq$ 8 | 0 | $\leq$ 14 | $\leq$ 8 | 0 | $\leq$ 10 | $\leq$ 8 | 0 | 2 |
| 3 | $\leq$ 100 | $\leq$ 24 | $\leq$ 2 | $\leq$ 20 | $\leq$ 16 | 0 | $\leq$ 28 | $\leq$ 16 | 0 | $\leq$ 20 | $\leq$ 16 | 0 | 3 |
| 4 | $\leq$ 200 | $\leq$ 32 | $\leq$ 4 | $\leq$ 30 | $\leq$ 16 | $\leq$ 1 | $\leq$ 43 | $\leq$ 16 | $\leq$ 1 | $\leq$ 31 | $\leq$ 16 | $\leq$ 1 | 4 |
| 5 | $\leq$ 300 | $\leq$ 48 | $\leq$ 6 | $\leq$ 50 | $\leq$ 24 | $\leq$ 2 | $\leq$ 71 | $\leq$ 24 | $\leq$ 2 | $\leq$ 51 | $\leq$ 24 | $\leq$ 2 | 5 |
| 6 | $\leq$ 500 | $\leq$ 64 | $\leq$ 10 | $\leq$ 80 | $\leq$ 24 | $\leq$ 4 | $\leq$ 114 | $\leq$ 24 | $\leq$ 4 | $\leq$ 82 | $\leq$ 24 | $\leq$ 4 | 6 |
| 7 | $\leq$ 1000 | $\leq$ 120 | $\leq$ 25 | $\leq$ 150 | $\leq$ 32 | $\leq$ 5 | $\leq$ 213 | $\leq$ 32 | $\leq$ 5 | $\leq$ 153 | $\leq$ 32 | $\leq$ 5 | 7 |
| 8 | $\leq$ 1500 | $\leq$ 200 | $\leq$ 40 | $\leq$ 225 | $\leq$ 48 | $\leq$ 8 | $\leq$ 320 | $\leq$ 48 | $\leq$ 8 | $\leq$ 230 | $\leq$ 48 | $\leq$ 8 | 8 |
| 9 | - | – | - | - | – | - | - | – | - | - | – | - | 9 |

Table 3.2: Guidelines for assignment of $QF_2$, based on interference statistics of the data and calibration files of the calibrators MRC0915-118, MRC1932-464 and MRC2211-172. $N_{int}$ is the number of interference points detected in the sidereal hour range used. The total number of points (*) for a data file is ≈3300 points, while there are 450 (≈8.2 min), 640 (≈11.7 min), 460 (≈8.4 min) points used for calibration file of by MRC0915-118, MRC1932-464 and MRC2211-172 respectively. $T_{max}$ is the maximum number of interference points occurring in a continuous stretch. $N_{HPBW}$ is the number of times interference is continuous for more than 16 sidereal seconds (Half Power Beam Width (HPBW) of the array in RA). For each point the integration period is ≈1.09 s.

affects all the baselines simultaneously[6], using a conjunction of Fourier filtering, Hampel filtering followed by a GUI based visual inspection (see Sec. 4.3.3). The details of the detected interference are stored in a centralized flag table for each visibility file.

The number of interference points and their duration is used to assign the $QF_2$ to a visibility file. A visibility file with short duration interference is preferred to the one with long duration interference stretches. The assignment of $QF_2$ for the data file and for the calibration files containing the observation of the calibrators MRC0915-118, MRC1932-464 and MRC2211-172 is decided as specified in the Table 3.2. The assignment of $QF_2$ has been decided based on our experience with the MRT database. During the assignment of $QF_2$ we check the required conditions to be satisfied in ascending order of $QF_2$. A visibility file with no interference is assigned $QF_2$ of 1. A visibility file having interference for more than half the total number of points considered is assigned a $QF_2$ of 9.

[6]In this chapter all the baselines refer to 480 baselines used for imaging unless specified otherwise.

**QF based on noise in the visibilities** :

The r.m.s noise at the output of a correlator is given by (Crane et. al., 1989),

$$\Delta S = \frac{1}{\sqrt{(\Delta\tau\Delta\nu)}} \times \sqrt{S^2 + \frac{S}{2}\left(\frac{2k_bT_{sys1}}{A_1} + \frac{2k_bT_{sys2}}{A_2}\right) + \frac{2k_b^2T_{sys1}T_{sys2}}{A_1A_2}} \tag{3.3}$$

where $\Delta\tau$ is the integration time, $\Delta\nu$ is the bandwidth of the signals being correlated, $S$ is the flux density of the main source in the antenna beams, $A_1$ and $A_2$ are the collecting areas of the two antennas whose outputs are being correlated, $k_b = 1.38 \times 10^{-23}$ JK$^{-1}$ is the boltzman constant, $T_{sys1}$ and $T_{sys2}$ are the system temperatures of the two antennas of the interferometer. The rms noise due to the flux density of a source can be neglected if,

$$S^2 + \frac{S}{2}\left(\frac{2k_bT_{sys1}}{A_1} + \frac{2k_bT_{sys2}}{A_2}\right) << \frac{2k_b^2T_{sys1}T_{sys2}}{A_1A_2}$$

At 150 MHz, the sky brightness temperature does not dominate the system temperature as it would at still lower frequencies. Both the receiver temperature ($T_{rcvr}$) and the sky background temperature ($T_{bg}$) contribute to the system temperature ($T_{sys}$).

$$T_{sys} = T_{bg} + T_{rcvr} \tag{3.4}$$

At MRT we correlate EW groups with the NS groups. They have different primary beams. Estimates show that on an average the $T_{bg}$ for both elements of the interferometer are in the range 150 K<$T_{bg}$<300 K. Assuming the receiver temperature, $T_{rcvr}\approx$300 K, $T_{sys}$=$T_{sys1}$=$T_{sys2}$, the source noise can be neglected if,

$$S << \frac{1}{2}\frac{k_bT_{sys}}{A_1A_2}\left\{\sqrt{(A_1+A_2)^2 + 8A_1A_2} - (A_1+A_2)\right\} \tag{3.5}$$

substituting $T_{sys} = 600$ K, $A_1 = 32 \times 4m^2$ and $A_2 = 4 \times 4m^2$, in the Eqn. 3.5 we get $S <<$ 9842 Jy. This condition is generally satisfied for all practical purposes at MRT, except in the case of active Sun. The rms noise when the source noise does not dominate is given by,

$$\Delta S = \frac{\sqrt{2}\,k_bT_{sys}}{\sqrt{A_1A_2}\,\sqrt{\Delta\tau\Delta\nu}} \tag{3.6}$$

Using Eqn. 3.6, we can calculate the noise expected in the measured visibilities by substituting the appropriate value of $T_{rcvr}$ and $T_{bg}$ depending upon region of the sky under consideration. The estimated value of rms noise for MRT interferometer, for 1 second integration and 1 MHz bandwidth, is $\approx$ 26 Jy. When a very strong source such as Sun is in the primary beam, the rms noise depends upon the flux density of the source. In case of Sun's transit, if we assume it were a point source for MRT and taking typical value of the

total solar flux as 5 sfu[7], using Eqn. 3.3 we get $\Delta S = 78$ Jy ($T_{sys}$=1800 K). Thus the rms noise can increase up to three times the normally expected value. In practice the actual increase depends upon the angular distance of the Sun, attenuation due to the primary beam response of the helix and Sun getting resolved by the array.

| S.No. | $\sigma_{ratio}$ | $QF_3$ |
|---|---|---|
| 1 | $1 \leq \sigma_r \leq 1.2$ | 1 |
| 2 | $1.2 < \sigma_r \leq \sqrt{2}$ | 2 |
| 3 | $\sqrt{2} < \sigma_r \leq \sqrt{3}$ | 3 |
| 4 | $\sqrt{3} < \sigma_r \leq \sqrt{4}$ | 4 |
| 5 | $\sqrt{4} < \sigma_r \leq \sqrt{5}$ | 5 |
| 6 | $\sqrt{5} < \sigma_r \leq \sqrt{6}$ | 6 |
| 7 | $\sqrt{6} < \sigma_r \leq \sqrt{7}$ | 7 |
| 8 | $\sqrt{7} < \sigma_r \leq \sqrt{8}$ | 8 |
| 9 | $\sqrt{8} < \sigma_r$ | 9 |

Table 3.3: Guidelines for assignment of $QF_3$, based on noise in the visibilities for data file as well as the calibration file. $\sigma_{ratio}$ is the ratio of rms noise measured ($\sigma_{meaus}$) in the observed visibilities to the expected rms noise ($\sigma_{theo}$) for the region of the sky under consideration.

The ratio of the measured noise in the visibilities to the expected noise (calculated using Eqn. 3.6 which does not include contribution due to source noise) is used to assign the $QF_3$ based on noise. The noise in the measured visibilities is estimated excluding all the interference points. For the data file the rms noise is calculated from the measured visibilities of the entire sidereal hour range while for the calibration file only the duration relevant for calibration is taken into account. The main principle while assigning the $QF_3$ based on the fact that the noise decreases by $\sqrt{n}$ times, where $n$ is the number of visibility files used to synthesize the image. The $QF_3$ assignment for the data file as well as for the calibration file is decided as specified in Table 3.3. A file with a measured rms noise within 20% of the expected rms is assigned a $QF_3$ of 1 while a file with measured rms noise more than $\sqrt{8}$ times the expected value is assigned a $QF_3$ of 9.

**QF assignment based on effect of the Sun on the visibilities** :

At MRT, generally night time observations are interference free and are most suitable for imaging. During the day time the effect of Sun which is a strong radio source becomes important due to wide primary beam of the helix. Since the position of the Sun changes in the sky with time, while we collect different Fourier components of the same portion of the sky over a period of time, we do not intend to image the Sun in the MRT survey. A solar cycle has a period of 11 years during which it goes through a maximum and minimum of Sun's activity. Fig. 3.1(a) shows the ISES[8] solar cycle Sun Spot Number (SSN) progression

[7]Solar Flux Unit; 1 sfu=$10^{-22}$ W m$^{-2}$Hz$^{-1}$=10,000 Jy.

[8]International Space Environment Service; http://www.ises-spaceweather.org/index.html

from January, 1994 to January, 2006 (part of solar cycle 22[9] and solar cycle 23[10]). The radio emission from the Sun (*e.g.* 10 cm flux) has been found to correlate well with the Sun spot number. Fig. 3.1(b) shows the ISES solar cycle F10.7 cm radio flux progression for the same period[11]. The duration of observations for the MRT survey lies during solar minima when the solar activity is relatively low. Sun's total solar flux at 150 MHz is typically 5 sfu and generally varies from 2 sfu to 20 sfu but occasionally can increase up to 200 sfu. The radio size of the Sun is about 34′×34′. Sun affects the visibilities mainly in two ways.

a. The increase in system temperature.

b. The Sun itself present in the images when it lies in the area of the sky to be imaged or causing artifacts in the images via its sidelobes or when it is transiting in the grating lobes.

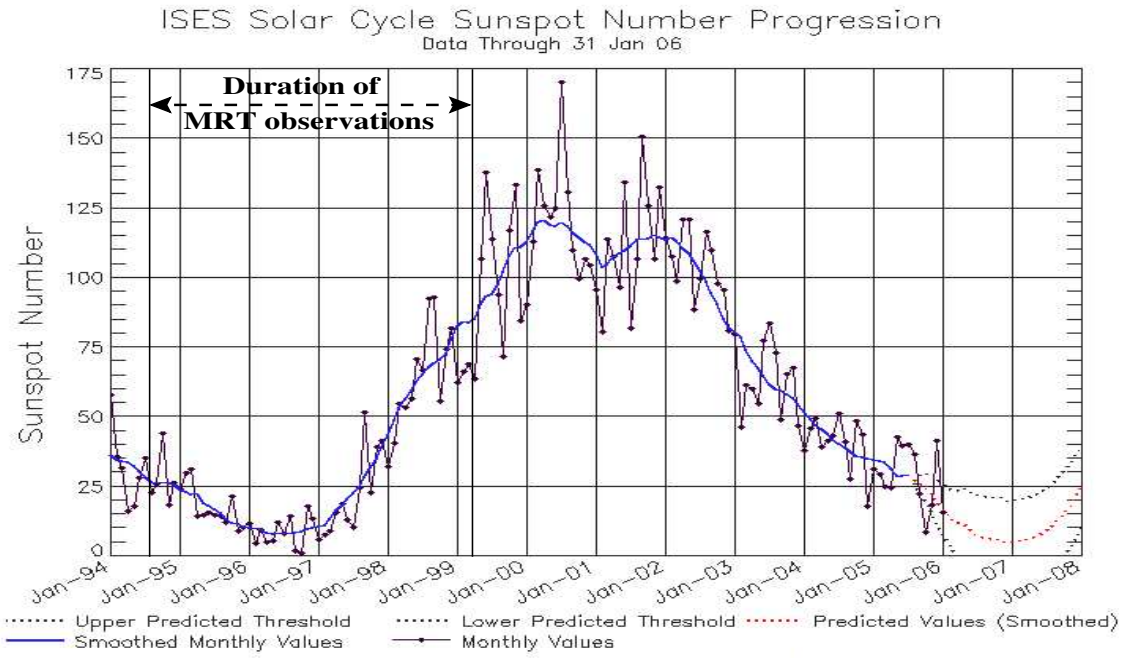
The increase in system temperature due to Sun is reflected in the noise measured in visibilities and has already been included in the assignment of $QF_3$. Thus we consider only aspect related to the strength of the Sun when it is present in the image, or strength of the artifacts when it affects via its sidelobes/grating lobes, to decide the $QF_4$. Before discussing the assignment of $QF_4$, we discuss the grating response of the array as it has to be taken into account to estimate the effect of Sun on the visibilities.

**Grating response of the array:** There are grating responses along the EW direction due to the helices being spaced 2 m ($d_{ew}$) apart in the EW and NS groups. These grating responses are at intervals of $\Delta l = \frac{\lambda}{d_{ew}}$, where $l$ is the direction cosine in the EW direction, $\lambda$ is the wavelength. These are in addition to the ones due to the sampling in the EW being equal to the group size (See Sec. 5.3.1). The strength of a source will be attenuated by the product of primary beam response of the helix and the combined grating response of the EW and NS group in the direction of the source. Fig. 3.2 shows the combined two-dimensional response of an EW and the NS group as a function of Hour Angle (HA) and declination. Fig. 3.3 shows the color image of one of the two lobes from the combined two-dimensional response of an EW and the NS group as a function of Hour Angle (HA) and declination. From the two-dimensional response, we note that the grating response peaks in at declination $\delta = 0^\circ$ and decreases for declinations away from it. The grating response peaks at an hour angle $H_g$=±81.°5. Although the grating response is in the direction where the theoretical beam response of the helix is expected to have an attenuation of about 30 dB, we still see Sun coming through these grating beams during solar activ-

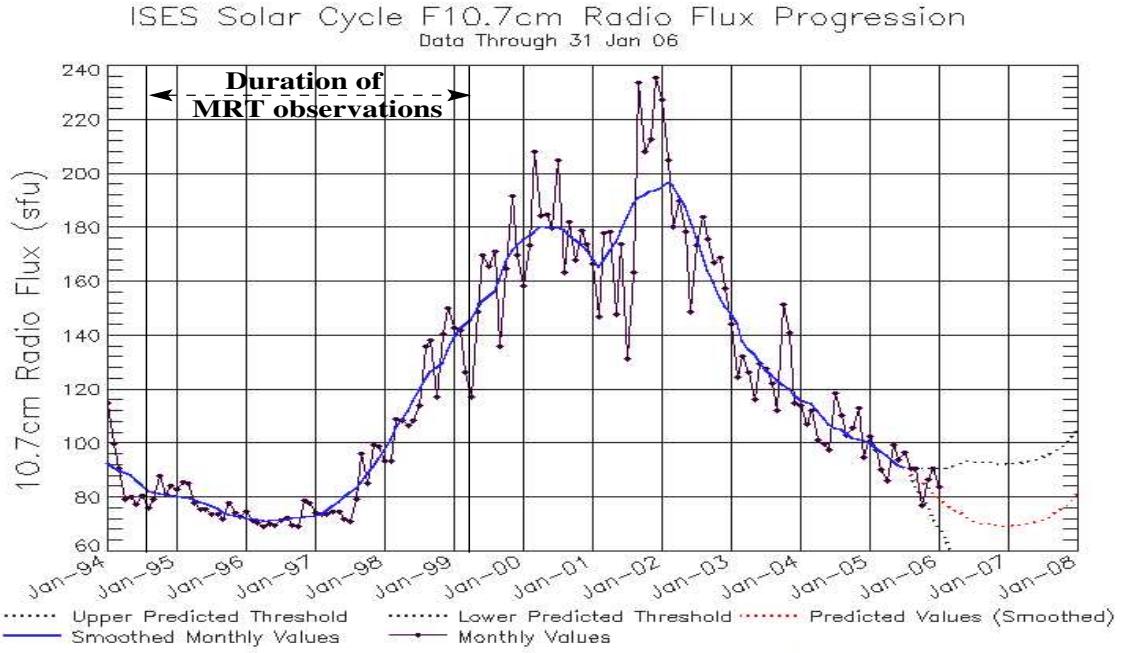
[9] Starts September, 1986; ends April, 1996. Minimum SSN : 12.3 and Maximum SSN : 158.5

[10] Starts May, 1996; ends December, 2005. Minimum SSN : 8.0 and Maximum SSN : 120.8

[11] source for the Fig. 3.1 is Space Environment Center (SEC), National Oceanic and Atmospheric administration (NOAA), USA (http://www.sec.noaa.gov/SolarCycle ; http://www.noaa.gov).




(a) ISES Solar Cycle Sunspot Number Progression. This plot displays SIDC monthly sunspot numbers, SIDC 13-month running smoothed sunspot numbers and the most recent forecast.




(b) ISES Solar Cycle F10.7cm Radio Flux Progression. This plot displays monthly Penticton 10.7 cm Radio Flux values, 13-month running smoothed values and the most recent forecast.

Fig. 3.1: A plot showing the solar activity from January, 1994 to January, 2006 as solar cycle sunspot number progression and F10.7cm radio flux progression. Most of the observations for the MRT survey were carried out during the period when the Sun's activity has been relatively low. Image source - Space Environment Center (SEC), National Oceanic and Atmospheric administration (NOAA), USA.

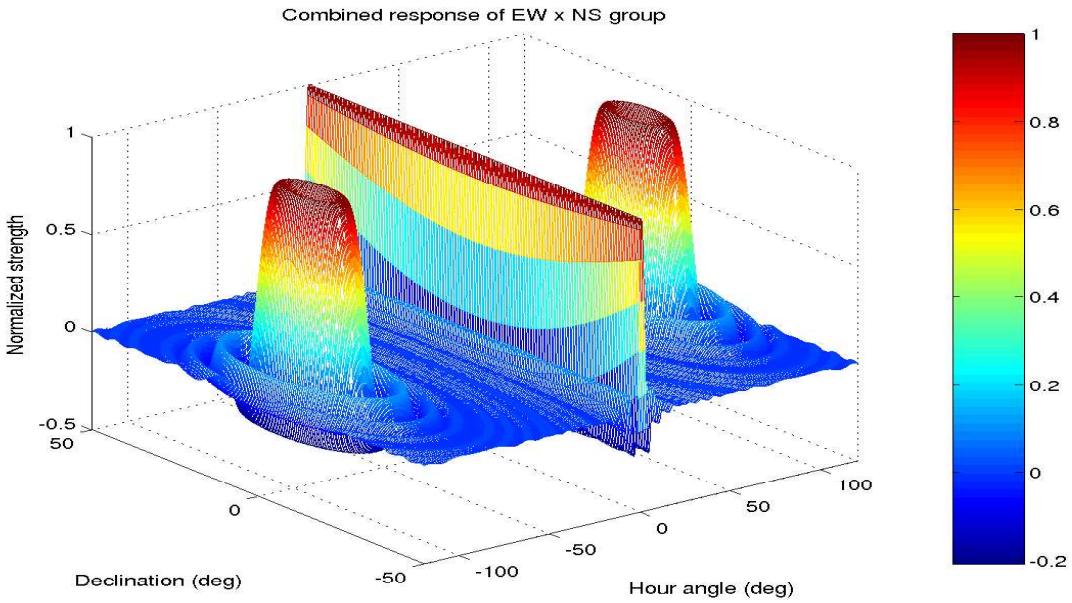



Fig. 3.2: The two dimensional grating response of the EW×NS as a function of HA from the meridian and declination of the source (without taking into account the attenuation due to the primary beam of the helix). The grating response peaks at an HA≈ ±81.°5 and 0° in declination. The FWHM of the grating lobes is ≈28° in HA and ≈22.°5 in declination.

ity especially when the Sun is within the declination range ±10°. Strong sources in this declination range will also contribute to the visibility measurements on short baselines, but this effect has not been observed so far possibly because we do not have many strong sources in this declination range. However, sources like the Sun and the Galactic plane in the declination region ±10°pose a problem and must be accounted for.

A source in the grating response does not appear at its true declination when imaging on the meridian. A source in the grating at an hour angle $H_g$ and declination $\delta_g$ has a direction cosine with the NS direction $m_g$ given by,

$$m_g = \cos\delta_g \cos H_g \sin\phi - \sin\delta_g \cos\phi \tag{3.7}$$

where $\phi$ is the latitude of the telescope. While imaging at meridian transit, i.e., HA=0°, the direction cosine $m$ is given by

$$m = \sin(\phi - \delta_i) \tag{3.8}$$

where $\delta_i$ is the declination at which the source appears in the image at the meridian. Therefore, the declination to which a source in the grating response gets imaged on the meridian is obtained by equating $m_g$ to $m$ and is given by,

$$\delta_i = \phi - \sin^{-1}(m_g) \tag{3.9}$$

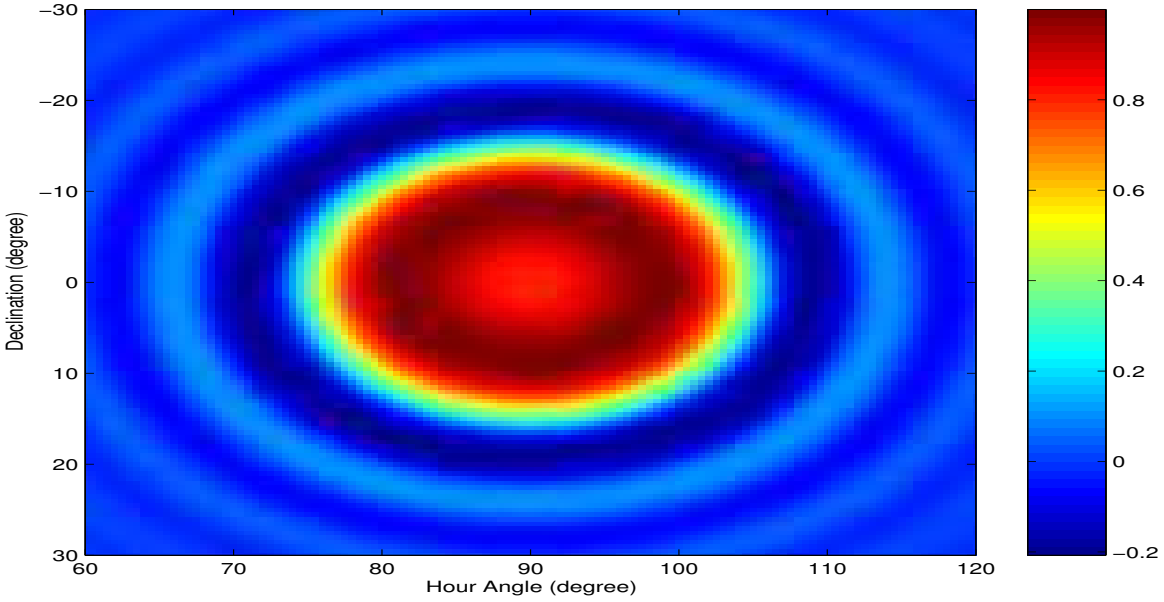



Fig. 3.3: A color image of one of the grating lobes from the two dimensional combined grating response of the EW and the NS group as a function of HA from the meridian and declination of the source (without taking into account the attenuation due to the primary beam of the helix). The grating response peaks at an HA of ≈81.°5 and 0° in declination. The FWHM of the grating lobes is ≈28° in HA and ≈22.°5 in declination.

Fig. 3.4 shows the image artifacts caused by the Sun transiting in grating response of the array in an allocation's image for the RA 18-19 hrs. The true position of the Sun on the day of observation of the visibility (JD2450529; date 22:03:1997) is RA 00:03hrs, Declination -0.°4 and its average total flux is 2.9 sfu. The expected position of the grating lobes in the image is at RA ≈18:37 hrs, Declination -17.°3 (estimated using Eqn. 3.8). The image clearly shows the grating lobe response at this expected position and also shows its aliased image at nearly the same RA and declination 1°. The strong artifacts corrupt the entire image. The effect of Sun via its sidelobes in the images has a very characteristic appearance, the grating lobe image appears elongated and is inclined to the RA axis. The lengthening is due to the sin(HA) effect. The change of declination in the image is due to the fact that for meridian transit imaging phases are compensated for HA=0°. As Sun moves in the grating lobe the phase difference it generates between any interferometer pair changes. Therefore in the image it will appear to change declination to compensate for that phase. The RAs and declinations which could be affected by the Galactic plane when all the baselines are used for imaging, are summarized in Table 3.4.

To estimate the effect of the Sun in the images, we calculate the strength of Sun as visible to MRT when observations fall between the sunrise and the sunset. The effect is more pronounced when the Sun is in grating lobes during solar maximum and when its

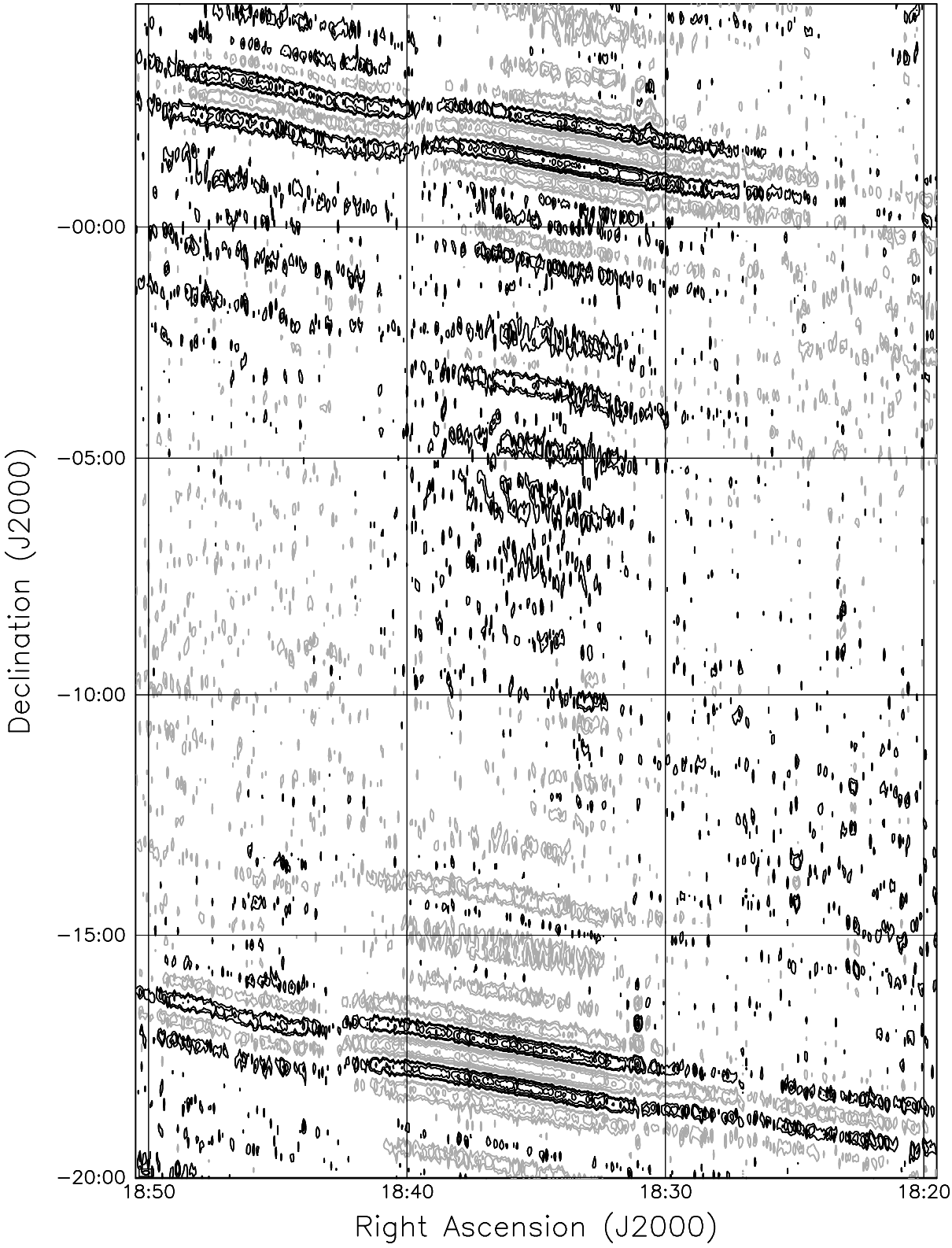



Fig. 3.4: The artifacts caused by the Sun transiting the grating response of the array in one allocation's image. The true position of the Sun on the day of observation of the visibility (JD2450529; date 22:03:1997), is RA=00:03hrs, declination=-0.°4 and its total flux is ≈2.9 sfu. The expected position of the grating lobes in the image is at RA ≈18:37 hrs, Declination ≈-17.°3. The image clearly shows the grating lobe response at this expected position. It also shows the aliased image at nearly the same RA of the grating response and declination≈1°. The strong artifacts corrupt the entire image. The contour levels are -40, -28, -20, -14, -10, -7, -5, -3, 2, 3, 5, 7, 10, 14, 20, 28, 40, 56, 80, 112 $\times\sigma$ where $\sigma$ is the rms noise (arbitrary units) in the image. The negative contours are represented in grey while the positive contours are represented in black.

| RA (B1950) when Galactic plane in $-10^\circ < \delta < 10^\circ$ | RAs (B1950) where this would corrupt the images. | Declination range (B1950) corrupted. |
|---|---|---|
| 06h28m to 07h10m | 01h00m to 01h42m & 11h56m to 12h38m | $-27^\circ \leq \delta \leq -8^\circ$ $-27^\circ \leq \delta \leq -8^\circ$ |
| 18h28m to 19h10m | 23h56m to 00h38m & 13h00m to 13h42m | $-27^\circ \leq \delta \leq -8^\circ$ $-27^\circ \leq \delta \leq -8^\circ$ |

Table 3.4: Table giving right ascensions and declinations which could be affected by the Galactic plane in the grating response (courtesy (Sachdev, 1999)).

declination is close to 0°. For a source at declination $\delta$=0°, a one dimensional cut of the fractional attenuation due to combined effect of interferometric beam pattern and the helix beam shape as a function of HA from the meridian is shown in Fig 3.5. We notice the enhanced response of the array when the hour angle is in the range ±5-6 hours.

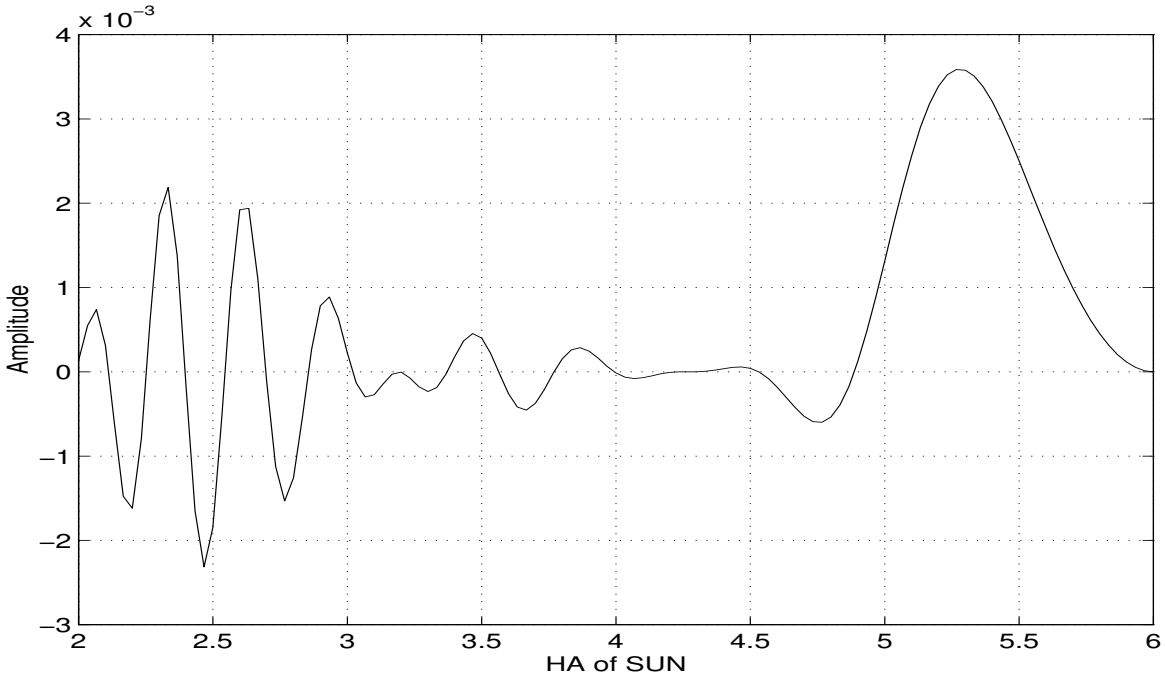



Fig. 3.5: Fractional attenuation of a source at $\delta = 0^\circ$ as a function of its HA from the meridian due to the response of interferometer formed by EW×NS and the attenuation due to helix primary beam response. The increase in the response due to grating lobes is easily seen around HA of 5.25 sidereal hour.

We obtained the daily averaged total solar continuum solar flux from the observations by the solar radio patrol group at Torun Centre for Astronomy NCU[12] (private communication Grazyna Gawronska) at 127 MHz. This daily averaged total solar flux at 127 MHz is shown in Fig. 3.6 from October, 1994 to February, 1999. The total solar flux at 151.5 MHz

[12]http://www.astro.uni.torun.pl/~gg/

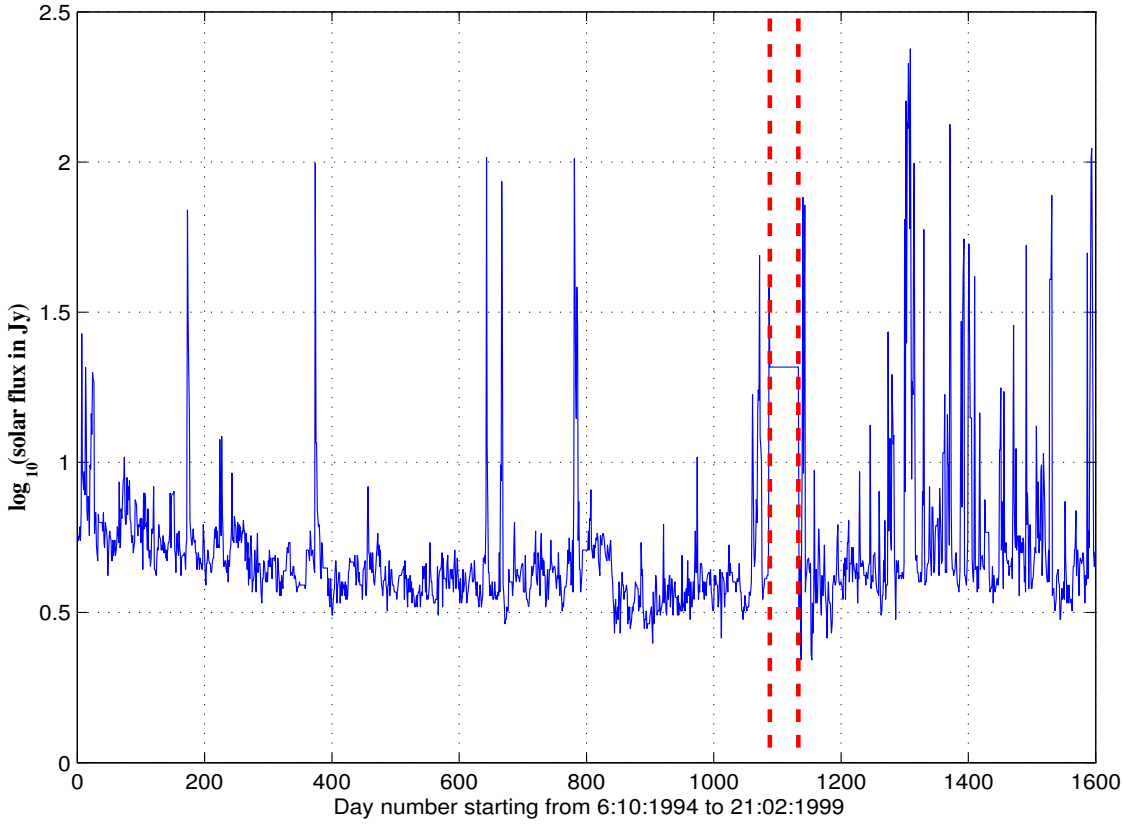



Fig. 3.6: The daily averaged total solar flux at 127 MHz from 6:10:1994 to 21:02:1999. The vertical dotted lines indicate the duration for which the solar flux observations were not available. The solar flux varies from day to day as shown and there are occasions of solar activity when the flux varies by a factor of up to 50.

was estimated by assuming a black body spectrum for Sun. Using the total solar flux, the two dimensional response of the array, position of the Sun on the day of observation, the resolution of the array for the allocation concerned, we estimate the maximum strength of the Sun which will affect the observed visibility. This is used to assign the $QF_4$ based on the effect of Sun on visibilities.

Data recorded during the night time is assigned a $QF_4$ of 1. When the data is observed during the twilight zone (See Sec. 3.5.1) it is assigned a $QF_4$ of 2. For the data observed during the day, the strength of the Sun as visible to MRT during observations is estimated. The assignment of $QF_4$ is based on the ratio of the maximum strength of the Sun as visible to the MRT to the noise expected in each allocation's image ($\approx$1 Jy) and is decided as specified in Table 3.5.

Our analysis revealed that contrary to our concerns there is a good fraction of data observed in the day time on which the effect of Sun is within acceptable limits (See Sec. 3.7). This generally happens when the HA of the Sun is within the range $\pm$3-5 hours and its declination is outside the range $\pm 10^\circ$.

| S.No. | R | $QF_4$ |
|---|---|---|
| 1 | Night time obsvn | 1 |
| 2 | Twilight zone obsvn | 2 |
| 3 | 0<R≤ 1 | 3 |
| 4 | 1<R≤ 2 | 4 |
| 5 | 2<R≤ 3 | 5 |
| 6 | 3<R≤ 5 | 6 |
| 7 | 5<R≤ 10 | 7 |
| 8 | 10<R≤ 20 | 8 |
| 9 | R > 20 | 9 |

Table 3.5: Guidelines for assignment of quality factor based on effect of the Sun in the visibilities. R=$\frac{F_\odot}{\sigma_{theo}}$, is the ratio of the strength of Sun as visible to the array ($F_\odot$) to the expected noise ($\sigma_{theo}$) in the image of an allocation.

### 3.4.2 Quality of a data or calibration file

The overall quality of a data or a calibration file is a function of the individual key parameters. It is obvious to represent the overall quality of a data file or a calibration file by combining the individual QFs based on the key parameters. There are several possible ways in which they can be combined and owing to no other additional information available the simplest way is to add them with equal weights. This has the implicit assumption that the key parameters are orthogonal (no cross terms) and are normalized in quality space. We have attempted to ensure this during the QF assignment process that the limits of the QFs within which a file is considered acceptable for imaging is same for all the key parameters (except for completeness whose QF limits for a file being acceptable is decided by our insistence to use only files which are complete for one sidereal hour for imaging).

Thus the overall quality factor of the data file $QF_{data}$ and the calibration file $QF_{calib}$ is a simple average of the QFs based on individual parameters.

### 3.4.3 The Total Quality Factor

The TQF represents comprehensively the quality of the combination of a data file and its calibration file, which decides if a data file is acceptable for imaging. The TQF is a simple average of $QF_{data}$ and $QF_{calib}$ which is given according to the Eqn. 3.1. Ideally the proximity of the calibrator to the observations of the data file should also be taken into account while calculating the TQF. We have found that the day to day rms variation per baseline in phase and amplitude is within ±10° and ±0.1 dB respectively. The rms variation of phases from one calibrator to another is within ±14° (see Sec. 5.2.3). Due to this system stability of the antennae amplitude and phases during uninterrupted observations, the proximity of the calibrator with the observed data file does not play a significant role and thus is not taken

into account while computing the TQF.

### 3.4.4 Thresholding

It is very obvious to expect that a visibility which ranks poorly based on some key parameters, should be avoided for imaging even if it has other QFs within acceptable limits. In order to take care of such cases thresholding is applied at the level of individual QFs ($QF_1$,$QF_2$,$QF_3$,$QF_4$), overall quality of the data file ($QF_{data}$) and its calibration file ($QF_{calib}$) and the total quality factor (TQF). The limits for thresholding are decided heuristically based on our experience with the MRT database. Although this does leave room for differences from person to person and hence inconsistency, it allows the subtle power of human vision to decide what 'should' be included as acceptable for imaging. It is to be noted that due to complete automatic nature of the algorithm it enables sufficient room for experimentation for judicious choice of the threshold limits. Having the histogram of the QFs displayed and realizing what threshold settings will give reasonable amount of data required for imaging, but at the same time ensuring that poor quality data does not pass through, has been the guiding principle behind fixing the thresholding limits. The first level of thresholding is applied as an upper cutoff on the QFs based on individual key parameters for both the data file and its calibration file as given below.

$$QF_1 \leq 2, \;\; QF_2 \leq 5, \;\; QF_3 \leq 5, \;\; QF_4 \leq 5 \tag{3.10}$$

This eliminates the possibility of a visibility file which ranks very poorly for one or more key parameters being selected for imaging. Once the $QF_{data}$ and $QF_{calib}$ are calculated for a data file and its calibration file according to Eqn. 3.1, a second level of thresholding is applied, for overall QF of the data file, $QF_{data}$ as well as for the calibration file $QF_{calib}$.

$$QF_{data} \leq 4, \;\; QF_{calib} \leq 4 \tag{3.11}$$

For a data file to be finally selected for imaging, the pair of data file and its calibration file which satisfy the first two levels of thresholding must meet an upper cutoff for TQF as given by[13].

$$TQF \leq 2.5 \tag{3.12}$$

[13]It is to be noted that Eqn. 3.11 and Eqn. 3.12 are not mutually independent. Eqn. 3.11 is strictly weaker than Eqn. 3.12, inasmuch as every assignment of values satisfying the latter also satisfies the former, but the converse is not true.

Only those pairs of data files along with their calibration file are considered for imaging which satisfy all the three conditions given in Eqn. 3.10, Eqn. 3.11 and Eqn. 3.12. In our experience we have found thresholding a vital tool and played an important role in eliminating such inconsistent cases, where a visibility ranked poorly in certain QFs but was within acceptable quality limits based on other QFs.

### 3.4.5 Prioritizing of visibility data

In order to rank a data file or a calibration file in decreasing order of quality, all the files which do satisfy the individual threshold limits according to Eqn. 3.10, are ranked in decreasing order of $QF_{data}$ or $QF_{calib}$. Those files which do not meet one or more thresholding criterion are ranked according to the following conditions in decreasing order of priority. These conditions have been decided based on our experience with the MRT data analysis.

1. Completeness threshold crossed by the visibility file.
2. Number of threshold limits crossed by the visibility file.
3. Interference threshold crossed by the visibility file.
4. Noise threshold crossed by the visibility file.
5. Effect of Sun threshold crossed by the visibility file.
6. $QF_{data}$ or $QF_{calib}$ of the visibility file.
7. The total number of interference points detected in the visibility file.

In order to rank combinations of different data files and their best calibration file, all the pairs which satisfy the thresholding criteria according to Eqn. 3.10 and Eqn. 3.11 are sorted in decreasing order of TQF for the combination. Those pairs which do not meet one or more thresholding criterion are ranked according to the following conditions in decreasing order of priority.

1. Completeness threshold crossed by the data file.
2. Completeness threshold crossed by calibration file.
3. Number of threshold limits crossed by the data file.
4. Number of threshold limits crossed by the calibration file.
5. Interference threshold crossed by the data file.
6. Noise threshold crossed by the data file.
7. Effect of Sun threshold crossed by the data file.
8. Interference threshold crossed by the calibration file.

9. Noise threshold crossed by the calibration file.
10. Effect of Sun threshold crossed by the calibration file.
11. The TQF of the combination.
12. Proximity between the observation of the calibration and the data file.
13. The total number of interference in the data file.

## 3.5 Automatic implementation

The heart of the data classification process is a program written in *Perl* programming language which accomplishes data classification automatically without any manual intervention. The approach described above is implemented in form of an algorithm in this standalone computer program whose flow chart is shown in Fig. 3.7. Perl was a natural choice due to its excellent text support, portability, speed and *pgperl* module to interface with PGPLOT graphics library.

It is important to note that the data classification software runs on the database of sum of magnitudes of visibilities on all the baselines rather than on the original visibility data, which helps to complete the entire data classification process faster. The program is first run to classify all the calibration files containing the observations of the three calibrators used. The procedure for assigning QFs to the calibration files is, as discussed earlier, similar to that of a data file except that only the duration relevant for calibration is considered for assigning the QFs and no TQF is calculated for calibration files. The classification process is carried out for the data files on a sidereal hour basis. For each one sidereal hour range, the program proceeds in ascending order of allocations and within each allocation, all the data files are processed one by one, in decreasing order of their completeness, i.e. to say that the files which are complete are taken up first, although this has in anyway, no bearing on the results of data classification process. Now we briefly describe the practical aspects of the various stages involved in data classification.

**The input files**: These act as the ground work required for data classification.

**Database of sum of magnitudes of visibilities:** A database consisting of *sum of magnitudes* of visibilities on all the baselines is generated for the visibility files in the entire survey data. This is an ascii file having sum of magnitudes and the corresponding sidereal time for each integration period.

**Database of interference:** Each file in the database of sum of magnitudes of visibilities on all the baselines is passed through RFI detection based on Fourier filtering and a flag table containing the RFI details is maintained.

**Noise information:** This file contains the expected noise as a function of RA.

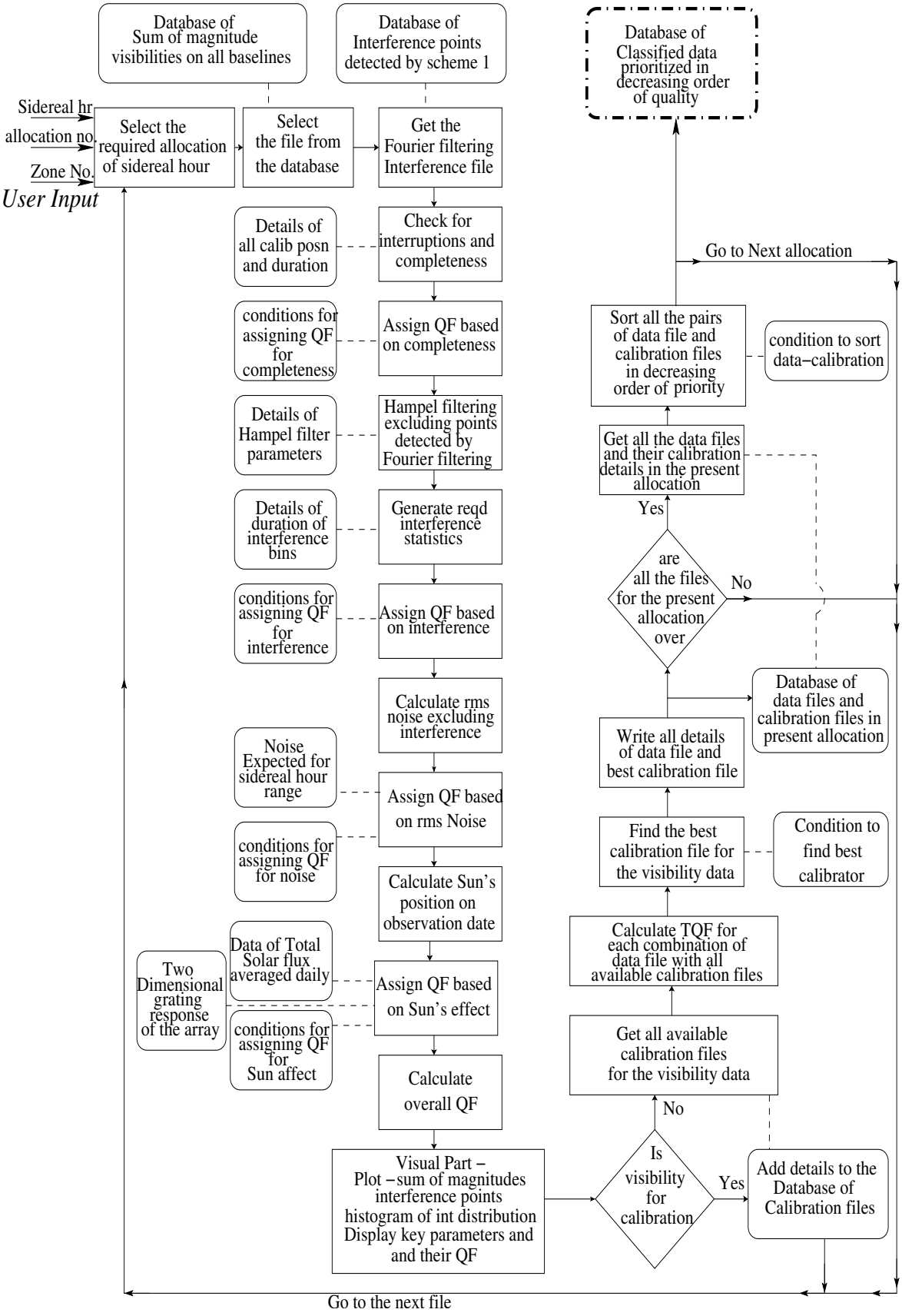



Fig. 3.7: Flow chart of the data classification program.

**Solar flux :** This ascii file contains the daily averaged total solar flux at 127 MHz for the entire span of ≈5 years of observation.

**Array response :** An ascii file containing the product of two dimensional grating response of the EW and NS groups and the 2-D primary response of the helix, as a function of hour angle and declination (used only for day time observations).

**Database of QFs for all the calibrators :** As described above in the first stage the program first classifies all the calibration files of the three calibrators used at MRT. The QFs of all the calibration files are estimated and a database containing all the required information including the QFs is prepared.

### 3.5.1 Calculation steps

The QFs are calculated as explained below.

**Calculation of $QF_1$ :** The first step is to check for any interruptions during the observations of the visibility file. The program calculates the percentage of visibility points present out of the total expected points in the file. Depending upon the % points present in the visibility files the program assigns the $QF_1$ accordingly as per specified conditions given in Table 3.1.

**Calculation of $QF_2$ :** For the visibility file under process, the corresponding file containing sum of magnitudes and the corresponding flag table file generated by the Fourier filtering are read. The visibility file is passed through a second stage of RFI detection using Hampel filtering after excluding the data points already detected by the Fourier filtering technique to detect occasionally remaining RFI, if any. This is accomplished via an external C program which is automatically invoked by the data classification program. Using all these interference points the program calculates the required statistics, namely the total number of interference points, the maximum duration of continuous interference and the number of times the continuous interference exists for more than HPBW of the full resolution synthesized images i.e. 16 sidereal seconds. A histogram showing the number of points in each interference bin is also generated (see Sec. 3.5.2). Subsequently the program computes $QF_2$ for data file or the calibrator as per specified conditions given in Table 3.2.

**Calculation of $QF_3$ :** The program calculates the rms noise in the measured visibilities using a moving box-car window after excluding the interference points. The ratio of the rms noise to the expected noise (using Eqn. 3.3) is used to compute the $QF_3$ as per specified conditions in Table 3.3.

**Calculation of $QF_4$ :** The program calculates the position ($RA_\odot$, $\delta_\odot$) of the Sun on the day of observation of the visibility file. The length of the day duration i.e. the sidereal hours elapsed between the sunrise and sunset ($H_d$), is calculated using the relation below (Gillard

and Holdway, 2004).

$$H_d = 2 \times 0.997 \times \frac{180}{\pi} \times \frac{1}{15} \times cos^{-1}\{-tan(\phi)\, tan(\delta_\odot)\} \tag{3.13}$$

where $\phi$ is the latitude of place (MRT) and the sunrise and sunset being defined at zenith distance 90°.5. The rise and set time of the Sun is calculated accordingly. The duration of twilight is calculated which is taken to occur between zenith distance of 90°.5 and 108°.

If the entire observation of the visibility file has been carried out in the night, such a file is assigned $QF_4$ of 1. In case some part or the entire observation of the visibility file has been carried out during twilight zone (but no part of the visibility should have been observed in the day), $QF_4$ is assigned a value of 2. In the case when some part or the entire observation for the visibility file has been carried out during the day time, the maximum effect of the Sun is calculated during the entire observation. For this, the daily averaged solar flux (at 151.5 MHz), the 2-D response of the array and the resolution depending upon the allocation number of the file under consideration are used to estimate the maximum strength of the Sun (or its artifacts) expected in the image. For the days when the daily averaged total solar flux is not available, it is estimated by interpolating the total solar flux on nearby days (overall there are about 200 days for which the daily averaged total solar flux is not available out of a total of about 1600 days). The strength of the Sun as visible to MRT array ($F_\odot$) for a given allocation is given by,

$$F_\odot = TF_\odot \times P_H \times R_{EW\times NS} \times \frac{\theta_{EW} \times \theta_{NS}}{\theta'_{EW} \times \theta'_{NS}} \tag{3.14}$$

where $TF_\odot$ is the total solar flux, $P_H$ is 2-D primary beam response of the helix, $R_{EW\times NS}$ is the 2-D response of the interferometer formed by EW and the NS arms, $\theta_{EW}$ and $\theta_{NS}$ are the angular resolution in RA and declination of the image of the allocation under consideration, while $\theta'_{EW}$ and $\theta'_{NS}$ being the radio size of the Sun along RA and declination. The $QF_4$ is assigned depending upon the ratio of the strength of the Sun as visible to MRT, to the expected noise in the image of the visibility file under progress, as per specified conditions in Table 3.5.

**Calculation of $QF_{data}$ or $QF_{calib}$ :** Once the quality factors based on the individual key parameters are calculated the overall quality of the data file ($QF_{data}$) or the calibration file ($QF_{calib}$) is given by a simple average of the individual QFs.

In case the file being processed is a calibration file, at this stage all the relevant details are written into the database of calibration files and the program jumps back to start to proceeds to the next calibration file. In case the file being processed is a data file, the program goes to the next step for calculation of TQF.

**Calculation of TQF :** Once the $QF_{data}$ is calculated for the data file the program finds out its all calibration files in the classified database of calibration files. When the observations have gone uninterrupted for more than one sidereal day, there may be more than one set of calibration files of each of the three calibrators used at MRT. The TQF is calculated separately for each pair i.e. combination of the data file with its each calibration file. All the calibration files for a given data file are sorted in decreasing order of quality according to the conditions specified in Sec. 3.4.5. The details of the data file and its all calibration files including the TQF for each combination are stored in a temporary database for the allocation under process and the next data file of the same allocation is taken up.

### 3.5.2 Visualization aid

The data classification program also provides visualization capabilities so that in case of any doubt the user is able to check all the details about a visibility file and satisfy himself about the correctness of the algorithm and its implementation in the program and may incorporate suitable changes if desired. Apart from this the visualization process helps the user to scan the entire database and gives him a feel about the quality of data in the database. It is to be noted that this visualization process is optional and the program can be executed in interactive or non-interactive mode. In the interactive mode the user can halt the program at any desired step and check different parameters and satisfy himself, while in the non-interactive mode he simply may or may not prefer to glance at the data while the program runs till completion. The program provides the following graphical displays.

a. Sum of magnitudes of visibilities on all the baselines as a function of RA as shown in Fig. 3.8.

b. Histogram of the number of interference points detected in different duration bins as shown in Fig. 3.9.

The visual inspection also helps in finding out how effective is the interference detection scheme. Few important details are also shown in the display windows including QFs.

### 3.5.3 The classified visibility database

Once the program has processed all the data files in the present allocation, it reads from the temporary database (for the allocation under process), details of each data file and its best quality calibration file. All the pairs of a data file and its best calibration file which satisfy the threshold criteria according to the Eqn. 3.10, Eqn. 3.11 and Eqn. 3.12 are sorted

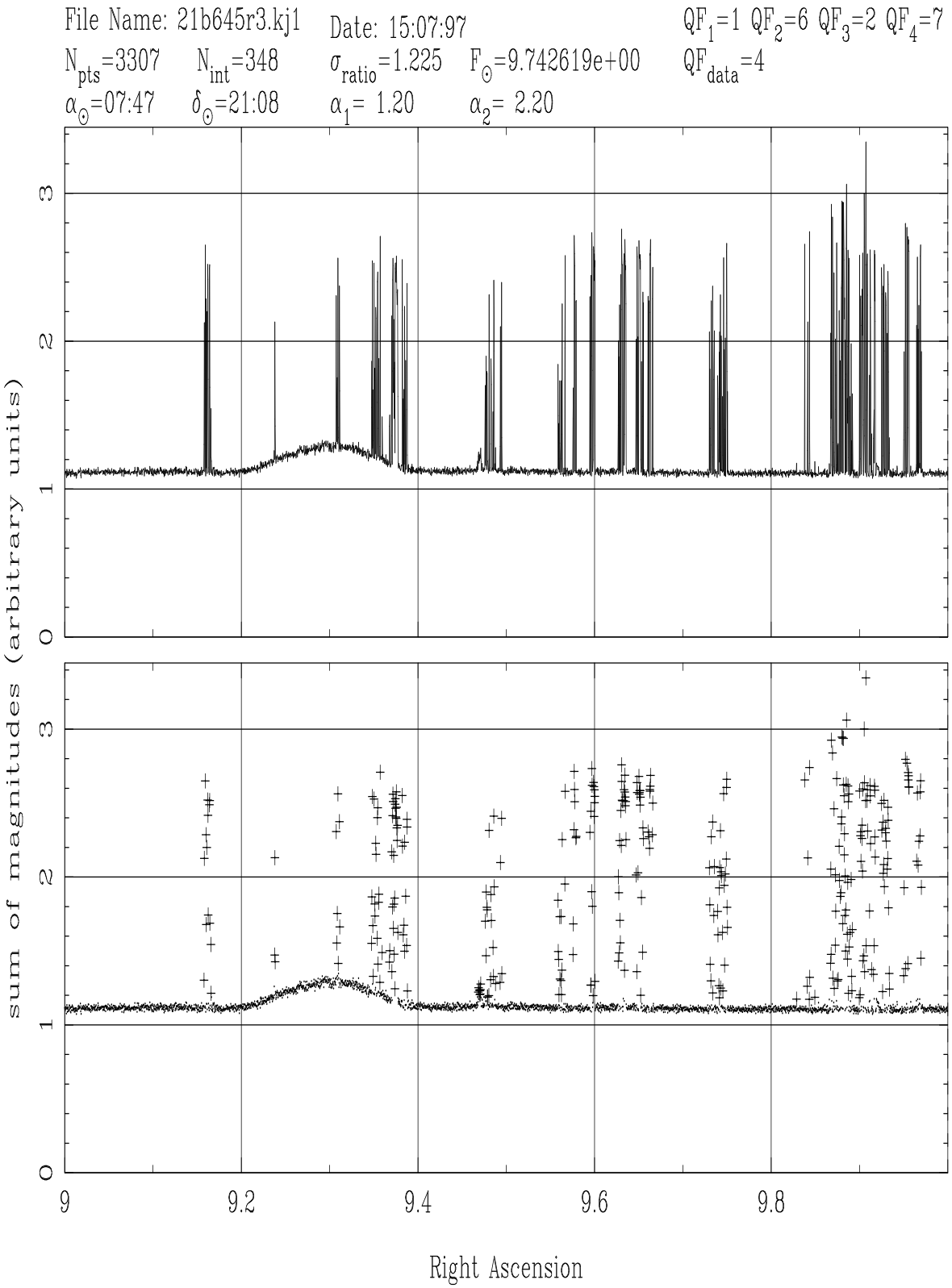



Fig. 3.8: Sum of magnitudes of visibilities on all the baselines as a function of RA for a data file. In the lower plot points detected as interference are shown by + while points not effected by interference are shown by dots. The comparison helps to ascertain that if the interference detection has been appropriate. Important details including the QFs are also shown in the plot.

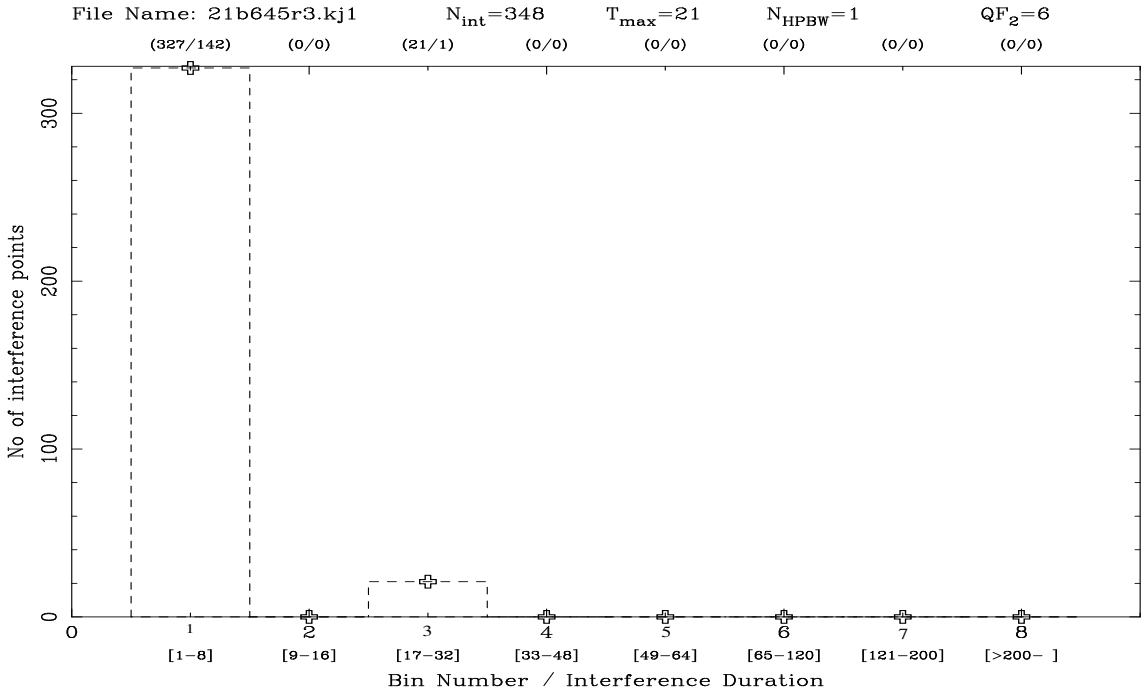



Fig. 3.9: Interference statistics of a visibility data (sum of magnitudes of visibilities shown in Fig. 3.8). On the X-axis the bin numbers and the continuous duration they represent is shown in brackets. On Y-axis the number of interference points detected in the respective bins are shown. The number of points and the number of times the continuous interference is detected in the respective bins is also shown just above each bin number within braces. $N_{HPBW}$ is the number of times there is continuous interference lasting for more than 16 sidereal seconds (HPBW of the array in RA). $T_{max}$ is the maximum continuous duration for which interference has been detected in this visibility file.

in ascending order of TQF. The pairs which do not satisfy one or more thresholding criteria are ranked according to the conditions specified in Sec. 3.4.5.

After all the pairs comprising of each data file along with its best calibration file are sorted in decreasing order of priority, the next step is to store the information into classified database of visibilities. All the data files are picked up one after another in the same order after they have been sorted in decreasing order of priority. From the temporary database, details of its all available calibration files are written down one after another. It is to be noted that all the calibration files of each data file have already been sorted in decreasing order of quality. At the end, the classified visibility database has information of all the data files along with their calibration files, classified, ranked and sorted in decreasing order of suitability for imaging, for the sidereal hour and allocation under process. The temporary database is deleted and the next allocation is processed. On the same lines all the allocations are processed one after another in increasing order of allocation.

Once all the allocations for the sidereal hour under consideration are processed we have the classified database of the visibilities, for the sidereal hour and delay zone under

consideration. The sample output of the classified database is shown in Table 3.6. The classified database is also available as a matrix, which can be imported in Matlab and queried using simple statements for analysis. In order to select good data for imaging the user can easily select one or more files from the database which pass the acceptance criteria.

| FNo. | File Name | $N_{tot}$ | $N_{int}$ | $\sigma_{ratio}$ | $F_{\odot}$ | $RA_{\odot}$ | $\delta_{\odot}$ | $QF_1$ | $Th_1$ | $QF_2$ | $Th_2$ | $QF_3$ | $Th_3$ | $QF_4$ | $Th_4$ | QF | $F$ | TQF | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***A*llocation : 31** | | | | | | | | | | | | | | | | | | | |
| 466 | 31a192r1.ms3 | 3306 | 0 | 1.01 | 0.000 | 01:46 | 11:01 | 1 | p | 1 | p | 2 | p | 1 | p | 1.25 | | | |
| 1 | 31a192r1.dj1 | 446 | 0 | 1.15 | 0.000 | 01:46 | 11:01 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.125 | y |
| 467 | 31d686r1.ss3 | 3307 | 2 | 1.07 | 0.000 | 10:17 | 10:38 | 1 | p | 2 | p | 2 | p | 1 | p | 1.5 | | | |
| 1 | 31d686r1.tt3 | 636 | 0 | 1.20 | 0.000 | 10:17 | 10:38 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.25 | y |
| 2 | 31d686r1.ww1 | 655 | 2 | 1.18 | 0.000 | 10:17 | 10:38 | 1 | p | 2 | p | 1 | p | 1 | p | 1.25 | 0 | 1.375 | y |
| 3 | 31d686r1.jj1 | 457 | 17 | 1.49 | 2.739 | 10:17 | 10:38 | 1 | p | 3 | p | 3 | p | 5 | p | 3 | 0 | 2.25 | y |
| 468 | 31a187r2.ns3 | 3305 | 0 | 1.05 | 0.000 | 01:31 | 09:36 | 1 | p | 1 | p | 2 | p | 2 | p | 1.5 | | | |
| 1 | 31a187r2.ej1 | 447 | 0 | 1.22 | 0.000 | 01:31 | 09:36 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.25 | y |
| 2 | 31a187r1.gj1 | 496 | 0 | 1.12 | 0.000 | 01:28 | 09:15 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.25 | y |
| 3 | 31a187r1.qt3 | 446 | 9 | 1.20 | 0.885 | 01:28 | 09:15 | 1 | p | 2 | p | 1 | p | 3 | p | 1.75 | 0 | 1.625 | y |
| 4 | 31a187r1.tw1 | 460 | 6 | 1.18 | 0.354 | 01:28 | 09:15 | 1 | p | 2 | p | 1 | p | 3 | p | 1.75 | 0 | 1.625 | y |
| 5 | 31a187r2.ot3 | 634 | 16 | 1.35 | 0.000 | 01:31 | 09:36 | 1 | p | 3 | p | 2 | p | 2 | p | 2 | 0 | 1.75 | y |
| 469 | 31d192r1.ss3 | 3304 | 4 | 1.04 | 0.000 | 01:46 | 11:01 | 1 | p | 2 | p | 2 | p | 1 | p | 1.5 | | | |
| 1 | 31d192r1.jj1 | 443 | 0 | 1.05 | 0.000 | 01:46 | 11:01 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.25 | y |
| 2 | 31d192r1.tt3 | 636 | 2 | 1.10 | 0.000 | 01:46 | 11:01 | 1 | p | 2 | p | 1 | p | 2 | p | 1.5 | 0 | 1.5 | y |
| 3 | 31d192r1.ww1 | 461 | 15 | 1.22 | 0.306 | 01:46 | 11:01 | 1 | p | 3 | p | 1 | p | 3 | p | 2 | 0 | 1.75 | y |
| 470 | 31a688r1.xs3 | 3307 | 25 | 1.65 | 0.000 | 10:24 | 09:57 | 1 | p | 2 | p | 3 | p | 1 | p | 1.75 | | | |
| 1 | 31a688r1.at3 | 636 | 0 | 1.21 | 0.000 | 10:24 | 09:57 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.375 | y |
| 2 | 31a688r2.bw1 | 461 | 0 | 1.22 | 0.000 | 10:28 | 09:35 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.375 | y |
| 3 | 31a688r1.yt3 | 636 | 1 | 1.55 | 0.000 | 10:24 | 09:57 | 1 | p | 2 | p | 3 | p | 1 | p | 1.75 | 0 | 1.75 | y |
| 4 | 31a688r1.dw1 | 461 | 1 | 1.60 | 0.000 | 10:24 | 09:57 | 1 | p | 2 | p | 3 | p | 1 | p | 1.75 | 0 | 1.75 | y |
| 5 | 31a688r2.mj1 | 449 | 100 | 1.54 | 7.611 | 10:28 | 09:35 | 1 | p | 7 | f | 3 | p | 7 | f | 4.5 | 2 | 3.125 | n |
| 6 | 31a688r1.oj1 | 449 | 140 | 1.25 | 5.859 | 10:24 | 09:57 | 1 | p | 9 | f | 2 | p | 7 | f | 4.75 | 2 | 3.25 | n |
| 471 | 31d186r1.ps3 | 3304 | 6 | 1.09 | 0.000 | 01:24 | 08:53 | 1 | p | 2 | p | 2 | p | 2 | p | 1.75 | | | |
| 1 | 31d186r1.gj1 | 448 | 0 | 1.12 | 0.000 | 01:24 | 08:53 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.375 | y |
| 2 | 31d186r1.tw1 | 460 | 0 | 1.08 | 0.342 | 01:24 | 08:53 | 1 | p | 1 | p | 1 | p | 3 | p | 1.5 | 0 | 1.625 | y |
| 3 | 31d186r1.qt3 | 635 | 12 | 1.15 | 1.343 | 01:24 | 08:53 | 1 | p | 2 | p | 1 | p | 4 | p | 2 | 0 | 1.875 | y |
| 472 | 31c208r1.bs3 | 3304 | 1 | 1.01 | 0.000 | 02:47 | 16:07 | 1 | p | 2 | p | 2 | p | 1 | p | 1.5 | | | |
| 1 | 31c208r1.ct3 | 636 | 0 | 1.23 | 0.000 | 02:47 | 16:07 | 1 | p | 1 | p | 2 | p | 1 | p | 1.25 | 0 | 1.375 | y |
| 2 | 31c208r1.qj1 | 449 | 2 | 1.10 | 0.000 | 02:47 | 16:07 | 1 | p | 2 | p | 1 | p | 2 | p | 1.5 | 0 | 1.5 | y |
| 3 | 31c208r1.fw1 | 461 | 0 | 1.04 | 0.301 | 02:47 | 16:07 | 1 | p | 1 | p | 1 | p | 3 | p | 1.5 | 0 | 1.5 | y |
| 473 | 31a187r1.ps3 | 3304 | 3 | 1.05 | 0.000 | 01:28 | 09:15 | 1 | p | 2 | p | 2 | p | 2 | p | 1.75 | | | |
| 1 | 31a187r2.ej1 | 448 | 0 | 1.22 | 0.000 | 01:31 | 09:36 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.375 | y |
| 2 | 31a187r1.gj1 | 448 | 0 | 1.12 | 0.000 | 01:28 | 09:15 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.375 | y |
| 3 | 31a187r1.qt3 | 635 | 13 | 1.20 | 0.858 | 01:28 | 09:15 | 1 | p | 2 | p | 1 | p | 3 | p | 1.75 | 0 | 1.75 | y |
| 4 | 31a187r1.tw1 | 461 | 9 | 1.18 | 0.305 | 01:28 | 09:15 | 1 | p | 2 | p | 1 | p | 3 | p | 1.75 | 0 | 1.75 | y |

Continued on next page...

Table 3.6 – Continued

| FNo. | File Name | $N_{tot}$ | $N_{int}$ | $\sigma_{ratio}$ | $F_{\odot}$ | $RA_{\odot}$ | $\delta_{\odot}$ | $QF_1$ | $Th_1$ | $QF_2$ | $Th_2$ | $QF_3$ | $Th_3$ | $QF_4$ | $Th_4$ | QF | $F$ | TQF | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 31a187r2.ot3 | 634 | 22 | 1.35 | 0.000 | 01:31 | 09:36 | 1 | p | 3 | p | 2 | p | 2 | p | 2 | 0 | 1.875 | y |
| 474 | 31c190r1.es3 | 3303 | 1 | 1.00 | 0.000 | 01:39 | 10:19 | 1 | p | 2 | p | 1 | p | 2 | p | 1.5 | | | |
| 1 | 31c190r1.ft3 | 636 | 0 | 1.14 | 0.000 | 01:39 | 10:19 | 1 | p | 1 | p | 1 | p | 2 | p | 1.25 | 0 | 1.375 | y |
| 2 | 31c190r1.iw1 | 461 | 40 | 1.08 | 0.385 | 01:39 | 10:19 | 1 | p | 5 | p | 1 | p | 3 | p | 2.5 | 0 | 2 | y |
| 475 | 31a184r1.ks3 | 3301 | 21 | 0.98 | 0.000 | 01:17 | 08:09 | 1 | p | 2 | p | 1 | p | 2 | p | 1.5 | | | |
| 1 | 31a184r1.bj1 | 447 | 0 | 1.29 | 0.000 | 01:17 | 08:09 | 1 | p | 1 | p | 2 | p | 1 | p | 1.25 | 0 | 1.375 | y |
| 2 | 31a184r1.lt3 | 637 | 11 | 1.13 | 2.933 | 01:17 | 08:09 | 1 | p | 2 | p | 1 | p | 5 | p | 2.25 | 0 | 1.875 | y |
| 3 | 31a184r1.ow1 | 462 | 120 | 1.28 | 1.356 | 01:17 | 08:09 | 1 | p | 7 | f | 2 | p | 4 | p | 3.5 | 1 | 2.5 | n |
| 476 | 31c186r1.bs3 | 3303 | 1 | 1.04 | 0.000 | 01:24 | 08:53 | 1 | p | 2 | p | 2 | p | 2 | p | 1.75 | | | |
| 1 | 31c186r1.ct3 | 636 | 0 | 1.23 | 1.343 | 01:24 | 08:53 | 1 | p | 1 | p | 2 | p | 4 | p | 2 | 0 | 1.875 | y |
| 477 | 31c191r1.cs3 | 3305 | 0 | 1.05 | 0.000 | 01:43 | 10:40 | 1 | p | 1 | p | 2 | p | 2 | p | 1.5 | | | |
| 1 | 31c191r1.dt3 | 636 | 50 | 1.15 | 0.000 | 01:43 | 10:40 | 1 | p | 5 | p | 1 | p | 2 | p | 2.25 | 0 | 1.875 | y |
| 2 | 31c191r1.gw1 | 461 | 58 | 1.18 | 0.380 | 01:43 | 10:40 | 1 | p | 6 | f | 1 | p | 3 | p | 2.75 | 1 | 2.125 | n |
| 478 | 31b686r1.ds3 | 3306 | 13 | 1.84 | 0.000 | 10:17 | 10:38 | 1 | p | 4 | p | 4 | p | 1 | p | 2.5 | | | |
| 1 | 31b686r1.et3 | 635 | 240 | 2.02 | 0.000 | 10:17 | 10:38 | 1 | p | 9 | f | 5 | p | 1 | p | 4 | 1 | 3.25 | n |
| 479 | 31a687r1.ss3 | 3306 | 11 | 1.14 | 0.000 | 10:20 | 10:18 | 1 | p | 2 | p | 2 | p | 1 | p | 1.5 | | | |
| 1 | 31a687r1.jj1 | 447 | 32 | 1.66 | 4.835 | 10:20 | 10:18 | 1 | p | 5 | p | 3 | p | 6 | f | 3.75 | 1 | 2.625 | n |
| 480 | 31c209r1.bs3 | 3302 | 0 | 1.08 | 0.000 | 02:51 | 16:24 | 1 | p | 1 | p | 2 | p | 1 | p | 1.25 | | | |
| 1 | 31c209r1.ct3 | 570 | 0 | 1.02 | 0.000 | 02:51 | 16:24 | 4 | f | 1 | p | 1 | p | 1 | p | 1.75 | 1 | 1.5 | n |
| 481 | 31c210r1.as3 | 3300 | 1 | 1.02 | 0.000 | 02:55 | 16:41 | 1 | p | 2 | p | 2 | p | 1 | p | 1.5 | | | |
| 1 | 31c210r1.bt3 | 583 | 0 | 1.02 | 0.000 | 02:55 | 16:41 | 3 | f | 1 | p | 1 | p | 1 | p | 1.5 | 1 | 1.5 | n |
| 482 | 31b185r1.ks3 | 3300 | 0 | 1.02 | 0.000 | 01:20 | 08:31 | 1 | p | 1 | p | 2 | p | 2 | p | 1.5 | | | |
| ***Allocation : 32*** | | | | | | | | | | | | | | | | | | | |
| 483 | 32a195r1.ls3 | 3308 | 2 | 1.05 | 0.000 | 01:57 | 12:03 | 1 | p | 2 | p | 2 | p | 1 | p | 1.5 | | | |
| 1 | 32a195r1.cj1 | 446 | 0 | 1.10 | 0.000 | 01:57 | 12:03 | 1 | p | 1 | p | 1 | p | 1 | p | 1 | 0 | 1.25 | y |
| 2 | 32a195r1.mt3 | 507 | 0 | 1.28 | 0.000 | 01:57 | 12:03 | 6 | f | 1 | p | 2 | p | 2 | p | 2.75 | 1 | 2.125 | n |
| . | | | | | | | | | | | | | | | | | | | |

Table 3.6: A sample of the classified database produced by the program for one allocation (31) of visibility files corresponding to the sidereal hour 18 to 19 hrs.

[14]

[14]The abbreviations used in Table 3.6 are : $N_{tot}$ is the total number of points in the visibility file for relevant duration, $N_{int}$ is the number of interference points detected, $\sigma_{ratio}$ is the ratio of measured noise to the expected noise, $F_{\odot}$ is the strength of Sun (Jy) as visible to MRT array, $RA_{\odot}$ and $\delta_{\odot}$ is the RA and declination of the Sun on the day of observation, $QF_1$, $QF_2$, $QF_3$, $QF_4$ are the QFs and $Th_1$, $Th_2$, $Th_3$, $Th_4$ convey whether the file passes the threshold limits for the individual key parameters (p-pass, f-fail) and QF is the overall QF of the data or the calibration file, F is the number of threshold limits failed for the combination of data file and its calibration file, TQF is the total quality factor and S shows finally a combination of data and calibration file is considered suitable for imaging or not (y/n).

| Class | Traditional approach (Number) | Automatic approach (Number) | Common Files (Number) | agreement (% ) |
|---|---|---|---|---|
| a | 105 | 115 | 92 | 80 |
| b | 57 | 54 | 38 | 70 |
| c | 210 | 203 | 192 | 95 |
| Total | 372 | 372 | 322 | 87 |

Table 3.7: Comparison between the manual approach and automatic data classification scheme when visibility files are classified in different groups (a) Good data (b) Poor data and (c) Bad data on a total of 372 visibility files. The agreement is given as the % of files in automatic scheme which are also classified by the traditional scheme in the same class. The comparison reveals that the human and automatic classification agree in more than 87% of the cases.

## 3.6 Comparison with the traditional scheme

To compare our results and check the correctness of the data classification algorithm, there does not exist any unique benchmark. Nevertheless, the results must agree to a good extent with that of the traditional approach. We compared our results with that of traditional approach based on human expertise using two schemes on about a total of 600 visibility files. Both the schemes for comparison and results are described below.

**Scheme - I :** In the first scheme each data file was classified in three categories namely, (a) Good data - definitely acceptable for imaging (b) Poor data - which is doubtful and (c) Bad data - which can definitely be rejected.

A total of 372 files were selected randomly from the MRT database. The automatic classification scheme was applied to the data files and suitable range of QFs were chosen to represent the three categories as above. All the data files were also classified in the three categories based on experienced human expertise. The results of comparison from the two approaches are shown in Table 3.7. The human and automatic classification agree in 87% of the cases.

**Scheme - II :** In this scheme the files were ranked in relative order of quality. We selected a total of 226 files from five allocations 5, 15, 30, 45 and 60 for data belonging to each of the four sidereal hour ranges 2-3 hrs, 6-7 hrs, 16-17 hrs and 23-24 hrs. All the files belonging to these allocations in the four sidereal hour ranges mentioned above were considered. Files within each allocation of a sidereal hour range were relatively ranked starting from 1 in decreasing order of quality using traditional approach. The relative rankings of the traditional approach were compared with the results of the automatic classification scheme.

The results of the comparison of the two approaches are shown in Fig. 3.10 and Fig. 3.11. Fig. 3.10 shows the relative rankings for files within each allocation of a sidereal hour range, as given by the traditional approach (denoted by circles) and the auto-

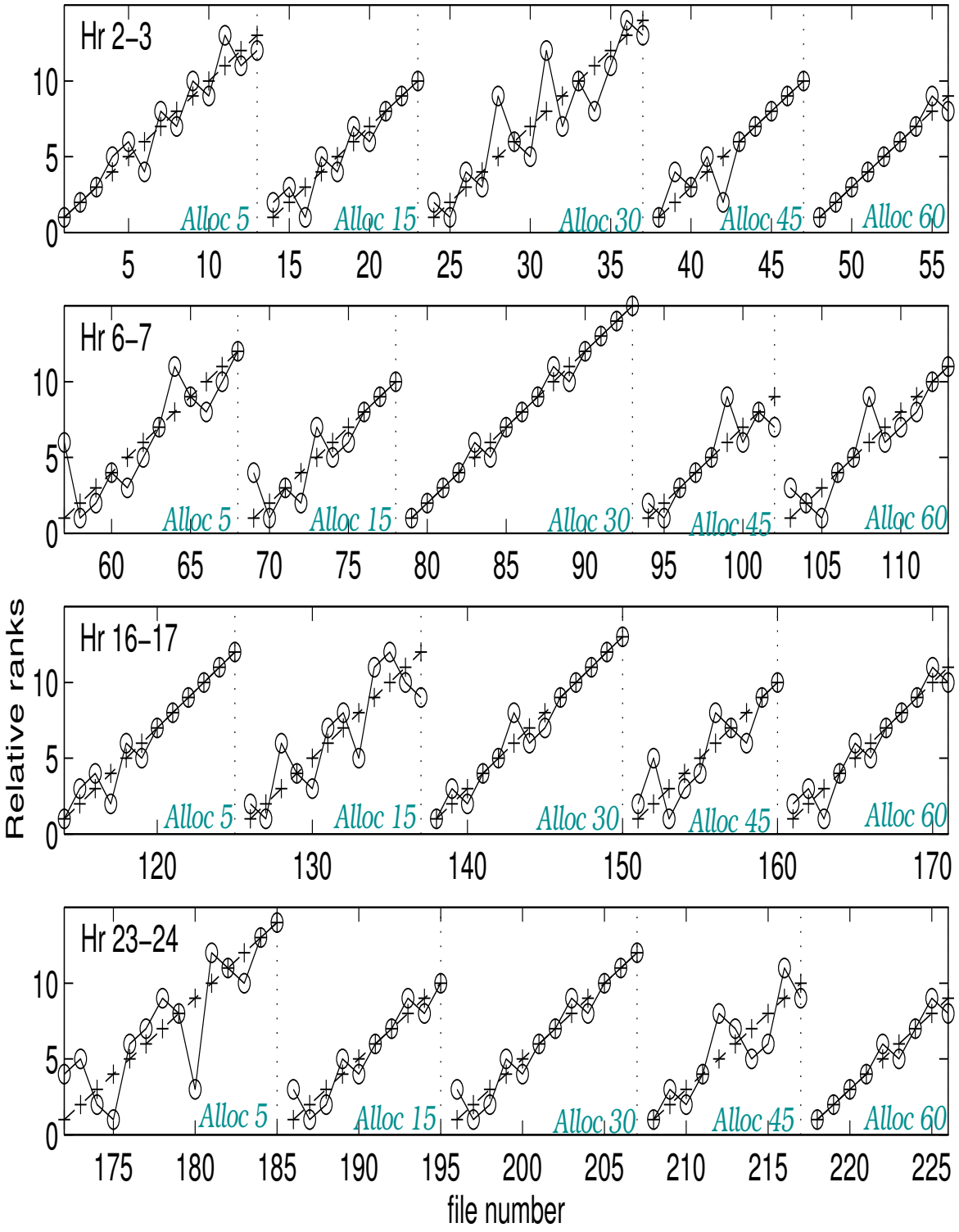



Fig. 3.10: The relative ranks within each allocation (5,15,30,45,60) for the visibility data files corresponding to the sidereal hours 2-3 hrs, 6-7 hrs, 16-17 hrs and 23-24 hrs as given by manual (marked as o) and automatic data classification (marked as +) on a total of 226 data files. The boundaries for each allocation is marked by vertical dotted lines. There is a good agreement between the manual and automatic ranking and there are only four cases for which the difference in relative approaches exceed ±3.

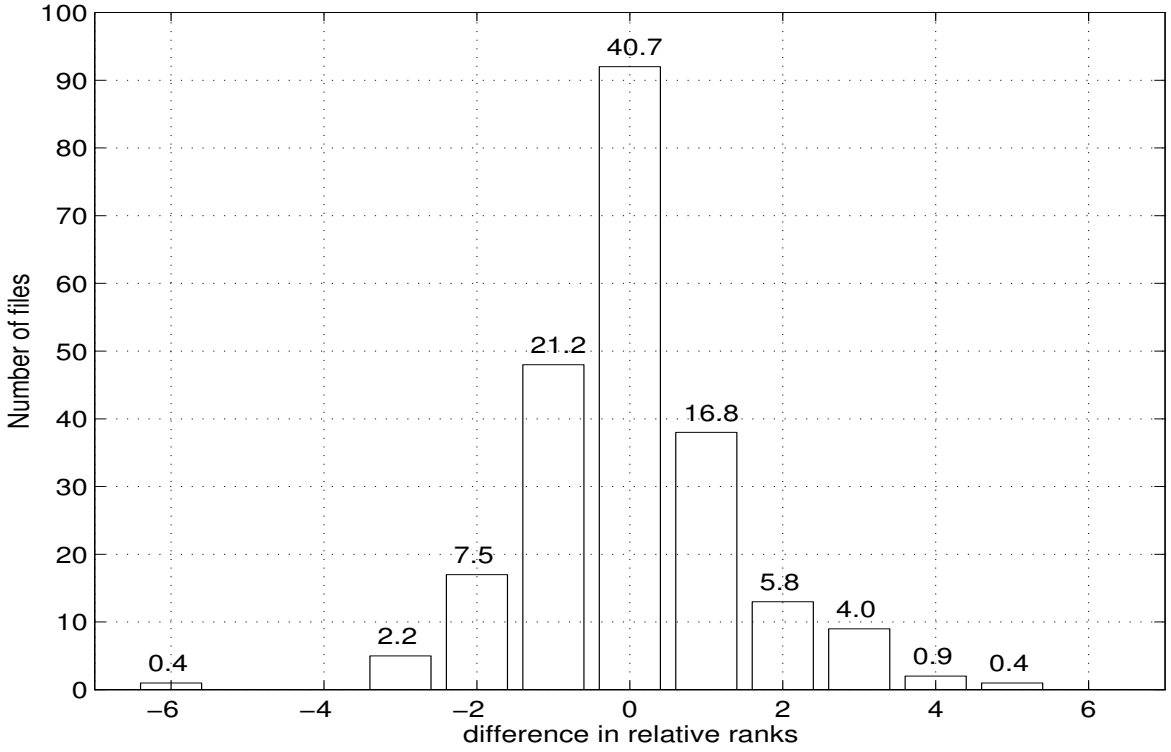



Fig. 3.11: Histogram of the difference in relative ranks given by the manual approach and automatic data classification. The % of the files falling within each bin is also indicated on the top of each bar. On an average, within a difference in the relative ranks of ±1, the agreement is 80%. The agreement increases to 92% when the difference in the relative ranks is considered within ±2.

matic classification scheme (denoted by +). The difference in relative ranks shows that on an average there is a good agreement and there are no large differences in the relative ranks by the two approaches. Fig. 3.11 summarizes the results in the form of a histogram of the difference in relative ranks between the traditional approach and the automatic data classification algorithm. On an average within a difference in the relative ranks of ±1 the agreement is 80%. The agreement increases to 92% when the difference in the relative ranks is considered within ±2. On an average there are 11-12 files (maximum 15) in each allocation for a sidereal hour range and an agreement of the relative ranks within ±2 can be considered as good. There are only four cases when the difference in relative ranks is outside ±3.

A careful analysis of 18 discrepant cases for which the difference in relative ranks exceeded ±2 was carried out. Our analysis revealed that for 9 such cases the automatic ranks were more appropriate. For 6 cases the ranking by the traditional approach was more appropriate as these files were effected by DC level shifts which is presently not taken into account in the automatic classification scheme. For 2 cases it was difficult to decide which of the two rankings was more appropriate. For 1 file there was interference due to satellite

which was not detected by the interference detection scheme and hence the traditional ranking was more appropriate.

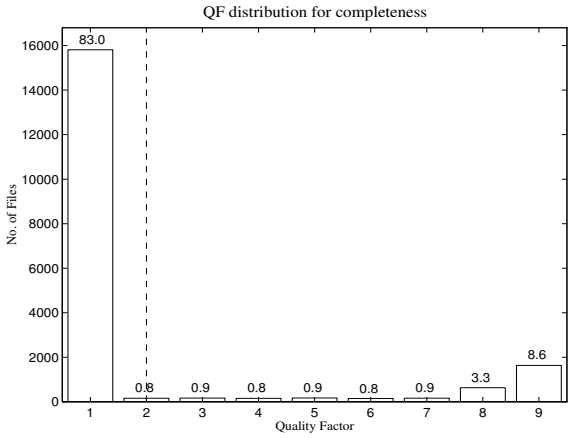
The comparison based on both the schemes clearly indicates that the automatic classification is objective, efficient and reliable. Even among the discrepant cases it is generally more appropriate. Thus the automatic data evaluation scheme can be successfully used for selecting good data for imaging.

## 3.7 Results

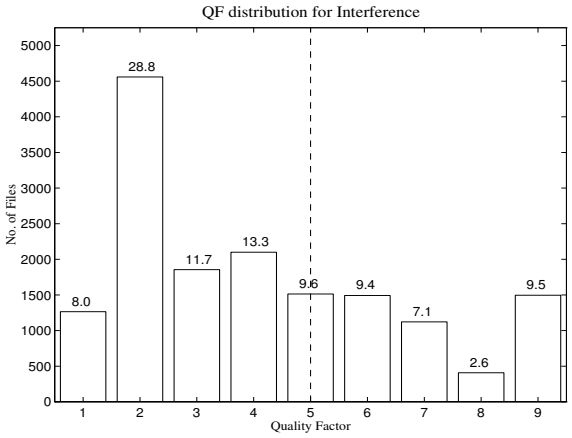
The data classification algorithm was used to classify the entire visibility database of MRT survey. One of the foremost interests regarding a large survey database is to know the distribution of quality of data. Fig. 3.12 shows the distribution of quality for all the data files (one delay zone) for the entire survey in terms of individual QFs, $QF_{data}$ and TQF. Nearly 84% of the data files cross the threshold for completeness which indicates that the observations are generally uninterrupted. Among the remaining files which are not complete there are only 7% files which have more than half the total number of expected data points. This clearly indicates that imaging on the basis of one sidereal hour range is justified and does not result in significant loss of data. For the analysis of distribution of QFs based on other individual key parameters only the data files which were complete were considered. The data files which cross the threshold for interference, noise and effect of Sun in the visibilities are 71%, 89% and 68% respectively. Nearly 80% of the data files cross the threshold for $QF_{data}$, while 58% of the files pass the threshold for TQF. About 40% of files cross all the individual thresholds, $QF_{data}$ and TQF. An important point to be noted is that there is ≈20% data observed in day time in which the effect of Sun is within acceptable limits.

Fig. 3.13 shows the distribution of quality for all the calibration files in the entire survey in terms of individual QFs and $QF_{data}$. Nearly 91% of the calibration files are complete. The calibration files which cross the threshold for interference, noise and effect of Sun in the visibilities are 78%, 88% and 77% respectively. About 85% of the calibration files cross the threshold for $QF_{calib}$. Another issue of interest is that on an average how many good acceptable calibration files are available for each data file. The results show that on an average there are ≈2.4 calibration files available for each data file. This also indicates that a calibration technique which would estimate the complex gain of the array using the multiple calibration files would be useful to improve the calibration. Presently the visibilities are calibrated using only the best calibration file among the ones available.

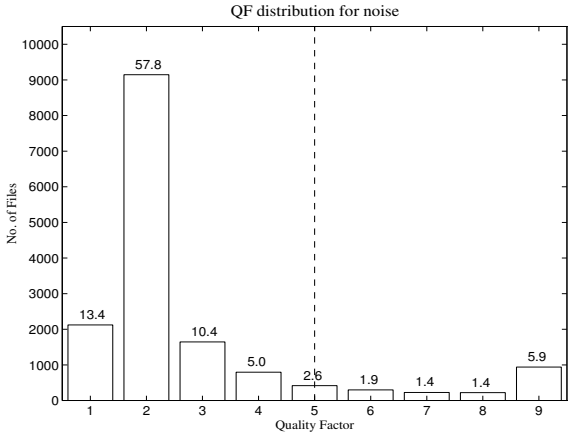
Another important result revealed by the automatic data classification algorithm is that



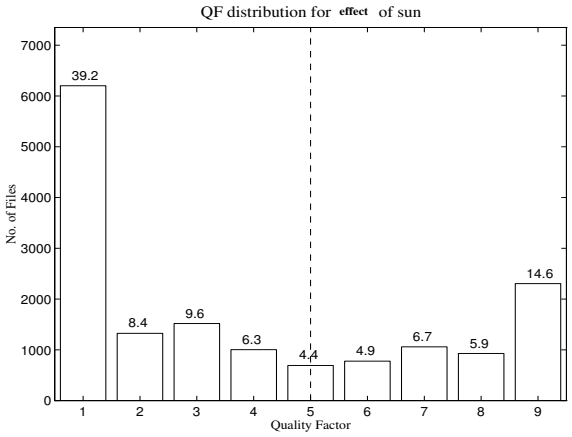

(a) QF distribution for completeness.



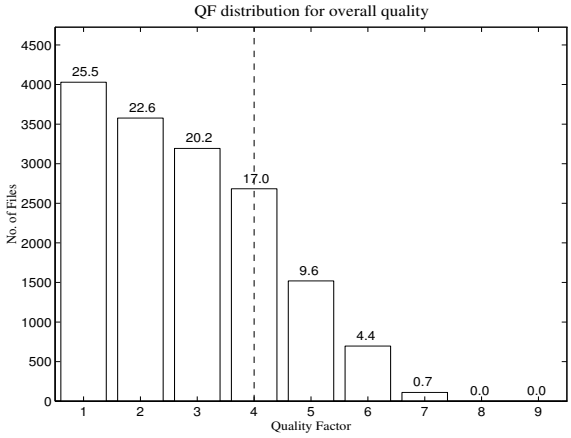

(b) QF distribution for interference.



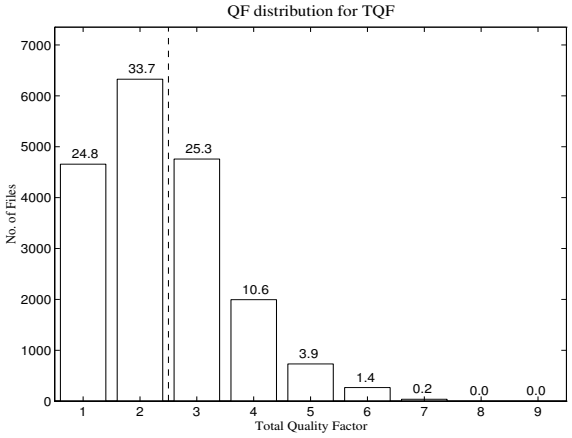

(c) QF distribution for noise.




(d) QF distribution for the effect of Sun.

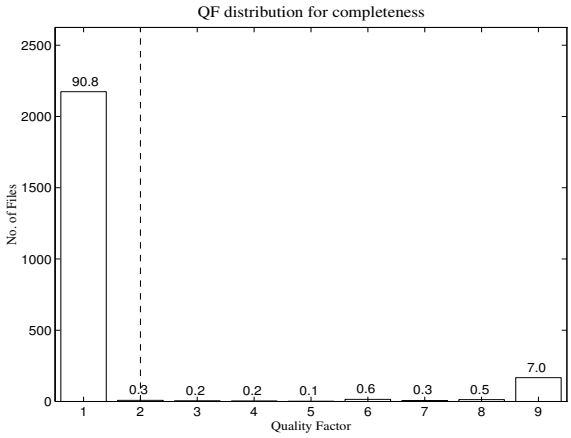



(e) QF distribution for $QF_{data}$.

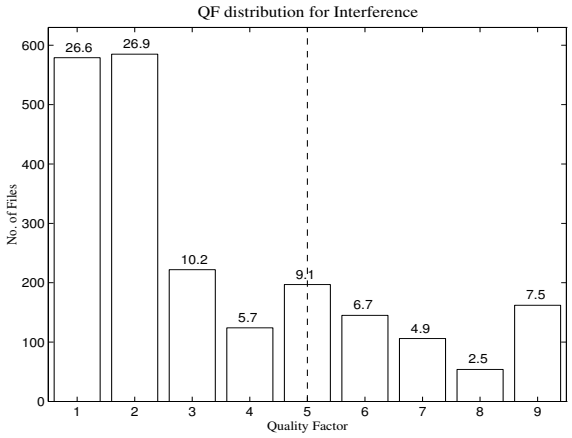



(f) Distribution of TQF.

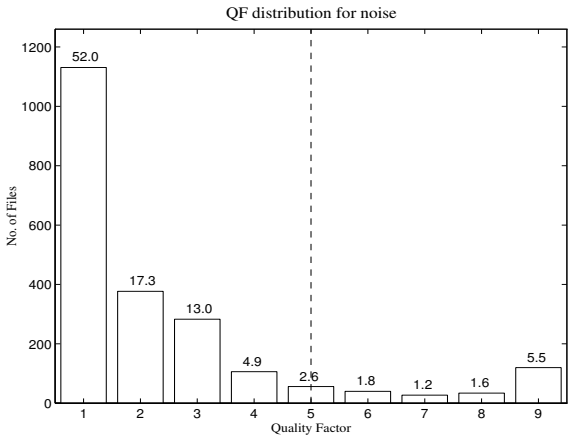
Fig. 3.12: Histograms of the distribution of QFs for data files for the entire survey (one delay zone). For the histogram showing the QF distribution for completeness (a) all the data files for the entire survey for one delay zone have been taken. For all other QF distributions (b to f), only those data files which satisfy the threshold for completeness have been considered. The threshold cutoffs are shown as dashed lines in the individual plots. The % of the files falling within each bin is also indicated on the top of each bar.



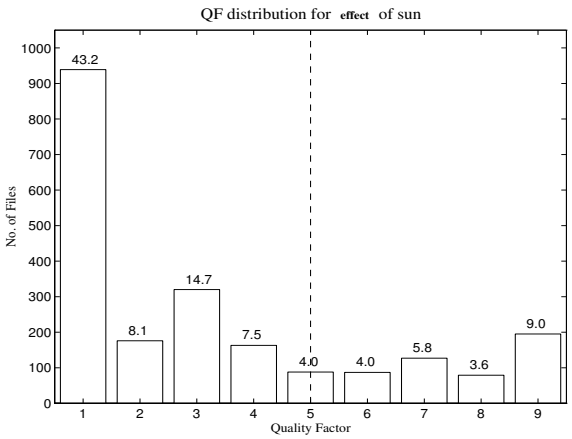

(a) QF distribution for completeness.



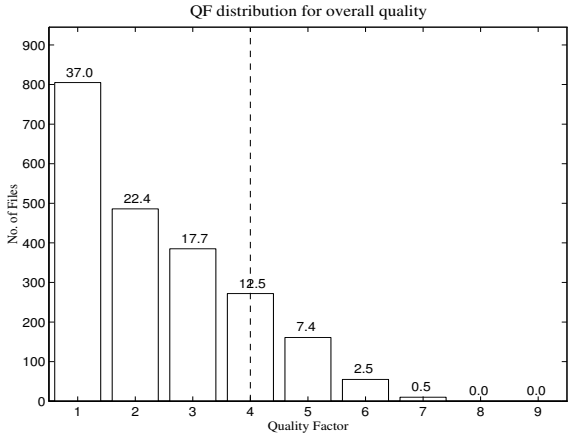

(b) QF distribution for interference.




(c) QF distribution for noise.




(d) QF distribution for the effect of Sun.




(e) QF distribution for $QF_{calib}$.

Fig. 3.13: Histograms of the distribution of QFs for the calibration files (of the three calibrators) in the entire survey data. For the histogram showing the QF distribution for completeness (a) all the calibration files for the entire survey for one delay zone have been taken. For all other QF distributions (b to e), only the calibration files which satisfy the threshold for completeness have been considered. The threshold cutoffs are shown as dashed lines in the individual plots. The % of the files falling within each bin is also indicated on the top of each bar.

nearly 15% of the data files cannot be imaged due to lack of a suitable calibration file. Implementing a calibration scheme such as field of view calibration, redundant baseline calibration etc. would facilitate usage of this good data. We also estimated the completeness of the *u*v coverage for each sidereal hour range expected in the final images, assuming that a pair of data and calibration file classified as good for imaging by the data classification algorithm will produce an image which is good and acceptable. The allocations for which good data was not available were very few. An important point to mention is that although imaging is carried out on sidereal hour basis, we need images of adjoining regions on both the sides in RA as guard zones to successfully deconvolve the dirty image near the boundaries. In view of this for each allocation's image the visibilities on both the sides of the sidereal hour range under consideration whenever available are also imaged separately and joined along RA. Thus for each allocation, we effectively need continuous data for nearly three sidereal hours. Due to this the actual number of good data files which were available for each allocation were about 30-40% lesser.

## 3.8 Discussion and possible applications

Classification of data for a large database which is a result of 20,000 hours of astronomical observations is an important but challenging and daunting task. The tedious and labor intensive method of editing visibility data has acted as a deterrent for obtaining the best quality images. The imaging of very large data sets require enormous efforts and the quality of image obtained depends upon the subjective satisfaction of the particular astronomer. The novel approach developed and implemented in form of data classification astronomical software has made the data classification objective, efficient and automatic which has brought down the time required for choosing good data to a minimum.

It is encouraging that even with a simple model, for automatic classification based on assigning a numerical index for each of the key criterion works very satisfactorily. The comparison of the results with the traditional approach gives us the necessary confidence to believe in the reliability of the data classification process. Further, the data classification process gives a comprehensive overview of the survey data quality. There does exist a lot of scope of the present analysis although even in its simple form it produces satisfactory results. In our approach we have assumed the key criteria to be orthogonal and weights for each of them have also been taken as equal. The weights can be better calibrated and it is obviously tempting to obtain them via simulation to match the results with that of traditional approach. But there is no definite way of ascertaining that the traditional approach is the best one and as shown in the analysis of the discrepant cases, the ranking given by

the automatic algorithm are usually more appropriate. Due to this, such an exercise to replicate the results of the traditional approach has not been attempted. The idea behind the present algorithm has been to apply the same logic as in traditional manual approach but define the range of parameters on a quantitative basis.

The data classification algorithm is based on sum of magnitudes of visibilities of all the baselines and does not take into account when only a few baselines are bad, or a few antennae are dead, or there is low level of interference still present in few baselines etc.. So a scheme which is based on the visibilities of the individual baselines may be more effective, although it would be more time consuming. The effect of ionospheric refraction etc.. have been assumed to be insignificant and thus not taken into account. The assignment of the QFs can also be refined to make the assignment process more objective. Another possible improvement is to take into account the level shifts in the visibility file. Fortunately, the number of files affected by level shifts are small (2-3%) and hence we have ignored this aspect at present.

An aspect which has to be kept in mind is that our discussions pertain to relative ranking of different data sets observed by the same instrument and then fixing the threshold levels. In such cases there may be factors which effect the data quality and may not have been taken into account in our scheme but since they either remain constant or change very slowly, they do not affect the results significantly.

**Possible applications :** The principles used in development of the framework for data classification are of very general nature. We believe the same approach with suitable modifications has the potential to be applicable to other data sets from synthesis telescopes. For an observatory where data recording goes on day to day, such an astronomical tool can act as a benchmark to monitor the performance of an observatory in terms of data quality. Data mining of the various properties of the data quality over a period of time can reveal important factors which are responsible for the deteriorating quality of the data and hence can be addressed in a more focused manner and eliminated if possible. For the Global Virtual Observatory, the tool can help in reducing the amount of data which should be made online, as data with very poor quality is very less likely to be in demand and hence it can reduce the online storage requirements. A user before requesting or downloading the actual data, may look at various properties of the data quality to decide if it will be useful to spend time on the data or not. This would also decrease the network traffic load.

There have been numerous efforts for automatic object classification in astronomy. In object classification we attempt to look for a pattern while in case of visibility data classification we are looking at combined effects of many intricately interdependent parameters. It would be interesting to investigate techniques used for object classification for classi-

fying the visibility data and compare our results. To our knowledge in radio astronomy the approach developed by us is the first of its kind especially in applying it to the visibility data. This should motivate us to improvise it and attempt its implementation for data sets of other synthesis telescopes. The good agreement of our results with the traditional results clearly demonstrate the success of the method developed. The future belongs to array type telescopes involving large number of baselines, number of frequency channels etc. and an astronomical tool based on such an approach may prove handy.

The developed data classification scheme was used to select good data for imaging. The next important step during data processing of RFI mitigation is discussed in the the next chapter.