VLA Archive Imaging Pilot

Loránt O. Sjouwerman
National Radio Astronomy Observatory

February 20, 2007

Abstract

This document describes the results of an automated pipeline imaging pilot on 5 and 8 GHz VLA data in the winter 1999/2000 B-array. In the pilot, 239 segments (one frequency setting in one proposal code during an observing day) were processed. At this time about 20-25% of the segments could not be processed by the pipeline. In total this resulted in 672 images of 395 unique sky positions (astronomical sources), where 63 positions were imaged at both 5 and 8 GHz. The setup and typical processing time for a B-array data segment and a handful of sources was about 10-15 minutes on an otherwise empty workstation, each typically yielding 90 MegaBytes of new data of which 30 MB in images. Processing of all VLA 1-50 GHz continuum data is estimated to take one to two person years and four to five TeraBytes of archive disk space. Separate memo's are a manual on the usage of the VLA pipeline in AIPS, and a manual (by Jared Crossley) on the operation of the automated pipeline.

Acknowledgments

Many thanks go to Eric for making some of my wishes come true, and to Amy, John, Ed and Jared for all their suggestions and other input.

Contents

1 Projects **1**

2 Results **2**

3 Common Problems **3**

4 Projections **4**
4.1 Software 4
4.2 Processing Time 5
4.3 Disk Space 5

5 Current Status **5**

1 Projects

The pilot program consisted of 5 and 8 GHz continuum data, taken with the VLA between 1999 Oct 29 and 2000 Feb 14, when the VLA was in its B-array configuration (109 days).

The NRAO archive was queried for eligible data sets in any of the five possible subarray settings, and a total of 377 VLA archive files was downloaded from 71 program (i.e., proposal) codes. Obvious operational programs (system

Table 1: Number of observations per day.

Month		Downloaded		Observ.
	Days	Observ.	Arch.Files	Day^{-1}
OCT	3	9	15	3.00
NOV	30	104	129	3.47
DEC	31	74	106	2.39
JAN	31	62	92	2.00
FEB	14	26	35	1.86
	109	275	377	2.52

startup, pointing, flux and baseline calibrations) and low frequency 4, and P band data were skipped, though programs called 'test', 'tst', etc., were included as they usually yield valid source images. An `observation` is defined as one or a series of VLA archive files contiguous in time with the same program code, which thus may include multiple frequency settings. A `segment` is one frequency setting of an observation and is the quantum of AIPS processing because it is calibrated and imaged in one process, i.e., a separate AIPS user number in the automated pipeline processing. For example, an observation from one program code that switches between sources at 5 GHz and 8 GHz (and any other frequency) sequentially or continuously will be split up in a 5 GHz and an 8 GHz segment.

The 377 VLA archive files downloaded consist of 275 different observations, about 2.5 per day, which in turn resulted in 357 segments of 5 or 8 GHz (3.3 per day, excluding the other bands). Some observations also contained segments in other continuum bands (21× at 1.4 GHz, 18× at 15 GHz, 32× at 22 GHz and 52× at 43 GHz during this 109 day interval). Although these segments in other observing bands were discarded for this pilot, they mostly would have processed fine. Of the 357 segments 118 (24%, 36 program codes) were abandoned because of one or more of the following reasons: it was a segment without a standard calibrator (33 times), it was a moving target (planet, mosaic, 12×), it was a spectral line (28×) or non-5/8 GHz continuum (48×) segment which got downloaded because it contained continuum pointing scans at 8 (or 5) GHz, or, it appeared to be a single dish VLBI observation (7×, phased array data were processed). Of these 118, 76 should not have been selected and downloaded (pointing and spectral line), so the total rejection rate for these reasons is not 33% but more like 15-20% (118-76 over 357-118). If the uncompleted segments are included (see below), the total rejection rate becomes 20-25%. The total number of processed segments was 239 (48 program codes - some appear both here and in abandoned usually because there were multiple observations for these program codes, though some segments lack the calibrator scans). Some of the abandoned segments probably were processable (planets, lacking standard calibrator), but not attempted for the pilot.

2 Results

A total of 222 segments were completed, most of which before further improvements to the pipeline code were made, with the images and calibrated data archived. Most segments were completed in a manner of minutes (mostly less than 10 minutes in B-array), and is depending on (array configuration,) number of sources, number of visibilities per source, source declination (etc., enabling multi-faceting) and source structure. No further analysis than a typical processing time is attempted here as modifications to the code during and after the pilot both on

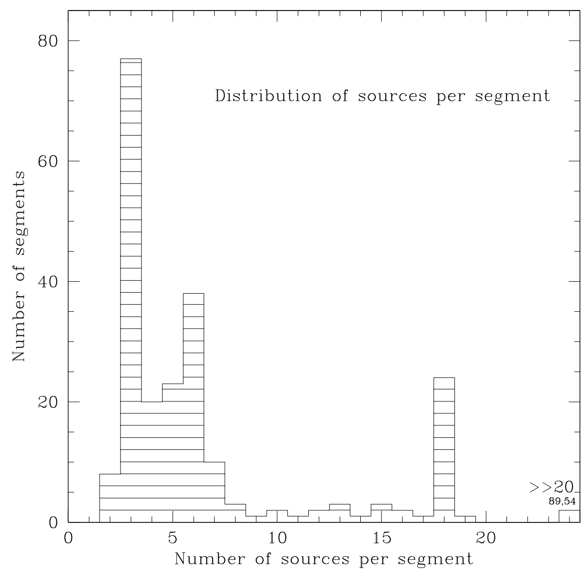

Figure 1: Distribution of number of sources per segment. The hunt for GRB counterparts and the two monitoring projects observing the same sources many times during the pilot show up as peaks in the distribution at 3, 6, and 18 sources, respectively. Two projects had many more than 20 sources (54 and 89), but typically people observe three to six sources per segment.

Table 2: Number of images per day.

Month	Images at 5GHz	Images at 8GHz	Total	Images Day^{-1}
OCT	97*	40	137	46.0*
NOV	51	115	166	5.5
DEC	42	81	123	4.0
JAN	45	66	111	3.6
FEB	91*	44	135	9.6*

*: These each include a single segment with many (89 and 54) sources skewing the statistics.

one hand increased the average and on the other hand decreased the average processing time. Later processing with other arrays confirm the intuition that A-array takes much more, whereas C-array and D-array take on average much less processing time. It turned out that 6 segments were doublets and another 11 needed some extra attention or adjustments to the pipeline code which was kept for later. The breakdown of number of (unique) images and such is given in the accompanying figures and tables. The images sic., etc., can be most easily accessed through http://www.aoc.nrao.edu/∼vlbacald . It is planned that eventually they will be available through the default NRAO archive portal, http://archive.nrao.edu .

3 Common Problems

Most problems relate to either bad data, bad observing conditions or bad pipeline constraints. In particular the automated flagging is not yet perfected and occasionally obviously bad data may not be flagged automatically. However, bad data is easily removed by interactive flagging programs and rerunning the pipeline with additional flagging information in the flagging table is trivial.

In some cases the automated flagging flags too much data, in particular in fast-switching observations. Or, when slewing of the telescope leaves too little data on the first (calibrator) scan, flagging can remove most or all of the data of those scans. If not enough data remains, calibration solutions may be too noisy or undetermined.

Other problems may occur because the input u, v-data file does not have the expected structure. For example, none of the standard flux calibrators observed has enough visibilities, the data contains multiple frequency-ID's or subarrays, or calibrator codes are missing. These problems usually are easily fixed by hand. Nothing can be done to correct for "unfortunate" observing conditions such as weather, bad pointing, faulty hardware, too few antennas in the subarray for proper imaging and observer errors.

In a few cases, a data set just cannot be calibrated, period! Though this turns out to be only a small fraction of this 5 and 8 GHz pilot (5%, 11/222), it probably is higher when the other bands are included (e.g., phase connection problems at higher frequencies).

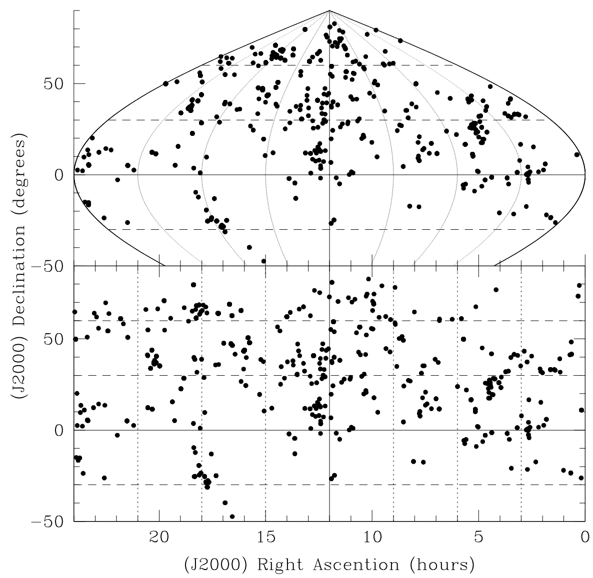

Figure 2: Distribution of sources on the sky. These are the unique sky positions in the pilot at 5 and/or 8 GHz - the current status can be found at http://www.aoc.nrao.edu/∼vlbacald .

4 Projections

4.1 Software

Apart from many unspecified improvements to the original code for VLARUN (the original AIPS VLA calibration and imaging pipeline procedure from the author) and specific fine-tuning to this C (5 GHz) and X (8 GHz) band B-array VLA data, new code was developed to streamline querying and downloading VLA archive files (with some input from John Benson), to blindly run and archive the resulting files (see Jared's document), and to collect and present the images to a prospective user through the web. Also some existing AIPS code was modified and new AIPS code was developed to improve non-interactive performance, in particular bad data flagging (with a big thank you to Eric Greisen), and by making calibrator models more readily available (Amy Mioduszewski). Except for ongoing improvements to general AIPS code itself, these software contributions by others are mostly a one time investment and, apart from further improving the pipeline with new capabilities, need not be continued for processing a larger part of the VLA data archive.

Currently, improvements are planned in the area of self-cal algorithms and dynamic range problems, as well as processing of "moving target" (i.e., solar system) data. Longer time scale improvements could include more robust flagging and perhaps polarization calibration. It is also possible to check and update calibrator positions and calibrator codes on the fly and perhaps create models for the calibrators prior to the target calibration. However the priority and implementation of part or all of this depends on local priorities.

Future new code could shorten processing time ("smarter tricks") as well as improve image quality (flagging, self-cal) or quality control (e.g., plotting diagnostics or sophisticated modeling).

4.2 Processing Time

The processing time includes querying the NRAO archive and downloading the data to the processing workstation (a few minutes) and the actual time to calibrate the data and make the images (several minutes). Not included is the time it takes to make diagnostic plots (as that was added later, another several minutes), exporting the data to the archive (a couple of minutes) and the assessment of the quality of the images (zero to infinity). The latter because it strongly depends on, e.g., updates to the code and the understanding of problems found. At this stage the results were analyzed interactively in AIPS, which takes a lot of time. Because the process is much more robust and streamlined now, it is more efficient to remove bad images from the archive than to include good ones. Currently the results are written directly to the archive where quality control is done after the fact, limiting the interactive process to starting the pipeline (a couple of minutes) and checking the results (another couple of minutes per segment).

Based on the 5 and 8 GHz B-array pilot, a typical segment takes 15 minutes to process, a total of about 100 hours for the 377 segments in this pilot. If the 123 segments at other bands were included (another 40%, or factor 1.4), and if translated to a full year it would take 460 hours or 20 days computer processing time on a single, otherwise empty workstation. Though this translates to 60 working days, the processing is done mostly in absentee. The whole VLA data archive, assuming all data sets run fine and as typical as the 5 and 8 GHz segments in this B-array, would the take 540 computer processing days (7 person years) but this is an absolute maximum as the effort is only in initiating the process and checking the results ($\sim$5 instead of 15 minutes). It can easily be done by one person full time on multiple machines in much less time. One to two person years on a hand-full of computers for the completion is probably more realistic and feasible.

4.3 Disk Space

The data archived from this 5 and 8 GHz continuum B-array configuration take up about 35 GigaBytes. This includes, next to the images, calibrated u, v-data and processing files. The images themselves take up about 30%, the calibrated u, v-data with their diagnostic plots about two-thirds, and the rest are processing text files. Assuming a 1.4 conversion factor for expansion to the whole 1 to 50 GHz range (see above), this translates to 50 GB per 4-month configuration, 165 GB per year or between 4 and 5 TeraBytes for the current, whole 28 years of VLA data. Of course, this assumes that this particular B-array values are a good estimate for these figures over all arrays.

5 Current Status

During and after this winter 1999/2000 VLA B-array pilot the pipeline was further improved and extended, although not yet perfect. Jared has since then continued to run the pipeline on all subarray and all configuration 1-50 GHz VLA archive data. Currently, as of mid-February 2007, we have about 16 000 images in the archive and filling it at a rate of about 2.5 GB per day! See: http://www.aoc.nrao.edu/~vlbacald .