The short FFT Data Base and the peak map for the hierarchical search of periodic sources, with the VIRGO detector

Federica Antonucci - Pia Astone - Sergio Frasca - Cristiano Palomba

Roma, 2/05/2007

Summary

- The short FFT data base
- Time domain disturbances: event's identification and removal
- Estimation of the average AR spectrum
- Construction of the peak map
- The log files

The Short FFT data base

The procedure relies on a data base of FFTs, computed from short stretches of data (short with reference to the observation time and to the effects of the Doppler shift due to the Earth motion). Each FFT in the data base contains a header, a *very short FFT*, which is the estimation of the average spectrum, needed in the identification of peaks, and the total power spectrum over sub-periods, important to get information on the stationarity of the data. The time duration T_{FFT} of one FFT must be such to have no effect due to the Doppler shift:

$$T_{max} = T_E \cdot \sqrt{\frac{c}{4\pi^2 \nu R_E}} \approx \frac{1.1 \cdot 10^5}{\sqrt{\nu}} \quad \text{[s]}$$

where ν, in Hz, is the source intrinsic frequency, T_E and R_E are the period and the radious of the Earth rotation at the latitude of the detector. The choice of T_{FFT} is a function of the frequency.

The Short FFT data base

- 500-2000 Hz, with $T_{FFT} \approx 1000$ s (this has been reduced by a factor 2, to reduce the computational burder);
- 125-500 Hz, with $T_{FFT} \approx 4000$ s;
- 31.25-125 Hz, with $T_{FFT} \approx 8000$ s;
- 0-31.25 Hz, with $T_{FFT} \approx 16000$ s.

Time domain disturbances: event's identification

The variance, hence the sensitivity of the detector, changes with the time. We change the threshold with the time ("adaptive threshold"). Let x_i be the samples. These data could be simply the output data, or data after a bandpass filter, or data filtered with a filter matched for delta-like signals. The background is estimated from the A-R mean of the absolute value and of the square of x_i

$$y_i = x_i + w \cdot y_{i-1} \quad (1)$$

$$q_i = x_i^2 + w \cdot q_{i-1} \quad (2)$$

with

$$w = e^{-\delta t/\tau} \quad (3)$$

δt is the sampling time and τ is the memory of the A-R mean, here in seconds. The normalization factor is $Z_i = [1 + w \cdot Z_{i-1}]$, with $Z_0 = 0$.

Time domain disturbances: event's identification

Mean and standard deviation are evaluated as:

$$\mu_i = \frac{y_i}{Z_i} \tag{4}$$

$$\sigma_i = \sqrt{\frac{q_i}{Z_i} - \frac{y_i^2}{Z_i^2}}\,. \tag{5}$$

The threshold is set on the critical ratio CR, defined as

$$CR = \frac{x_i - \mu_i}{\sigma_i} \tag{6}$$

The memory time τ depends on the apparatus. We set it to 600 s,and the CR_{thr} to 6. The procedure makes use also of the concept of *dead time*,minimum time between two events. It depends on the noise and the expected signal. We used 1 s.

Time domain disturbances: event's identification

Once we have defined the *adaptive threshold* the procedure works as follows:

- when the signal x_i goes over the threshold ($CR > CR_{thr}$), an event begins;
- the event ends after the signal has remained below the threshold for a time longer than the *dead time*;
- the event is characterized by various parameters, the ones which matter here are its *beginning time* and *duration*, defined as time above threshold subtracted the dead time, while *time of the maximum amplitude* and *maximum amplitude* are not important here.

Time domain disturbances: event's removal

Once an event has been identified, we clean the data with the removal procedure. This requires the set up of another parameter, which we call the *edge* and indicates how many seconds before and after the event are used in the cleaning of the data. In the examples here we have used 0.15 s.

- Data from the *beginning time* up to the (*beginning time + duration*) are set to zero;
- data from the time (*beginning time - edge*) are linearly set toward zero, while data from the time (*beginning time + duration*) are linearly set toward the value at (*beginning time + duration + edge*).

Summarizing: the data which define the event are set to zero, and a linear connection with values around the event is applied.

The procedure to estimate the average spectrum

A good estimator should have the following properties:

- if peaks in the frequency domain are present, the estimator should not be affected by the peaks. This should be as much as possible independent on the SNR of the peak;
- if the noise level varies, either slowly or rapidly, the estimator should be able to follow the noise variations.

We refined the procedure, with the use of an autoregressive estimation (AR) of the average of the spectrum, with the basic idea of a “clean estimator”.

The procedure to estimate the average spectrum

For a given spectrum, with frequency resolution δ_ν and absolute value of each sample x_i, where $i = 1, N$ and N the lenght of the FFT, the estimator of the average μ_i is obtained with the following recursive equation:

$$y_i = x_i + w \cdot y_{i-1} \,, \tag{7}$$

where

$$\begin{aligned} w &= e^{-\delta_\nu/\tau} \\ Z_i &= 1 + w \cdot Z_{i-1} \end{aligned}$$

and

$$\mu_i = \frac{y_i}{Z_i}$$

Z_i is the normalization constant ($Z_0 = 0$).

The procedure to estimate the average spectrum:

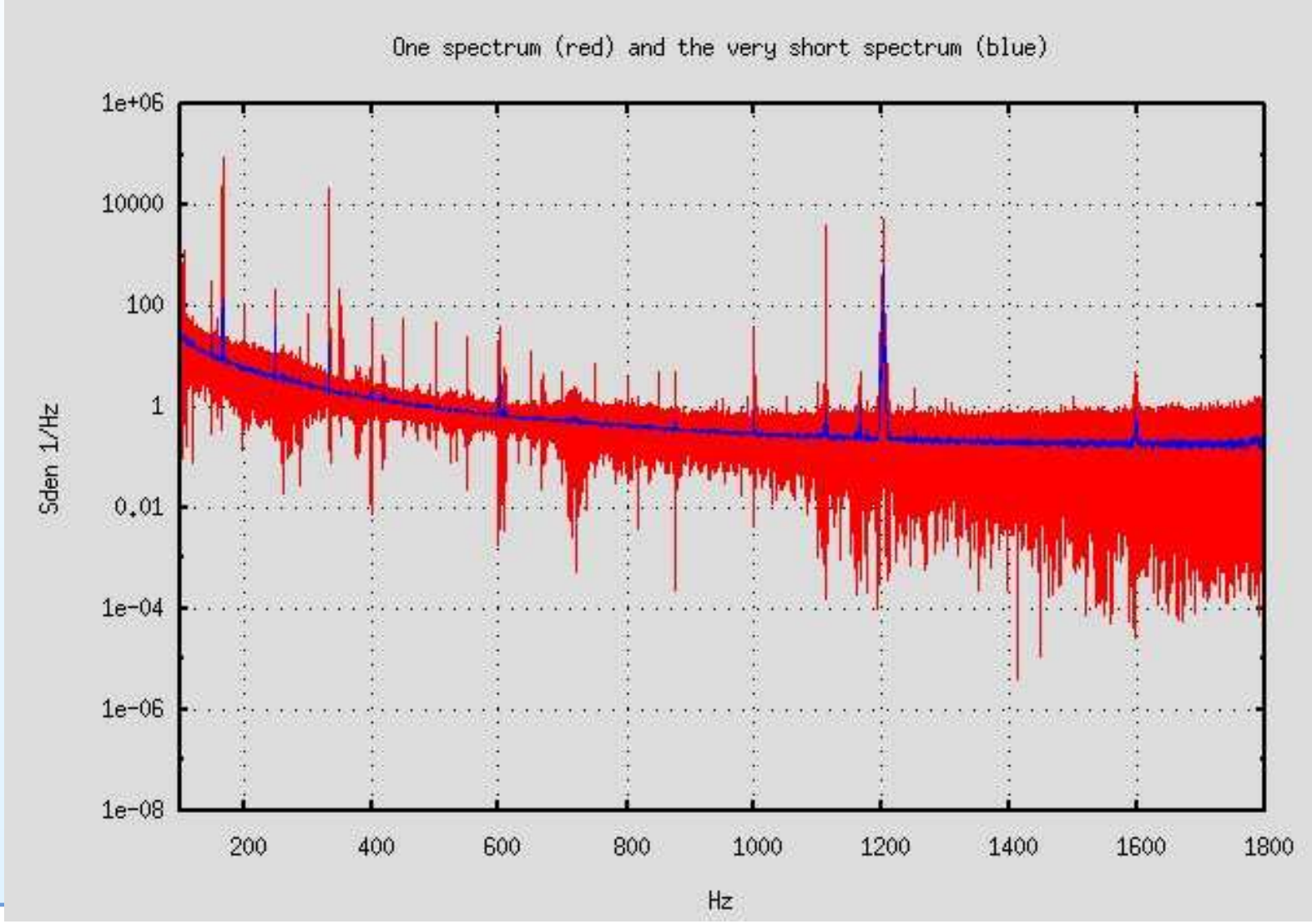
$w = e^{-\delta_\nu/\tau}$ is the memory of the process by means of the constant τ, which has the dimensions of a frequency, that is Hz.
This is a first parameter which has to be chosen.
Its choice depends on the characteristics of the noise.
The application of a “clean procedure”, useful for an efficient removal of peaks from the average, requires the definition of two additional parameters: a threshold V_{max} and a maximum age A_{max} (in Hz). It works as follows:

The procedure to estimate the average spectrum:

- while $r = x_i/\mu_{i-1} < V_{\mathsf{max}}$ the new datum x_i is used to evaluate the actual mean μ_i and the age A of the process is set to zero;
- when $r = x_i/\mu_{i-1} \geq V_{\mathsf{max}}$ the new datum x_i is not used to evaluate the actual mean μ_i and the age A of the process is incremented of δ_ν. This eliminates or reduces the effect of peaks from the estimation;
- if the age (in Hz) $A > A_{\mathsf{max}}$, we decide that the characteristics of the noise changed and thus we have to go back of a number of samples $n = A/\delta_\nu$, and begin a new evaluation of the mean, restarting from zero at the sample $(i-n)^{th}$. This is needed to face with all those situations when the noise is highly non-stationary, with abrupt changes of the level of the floor.

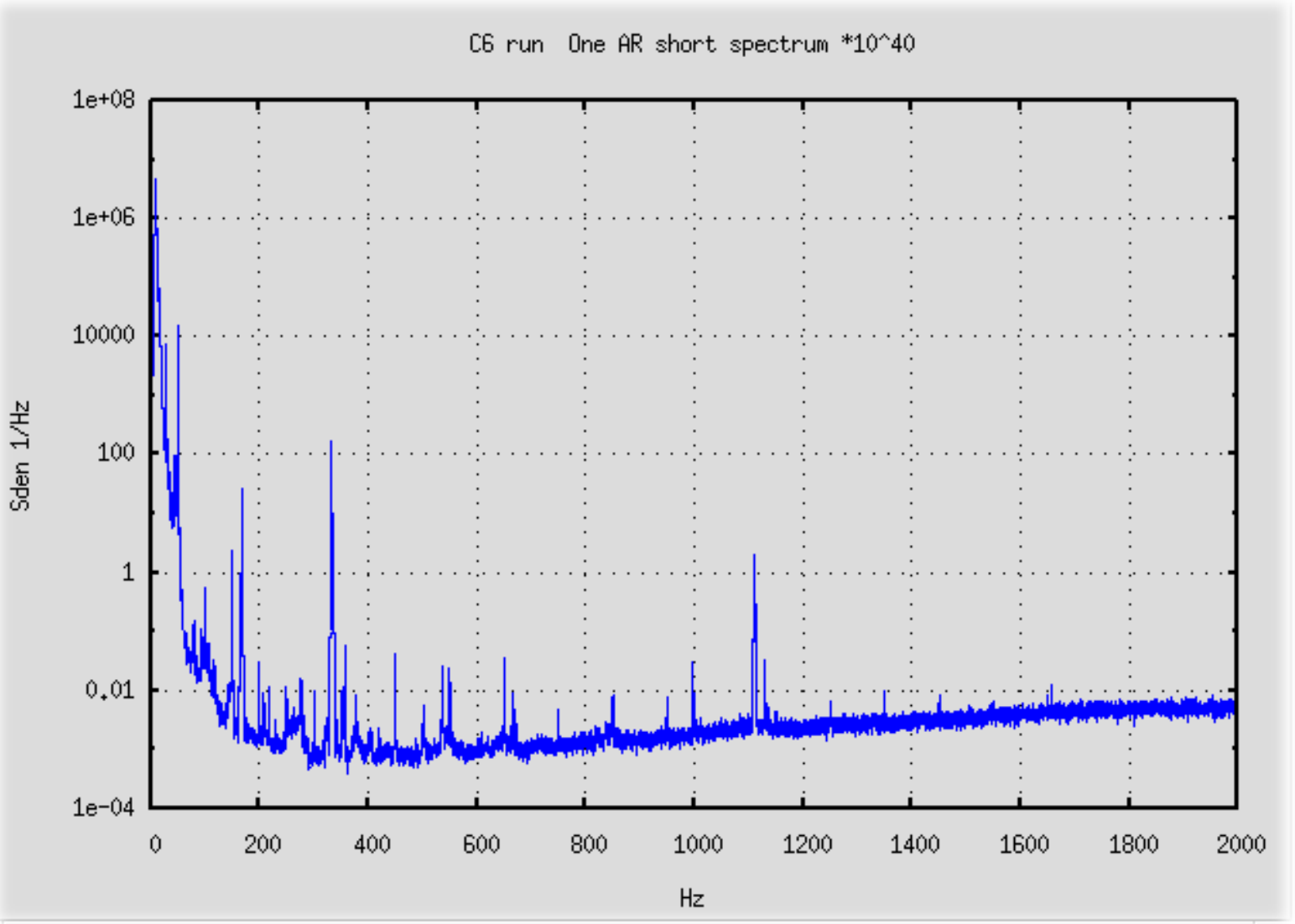
The procedure to estimate the average spectrum.

$\Rightarrow$ Example using C6 data. $T = 524.288$ s



The procedure to estimate the average spectrum.

$\Rightarrow$ Example using C6 data. $T = 1048.6$ s



Construction of the peak map

- The construction of the peak map, first step of the hierarchical procedure, is a very delicate point, as it influences the sensitivity of the next steps. A potential candidate, which is skipped at this stage due to an inaccurate or non optimal construction of the peak map, will never be recovered. The peak map construction starts with the ratio $\mathcal{R}$ of the spectrum to its AR estimation. On this function, we set a threshold at the level of $\mathrm{SNR}_{thr} = \sqrt{2.5}$. All the data which cross the threshold and are local maxima are then registered into a new file, from which we can produce a time-frequency plot which we call the *peak map*. This new file contains the information on the peaks (beginning time of the FFT, frequency bin of the peak,ratio,v_x, v_y, v_z), which is the information needed to the next step of the analysis.

The peak map

→ The code which contains all the utilities to evaluate, and thus simulate or remove, Doppler effect from sources to a detector on Earth, based on JPL ephemerides file and NOVAS utilities, is ready and is called the **PSS-astro library**, part of the libraries developed by the Virgo group in Rome.

C6 data: creation of the FFT data base

$\Rightarrow$ presently: $T = 1048.6$ s

- Job run on one workstation at CNAF.
  Start 12.40, 26 october 2005 End 19.30, 26 october 2005
- Band: 0-2000 Hz;
  Group C6
  Input files (N/GB): 255 / 17.8 GB
- FFT length: 4194314
  FFT mode: overlapped by the half, flat top- cosine window
  Short power spectrum reduction: 128
- FFT number: 2287
  Files created (N/GB):23 / 35.8 GB
  typical file length: 1.6 GB (100 FFT)

C6 data: creation of the data base

$\Rightarrow T = 1048.6$ s

- High pass filter, before the events cleaning, $f_T = 100$ Hz.
- Big events cleaning parameters:
  tau: 600 s; cr:6; deadtime: 1 s; edge: 0.01 s.
- AR spectral estimate parameters:
  ratio min: 1.5811; tau Hz: 0.02 Hz; maxage: 0.02 Hz;
  threshold to veto: 2.5;
- Number of cleaned events (in time): 8109
  Total time vetoed: 2000 s
  (in practise is the half, due to overlapping)

C6 data: creation of the peak map

$\Rightarrow T = 1048.6$ s

- Peak map creation: job run on one workstation at CNAF
  Start: 10.40, 27 october 2005- End: 14:00, 27 october 2005
- files produced: 4
  (1 every 6 sfdb files-600 FFTs, the last has 487 FFTs)
  typical file length: 1.4 GB
  peakmap-c6 1 peak number: 134152922
  peakmap-c6 2 peak number: 131390087
  peakmap-c6 3 peak number: 129360170
  peakmap-c6 4 peak number: 98137572
- Total and average (1 FFT) peak number:
  total=493040751 ; average=2.1558e+05

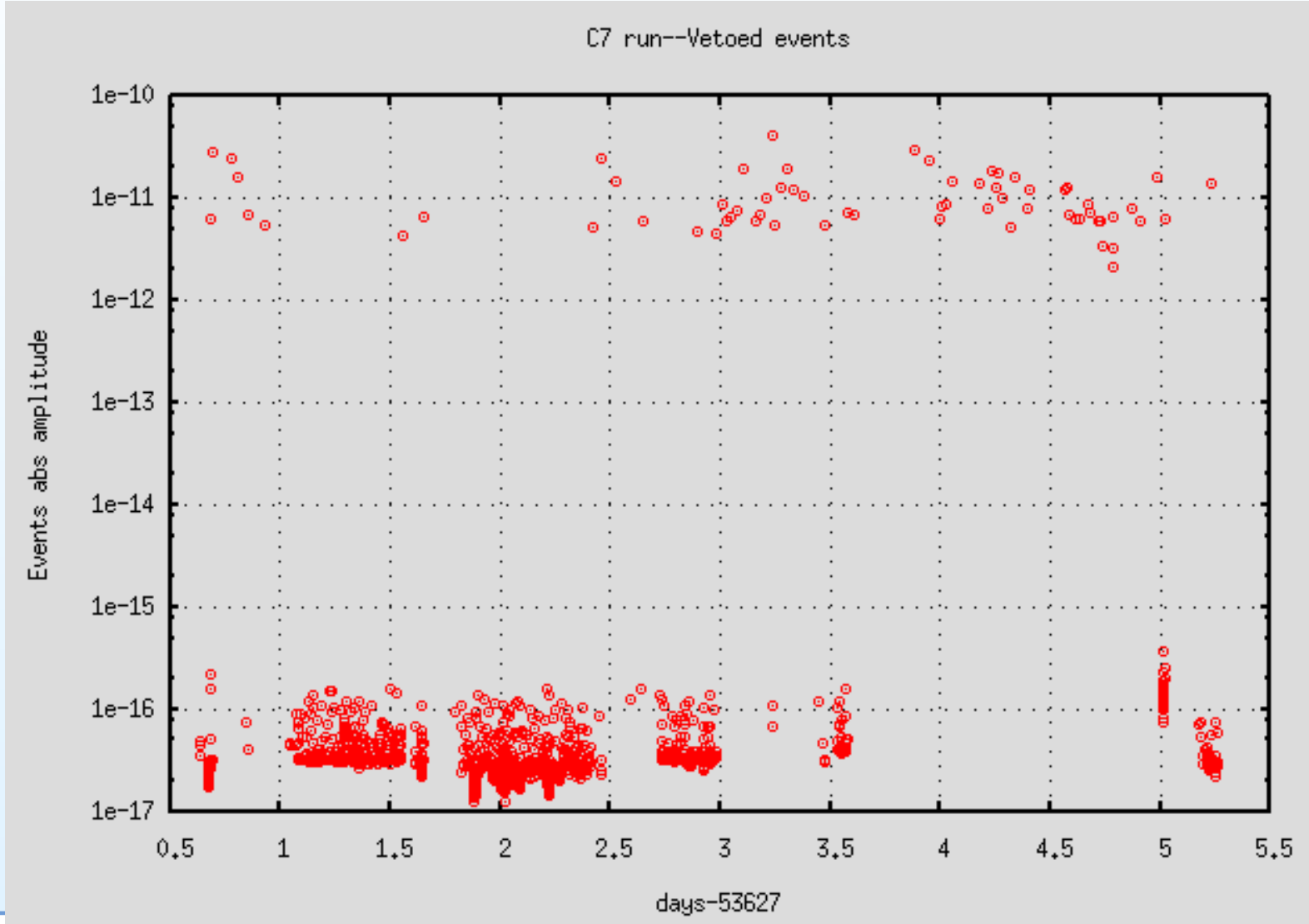
C7 data: creation the FFT data base

$\Rightarrow T = 1048.6$ s. Number of FFTs=556

- High pass filter, before the events cleaning, $f_T = 100$ Hz.
- Big events cleaning parameters:
  tau: 600 s; cr:6; deadtime: 1 s; edge: 0.01 s.
- AR spectral estimate parameters:
  ratio min: 1.5811; tau: 0.02 Hz; maxage: 0.02 Hz; threshold to veto: 2.5;
- Number of cleaned events (in time): 1963.
  Total time vetoed: 722 s (the half, due to overlapping)
  Median of the events duration: 0.046 s.
  Mean: 0.368 s ; Std= 2.0 s;
  Median of the vetoed events abs(amplitude)= $2.99 \cdot 10^{-17}$
  Max of the vetoed events abs(amplitude)= $0.4 \cdot 10^{-10}$

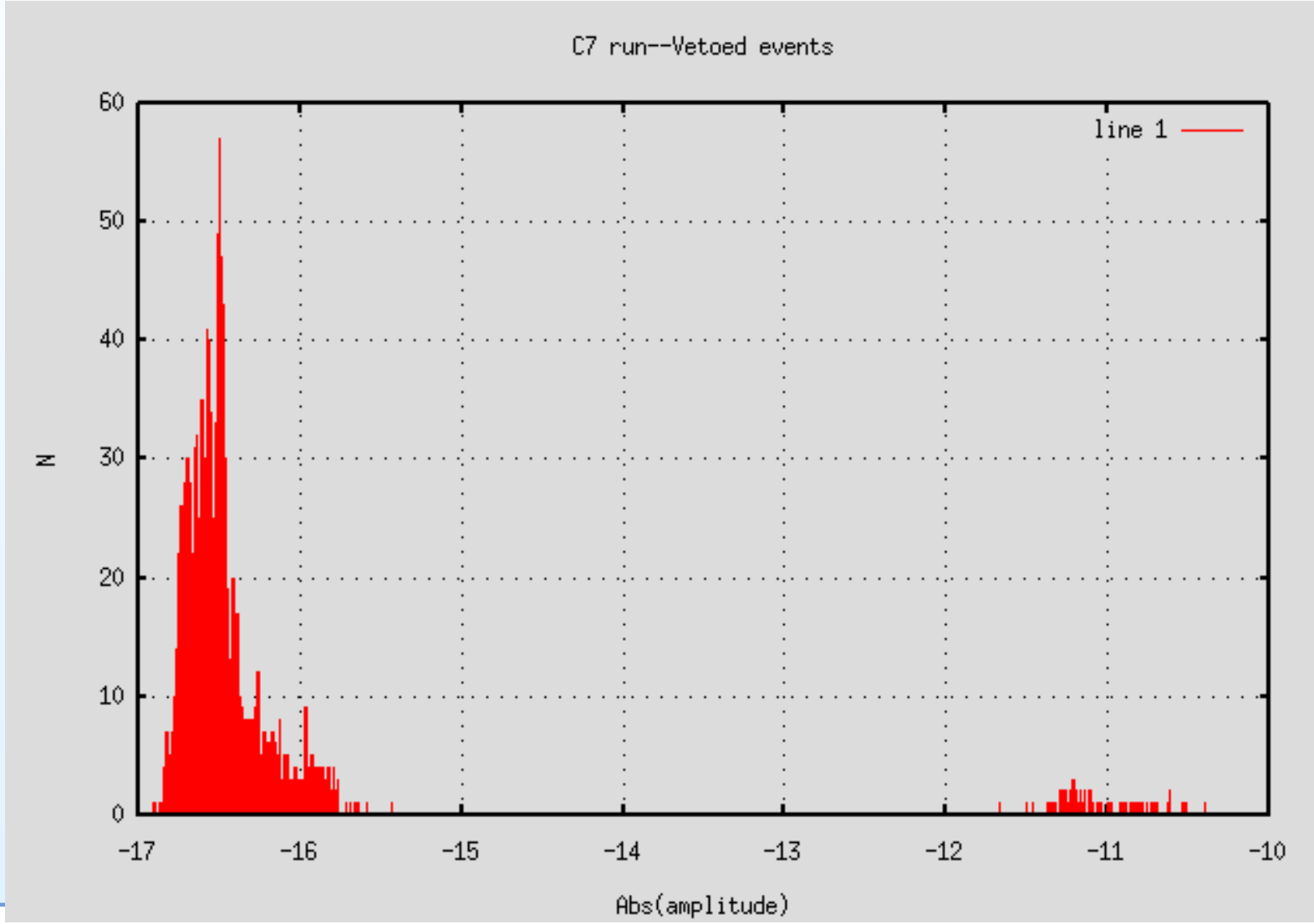
C7 cleaned events, in time domain

⇒ Abs value of the cleaned events



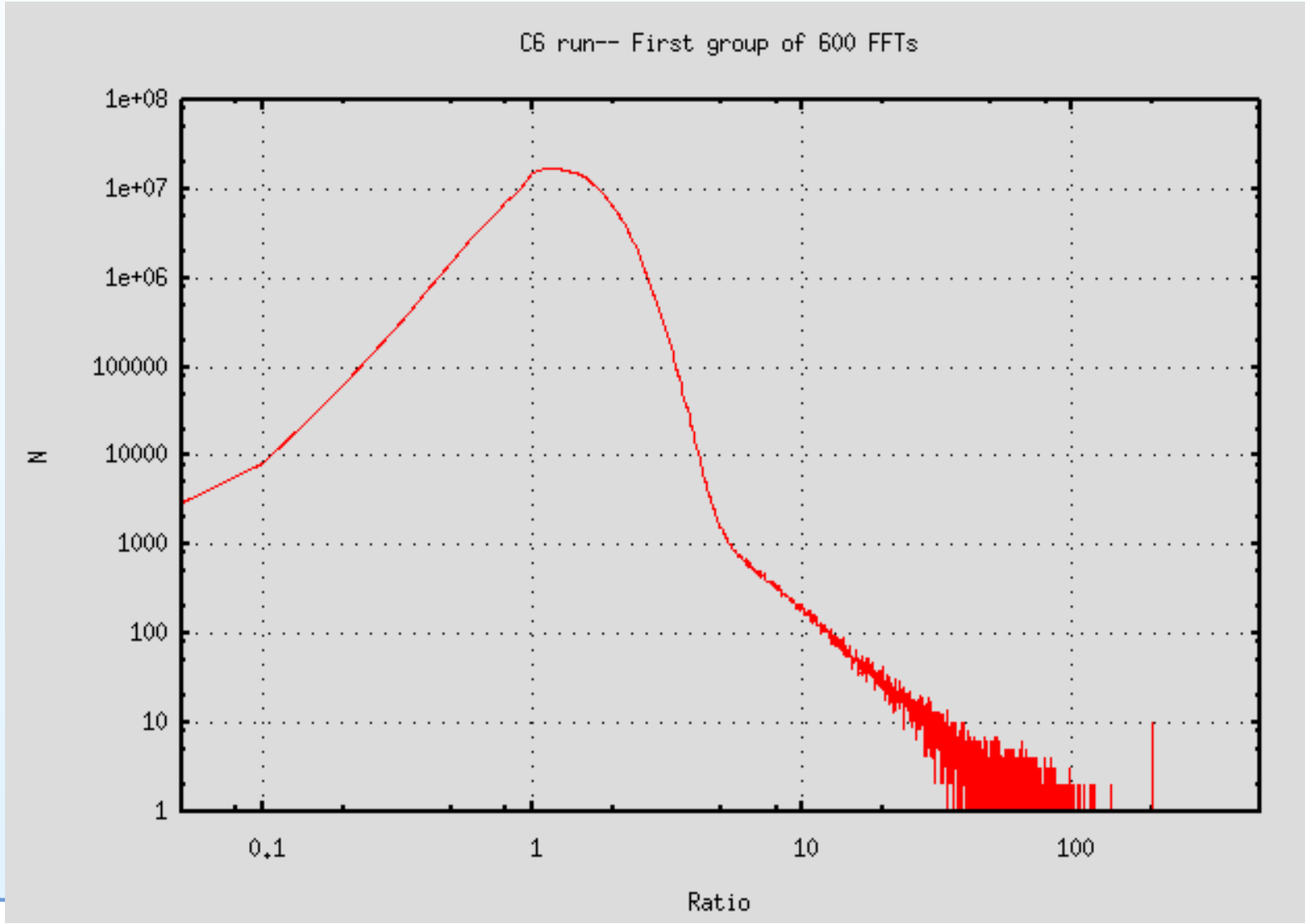
C7 cleaned events, in time domain

⇒ Histogram Abs value of the cleaned events



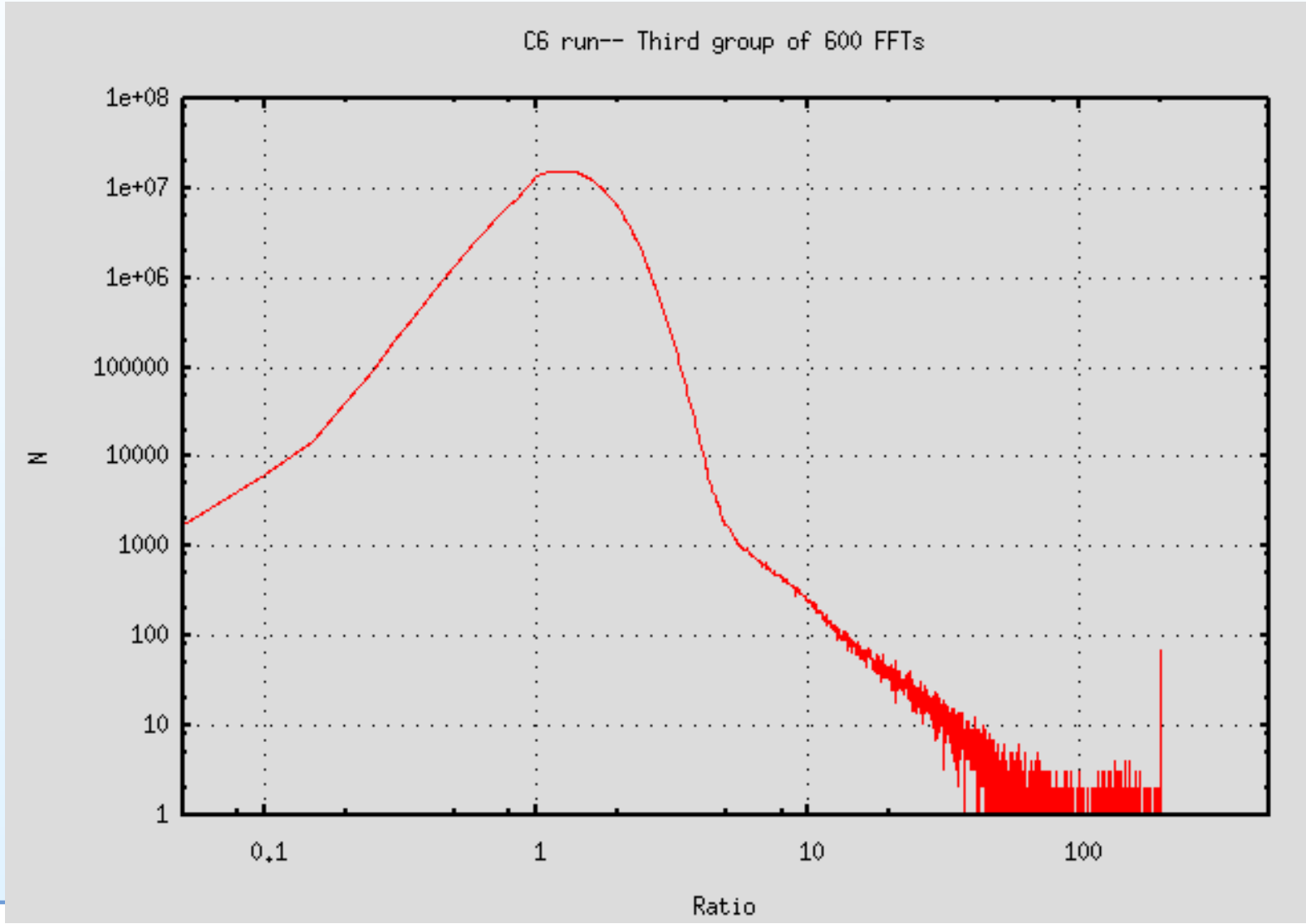
Histograms of the Ratio, without threshold

$\Rightarrow$ C6 data (first group). Very good agreement with expected



Histograms of the Ratio, without threshold

⇒ C6 data (third group).



Log files

$\Rightarrow$ Production of Log files with relevant information,
using a standard format: every process writes a Log file
(library already written by S. Frasca)

- Name of the Log file (e.g.)
  crea_sfdb_20060201_141501.log
- Info written in the log file:
- PSS crea_sfdb job log file
  started at Wed Feb 1 14:15:01 2006
  INPUT :
  /mnt/virgo/16/data2005/astone/pss/virgo/sd/sds/c7/VIR_h
  rec50Hz_20050914_151309_.sds
  First file of the run
  OUTPUT : VIR_hrec50Hz_20050914_151309_.SFDB
  The first SFDB file opened

Log files

- ! even NEW: a new FFT has started
  ! PAR1: Beginning time of the new FFT
  ! PAR2: FFT number in the run
- ! even EVT: time domain events
  ! PAR1: Beginning time, in mjd
  ! PAR2: Duration [s]
  ! PAR3: Max amplitude*EINSTEIN
- ! even EVF: frequency domain events, with high threshold
  ! PAR1: Beginning frequency of EVF
  ! PAR2: Duration [Hz]
  ! PAR3: Ratio, in amplitude, max/average
  ! PAR4: Power*EINSTEIN**2 or average*EINSTEIN
  (average if duration=0, that is when age>maxage)

Log files

- ! stat TOT: total number of frequency domain events
- (PAR) GEN_BEG = 53627.634131944
  (PAR) GEN_NSAM = 2.09715e+06
  (PAR) GEN_DELTANU = 0.000953674
  (PAR) GEN_FRINIT = 0
  ! GEN_BEG is the beginning time (mjd)
  ! GEN_NSAM the number of samples in 1/2 FFT
  ! GEN_DELTANU the frequency resolution
  ! GEN_FRINIT the beginning frequency of the FFT

Log files

- (PAR) EVT_CR = 6
  (PAR) EVT_TAU = 600
  (PAR) EVT_DEADT = 1
  (PAR) EVT_EDGE = 0.15
  ! EVT_CR is the threshold
  ! EVT_TAU the memory time of the AR estimation
  ! EVT_DEADT the dead time [s]
  ! EVT_EDGE seconds purged around the event

Log files

- (PAR) EVF_THR = 2.5
  (PAR) EVF_TAU = 0.02
  (PAR) EVF_MAXAGE = 0.02
  (PAR) EVF_FAC = 2
  ! EVF_THR is the threshold in amplitude
  ! EVF_TAU the memory frequency of the AR estimation
  ! EVF_MAXAGE [Hz] the max age of the process. If age>maxage the AR is re-evaluated
  ! EVF_FAC is the factor for which the threshold is multiplied, to write less EVF in the log file

Log files

- –> NEW > 53627.634131944 1
- –> EVT > 53627.634131944 0.001 -4781.29
  –> EVT > 53627.634418134 0.0045 -4317.92
  –> EVT > 53627.634968134 0.00475 -953.01
  .....
  .....

Log files

⇒ Production of Log files, for every job submitted

- –> EVF > 3.206253052 0.0181198 10.1963 2.95991e+11
  –> EVF > 19.412040710 0.0162125 6.22908 2.25638e+11
  –> EVF > 21.213531494 0 5.93966 2.10081e+06
  –> EVF > 25.011062622 0.0190735 8.08396 2.17784e+11
  –> EVF > 29.767036438 0 8.26272 868604
  .....
- »> TOT > 894
  –> NEW > 53627.668330000 2
  –> EVT > 53627.674398148 76.5743 -7877.92
  ....
- ....
  –> EVF > 1929.999351501 0.00286102 6.63324
  2.59543e+06
  »> TOT > 15590
  stop at Wed Feb 1 19:37:00 2006

DONE:

$\Rightarrow$ On C6, C7 data and on the WRs data

- production of the FFTs Data Base;
- production of the peakmap files;
- production of the Log files.
- To be done in the next days: A detailed study of Log files to characterize/understand the quality of the data, for the search of continuous signals (with possible contributions to the the detector characterization).