Turun yliopisto
University of Turku

Tuorla Observatory
University of Turku

Statistics

Prologue/Esipuhe

hlehto@utu.fi

Very basic stats

Testing hypothesis:

Accept or
Reject
Critical limits 0.05,0.01 0.001
One or two sided?

- Test single r.v. (t-test)
- Test (r.v)^2 (χ^2-test, analysis of variances)
- Test r.v./$\sqrt{(\chi 2/DF)}$ (t-test)
- Test $(\chi 2/DF_1)/(\chi 2/DF_2)$ (F(k,m) -test)

- DF, DOF, ν (usually N (minus) parameters needed for fit)

- Discrete: Poisson, Beta and hypergeometric distributions.

- In a (x,y) data set, you can tell if a constant value is accepted (eg. $p<0.01$)
- Or a linear slope is accepted (eg. $p<0.01$)
- You can even check if a 2nd order polynomial gives you a good fit (e.g $p<0.001$)

- But you cannot tell which one of these is the best fit. You cannot compare directly the p's above in the frame of classical analysis

END OF PROLOGUE!

Turun yliopisto
University of Turku

Long datasets, noise, sampling

Harry J Lehto
University of Turku
Department of Physics and Astronomy
Tuorla Observatory

~10^2-~10^4 data points, stationary, some white noise, evenly sampled

- Not heavy on computer (N not too large)
- Get reasonable statistics (N not too small)
- Easy to visualize (fits on one screen)
- Noise well behaving (white noise does not mean it is Gaussian!)
- No correlated noise
- Data set is well behaving (in general)
- Rigorous mathematics works generally

Analysis

- $\langle x \rangle$, s^2, DF
- Wavelets

  (assume n→ infty)
- Structure functions
- FFT, DFT

$\langle x \rangle$,s^2, DF

- $\langle x \rangle$, DF = n−1
- s^2, DF = n−2
- Sample mean and standard deviation. What are the true mean and standard deviations?
- Note:
  - no assumptions on the properties of x, or s^2.
  - $\langle x \rangle$, and s^2 have Gaussian distributions due to central limit theorem (when N is large enough).
  - DF usually straightforward

FFT, DFT

- Breaking time series into Fourier components

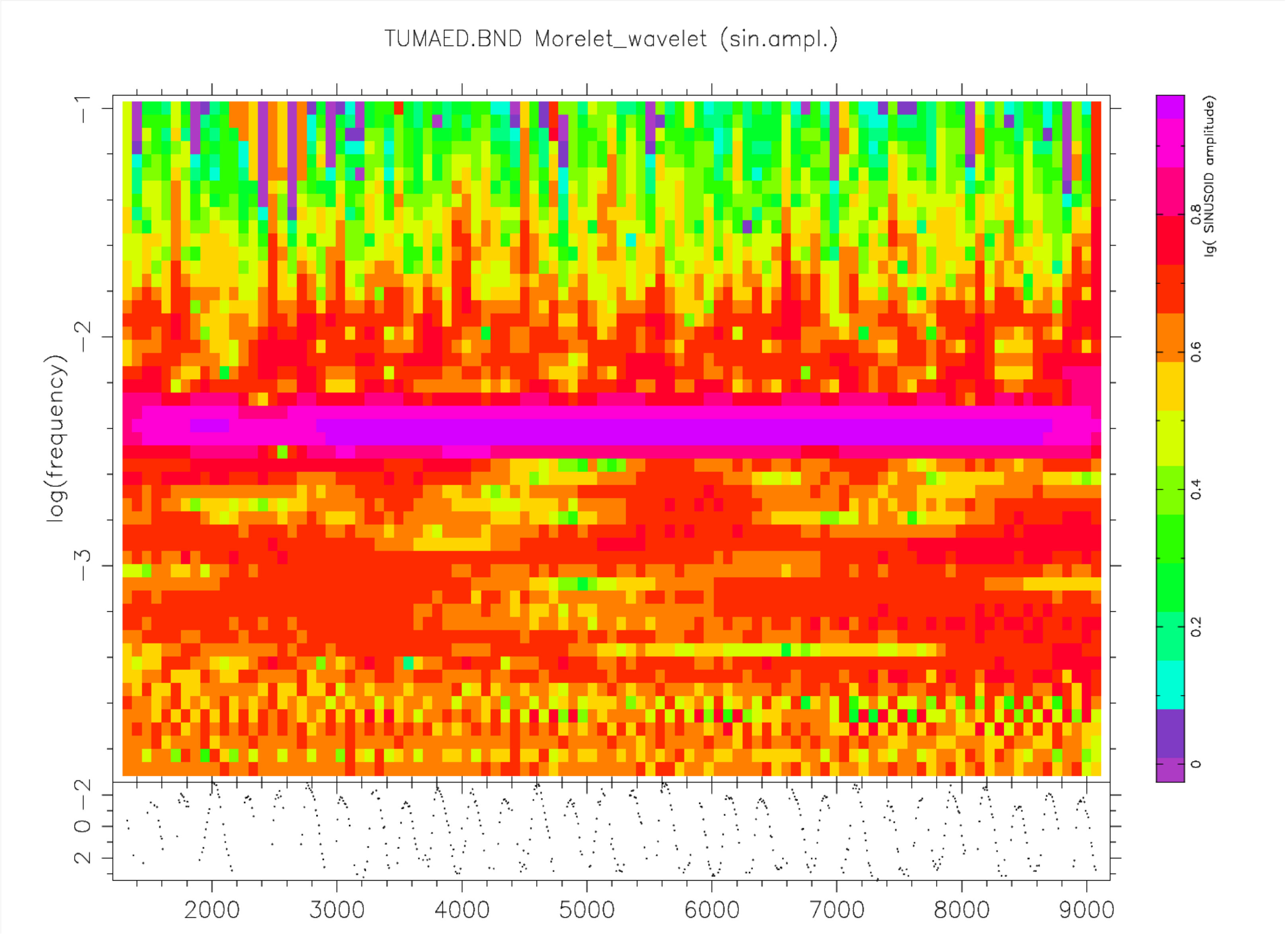
$$F(\nu)=\int_{-\infty}^{\infty} f(t)e^{-j2\pi\nu t}dt$$

- For discrete cases

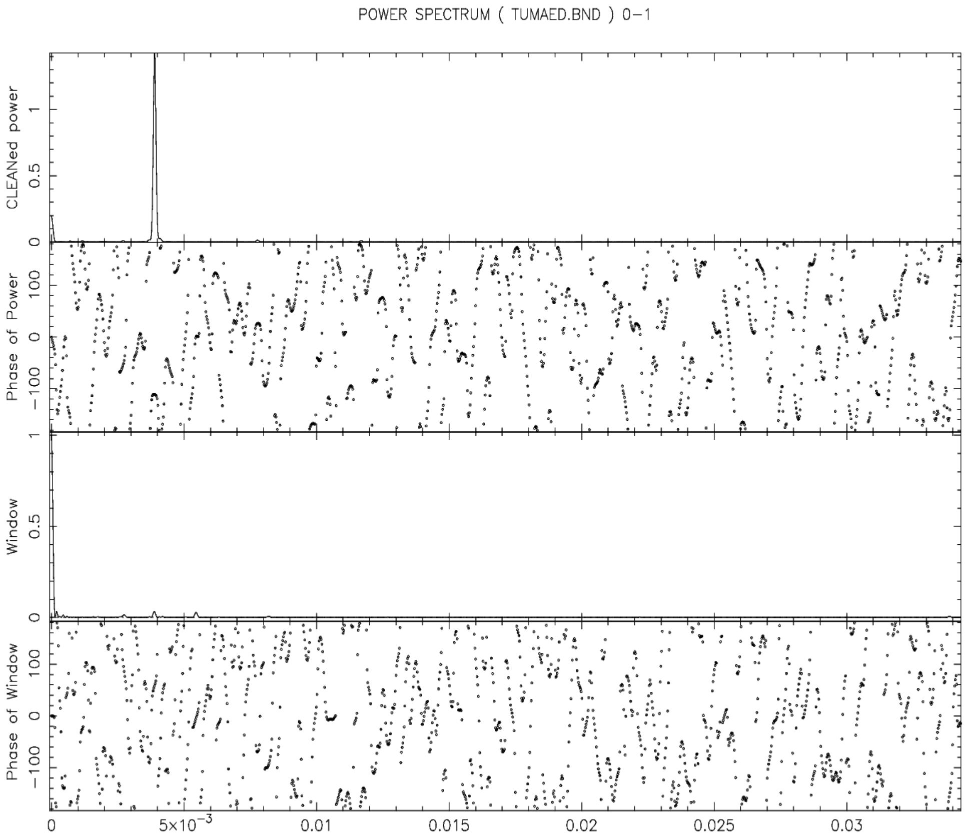
$$F(\nu_i)=\Sigma_{\text{all k}} f_k e^{-j2\pi\nu_i t_k}dt_k$$

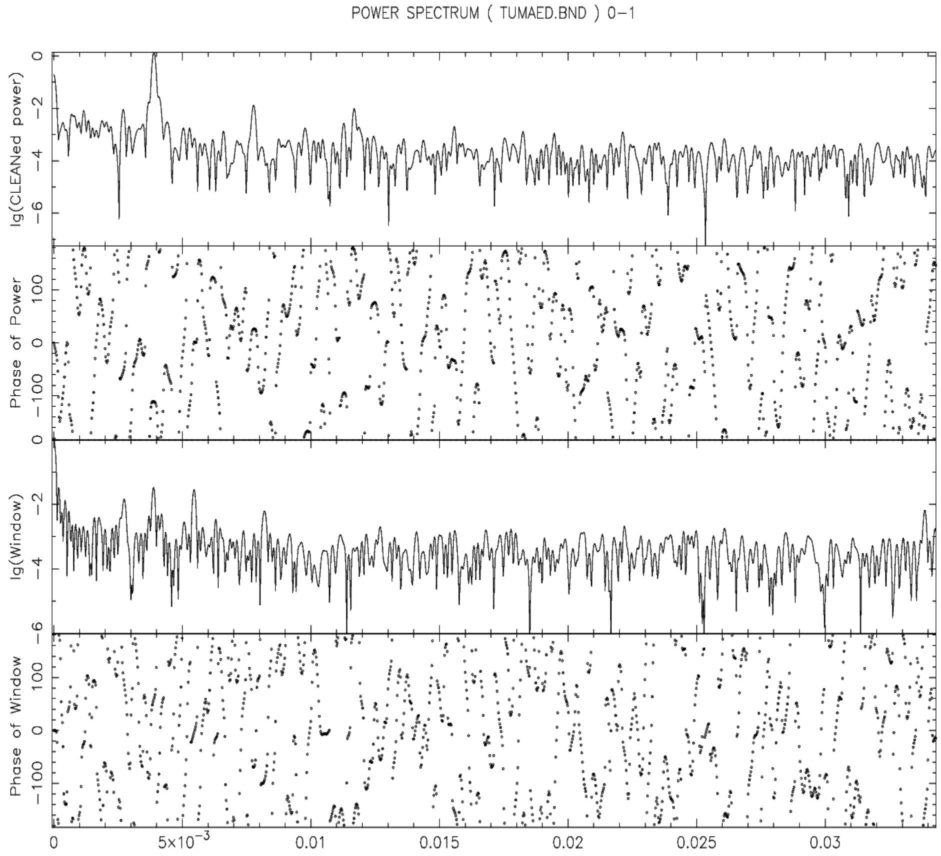
  - Reversible, with suitable sampling
- Several variations. Some more suitable for noise and some signal searching.
- DFT is a *n*log*n* process, FFT is a n^2 process
- Stationarity expected in principle
- Power spectrum = F(*ν*)F*(*ν*) – phase conserving,
  - or as Fourier transform of autocovariance function – phase destroying.

T Uma /SVSO


TUMAED.BND Morelet_wavelet (sin.ampl.)
log(frequency)
−1
−2
−3
−2
0
2
2000
3000
4000
5000
6000
7000
8000
9000
lg(SINUSOID amplitude)
0.8
0.6
0.4
0.2
0

TUMa


POWER SPECTRUM (TUMAED.BND) 0–1
CLEANed power
0
0.5
1
Phase of Power
−100
0
100
1
Window
0
0.5
Phase of Window
−100
0
100
0
5×10^{-3}
0.01
0.015
0.02
0.025
0.03


POWER SPECTRUM (TUMAED.BND) 0–1
lg(CLEANed power)
0
−2
−4
−6
Phase of Power
−100
0
100
0
lg(Window)
−2
−4
−6
Phase of Window
−100
0
100
0
5×10^{-3}
0.01
0.015
0.02
0.025
0.03

Structure functions

- $D(\tau)=\langle(x(t+\tau)-x(t))^2\rangle$,
- Used to determine the type of (red) noise
- Does not need to be stationary
- $D(\tau)=\langle(x(t+2\tau)-2x(t+\tau)+x(t))^2\rangle$

  Second order structure function

  = Allen variance

D(τ)

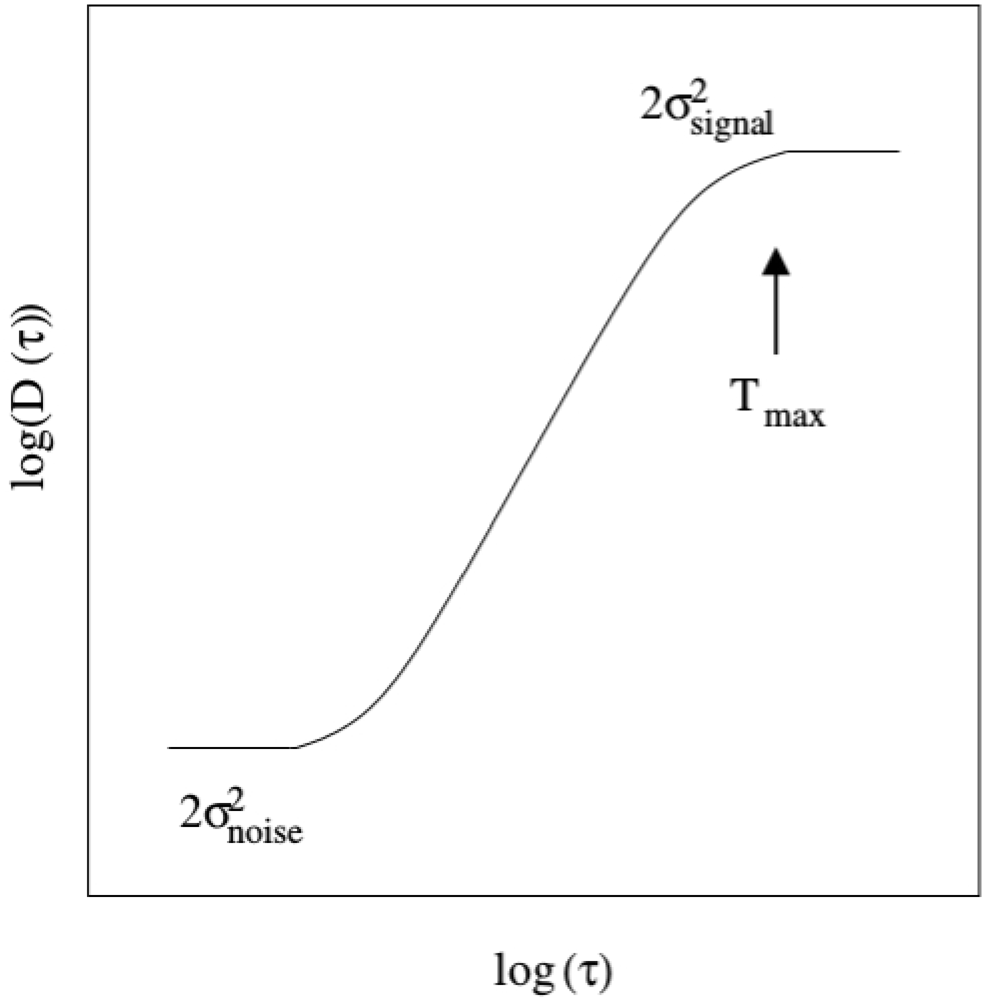


Fig. 1. Ideal structure function.

Adopted

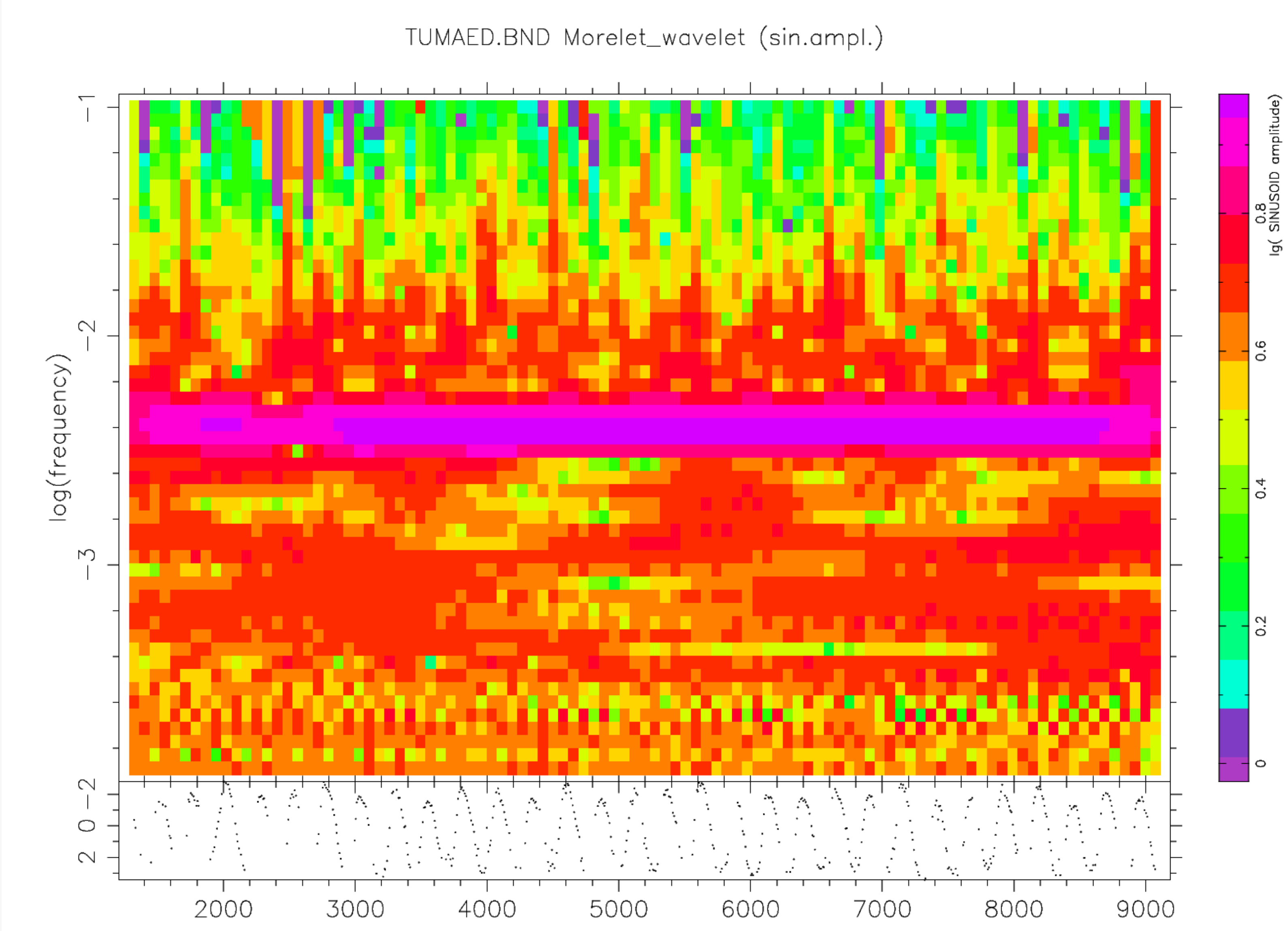
Hovatta et al A&A 469, 899–912 (2007)

Wavelets

- Local analysis – as a function of time location and timescale
- $g^*(f,\tau) = \exp(-icf(t-\tau)-(f(t-\tau))^2)$ *transform*
- $W(f,\tau) = f\cdot(S^2(f,\tau)+C^2(f,\tau))$ *power*
- Wavelet transform is reversible with suitable sampling

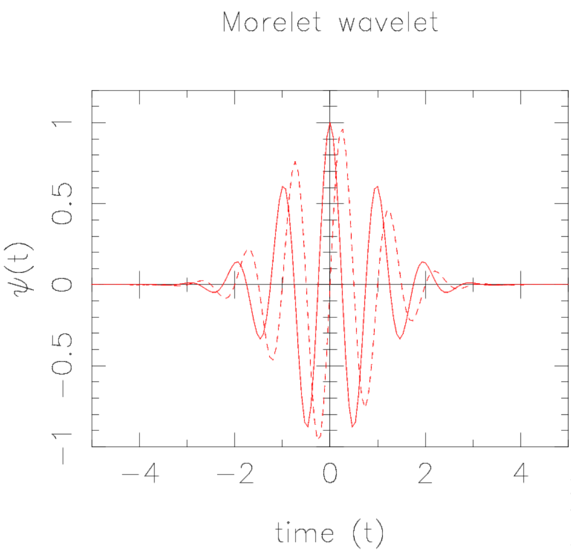
Turun yliopisto
University of Turku

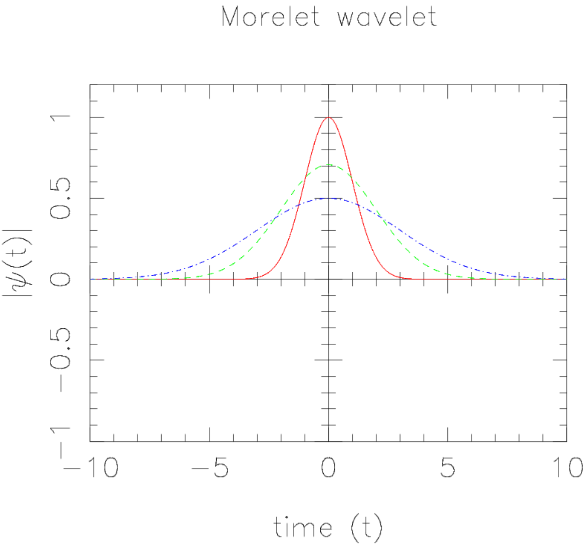
T UMa



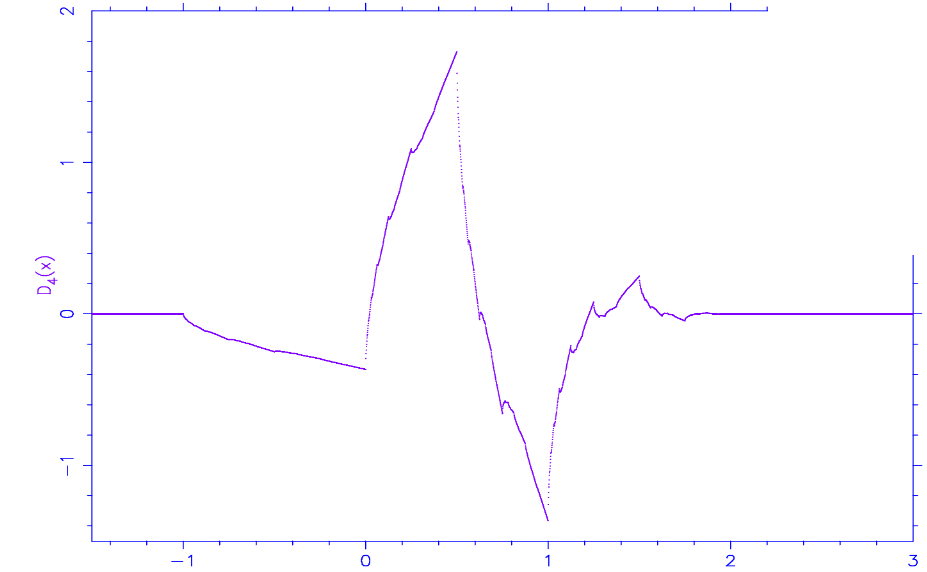
Data from SVSO/URSA
variable star observers
data base

Morelet and D_4 wavelets





Daubechies wavelet D_4



Turun yliopisto
University of Turku

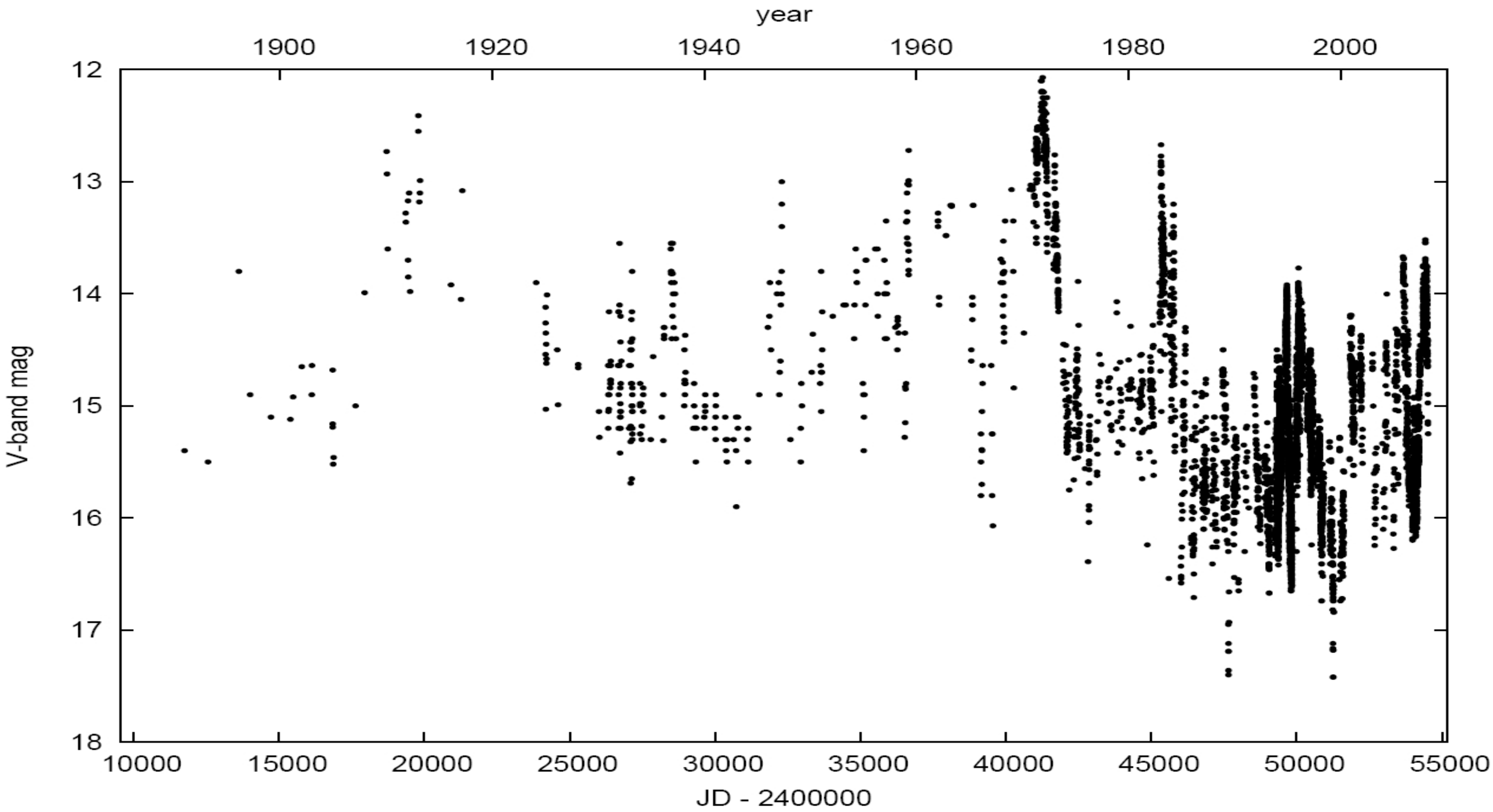
Linear regression

- Special example:
  - Easy to do with a calculator
  - Note the least squares linear regression is valid when
    - x is known in principle exactly
    - y is a random sample drawn from a Gaussian sample centered on the true y – usually from the same Gaussian distribution.
    - But this has potentially some serious problems.

Some real cases and challenges

Turun yliopisto
University of Turku

Sample average and variance: OJ287



Tuorla obswervatory
OJ287 database

Turun yliopisto
University of Turku

DFT

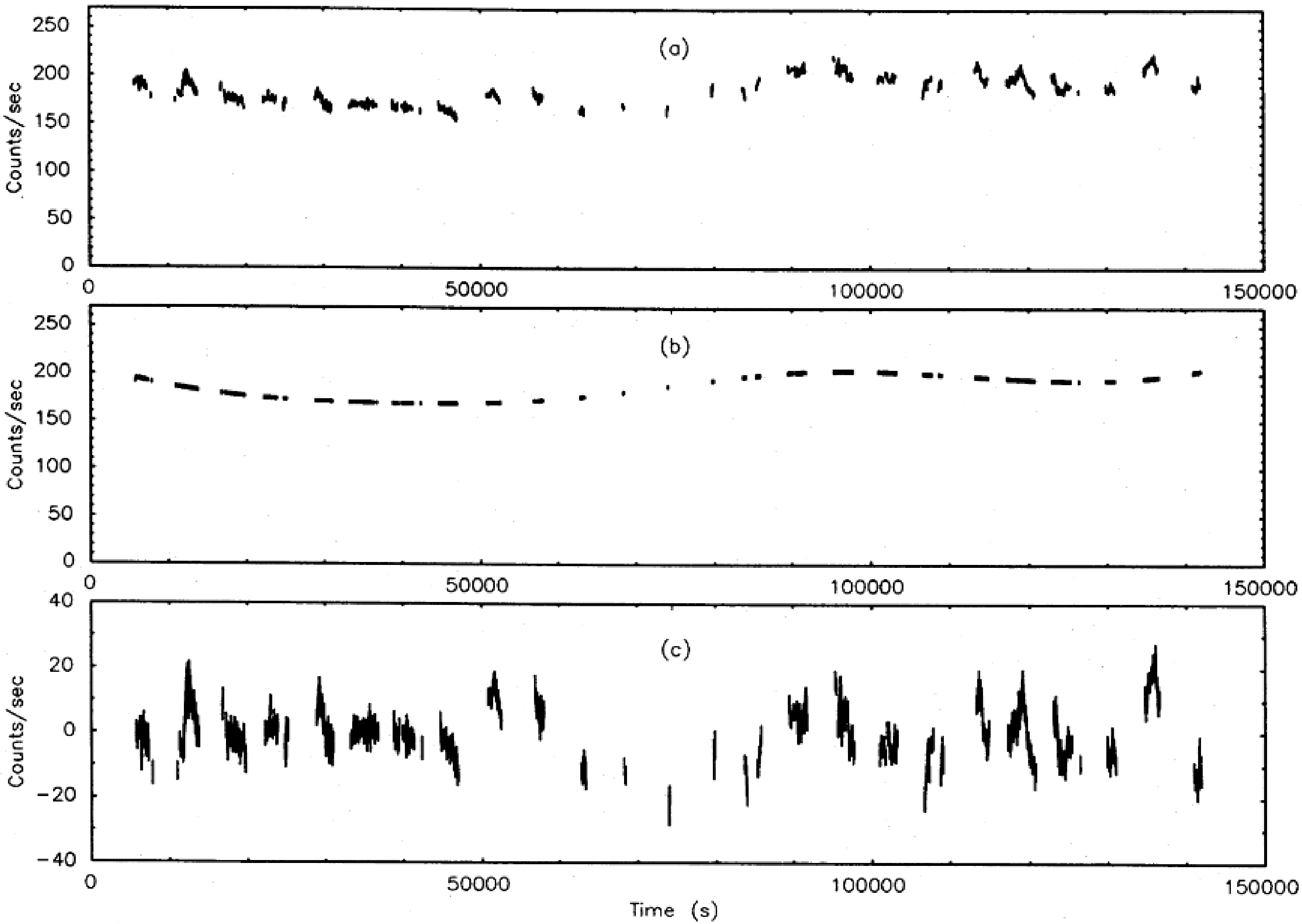
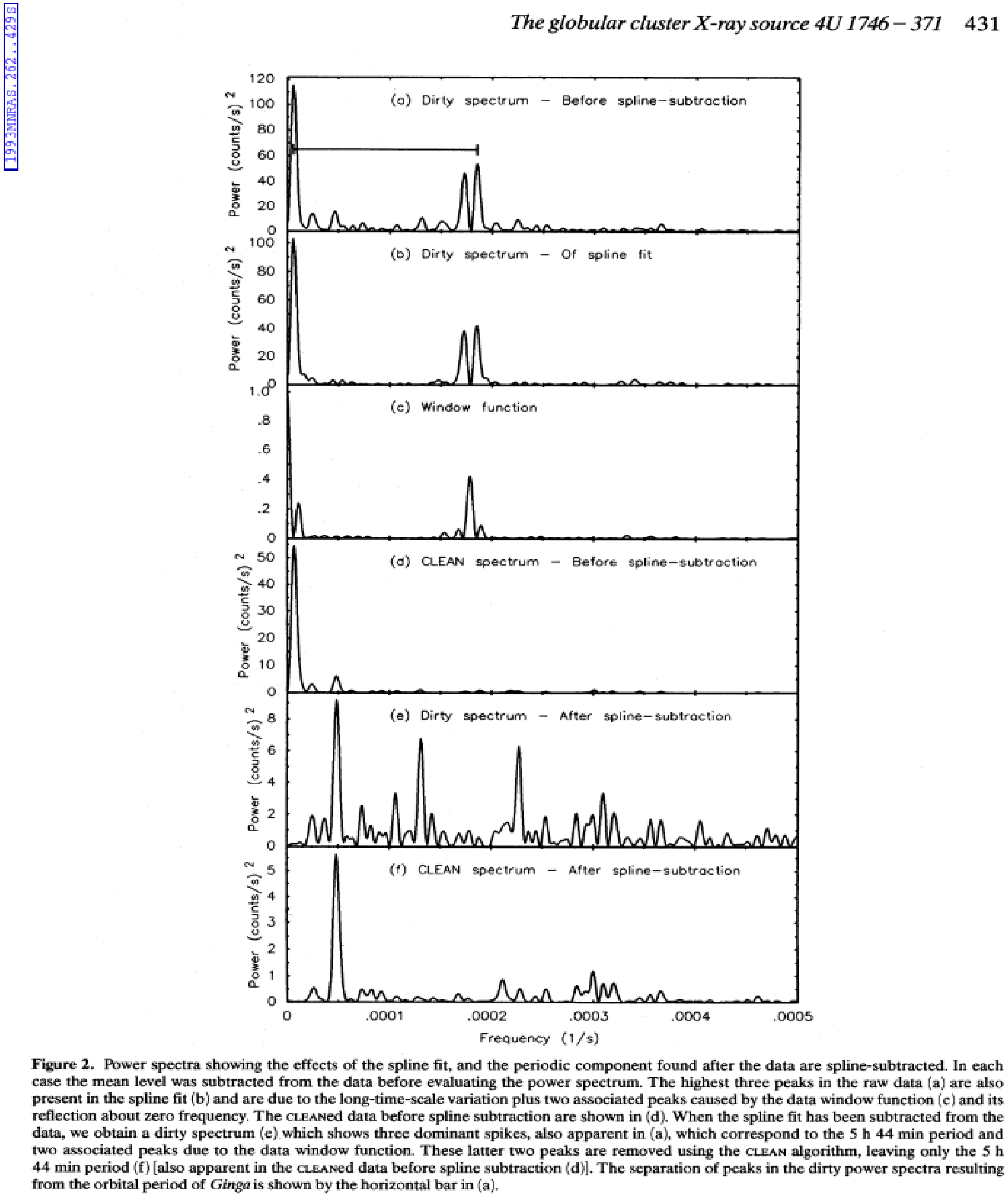


Figure 1. *Ginga* time series of 4U 1746 − 371 (2–17 keV): (a) after background subtraction and aspect correction; (b) a spline fit to the data; and (c) with the long-time-scale spline subtracted.

Turun yliopisto
University of Turku

DFT/CLEAN

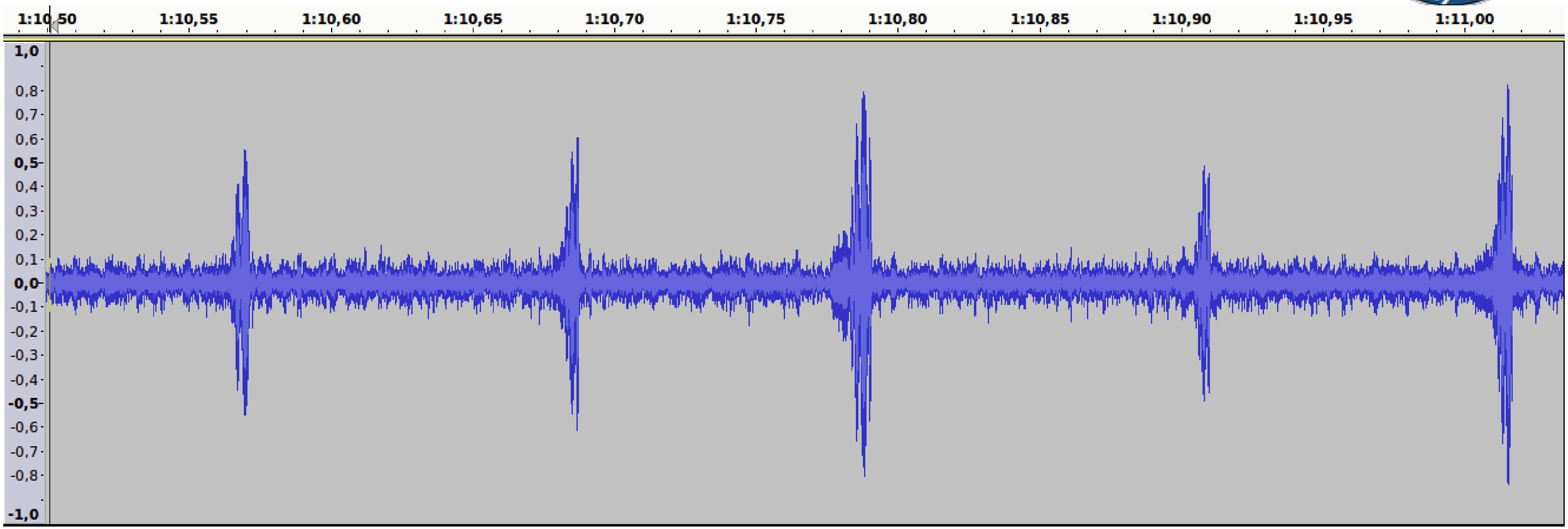


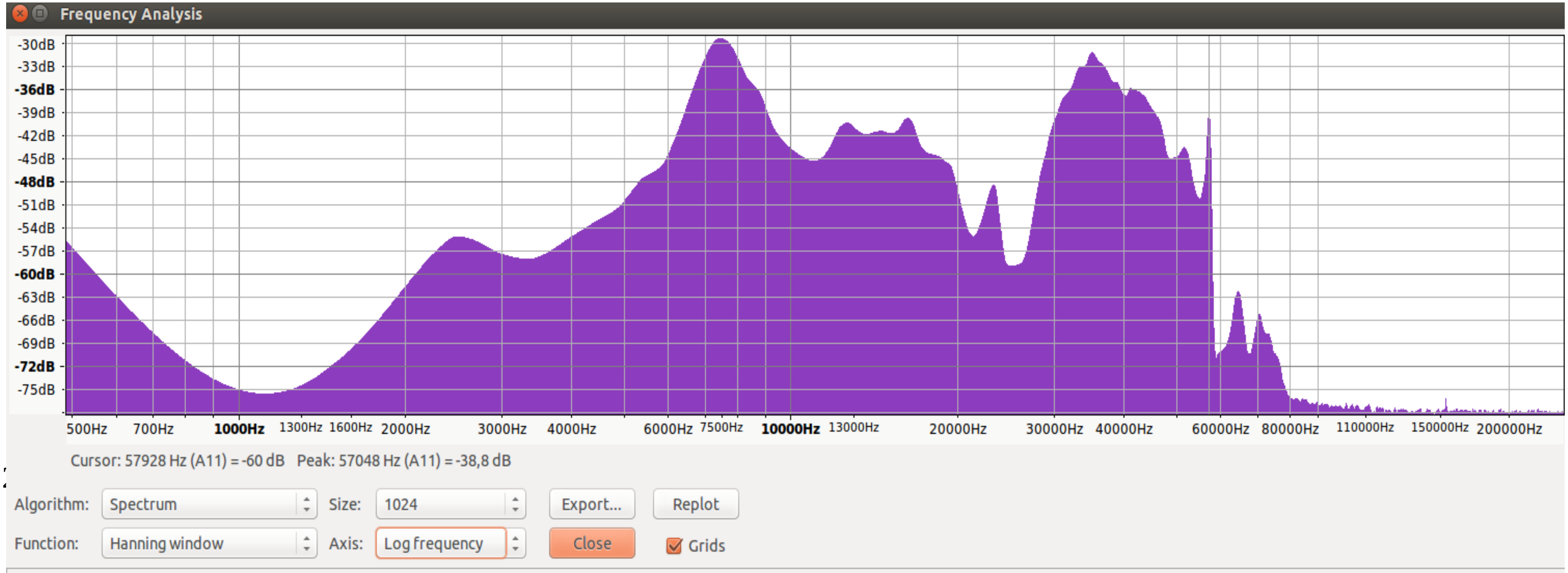
The globular cluster X-ray source 4U 1746 – 371 431

Figure 2. Power spectra showing the effects of the spline fit, and the periodic component found after the data are spline-subtracted. In each case the mean level was subtracted from the data before evaluating the power spectrum. The highest three peaks in the raw data (a) are also present in the spline fit (b) and are due to the long-time-scale variation plus two associated peaks caused by the data window function (c) and its reflection about zero frequency. The CLEANed data before spline subtraction are shown in (d). When the spline fit has been subtracted from the data, we obtain a dirty spectrum (e) which shows three dominant spikes, also apparent in (a), which correspond to the 5 h 44 min period and two associated peaks due to the data window function. These latter two peaks are removed using the CLEAN algorithm, leaving only the 5 h 44 min period (f) [also apparent in the CLEANed data before spline subtraction (d)]. The separation of peaks in the dirty power spectra resulting from the orbital period of *Ginga* is shown by the horizontal bar in (a).

FFT

- Fill small gaps
- Interpolate
- Taper
- No good solutions for interpolation
- Pathological features if trying to recover the original attempted
- Works well for 2^n data points

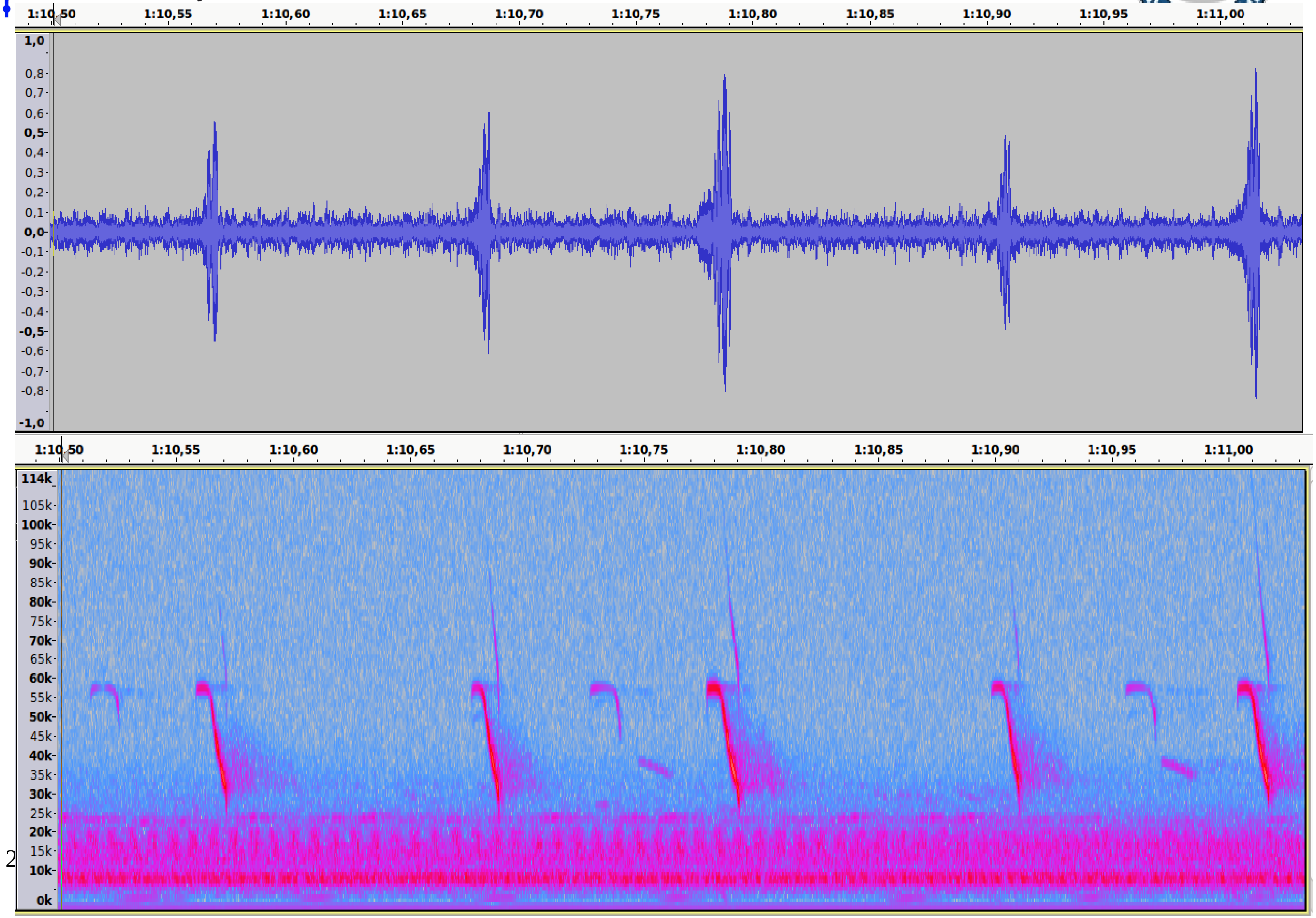




Cursor: 57928 Hz (A11) = -60 dB   Peak: 57048 Hz (A11) = -38,8 dB

Algorithm: Spectrum   Size: 1024

Function: Hanning window   Axis: Log frequency   ☑ Grids


1:10,50
1:10,55
1:10,60
1:10,65
1:10,70
1:10,75
1:10,80
1:10,85
1:10,90
1:10,95
1:11,00
1,0
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
-0,2
-0,3
-0,4
-0,5
-0,6
-0,7
-0,8
-1,0
1:10,50
1:10,55
1:10,60
1:10,65
1:10,70
1:10,75
1:10,80
1:10,85
1:10,90
1:10,95
1:11,00
114k
105k
100k
95k
90k
85k
80k
75k
70k
65k
60k
55k
50k
45k
40k
35k
30k
25k
20k
15k
10k
0k

Structure function

910 T. Hovatta et al.: Statistical analyses of long-term variability of AGN at high radio frequencies

A&A 469, 899–912 (2007)

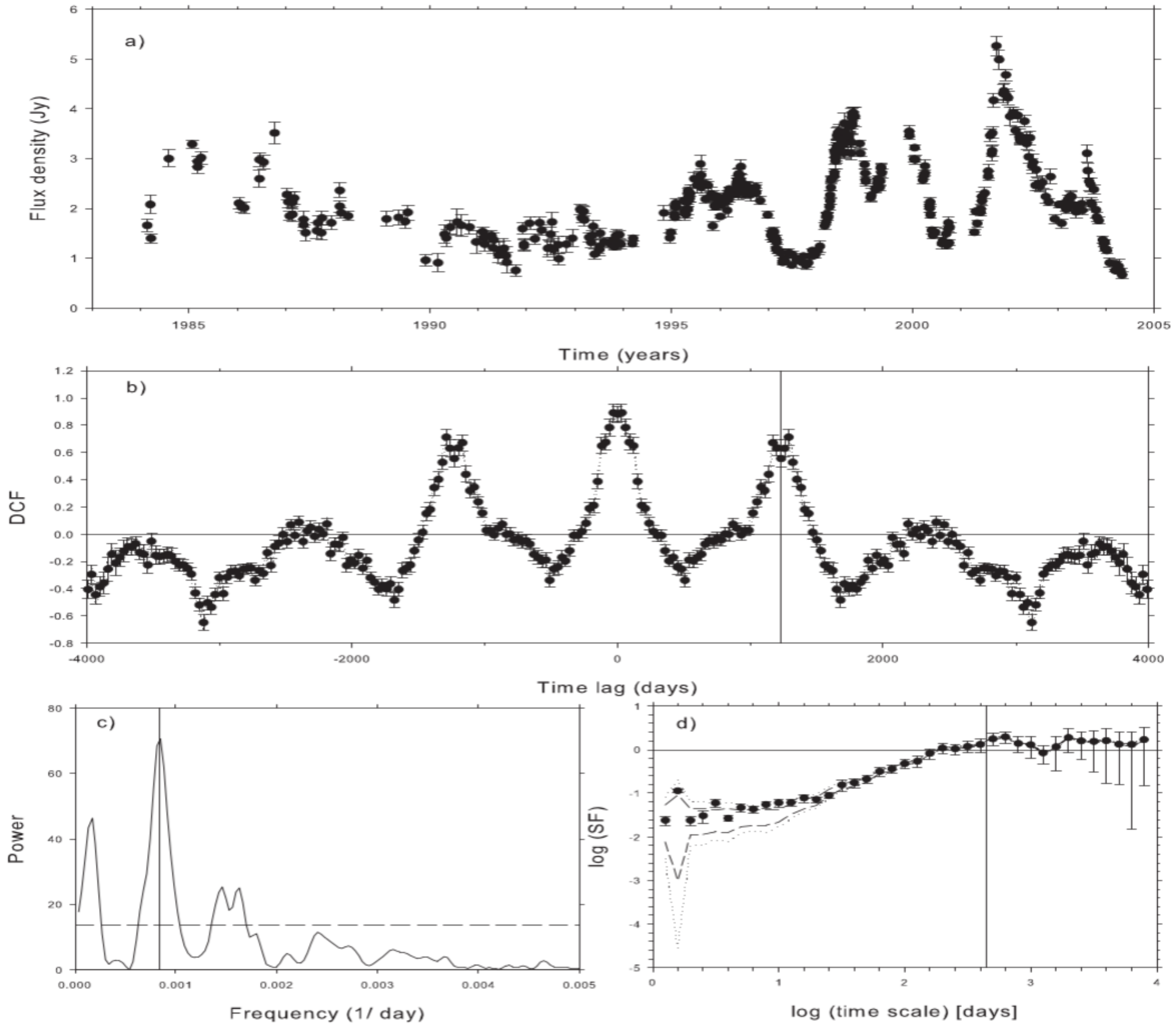


Fig. 11. Analyses of the HPQ source 1156+295 at 22 GHz. **a)** Flux Density curve. **b)** The discrete correlation function. The 99.5% significance level is shown with dotted line. **c)** The Lomb-Scargle periodogram. Dashed line shows the false-alarm probability. **d)** The structure function. Dashed and dotted lines show the 97.5% and 99.5% significance levels. Time scales obtained with each method are marked by vertical lines. The most significant spike of the periodogram is at time scale of 3.29 years, which is 0.2 years shorter than the first correlation in the DCF at 3.49 years. The SF gives a time scale of 1.21 years.

Structure function

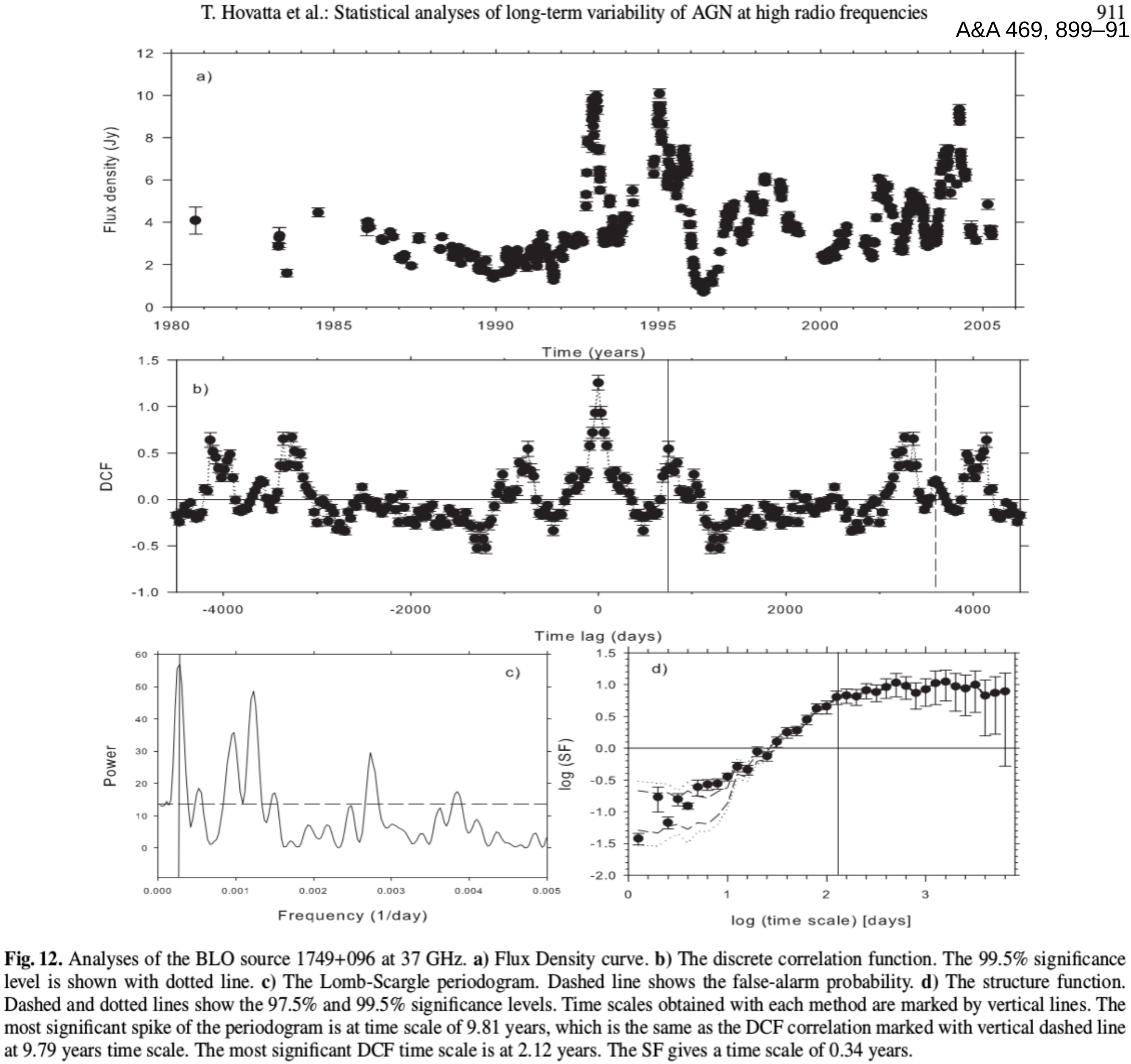
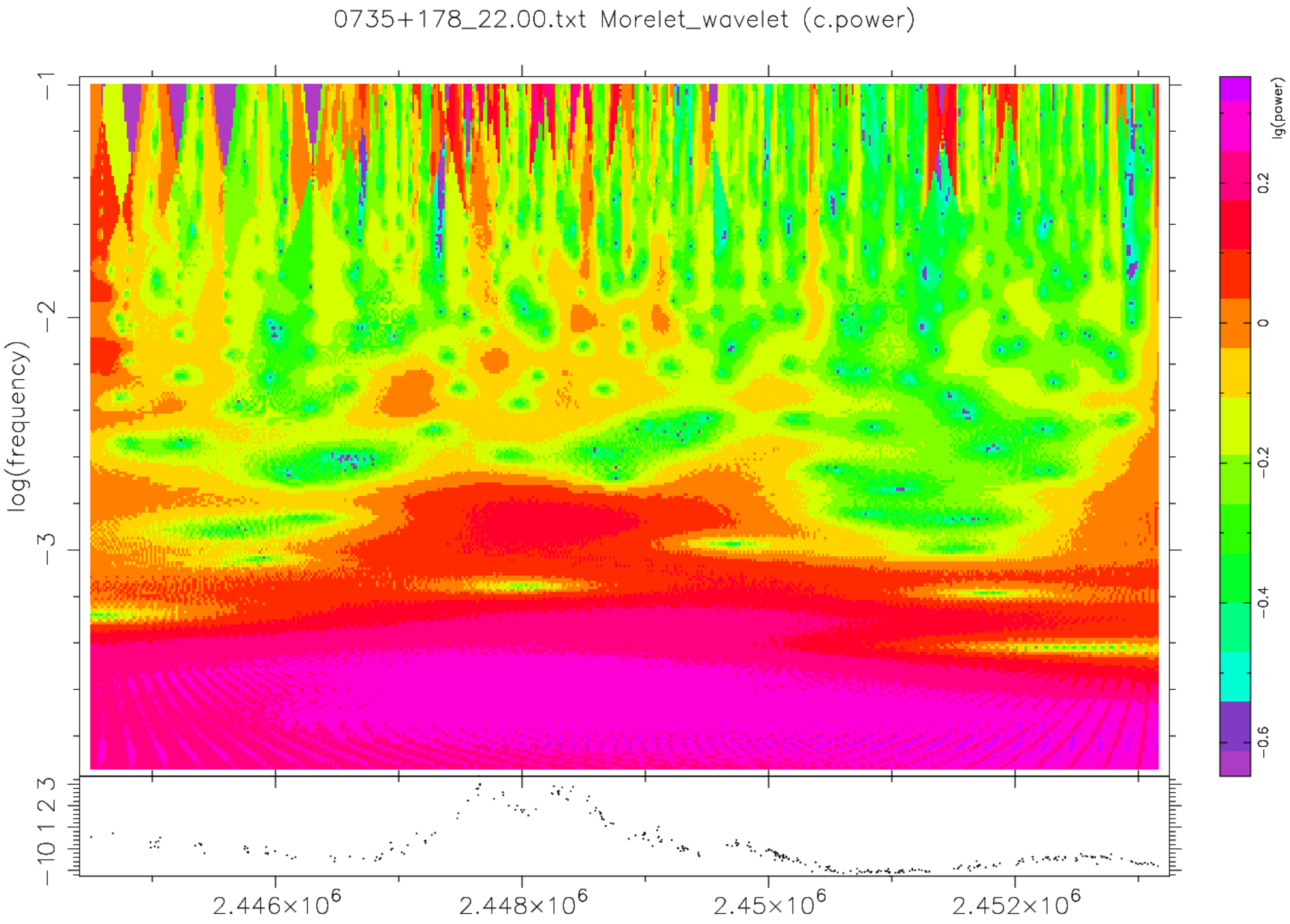

T. Hovatta et al.: Statistical analyses of long-term variability of AGN at high radio frequencies 911

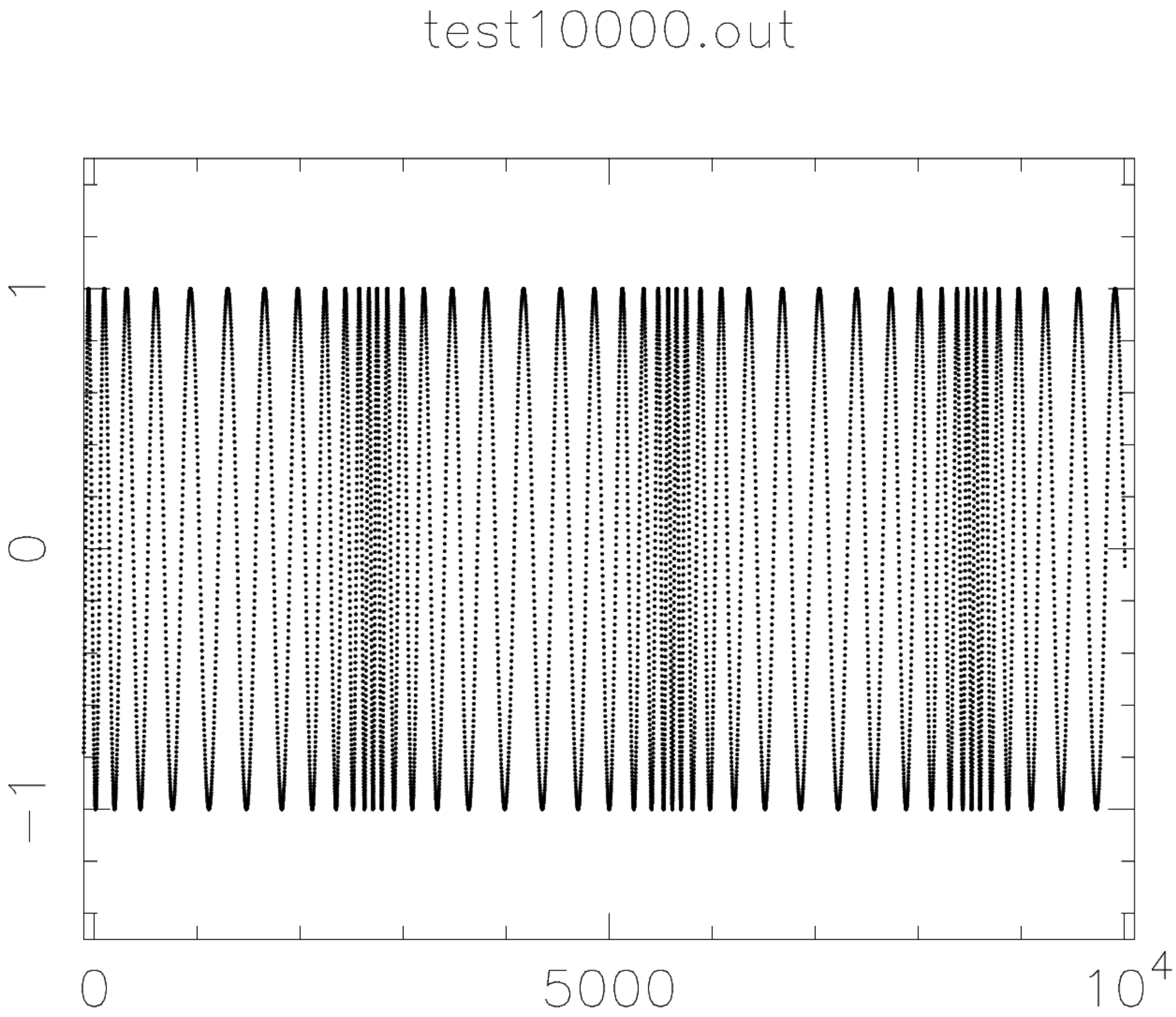
Fig. 12. Analyses of the BLO source 1749+096 at 37 GHz. **a)** Flux Density curve. **b)** The discrete correlation function. The 99.5% significance level is shown with dotted line. **c)** The Lomb-Scargle periodogram. Dashed line shows the false-alarm probability. **d)** The structure function. Dashed and dotted lines show the 97.5% and 99.5% significance levels. Time scales obtained with each method are marked by vertical lines. The most significant spike of the periodogram is at time scale of 9.81 years, which is the same as the DCF correlation marked with vertical dashed line at 9.79 years time scale. The most significant DCF time scale is at 2.12 years. The SF gives a time scale of 0.34 years.

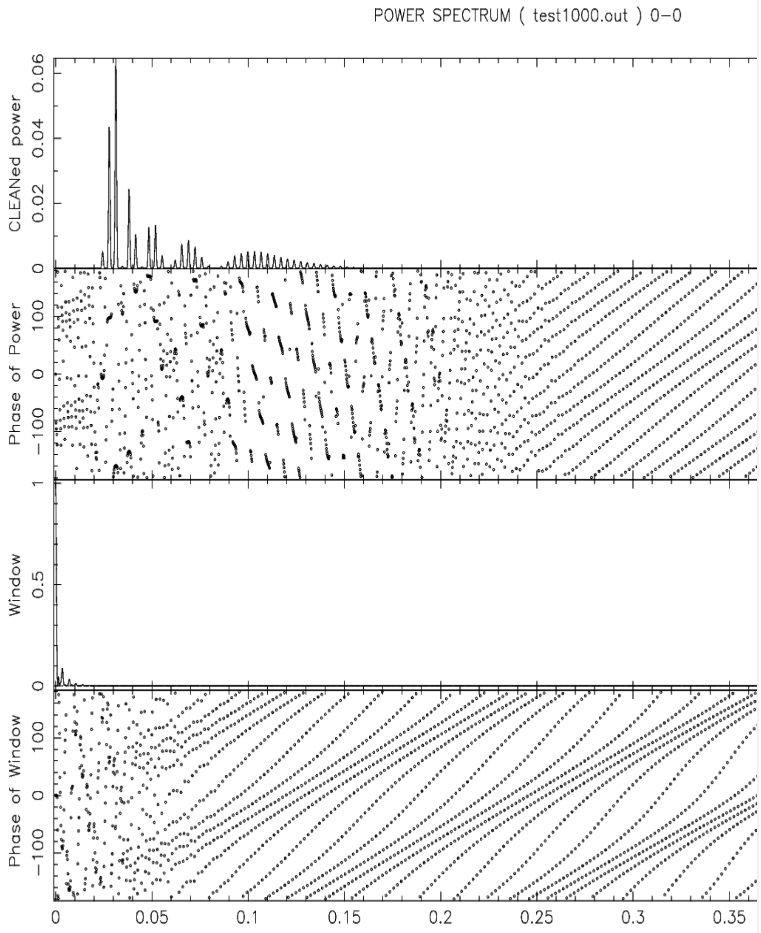
A&A 469, 899–912 (2007)

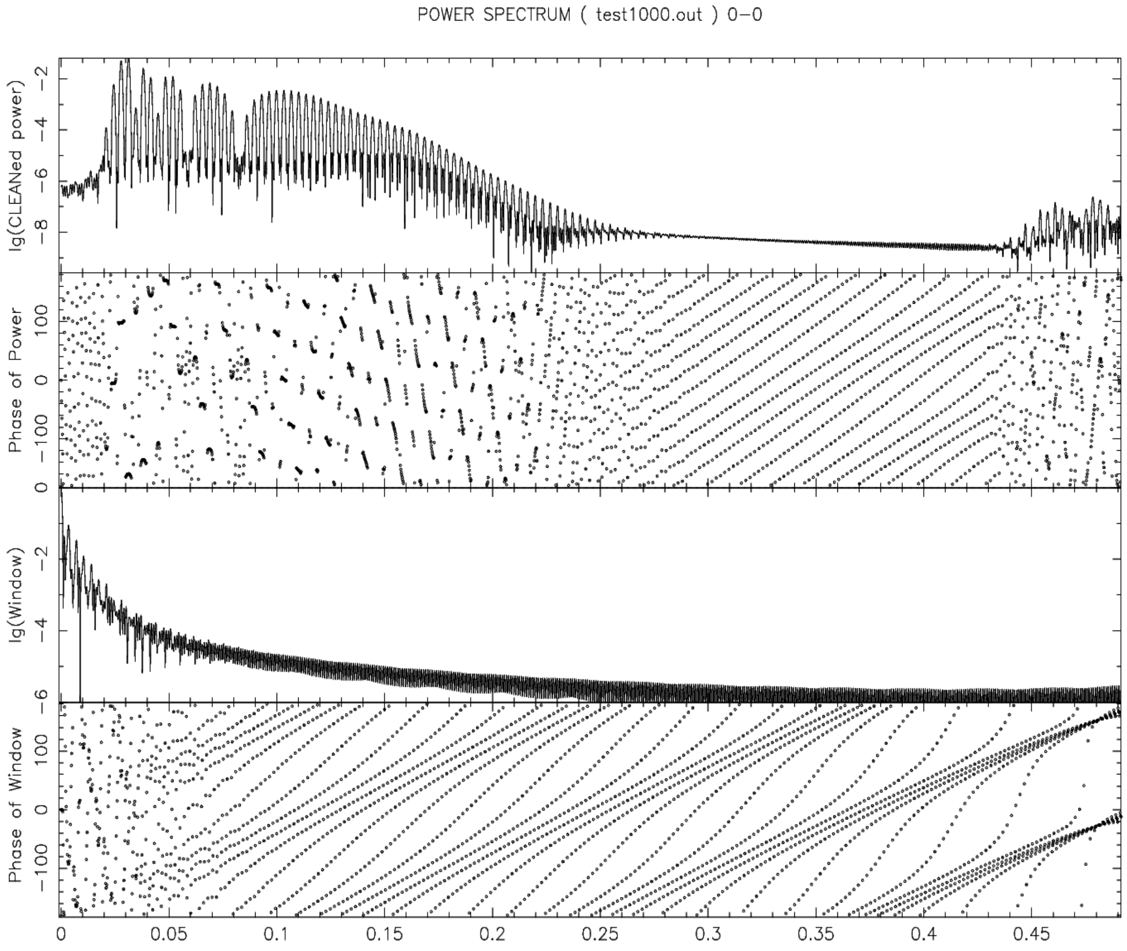
Wavelets

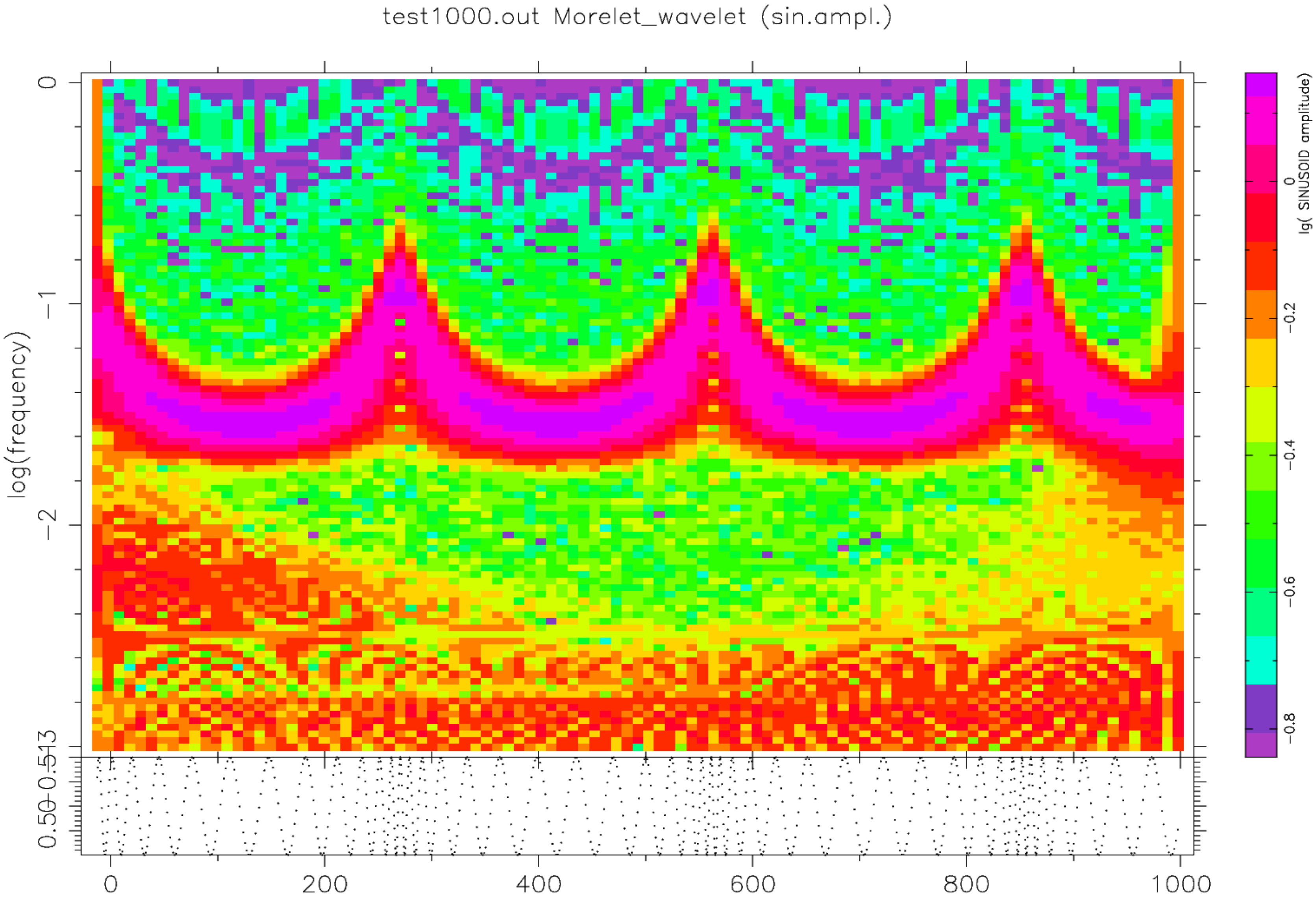

0735+178_22.00.txt Morelet_wavelet (c.power)
log(frequency)
−1
−2
−3
lg(power)
0.2
0
−0.2
−0.4
−0.6
3
2
1
0
−1
2.446×10^6
2.448×10^6
2.45×10^6
2.452×10^6

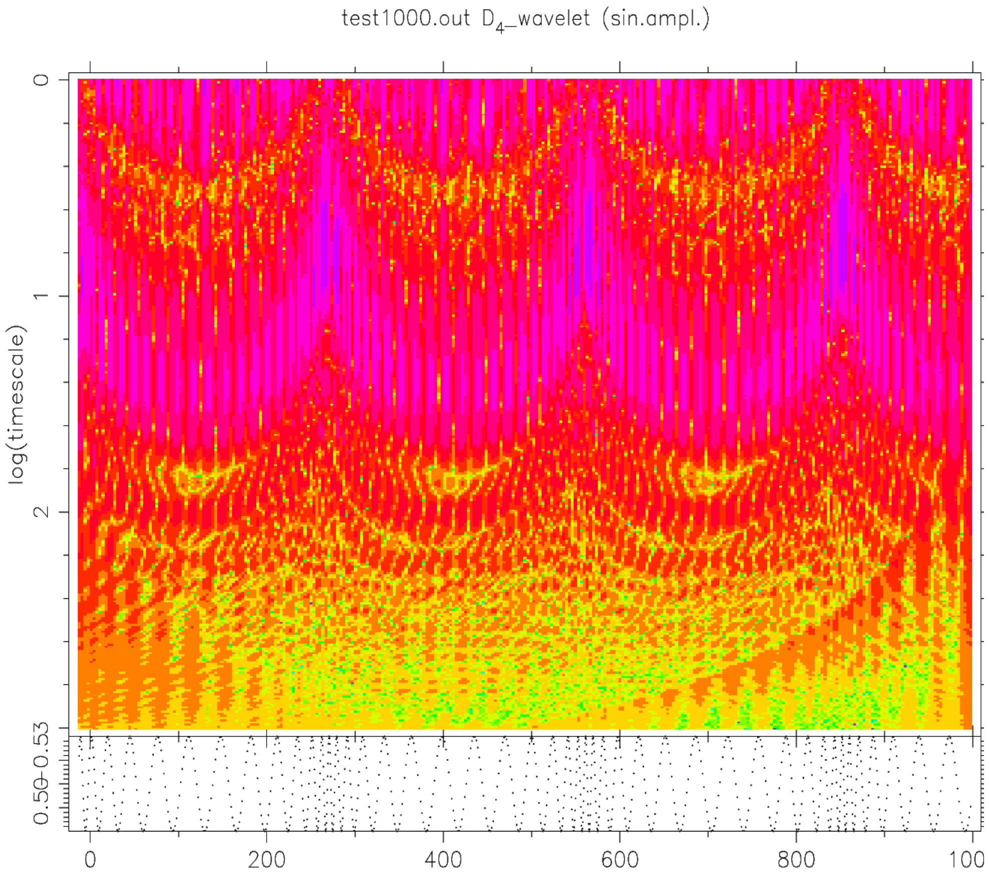
Variable oscillating frequency



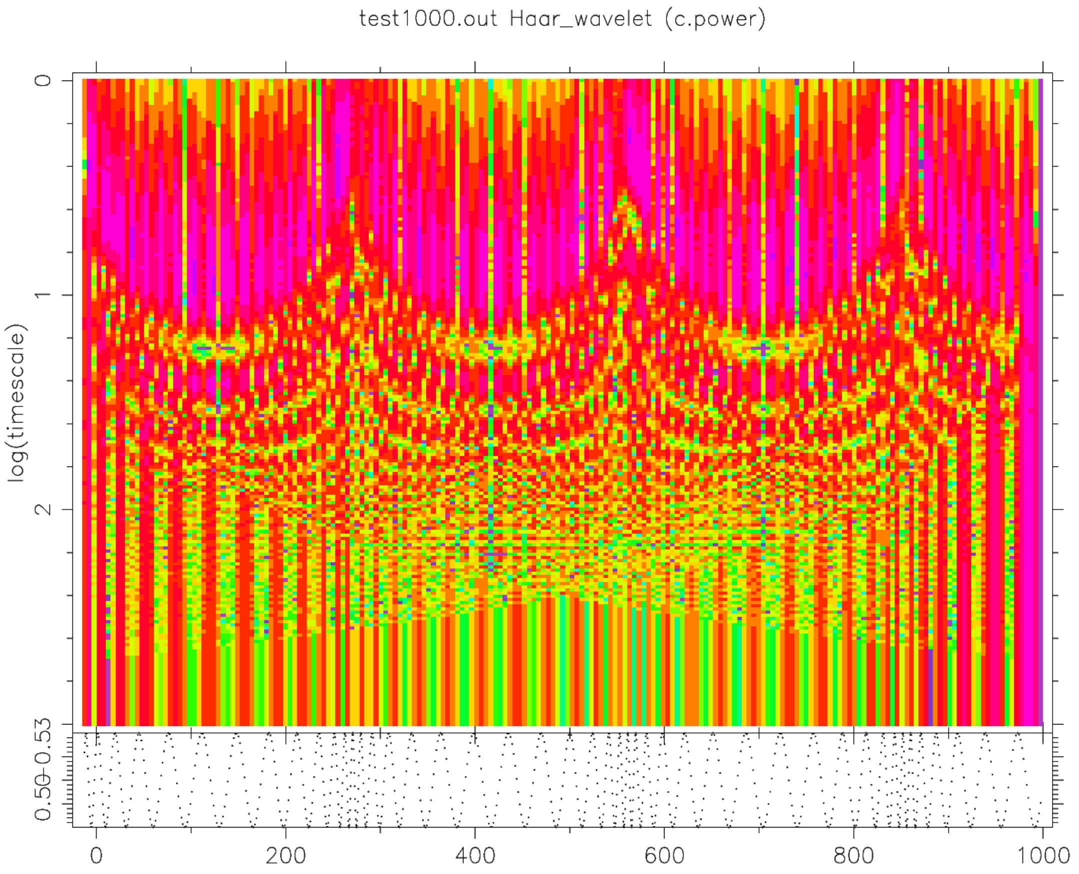

POWER SPECTRUM (test1000.out) 0–0
CLEANed power
0
0.02
0.04
0.06
Phase of Power
−100
0
100
Window
0
0.5
1
Phase of Window
−100
0
100
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35


POWER SPECTRUM (test1000.out) 0–0
lg(CLEANed power)
−8
−6
−4
−2
Phase of Power
0
−100
0
100
lg(Window)
−6
−4
−2
Phase of Window
−100
0
100
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45


test1000.out Morelet_wavelet (sin.ampl.)
log(frequency)
0
−1
−2
−3
0.50
−0.513
0
200
400
600
800
1000
lg(SINUSOID amplitude)
0
−0.2
−0.4
−0.6
−0.8



test1000.out D_4_wavelet (sin.ampl.)



test1000.out Haar_wavelet (c.power)

Linear regression

- There are at least 6 ways of doing a linear regression
- Sometimes you have to linearize first
- OLS(x|y) is the normal ordinary least squares, but you can have also
  - OLS(y|x)
  - Bisector
  - Geometric mean of OLS(y|x) and OLS(x|y)
  - Arthmetic mean of OLS(y|x) and OLS(x|y)
  - orthogonal minimization min(x^2+y^2)
  - Note that all these pass through $\langle x \rangle$, $\langle y \rangle$ and only the slope varies
  - Note that all these have the same correlation coefficient (which really does not depend on the fit in any way)

Bayes

- Now go to another pdf and then back to this...

SEPARATING NON-ORGANIC AND ORGANIC PEAKS

Harry Lehto and Boris Zaprudin

Tuorla Observatory, Dept of Physics and Astronomy,
University of Turku

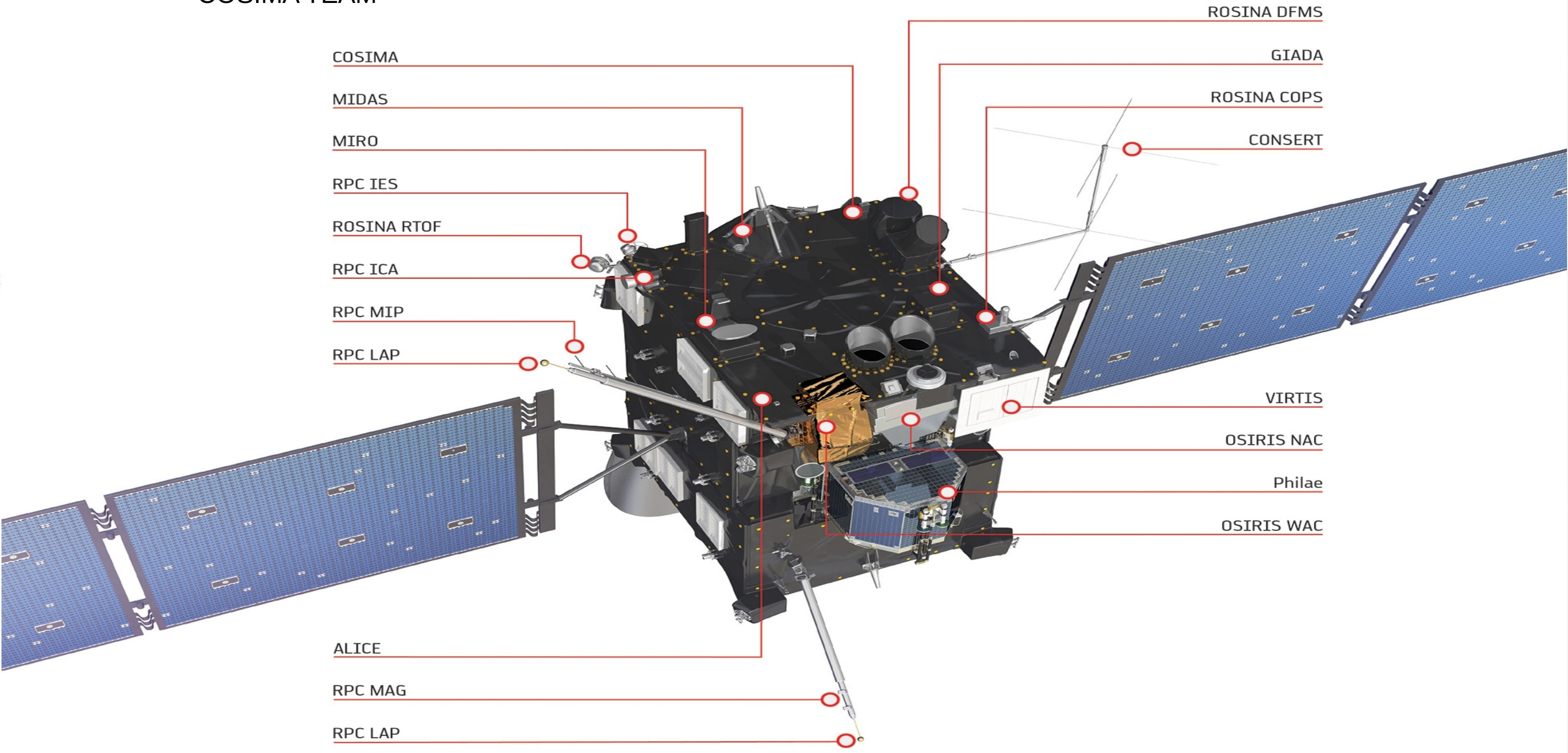
Kirsi Lehto	Biochem, Univ of Turku
Johan Silén	FMI
Tuomo Lönnberg	Biochem, Univ of Turku

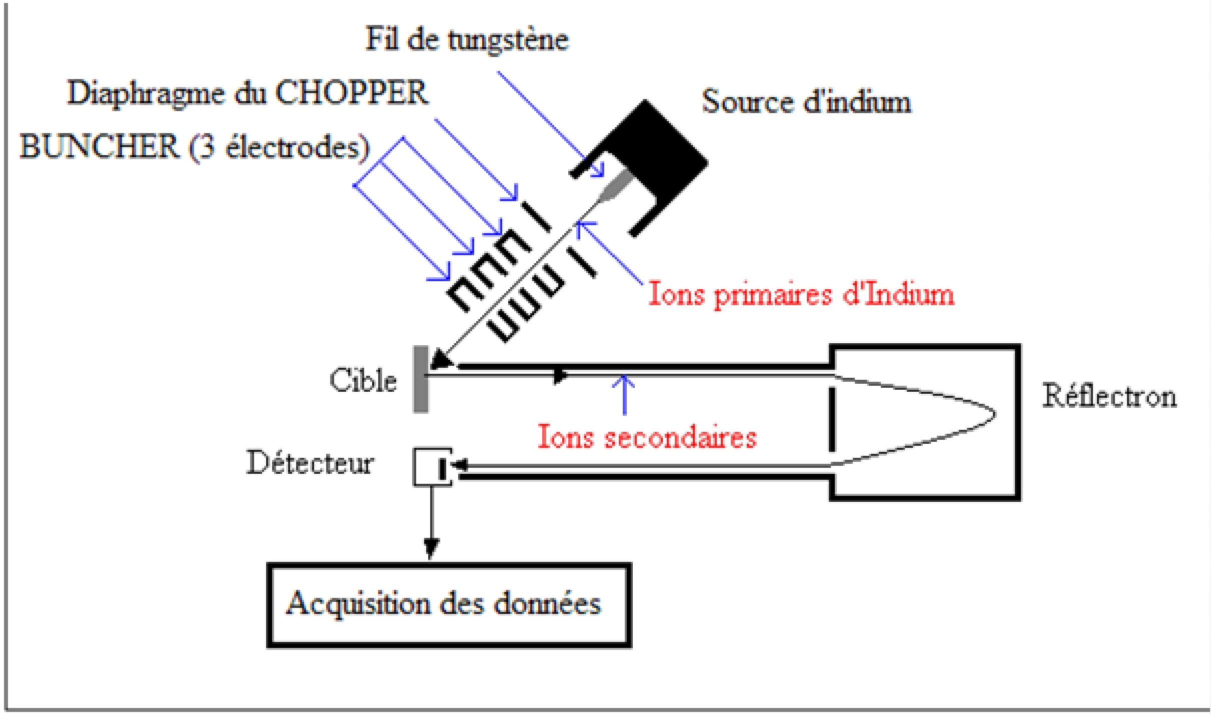
Adopted from Glurns, Italy talk May 2014
And pictures from:

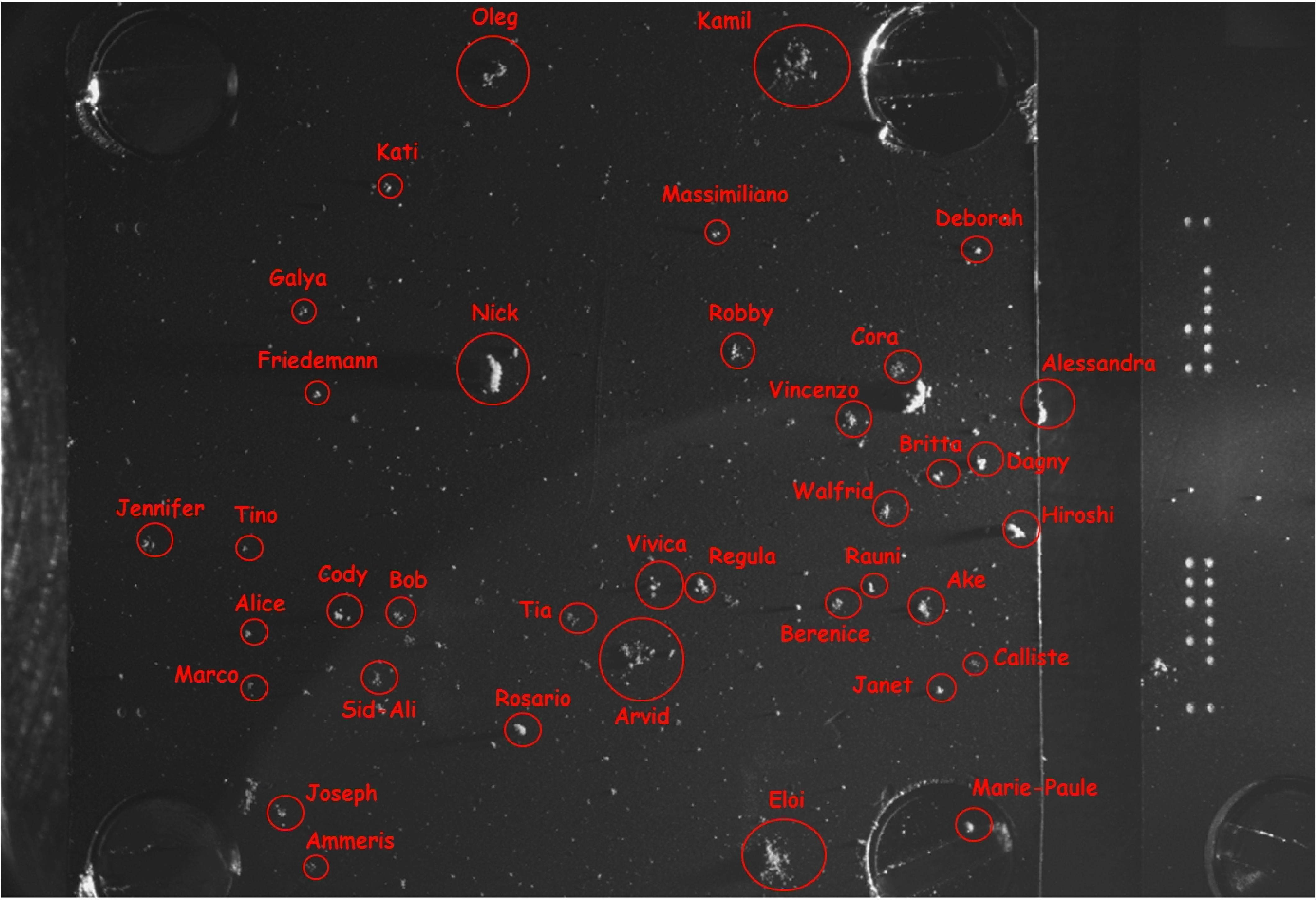
Lehto et al. Geosci. Instrum. Method. Data Syst., 4, 139-148, 2015
http://www.geosci-instrum-method-data-syst.net/4/139/2015/
doi:10.5194/gi-4-139-2015 , which see!

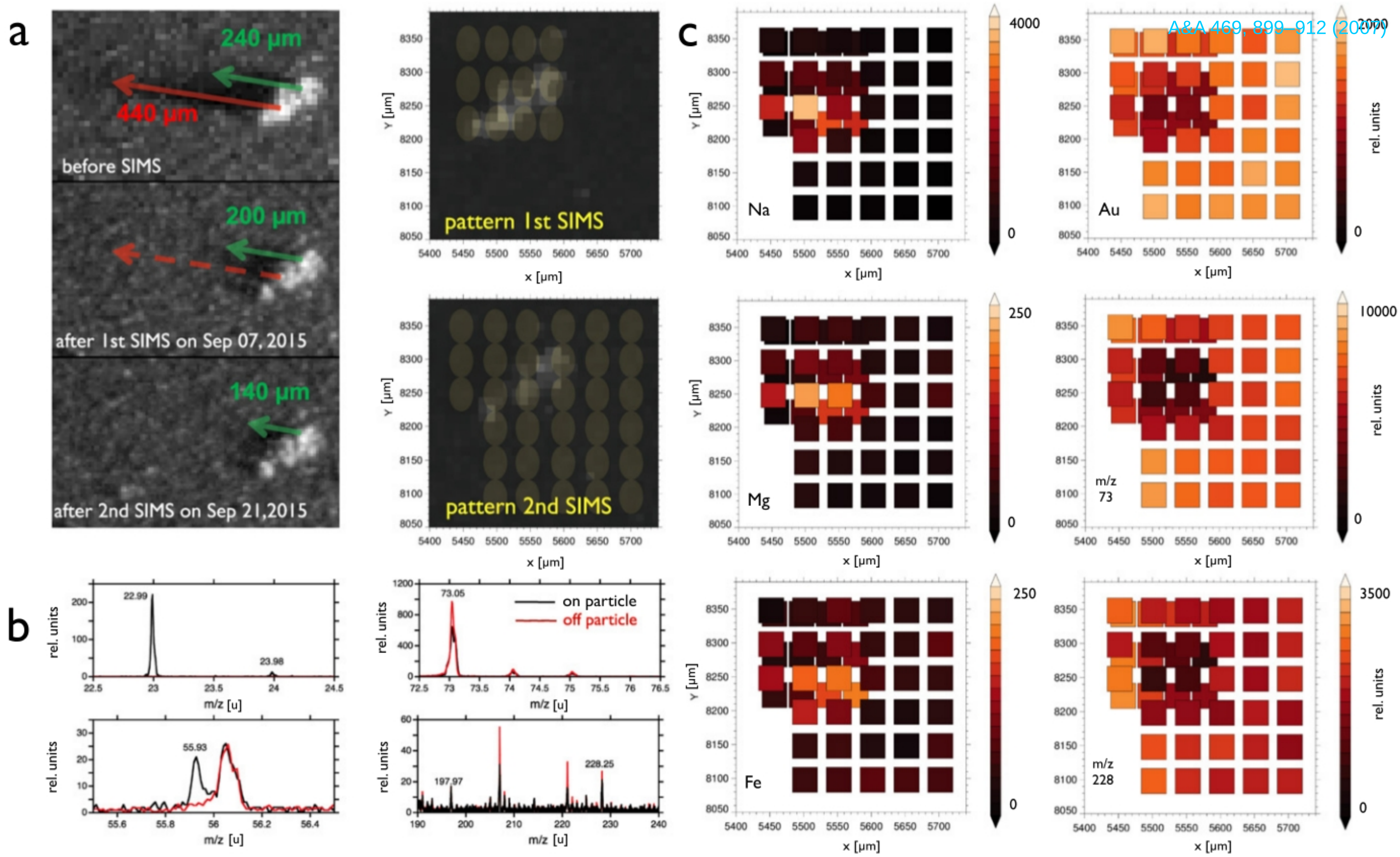
COSIMA TEAM

MPS CSNSM UNIBW TUORLA IWF IAS ESA BUW
ROSETTA
COSIMA
MPE LPC2E LCM FMI UTU LISA UOFC VH&S


COSIMA
MIDAS
MIRO
RPC IES
ROSINA RTOF
RPC ICA
RPC MIP
RPC LAP
ROSINA DFMS
GIADA
ROSINA COPS
CONSERT
VIRTIS
OSIRIS NAC
Philae
OSIRIS WAC
ALICE
RPC MAG
RPC LAP


Fil de tungstène
Diaphragme du CHOPPER
Source d'indium
BUNCHER (3 électrodes)
Ions primaires d'Indium
Cible
Réflectron
Ions secondaires
Détecteur
Acquisition des données


Oleg
Kamil
Kati
Massimiliano
Deborah
Galya
Nick
Robby
Cora
Friedemann
Alessandra
Vincenzo
Britta
Dagny
Walfrid
Jennifer
Tino
Hiroshi
Vivica
Regula
Rauni
Cody
Bob
Ake
Alice
Tia
Berenice
Calliste
Marco
Janet
Sid-Ali
Rosario
Arvid
Joseph
Eloi
Marie-Paule
Ammeris



A&A 469, 899–912 (2007)

Figure 3. Optical images and time-of-flight secondary ion mass spectrometry (TOF-SIMS) spectra and distribution images of particle Donia. (a) Images taken before and after two TOF-SIMS analyses applying high electric fields perpendicular to the target. Top left image shows the particle after collection, middle one after the first and the bottom one after the second measurement sequence. Particle Donia lost elevated parts as a consequence of the TOF-SIMS analysis as indicated by the length of the particle cast shadows. The panels to the right show the footprints of the two sequential SIMS matrix scans. (b) TOF-SIMS mass spectra of the particle and of the Au-target are shown in the lower left panels. (c) Color coded elemental x–y maps derived from the TOF-SIMS spectra for Na^+, Mg^+, and Fe^+, as well as $m/z = 73.05$ u (PDMS), Au^+, and $m/z = 228.25$ u.

Hilchenbach, M et al : ApJL 469, 816:L32 (6pp) (2016)

BAYESIAN METHOD – SIMPLIFIED IDEA

You know something:

Measured data*
You know your noise distribution (poisson*)
You know that your peaks are
in a well tuned instrument
In shape close to Gaussian*
Can have multiple peaks*
Usually of similar width in tof*
Usually concentrated close to integer mass*
Usually have a couple of peaks 0-4.
Usually you can make a reasonable guess
from the observed data

Given the data you find out optimal parameters for your models
Note you DO NOT KNOW what is the ”true” data.

Traditional analysis is that your observations are the ”true data”,
find the optimum model.

Bayesian method tries to find the model that best explains what you
have observed, and realized that your data is just a reflection of the reality.

For our data:

Baseline + several line shapes(height, width, center) and
modelled noise which depends (or not) on the
points above by some functional form
calculate the probability of an observed data point.

Sum all observed probabilitties to get p_i
If larger than p_{i-1} from previous round accept this round.
If p_i worse than pervious p_{i-1} then calculate
q, a random point from U[0,1).
If $q > p_i/p_{i-1}$, then accept as a new guess.
Change your parameters a little bit and return to start of the loop

Continue until convergence has been achieved for given models.

Apply Occam's razor to estimate the validity of different models.

What you don’t get or do with Bayesian approach

-Have negative nonsense values

-Calculate your limits with
Gaussian distribution

-Assume or claim symmetric errors
or confidence limits

-Claim that this as a
”whatever sigma detection”, but

rather you use confidence limits calculated
in a correct way.

Get a model

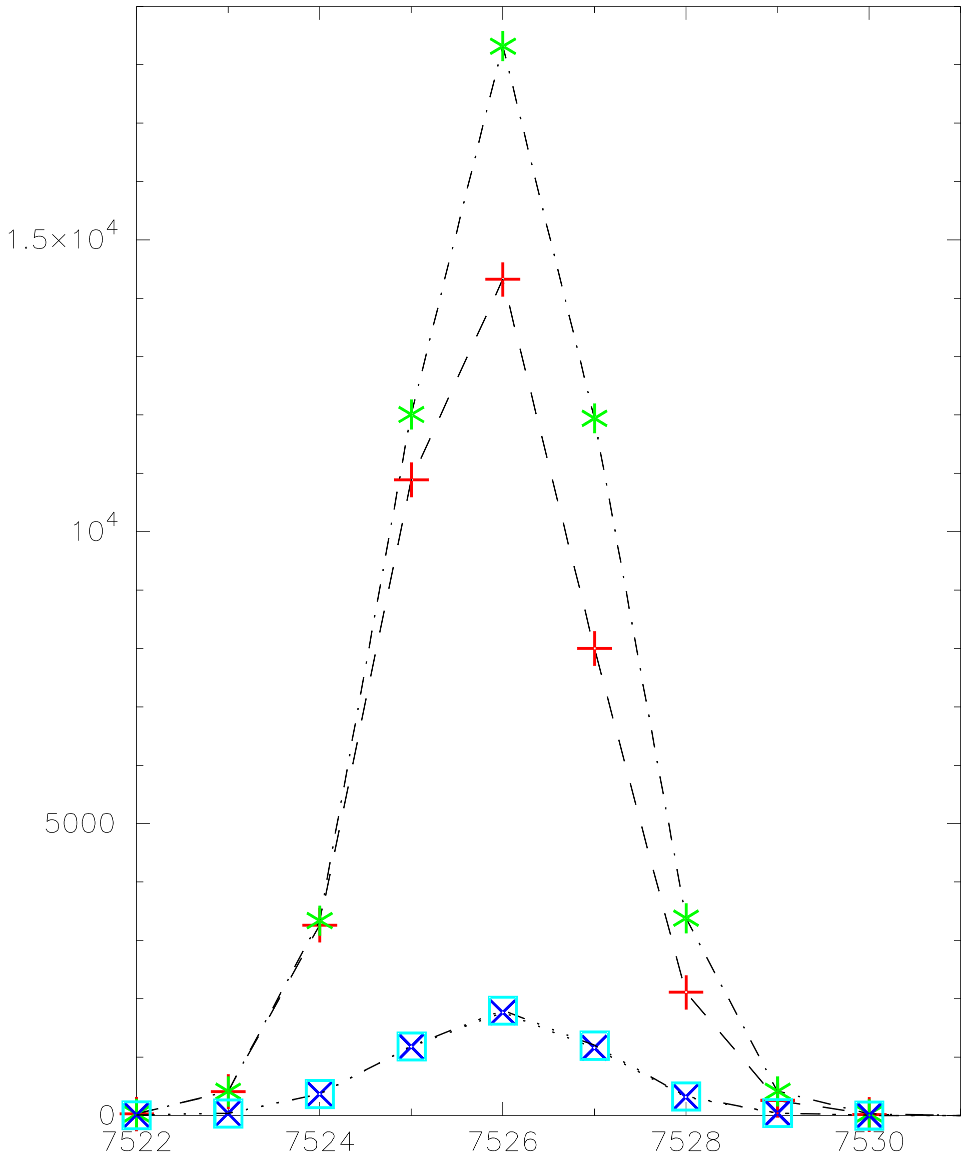
Include deadtime effects and other distortions

Do not add noise to the data points

Calculate with what likelihood your model explains the observed data with the observed confidence limits

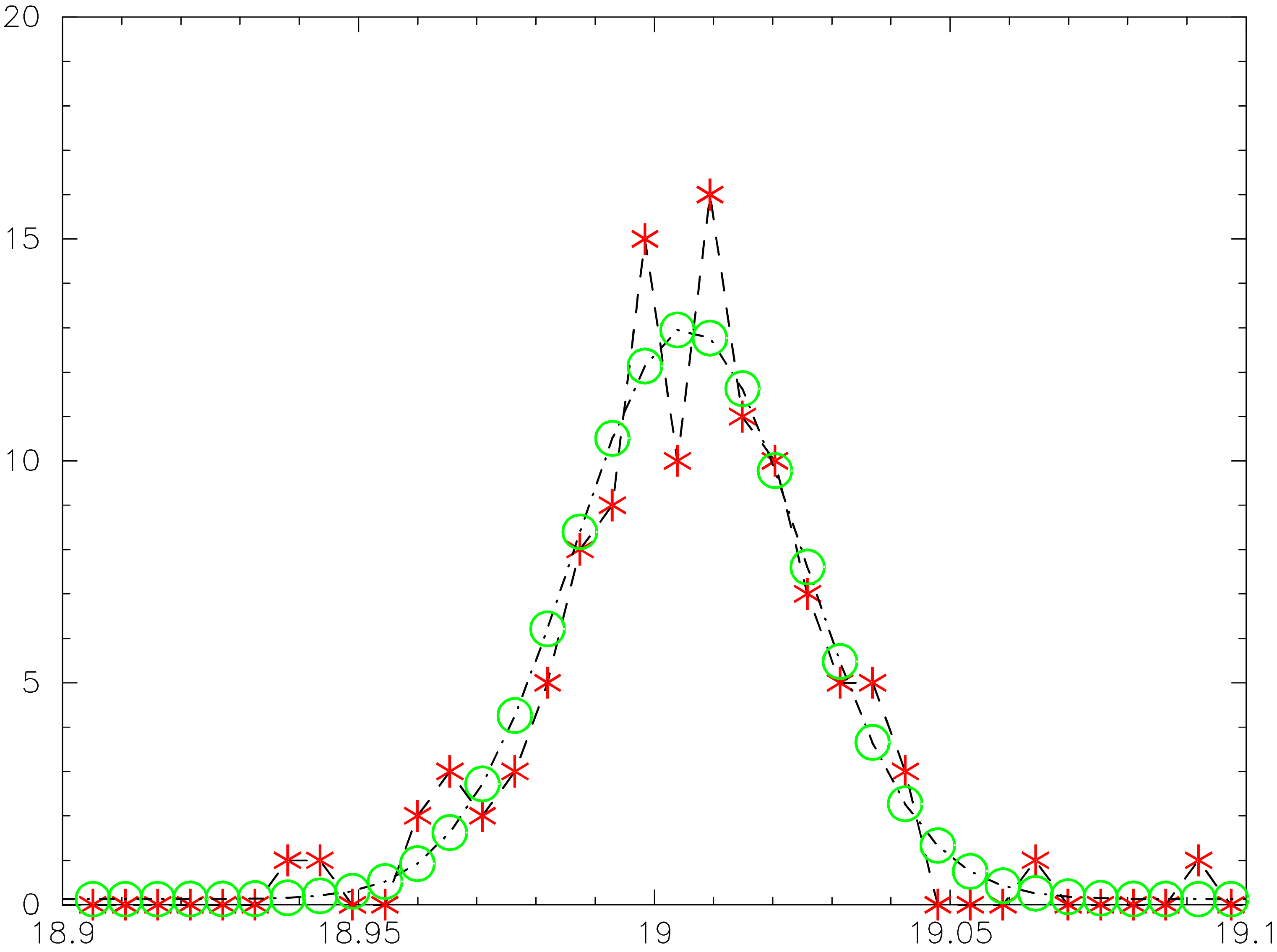
Using the correct measuring errors and binning.

Next a couple of examples.



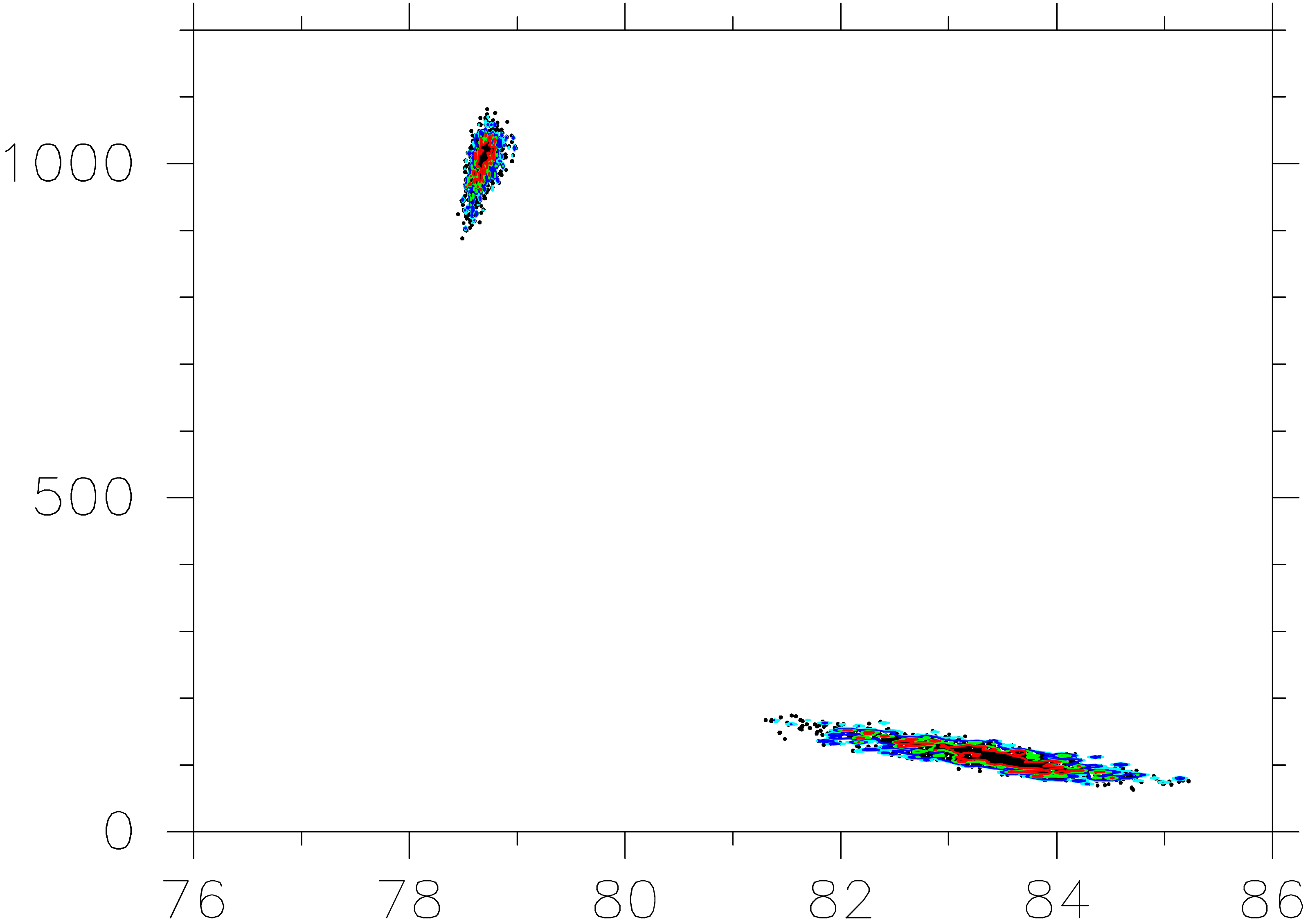
DEAD TIME EFFECTS

Number of count/shots
0.5 stronger line
0.05 fainter line



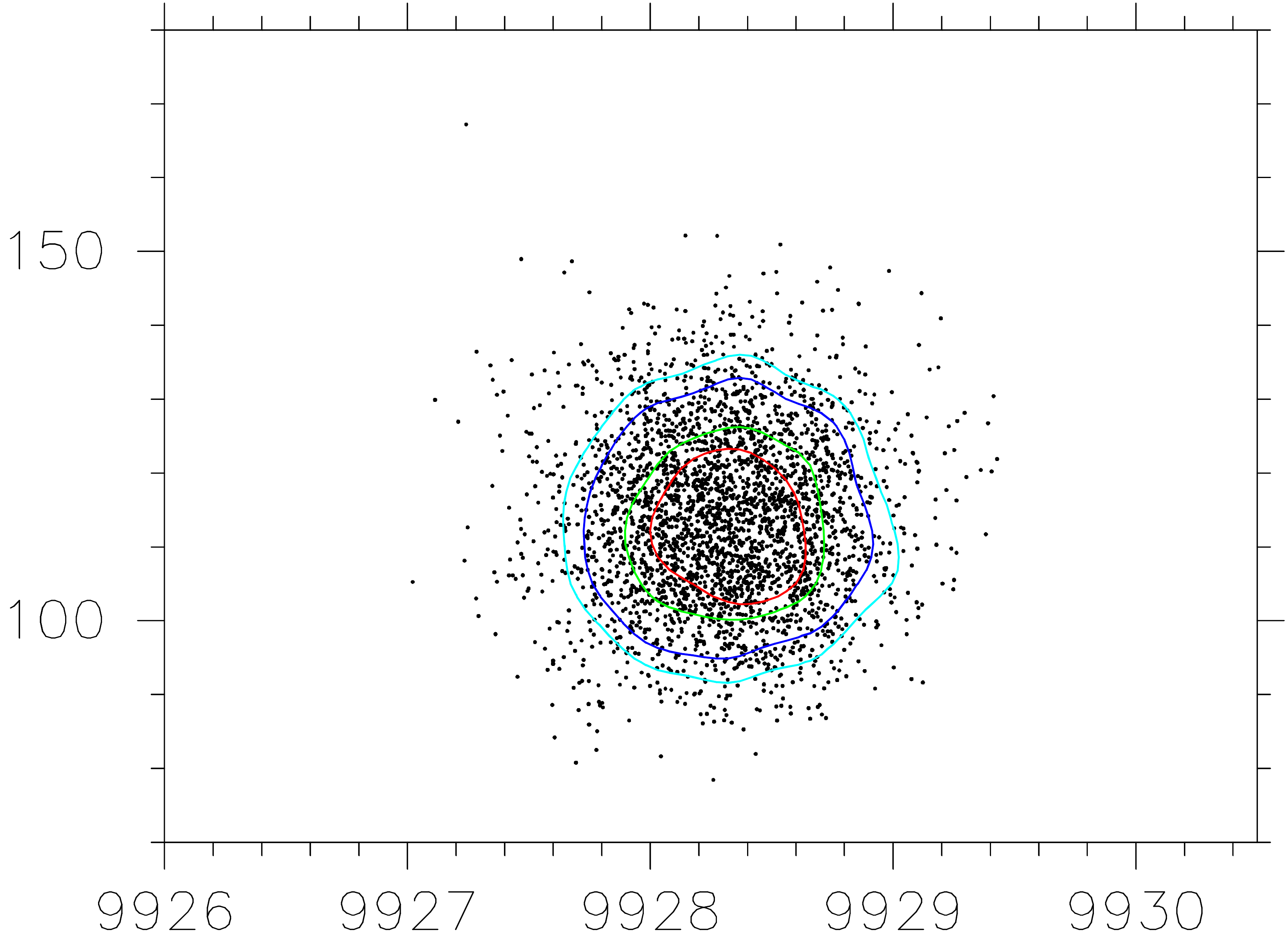
CS 2D8 20100509T194035 SP P.TAB

19.0130 ?: F+ 18.9984 Da,
HDO+19.052 Da, ?

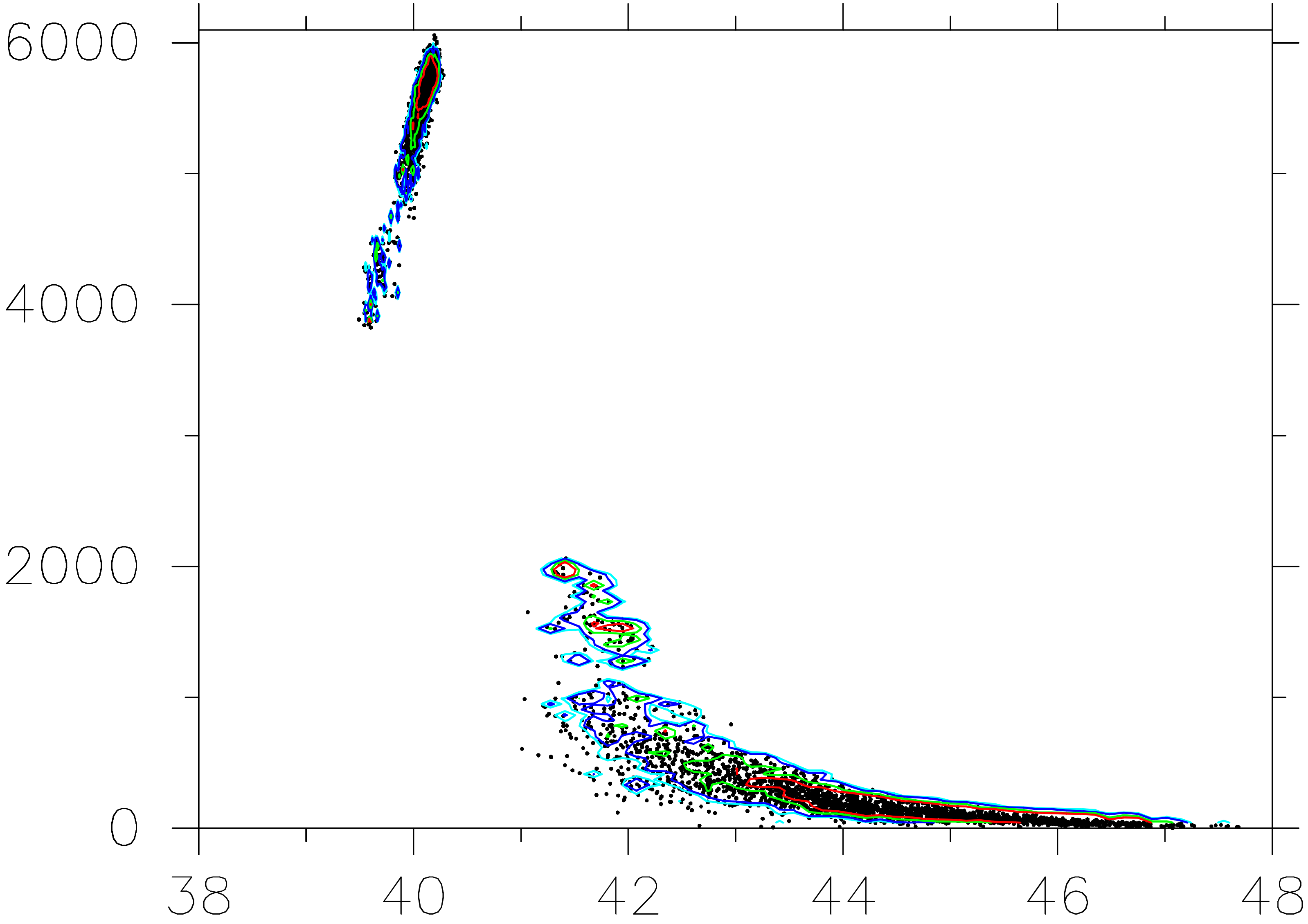


^{26}Mg,$^{12}C_2H_2$:   1000 counts vs 100 counts

25.982593,26.015650



CS 2D8 20100509T194035 SP P.TAB.



Mass 100.00, 100.035, separation 2.7, 5200, 520