Localization and estimation of jumps, hidden frequencies and other perturbations in the Earth tide residues

Angel Venedikov[1†)], Bernard Ducarme[2)]
1) Geophysical Institute & Central Laboratory of Geodesy, Sofia.
2) Research Associate NFSR, Royal Observatory of Belgium. ducarme@oma.be

† Prof. Angel Petkov Venedikov unexpectedly passed away on December 1st 2007

1. Introduction.

The paper deals with the non-tidal component $Y(t)$ of some SG data. $Y(t)$ (t is time) is determined as drift of the data through the tidal analysis by the VAV program (Venedikov et al., 2003, 2005). The analysis takes into account all tidal constituents, including the LP tides. Thus $Y(t)$ is expected to be free of all tidal signals.

The aim of the investigation is finding non-tidal signals. They can be useful, e.g. as potential or eventual earthquake and volcano precursors, as well as non-useful or parasite components of the data, which should be removed from the data.

The investigation discussed here is an application of our new program for regression analysis called M-LEVEL. It can deal with the following signals:

(i) Sudden jumps or very fast displacement of the curve $Y(t)$ and sudden or very fast changes in the slope of the curve $Y(t)$, i.e. in the derivatives of $Y(t)$.

(ii) Intervals of $Y(t)$ with anomalies or perturbations and

(iii) Non-tidal waves with known and unknown frequencies.

These items have been successfully applied in (Ducarme et al., 2006a, b). Here a series of superconducting gravity (SG) data is studied in a somewhat more advanced way.

2. General model of the non-tidal data $Y(t)$.

The VAV analysis is made by using filters of 24^h, moved also by 24^h. Each application of the filters provides one value of $Y(t)$. Due to this the unit of the time t used is 1 day. Thus the set of data generally looks like

$$Y(t), t = t_1, t_2, \ldots t_N, \text{ where } t_{i+1} - t_i \begin{cases} = 1 \text{ day when there is not a gap} \\ > 1 \text{ day when there is a gap} \end{cases} \quad (1)$$

In principles, the gaps are not an obstacle, because the processing is made by the Least Square Method (LSM) which allows avoiding artificial creation of data through interpolation. We shall recall the simple principle of dealing with gaps by LSM: LSM deals with the really existing data and it does not deal with data, which do not exist or artificially created.

As shown by Figure 1, the tidal component dominates considerably the non-tidal $Y(t)$, i.e. it should be very carefully eliminated. We hope that this is well done by the tidal analysis program VAV

$Y(t)$ can be conceived as a kind of a mean level of the data, in a way like the mean sea level. Here it looks as a a smooth curve, nearly a constant.

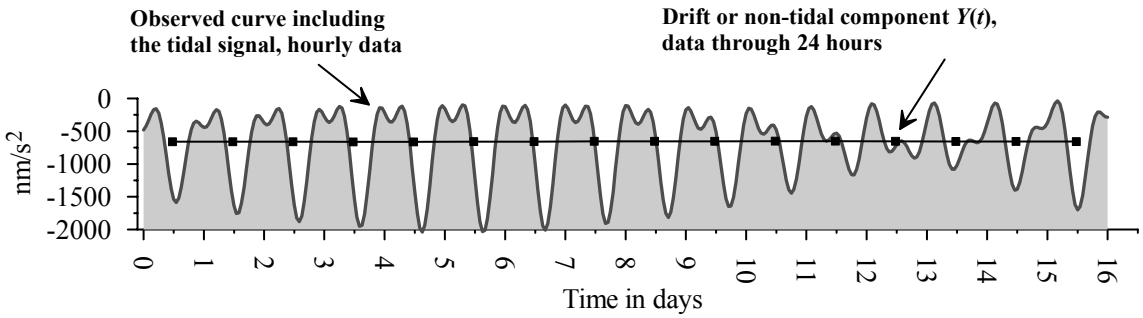

Figure 1. Sample of observed super conducting gravity (SG) tidal data with the non-tidal component $Y(t)$ (black points through 24 hours), station Moxa.

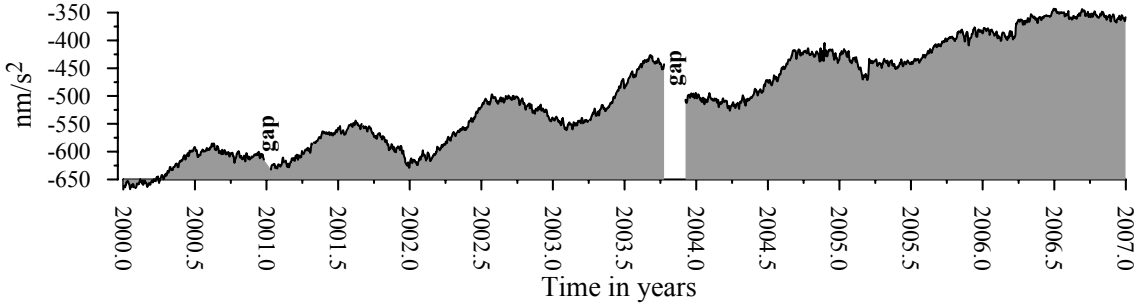

Figure 2. Sample of the non-tidal component $Y(t)$ in another scale, SG station Moxa.

Figure 2 shows, unlike the Figure 1, that actually $Y(t)$ represents a strongly variable phenomenon. Visually, it can be revealed an annual wave plus a proper drift, raising steadily the curve, as well as high frequency oscillations which can be considered as a noise.

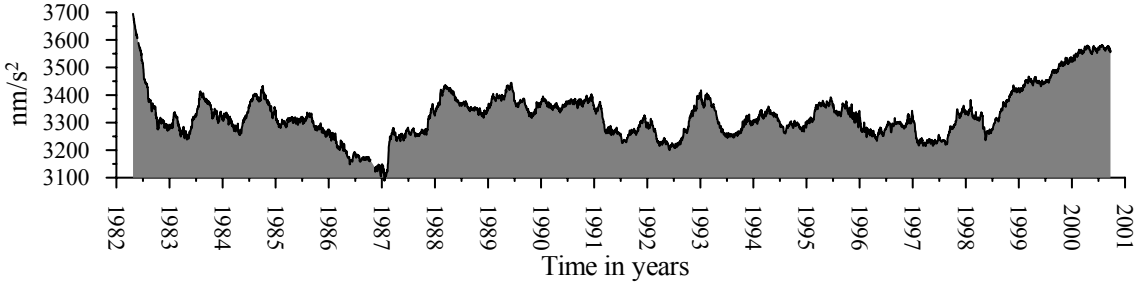

Figure 3. Complete series of the largest and oldest series of the non-tidal $Y(t)$ data from the SG station Brussels.

Figure 3 shows a more sophisticated and intricate series of $Y(t)$ data. It is possible visually to distinguish an annual component, but not such a clear drift with nearly constant slope as in Figure 2.

The task is to find a regression model, well approximating $Y(t)$, i.e. the mean level which includes various non-tidal components or signals

The initial model used here is

$$Y(t) = \sum_{j=1}^{\mu} h_j \cos\left(\varphi_j + \omega_j t\right) + \sum_{k=0}^{K} a_k (t - t_C)^k + \varepsilon_t \qquad (2)$$

where t_C is a central point of the data.

The following parts 2.1 & 2.2 discuss the first two terms of this model. Part 2.3 will discuss some statistical criteria, which are used to estimate some of We shall return to the last term, the noise ε_t, in part XXX.

2.1. Periodic components.

The first term in (1) represents the periodic signals or, may be more precisely, it approximates quasi-periodic signals. The amplitudes h_j and the phases φ_j are unknowns, estimated by the program M-LEVEL, with an application of LSM. Generally, the frequencies ω_j may be also unknowns. Nevertheless, in some of the examples M-LEVEL uses a priori defined values of ω_j, in order to simplify the presentation. Namely the frequencies ω_j expressed in cpy (cycles/year) and the periods in days applied here are given in Table 1.

Table 1. A priory fixed periodic components of the data, used in some of the following examples

ω_1 = 1 cpy, period = 1 year of 365.25 days: annual wave,
ω_2 = 2 cpy, period = 1/2 year: semiannual wave,
ω_3 = 3 cpy, period = 1/3 year: third-annual wave and
ω_4 = 0.842061 cpy, period = 433.76 days: Chandler wave.

In the last parts of the paper (section 5), ω_j are considered as unknown parameters, estimated by a specially developed technology.

Remark about the unit of the frequencies ω_j: In all discussions and results about ω_j the unit used is cpy. However, formally, in expressions like 2, the unit of ω_j should be radians/unit of time. We shall use one and the same denotation ω_j, no matter what is formally the unit used.

2.2. Polynomial component.

This is represented by the second term of (2). In the case of ocean tidal data, this term may be considered as approximation of a variable mean sea level, free of periodic signals. In the case of gravity data, this term can be considered as an approximation of a drift, also free of periodic signals.

Actually, (2) includes the most simple variant of a polynomial approximation. Theoretically, such an approximation is legal when and only when the approximated function of the time is continuous with continuous derivatives.

If Figure 3 is carefully studied, it can be stated that at the start, as well as nearly year 1987 there is a rather fast displacement of the curve, which may be considered as a jump or a

discontinuity of the approximated function. At about July 1999 there is a rather fast change of the general slope of the curve, which has to be considered as a discontinuity of the derivatives. It is of course possible that there are more points of discontinuities which cannot be visually distinguished. Here and further we shall call such a point as D-point.

In the case of 3 D-points, say T_1, T_2 & T_3 they partition the data interval in 4 segments, say $S_1, S_2\ S_3$ & S_4. Then the discontinuity problem is solved by using the model (2) with different polynomials in the different S_i, as shown by the expressions (3).

$$
\begin{aligned}
Y(t) &= \sum_{j=1}^{\mu} h_j \cos(\varphi_j + \omega_j t) + \sum_{k=0}^{K} a_{k1}(t - t_C)^k + \varepsilon_t \quad \text{for } t \in S_1, \text{ i.e. before } T_1 \\
Y(t) &= \sum_{j=1}^{\mu} h_j \cos(\varphi_j + \omega_j t) + \sum_{k=0}^{K} a_{k2}(t - t_C)^k + \varepsilon_t \quad \text{for } t \in S_2, \text{ i.e. between } T_1 \text{ and } T_2 \\
Y(t) &= \sum_{j=1}^{\mu} h_j \cos(\varphi_j + \omega_j t) + \sum_{k=0}^{K} a_{k3}(t - t_C)^k + \varepsilon_t \quad \text{for } t \in S_3, \text{ i.e. between } T_2 \text{ and } T_3 \\
Y(t) &= \sum_{j=1}^{\mu} h_j \cos(\varphi_j + \omega_j t) + \sum_{k=0}^{K} a_{k4}(t - t_C)^k + \varepsilon_t \quad \text{for } t \in S_4, \text{ i.e. after } T_3
\end{aligned}
\tag{3}
$$

Here are 4 sets of different unknown polynomial coefficients a_{ki}, $(i = 1,2,3,4)$, while the unknowns in the periodic terms remain one and the same.

It is obvious, how this scheme can be used for an arbitrary number of D-points and segments.

2.3. Statistical criteria.

When the D-points and the frequencies are known, the estimation of the remaining unknowns, a_{ki}, h_j & φ_j by the LS is a trivial problem. When the D-points and the frequencies are considered as unknowns, the problem is much more sophisticated, because these unknowns take part in a non-linear way.

Our solution of the problem is based on using variations of this kind of unknowns until finding optimum value of a given statistical criterion. D-points

For finding the D-points and the frequencies the Akaike information criterion (Sakamoto et al., 1986) is used, namely

$$\mathrm{AIC} = N(\log \pi + \log \sigma^2 + 1) + 2(N_x + 1) \tag{4}$$

where N is number of the data, N_x is the number of all unknowns and σ^2 is the estimated variance of the data. The optimum value of AIC is its minimum for various models of the data.

In the case, not yet consider, of finding anomalies or perturbations of the data, AIC is not suitable. In this case we use as a criterion the MS error of one of the unknowns, mainly the MS error of the amplitude of the annual component.

3. Finding D-points.

The subscripts of the D-points $T_1,...T_4$ are arranged depending on the depth of the minimum, T_1 being the deepest one.

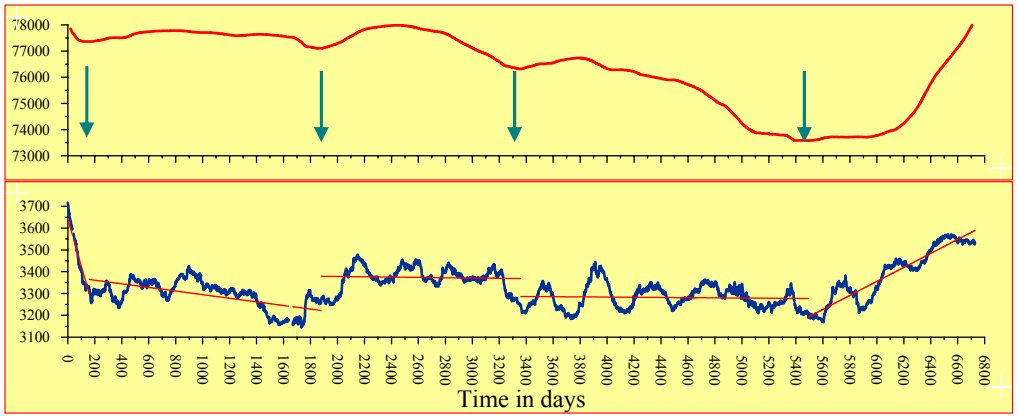

Figure 4. upper: Detection of the minima of the AIC value (arrows) for the data set Brussels, 21.04.1982 - 21.09.20000,

lower: Polynomials $P_1(T)$ (red lines) with 4 D-points AIC=67,574; σ_h = 0.660 nm/s^2

The D-points T_i determined in this way, after an initial creation of pictures like Figure 4, are only initial values. If we need a higher precision, the D-points have to be defined more accurately in an iteration procedure. For the case of 3 D-points this can be done following the following simple algorithm.

1. Fixed T_1 & T_2 $\rightarrow$ let vary T_3 till finding new minimum of AIC.
2. Fixed T_1 & the new T_3 $\rightarrow$ let vary T_2 till finding new minimum of AIC.
3. Fixed the new T_2 & T_3 $\rightarrow$ let vary T_1 till finding new minimum of AIC.
4. Looking for new minima of AIC in the segments defined by the final T_1, T_2 & T_3.
5. If necessary, the procedure can be re-started from item 1.

We find four discontinuities and the adjustment in 5 blocks with polynomials of degree one provides an important improvement with a decrease of both statistical criteria: AIC=67,574; σ_h = 0.660 nm/s^2.

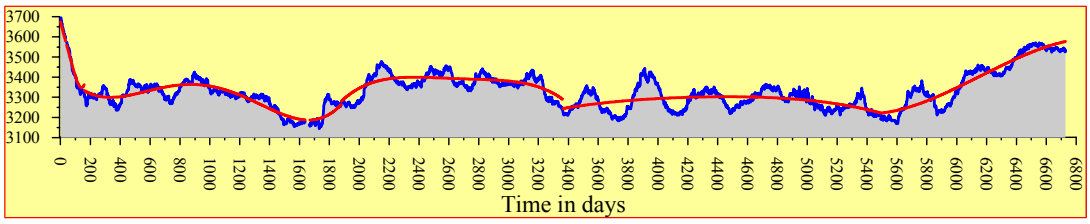

Figure 5. Drift adjustment using a fourth degree polynomial on each block.
Power K = 4: AIC = 62,620, σ_h = 0.471 nm/s^2

Is it possible to get a further improvement by adjusting the degree K of the polynomials? An optimal solution is found with the degree 4 (AIC = 62,620, σ_h = 0.471 nm/s^2, Figure 5).

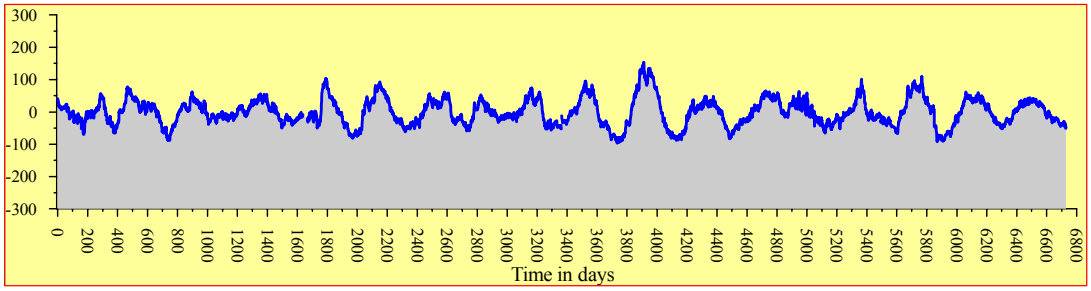

Figure 6. Final representation of the harmonic part of the signal after subtraction of the drift representation in Figure 3.

The harmonic part of the signal is obtained by subtracting the drift representation from the Y(T) series (Figure 6). It is dominated by the annual component and the pole tide signal (section 5 and Ducarme et al., 2005)

4. Detection of anomalous data

Detecting the anomalies in the tidal data is interesting for two reasons:

(i) eliminate them in order to improve the precision and

(ii) using them as particular non-tidal signals, eventually as precursors.

The AIC criterion is only applicable if the data set remains unchanged. For the anomalous data detection and rejection we can only rely on the MSD reduction. In the examples hereafter the statistical criterion used is: σ_h (MSD of the amplitude of the annual wave with frequency 1 cpy) minimum.

Let V1 & V2 be two variants of the processing or/and of the data used.

If $\sigma_h(V2) < \sigma_h(V1)$, the variant V2 should be preferred by the simple reason that V2 provides a higher precision.

In this sense σ_h is used in the same way as the criterion AIC of Akaike above. Actually, there are seldom seen contradictions between these two criteria. Nevertheless, in the case of different eliminations of data, AIC cannot be used, because it should be applied on one and the same sets of data.

Let:

$\sigma_h(0)$ be the value of σ_h when all data in a series are processed,

(t_a, t_b) be a time interval of the data with a central point $t_{ab} = (t_a + t_b)/2$ and

$\sigma_h(t_{ab})$ be the value of σ_h when the processing eliminates the interval (t_a,t_b).

In this sense (t_a, t_b) will be called E-interval (elimination interval).

It should be expected $\sigma_h(0) < \sigma_h(t_{ab})$ because the elimination of the E-interval decreases the quantity of the data and thus decreases the analysis precision.

Hence, if on the contrary $\sigma_h(t_{ab}) < \sigma_h(0)$, this is a serious indication that the E-interval (t_a, t_b) contains some anomalous data or it is a part of an anomaly.

The idea used is the following. We deal with an E-interval (t_a, t_b) with variable limits t_a and t_b. Initially, the E-interval covers m = (b – a + 1) = 11 days. We let run the limits t_a and t_b by increasing both t_a and t_b by steps of 1 day. This is made so that the moving (t_a, t_b) runs along the whole set of the data.

For every position of (t_a,t_b) the program applies the analysis, the data in the interval (t_a, t_b) being excluded (eliminated). In such a way we get our criterion

$\sigma_h = \sigma_h(t_{ab})$ as function of the central point $t_{ab} = (t_a + t_b)/2$

If at t_{ab} in which $\sigma_h(t_{ab})$ has relative minimums, the corresponding interval is almost certainly anomalous and it should be eliminated.

Let us consider the Membach data set 04.08.1995 – 03.12.2006 (Figure 7). The curve of $\sigma_h(t_{ab})$ as function of $t_{ab} = (t_a + t_b)/2$ (central point of the E-interval), displayed in Figure 8, shows 5 relative minimums A1, … A5, clearly indicating anomalous data, e.g. the strongest minimum A1 shows a strong anomaly at the beginning, which is usual.

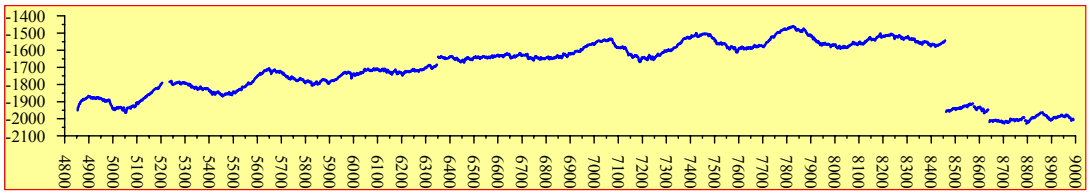

Figure 7. Membach data set 04.08.1995 – 03.12.2006

If we give a closer look to the E-interval A4, we see a rather complex minimum close to the E-interval ($t_a = 7185^d$, $t_b = 7195^d$) in Figure 9.

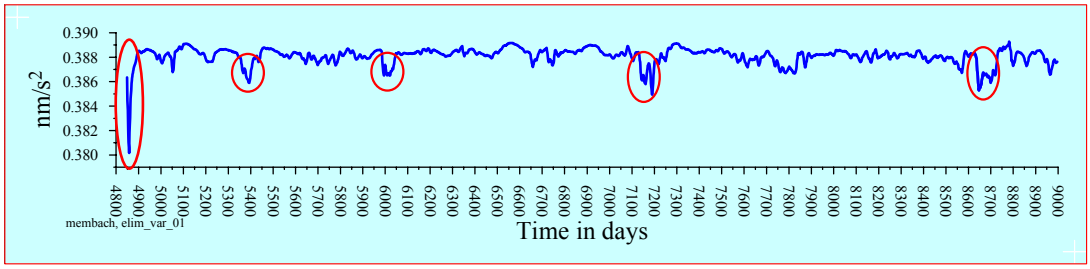

Figure 8. Curve of $\sigma_h(t_{ab})$ as function of $t_{ab} = (t_a + t_b)/2$ (central point of the E-interval) with 5 relative minimums A1, … A5, clearly indicating anomalous data.

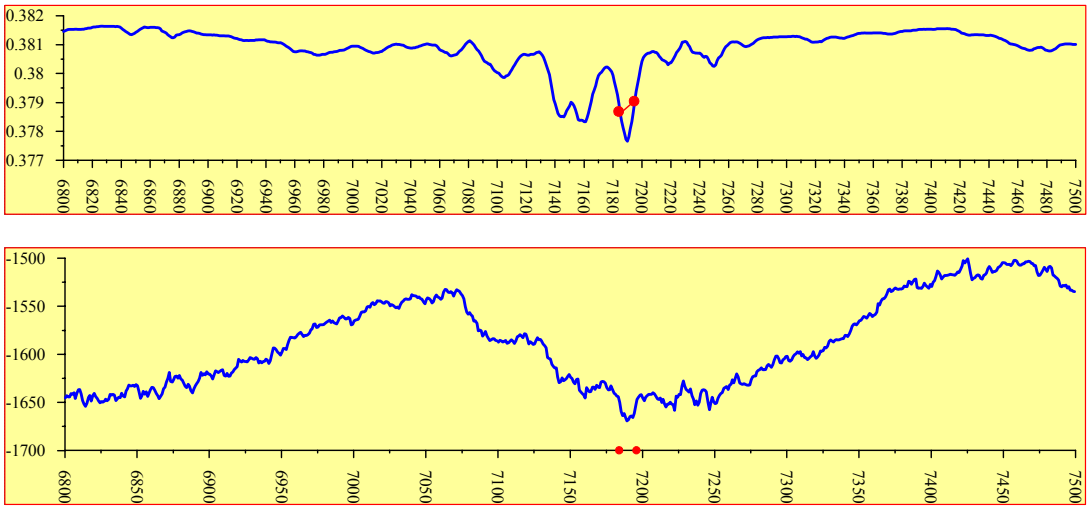

Figure 9. Details around the A4 minimum.
Upper curve: variations of $\sigma_h(t_{ab})$ as function of $t_{ab} = (t_a + t_b)/2$
Lower curve: Y(T) series
Red dots indicate the position of the 10 days interval corresponding to the lowest minimum.

After an initial E-interval is determined, as shown here, JUMP_07 allows variations of the limits t_a & t_b, in opposite directions, in order to find eventually further reduction of the criterion σ_h and thus a modulation of the E-interval for a better precision.

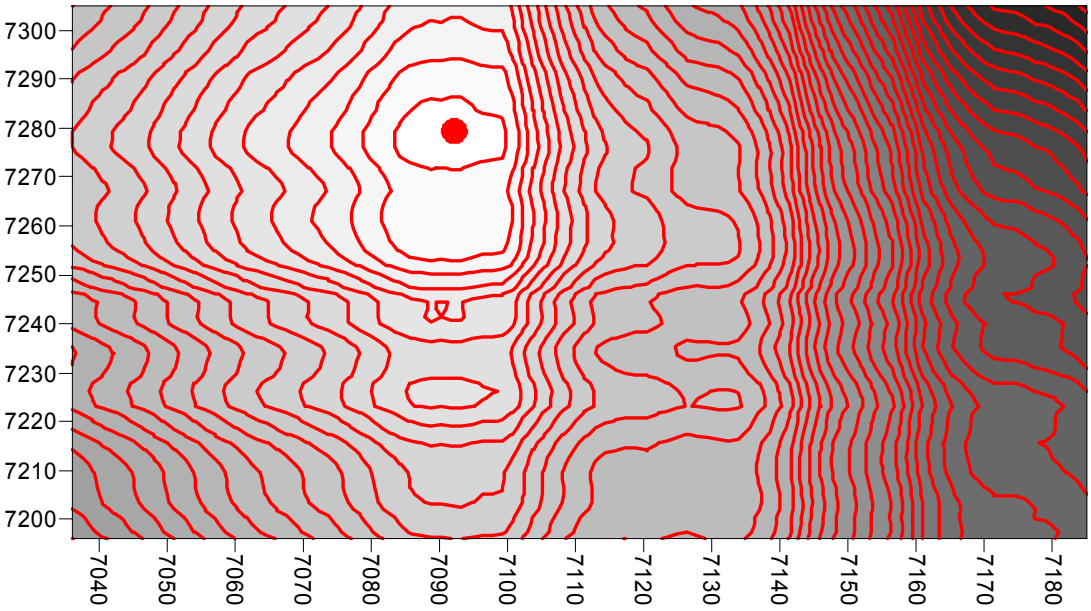

Figure 10. Minimum of $\sigma_h(t_a, t_b)$ at $t_a = 7092^d$ & $t_b = 7276^d$; this means that the E-interval (7092^d, 7276^d) may be eliminated in order to improve the precision.

The variations can be imposed independently on the limits t_a & t_b or symmetrically on both limits. Through the mutual variation we get σ_h in a 3-D space, i.e. $\sigma_h = \sigma_h(t_a, t_b)$ as a function of the two variables t_a and t_b (Figure 10).
A clear minimum is obtained for t_a equal day 7092and t_b equal day 7276. The E-interval (7092^d, 7276^d) may be eliminated in order to improve the precision, although this is a rather large (185 days) interval. Finally the 5 intervals shown in red in Figure 11 were eliminated.
The reduction of the MSD is important, from $\sigma_h = 0.380$ nm/s^2 (all data, without eliminations) to $\sigma_h = 0.317$ nm/s^2. Such a raise of the precision could be obtained only if the length of the data is increased in the ratio $(0.380/0.317)^2$, i.e. 44%.

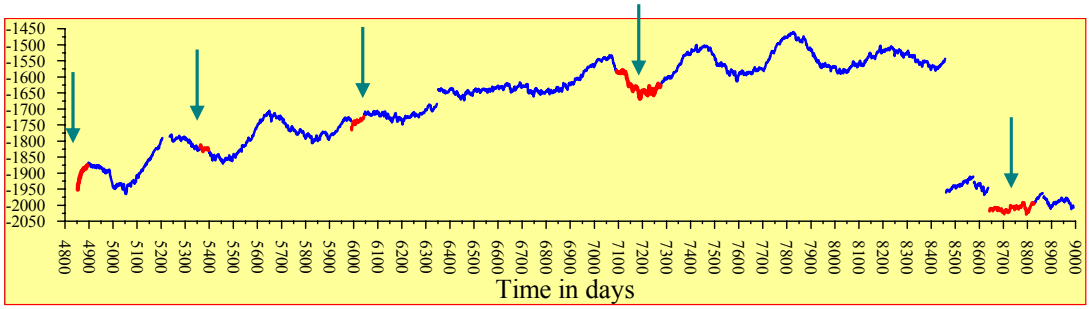

Figure 11. The five eliminated intervals are shown in red

5. Search for non-tidal waves with unknown frequencies

Here also we apply an automatic search algorithm constrained by the AIC criterion.
Instead of looking for the maximum of the amplitude $h = h(\omega)$ as a function of ω, as it is done in the spectral analysis, we shall consider the criterion AIC = AIC(ω) as a function of ω, looking for its minimum. So to say, instead of the spectrum h(ω) we shall deal with the AIC (ω) spectrum.
Let us suppose that we know *m* frequencies ω, i.e. we can solve by LSM the system

$$Y(T) = \sum a_k T^k + \sum (\alpha_j \cos \omega_j T + \beta_j \sin \omega_j T), T = T_1, \ldots T_n \qquad (0.1)$$

Let us call the corresponding value of the Akaike Information Criterion AIC_m.
In a next step we add a new wave with variable ω

$$Y(T) = \sum a_k T^k + \sum (\alpha_j \cos \omega_j T + \beta_j \sin \omega_j T) + \alpha \cos \omega T + \beta \sin \omega T \qquad (0.2)$$

Let us vary the value of ω by steps Δω and for every value of ω we get $AIC_{m+1}(\omega)$.
When we get for a given value $\omega=\omega_{Min}$:

$$AIC_{m+1}(\omega_{Min}) = \text{Minimum}$$

and

$$AIC_{m+1}(\omega_{Min}) < AIC_m,$$

we accept ω_{Min} as a new frequency ω_{m+1} and the procedure can be repeated iteratively

.

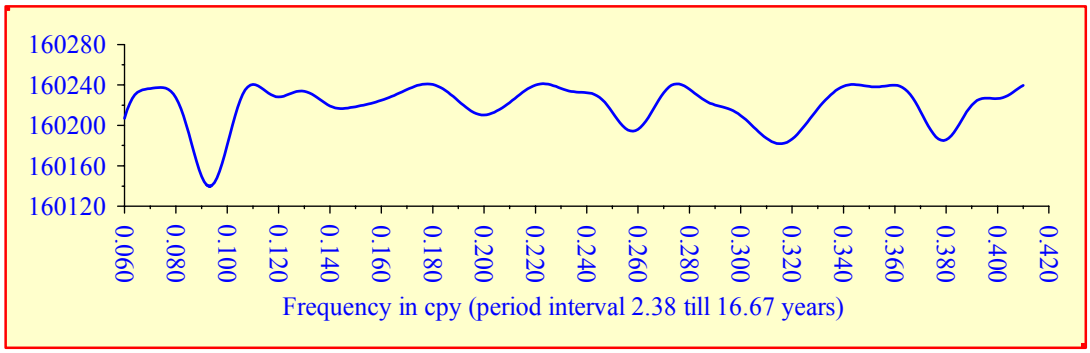

Figure 12. Station Cananéia (Brazil): variation of AIC as a function of a moving frequency showing a minimum at 0.093124 cpy, period = 10.74 years

As an example let us consider first the research of a very low frequency non-tidal component in a 50 year tide gauge record at Cananéia (Brazil) (Ducarme et al., 2006b). An additional harmonic constituent was search between 0.06cpy (16.67y) and 0.42cpy (2.38y). Several relative minima were found, with an absolute one at 0.093124 cpy (period = 10.74 years), corresponding to the solar cycle (Figure 12).
When the existence of two close frequencies is suspected, a simultaneous search is possible that can be represented in a three dimensional space. The example given below illustrates the simultaneous fit of the annual period and the Chandler one on the Brussels data of Figure 6. An absolute AIC minimum is found in Figure 13 at a point corresponding to periods (365.12, 433.76 days) or frequencies (1.00037, 0.842061 cpy).

6. Conclusions

A program M-LEVEL has been developed to exploit fully the possibilities of the VAV tidal analysis program. Starting from the daily values Y(T) of the residues obtained automatically from VAV by subtracting the full tidal signal, it allows to detect anomalies and jumps in the data as well as unknown non-tidal frequencies. The procedure is based on the use of statistical criteria to find the optimal regression model or the best part of a data set. The Akaike Information Criterion (AIC) is used in parallel with the Mean square Deviation (MSD) of the adjusted parameters. The best solution coincides with the lowest value of these statistical quantities. The procedure is very flexible.

Bibliography

Ducarme B., van Ruymbeke M., Venedikov A.P., Arnoso J., Vieira R., 2005.
Polar motion and non tidal signals in the superconducting gravimeter observations in Brussels.
Bull. Inf. Marées Terrestres, 140, 11153-11171.
Ducarme B., Venedikov A.P., Arnoso J., Chen X.D., Sun H.P., Vieira R., 2006a.
Global analysis of the GGP superconducting gravimeters network for the estimation of the pole tide gravimetric amplitude factor. *Proc. 15th Int. Symp. On Earth Tides, J. of Geodynamics, 41, 334-344.*
Ducarme B., Venedikov A.P., de Mesquita A.R., De Sampaio França C.A., Costa D.S., Blitzkow D., Vieira R., Freitas S.R.C., 2006b. New analysis of a 50 years tide gauge record at Cananéia (SP-Brazil) with the VAV tidal analysis program. *Dynamic Planet, Cairns, Australia, 22-26 August, 2005. Springer, IAG Symposia, 130, 453-460.*

Sakamoto Y., Ishiguro M., Kitagawa G. 1986. Akaike information criterion statistics, *D. Reidel Publishing Company, Tokyo,* 290 pp.
Venedikov, A. P., Arnoso, J., Vieira, R. (2003). VAV: a program for tidal data processing. *Comput. Geosci., 29, 487-502.*
Venedikov, A. P., Arnoso, J., Vieira, R. (2005). New version of the program VAV for tidal data processing. *Comput. Geosci., 31, 667-669*

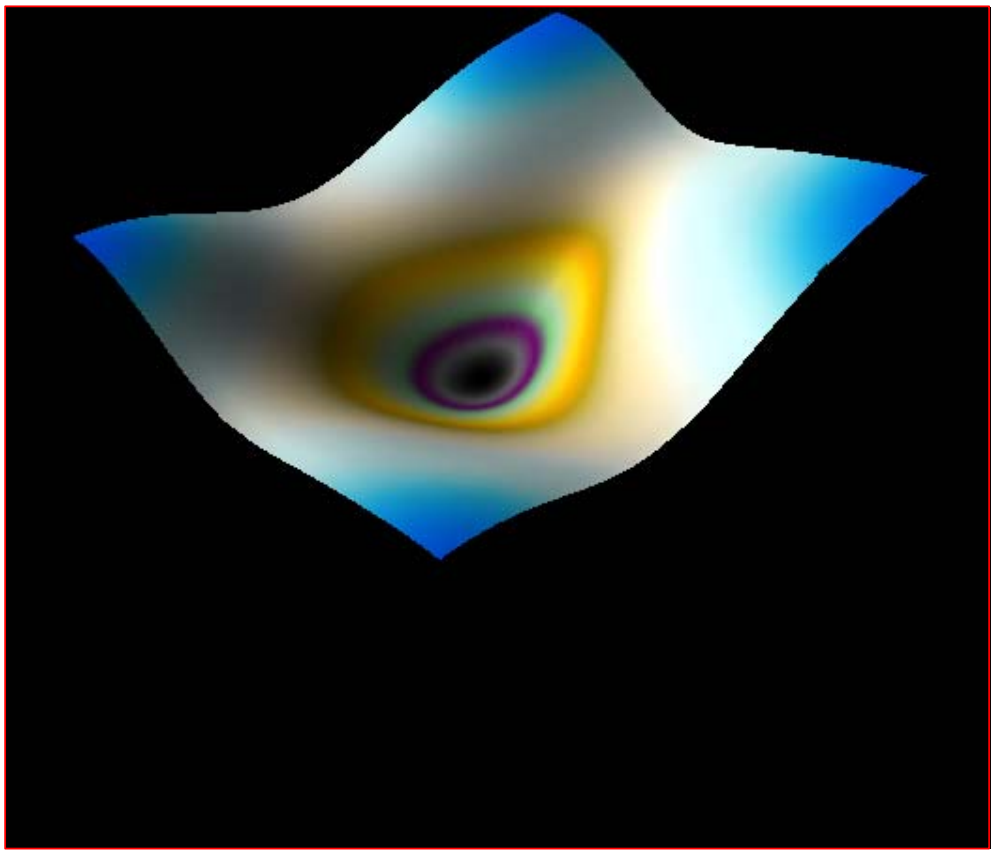

Figure 13. Three dimension graph showing the existence of two close frequencies in the Brussels data, an annual one at 1.00037cpy (365.12 days) and the Chandler one at 0.842061cpy (433.76 days)