DETECTION OF OUTLIERS IN TWSTFT DATA USED IN TAI

A. Harmegnies, G. Panfilo, and E. F. Arias
International Bureau of Weights and Measures (BIPM)
Pavillon de Breteuil F-92312 Sèvres Cedex, France
E-mail: ***aharmeg@bipm.org, gpanfilo@bipm.org, farias@bipm.org***

Abstract

This paper describes a new filtering technique used to detect and eliminate outlying data in two-way satellite time and frequency transfer (TWSTFT) time links. In the case of TWSTFT data used to calculate International Atomic Time (TAI), three main problems have to be considered:

- ***the difficulty of recognizing outliers from useful data;***
- ***the need to avoid deleting useful data;***
- ***that TWSTFT links can show an underlying slope which renders the standard treatment more difficult.***

Using phase and frequency filtering techniques, a new way of detecting outliers, while avoiding detection of useful data, has been developed and implemented at the BIPM to clean TWSTFT data.

INTRODUCTION

Each month, the BIPM Time, Frequency, and Gravimetry Section produces International Atomic Time (TAI) and Coordinated Universal Time (UTC) from data contributed by almost 70 laboratories. This involves about 370 atomic clocks linked by various techniques. Most time links (85%) are computed from multi-channel GPS receivers, either single- or dual-frequency; 14% of the links are from TWSTFT **[1,2]** observations in Europe, North America, and the Asia-Pacific region. In September 2009, a new technique based on the carrier phase combined with the code of the GPS signal (Precise Point Positioning, GPS PPP **[3]**) was introduced into *BIPM Circular T* **[4]**. The BIPM treats, in total, TWSTFT observations from 20 laboratories, half of which currently are under study prior to their inclusion in the routine TAI calculation. Several TWSTFT links are affected by outlying data; to ensure safe handling of the data, a new cleaning technique has been developed at the BIPM. Although the data-cleaning technique has been developed for application to TWSTFT links, it could equally be adapted to other kinds of time links. Removing the outliers on TWSTFT links is a challenge for a number of reasons: it is difficult to recognize outliers from useful data; the TWSTFT links may show an underlying slope which complicates the standard treatment **[5-10]**; and finally, the number of TWSTFT data points is rather low. TWSTFT time links routinely provide 12 data points per day (one measurement every 2 hours in most cases, and every hour in the case of Asia-Europe time links); this number is rather low compared to the number of measurements for GPS time links (about 100 measurements per day).

The first part of this paper illustrates the various approaches tested using phase and frequency data. The second part shows the results obtained using the new filtering technique.

THE OUTLIER DETECTION TECHNIQUE

The techniques most commonly used to detect outliers in data used in the calculation of TAI **[5]** are based on different statistical estimators applied to the phase data. However, TWSTFT links may show an underlying slope that makes the standard treatment more difficult and increases the risk of removing useful data. In this case, a better approach is to consider both phase and frequency values.

After testing various approaches, we concluded that mixing different methods increases the reliability of the filter. A first step called "Rough cleaning" consists of detecting very large outliers by using a moving average on the phase data and observing the residuals between the real and the filtered values. The second step consists of a more refined detection, making use of two different mathematical methods:

1. the moving average applied to phase data;
2. the Median Absolute Deviation (MAD) estimator applied to the frequency data.

A data point is removed only when both techniques identify it as an outlier.

This new outlier detection process gives satisfying results when applied to TWSTFT links without removing too many data and complements the existing cleaning tools developed for time link data **[5]**. In the next sections, the mathematical techniques used in the algorithm for outlier detection are presented.

FREQUENCY FILTERING

The mean and standard deviation, standard estimators used to characterize the properties of data set, can be really affected by outliers **[7]**. An estimator more resistant to outliers is the median. In this study, we consider a robust estimator, based on the median, called Median Absolute Deviation (MAD) **[6-8,10]**, which is frequently used in data sets affected by outliers. The MAD is defined as the median of the absolute deviations from the data median:

$$MAD = median_i \left(\left| X_i - median_j (X_j) \right| \right) \tag{1}$$

The classical "standard deviation" can be estimated using the MAD [6]:

$$\hat{\sigma}_{MAD} = K \cdot MAD \tag{2}$$

where K is a constant scale factor which depends on the type of the distribution: for normal distribution of data, $K \approx 1.4826$ **[6,7]**. Consequently, by choosing K = 1.4826 the expected value of $\hat{\sigma}_{MAD}$ is equal to the standard deviation σ for normally distributed data.

To test the robustness of the MAD with respect to the standard deviation in the case of TAI calculation, a test was performed using the TWSTFT data reported to the BIPM in February 2009 (hereafter, 0902). We calculate the classical standard deviation and the estimated standard deviation using the MAD of the frequency data for all TWSTFT links. The MAD was found to be more robust than the classical standard deviation. Figure 1 shows the histogram of the standard deviation of the links (in blue) and the standard deviation calculated by using the MAD (in red).

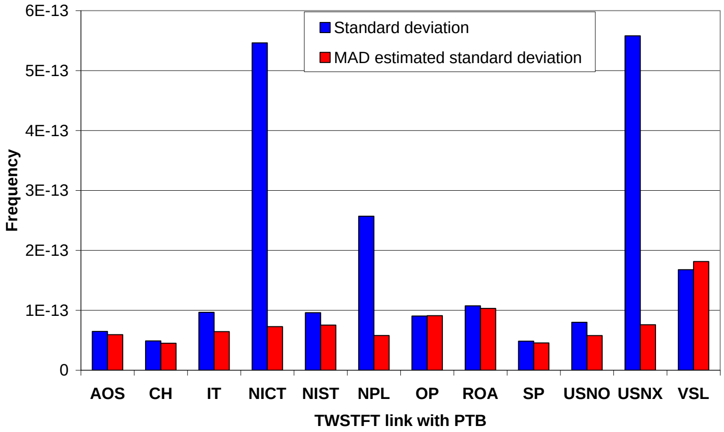

Figure 1. Comparison between the standard deviation and the standard deviation estimated using the MAD for TWSTFT data of February 2009.

In month 0902, outliers are mainly present in the USNO-PTB X-band link (USNX), in the NICT-PTB link (NICT), and in the NPL-PTB link (NPL). The behavior of the statistical tools is different in each case, but in each case the MAD was found to be more resistant to outliers.

According to Sesia and Tavella **[8]**, the MAD is the most widely used filter to detect and remove outliers. To create the filter, a threshold *t* has to be defined so that a value X_i is considered an outlier if:

$$\left| X_i - median_j(X_j) \right| > t \cdot \hat{\sigma}_{MAD} \tag{3}$$

In the case of our filter, the threshold was defined as *t* = 3, which corresponds to about 1% of outliers if the data are normally distributed **[7]**.

This filter is applied to frequency data derived from phase data using the well-known relations:

$$y(t) = \frac{d\,x(t)}{dt} \qquad\qquad y_i = \frac{x_{i+1} - x_i}{\tau_i} \tag{4}$$

where *y* (*t*) is the frequency value and *x* (*t*) is the phase value at time *t*. TWSTFT data are usually sampled every hour (or every 2 hours), so τ = 1 hour (or 2 hours) (standard time interval). A time interval τ_i bigger than τ between two successive observations indicates a hole in the data. In such cases, *y*(*t*) will be affected and the value will be wrong. For this reason, holes in phase data are detected before computing the MAD and only frequency data not corresponding to a hole are used to calculate the MAD (1). Once the MAD is determined, the complete data set is treated to define outliers with respect to the calculated MAD. The frequency values can be obtained by considering different values of τ_i.

The use of frequency data leads to several difficulties linked to the identification of the corresponding phase data that generate frequency outliers. Several examples are shown in the four plots in Figure 2.

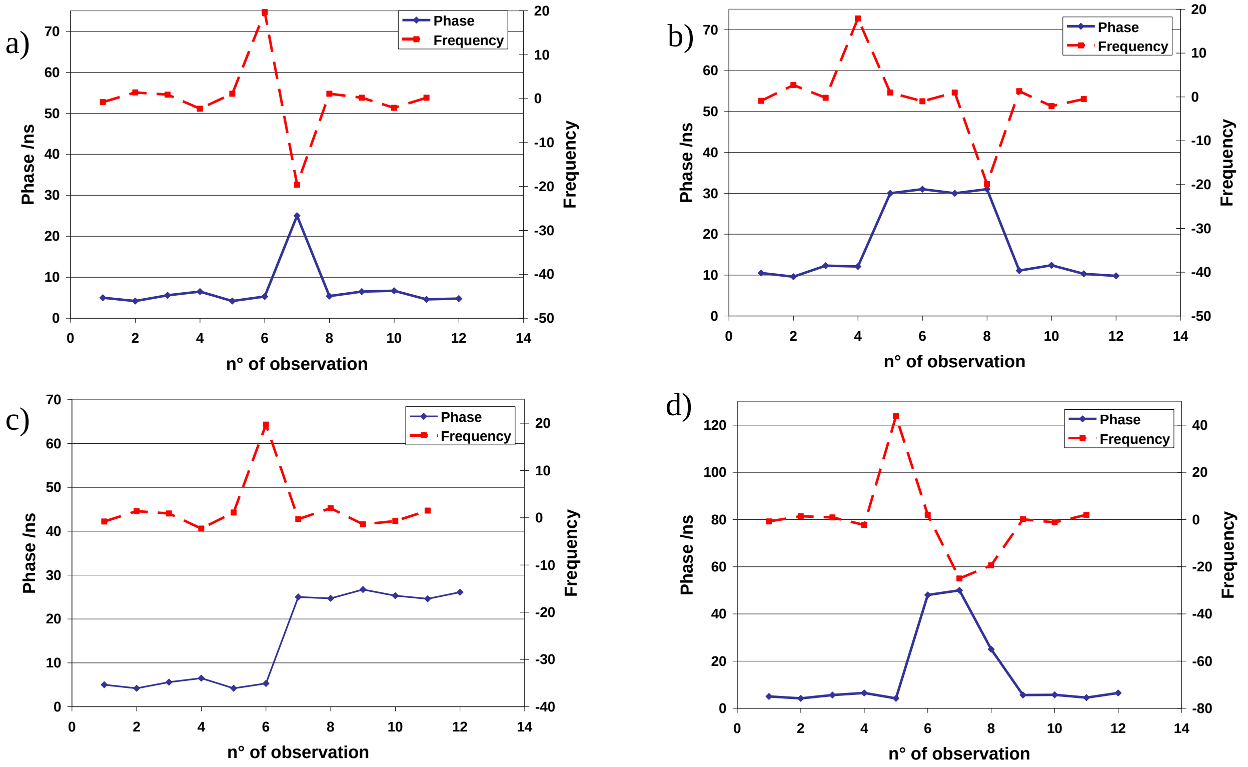

Figure 2. Representation of specific behaviors in phase and corresponding frequency data: (a) a single phase outlier; (b) successive phase outliers; (c) time step in phase data; and (d) a more complex case with successive series of outliers.

Sesia and Tavella **[8]** consider that when frequency outliers are detected, the two corresponding phase data should be eliminated. It is obvious that, in such cases, more phase data than necessary are cleaned, considering that two phase data are used to produce one frequency value. This solution was not applicable to our case because of the small quantity of data in TWSTFT time links.

The frequency data have to be carefully checked to detect the corresponding phase data outlier. We analyzed a number of data after the outlier detection and checked the sign of the frequency jump. If a time step occurs (Figure 2 (c)), the frequency outlier data corresponding to the end of the observation period does not exist and, thus, no data are modified. In the case that frequency outliers corresponding to successive phase outliers are found, all points are considered as outliers (Figure 2 (b)).

This method is very important in order to distinguish between successive phase outliers and a time step. If we remove the two phase data corresponding to the frequency outlier, we might remove useful data or miss real outliers.

Figure 3 shows as an example the case of the USNO-PTB X-band link. The filter is efficient in the presence of a single outlier, but if several close outliers appear, then the filter fails, such as for the data prior to MJD 54869.5. By checking the frequency values with the sign, the filter works very well.

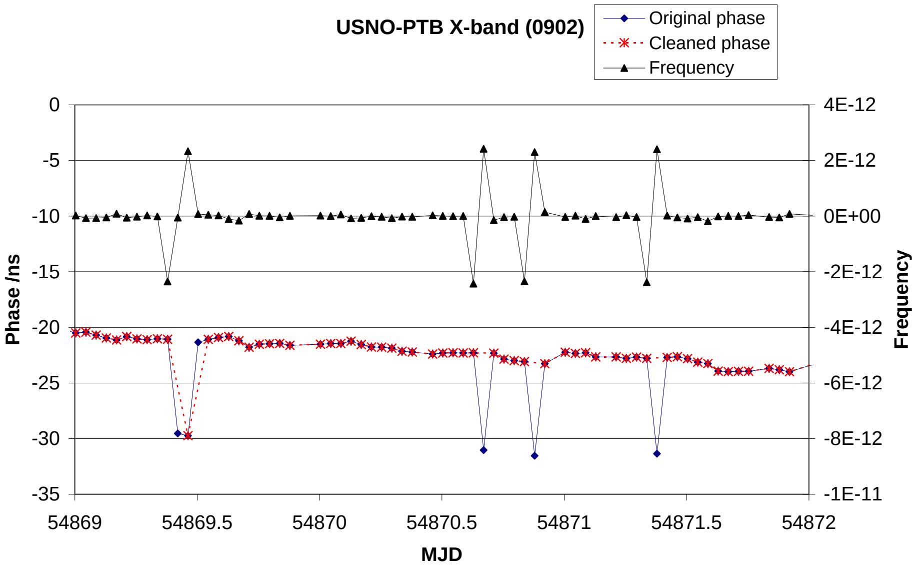

Figure 3. Original and cleaned data for the USNO-PTB X-band link. The blue line (♦) represents the real phase data, the black line links the frequency data (▲), and the red dots (×) shows the cleaned data.

Figure 2(d) shows a more complicated situation where the interpretation of the behavior of the frequency values becomes difficult. Without examining the phase values, there is no way of detecting whether the last datum of a short time jump has a useful ("normal") value. Since the goal is avoid the deletion of useful data, a second filter on the phase data is added to check the correct outlier detection. This second filter is described in the next section.

PHASE FILTERING

To ensure the removal of only real phase data outliers, a second test is performed on the phase data. We filter the data with the moving average technique, and detect the outliers by analyzing the residuals between the filtered and real data.

The width of the window used in the phase filtering can be adopted depending on the noise presents in the data set. We decided to use 12 data in the moving average, but a smaller size can be required if the drift of the phase is high; and the window can be enlarged if large data holes exist.

A maximum accepted value of the residuals between real and smoothed data must be fixed. To determine this value, the cumulated frequencies of the residuals were calculated for different time links in different periods. The results are reported in Figure 4. It is seen, for example, that the NICT-PTB time link for two different months shows very different results and it is therefore difficult to set a default value perfectly adapted to every situation. We chose to set the residual threshold Z equal to 2 ns because most residuals are located between –2 ns and +2 ns independently of the link used. However, this parameter can be adapted to special cases to increase the efficiency of the filter.

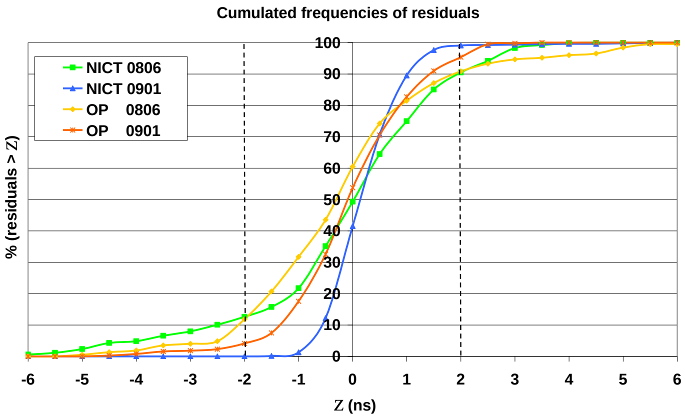

Figure 4. Cumulated frequency of residuals between the filtered and real data for different time links.

This filter can be used for outlier detection and the results in most cases will be satisfactory. However, in the calculation of time links for TAI, we are often faced with very complex situations where the detection of useful data is not straightforward. One example is reported in Figure 5, showing the results obtained by applying the phase filter to the NICT-PTB time link (data period 0806). In this situation, we should avoid using these data in the TAI calculation and consider an alternative technique. We have retained this case as a good example to check and refine the outlier detection technique.

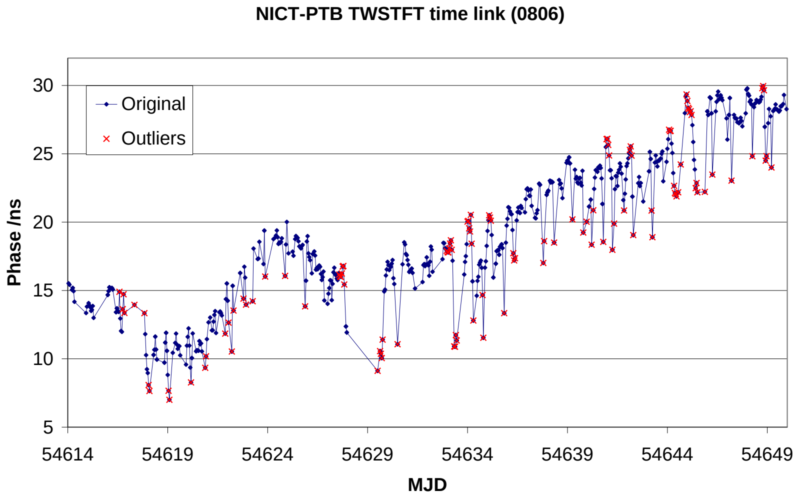

Figure 5. The blue line represents real data and the red stars (×) represent detected outliers.

In this particular case, with threshold equal to 2 ns, the phase filter detects more outliers than needed. The use of frequency and phase filtering together avoids the detection of useful data. Thus, both frequency and phase filters have been combined to establish a refined way of detecting outliers: if a datum is considered as an outlier for both filters, then it is removed from the data set.

During tests of the algorithm on different data sets, it was found that too many outliers in the same region, or the presence of very large outliers, could lead to wrong results, so that sometimes useful data were considered to be outliers and, conversely, some outliers were considered to be useful data (see Figures 6 and 7). In particular, Figure 6 shows the USNO-PTB Ku-Band time link (related to the period 0806). Figure 7 plots the results after the application of the composed filter. Zooms of the Figures 6 and 7 on a specific date show the effect of the filter with too many close outliers. This effect is due to the fact that the phase filtering follows the behavior of the pits and is very sensitive to the presence of outliers. It was, therefore, appropriate to add a rough phase filter as an initial step in the outlier detection process.

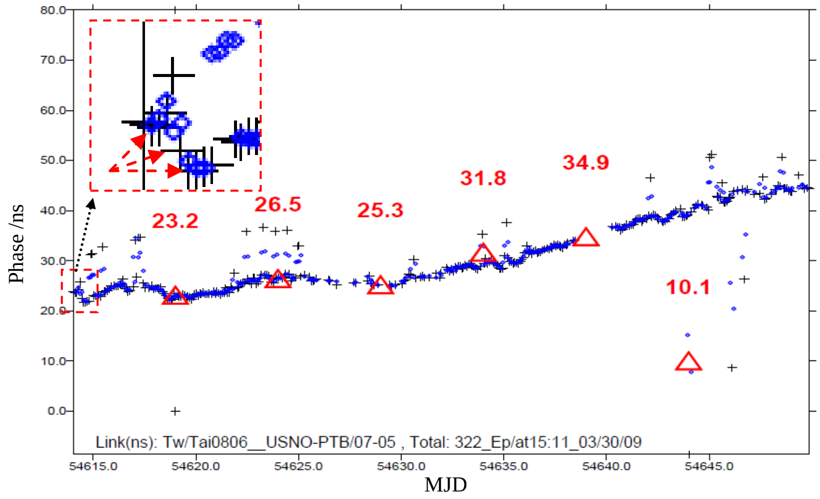

Figure 6. USNO-PTB Ku-Band time link for month 0806. The black crosses (+) represent real data, the blue dots (○) the smoothed data, and the red triangle (Δ) the phase values on standard dates.

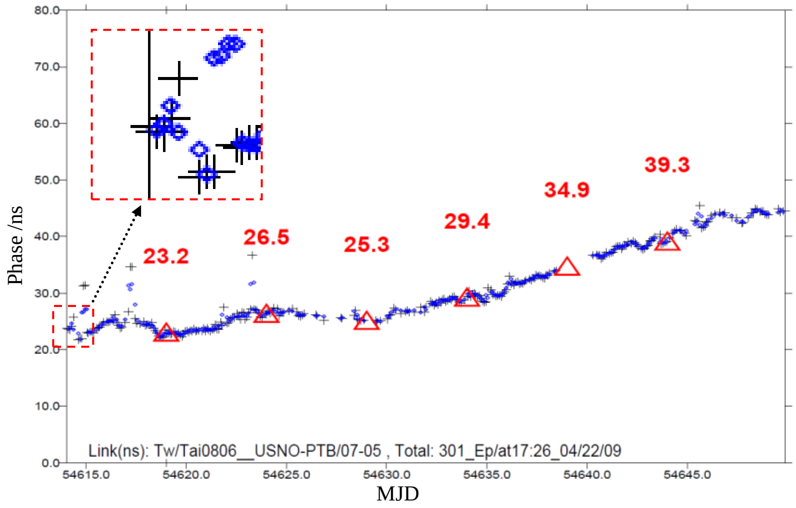

Figure 7. Cleaned link USNO-PTB Ku-Band (0806). The black crosses (+) represent real data, the blue dots (○) the smoothed data, and the red triangle (Δ) the phase values on standard dates.

ROUGH PHASE CLEANING

To apply a preliminary rough cleaning of the phase data, we applied a phase filter using a moving average and analyzed the residuals using a threshold of 3×Z (Z = 2 ns) **.**

Figure 8(a) shows the data obtained using this rough filter on the USNO-PTB Ku-band data presented in Figure 6, and the final results Figure 8(b) obtained using the complete filter.

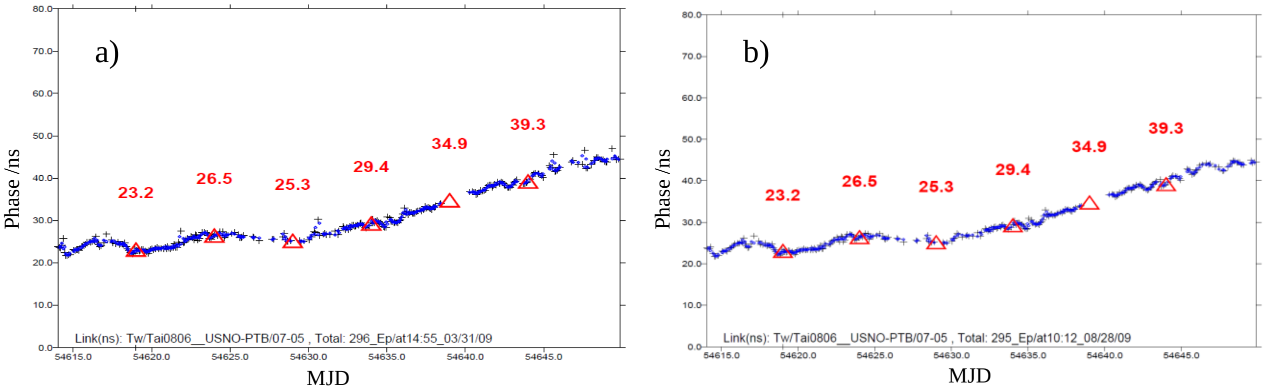

Figure 8. Result of the rough cleaning (a) and complete filter (b) on USNO-PTB Ku-Band link data (0806). The black crosses (+) represent real data, the blue dots (○) the smoothed data, and the red triangle (Δ) the phase values on standard dates.

In this example, the rough cleaning was effective, even if used alone (Figure 8(a)). Applying the refined filter to the rough-cleaned data enables the removal of the remaining outliers (Figure 8(b)). Furthermore, no useful points were removed, during this data cleaning process.

We conclude that a combination of the data-cleaning methods presented in the previous sections gives satisfactory results and no useful data are removed. This solution has been tested on different data sets and the results are presented in the next section.

EXAMPLES

The efficiency of the new data-cleaning technique has been tested on various data sets. We chose the data of June 2008 (0806), since there was a change in satellite affecting TWSTFT data; in particular, we analyzed the time links NICT-PTB and NIST-PTB, where the presence of outliers was very important. We checked the algorithm also on another data set corresponding to the month 0902, considered as a "classical month."

As described above, the cleaning parameters used in the filter are:

- the maximum value for the residuals is set equal to 2.0 ns
- the width of the window for moving average smoothing is chosen to be 12 points
- the maximal number of successive outliers that can be detected is defined as 12 points.

Classical Situations (Month 0902)

The USNO-PTB X-band link usually needs outlier detection; in 0902, it presents several isolated outliers. Figure 9(a) shows the original USNO-PTB X-band link for the period 0902 and, in Figure 9(b), the corresponding cleaned data. We conclude that this cleaning technique is efficient for isolated outliers, since it does not remove any useful data.

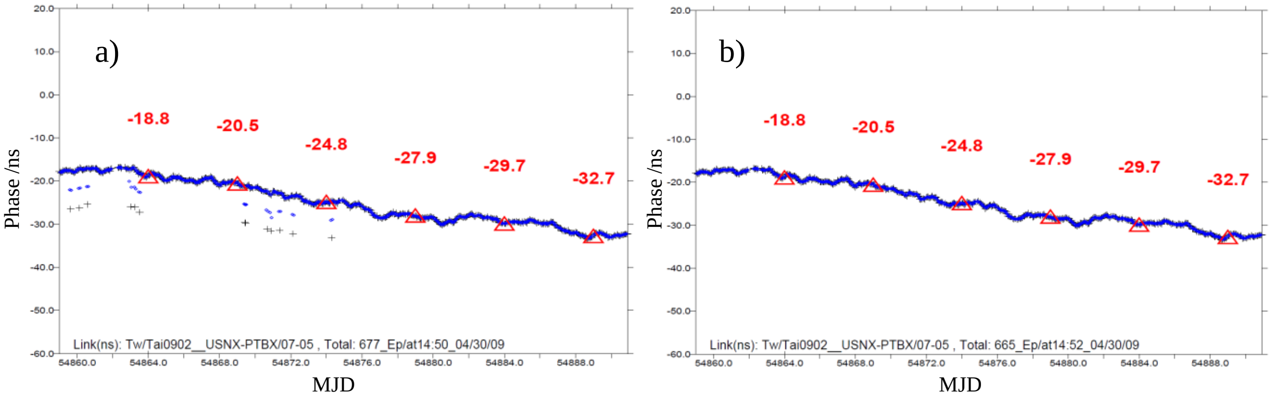

Figure 9. The original USNO-PTB X-band data for 0902 (a) and the cleaned data (b). The black crosses (+) represent real data, the blue dots (○) the smoothed data, and the red triangle (Δ) the phase values on standard dates.

Particular Cases (Month 0806)

Due to a change of satellite on June 2008, many TWSTFT time links were affected by a large number of outliers; these data are interesting for testing the efficiency of the new filter. Two very noisy time links were chosen to perform the tests: the NICT-PTB link, which shows a combination of data holes and outliers, and the NIST-PTB link, which shows a large number of outliers.

The NICT-PTB link is an interesting and anomalous case, because it contains a hole surrounded by outliers and in addition shows a very high noise. In this case, it is not simple to distinguish between the outliers and the real data, and it provides a challenging test of the algorithm. The filter was applied two successive times to the NICT-PTB link (Figure 10(a)) with more restrictive parameters (residual threshold of 1.5 ns instead of 2 ns). The link was significantly improved (Figure 10(b)), but the result cannot be perfect due to the low quality of the original data.

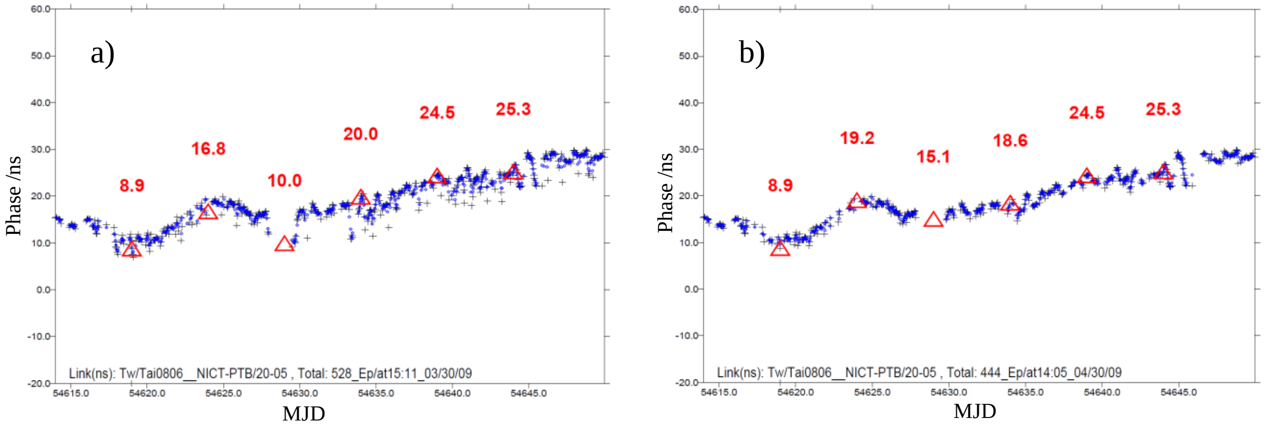

Figure 10. The original NICT-PTB data for 0806 (a) and the corresponding cleaned data (b). The black crosses (+) represent real data, the blue dots (○) the smoothed data, and the red triangle (Δ) the phase values on standard dates.

The NIST-PTB time link related to the period 0806 (Figure 11(a)) represents another complex case, similar to USNO-PTB Ku-band data presented in the previous paragraph, but with a larger number of outliers.

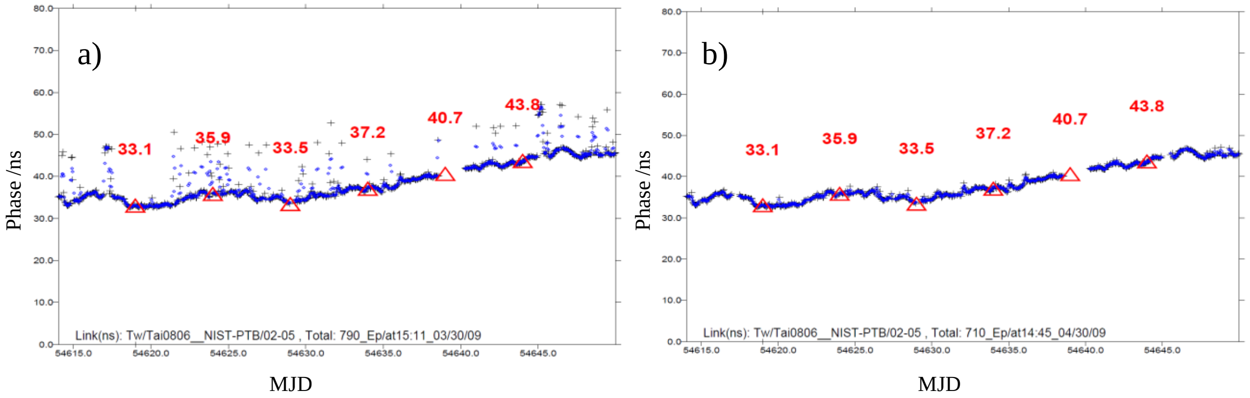

Figure 11. The original NIST-PTB link data (0806) (a) and the corresponding cleaned data (b). The black crosses (+) represent real data, the blue dots (○) the smoothed data, and the red triangle (Δ) the phase values on standard dates.

The cleaned data reported in Figure 11(b) were obtained by running the algorithm twice; outlying points not removed after the first run were cleaned during the second run.

CONCLUSION

A new technique to detect and eliminate outliers from TWSTFT data sets has been developed at the BIPM. The filter consists of a combination of two different mathematical methods to detect outliers; in a first step, a rough cleaning is applied to the phase data to eliminate significant outliers that can affect the efficiency of a more refined filter, and in the second step two filtering techniques are applied separately to the frequency and phase data. A point is considered to be an outlier, and is consequently eliminated, only if both techniques detect it is as such.

We also conclude that there is no a “unique recipe” for outlier detection, the results depending on each particular situation. In the calculation of time links for TAI, the premise is to preserve the data ensemble and to avoid deleting useful data. To remain flexible, the filtering process has several parameters that can be adapted, and the possibility exists of running the algorithm twice to improve the results.

However, the filter can also be run in automatic mode in most cases; the change of satellite affecting the TWSTFT data represents a special situation.

After validation of the results, this filtering for outlier removal in TWSTFT links has been implemented for the calculation of time links in the BIPM Time, Frequency, and Gravimetry Section in order to improve and refine calculation of TAI. The process can of course equally be used on any other data set.

REFERENCES

[1] D. W. Hanson, 1989, “*Fundamentals of two-way time transfers by satellite*,” in Proceedings of the 43rd Annual Symposium on Frequency Control, 31 May-2 June 1989, Denver, Colorado, USA (IEEE 89CH2690-6), pp. 174-178.

[2] D. Kirchner, 1991, “*Two-way Time Transfer via Communication Satellites*,” **Proceedings of the IEEE, 19,** 983-990.

[3] G. Petit and Z. Jiang, 2008, “*Precise point positioning for TAI computation*,” **International Journal of Navigation and Observation,** Article ID 562878, doi:10.1155/2008/562878.

[4] BIPM Circular T (monthly), *http://www.bipm.org/jsp/en/TimeFtp.jsp?TypePub=publication*.

[5] Z. Jiang, W. Lewandowski, and H. Konaté, 2009, *"TWSTFT Data Treatment for UTC Time Transfer,"* BIPM internal report.

[6] B. Parry, 2004, *“Evaluation of Outliers in Metrological data,”* **CAL LAB: The International Journal of Metrology,** Jan/Feb/Mar, 31-37.

[7] R. Pearson, 2002, *“Outliers in Process Modelling and Identification,”* **IEEE Transactions on Control Systems Technology**, **CST-10,** 55-63.

[8] I. Sesia and P.Tavella, 2008, *“Estimating the Allan variance in the presence of long periods of missing data and outliers,”* **Metrologia, 45**, 134-142.

[9] D. F. Vecchia and J. D. Splett, 1994,*“Outlier-Resistant Methods for Estimation and Model Fitting,”* in Proceedings of the International Workshop on Advanced Mathematical Tools in Metrology, Series on Advances in Mathematics for Applied Sciences, Vol. 16 (World Scientific, Hackensack, New Jersey), pp. 143-154.

[10] K. Pearson, 2001, *“Exploring process data,”* **Journal of Process Control**, **11**, 179-194.