26 May 2009

Improvement in Two-Way time links treatment: New outliers detection technique.

A. Harmegnies, G. Panfilo

Abstract

This report summarizes the new TW filtering technique, presenting mathematical methods used to detect and eliminate the outliers in TW time transfer technique data.
Considering different techniques studied for the outlier's detection in literature we encountered two different problems:

- the difficulty on the outlier's detection in this kind of data (outlier values are not far from the good values)
- the necessity to avoid deleting good data considering the TW data are collected every two hours.

So the new technique implemented to clean Two-way time links has been set up with the main idea of detecting outliers and to avoid deleting good data. Some results are shown at the end of the document.

I) MATHEMATICAL POINT OF VIEW

The most common techniques used to detect outliers are based on the use of different statistical estimators applied to the phase data. However, in the case of Two-way time links, considering that sometimes some drifts are present in the link, the use of these statistical techniques would not be precise enough and the risk of removing good data becomes elevate. For this reason, we decided to use also frequency values to detect outliers, to be more precise. Both techniques have been combined to proceed in two different steps:

1. In the first step we clean the phase data from the big outliers by using filtering technique applied to the phase data. In this way we avoid to compromise the outlier's detection by using a more cleaned data.
2. In the second step we refine the outlier's detection by using the frequency data and the phase data. If both observations on phase and on frequency give a point as outlier, it will be removed with no doubt about the fact that it is a real outlier.

Both filtering techniques used on frequency and phase data are described afterwards.

1.1. Phase filtering :

Rough cleaning

First to apply more complex filters, it is applied a rough cleaning to clean obvious phase outliers. In this way, both filters are used on data already cleaned thanks to a moving average smoothing on phase data described further (with an upper limit of detection for outliers to ensure removing only real outliers).

Moving average smoothing:
A second phase filter is applied by the following way:
A moving average smoothing is computed, and then, residuals are determined:

$$residuals = \left| x_i - x'_i \right|$$

If $residuals > X\ ns$ then the point is considered as an outlier.

In order to avoid removing good data, it has been decided to add another condition to remove a point: its frequency value is also tested.

1.2. Frequency filtering:

For this new cleaning technique, the statistical tool used for frequency outliers' determination is the **Median Absolute Deviation**, which is defined as followed [1-4]:

$$MAD = median_i(\left| X_i - median_j(X_j) \right|)$$

The MAD is a robust statistic that is more resilient to outliers in a data set than the standard deviation because of based on median.

The estimation of the "standard deviation" can be determined using the MAD:

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

where K is a constant factor scale, which depends on the type of the distribution and for a normal law distribution of data K is taken to be ≈ 1.4826 [1]. By choosing K=1.4826 the expected value of $\hat{\sigma}_{MAD}$ is equal to the standard deviation σ for normally distributed data.

According to Sesia and Tavella [2], the MAD is the most widely filter used to detect and remove outliers. For this, a threshold t has to be defined, and we define X_i as outlier if:

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

In our case, the threshold has been defined as t = 3, that is corresponding to about 1% of outliers if the law followed is normal [2].

1.3. Short description of the new cleaning principle:

Data determined as outliers with rough cleaning are removed and those determined as outliers in phase **and** frequency thanks to both techniques are removed too.
Results are shown is the following part.

II) RESULTS

On good time links, the technique does not remove any data as we expected.
But in case of wrong links some results of cleaning are shown below (example for TW links of June 2006):

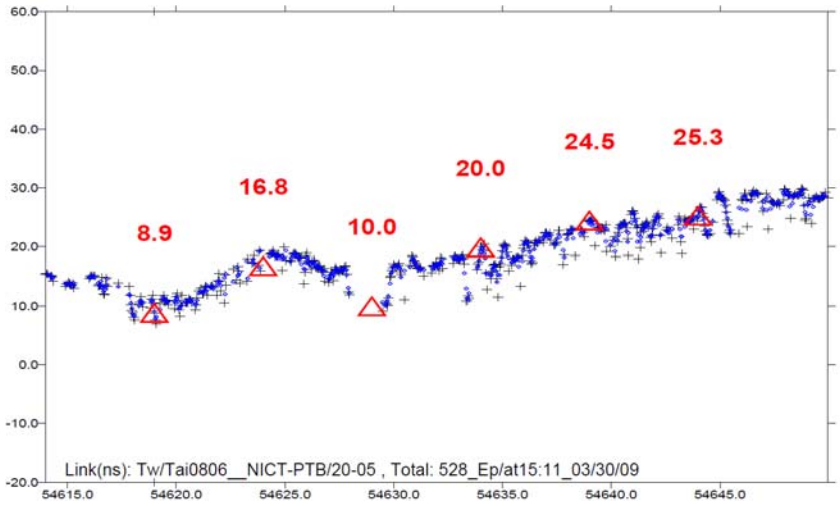

Figure 1: *NICT-PTB link (June 2006) - Original.*

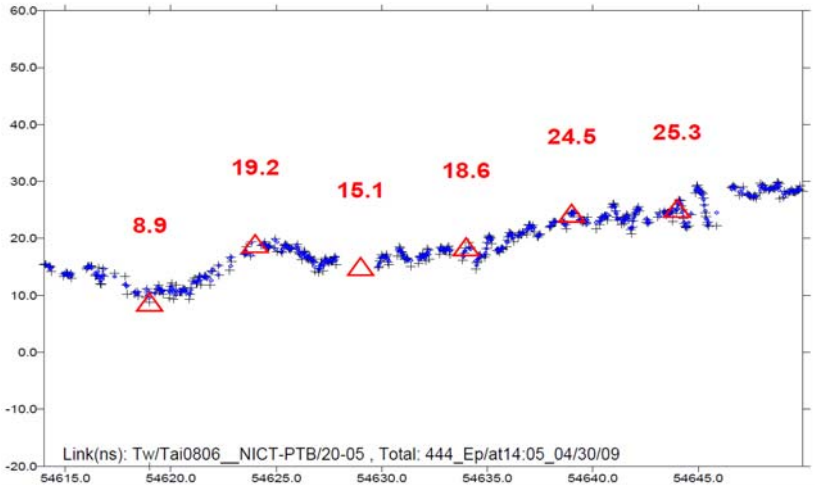

Figure 2: *NICT-PTB link (June 2006) – Cleaned.*

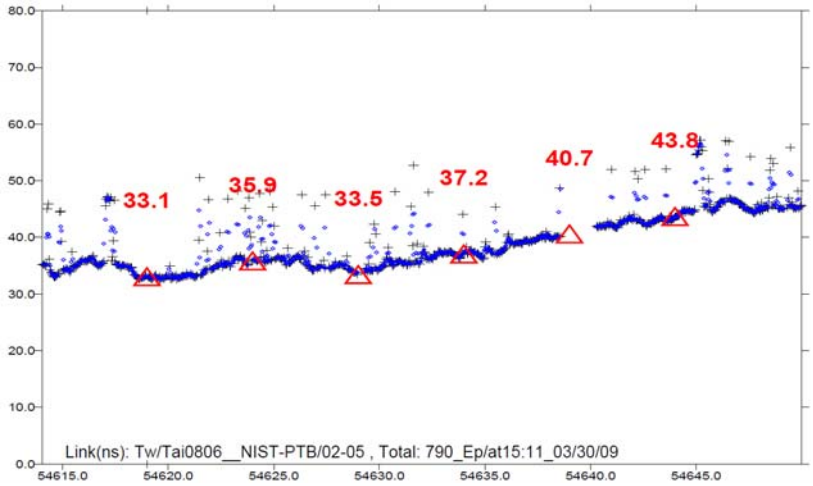

Figure 3: *NIST-PTB link (June 2006) - Original.*

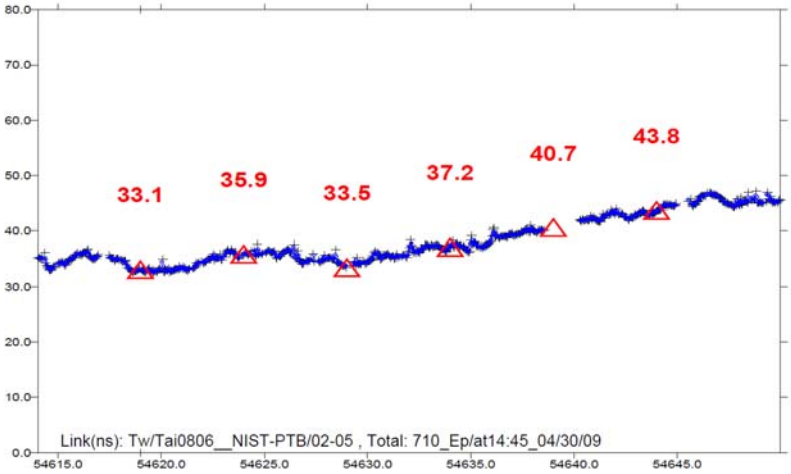

Figure 4 : *NIST-PTB link (June 2006) – Cleaned.*

CONCLUSION

The new technique established to clean TW data answer to all that we were expected: it does not eliminate any correct data and tests performed with this new technique show that this technique gives good result independently of the time link used.

REFERENCES

[1] http://en.wikipedia.org/wiki/Median_absolute_deviation.
[2] Sesia I., Tavella P., “Estimating the Allan Variance in the presence of long periods of missing data and outliers”, Metrologia 45, 2008.
[3] Pearson K., “Outliers in Process Modelling and Identification”, IEEE, 2002.
[4] Vecchia D.F., Splett J.D., “Outlier-Resistant Methods for Estimation and Model Fitting”, Advanced Mathematical Tools in Metrology, Oct. 1993.