Computer-aided quality assurance of high-resolution digitized historic tide-gauge records

Hartmut HEIN, Ulrich Barjenbruch, Christoph Blasi, Stephan Mai
German Federal Institute of Hydrology
Topic: J Accurate hydrodynamics

INTRODUCTION

In the recent era of strong discussion of climate change there is the growing need of historical data of observed sea level as the basis for different purposes from daily practical routines like navigation, operational modeling for storm-surge warnings, fundamental statistics and research on climate change itself. In a changing world long-term series of observations are essential to make decisions for the management of coastal waterways. Nevertheless, the data for all these purposes must undergo the restrict quality assurance to avoid incorrect planning or also for the sustainability of coastal managing. Future analysis of the data, allows to investigate in changes of astronomic tides and wind surges. High resolution data are one condition sine qua non for the basic understanding of both, the natural and man-made changes in coastal waterways.

During the recent meeting of the GLOSS Group of Experts (7-11 November, 2011, Paris) the rescue of tide gauge data which currently stored in non-computer forms (sheets, tabulations, etc.) was addressed (Circular Letter, IHB File No S3/2705). Our study reports the difficulties connected with the digitalization of tide gauge data in paper form. The crucial challenge is situated in the quality control of the data. Generally, these data are so extensive that automatic methods, so called Computer-aided quality assurance (CAQ), must be used to identify failed digitization, data gaps or distortions of water levels.

Analogues sea-level records include many ambiguities and errors in the time series, which may disturb the automatic data processing. We analyze the suitability of methods to detect this uncertainties. Different statistical methods like spectral analysis or fuzzy-logic show good results in detecting outliers and also filling gaps. Gaps in time series are a serious problem, and every method that is used to fill them can give only estimations. For the time being, the fuzzy-logic methods investigated to find a suitable solution.

DATA SETS

In the archives of the German Federal Waterways and Shipping Administration about 10000 years of paper sheets from historic automatically but analogically recording tide gauges wait to be recovered and digitized. This is one of the biggest sea level archive worldwide, and much effort is necessary to make it available to science and management. With this dataset it is possible to reconstruct the whole tide curves of about 100 tide gauge stations - some of them started to work in the 19th century.

Table 1: Tide Gauges

Tide gauge	**Location** (DHDN Bessel 1841)	**Digitalizited Years**
Borkum Südstrand	5.938.584,00; 2.543.850,72	1949 - 1985
Emden Neue Seeschleuse	5.912.318,32; 2.579.065,03	1949 - 1982
Mellumplate	5.960.449,00; 3.440.237,00	1964 - 1990
Bremerhaven Alter Leuchtturm	5.934.916,00; 3.471.446,00	1965 - 1972

For this study we use a test data set of O(100) years to detect general challenges of the digitalization; only to give two numbers: 100 years means more than 3000 paper sheets of more than 5 million data points. The four tide gauge stations, their location and the related years are given in table 1. For the given years equidistant points with a distance of ten minutes were digitized - if the paper sheets were found in the archive.

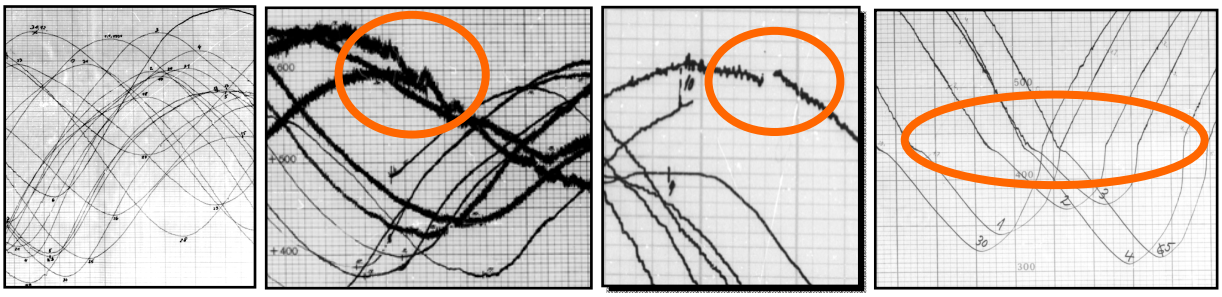

Figure 1: Example snippets from paper sheets to demonstrate the challenges of the digitalisation.

Figure 1 shows some examples from paper sheets, which were digitized. Obviously, several challenges in the digitalization process can be seen. First, there is not only one tide-curve, but several overlaying lines, which must be assigned. Secondly, on historic paper sheets wind-waves are not every time damped mechanically in the writing process, so the line mutates into a thick doodle. Several cut-offs can be found. We detected also curios deformations of the drawn lines. For example, on the outer right side of Figure 1 pondage of water during low-water is shown, this is, however, not the water-level which is representative for the overall region near the tide gauge.

One critical point for the quality control is that the sea level in the German Bight is dominated by astronomical tides (mainly semidiurnal, M2). They are in general deterministic, but they also change with time (Müller, . The starting point therefore is that we must take into account both, the deterministic and the stochastic part of the tide gauge observations.

OUTLIER TEST

The first test which is applied is that for outliers, which are detected by the mean of a nearest neighbor running standard deviation test on the first deviation of the dataset. These new time-series of the running mean standard deviation is tested for outliers following the Thompson rule (Mueller et al., 1973). Because of the tidal signal in the data this stepwise procedure provides to assign tidal self–oscillations as outliers. We done this procedure for both, the time values and the water levels.

Figure 2 shows results of the estimations of outliers of the water level. Because of the inherent tidal signal in the data set the values change with a constant frequency. Nevertheless, some peaks can be seen visually. Such values, which are a kind of suspect by the means of the standard deviation must be flagged. We use a percentile based model to detect possible outliers (orange dots in figure 2). With these method O(500) outliers were found in the dataset. Two typical examples could be seen in Figure 2b, 2c. Figure 2b shows a sudden jump in the time series, which may be related to the wrong assignment of a tide curve. Figure 2c shows an example which demonstrate the expose of the mechanics of the tide gauge. This yields the next challenge, the detection and the closure of gabs.

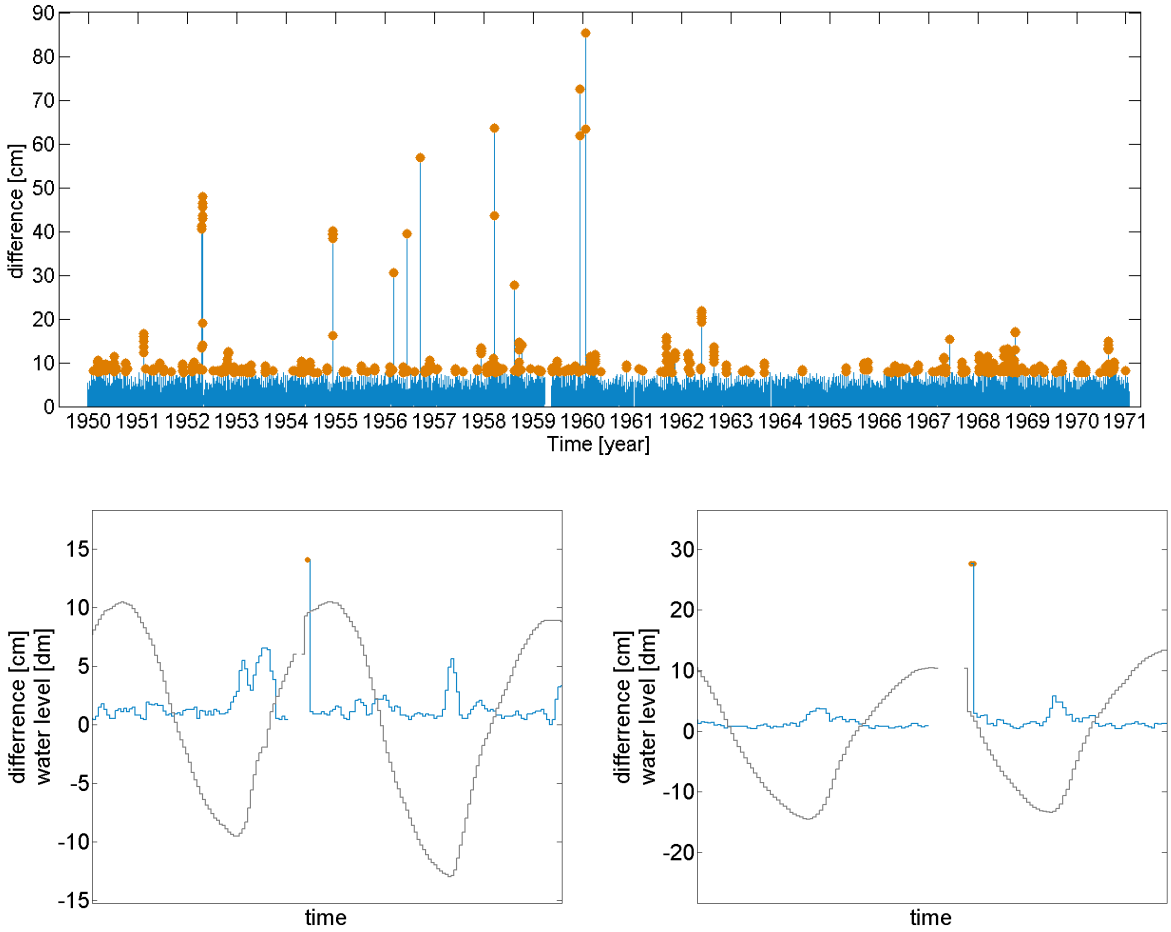

Figure 2: a) Marked automatic detected outliers. b), c) Examples for outliers. [Units are cm, except for the tide curve: 1o cm and Reference gauge-zero –500 cm]

GABS

Frequency based closure of gabs

Notwithstanding, the data have several gaps, almost for two reasons: 1. no sheet was found, or 2. because of gabs in the observations on the sheet. Figure 3 shows two typical examples for gaps. In Figure 3a a coastal hydrologist has manually inserted some curves for high and low waters. Figure 3b shows some small gabs, which are related to short interruptions of the curve on the sheet.

To reconstruct the data inside the gabs we use the so called Lomb-Scargle Periodogram (LSP; (Lomb, 1976); Scargle, 1982), which was introduced to detect sinusoidal signals in noisy unevenly sampled time series, so it can be useful also for tide gauge time series with gabs. In practice, the first step of the method is to remove the mean value of original time series from each observation. Then the Lomb-Scargle periodogram is equivalent to a linear least-squares fit of sine and cosine model functions to the time series. In detail:

$$P(f) = \frac{1}{2s^2}\left(\frac{\left[\sum_i (y_i - \bar{y}) \cos 2\pi f (t_i - \tau)\right]^2}{\sum_i \cos^2 2\pi f (t_i - \tau)} + \frac{\left[\sum_i (y_i - \bar{y}) \sin 2\pi f (t_i - \tau)\right]^2}{\sum_i \sin^2 2\pi f (t_i - \tau)} \right) \qquad (1)$$

Here the time constant of the frequency is defined with:

$$\tan(4\pi\tau) = \frac{\sum \sin(4\pi f t_i)}{\sum \cos(4\pi f t_i)} \qquad (2)$$

Hocke and Kämpfer (2009) used the LSP to compute the LSP of unevenly sampled time series and reconstructed the missing values in an astrophysical series from the amplitude and phase information of the dominant frequencies. Muller and MacDonald (2000) used the LSP to show the relation between ice-ages and astronomy. Figure 3 shows two example results of closed gabs. In the first example we use the hand-drawn lines on the sheet originated from a former hydrologist (which are part of the digitalization) as an additional non-observed information. Visually in both cases the reconstructed curves fit sufficient well. The handmade insertions of the former hydrologist fit quite well with the reconstructed curve.

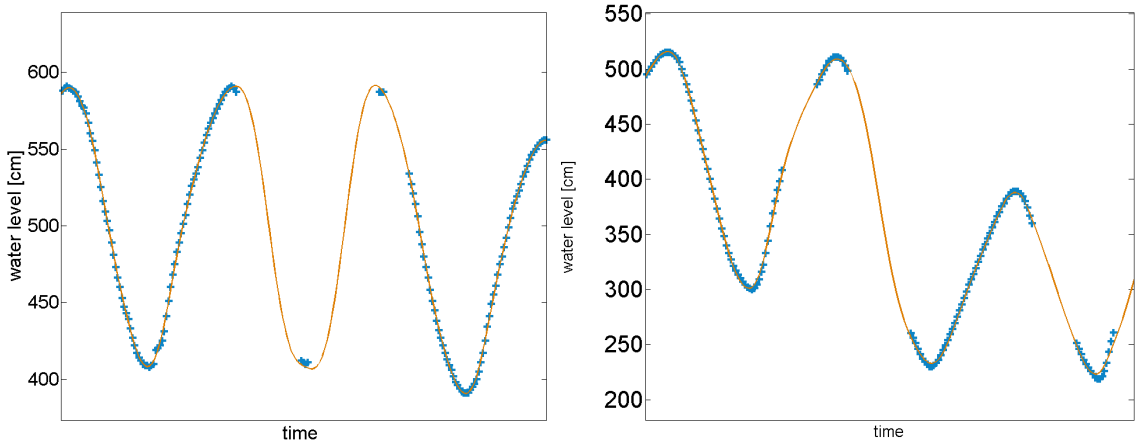

Figure 3: Examples for filled gabs with the LSP approach. Blue: digitised values. Orange: Reconstruction.

Fuzzy based closure of gabs

Next, we use the fuzzy logic approach to fill gabs inside the time series, this methods is only possible if two comparable tide gauge data set are available. But if there are, fuzzy logic is a quite effective method to fill gabs of time series. Thereby, a fuzzy inference system simulates the behavior of the sea level system by means of "if-then" rules of correlations in the different gauge data. Fuzzy logic based on fuzzy sets and membership functions, which depict the tide gauge measurements to fuzzy sets and towards suitable logical operations on these quantities and their inference. In the sense of water levels Hein (2011) uses a fuzzy inference system for backward prediction of water levels and also for quality control and find quite well result. Thereby, fuzzy logic is also excellent with an additional quality control of time series data: with a comparison between trained and measured time series we see rather good any discontinuities, outliers and systematic errors in single time series.

Figure 4 shows the performance of the fuzzy logic approach. In the middle of the figure not only a gab, but a complete non-sense cluster of digitized values can be seen. The fuzzy logic uses values from an other tide gauge for the reconstruction, because of this there is no influence on the reconstruction by this chuck of uncertain data. However, the reconstruction work fine if we use it against the propagation of the tide curve into an estuary, but not so well in the opposite direction. We explain that with the need of additional information from topographic induced self oscillating effects in the estuary, which take commonly a non-linear form.

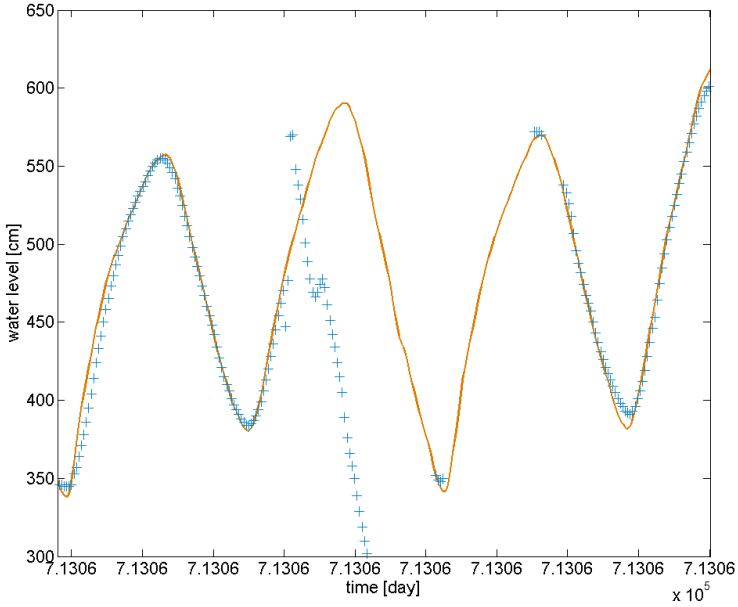

Figure 4: Example for a filled gab with the fuzzy logic approach. Blue: digitised values. Orange: Reconstruction.

UNCERTAINTIES

By uncertainties in long-term observations, there is a serious danger, that there are erroneous interpretations analysis of sea level data (Hein et al., 2010). Sometimes it is difficult to assess the data quality of long-term observations of the tide gauges, which provide a remarkable contribution to the current discussion to the possible acceleration of the sea level rise.

The one sigma mean standard deviation calculated in the procedure during the detection of outliers is O(2.5 cm). However, we must understand this value as a subjective number, representing only the difference between the value from the digitalization and that from an idealized tidal curve. But, the value works well to conclude that the sheets in general are good basis for accurate digitalization. The one sigma standard deviation of the reconstruction by fuzzy logic is O(8 cm) for one single digitized value. This uncertainty contains two parts, the uncertainty of the digitalization itself and the uncertainty of the reconstruction.

CONCLUSION

Our study reports the difficulties connected with the digitalization of tide gauge data in paper form. The crucial challenge is situated in the quality control of the data. Generally, these data are so extensive, that automatic methods must be used to identify failed digitization, data gaps or distortions of water levels. In this study we present several methods which are the base for a new degree of automations in the quality assurance. As well frequency based methods as stochastic methods - by the means of fuzzy logic - can be used to close gabs in the datasets. If more than one tide gauge was digitised we prefer the later method.

Both mentioned methods are reasonable useful to detect (subjective) uncertainties of the digitalisation. However, more robust methods are necessary for the final homogenisation of the data sets. We prefer to remove suspect observations generously from the dataset and use reconstructed instead of them. Moreover this should be done with the combination of the two methods we

presented in this study, to take account for both, the deterministic and the stochastic part of the tide gauge observations. The next step must be the analysis of breakpoints to determinate long-term uncertainties in the data set. In a recent study we present some reasonable methods to do so (Jennings, 2012).

REFERENCES

Hein, H., Weiss, R., Barjenbruch, U., Mai, S. 2010. "Uncertainties of tide gauges & the estimation of regional sea level rise". Extended abstract, Hydro 2010, Warnemünde.

Hein, H., Barjenbruch, U., Mai, S. 2011. "What tide gauges reveal about the future sea Level", Aqua Alta 2011, Hamburg, http://acqua-alta.de/fileadmin/design/acqua-alta/pdf/abstracts/paper/13_10/Hein_Harmut_full_papers.pdf.

Hocke, K. and Kämpfer, N. 2009. "Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram", Atmos. Chem. Phys., 9, 4197–4206, doi:10.5194/acp-9-4197-2009.

Jenning, S. Hein, H., Mai, S.; Schüttrumpf, H., 2012. "Breaks and long term trends of the tidal characteristics in the southern German Bight", International Conference of Coastal Engineering, Santander, 2012.

Muller, R.A., MacDonald, G.J. 2000: "Ice Ages And Astronomical Causes. Data, Spectral Analysis and Mechanisms", Chichester, UK: Praxis Publishing Ltd - ISBN 1-85233-634-X.

Mueller, P.H., Neumann, P. and Storm, R. 1973. "Tafel der mathematischen Statistik", VEB Fachbuchverlag, Leipzig.

Müller, M. 2011. "Rapid change in semi-diurnal tides in the North Atlantic since 1980". Geophysical Research Letters, 38, L11602, 6 PP., 2011 doi:10.1029/2011GL047312

Lomb, N.R. 1976. "Least-squares frequency analysis of unequally spaced data". Astrophys Space Sci 39: 447–462.

Scargle, J.D., 1982. "Studies in astronomical time series analysis. II. Statistical aspects of unevenly spaced data". Astrophys J 302: 757–763.

CONTACT DETAILS

Hartmut HEIN
German Federal Institute of Hydrology (BfG)
Am Mainzer Tor 1
56068 Koblenz
GERMANY
COUNTRY
Phone: +49-261-1306-5226
Email: hein@bafg.de
WEB: http://www.bafg.de/nn_222644/M1/DE/06__Mitarbeiter/hein/hein__node.html?__nnn=true

ACKNOWLEDGEMENT

The results of this study are from the "KLIWAS" research program funded by the German Federal Ministry of Transport, Building and Urban Development. The authors thank the German Federal Waterways and Shipping Administration for the operation of the tide gauges, which is a more and more meaningful job in the era of climate change.