#16334

Generalized Quality Control Approach for Raw Data

R.J. Foreman (richard.foreman@ul.com), N. Hansen, T. Neumann

UL International GmbH, Wilhelmshaven, Germany

Introduction

Data associated with wind energy come from diverse sources, encompassing time scales from a fraction of a second (e.g., accelerometers, electrical current) to years (e.g., corrosion data). The following quality-control approach for raw data (pre-calibrated) has been developed around long-term offshore wind-energy-measurement campaigns in the North Sea, including meteorological data from the FINO1 platform, and structural data (e.g., strain gauges) from the alpha ventus wind farm as part of the RAVE project (Fig. 1).

Figure 1: Long-term offshore-measurement campaigns involve a wide range of sensors tracking the meteorological, oceangraphic and structural effects within wind farms.

Example

Non-physical behaviour, such as spikes and "drop-outs", may alter the statistical reliability of time series, or even deem the measurements unusable. Figure 2 gives an example of archetypal single-point spikes in a signal. Here, we describe a general procedure for correcting the signal and the evaluation of the data quality.

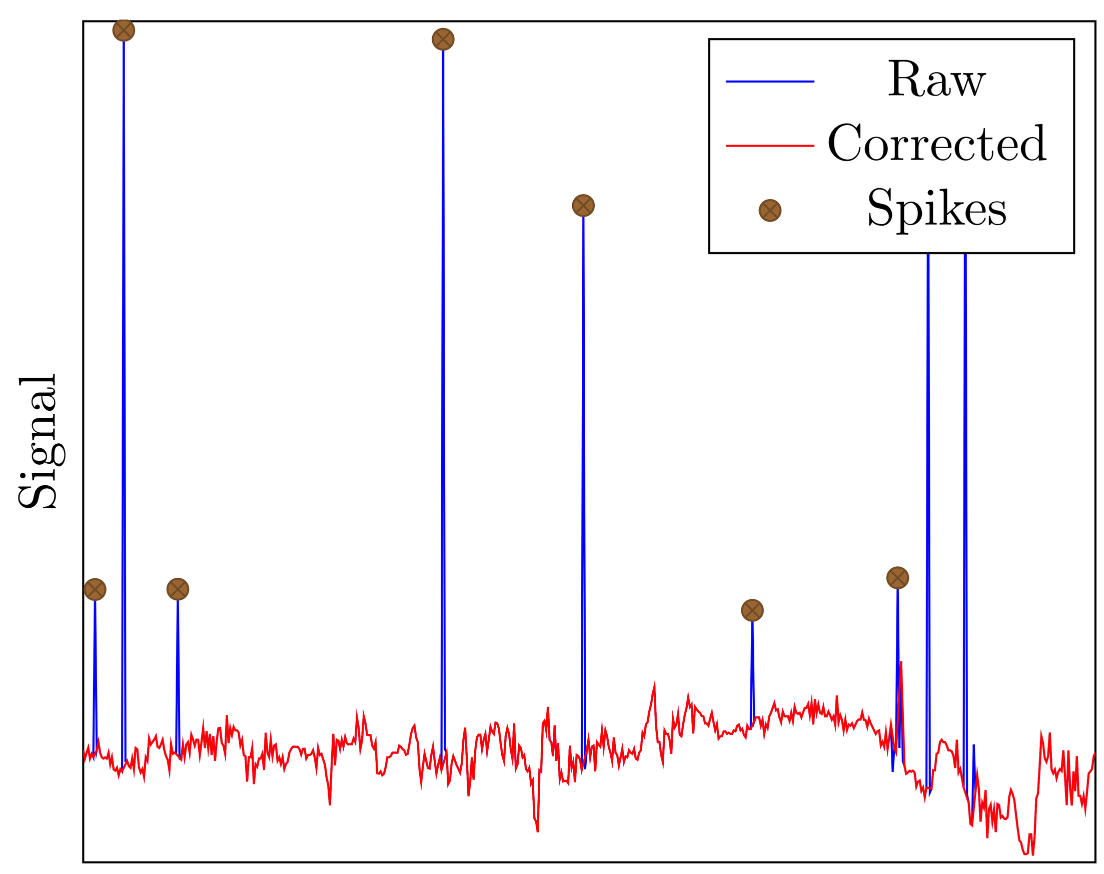

Figure 2: Signal containing spikes in comparison with the corrected signal. Outlier or spike events are indicated by points.

Event Detection

To detect non-physical behaviour, first evaluate the differenced signal so as to remove signal non-stationarity. Differenced points outside a specified number of standard deviations within a specified window of a fraction of a length of the time series are then replaced by linear interpolation. The window is then shifted through the time series until no further outliers are found. Below are examples of the spike detection, including of the differenced signal shown in Fig. 2. Here, a spike is defined as an isolated outlier with a positive (negative) value in the differenced signal followed immediately by a negative (positive) value as indicated in Figs. 3 and 5. The procedure is also able to detect drop-outs in the signal by detecting large step changes, whose presence are revealed by evaluating the correlation coefficient (Eq. 1).

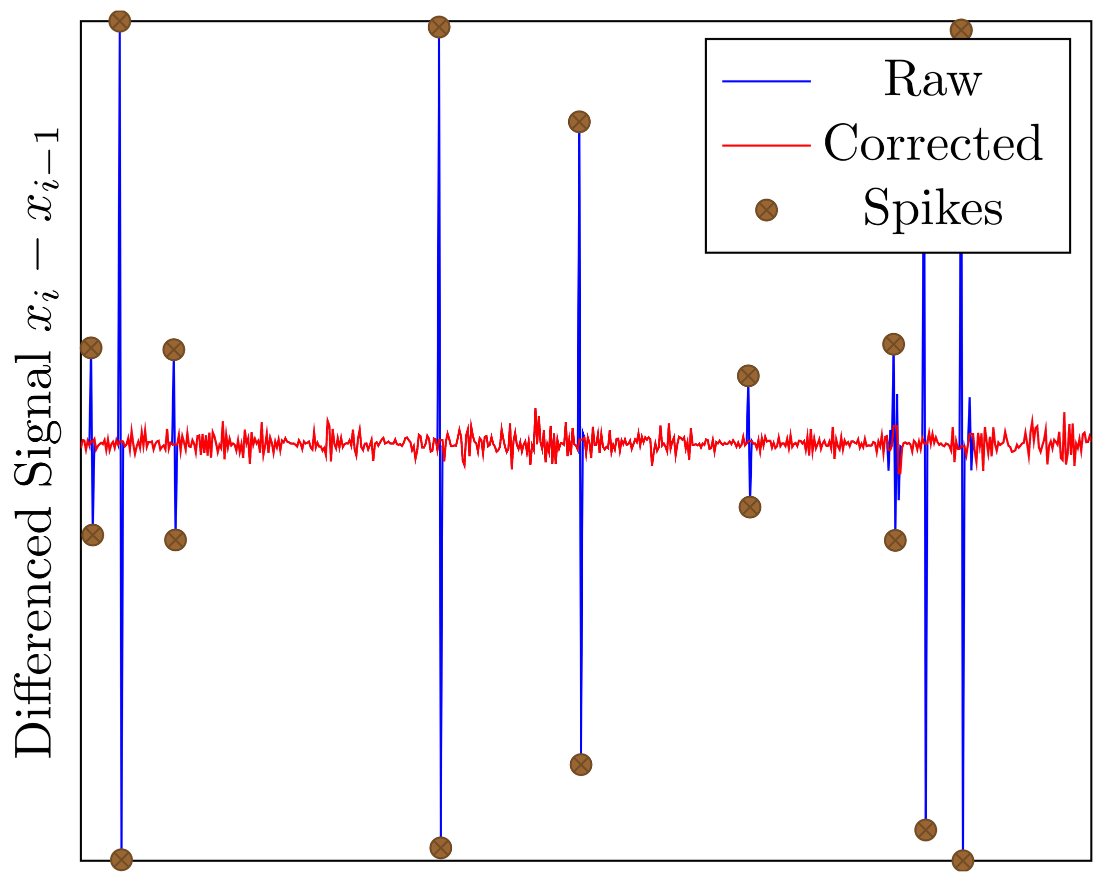

Figure 3: The differenced $(x_n - x_{n-1})$ signal shown in Fig. 2 compared with the corrected (despiked) signal and the identified spikes with consecutive minimum/maximum values.

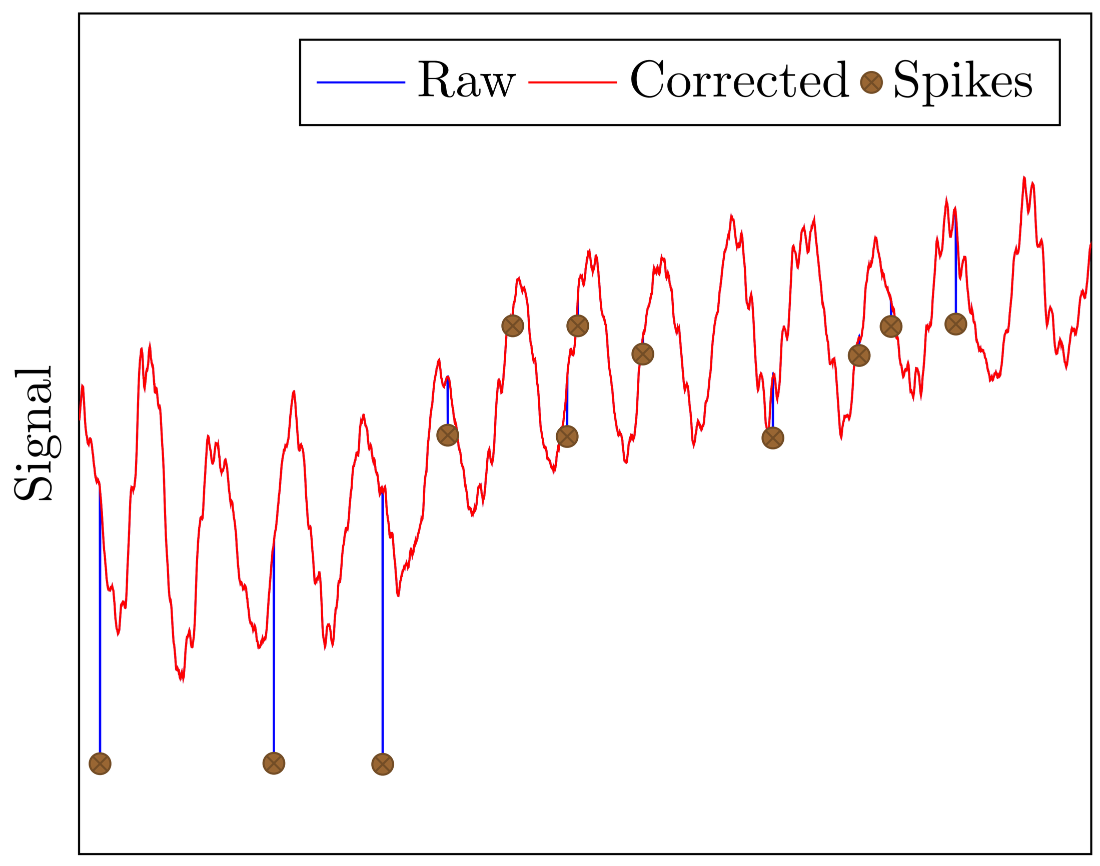

Figure 4: Signal containing spikes embedded within the sinusoidal behaviour compared with the corrected signal. Plausible spike events are indicated.

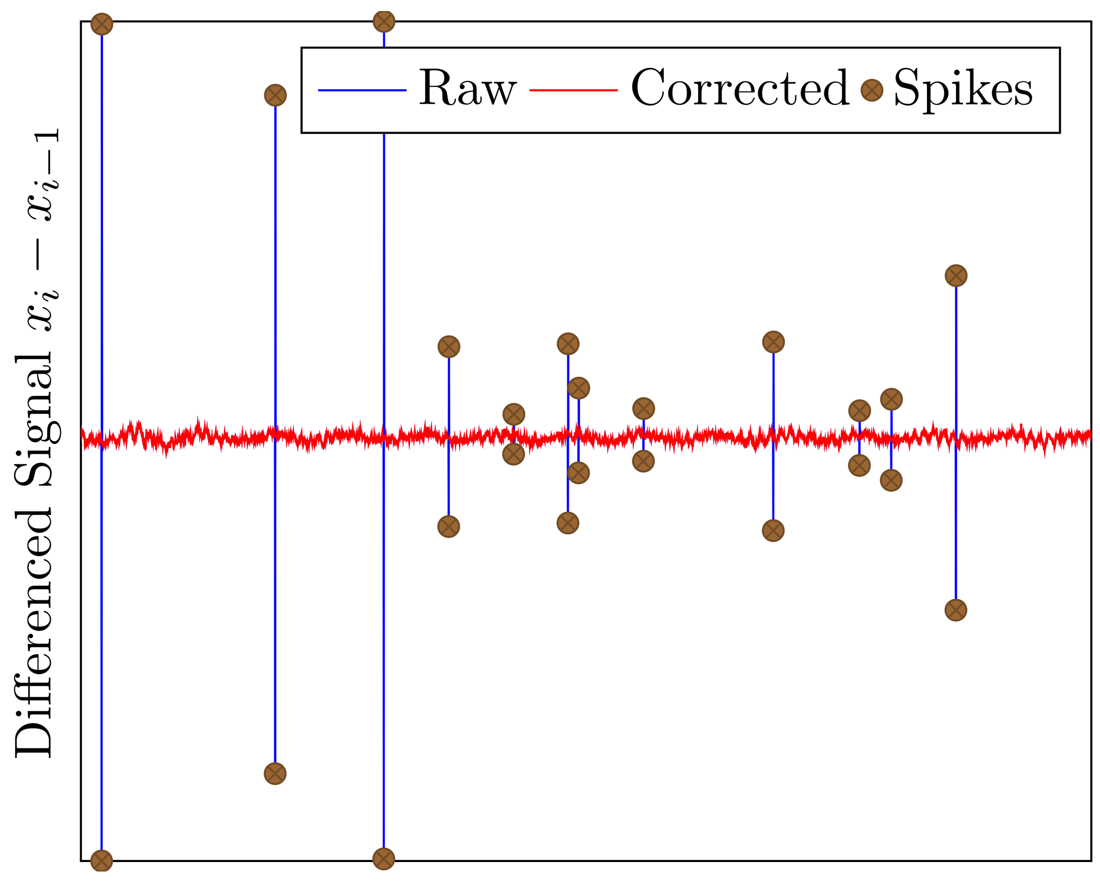

Figure 5: The differenced $(x_n - x_{n-1})$ signal shown in Fig. 4 compared with the corrected (despiked) signal and the identified isolated spikes.

Data Evaluation

The results of the algorithm proceed through the evaluation of the number of spikes n and the correlation coefficient,

$$r = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}, \quad (1)$$

between the raw and corrected signal (Fig. 6), where n is the size of the raw x and corrected (e.g., despiked) y time series, with their corresponding mean μ and standard deviations σ. The value of r varies from -1 to 1, representing perfect negative or positive correlation, respectively, with a value of zero implying no linear correspondence. For example, a signal whereby no events are detected is essentially compared with itself, giving $r = 1$.

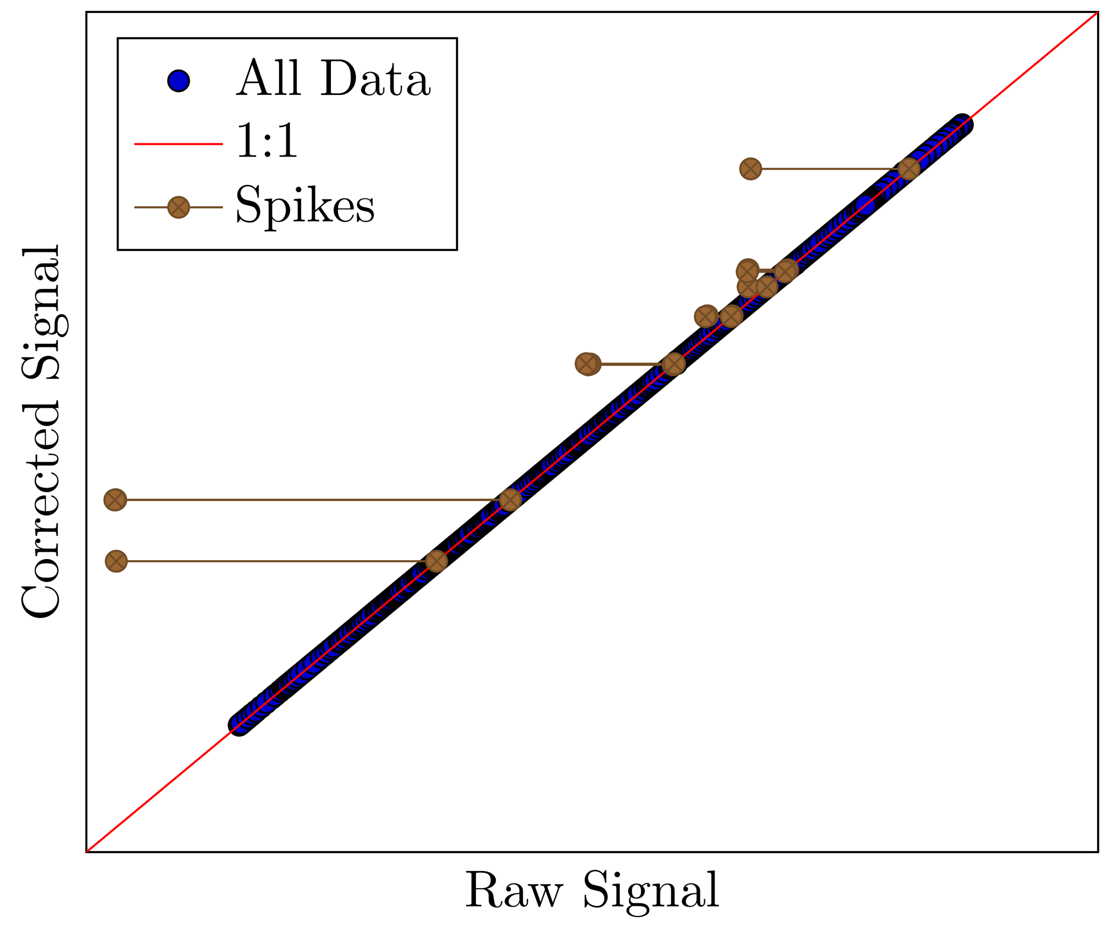

Figure 6: Correlation between the raw and corrected signals shown in Fig. 4, with the spikes indicated relative to the 1:1 line $(r = 1)$, where the correlation coefficient r defined by Eq. 1 is used to evaluate the quality of the raw signal. Here, $r = 0.74$.

We wish to filter data with low values of r and many spikes n $(r < 1,\ n > 0)$, while avoiding discarding otherwise error-free data. For example, rejecting more data as $r \to 1$ reduces the data availability without necessarily finding problematic data; similarly, the data availability is lower if all data with at least one or two spikes is rejected. The shape and intersection of these two curves as a function of the data availibility indicates plausible thresholds for filtering data (Fig. 7).

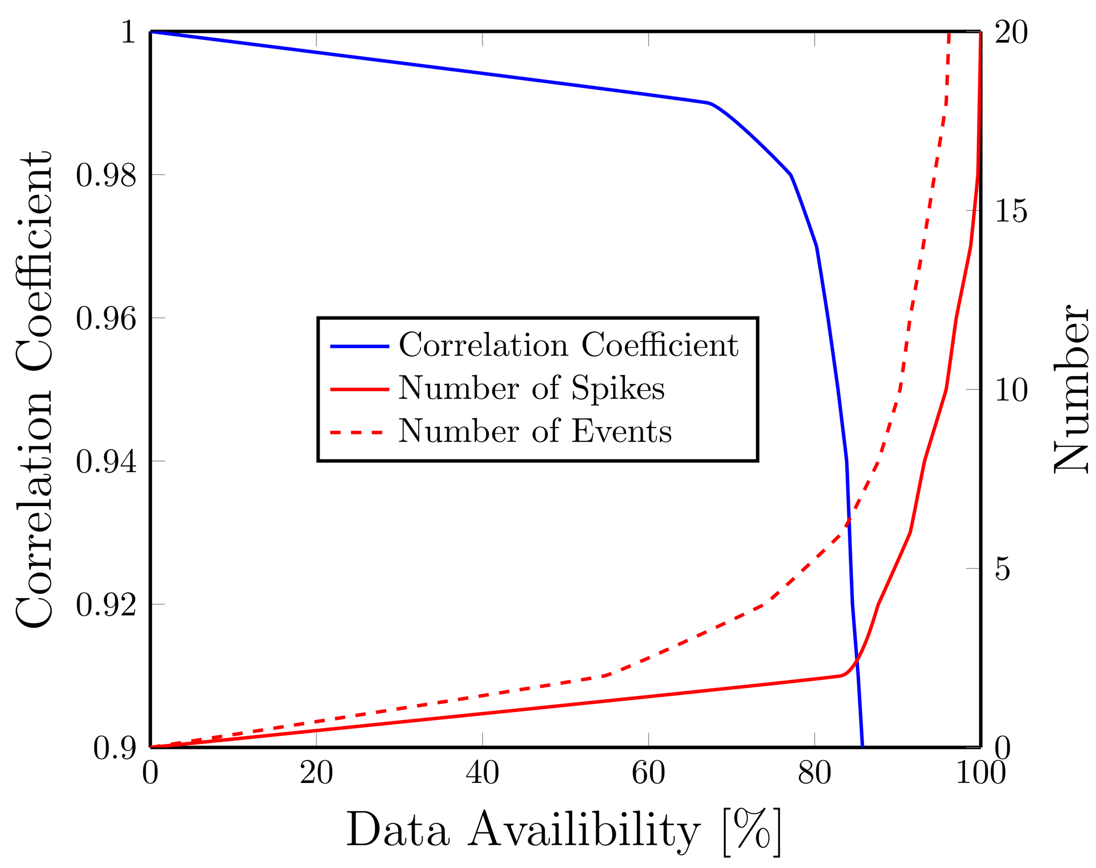

Figure 7: Quality-controlled dataset whose availibility is based on filtering data for values $< r$ (left) and $> n$ (right).

Acknowledgements This work was supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) through Projektträger Jülich (PTJ) and the German Federal Maritime and Hydrographic Agency (BSH) as part of the quality-control work package of the RAVE project (RAVE Offshore Service 0325026D) together with the BSH (M. Wilms) and DNVGL (H. Link, A. Venu).