Journal of Petroleum Exploration and Production Technology (2021) 11:819–837
https://doi.org/10.1007/s13202-020-01024-x

ORIGINAL PAPER-PRODUCTION ENGINEERING

Drilling data quality improvement and information extraction with case studies

Suranga C. H. Geekiyanage[1] · Andrzej Tunkiel[1] · Dan Sui[1]

Received: 8 June 2020 / Accepted: 13 October 2020 / Published online: 2 November 2020

Abstract

Data analytics is a process of data acquiring, transforming, interpreting, modelling, displaying and storing data with an aim of extracting useful information, so that decision-making, actions executing, events detecting and incidents managing can be handled in an efficient and certain manner. However, data analytics also meets some challenges, for instance, data corruption due to noises, time delays, missing and external disturbances, etc. This paper focuses on data quality improvement to cleanse, improve and interpret the post-well or real-time data to preserve and enhance data features, like accuracy, consistency, reliability and validity. In this study, laboratory data and field data are used to illustrate data issues and show data quality improvements with using different data processing methods. Case study clearly demonstrates that the proper data quality management process and information extraction methods are essential to carry out an intelligent digitalization in oil and gas industry.

Keywords Drilling data · Quality improvement · Information extraction

Introduction

While all other industries are aligned with digital evolution, oil and gas operations have also been taking advantage of the importance of digital and automatic technique transformation. Oil and gas industry is arguably in a new wave of digital oilfields, with a growing consensus toward intelligent and digital operations, and predictive maintenance. In recent years, hot topics such as digitalization, automation, artificial intelligence, drilling robots, deep learning, digital twins and big data have evolved from being envisions on the paper to state of the art solutions, expected to revolutionize drilling efficiency and safety.

Recently, the growing interests and trends in the oil and gas industry coupled with new intelligent sensing technologies have resulted in an overwhelming amount of data in need of having useful and valuable information in surface and down-hole environment, improving real-time decision support, enabling precise control of drilling processes, mitigating drilling incidents, optimizing drilling processes and providing visibility of wellbore conditions for real-time drilling operations, see (Thonhauser 2018; Saputelli 2020; Rassenfoss 2020; Donnelly et al. 2020; Dursun et al. 2014; Lu et al. 2017; Aibar et al. 2018). However to realize the full potentials and deal with the challenges/issues of data, as well as to develop digital, automatic and intelligent data management processes, some research questions are raised (Hegde and Gray 2017; Thonhauser 2004; Nybo and Sui 2014; Saptawati and Nata 2015). Among them, two main discussions are:

- how to develop proof-of-concept technologies/methodologies to support data acquisition, data management and processing in oilfields;
- how to precisely interpret data to provide useful and valuable information. At present, big data with its high quality becomes an essential part of digitalization. However, data quality challenges are one big obstacle of digitalization development and vary from case to case, for instance:
- **Dataset availability** Data is saved in different formats and sources. Challenges regarding data integration, availability, usage, storage, visualization and database development always exist. Producing data hub/ database in

✉ Dan Sui
dan.sui@uis.no

[1] Energy and Petroleum Engineering Department, University of Stavanger, Stavanger, Norway

an easily accessible format with additional explanation information is desired.

- **Right data** In different drilling scenarios, data used and selected for analysis can be different. It is important to identify, select and use right data with respect to predefined objectives or scenarios.
- **Data quality** The issues related to systematic/random/gross errors due to sensors failure, malfunction, incorrect calibration, user entry errors, sampling frequencies, corruptions and so on are often met, see the discussions in Bello et al. (2017); Dickson (2014); Temer and Pehl (2017); David (2016); Nybo et al. (2012). High quality is desired to provide valuable and useful information. Data filtering, cleansing, outlier removal and data correction are necessary steps to improve data quality.
- **Data structure** One problem is that a large amount of data (for example in drilling daily reports) is "unstructured" or "semi-structured", which means it is difficult or costly to extract data or routinely query and analysis.
- **Data diversity** In some cases, substantial amount of historical data does not possibly cover all situations and provide all information, simply because certain feasible and relevant combinations of events may not have occurred. It motivates the use of simulations from sophisticated models or experiments that generate huge amounts of data augmenting the historical (logged) data, and making data analysis necessary.
- **Data versioning** is another hidden challenge associated with the drilling data. Raw real-time data, edited real-time data and memory (historical) data are gathered and stored during drilling. Moreover, the volume of the data produced over the time accumulates and grows. The questions "should an operator store all above categories of the data or a selected category for a period?" needs more attention.
- **Data streaming** Down-hole-to-surface communication and connectivity issue is an industry specific data streaming challenge. In addition, drilling digitalization must also address the requirement of batch or continuous processing of the data and distribution to multiple targets in real time (i.e. a distributed solution or a centralized hub data processing challenge).
- **Multiple data sources** Data is collected from multiple sources in real time which causes some common challenges to merge the data. Since data could originate from either surface sensors, down-hole sensors, control system outputs or manual inputs that describe the operation, all data should be synchronized with a common time reference. As an example, the clock time in every microcontroller or PC, varies slightly. If the sampling frequency is low, for instance, 10 Hz, there is typically only a need to calibrate each system ahead of the operation and in regular intervals to prevent that the clock times get out of synchronization. If, however, one is working with data sampled at hundreds if not thousands of Hz (number of samples per second), only a small offset in synchronization could cause the data to become highly inaccurate once the data gets merged with data from other sources. One solution to this problem is to transmit a common pulse to all sensors or microcontrollers, requesting measurements from all sources simultaneously.
- **Data calibration** Another challenge when aggregating data from several sources is calibration. Before data logging begins, all systems should get calibrated to ensure that data from each operation has the same base value, unit and threshold. One example could be if the hook load gets measured and calibrated for one bottom hole assembly (BHA) configuration, and the hook load is not updated for another configuration for a later operation. In such case, the data can be merged if the user is aware of the variations. There is, however, no way that the computer can automatically work with such differences in the data, unless the variations in the data are inserted as metadata that the computer can access and use to correct the data. Data management, processing, interpretation, modelling and applications require systematic procedures, hierarchical services and management infrastructure to solve the challenges from data volume, velocity, variety and resultant complexity. Some good approaches to identify and improve above-mentioned data quality issues are recognized and presented in Mathis and Thonhauser (2007); Ouyang and Kikani (2002), for instance:
- **Range check** upper and lower boundary check and resampling of data to form a uniform sampling interval can be introduced as a good solution for invalid data issues.
- **Gap filling** algorithm capable of interpolating or extrapolating data within a set of constraints to improve sampling frequency errors and missing data points (or null values) is an option for improving incomplete or inconsistent data quality challenges.
- **Outlier removal and noise reduction** via filtering (digital or circuit) is another option for improving data accuracy. Examples for such filter algorithms are moving average filter, low-pass filer and median filter.
- **Data redundancy** challenge is addressed using technique known as data assimilation (Lewis et al. 2006). It is essentially a model-based data assimilation and prediction algorithm, which is capable of self-correcting parameters to minimize an error of a measured and expected variable during real-time operation. Data validation and reconciliation (DVR) is the other popular method to minimize measurement errors using model correlations (Sui et al. 2018; Stanley and Mah 1981). DVR allows unmeasured variables to be estimated based on combined information from process measurements

and models. In this paper, the approaches to evaluate pre-data quality to identify data issues such as missing or incomplete data, non-standard or invalid data and redundant data are presented. Then, the implementations of different data quality managing practices such as filtering, data assimilation and data reconciliation to improve data accuracy and discover useful information are introduced. Thirdly, a post-data quality evaluation and information interpretation is presented, which is conducted to assure data quality, enhance the system performance and extract knowledge/information for modelling. Finally, some results are given to illustrate our proposed methods and algorithms.

Data issues

These main data quality issues are listed below:

- **Invalid data** For the drilling data captured from drilling systems, invalid data is the data measured outside of the specific sensors measurement range.
- **Inaccurate data** Inaccurate data is recorded due to the random noise generated by the sensory equipment, electrical cables/ electromagnetic interference of other nearby equipment, power fluctuations or due to calibration fails of imperfect sensors. This is identified as white noise and outlier problem. An example of continuous azimuth data with white noise and system disturbances is given in Fig. 1.
- **Incomplete data (missing data)** There can be a number of reasons for why data is missing in a dataset. One possible reason is when different sensors get sampled with varying sampling frequencies, for instance, 10 Hz for one sensor and 20 Hz for the other. Another common cause could be hardware (electrical) failure, where the signal is lost for a short duration of time. One example of continuous inclination data with big gaps is shown in Fig. 2.
- **Inconsistent data** Data arrival with inconsistent sampling frequencies, or random sampling frequency fluctuations are also observed in recorded datasets, causing an incomplete and inconsistent data issues.
- **Redundant data** Figure 3 shows that the inclination data is measured and estimated by using two different mechanisms (measurement while drilling (MWD) and compact roto sonic (CRS)). At a given time, due to white noise or other disturbances (vibrations), these two measurements are possibly not same.

Data quality improvement and validation

Re-sampling, gap filling and range checking

To handle missing data, interpolation or extrapolation of data to fill in missing data points, i.e. gap filling algorithms based on gap time vs. sampling frequency input to improve inconsistent sampling frequency is considered first. If the gap is smaller than the accepted gap time, interpolation between edge points is implemented. Otherwise, if the gap is bigger than the defined gap time, extrapolation based on last two available data points is used to fill-in the blanks.

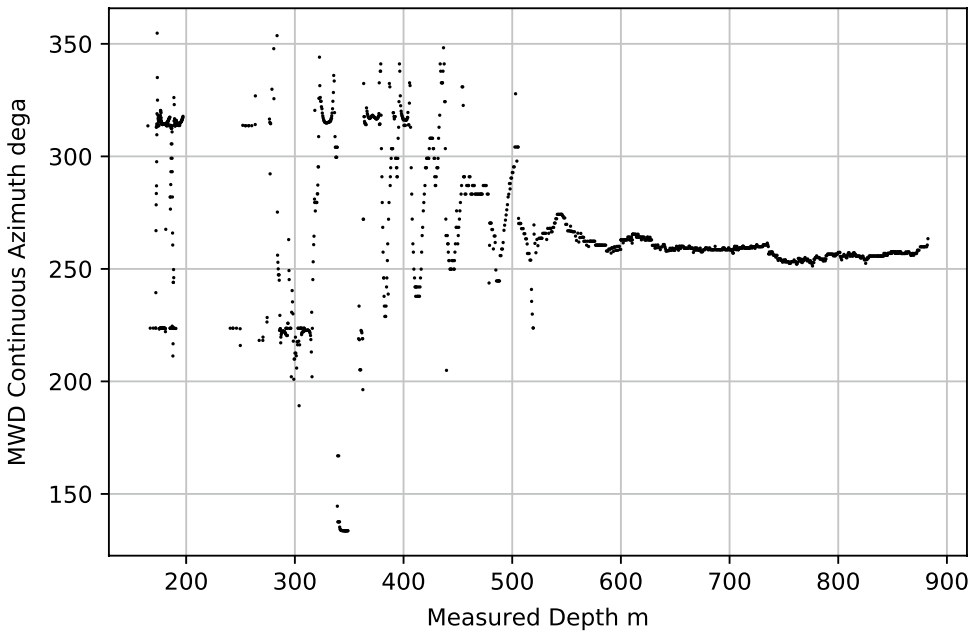


Fig. 1 Raw continuous azimuth data with measured noise

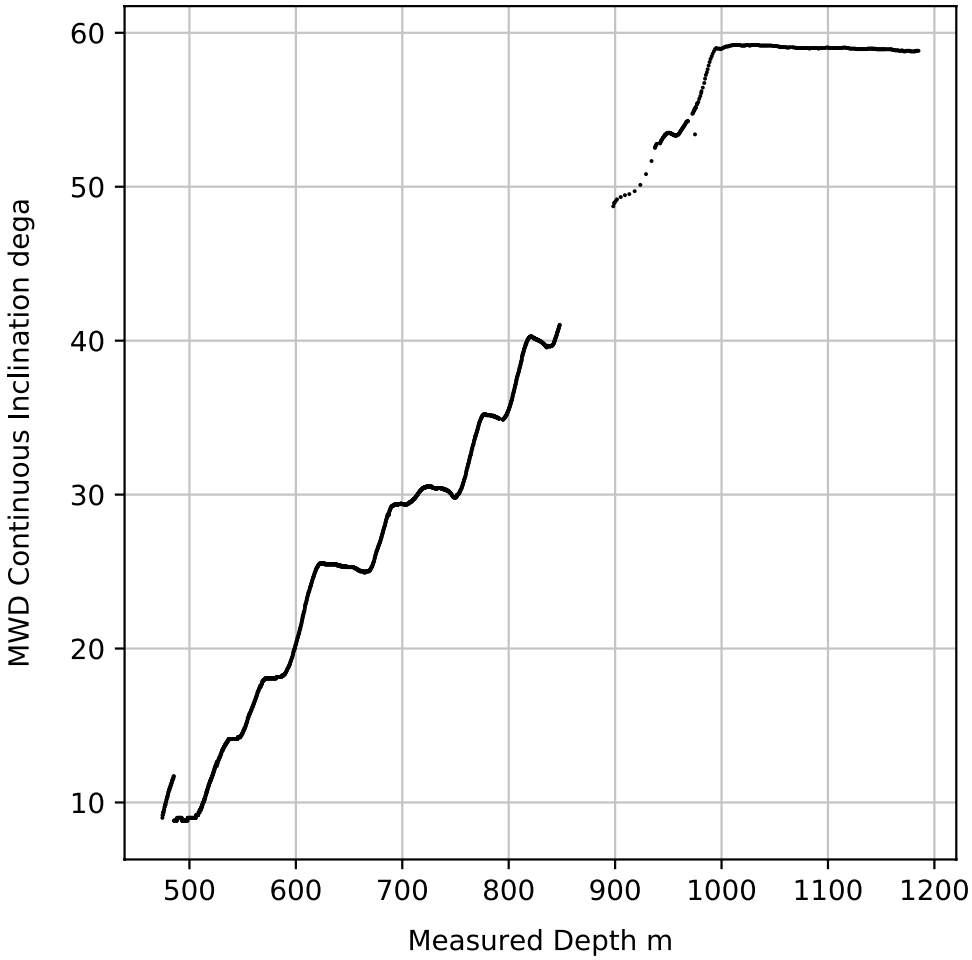


Fig. 2 Raw continuous inclination data with big gaps

To handle invalid data, a range check is performed to identify invalid values and replace them with either null values or interpolated/ extrapolated data.

Verifying average (or median) and standard deviation of the dataset before and after re-sampling or interpolating/ extrapolating to be same is used as a verification method to see that the process does not affect results negatively. Implementation of a gap filling/ data rejecting mechanisms also provides the ability to count number of data points that had to be fixed or rejected during a period (Mathis et al. 2006). More detailed discussions of gap filling on field data will be given in "Results" section.

Filtering

Digital filters to remove white noise and outliers are considered next. Many different types of filters are available for this purpose. In fact, properties and advantages of one filter over the others should be understood before its implementation.

Moving average filter (MAF) is simple, yet powerful tool used for data filtering. Generalized formula for moving average filter in discrete time domain is given below:

$$y(t) = \frac{1}{M}\sum_{j=0}^{M-1} x(t-j) \tag{1}$$

where x is the raw data, y is the processed data after filtering, t is the time coordinate and the index j corresponds to the number of convolution steps for a data point at time t, M is the number of data point span considered for the average taking. The MAF is very practical for engineers since it does not require a frequency analysis for implementation (Smith 1997). Low-pass filter (LPF) is an improvement to the MAF where better noise removal is achieved over a certain frequency selected (Smith 1997). An ideal LPF allows passing of all frequencies of the incoming signal below this defined cut-off frequency. The discrete time, digital LPF equation (first-order) is given below:

$$y(t) = ax(t) + (1-a)y(t-1) \tag{2}$$

where

$$a = \frac{t_s}{t_s + \tau}. \tag{3}$$

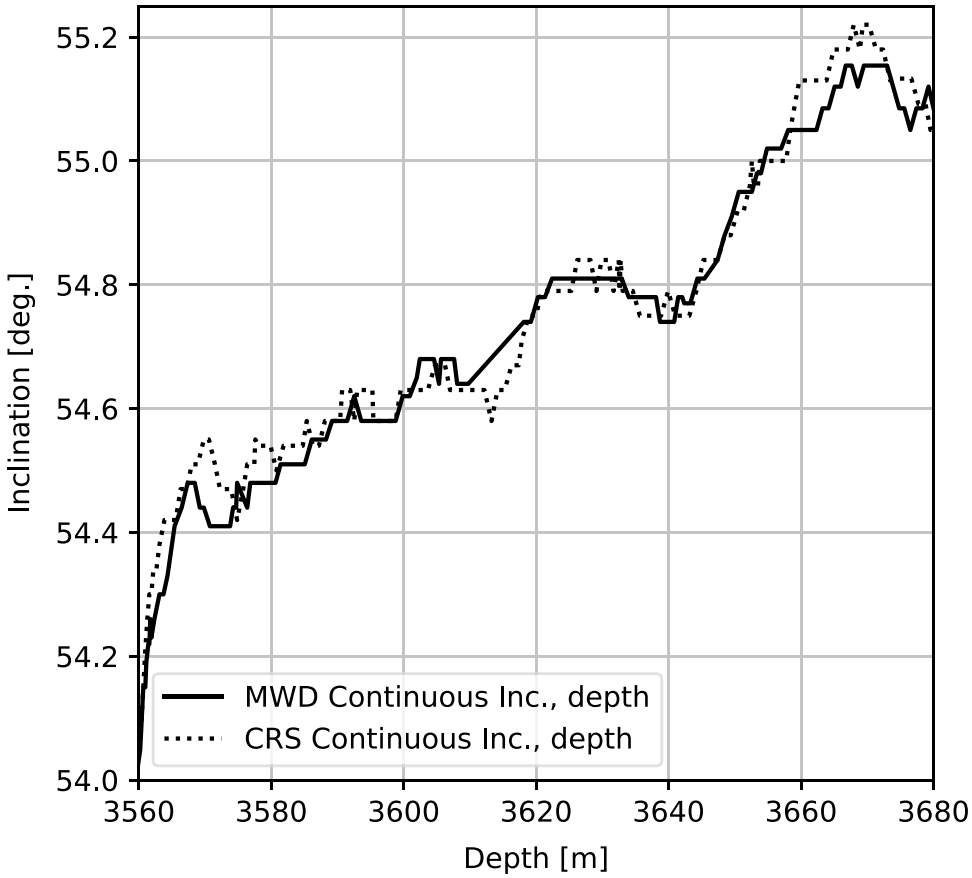


Fig. 3 Inclination measurements from two different systems

For the LPF, t_s is the sampling rate and τ is a design parameter. Normally, it is selected by $\tau = \frac{1}{\omega_c}$ where ω_c is the cut-off frequency, a boundary in a system's frequency response at which energy flowing through the system begins to be reduced rather than passing through. Generally with the higher τ, the more noises can be removed from the original data. However the higher τ most likely causes the smaller a, in turn the processed data $y(t)$ more relies on $y(t-1)$, see Eq. (3), which could lead to the information/time delay due to little new information passing with the filter. In addition to the above-mentioned first-order LPF, a second-order LPF is also examined for data filtering. The discrete-time second-order LPF equation is given below:

$$y(t) = b_1 x(t) + b_2 y(t-1) + b_3 y(t-2) \tag{4}$$

where

$$b_1 = \frac{\alpha^2 t_s^2}{1 + \alpha\beta t_s + \alpha^2 t_s^2},\ b_2 = \frac{2}{1 + \alpha\beta t_s + \alpha^2 t_s^2},\ b_3 = \frac{\alpha\beta t_s - 1}{1 + \alpha\beta t_s + \alpha^2 t_s^2}, \tag{5}$$

where β and α are design parameters. Advantage of the second-order LPF over the first-order LPF is that it provides two controlling parameters (α, β) for improving the performance. Therefore, not only time delay of the filtered signal but its amplitude can be adjusted using these parameters. However, the more design parameters are considered, the more complicated the filter becomes. Besides the above-mentioned LPFs, there are many other different low-pass filters, like Butterworth filter, Bessel filter, Chebyshev filter, etc., that are good solutions to filter out the noises, see (Smith 1997). For the most cases, the first-order low-pass filter is sufficient to remove noises from drilling data. To assess the quality (accuracy) of the filtered data set, mean and standard deviation of the data set before and after is used as an evaluation criteria, see the more discussions in our result section.

Outlier removal

Outliers are ones that are situated away from the main observation window. An important factor to consider before removing outliers is to find out whether they consist of relevant information or are the result of noises. In some datasets, for example, when dealing with kick detection or stuck pipe detection, the important information could be apparent in the outlying points. In our work, several techniques like the mean filter and the median filter have been evaluated for optimal outlier removal. The interquartile range (IQR)

method has been identified as the most optimal when dealing with outliers, see the detailed introduction of the IQR approach in Jiawei and Susanto (2019).

Data assimilation

Data assimilation approach is considered for solving the redundant data issues. In cases, where two sensors measuring the same parameter or where a parameter can be both measured and calculated, data assimilation is a powerful tool to get a better estimation (Lewis et al. 2006). If Gaussian distribution is followed by both time series data sets, each can be represented by its mean value, $\bar{m}_1$ and $\bar{m}_2$, and standard deviation σ_1 and σ_2. Assuming independence between two measurements x_1 and x_2 measured at the same time, a linear unbiased estimator $\hat{x}$ calculated by data assimilation, based on above measurements can be written as follows:

$$\hat{x} = a_1 x_1 + a_2 x_2 \tag{6}$$

where

$$a_1 + a_2 = 1.$$

Therefore, the variance of the estimated value becomes

$$var(\hat{x}) = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2. \tag{7}$$

Data assimilation method aims to find an optimal a_1 and a_2, such that $var(\hat{x})$) is minimum. To achieve it, the derivative of $var(\hat{x})$) with respective to a_1, a_2 should be set to be zero, or

$$\frac{d[var(\hat{x})]}{da_1} = 0. \tag{8}$$

Then after some derivations, we have

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \tag{9}$$

and

$$var(\hat{x}) = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \tag{10}$$

From the above discussions, the estimated/assimilated data, $var(\hat{x})$, always results in a better variance than that of either x_1 or x_2.

Data validation and reconciliation

DVR is an advanced technology which uses process information and mathematical methods in order to automatically correct raw measurements, estimate model parameters/unmeasured variables in industrial processes. The use of the DVR allows for extracting accurate and reliable information from raw measurement data and produces a consistent set of data representing the most likely process operation. The models used in the DVR are normally based on conservation laws of nature and can be either dynamic or static.

Data reconciliation can be formulated by a constrained weighted least squares optimization problem, where the measurement errors are minimized with model constraints. Given n measurements, the DVR can mathematically be expressed as an optimization problem of the following form:

$$\min_{y^*, x} J(x, y^*) = \sum_{i=1}^{n} \left(\frac{y_i^* - y_i}{\sigma_i}\right)^2 \tag{11}$$

subject to

$$f_m(x, y^*) = 0, \tag{12}$$

$$g_m(x, y^*) \leq 0, \tag{13}$$

where y_i is the raw measurement value of the i-th measurement, $y^* = \{y_1^*, \ldots, y_n^*\}$, y_i^* is the reconciled value of the i-th measurement, x is a vector of estimates for unmeasured values of the process and σ_i is the standard deviation of the i-th measurement. f_m is a vector describing the functional form of model equality constraints and g_m is a vector describing the functional form of model inequality constraints which include simple upper and lower bounds. Solving this optimization problem provides simultaneously the measurement error corrections and the estimates for unmeasured variables.

Data management flow

The flow chart for data quality management process implemented is summarized in Fig. 4. Improvement of consistency, completeness and reliability of operational data while maintaining data accuracy, availability and validity (amplitude, average and frequency/time delay) within defined boundaries, are considered as main objectives of this process.

Information extraction

In this section, an example of parameter identification is presented to illustrate how to exact the hidden information from measured data. Here, a simple drill string dynamic model is considered which is represented as a spring-mass system (Thomson 1996). It is assumed that the axial motion is independent on torsional or lateral motion. Mass of the system is assumed concentrated to a centre of gravity residing within the drill string and bottom hole apparatus.

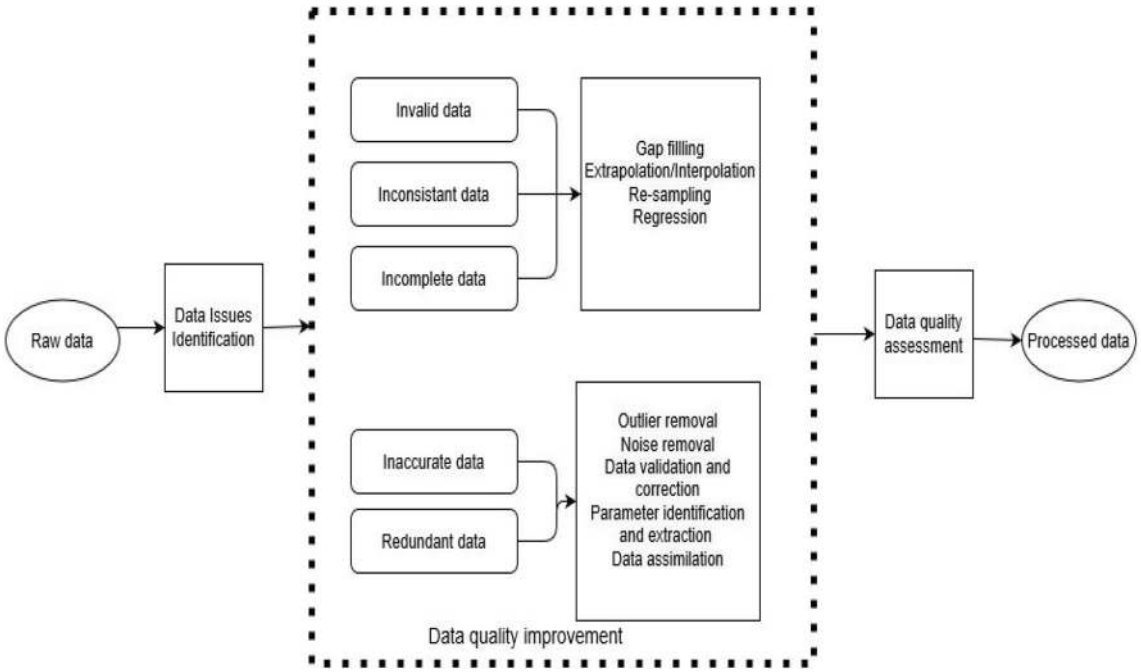


Fig. 4 Data management flow chart

Considering the momentum balance, the system can be easily expressed as one ordinary differential equation shown below:

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = f(t). \tag{14}$$

where $x(t)$ and $f(t)$ are the displacement and external force loaded on the subject at time t; m, c, k are the mass, damping and spring coefficient, respectively. Considering the initial conditions $x(0) = x_0, \dot{x}(0) = v_0$, the general solution is given in Thomson (1996) for the underdamped case $(0 < \zeta < 1)$ as:

$$x(t) = e^{-\zeta\omega_n t}(x_0\cos(\omega_d t) + \frac{v_0 + \zeta\omega_n x_0}{\omega_d}\sin(\omega_d t)) + \int_0^t h(t-\tau)f(\tau)d\tau, \tag{15}$$

where ω_n is the natural frequency given as

$$\omega_n = \sqrt{\frac{k}{m}} \tag{16}$$

and ζ is the damping ratio that describes the system dynamics, defined as

$$\zeta = \frac{c}{2m\omega_n}, \tag{17}$$

and ω_d is the damped frequency of the system. In general, the system dynamics can be divided into three cases: underdamped case $(0 < \zeta < 1)$ where the system oscillates with the amplitude gradually decreasing to zero; critical damped case $(\zeta = 1)$ where the system returns to equilibrium as quickly as possible without oscillating; and overdamped case $(\zeta > 1)$ where the system returns to equilibrium without oscillating. For an underdamped system $(0 < \zeta < 1)$,

$$\omega_d = \omega_n\sqrt{1-\zeta^2} \tag{18}$$

In (15), $h(t)$ is the axial unit impulse response function (UIRF) of the system, given as (Thomson 1996):

$$h(t) = \frac{e^{-\zeta\omega_n t}}{\omega_d}\sin(\omega_d t). \tag{19}$$

The phase angle, ψ, of the system is:

$$\psi = tan^{-1}\left(\frac{\omega_d x_0}{v_0 + \zeta\omega_n x_0}\right). \tag{20}$$

Typically, the system dynamics depends on systemic parameters m, k and c and the external force f on the subject. For the drill string system, the mass of the pipe m can be easily calculated if the material of the pipe is known. However, the calculation of the spring coefficient k has been influenced by many uncertain factors, like pipe size and length variations. Similarly, it is also difficult to determine the damping coefficient c, which depends on several coupled factors, like hydraulic viscous forces, mechanical viscous forces, side forces, bending forces and so on.

In the following, one approach is presented to exact ζ and ω_d (in turn, k and c can be easily determined based on values of ζ and ω_d) from the measurement. First, two arbitrary peak co-ordinates: (t_i, x_i) and (t_{i+n}, x_{i+n}) can be selected for calculating the damped period, t_d,

$$t_d = \frac{|t_{i+n} - t_i|}{n}, \tag{21}$$

where n is the number of peaks between peaks (x_i, x_{i+n}). Then, it is easy to calculate ω_d since $t_d = 2\pi\omega_d$, or

$$\omega_d = \frac{2\pi}{t_d}. \tag{22}$$

Following (15), by taking the amplitude of these two points (amplitude logarithmic decrements), we have,

$$\ln(\frac{x_i}{x_{i+n}}) = n\zeta\omega_n t_d. \tag{23}$$

Solving (21) and (22) , ζ, ω_d are obtained. Following (20), the phase angle is then calculated. It is clear that the selection of data points: (t_i, x_i) and (t_{i+n}, x_{i+n}), has a big impact on above parameters estimation. Hence, a numerical approach using nonlinear least square method is proposed below to calculate the best-fit parameter values. It is assumed that a model $\tilde{f}(t)$ given below represents the external force, where $\tilde{f}(t)$ is shown as

$$\tilde{f}(t) = F_0 e^{-\zeta\omega_n t}\cos(\omega_d t + \psi), \tag{24}$$

and F_0 is the initial force. For the underdamped case, the optimal cost function, J, subjected to constraints: $0 < \omega_d < \omega_n$ and $0 < \zeta < 1$, is formulated as

$$min_{\zeta,\omega_n} J(\zeta, \omega_n) = \sum_{i=1}^{N} |\tilde{f}(t_i) - f(t_i)|^2, \tag{25}$$

where $f(t_i)$ is the measurement force at time t_i and N is the number of measurement points in the data set. By solving this optimization problem, the optimal parameters ζ, ω_n are obtained that will describe the system dynamics by using (15). The results and discussions are given in the next section.

Results

Laboratory data

In this case, a laboratory-scale fully automated drilling rig (Løken et al. 2018; Khadisov et al. 2020) developed and equipped with a state-of-art sensors collection, is used as a case example to illustrate the data issues and demonstrate the proposed approaches for data quality improvement. Various drilling scenarios can be simulated on the rig, for instance, normal operations, overpull, string/bit washout, vibrations, etc., and the response of the system can be recorded by the data acquisition system. Such data carries valuable information; however, to retrieve it strong data analytics skills are required. Having such unique drilling rig allows us to conduct multiple experiments in a laboratory with minimum costs and creates possibilities to develop, test and validate the data analytics methods to identify and react to the common problems occurring during drilling. The detailed information about the rig structure, its software and control system was given in Løken et al. (2018); Khadisov et al. (2020); Løken et al. (2019, 2020). It is observed that use of PLC (programmable logic controller) type data acquisition systems can mitigate some of the discussed challenges, in laboratory scale rig, for instance, missing data and inconsistent sampling intervals. Therefore, results related to noise filtering, data assimilation and parameter estimations are shown and discussed below.

Data quality improvement

The sampling rates for the sensors were estimated in between 30 and 100 Hz. This was selected based on the available data storage capacity, memory, required controller reaction time, real-time computational capacity and data quality (preprocessing) required prior to decision making.

Results obtained from experimental tests using proposed approaches are analysed and discussed in this section. First, all logged data is re-sampled, the range is checked, and gaps are filled to get an even time spacing between samples and to remove invalid/inaccurate data. Then using the MAF, for the weight on bit (WOB) measurement, noise and outlier removal is examined with the different filter window sizes (Fig. 5). Results from Table 1 clearly show a reduction of σ (standard deviation) with the increasing filter window size. Selection of M also affects the time delay of the output signal, see Fig. 5. The larger M, the smoother the processed data curve. However, the larger M also leads to the time-delay issue. Figure 6 shows the case when $M = 8$. It is obvious that the processed data (in black) is delayed compared with raw data (in red), but most of noises are removed from the raw data. Hence, a trade-off between computational time and accuracy required has to be balanced when selecting M.

Next, the WOB data processed by the first-order LPF is analysed. The selection of τ is assessed using frequency analysis. Figure 7 and Table 2 represent the filtered results after the first-order LPF for the WOB data. It is clear that selection of τ has a clear impact on filtered results regarding amplitude and accuracy. The larger τ, the smoother the filtered data curve. However, the time delay of the filter is increased with increasing τ, see Fig. 7. Figure 8 shows the filtered data when $\tau = 1.1$, where the delay can be easily

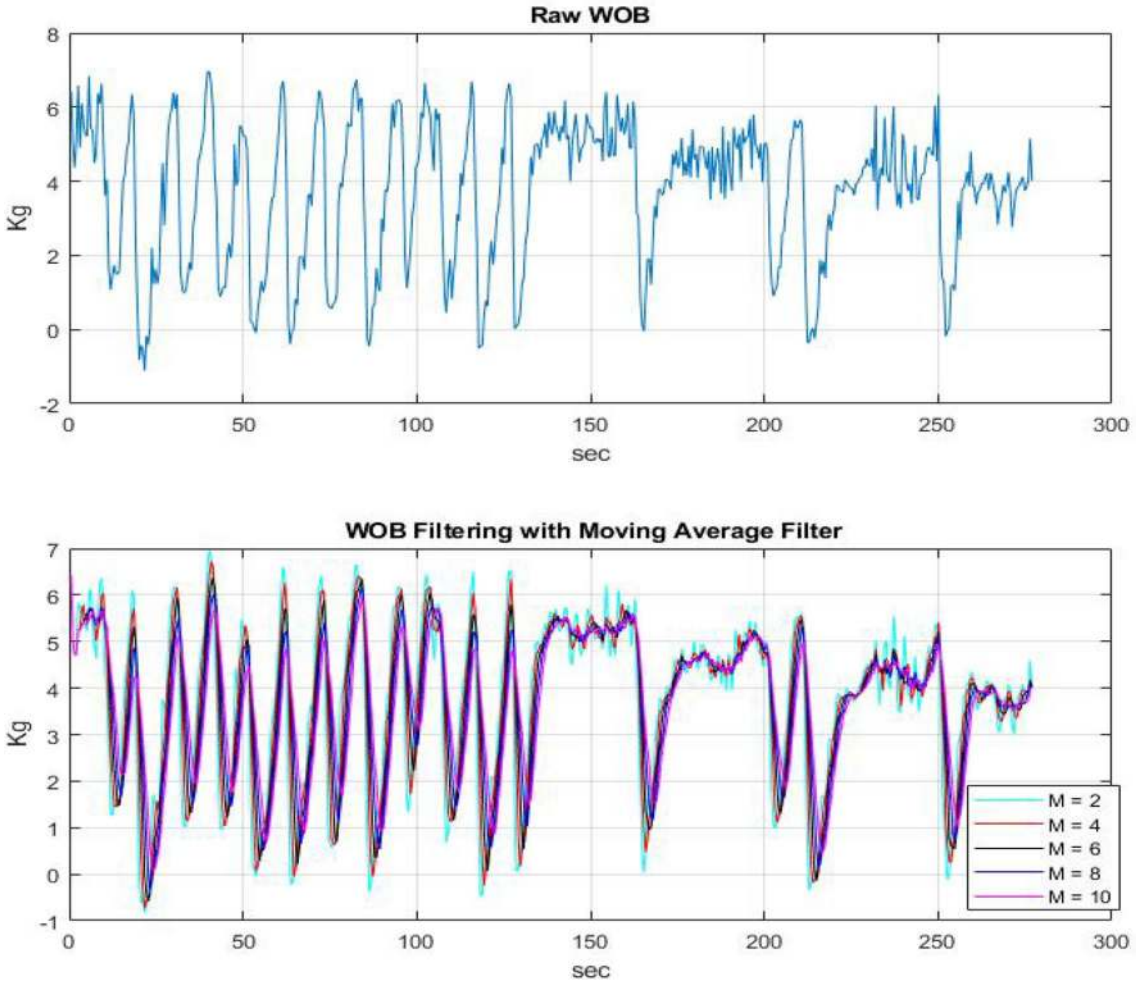


Fig. 5 WOB filtered data comparison with different window sizes

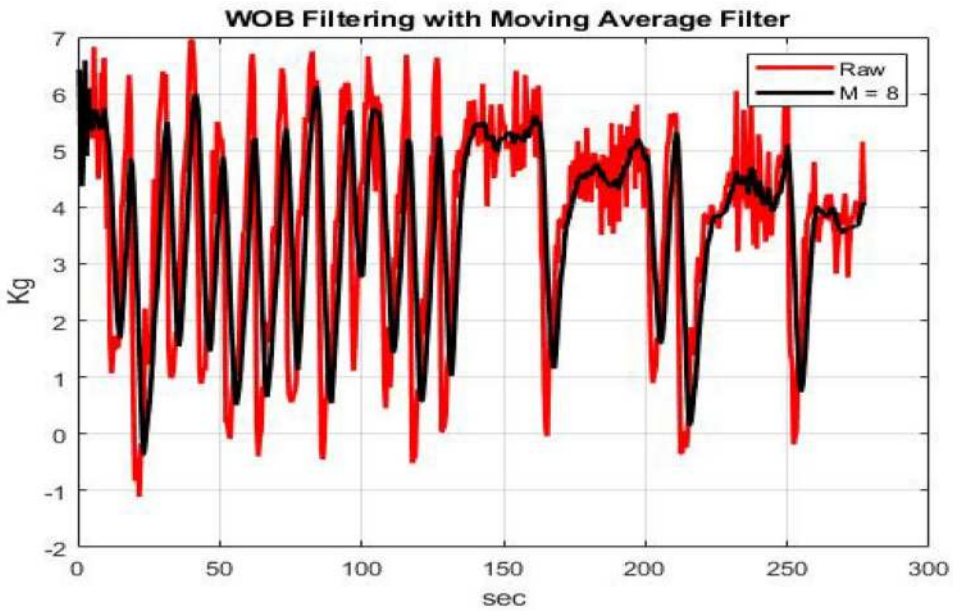


Fig. 6 WOB filtered data comparison with M=8

Table 1 WOB filtered statistical data comparison with MAF different window sizes

Parameter	Values	$M = 2$	$M = 4$	$M = 6$	$M = 8$	$M = 10$
μ	3.7083	3.7082	3.7091	3.7111	3.7133	3.7159
σ	1.8779	1.8188	1.7120	1.5925	1.4669	1.3407

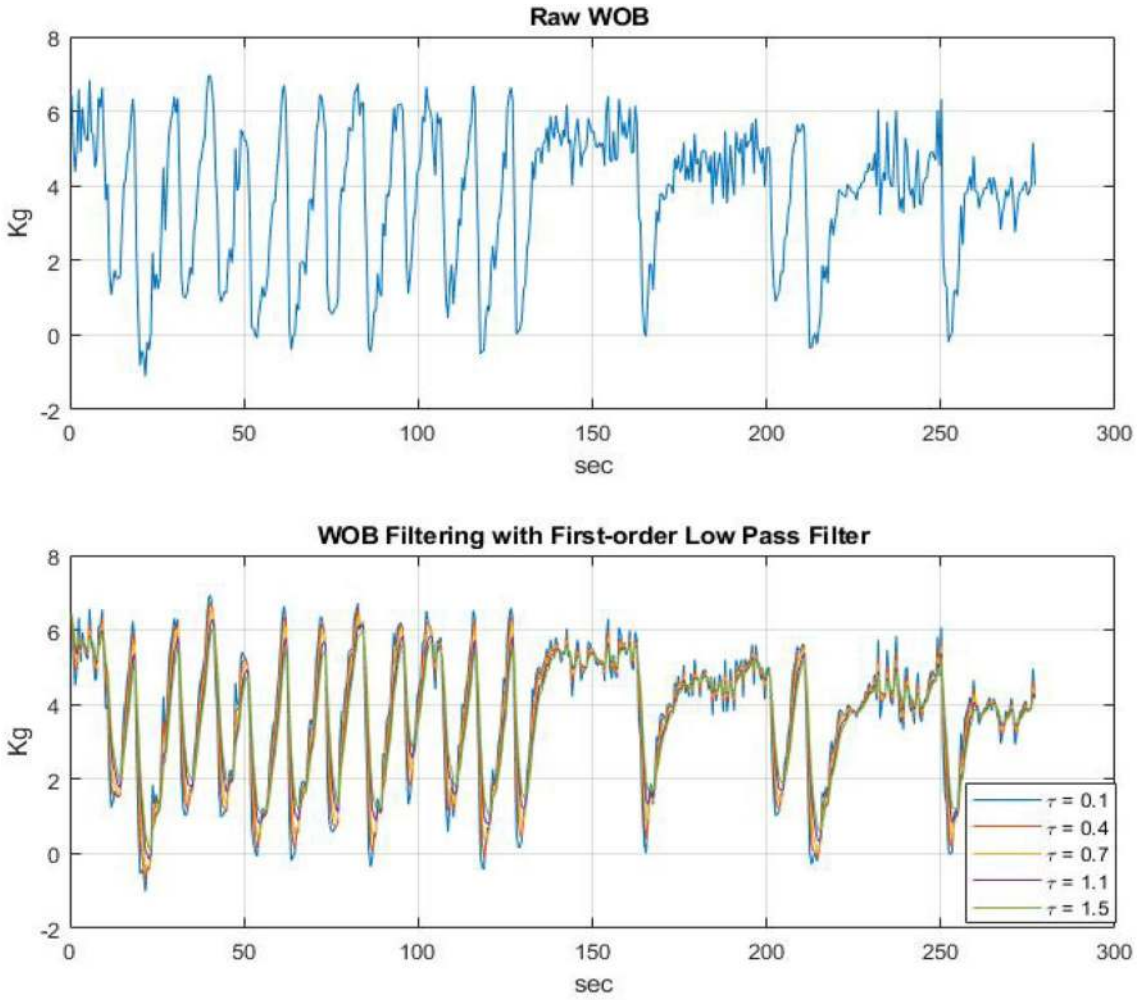


Fig. 7 WOB filtered data comparison with different τ

observed. Hence, similarly to the MAF, there is a trade-off between time delay and accuracy of the data for the first-order LPF application.

Then, the second-order LPF is considered. The advantage of selecting the second-order LPF over the first-order LPF is that it provides two controlling parameters (α, β) for users to balance the trade-off between delay and noise-removal. Therefore, not only time delay of the filtered signal but its amplitude can be adjusted using these parameters. Figure 9 and Table 3 illustrate the results of varying β with the constant $\alpha = 2$. Normally the larger β, the more noises are filtered. Compared with the first-order LPF, the delay becomes better, see Fig. 10. Figure 11 and Table 4 show the results of varying α with the constant $\beta = 0.4$. The larger α, the larger amplitude of filtered data. From 9, it shows that the amplitude of filtered data is larger than the one of raw data when $\alpha = 1, 1.5$. Figure 12 shows the filtered data with $\alpha = 2$, where the amplitude of the filtered data can be kept close to the raw data.

Data assimilation of two torque sensor readings is considered using two data sets with nearly same average. Assimilated data points' deviation from two data sets is given under Fig. 13 and Table 5. Results confirm that final estimate has better variance than the two measurements and it is more depended upon the measurements with least variances.

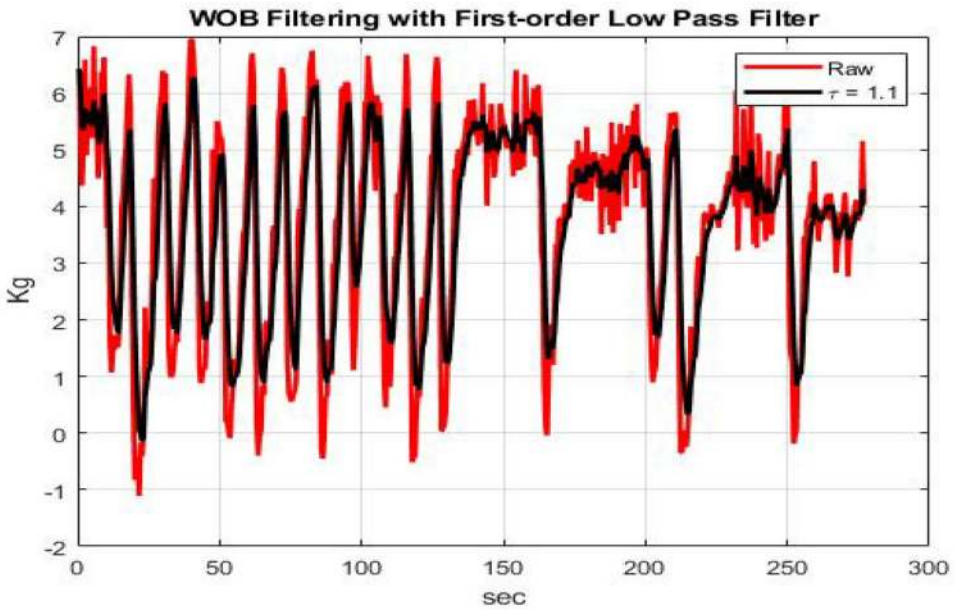


Fig. 8 WOB filtered data comparison with $\tau = 1.1$

Table 2 WOB filtered statistical data comparison from the first-order LPF with different τ

Parameter	Values	$\tau = 0.1$	$\tau = 0.4$	$\tau = 0.7$	$\tau = 1.1$	$\tau = 1.5$
μ	3.7083	3.7091	3.7114	3.7138	3.7171	3.7206
σ	1.8779	1.8379	1.7344	1.6360	1.5142	1.4082

Model identifications

In the small-scale rig, axial and transverse vibrations were dominant compared to the torsional oscillations. We concluded this was due to the short length of the drill pipe and/or its eccentricity from exact vertical axis. Moreover, our rotational system uses a brushless commercial motor with a robust RPM controller. Therefore, torsional vibrations were not observed frequently.

Model parameters that are calculated from the axial vibration model given in "Information extraction" section based on load cells measurement are summarized in Table 6. Frequency analysis of WOBs data validates the estimated value of ω_n. The natural frequency of WOBs is around 30.6 Hz.

By observing simulation results from Fig. 14, it is clear that a linear input of $f(t)$ will result in a linear output behaviour of bit position. The presence of an initial/ final velocity (a nonzero force at start or end of operations) will trigger an under-damped transient response. This is because system response to an initial velocity is same as its response to first impulse of an impulse series, although no other initial conditions are present. Figure 6 illustrates bit position behaviour during a bit bouncing event or under heave in offshore drilling. Similar bit position and off-bottom WOBs measurement behaviour are observed in Fig. 15.

For field data, there exist more challenges than the laboratory-scale rig data management challenges. For example, time-delay, sensor malfunctions, user entry errors, no communications and corruptions. Nonetheless, it is clear that some data quality challenges are observed independent of the scale of operations. Solving such data quality challenges can be studied and experimented in laboratory-scale cost effectively.

Field data

Volve field data, published by Equinor in 2018, is a valuable source of real drilling data, making it possible to evaluate methods derived from laboratory data. The available logs are coming from the field that was operational in the North Sea from 2008 until 2016. The published dataset contains seismic data, production logs, drilling daily reports, reservoir models, geophysical interpretations, real-time drilling data and more. In this study, drilling logs converted from original WITSML (wellsite information transfer standard markup language) data to CSV (comma-separated values) files were used, a process described in detail in Tunkiel et al. (2020). As a case study well F5 was used, drilled using Schlumberger PowerDrive RSS tool. Inspecting the available inclination data, a number of issues have been identified that can be solved or at least mitigated using methods described in this paper. Inclination data is plotted in Fig. 16.

Outliers are clearly seen in the inclination data recorded both by the MWD and the PowerDrive tools. In case of

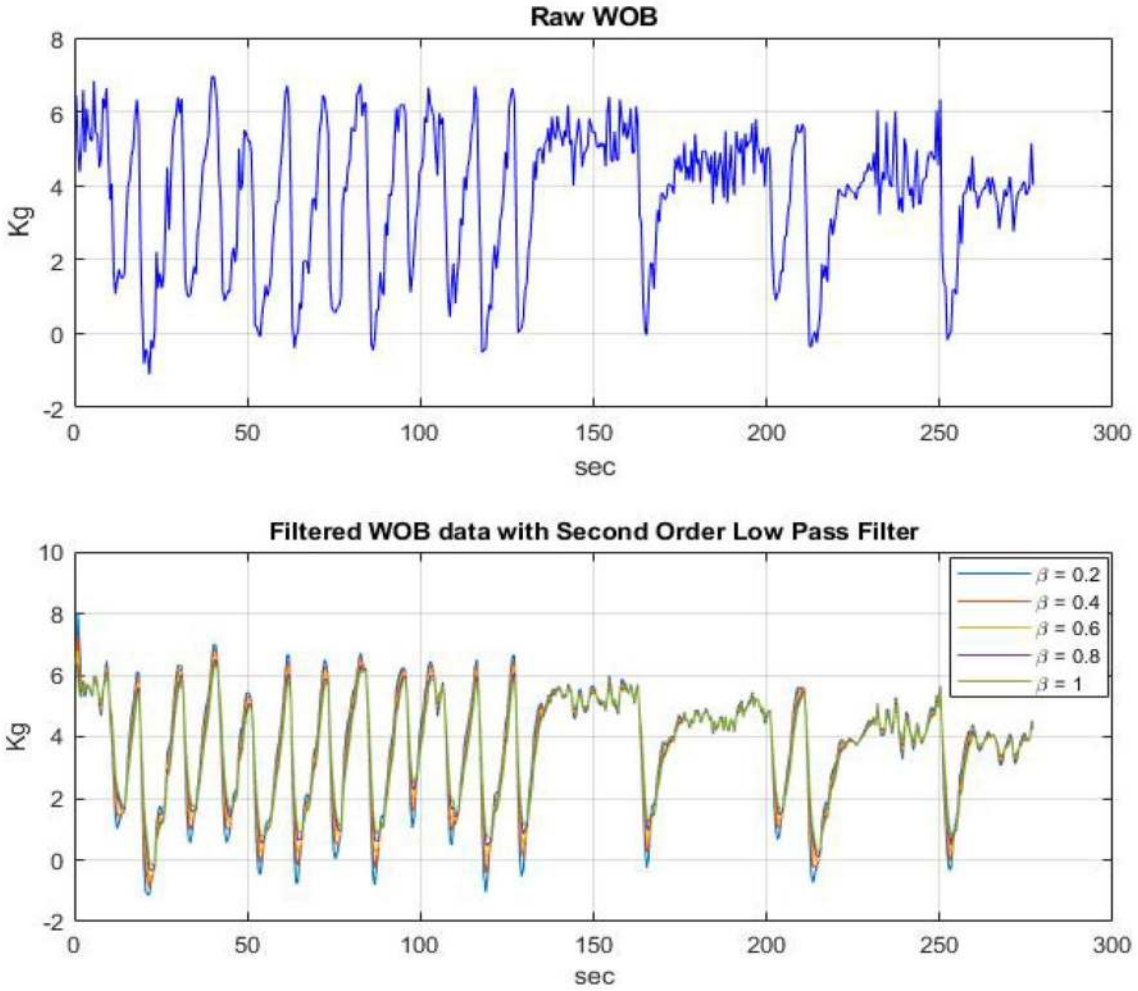


Fig. 9 WOB filtered data comparison with different β

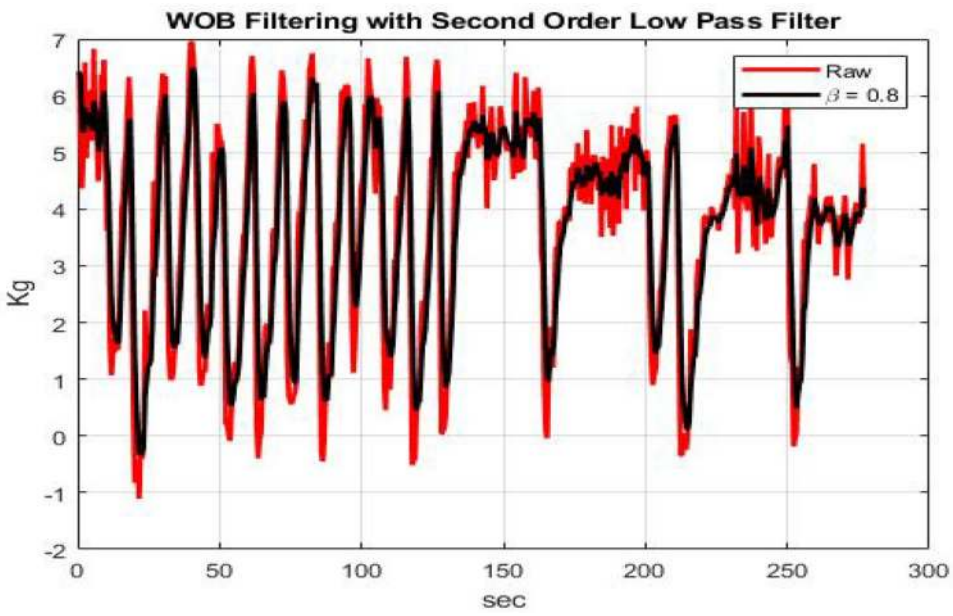


Fig. 10 WOB filtered data comparison with $\beta = 0.8$

Table 3 WOB filtered statistical data comparison from the second-order LPF with different β

Parameter	Values	$\beta = 0.2$	$\beta = 0.4$	$\beta = 0.6$	$\beta = 0.8$	$\beta = 1$
μ	3.7083	3.7187	3.7179	3.7172	3.7165	3.7159
σ	1.8779	1.8989	1.7926	1.6994	1.6161	1.5414

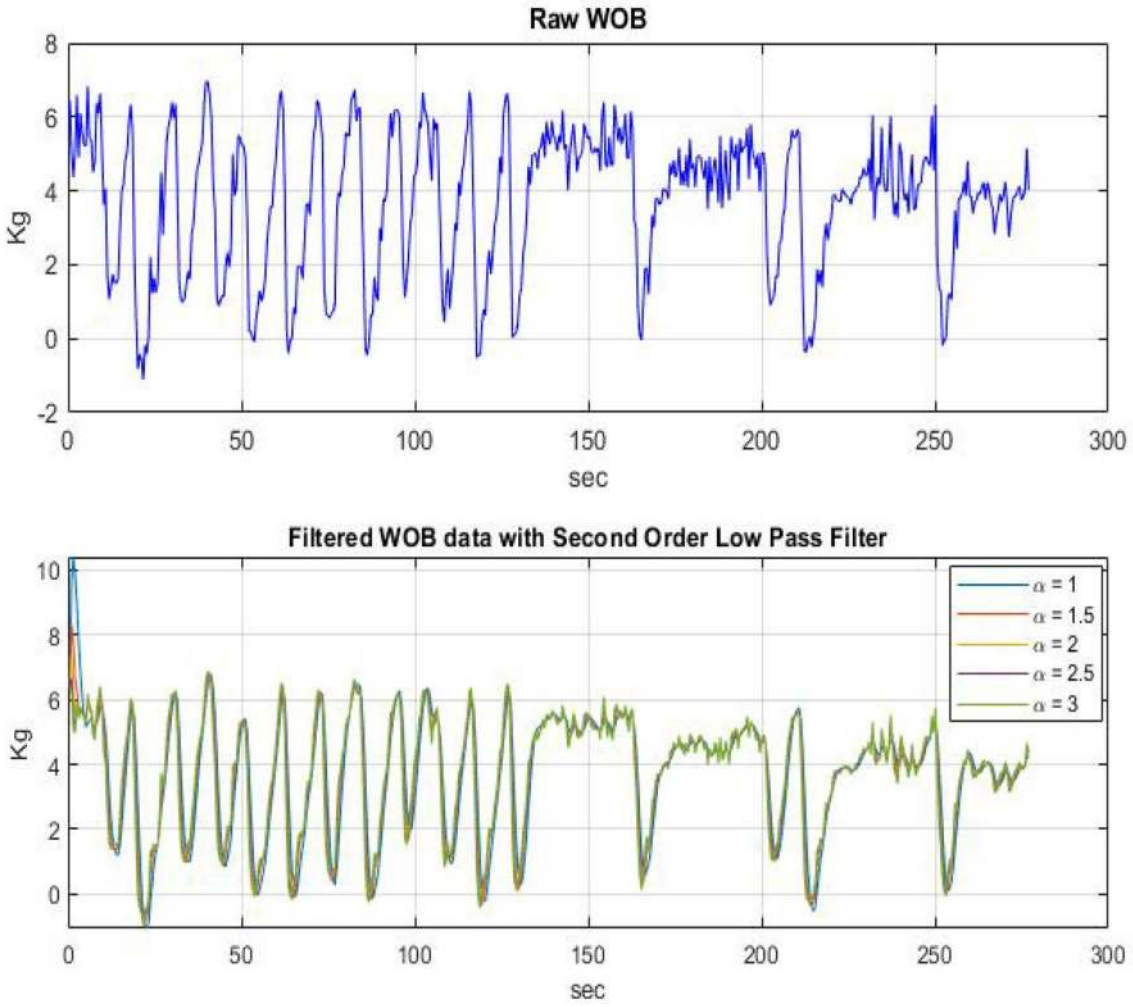


Fig. 11 WOB filtered data comparison with different α

the MWD tools, there are multiple individual points outside of the continuous trend. This is likely due to data transmission errors of data storage corruption. Different issues are connected to the readings from PowerDrive. At three distinct depths, there are multiple readings of inclination between zero and the correct value. This may happen when the bottom hole assembly was tripped in or out while recording inclination data. While this in itself is not an issue, the log was improperly merged at some stage assigning all the readings made through tripping to one depth value. Alternative explanation is that excessive rotation or vibration negatively affected the sensor responsible for inclination reading, resulting in incorrect data being recorded. No matter the root cause, the resultant log quality needs improvement.

Real-time logs often contain gaps in data. These gaps can be divided into four distinct categories as shown in Fig. 17. These are based on the quantity of continuous gaps

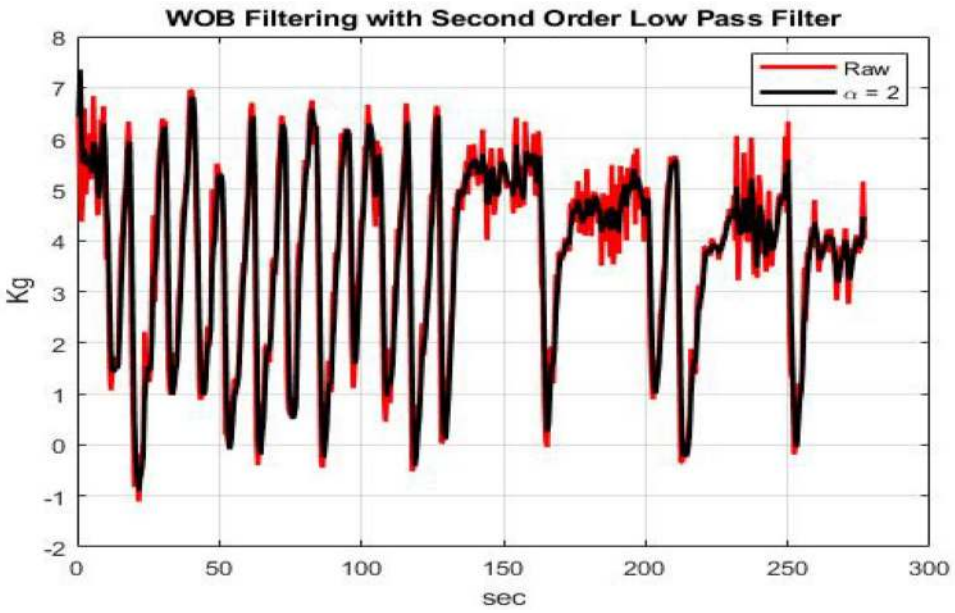


Fig. 12 WOB filtered data comparison with $\alpha = 2$

Table 4 WOB filtered statistical data comparison from second-order LPF with different α

Parameter	Values	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$	$\alpha = 3$
μ	3.7083	3.7496	3.7260	3.7179	3.7143	3.7123
σ	1.8779	1.8586	1.8038	1.7926	1.7929	1.7967

Table 5 Assimilated and raw data properties for sensor fusion

Parameter	T_1	T_2	T_e
μ	2.0777	2.1295	2.1036
σ	0.7260	0.7248	0.7235

(HQ—high quantity, LQ—low quantity), and percentage of dataset occupied (HP—high percentage, LP—low percentage). Note that this proposed method of classifying gaps is related to continuous data only, such as drilling logs. Non-series type of data, such as customer database, car fleet database, cannot be classified this way.

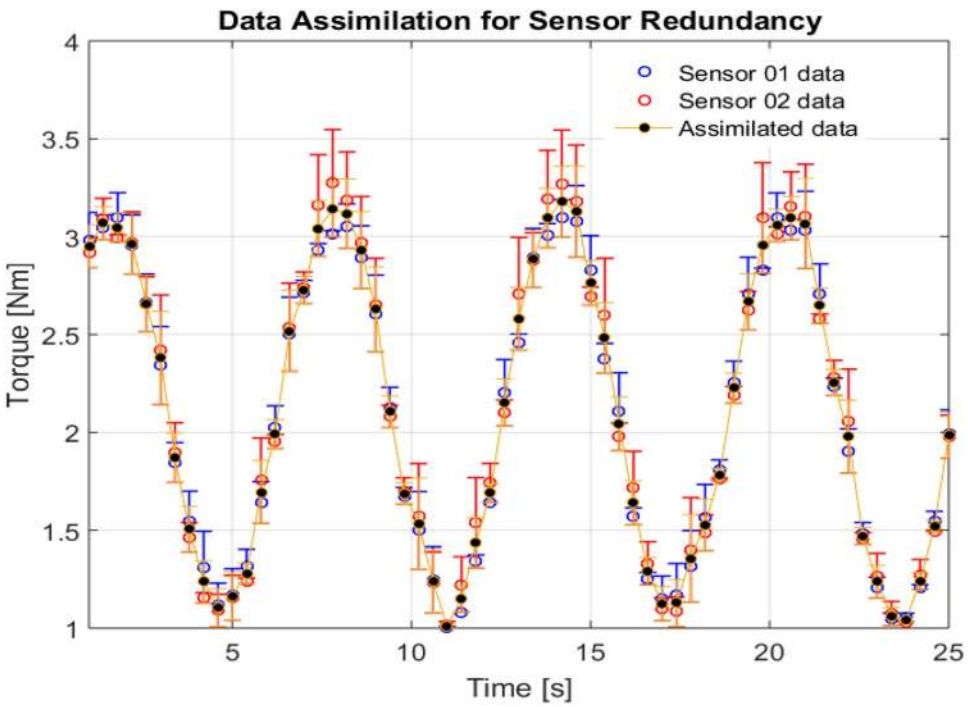


Fig. 13 Data assimilation for torque measurements

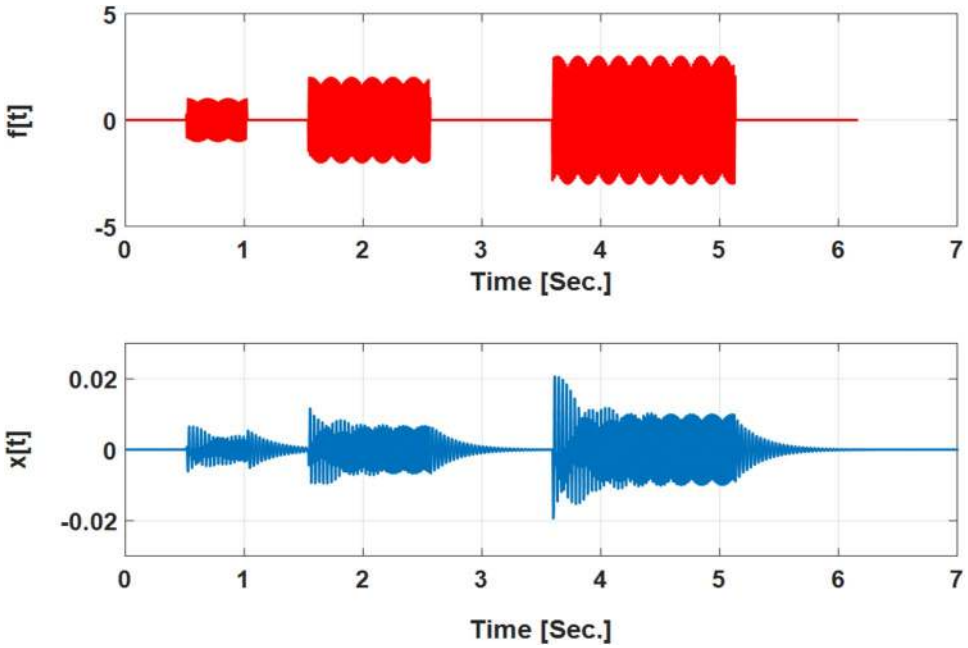


Fig. 14 Periodic WOB at natural frequency and bit position response

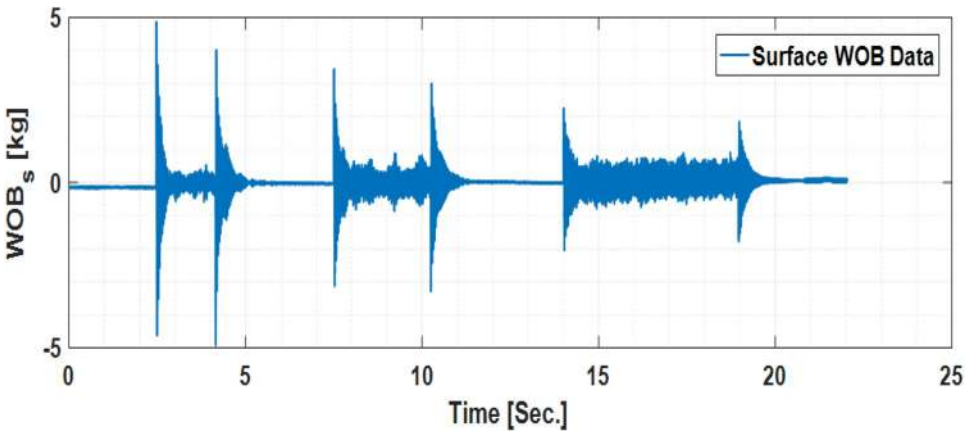


Fig. 15 Off-bottom WOBs measurement from small-scale rig under three different actuator speeds

Table 6 Estimated parameters

Parameter	μ	σ
ω_d	192.663	12.872
ζ	0.0213	0.008
ω_n	192.713	12.866
k/m	943.331	133.375
c/m	8.150	3.102

HQLP—high quantity of gaps that occupy relatively low percentage of the data are very common in real-time drilling logs. The investigated logs had issues at a much smaller scale, with multiple missing values spanning from one to few dozens of rows. Various sensors report data at different times and different frequencies; values transmitted through mud pulse telemetry are particularly susceptible to this issue, as a complete cycle of uplinking data may take over few minutes. There are at least two different approaches to filling these small-scale gaps. The basic method is to forward fill values forward whenever a missing value cell is encountered. This is consistent with the logic, that if a new reading

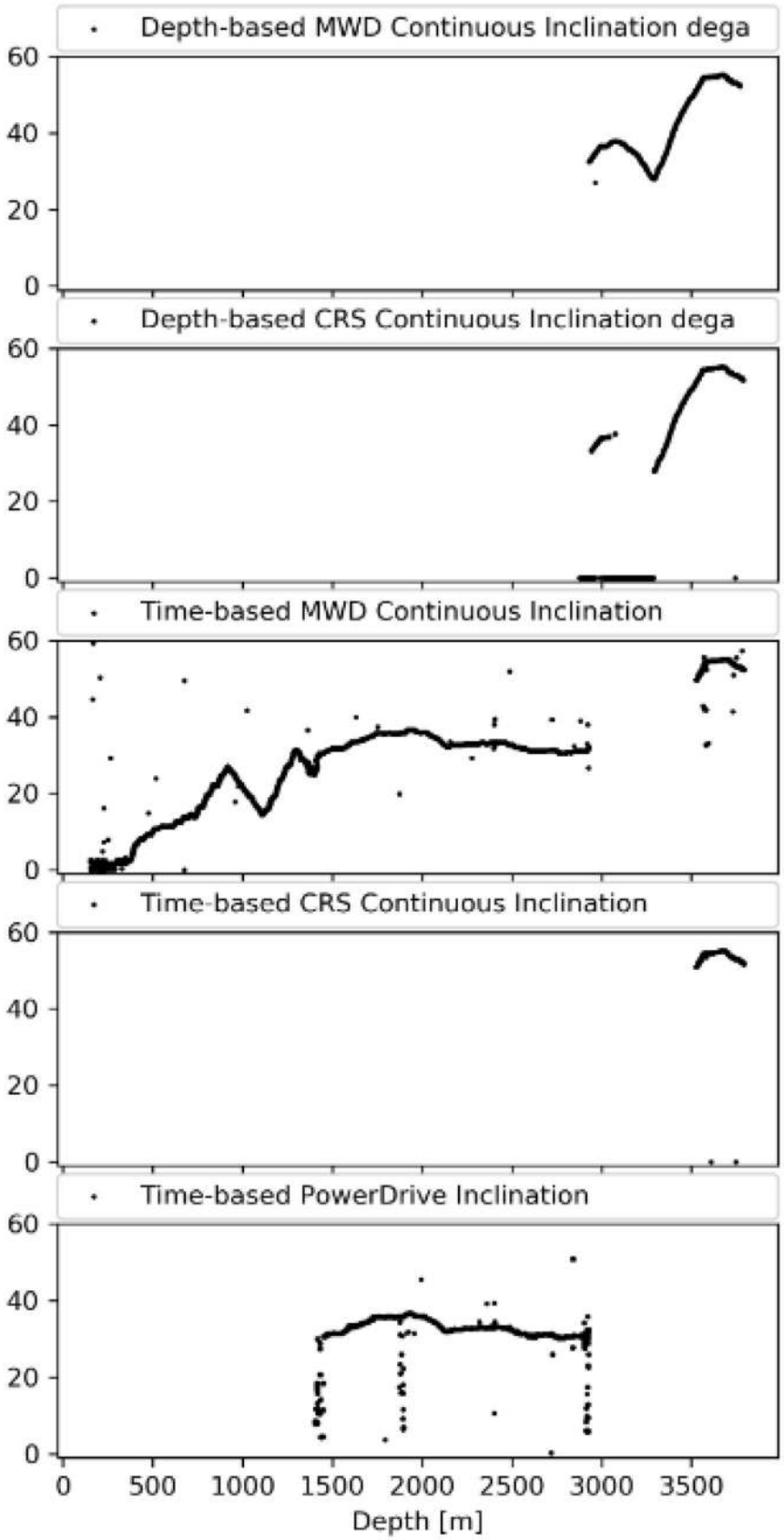


Fig. 16 Inclination data with gaps

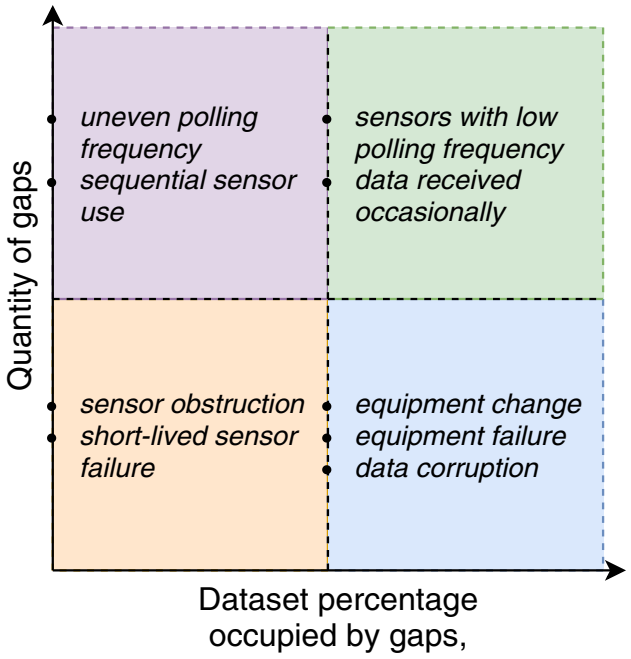


Fig. 17 Missing data category

is not available, the old value is considered still valid, see Fig. 18 as an example for forward filled data.

Alternative method is to perform a linear interpolation using the last available value before the gap and the first available value. This method may increase the apparent polling frequency and is not suitable for discrete data. Additional drawback is that such approach cannot be applied to real-time data; it is possible only after a given gap is "closed" with a new, correct value, leading to delay in data, see Fig. 19 as an example for interpolated data.

LQHP gaps occupy a significant portion of the dataset with the gaps being long and continuous. A good example of such gaps is data in Fig. 16, where significant percentage of different log is missing. This is typically caused by equipment change, sensor failure or data corruption. Filling such gaps requires bespoke solutions that will differ from log to log. It may be possible, that a certain reading is duplicated by a different equipment—for example, where MWD provider changed mid-well, the same data will exist as different attributes. Data can be restored using machine learning methods, given that correlations exist between the missing

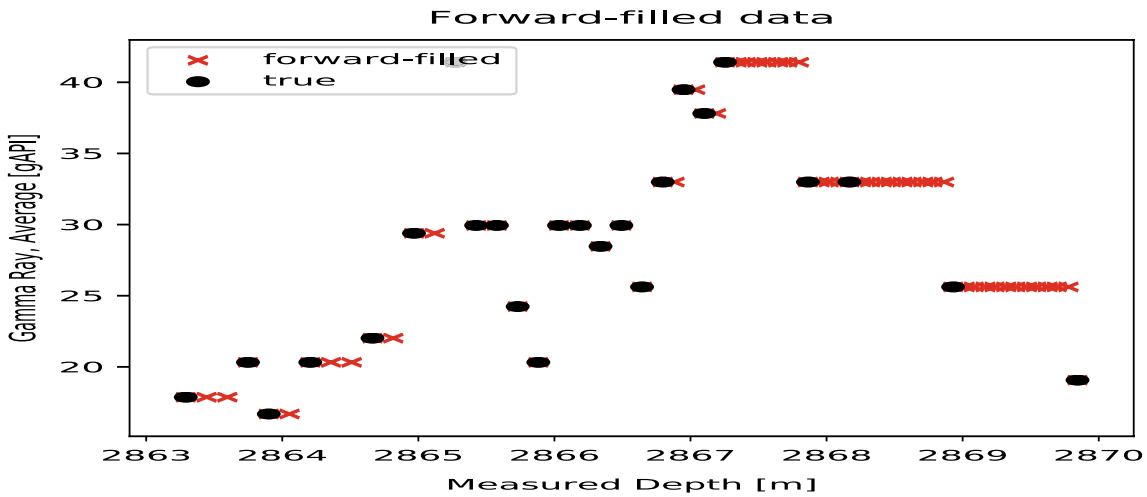
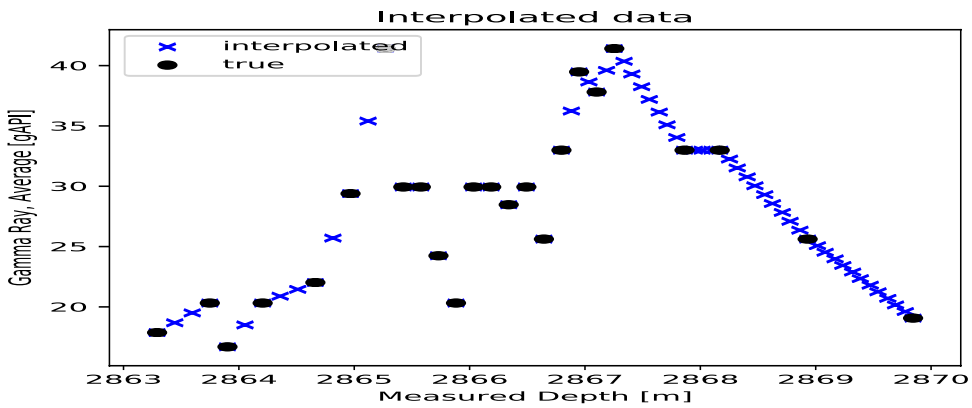


Fig. 18 Forward fill values

Fig. 19 Interpolated data



and the remaining parameters. Additionally case-specific solutions may be possible, such as when missing inclination data from drilling operation can be filled with data recorded while tripping, recording readings from the given section of a well. Often however, LQHP gaps cannot be filled.

HQHP gaps can be identified, when a certain parameter is logged very rarely compared to other parameters. This may be by design, when a parameter is of limited interest, and/or containing data of significant inertia. Interpolation is a good candidate of gap filling technique for this category. LQLP gaps are typically easiest to fill with machine learning methods. Small, sparse gaps suggest intermittent continuous short-lived sensor failures, or sensor obstruction, as it may be the case in motion-capture technology. Having most of the dataset for training is likely to produce a robust model. Methods typical for LQHP gaps can be used here as well. As a last resort, the data can simply be abandoned if the percentage of dataset lost is small, and the location in the log is of little interest.

Conclusion

This paper proposes a systematic approach to improve drilling operational data reliability and consistency while preserving data accuracy and validity. It also includes a summary of several drilling data quality challenges and methods to improve such quality issues. The data quality issues have been identified, improvement approaches have been investigated, and results have been then analysed to verify the enhancement of data quality.

Although one case study that utilizes laboratory data may not directly reflect the data quality situation of a standard rig operating in the field (due to the involvement of different service companies and additional quality issues, which are related to data transmission and handling data via a sequence of different data systems from source to consumer), observations are made to several semantic data quality challenges. In addition, several hidden data management challenges are emphasised. Therefore, laboratory-scale drilling and data management can be considered as a useful tool to identify drilling data challenges to speed-up drilling digitalization.

Funding No particular funding for this project.

References

Aibar AH, Alotaibi BM, Asfoor HM, Nefai MS, Aramco S (2018) A journey towards building real-time big data analytics environment for drilling operations: challenges and lessons learned. SPE Kingdom Saudi Arabia Annual Tech Symp Exhibit. https://doi.org/10.2118/192285-MS

Bello O, Yang D, Lazarus S, Wang XS, Denney T (2017) Next generation downhole big data platform for dynamic data-driven well and reservoir management. SPE Reserv Charact Simul Conf Exhibit. https://doi.org/10.2118/186033-MS

David RM (2016) Approach towards establishing unified petroleum data analytics environment to enable data driven operations decisions. SPE Annual Tech Conf Exhibit. https://doi.org/10.2118/181388-MS

Dickson D (2014) Removing risk and cost from remote operations through intelligent practices. SPE Intell Energy Conf Exhibit. https://doi.org/10.2118/167852-MS

Donnelly J, Wilson A, Feder J, Jacobs T (2020) Focus on sustainability, digital transformation at 2019 atce. J Petrol Technol. https://doi.org/10.2118/1119-0040-JPT

Dursun S, Duman K, Tuna TM, Ding J (2014) A workflow for intelligent data-driven analytics software development in oil and gas industry. SPE Annual Technical Conf Exhibit Amsterdam. https://doi.org/10.2118/170859-MS

Hegde C, Gray KE (2017) Use of machine learning and data analytics to increase drilling efficiency for nearby wells. J Nat Gas Sci Eng 40:327–335. https://doi.org/10.1016/j.jngse.2017.02.019

Jiawei Y, Susanto R (2019) Outlier detection: how to threshold outlier scores? 12: 1–6

Khadisov MA, Hagen HP, Jakobsen AS, Sui D (2020) Developments and experimental tests on a laboratory-scale drilling automation system. J Petrol Explor Prod Technol 10:605–621. https://doi.org/10.1007/s13202-019-00767-6

Lewis JM, Lakshmivarahan S, Dhall S (2006) Dynamic data assimilation a least squares approach. Cambridge University Press, Cambridge

Løken EA, Løkkevik J, Sui D (2020) Data-driven approaches tests on a laboratory drilling system. J Petrol Explor Prod Technol. https://doi.org/10.1007/s13202-019-00767-6

Løken E, Geekiyanage S, Sui D (2019) Small-scale autonomous drilling development for drilling digitalization. Oil Gas European Magazine

Løken E, Trulsen A, Holsaeter AM, Wiktorski E, Sui D, Ewald R (2018) IFAC-OOGP

Lu P, Liu H, Serratella C, Wang X (2017) Assessment of data-driven machine learning techniques for machinery prognostics of offshore assets. Offshore Technol Conf. https://doi.org/10.4043/27577-MS

Mathis W, Thonhauser G (2007) Mastering real-time data quality control–how to measure and manage the quality of (rig) sensor data. SPE/IADC Middle East Drilling Technol Conf. https://doi.org/10.2118/107567-MS

Mathis W, Thonhauser G, Wallnoefer G, Ettl J (2006) Use of real-time rig sensor data to improve daily drilling reporting, benchmarking and planning–a case study. Intell Energy Conf Exhibit. https://doi.org/10.2118/99880-MS

Nybo R, Sui D (2014) Closing the integration gap for the next generation of drilling decision support systems. Soc Petrol Eng SPE Intell Energy Int. https://doi.org/10.2118/167864-MS

Nybo R, Froyen J, Lauvsnes AD, Korsvold T, Herbert MC, Choate M (2012) The overlooked drilling hazard: decision making from bad data. SPE Intell Energy Int. https://doi.org/10.2118/150306-MS

Ouyang L, Kikani J (2002) Improving permanent downhole gauge (pdg) data processing via wavelet analysis. Euro Petrol Conf. https://doi.org/10.2118/78290-MS

Rassenfoss S (2020) Succeeding at petroleum engineering in a digital age. J Petrol Technol. https://doi.org/10.2118/0320-0034-JPT

Saptawati GAP, Nata GNM (2015) Knowledge discovery on drilling data to predict potential gold deposit. Int Conf Data Softw Eng 71:143–147. https://doi.org/10.1109/ICODSE.2015.7436987

Saputelli L (2020) Technology focus: data analytics. J Petrol Technol. https://doi.org/10.2118/1019-0061-JPT

Smith SW (1997) The scientist engineer's guide to digital signal processing. California Technical Publishing, San Diego

Stanley GM, Mah RSH (1981) Observability and redundancy in process data estimation. Chem Eng Sci 36:259–272. https://doi.org/10.1016/0009-2509(81)85004-X

Sui D, Sukhoboka O, Aadnoy BS (2018) Improvement of wired drill pipe data quality via data validation and reconciliation, vol 15. Cambridge University Press, Cambridge, pp 625–636. https://doi.org/10.1007/s11633-017-1068-9

Temer E, Pehl HJ (2017) Moving toward smart monitoring and predictive maintenance of downhole tools using the industrial internet of things iiot. Abu Dhabi Int Petrol Exhibit Conf. https://doi.org/10.2118/188382-MS

Thomson W (1996) Theory of vibration with applications. CRC Press, Boca Raton

Thonhauser G (2004) Using real-time data for automated drilling performance analysis. Oil Gas Euro Magaz 120:170–173

Thonhauser G (2018) Guest editorial: digital drilling disruption–understand downhole, gain control. J Petrol Technol 70(11):14–15

Tunkiel AT, Wiktorski T, Sui D (2020) Drilling dataset exploration, processing and interpretation using volve field data. Proceed Int Conf Offshore Mech Arctic Eng