Journal of Petroleum Science and Engineering 196 (2021) 108069

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: http://www.elsevier.com/locate/petrol

Reference dataset for rate of penetration benchmarking

Andrzej T. Tunkiel [a,*], Dan Sui [a], Tomasz Wiktorski [b]

[a] Department of Energy and Petroleum Engineering, Faculty of Science and Technology, University of Stavanger, 4036 Stavanger, Postboks, 8600, Forus, Norway
[b] Department of Electrical Engineering and Computer Science, Faculty of Science and Technology, University of Stavanger, 4036 Stavanger, Postboks, 8600, Forus, Norway

ARTICLE INFO

Keywords:
Rate of Penetration
Machine learning
Benchmarking
Volve Dataset

ABSTRACT

In recent years, there were multiple papers published related to rate of penetration prediction using machine learning vastly outperforming analytical methods. There are models proposed reportedly achieving R2 values as high as 0.996. Unfortunately, it is most often impossible to independently verify these claims as the input data is rarely accessible to others. To solve this problem, this paper presents a database derived from Equinor's public Volve dataset that will serve as a benchmark for rate of penetration prediction methods. By providing a partially processed dataset with unambiguous testing scenarios, scientists can perform machine learning research on a level playing field. This in turn will both discourage publication of methods tested in a substandard manner as well as promote exploration of truly superior solutions. A set of seven wells with nearly 200–000 samples and twelve common attributes is proposed together with reference results from common machine learning algorithms. Data and relevant source code are published on the pages of University of Stavanger and GitHub.

1. Introduction

Research review related to machine learning (ML) models within petroleum is problematic. To fully evaluate a proposed method both data and exact description of the method are necessary, which is regrettably rare in the field. Data is often withheld on the grounds of confidentiality and there is little pressure to release source code behind presented methods. It leaves the scientific discourse susceptible to errors or cheating (Davey Smith and Ebrahim, 2002), where results might be artificially inflated. Most researchers consider that there is currently a reproducibility crisis in science (Baker and Penny, 2016), with 70% of polled scientists admit to trying and failing to reproduce experiments.

Introduction of a standardized real-time well log dataset has a capacity to transform research in data-driven methods related to drilling. It has high potential to spark a healthy competition between researchers striving for better performance. Well known datasets such as MNIST promoted competition, and facilitated research and knowledge sharing in the field of handwriting recognition. Additionally, having source data available makes paper authors accountable, as any dishonest practices will be easily discoverable when others attempt to reproduce the results.

Building upon Volve dataset (Equinor, 2018), made public by Equinor in 2018 on a very permissive license,[1] and previous data preparation work (Tunkiel et al., 2020), this paper proposes a standardized dataset of seven wells containing twelve commonly logged attributes for ROP prediction purposes. In addition to data itself, three benchmarking scenarios are proposed, with specific metrics attached to them, ensuring that future results are comparable with each other. To establish a point of reference results from a number of basic algorithms are presented, together with complete source code[2] needed for result replication as well as a starting point for other researchers. We hope that it will enable reproducibility, competition, and cooperation between the researchers elevating the quality of papers published in the field. It also has potential of lowering the entry point for data-driven methods in drilling as well as attracting talent from outside of petroleum field.

2. Existing problems and potential pitfalls

2.1. Rationale behind rate of penetration prediction and methodology

2.1.1. Purpose and goal

Drilling of oil wells is very expensive. IHS reports day rates of semisubmersible, and jack-up rigs, as well as drillships (IHS Markit,

* Corresponding author.www.ux.uis.no/~atunkiel/
E-mail address: andrzej.t.tunkiel@uis.no (A.T. Tunkiel).
[1] Creative Commons BY-NC-SA 4.0, https://creativecommons.org/.
[2] https://github.com/AndrzejTunkiel/USROP.

https://doi.org/10.1016/j.petrol.2020.108069
Received 24 June 2020; Received in revised form 10 September 2020; Accepted 20 October 2020
Available online 24 October 2020

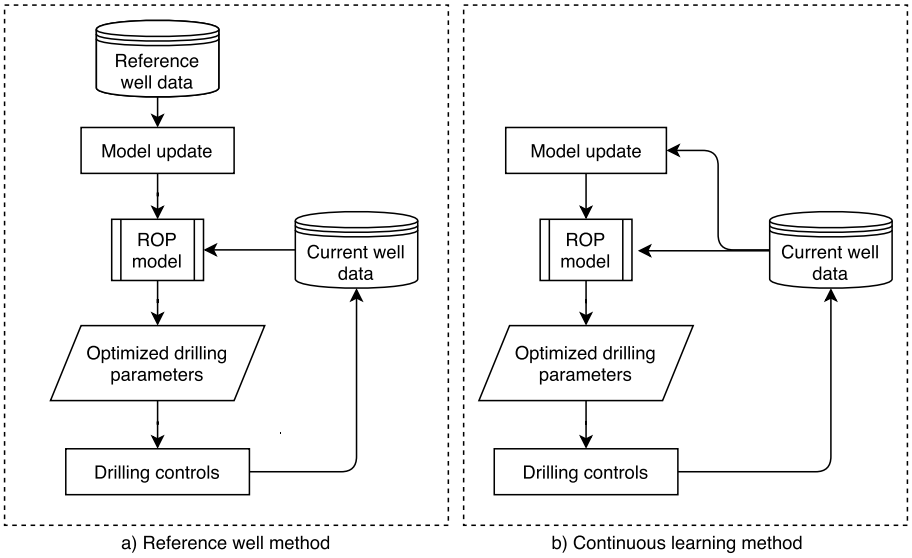


Fig. 1. ROP optimization method.

0000) from mid 2017 up to present day. At the time of this writing[3] average rates vary between 40,000 and 300,000 USD per day of drilling, therefore increasing the ROP translates directly to savings for the operator.

Additional area of potential savings is optimization of the mechanical specific energy (MSE) to ROP ratio required to drill. While the energy itself is not a significant portion of the drilling cost, its excess has to be released as effects other than cutting of the rock, as vibrations and heat, leading to higher tool wear and increased likelyhood of failure. This in turn can extend the drilling time due to additional tripping operations to exchange equipment, or fishing operations to retrieve equipment lost in hole.

2.1.2. Assumptions

ROP prediction model is necessary for ROP optimization. This historically is done both though analytical as well as data-driven models. Both approaches require reference data, to either find model constants (Rahimzadeh et al., 2011) or to train the machine learning model. The closer the reference drilling is to future drilling, the smaller expected error between reality and the model should become. This also means, that reported accuracy of a model is only applicable to drilling at the same level of similarity, in terms of equipment, lithology, procedures, depth, etc. Reference, or training, data, can be taken from neighbouring wells, or from the currently drilled well in continuous learning scenario (Liu, 2017), where a model is created *on the go*.

2.1.3. ROP optimization process

ROP optimization can be performed in multiple ways. Fig. 1 presents a basic outline of two general approaches to such an optimization process. ROP model is used to determine optimal drilling parameters, related to a chosen optimization problem, be it minimizing drilling time, minimizing MSE used to drill or others. This in turn is to set typical drilling inputs, such as drill bit RPM or weight on bit. The difference between a reference well and continuous learning method is indicated in subfigures a) and b). A reference well can be used, which is similar to the well drilled, a). This allows for model development to happen offline and is done once, and is in general simpler to deploy; a reference well is however required, which may or may not be available. Alternatively, the model can be created *on the go* in a continuous learning fashion, as shown in the subfigure b). This requires model development to happen while drilling, which is more difficult to implement due to computing equipment and skills necessary. Additional drawback is that the initial dataset is very small, and empty at the very start, which is detrimental to machine learning training process. Temporary model can be used, or a *warm-up* period exists where there is no model present. This is countered by the fact that the data is much more closely related to drilling at hand, improving the performance. In both approaches the ROP model is used to calculate the optimized drilling parameters that are then applied to the drilling controls related to parameters such as weight on bit, drill string RPM, mud flow etc. Method choice would depend on data availability, and other local constraints. Hybrid methods are also possible, and best practices are yet to be established.

2.2. Data availability

Initiating this research was a review of a number of recent papers (Ahmed et al., 2019a, 2019b; Hegde and Gray, 2017, 2018; Hegde et al., 2015, 2017; Soares and Gray, 2019; Sabah et al., 2019; Han et al., 2019; Shi et al., 2016; Mantha and Samuel, 2016; Eren and Ozbayoglu, 2010, 2011; Soares et al., 2016; Amar and Ibrahim, 2012; Bourgoyne and Young, 1999; Yi et al., 2014; Jiang and Samuel, 2016) that aimed at rate of penetration (ROP) prediction. Nearly none provided the data; we were able to identify few exceptions - two papers (Amar and Ibrahim, 2012; Eren and Ozbayoglu, 2011) used data originally published in 1974, of which 30 samples (Bourgoyne and Young, 1999) were directly quoted in the reference paper. Another identified paper provided 25 samples (Yi et al., 2014) that were subsequently re-used by others (Jiang and Samuel, 2016). This means that the existing publications are either not reproducible or use a very small size dataset that does not meet the standards of modern machine learning research. For comparison, a popular open MNIST (Deng, 2012) dataset, a Modified National Institute of Standards and Technology database of handwritten digits contains 70,000 samples, which was later extended (Cohen et al., 2017) to 300,000 samples. These datasets contributed to over 20,000 publications over 20 years since its inception.

2.3. Source code availability

Source code availability among reviewed papers was even poorer - none of the reviewed papers shared it. While mathematical description of the basic underlying method was common, it is rarely sufficient to reproduce results. Simple Multi-Layer Perceptron (MLP) in a popular Keras implementation has to be described at least by number of layers

[3] August 2020.

with number of neurons, activation of each layer, bias, kernel initializer, kernel regularizer, bias regularizes, activity regularizer, kernel constraint, bias constraint, training loss function, batch size, epoch count, an optimizer selection together with specific internal parameters such as learning rate, beta 1, beta 2, epsilon, and amsgrad status in case of the default adam optimizer. While most of those parameters have defaults, they differ between libraries and even version of the libraries themselves, making reproduction or result verification impossible if configuration information is incomplete. This problem can, and often is, solved via source code publication, often with a random seed fixed to achieve exactly the same results as published. A popular website *paperswithcode.com* tracks publications with source code available, as well as keeps leaderboards tracking performance improvement on selected datasets. This website has no drilling related papers listed. It is difficult to improve the state-of-the-art if it is not possible to reproduce it, making the goalposts invisible.

2.4. Incorrect data split

Sequential nature of real-time drilling data makes it susceptible to mistakes, for example related to train/test data split. One of the commonly logged attributes is measured depth, which with wrong data split, can inadvertently provide the model with information it should not have. Consider data where the only input attribute consists of evenly spaced values from 0 to 1. The target value is a random walk with step distance taken from normal distribution. There is absolutely no correlation between the input and the output. Yet if a random test/train split is applied, a common practice in ML, together with an off the shelf regression algorithm the resulting prediction R^2 score is 0.9948. This is because machine learning algorithm will *learn* that for input 0.40 target is 55 and for input 0.42 target is 57, then, when asked for prediction for input 0.41 accurate answer is very easy (most likely about 56). Full implementation showing the problem in the Appendix as Listing 1.

Even without the depth attribute there may be enough information for the model to correctly recognize that a certain sample is from the same area as the samples used for training and infer a correct output. Consider an absurd notion - Measured Depth prediction based on Surface Torque and Rotary Speed. While one can argue, that there would be some correlation, at least with the Surface Torque, these attributes are surely insufficient for an accurate prediction. Performing exactly this exercise, applying a random train/test split, it is possible to achieve impressive R^2 score of 0.946 for measured depth prediction using an off the shelf Gradient Boosting Regressor with default parameters - source code in Listing 2. With automated best model selection, testing portion reduced to 10%, and a malicious random seed selection, R^2 of 0.998 was achieved, or 1.00 if rounded to two decimals.

Using random train/test split may be used if the goal of the model is to interpolate existing data or explore the relationships existing in specific well. This however has to be done with caution and with understanding that spurious correlations will be utilized by the model and it should not be used to predict values for other wells. For example, if drilling was unusually slow at Measured Depth of 1,040m and 1,050m, such model will correctly predict that it was also slow at MD of 1,045m. This however is not an indication that in a different well ROP will also be low at this specific depth, even though the model is likely to indicate that.

In Scikit-Learn (Pedregosa et al., 2011) library an appropriate function exists for splitting sequence-type data; it is *sklearn.model_selection.TimeSeriesSplit*. While the name suggests that it is meant specifically for time series, it is also the correct tool to use for depth series type of data, or any sequential data where there is dependence between consecutive measurement points. This function divides the data into even, continuous splits. In the *k*th split, it returns first *k folds* as train set and the $(k+1)^{th}$ *fold* as test set. This creates multiple train/test sets in a structure suitable for continuous learning methods.

2.5. Easy target problem

Different problems can arise due to specific data selection and scoring. Consider data where the only input is a random number with average of 20 and standard deviation of 2. The target attribute is a random number with average of 50 and standard deviation of 3, so it ranges approximately between 40 and 60. No correlation exists between these two attributes, yet again, average error between the prediction and ground truth is only 4.8%. This is result of a very easy target, where simply guessing at the average, here 50, and ignoring the input will yield such an impressive result. The example code is shown in Listing 3.

While all these described shortcomings do not necessarily mean that any of the published papers have flawed methodology or doctored results. It is nevertheless a fact that the bulk of published papers on data-driven ROP prediction is done on undisclosed data, using methods that are not described sufficiently for reproduction, and therefore present results that are impossible to verify. One cannot improve on the existing research without reproduction or at minimum - a presence of a unified benchmark.

3. University of Stavanger Rate of Penetration (USROP) dataset

3.1. Data source - Volve dataset

In 2018 Equinor published data related to the Volve field located off the coast of Norway. It was in operation in years 2008–2016 and in total produced 63,000,000 bbl of oil. The dataset contains data related to geoscience, production, seismic, reservoir modelling and drilling. What made the adoption of the data for research relatively slow, is that the data is provided without any preprocessing that would make it more accessible. To facilitate research related to drilling, independent work was done to convert the real-time drilling logs from segmented WITSML files into compact CSV files (Tunkiel et al., 2020). This makes the data much easier to handle for the purpose of data science. The files are made available for download from the pages of University of Stavanger.[4] The Volve dataset is published on Creative Commons BY-NC-SA 4.0 licence, which allows anyone to modify and re-publish the data as long as the original source is attributed, it is done in a non-commercial fashion, and that the license is retained.

3.2. The data

The real-time drilling data from the Volve dataset is vast and allows for research in different sub-fields of drilling. At the same time the logs are of varying quality, containing missing data and different attributes. Significant clean up pre-work is required before feeding the data into ML algorithms, which can be to a high extent arbitrary. To solve this issue a curated subset of Volve is necessary. Based of analysis of the logged attributes total of seven wells were selected. The selection criteria was the completeness of the data and common attributes logged. Additionally only the depth-based real-time WITSML logs were used as opposed to time-based logs. Rationale behind this selection was that the time-based data would require additional processing which is currently outside of the scope of this study. The following wells were selected:

- Norway-NA-15_47_9-F-9 A depth
- Norway-Statoil-15_47_9-F-7 depth
- Norway-StatoilHydro-15_47_9-F-14 depth
- Norway-StatoilHydro-15_47_9-F-15 depth
- Norway-StatoilHydro-15_47_9-F-15S depth
- Norway-StatoilHydro-15_47_9-F-5 depth
- Norway-StatoilHydro-15_47_9-F-9 depth

[4] http://www.ux.uis.no/~atunkiel/.

Table 1
USROP dataset well depth reference.

Filename	Starting Measured Depth[m]	Final Measured Depth[m]	Available length[m]	Sample count
USROP_A 0 N-NA_F-9_Ad.csv	491	1206	715	13,746
USROP_A 1 N–S_F-7d.csv	301	634	332	6389
USROP_A 2 N-SH_F-14d.csv	988	3466	2478	47,645
USROP_A 3 N-SH-F-15d.csv	1306	4065	2759	53,041
USROP_A 4 N-SH_F-15Sd.csv	1401	4090	2689	51,708
USROP_A 5 N-SH-F-5d.csv	2828	3792	964	18,548
USROP_A 6 N-SH_F-9d.csv	225	634	408	7851

In terms of available attributes, the focus was on commonly logged data to promote models for wide application. Following attributes were selected: Measured Depth [m], Weight on Bit [kkgf], Average Standpipe Pressure [kPa], Average Surface Torque [kN.m], Rate of Penetration [m/h], Average Rotary Speed [rpm], Mud Flow In [L/min], Mud Density In [g/cm3], Diameter [mm], Average Hookload [kg], Hole Depth (TVD) [m], USROP Gamma [gAPI].

It is necessary to understand, that data in Volve, as well as generally in drilling rigs, is not collected in a standardized manner. Different equipment is used and operated using different procedures. This is often the reality of the oilfield operations and a method that is meant for future field deployment should be robust enough to overcome those challenges. Alternatively, all equipment could be properly and identically calibrated, as highlighted by (Hegde et al., 2019). This is an important distinction to note when comparing methods and results.

Minimal processing was done to the attributes to preserve original data. This is necessary as the drilling logs often contain erroneous, non-physical values. There may be sentinel values to indicate no reading (typically −999), corrupted values coming through the mud-pulsing system, and others. Samples containing *Weight on Bit* values below 0 and above 35 were truncated. The same way rows with *Mud Density In*, *Mud Flow In*, and *Average Surface Torque* values below zero were removed, as well as with *Rate of Penetration* values above 100 and *Average Standpipe Pressure* above 25,000. *Diameter* refers to the nominal wellbore diameter. Forward and backward filling was used to fill in the small gaps in the data resulting from uneven logging frequency of different equipment.

There was no unified gamma reading between all the wells, hence a new attribute, *USROP Gamma*, was introduced. It contains data logged under different names and different equipment, sometimes even within the same well. Source code is provided for exact explanation of which gamma related attribute was used in each case. The dataset was balanced in terms of samples per measured depth available. The goal was to remove the variability in the polling rate, so that a given length of a well has the same weight in terms of error, when it is calculated as a total of multiple wells. Well with fewest samples per available depth was identified and sample count of other wells was reduced through random sampling to match the identified value. Additionally, it is typically beneficial for ML approaches to balance the dataset anyway. It prevents error minimization algorithms from being overwhelmed by an over-represented value. In simple terms, an algorithm differentiating between cats and dogs will be best trained when the dataset is split evenly between these two classes. If cats were to be over-represented in a ratio of 9:1, algorithm may settle on a local minimum where everything is a cat with 90% accuracy. Complete source code used for data preparation is made available on GitHub.[5]

The described pre-processing results in data from seven wells with total available measure depth ranging from 332m to 2,759m. This translates to 6389 and 53,041 data samples respectively with total of 198,928 samples and 10,346 m of measured depth among all the wells. Table 1 provides the exact breakdown of those values. The names of the well were changed so that they are easily identifiable, such as *USROP_A 2 N-SH_F-14d.csv*. *USROP* stands for ***U***niversity of ***S***tavanger ***R***ate of ***P***enetration. ***A*** refers to the revision of the dataset. 2 is a short well identifier, and ***N-SH_F-14d*** refers to original CSV file titled ***N***orway-***S***tatoil***H***ydro-15_47_9-***F-14 d***epth.csv. For brevity, this paper henceforth will refer to the wells simply by their consecutive number, such as *well 2*.

To provide a general insight in how the quality of the data well 2 is reproduced in Appendix as Fig. 10. This chart is for reference only to provide a general feel for the dataset for potential researchers. Some outliers are present, changes in equipment are visible both in terms of well diameter, as well as gamma reading, which abruptly changes.

Among the attributes available in the presented dataset one can notice that lithology information is not present. This was done because such data is not always immediately available while drilling, and it is actually missing from the Volve dataset for a number of wells. The lack of lithology information will therefore promote creation of more universal models. There is still a possibility of an ROP prediction model that identifies lithology through unsupervised methods, such as clustering first, and then applies this knowledge to the testing dataset. Such approach potentially makes the model more robust and applicable to more operations. Nevertheless, further work is planned on creating additional curated dataset that would contain lithology information, bit wear, time, and other parameters.

It is also possible to develop models that work on additional predefined internal splits, for example dividing data further by the wellbore diameter, so that data for a given well section is evaluated by a model created only on training data from wellbores with an identical or a similar diameter.

3.3. Additional information

Volve dataset contains a plethora of information about the field and all the relevant operations, ranging from daily reports and production logs to reservoir models and seismic data. Due to volume of that additional information it was not pre-processed as a part of this paper. Some of that information still may be used indirectly as a hint or idea driving model creation, or directly, where additional data is used to potentially significantly reduce prediction error. Also in this case the USROP dataset will be useful as a reference point. When proposing an improvement, for example inclusion of seismic data to improve ROP prediction, the results may be directly benchmarked against the state-of-the-art models of other researchers. Today, if such improvement is proposed a *before-and-after* comparison is often questionable, as the *before* state is provided by the same researcher due to lack of comparable results available. This, consciously or not, may not be the best effort as one focuses on the new and improved model.

3.4. Score metrics

Not only data needs to be standardized in order to evaluate differences in models' performance, but unification of the way the results are calculated is also needed. Mean Absolute Error (MAE) is suggested as the key metric in case of the ROP prediction.

[5] https://github.com/AndrzejTunkiel/USROP.

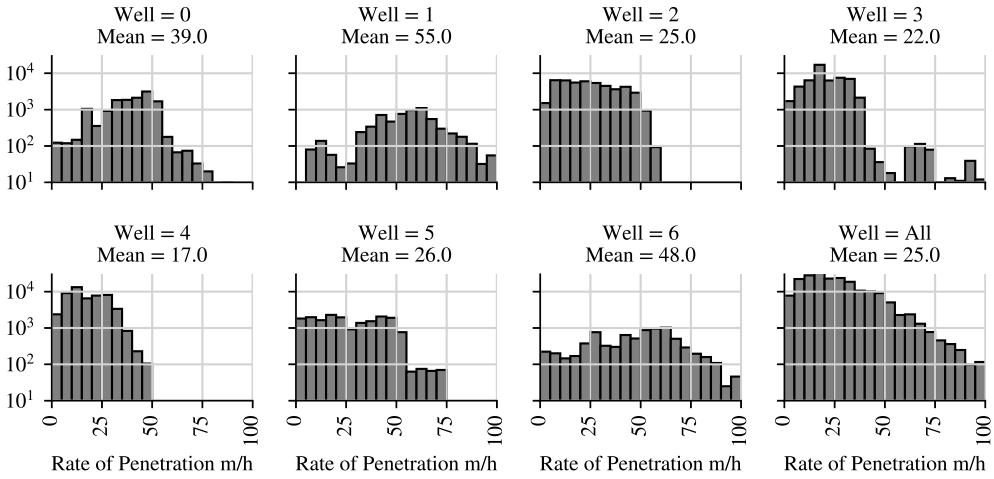


Fig. 2. ROP distribution of USROP wells.

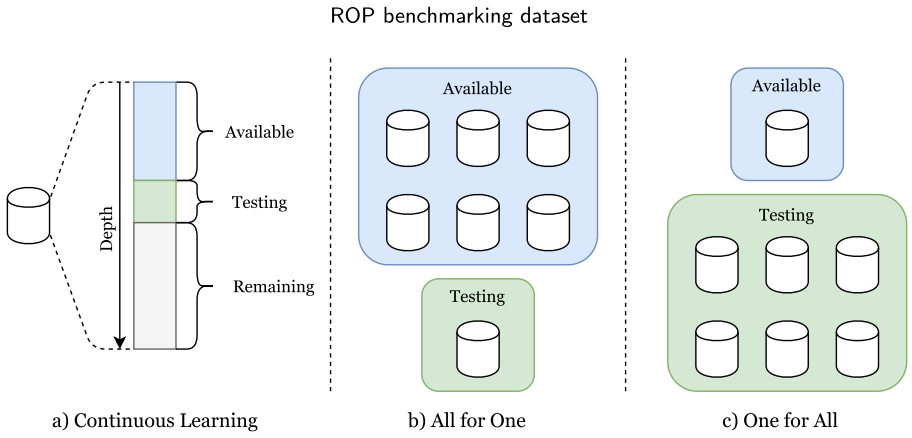


Fig. 3. Proposed scenarios for model validation.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{1}$$

Where n is the sample count, t is the consecutive sample number, A_t is actual value at sample t, and F_t is the forecast value for sample t. Rationale behind this choice is that ROP modelling is mainly done for drilling time optimization, where the interest is in the cost per meter drilled. A given value of error, for example 10 m/h, will be of roughly the same significance for an operator whether the true value is 30 m/h or 80 m/h.

Commonly used alternative to MAE is Mean Absolute Percentage Error (MAPE)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{2}$$

This heavily penalizes errors for low ROP values. In case of an error of 10 m/h in a very slow section of a well, like 0.1 m/h, becomes a 10,000% of error. In practice ROP can be very close to zero, or zero, on some datapoints, making the problem even more extreme generating error values at infinity. Manual removal of such datapoints is possible, but would make the results significantly influenced by arbitrary decisions. This was the deciding factor behind selecting MAE as the key metric. Supplemental metric is proposed to indicate the error value related to the absolute value of ROP without the infinite error problem - Weighted Mean Absolute Percentage Error (WMAPE)

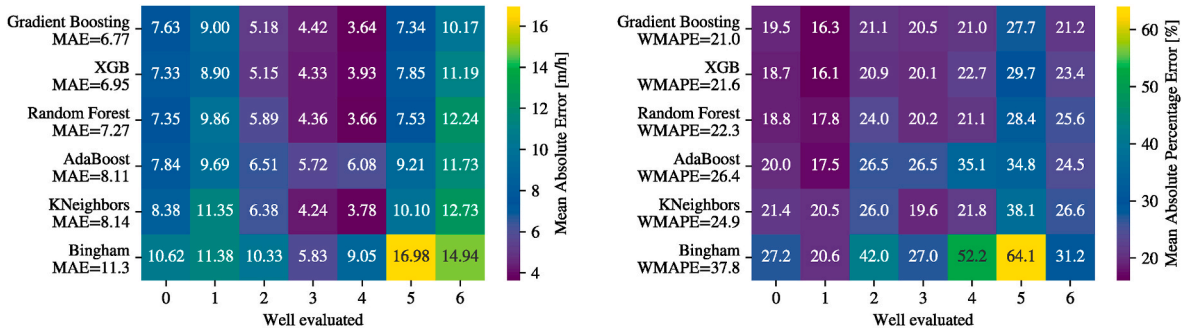
$$\mathrm{WMAPE} = \frac{\sum_{t=1}^{n}|A_t - F_t|}{\sum_{t=1}^{n}|A_t|} \tag{3}$$

This way of calculating error gives an indication of the scale of the error related to the complete well without the infinite error problem, as it avoids dividing by values close to zero. WMAPE is used as a supplemental metric in this paper.

Note that this is related only to model evaluation, and other metrics may be better for the purpose of training the model. Care must be taken when evaluating total MAE as it cannot be simply averaged between different wells, as the sample count is not identical. The best approach is to store absolute error values per sample from evaluating all the iterations and calculate mean at the end. To better understand distribution of ROP values histograms are provided in Fig. 2. Note that the sample counts are displayed in logarithmic scale.

3.5. Defined scenarios

Three scenarios are proposed to evaluate ROP prediction models **that reflect both reference well as well as continuous learning methods discussed earlier.** First, **Continuous Learning** scenario is suggested as a particularly attractive approach for real-time prediction. Each well is evaluated separately; initially first 30 m are available for training and validation to evaluate next 30 m. Next iteration considers first 60 m as available and subsequent 30 m for testing. After that 90 m are taken and so on. Note that the last testing section will necessarily be smaller than 30 m, as the total length of wells is not a multiple of 30.



(a) Mean Absolute Error (b) Weighted Mean Absolute Percentage Error

Fig. 4. Continuous learning scenario results for individual wells and methods, heatmap.

After processing the mean absolute error is reported for complete well taking into account testing scores from all iterations. Distance of 30m (100 ft) was selected as a typical length of a stand in a common triple drilling rig. To practically implement the train/test split in a unambiguous way, the split should done be every**577**[6] depth-sorted samples, which is equivalent to approximately 30 m of Measured Depth in USROP dataset. This implementation works around the problem of minor gaps in the data that can potentially cause different results in different implementations.

Second proposed scenario is **All for One**,[7] when all but one well is available for training and validation and one complete well is used for testing. This allows for seven different iterations with different well left for testing as cross-validation, and calls for one final MAE score from all the wells combined. Lastly, a **One for All**[8] scenario is proposed, where only one well is available, and all other wells are evaluate based on this model. As with the second scenario, this results with 7 train/test iterations acting as cross-validation. All three scenarios are shown symbolically in Fig. 3.

Note that only the split into available and testing data is fixed for each scenario. When methods like early stopping or dynamic model selection are used, the available data has to be split into training and validation data. All these samples have to be taken from the dataset designated available in the current iteration. It is also highly recommended that all publications related to ROP prediction based on this dataset share the relevant source code for research reproducibility and verification.

Note that all iterations in *All for One* and *One for All* scenarios are independent, hence developed algorithms should work on data only from a given iteration. In case of *Continuous Learning* scenario, each well is considered independent, but the sequence of expanding the dataset through drilling has to be maintained.

Referring to the data-split discussion in the *Incorrect data split* in the *Existing problems and potential pitfalls* section, it is worth highlighting that the proposed scenarios are *de facto* predefined, custom data-splits. Continuous Learning scenario is a variant of *TimeSeriesSplit* implementation, with fixed depth step instead of fixed split count. All for One and One for All explicitly splits the dataset into training and testing by well. Validation dataset is not specified, and it can be taken from the training data if so desired.

4. Reference results

Reference results are provided as a starting point for the basic algorithms' performance, as well as to gauge the available room for improvement. Source code to replicate the results is provided on GitHub.[9] Tested algorithms were mostly sourced from the Scikit-Learn library (Pedregosa et al., 2011): Gradient Boosting Regressor, Random Forest Regressor, AdaBoost Regressor and K-Nearest Neighbors Regressor. Additionally XGBRegressor was used from the popular XGBoost library (Chen and Guestrin, 2016). Additionally, results for classical approach - Bingham model (Bingham, 1964) developed in 1964 - are provided as well.

4.1. Bingham model

This is a popular model developed through laboratory testing. This is a common model acting as a reference point when suggesting improved ROP prediction methods in published research.

$$ROP = K\left[\frac{WOB}{D_b}\right]^a N \tag{4}$$

where K is constant accounting for formation strength, WOB is weight on bit, D_b is bit diameter, a is bit weight exponent and N is the rotary speed. The constants K and a were established based on minimizing the mean square error between the predicted and true value in the testing dataset, using the same training/testing data splits as done for the data-driven methods. A least squares optimization algorithm was used from SciPy library (Virtanen et al., 2020), where parameters K and a were selected such, that mean squared error (MSE), between real ROP values and calculated ROP values is smallest. This stands in apparent conflict where results are evaluated based on different metric. Our results however showed that MSE based optimization algorithms worked best producing lowest MAE and WMAPE results.

More advanced methods such as neural networks, 1D convolutions, recurrent models, automatic model selection such as TPOT library (Olson et al., 2016), ensemble results, additional pre-processing, and other approaches are possible. They are likely to yield superior results, however multiple separate publications will be needed to fully explore the ever-evolving landscape of machine learning methods that can be applied to the USROP dataset.

4.2. Continuous learning

Continuous learning is an attractive practical approach that can yield good results. It does not require information from reference wells making it possible to implement it without prior knowledge or data from a given field. It also does not suffer from typical problems such as changes in logged attributes or differences in equipment used between

[6] Total samples divided by total meters times $30.198928/10346{\cdot}30 = 576.8 \approx 577$

[7] Training on All, testing for One.

[8] Training on One, testing for All.

[9] https://github.com/AndrzejTunkiel/USROP.

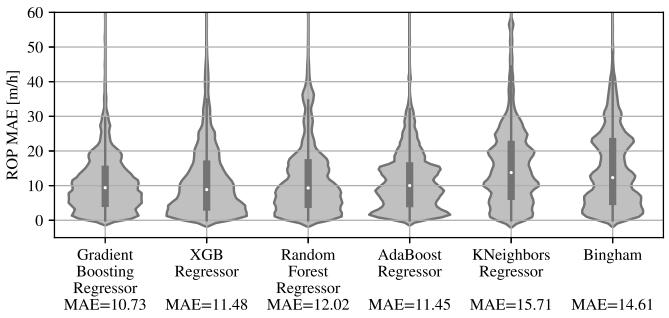


Fig. 5. All for One results per method, violin plots.

the wells and varying equipment calibration. Reference results for this scenario are presented in Fig. 4. The main proposed metric, MAE, is shown as subfigure 4a. It is presented as a heatmap, where cells correspond to various well and method combinations, and the color represents the error value. There is a clear difference in performance between the wells, with wells 1 and 6 typically being the most poorly modelled across all applied methods, and with best results for wells 3 and 4. There are two key factors that explain this finding: wells are different lengths, with longer ones allowing for nominally bigger training datasets, and different average ROP values (ref. Table 1 for well's size, and Fig. 2 for average ROP values). To compensate for differences in ROP, alternative heatmap is reproduced, as a subigure 4b, where results are normalized in terms of average ROP of a well and shown as WMAPE. Note that this is an error shown as a percentage of *average* ROP of a well, not ROP of a given sample, as in MAPE. This means that error of *x m/h* has the same value across a specific well, but it will change between wells.

For both MAE and WMAPE metrics the best overall score was achieved for Gradient Boosting Regressor of Scikit-Learn library. Bingham model, the classical approach, scored worst in all but one well, where it was placed next to last beating the AdaBoost Regressor, albeit only for MAE score. What is worth noting is that depending on the metric used a well can be *easy* or *difficult*. This is the case with well 1 and 6, where the error is high when calculating MAE, compared to other wells, but is low when using WMAPE. Refering back to Fig. 2, these are wells that have highest mean ROP. In practice both those metrics carry valuable information about the performance, and while there are differences between the wells, the overall rating of the methods remains mostly unchanged.

4.3. All for one

Reference results for All for One benchmark are shown in Fig. 5 as a violin plot. This type of chart is a merger of a histogram and a box plot, providing high resolution results at a glance in a compact format. The width of the light grey portion referencing the sample count for a given value on the y-axis, and is smoothed out via kernel density estimation. Note that the dot in the center of each figure represents median absolute error, while the key metric is mean absolute error. The dark grey bar spand 15th and 75th percentile. Best value in terms of MAE was achieved by Gradient Boosting Regressor of Scikit-Learn library. All defaults were kept for this algorithm, hence the outcomes are not fully representative of its potential. The results in terms of the well-by-well split are presented in as a heatmap in Fig. 6a. It is possible to see how the tested algorithms behaved on individual wells. Note that different algorithms may perform better for different wells. While Gradient Boosting Regressor has the lowest overall error, alternative approaches worked better when testing wells 0, 1, 3, 4, and 6. This suggests that using multiple models, such as an ensemble approach, is likely to yield improvements. What is particularly noteworthy is that the classical Bingham model scored best for wells 2 and 3. Additional heatmap with WMAPE metric is shown in Fig. 6b. As it was the case in Continuous Learning scenario, this metric changes for which well ROP prediction was done well or poorly, but the overall rating of tested algorithms stays the same. What becomes highlighted however, is the fact that Bingham model's error is over 100% for well 4, which is significantly higher than other methods.

Investigating results in more detail one can notice that Continuous Learning approach typically yielded better results than All for One. While this is true for all models, it is not true for all wells. Random Forest Regressor in All for One scenario achieved better results than any model in Continuous Learning approach. While this is only a preliminary result, it suggests that different modelling approaches may be optimal for different wells.

4.4. One for All

The One for All benchmark is significantly more difficult, as it is trained on one well only in contrast to six wells in All for One variant. This is visible clearly in inferior results seen in Fig. 7. In this benchmark, again, it is the Gradient Boosting Regressor that shows the lowest MAE;

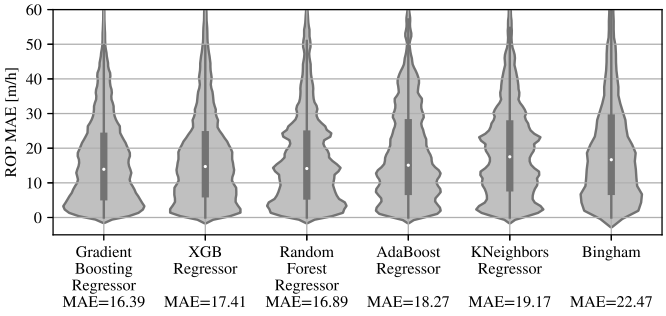
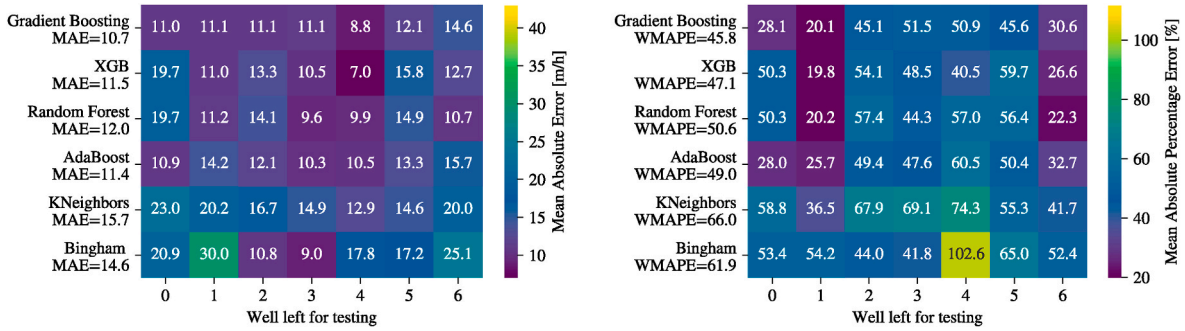


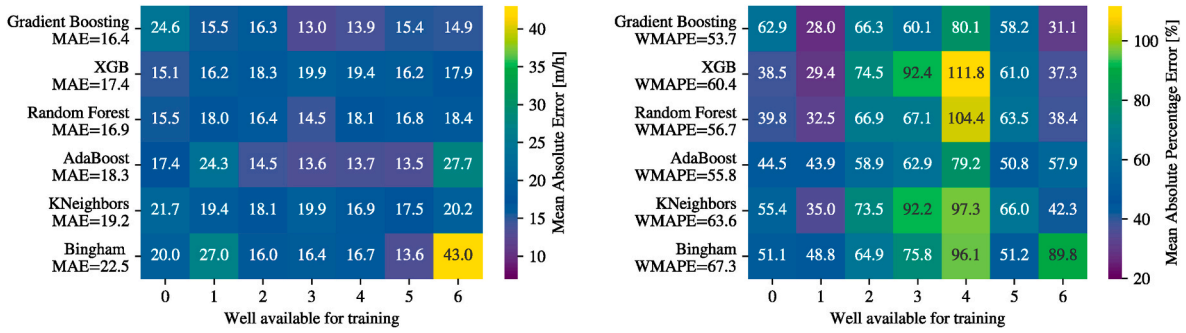
Fig. 7. One for All results per method, violin plots.



(a) Mean Absolute Error

(b) Weighted Mean Absolute Percentage Error

Fig. 6. All for One results for individual wells and methods, heatmap.



(a) Mean Absolute Error (b) Weighted Mean Absolute Percentage Error

Fig. 8. One for All results for individual wells and methods, heatmap.

albeit the value is 49% higher than in All for One scenario. The heatmap in Fig. 8a shows that, surprisingly, the best algorithm showed worst results for well 0 out of all tested algorithms. Note that the average MAE is calculated per sample basis, hence good results in biggest wells, here 3 and 4, have the highest weight. The traditional Bingham model again showed mixed performance scoring well for some wells (5) and poorly for others (6). In calculations done for WMAPE, Fig. 8b, results are similar with Bingham model performing worst, but uncovering that relatively it was the well 4 where it was most off-target, as that well was drilled more slowly, increasing the percentage error.

5. Discussion

5.1. Comparison against a flawed methodology

The reference results for all three proposed scenarios significantly differ from numbers presented in previous research related to ROP prediction using machine learning. The absolute error is much higher, and improvements over the Bingham model are only modest. First basic reason for this situation is that presented results are not indicative of the state of the art, but act as an indication of performance of the off-the-shelf algorithms applied without any tuning. This paper does not aim at developing an ROP prediction model, but to facilitate comparative research in this domain.

The second reason for relatively poor results is that all three proposed scenarios were developed to be realistic and representative of the overall performance of different models. Proposed dataset consists of drilling operation through various lithologies with no explicit attribute identifying them, and using different eqiupment. This is representative of real-life drilling, where such information is not always readily available. It was found to be very common in related publications that they are often limited the ROP prediction models in specific lithology.

To underline the perceived performance differences potentially stemming from different data split an exercise was performed, where USROP dataset as a whole, all wells joined together, was split into train/test portions in *9:1* ratio, with data rows randomly assigned to those subsets. Such methodology is flawed and does not represent real-life performance, a problem indicated in the second section of this paper. This approach was found to be in use in at least one of the reviewed papers, unsurprisingly reporting very good results. Using an untuned Gradient Boosting Regressor we achieved MAE value of *4.75*, and $R^2 = 0.82$, better than all benchmarked algorithms in all the USROP proposed scenarios, and approximately half the error of the best global result. To further improve the score, TPOT library (Olson et al., 2016) was used to automatically search for best performing off-the-shelf algorithm to replicate assumed reasonable best effort in making an ROP model. This resulted in an algorithm with MAE under *0.3* m/h. Further stretching the apparent performance by applying the train/test split with different random seeds it was possible to identify a specific random sampling

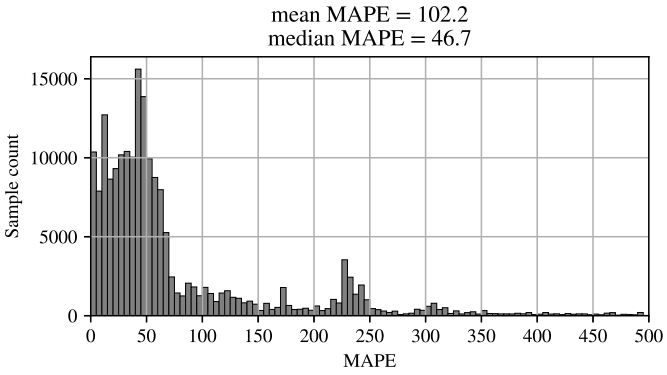


Fig. 9. Mean Absolute Percentage Error, calculated as per sample percentage (MAPE).

pushing the MAE down to *0.265* m/h and a near-perfect R^2 value of *0.999*. While those numbers may look impressive, they are just a result of a flawed methodology and do not translate to practical application.

5.2. Bingham reference results

Among the tested algorithms in this paper the Bingham model is both most widely used by other researchers, and the one implemented most unambiguously. This allows one to compare the USROP dataset to other datasets via proxy of this model. Previous work on undisclosed dataset provided by Marathon (de Salles Soares, 2015) evaluated the Bingham model both on individual lithologies as well as on entire dataset. The overall MAPE ranged between 23% and 43% depending on the method used to identify the coefficients. This is in line with results from the Continuous Learning scenario, where in USROP dataset the WMAPE was between 20% and 103% when inspecting individual wells in different scenatios. Another research (Soares et al., 2016) showed Bingham model achieving MAPE between 33% and 40%, again in line with results from USROP dataset. It is worth noting that these metrics are similar, yet not identical, since WMAPE in our paper refers to percentage of average ROP to avoid results being overwhelmed by moderate nominal error at very low ROPs. Methods in calculating the coefficients of the Bingham model also vary, with the quoted papers using a 3rd party software, making comparisons difficult. For the sake of comparison MAPE was also calculated per sample for All for One scenario and shown in Fig. 9. The mean error is 102%, however as expected, the histogram clearly shows that the bulk of actually expected error is between 0% and 60% with outliers inflating the average. Truncating these would significantly reduce the overall MAPE. Those results are in-line with the other quoted papers, suggesting that USROP dataset does not significantly differ in terms of ROP prediction difficulty.

Listing 1: Train/Test split failure

```
import numpy as np
from sklearn.ensemble import GradientBoostingRegressor
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
np.random.seed(42)

X = np.linspace(0,1,10000)
y = np.cumsum(np.random.normal(size=10000))

reg = GradientBoostingRegressor(random_state=42)

X_train, X_test, y_train, y_test = train_test_split(
    X[:,np.newaxis], y, test_size=0.33, random_state=42)

reg.fit(X_train,y_train)

print(reg.score(X_test,y_test))
```

5.3. Future work

Volve dataset, which is the source for the USROP dataset presented here, contains significant amount of additional data that was not included. This includes, but is not limited to, time-based data, pit volume, trip tank volume, block position, bit depth, mud type, mud name, mud properties, daily drilling reports, seismic data, lithology data, production data, and more. Significant effort is necessary to make these data seamlessly available for the purpose of machine learning or more general data science. Operations in the Volve field were not recorded using unified parameter set, and therefore making a curated dataset is necessarily a balance between the amount of parameters included and the amount of wells that contain all the selected parameters.

Immediate work that is planned by authors is to either expand, or create additional dataset meant for ROP prediction that includes lithology data, as well as parameters currently missing that are necessary for implementation of Bourgoyne and Young ROP model (Bourgoyne et al., 1986); these include jet impact force, pore pressure gradient, fractional bit tooth wear and threshold bit weight per inch of bit diameter at which the bit begins to drill. Other researchers are encouraged to create curated (sub)datasets for the specific problems they are working on, what will facilitate and accelerate research in the respective domain.

6. Conclusion

The key novel aspect of presented work is the creation of a reference, pre-processed, and simple to use ROP prediction dataset with specific challenges to solve has high potential to become a catalyst for higher quality research. It is a necessary step to evaluate what is the current state-of-the art in ML applied to drilling. The proposed three scenarios are related to real-life situations and aim to be representative of field deployment. As the reference dataset is based on Volve data, it is possible to extract further information about the wells used, allowing for more informed modelling decisions, background information, as well as further expansion towards more reference datasets. It allows researchers to propose inclusion of specific new information and to have a universal benchmark to show the achieved improvement.

Additionally, we found that when the data-split is performed in a realistic manner, suitable for field deployment, the standard ML algorithms perform worse than in previously published studies. Presented reference results shed a light on the performance of data-driven methods, where, depending on the methodology and specific well, they may be both vastly superior and sometimes inferior to the classical approach of physics-based models. This further confirms the need for bigger reference datasets, which enable robust evaluation of the accuracy of developed models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Achnowledgements

We would like to express gratitude to Equinor for providing funding for this research through Equinor Akademia Program.

A Appendix

Nomenclature

A_t	actual value at sample t
F_t	forecasted value at sample t
n	sample count
t	consecutive sample number
bbl	barrel (unit)
CSV	comma separated values
MAE	mean absolute error
MAPE	mean absolute percentage error
MD	measured depth
ML	machine learning
MLP	multi-layer perceptron
MSE	mechanical specific energy
ROP	rate of penetration

RPM	revolutions per minute (unit)
USROP	University of Stavanger Rate of Penetration (dataset)
WITSML	wellsite information transfer standard markup language
WMAME	weighted mean absolute percentage error
XGB	extreme gradient boosting

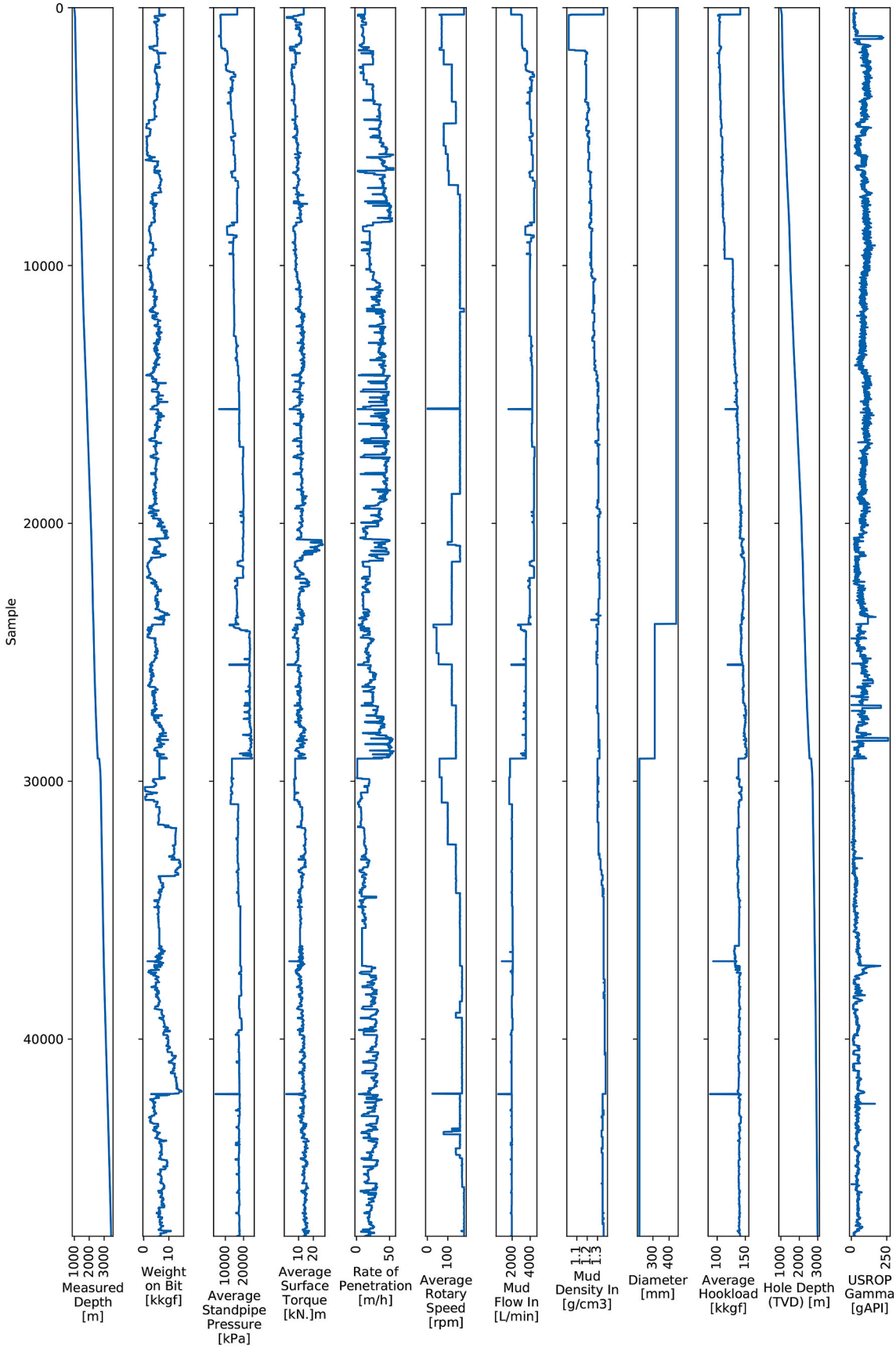


Fig. 10. All parameters of well 2, for reference only.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petrol.2020.108069.

Author statement

Andrzej Tunkiel: Conceptualization, Methodology, Software, Data curation; Dan Sui: Formal analysis, Writing - review & editing, Supervision; Tomasz Wiktorski: Writing - review & editing, Supervision

References

Ahmed, A., Ali, A., Elkatatny, S., Abdulraheem, A., 2019a. New artificial neural networks model for predicting rate of penetration in deep shale formation. Sustainability 11, 6527. https://doi.org/10.3390/su11226527. https://www.mdpi.com/2071-1050/11/22/6527.

Ahmed, O.S., Adeniran, A.A., Samsuri, A., 2019b. Computational intelligence based prediction of drilling rate of penetration: a comparative study. J. Petrol. Sci. Eng. 172, 1–12. https://doi.org/10.1016/j.petrol.2018.09.027.

Amar, K., Ibrahim, A., 2012. Rate of penetration prediction and optimization using advances in artificial neural networks, a comparative study. In: IJCCI 2012 - Proceedings of the 4th International Joint Conference on Computational Intelligence, pp. 647–652. https://doi.org/10.5220/0004172506470652.

Baker, M., Penny, D., 2016. Is there a reproducibility crisis? Nature 533, 452–454. https://doi.org/10.1038/533452A.

Bingham, M., 1964. A New Approach to Interpreting Rock Drillability. The Petroleum Publishing Co.

Bourgoyne, A.T., Young, F.S., 1999. A multiple regression approach to optimal drilling and abnormal pressure detection. SPE Repr. Ser. 14, 27–36. https://doi.org/10.2118/4238-pa.

Bourgoyne Jr., A.T., Millheim, K.K., Chenevert, M.E., Young Jr., F.S., 1986. Applied drilling engineering 2.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785. https://dl.acm.org/doi/10.1145/2939672.2939785.

Cohen, G., Afshar, S., Tapson, J., van Schaik, A., 2017. EMNIST: an Extension of MNIST to Handwritten Letters. Arxiv preprint. http://arxiv.org/abs/1702.05373.

Davey Smith, G., Ebrahim, S., 2002. Data dredging, bias, or confounding. Br. Med. J. 325, 1437–1438. https://doi.org/10.1136/bmj.325.7378.1437.

Deng, L., 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process. Mag. 29, 141–142.

Equinor, 2018. Volve field data (CC BY-NC-SA 4.0). https://www.equinor.com/en/news/14jun2018-disclosing-volve-data.html.

Eren, T., Ozbayoglu, E., 2011. Real-time drilling rate of penetration performance monitoring. In: Offshore Mediterranean Conference and Exhibition, 23-25 March, Ravenna, Italy., Offshore Mediterranean Conference. OMC-2011-076.

Eren, T., Ozbayoglu, M.E., 2010. Real time optimization of drilling parameters during drilling operations. In: SPE Oil and Gas India Conference and Exhibition. Society of Petroleum Engineers. https://doi.org/10.2118/129126-MS. http://www.onepetro.org/doi/10.2118/129126-MS.

Han, J., Sun, Y., Zhang, S., 2019. A data driven approach of ROP prediction and drilling performance estimation. In: International Petroleum Technology Conference 2019. IPTC 2019. https://doi.org/10.2523/iptc-19430-ms.

Hegde, C., Daigle, H., Millwater, H., Gray, K., 2017. Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models. J. Petrol. Sci. Eng. 159, 295–306. https://doi.org/10.1016/j.petrol.2017.09.020. https://linkinghub.elsevier.com/retrieve/pii/S0920410517307258.

Hegde, C., Gray, K., 2018. Evaluation of coupled machine learning models for drilling optimization. J. Nat. Gas Sci. Eng. 56, 397–407. https://doi.org/10.1016/j.jngse.2018.06.006.

Hegde, C., Gray, K.E., 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.

Hegde, C., Millwater, H., Pyrcz, M., Daigle, H., Gray, K., 2019. Rate of penetration (ROP) optimization in drilling with vibration control. J. Nat. Gas Sci. Eng. 67, 71–81. https://doi.org/10.1016/j.jngse.2019.04.017.

Hegde, C., Wallace, S., Gray, K., 2015. Using trees, bagging, and random forests to predict rate of penetration during drilling. In: Society of Petroleum Engineers - SPE Middle East Intelligent Oil and Gas Conference and Exhibition. Society of Petroleum Engineers. https://doi.org/10.2118/176792-MS. http://www.onepetro.org/doi/10.2118/176792-MS.

IHS Markit, . Offshore Rig Day Rate Index. URL: https://ihsmarkit.com/products/oil-gas-drilling-rigs-offshore-day-rates.html.

Jiang, W., Samuel, R., 2016. Optimization of rate of penetration in a convoluted drilling framework using ant colony optimization. In: SPE/IADC Drilling Conference, Proceedings. Society of Petroleum Engineers (SPE). https://doi.org/10.2118/178847-ms.

Liu, B., 2017. Lifelong machine learning: a paradigm for continuous learning. Front. Comput. Sci. 11, 359–361. https://doi.org/10.1007/s11704-016-6903-6.

Mantha, B., Samuel, R., 2016. ROP optimization using artificial intelligence techniques with statistical regression coupling. In: Proceedings - SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers (SPE). https://doi.org/10.2118/181382-ms.

Olson, R.S., Urbanowicz, R.J., Andrews, P.C., Lavender, N.A., Kidd, L.C., Moore, J.H., 2016. Automating biomedical data science through tree-based pipeline optimization. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics. Springer Verlag, pp. 123–137. https://doi.org/10.1007/978-3-319-31204-0_9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rahimzadeh, H., Mostofi, M., Hashemi, A., 2011. A new method for determining Bourgoyne and Young penetration rate model constants. Petrol. Sci. Technol. 29, 886–897. https://doi.org/10.1080/10916460903452009. http://www.tandfonline.com/doi/abs/10.1080/10916460903452009.

Sabah, M., Talebkeikhah, M., Wood, D.A., Khosravanian, R., Anemangely, M., Younesi, A., 2019. A machine learning approach to predict drilling rate using petrophysical and mud logging data. Earth Sci. India 12, 319–339. https://doi.org/10.1007/s12145-019-00381-4.

de Salles Soares, C.M., 2015. Development and Applications of a New System to Analyze Field Data and Compare Rate of Penetration (ROP) Models. Ph.D. thesis. University of Texas at Austin.

Shi, X., Liu, G., Gong, X., Zhang, J., Wang, J., Zhang, H., 2016. An efficient approach for real-time prediction of rate of penetration in offshore drilling. Math. Probl Eng. 2016, 3575380. https://doi.org/10.1155/2016/3575380.

Soares, C., Daigle, H., Gray, K., 2016. Evaluation of PDC bit ROP models and the effect of rock strength on model coefficients. J. Nat. Gas Sci. Eng. 34, 1225–1236. https://doi.org/10.1016/j.jngse.2016.08.012.

Soares, C., Gray, K., 2019. Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models. J. Petrol. Sci. Eng. 172, 934–959. https://doi.org/10.1016/j.petrol.2018.08.083.

Tunkiel, A.T., Wiktorski, T., Sui, D., 2020. Drilling dataset exploration, processing and interpretation using Volve field data. In: Submitted to Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering - OMAE.

Virtanen, et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Yi, P., Kumar, A., Samuel, R., 2014. Real-time rate of penetration optimization using the shuffled frog leaping algorithm (SFLA). In: Society of Petroleum Engineers - SPE Intelligent Energy International 2014. Society of Petroleum Engineers (SPE), pp. 116–125. https://doi.org/10.2118/167824-ms.