SPE-191601-18RPTC-MS

Application of Machine Learning for Oilfield Data Quality Improvement

Alla Andrianova, Maxim Simonov, Dmitry Perets, Andrey Margarit, and Darya Serebryakova, Gazpromneft Science & Technology Centre; Yuriy Bogdanov, Semen Budennyy, Nikita Volkov, Artem Tsanda, and Alexander Bukharev, MIPT Engineering Center

This paper was prepared for presentation at the SPE Russian Petroleum Technology Conference held in Moscow, Russia, 15–17 October 2018.

Abstract

The paper describes the principal possibility of using machine learning methods for verifying and restoring the quality of oilfield measurements. Basic methods for screening incorrect values have been given and approaches for solving three problems have been recommended:

- Correctness analysis of well logging data
- Quality control of physical and chemical fluid properties (PVT-studies)
- Separation between the base production and effect from well interventions (WI) to predict the performance of hydraulic fracturing (frac).

The main deliverable is a set of algorithms based on machine learning methods, which allows to automatically process large volumes of field data. A number of approaches is proposed, including using modern methods of machine learning, to restore the missing values and the quality of algorithms operation.

Introduction

Automation of data processing and preparation processes is one of the priority areas for introducing cognitive technologies in Gazpromneft [1]. The company, following a modern industry trend, is committed to digitalization, but faces a number of limitations related to the quality of input data. The lack of structure in the representation and storage of field data, low sampling frequency, large volume, and low reliability make it difficult to make quick decisions. Solutions based on inaccurate information can lead to critical consequences, therefore it is necessary to check the data for correctness and consistency in a timely manner. In conditions of exponential growth in the amount of incoming information, it is necessary to develop new integrated approaches aimed at analyzing large volumes of information using machine learning methods [2].

A relatively simple class of models is applicable to solving the problem of removing data outliers which is relevant for measuring bottomhole pressure measurements. The outlier screening is implemented by the Hampel method: the signal is analyzed with a specified window size (the number of neighbors is defined) within which the outliers are the points lying on several values of the standard deviation from the median value, after which they are replaced by the average of the neighboring points. In order to understand the nature of the outliers from the point of view of classical statistics, the figures below show the screening results of downhole pressures of two wells (Figure 1, Figure 2). For each well, histograms with the bottomhole pressure distributions are shown. In the first case, there is a clear peak on the right which corresponds to a "fence" of outliers lying above the main trend.

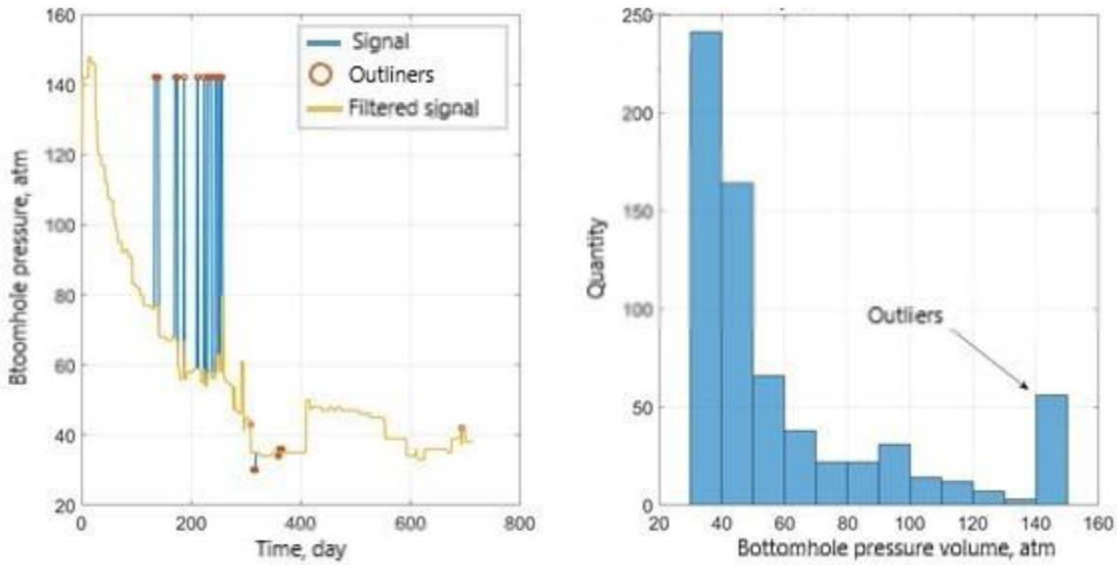

Figure 1—Removing outliers from bottomhole pressure measurement for production well A.

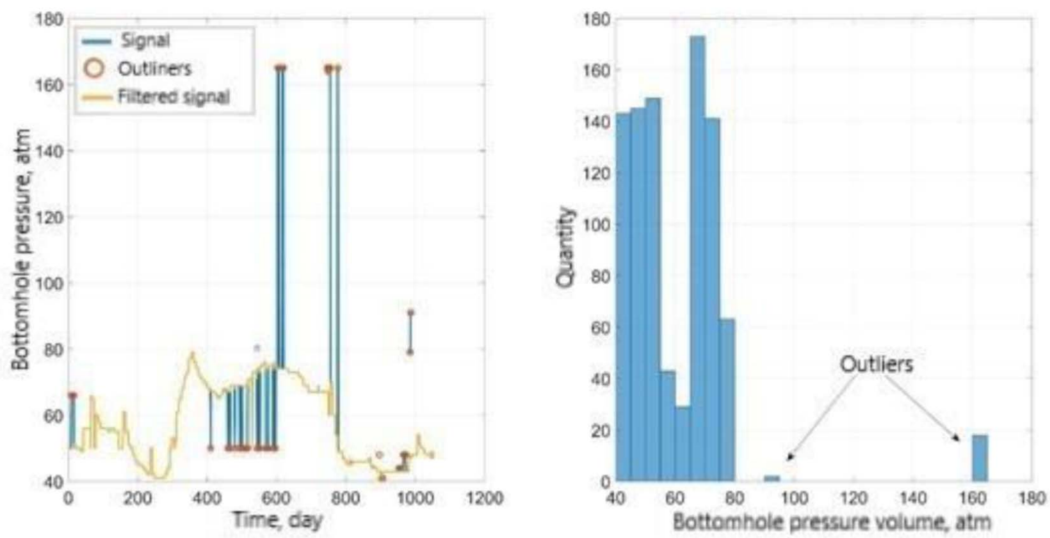

Figure 2—Removing outliers from bottomhole pressure measurement for production well B.

However, there is often a need to go beyond the classical statistical methods for data screening. Examples of such tasks are: analysis of logging data quality, analysis of the quality of PVT-properties, analysis of the estimation quality of a well production base. In the next sections, the above cases will be considered in more detail.

Logging Data Quality Analysis

Taking key decisions on field development directly depends on the interpretation of geological exploration data (well logging). Interpretation is performed by the subdivisions of the field operator and external contractors, and depends on the experience of the geology expert and the knowledge of the particular region/ target. The geological data contains many uncertainties, which complicate the interpretation and makes it probabilistic. To reduce uncertainties, it is logical to use additional independent expert reviews. It can be an independent expert system built on machine learning algorithms. In cases where the machine interpretation differs from the expert's interpretation, in the future, when designing the development, additional studies or examination of existing data will be required.

The task of restoring and evaluating the quality of data is solved in the context of the restoration of the lithology data interpretation and the search for reservoirs. The proposed approach is relatively universal and can be used to restore other both categorical and numerical data.

The published sources cover the approaches to log data processing and interpretation. Kormaksson, 2015, [26], offers an automatic method for identifying reservoirs using data from vertical wells, taking into account their spatial position relative to other wells. To highlight the informative features, the fPCA (Functional PCA) method is used, a functional method for selecting the main components. Pejman Tahmase, 2017, [27], addressed the exploration problem of shale gas reservoirs using the wellbore studies data. As a prediction algorithm, the author used an artificial neural network with optimization by a genetic algorithm. Leila Aliouane, 2014, [28], proposed a method of building a Poisson's ratio map for further use when determining the drilling direction and gathering rock characteristics information. Ritesh Kumar Sharma, 2015, [29], studied the problem of separating thin interlayers by logging data and the influence of a tool characteristic function on the recorded parameters.

The task of determining the lithology type and search for a reservoir by log data is significantly different from the classical classification problem. First, the number of logging studies carried out in all wells is limited. Second, the conclusion about the presence of a reservoir is made on the basis of complex information, including the spatial position of a potential reservoir, the distribution of properties within it, and the presence of "traps." To obtain a satisfactory result, two problems must be solved: classification in conditions of missing data, use of over- and under-lying intervals to predict net reservoir at a particular depth.

To solve the first problem, the authors propose to use a classifier capable of operating under conditions of missing data. At the moment, such a problem has been solved in the concept of gradient boosting over selection trees and implemented in XGBoost libraries.

To solve the second problem, the authors propose to apply a segmentation problem, which will greatly reduce the influence of noise in the input data and simultaneously take into account the large depth intervals. Among the algorithms examined, which show the most qualitative and stable results of segmentation, a one-dimensional convolutional neural network can be identified. At the same time, it is proposed to use the F1 quality metric (Dice-S⊘rensen coefficient).

The architecture of a one-dimensional convolutional neural network represents the encoder-decoder scheme (Figure 4), built on the basis of the network proposed [30], where one half is created to extract useful attributes from the data (encoder), and the second - for building the result of segmentation (decoder). The peculiarity of the architecture is a "passthrough" of layers between the encoding and decoding layers, which allows to reproduce the boundaries of segmented objects more accurately.

The final result is proposed to be built on the basis of the classifier prediction (XGB), built on a posteriori probabilities produced by the XGB model and a one-dimensional convolutional neural network.

The well logs from more than 200 vertical wells from a single field were used as experimental data. The data characteristics are shown in the figures below.

The source data features principal for models learning are as follows:

- The classes ("reservoir"/"non-reservoir") are not balanced (Figure 3)
- Only five log suites were run in most of the wells (Figure 4)
- Most logs contain outliers, the result of screening is shown in Figure 5.

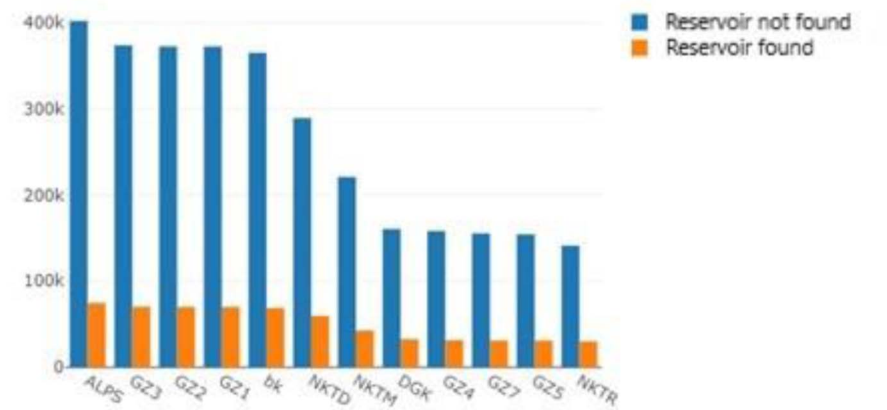

Figure 3—Description of the data used for the experiment. The columns correspond to different logging studies (the height is equal to the number of filled values).

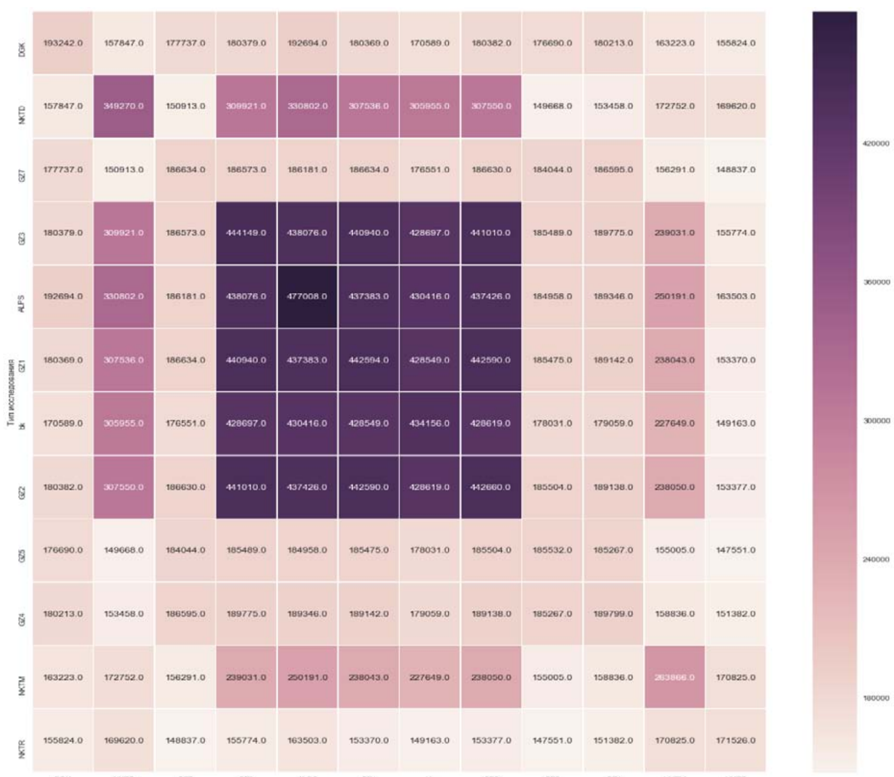

Figure 4—Description of the data used for the experiment. The columns and rows of the table correspond to the logging studies, cells show the number of points (depths) with both logs run simultaneously.

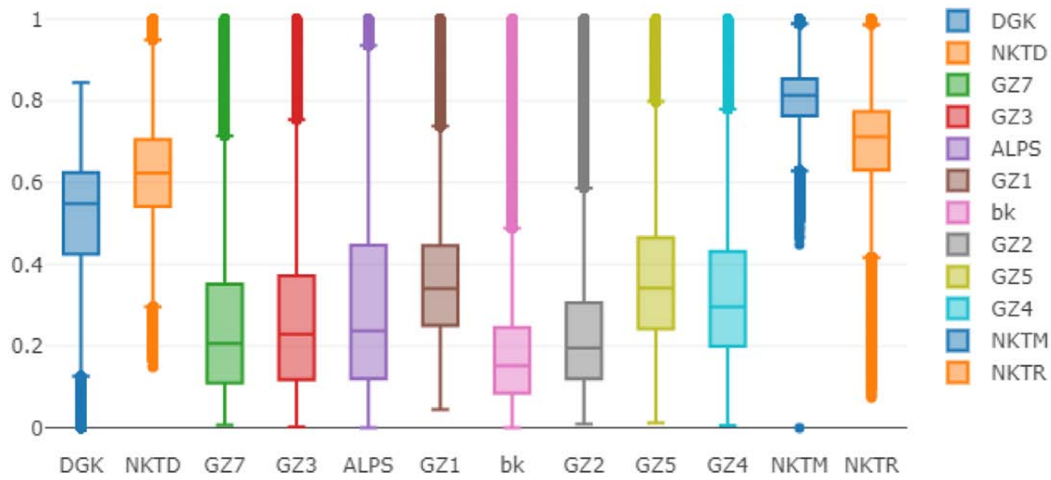

Figure 5—The distributions of values used for learning. The highlighted columns correspond to IQR (interquartile range), "whiskers" – to 3*IQR.

After preprocessing the data from the sample collection, standard procedures were performed: the wells were randomly divided into the main, validation, and test sample collections. The breakdown amounted to (approximately) 80% for the main, 20% for the validation, and 1% for the test samples (1 well).

The results of reservoir separation for the XGB model and models of the convolutional neural network are shown below for Well 102. The samples for learning the XGB model and the one-dimensional convolutional neural network were formed differently because of the different logic for processing the missing values for different algorithms:

- For the XGB model, all logs were used as input data, the missing values were not filled
- For the one-dimensional convolutional neural network, all logs were also used, but the missing values were filled with zeros, the DGK and NKTR logs were inverted.

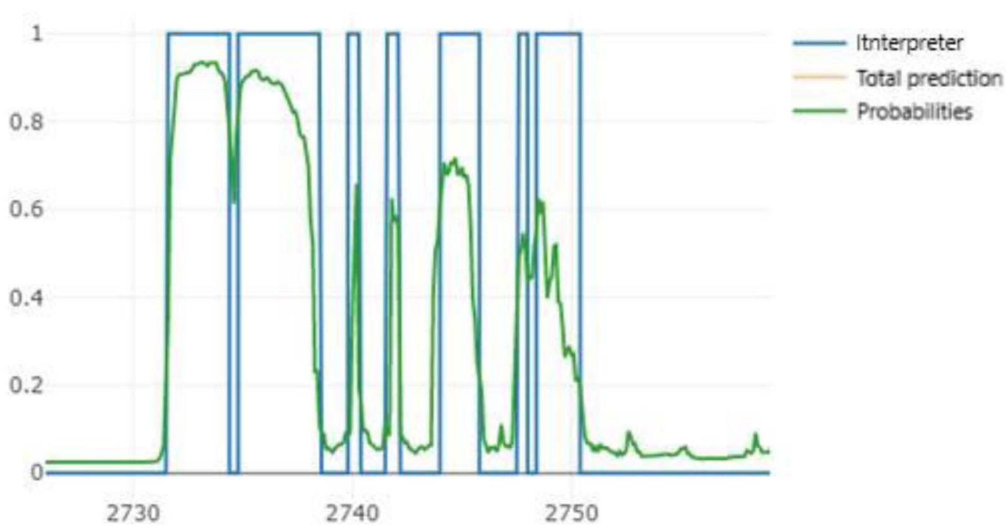

Figure 6—The reservoir presence probability (a priori) function (green curve), built on the basis of the XGB algorithm, and the initial interpretation (blue curve).

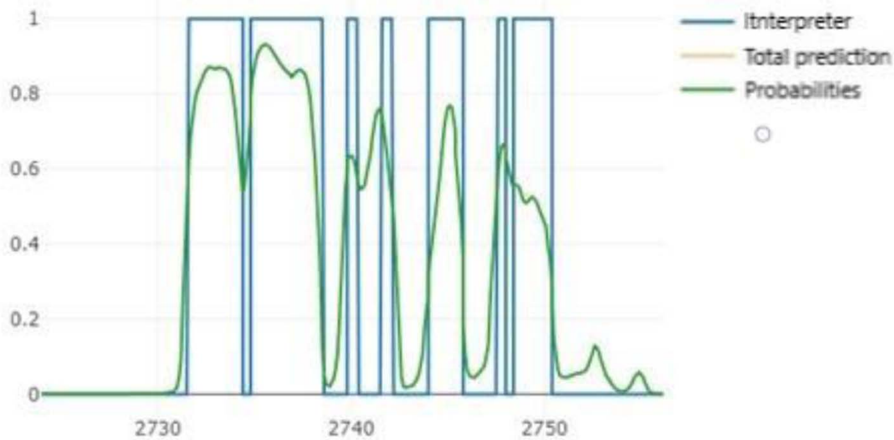

Figure 7—The function, similar to the one presented above, built based on the algorithm of the one-dimensional convolutional neural network.

The figure below shows an example of the algorithm to determine oil-saturated layers, compared with the results of the expert's interpretation. Note that in most cases there is a coincidence of results, regions 1, 2, and 3 mark the discrepancy regions. These discrepancies can not define the correctness of the algorithm or the level of the expert. These regions are needed for a closer attention of experts.

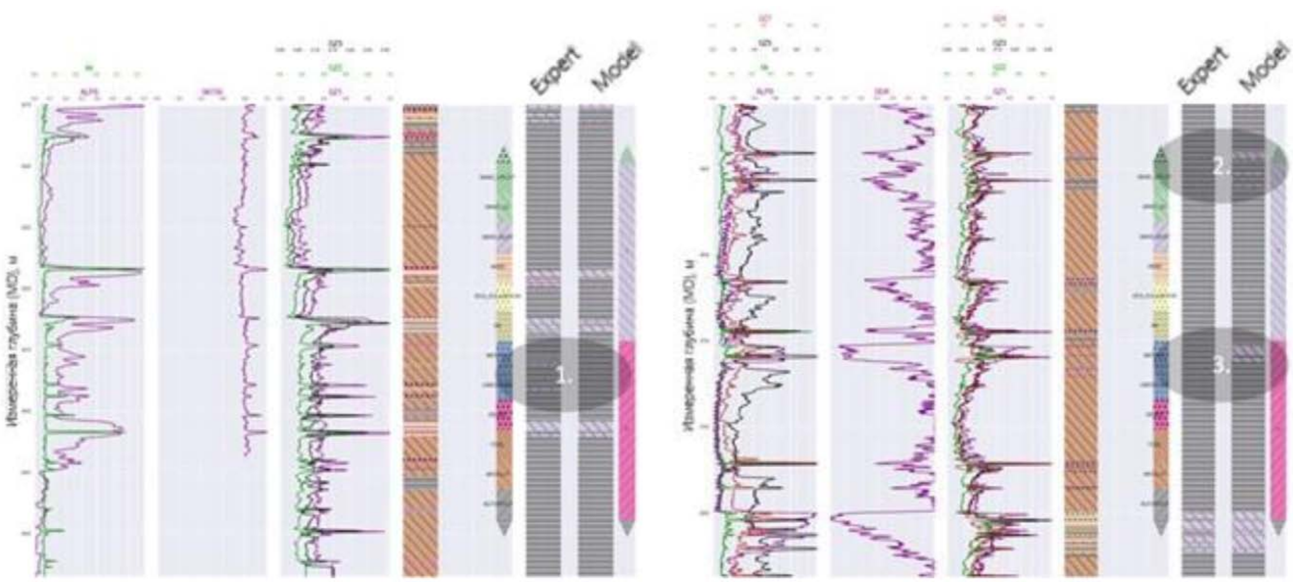

Figure 8—Regions for closer attention when interpreting well logs.

The real wells demonstrate the possibility of logging automation by machine and in-depth learning methods. The implemented algorithm acts as an independent expert: in the case of a strong discrepancy between the results of the expert's interpretation and the results obtained by the trained algorithm, an additional analysis of the well will be required.

The quality of the machine learning algorithms is directly affected by the quality of input data, thus, the quality of reservoir presence predicted with the help of the convolutional neural network can be used to judge the quality of the log data.

PVT properties analysis

The justification of the component composition and PVT properties of reservoir oil is one of the most important conditions for increasing the reliability of reserves estimation and the efficiency of field development design. Currently, there is no single regulatory document describing the procedure for laboratory studies of formation fluids (oil, gas, and water) [3]. There is also no single form of presentation of the studies results of formation fluid properties.

There is a number of methods for assessing the representativeness [4]:

- Check the tightness of sampling chambers, the presence of free water, compare the oil bubble-point pressure with the separation pressure at the separation temperature, etc.
- Apply the Hoffman-Crump-Hocott method [5], based on the correlation of equilibrium constants.
- Determine the representativeness of samples by the criterion of contamination with technological fluids used in drilling, perforation, and pre-production of the well.

In conditions where only raw data is available, the methods described above cannot be applied, and therefore, it becomes necessary to create algorithms for identifying potentially incorrect values using statistical methods involving analog fluids data. Figure below shows the PVT-properties quality restoration process.

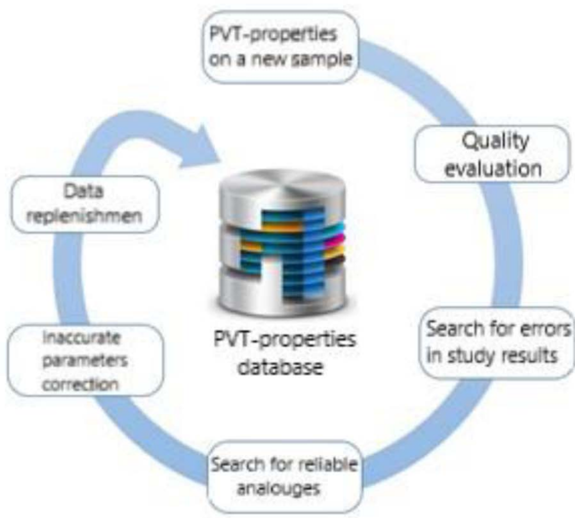

Figure 9—Analysis of PVT-data quality

For practical application of the PVT fluid properties studies, a total of 100 reports have been digitized and brought to a single format, containing a total of more than 300 fluid samples studies. Each report includes the results of a study of several reservoir oil and/or gas samples taken from a particular well. An algorithm for assessing the degree of reliability of the study results has been developed. A matrix of fluid similarity from the point of view of their PVT properties has been built.

Reliability of measurements of fluid properties analysis

To evaluate the validity of the fluid properties data, it is proposed to use a "value anomaly index" which is a combination of the results of outliers determination and the accuracy of value recovery using a nonlinear regression model which was based on the XGBoost model class. This approach allows to take into account both the value deviation from the sample median, and the degree of dependence of the studied value on the other parameters of the same object. If, with the selected threshold, this measurement is an outlier and is weakly dependent on other parameters for the object, that is, it is poorly predicted by the model, then the anomaly index for the given value will be close to 1 (lies in the range from 0 to 1), respectively, the lower the index, the more reliable is the value. Figures below show an example of visualization (Figure 3).

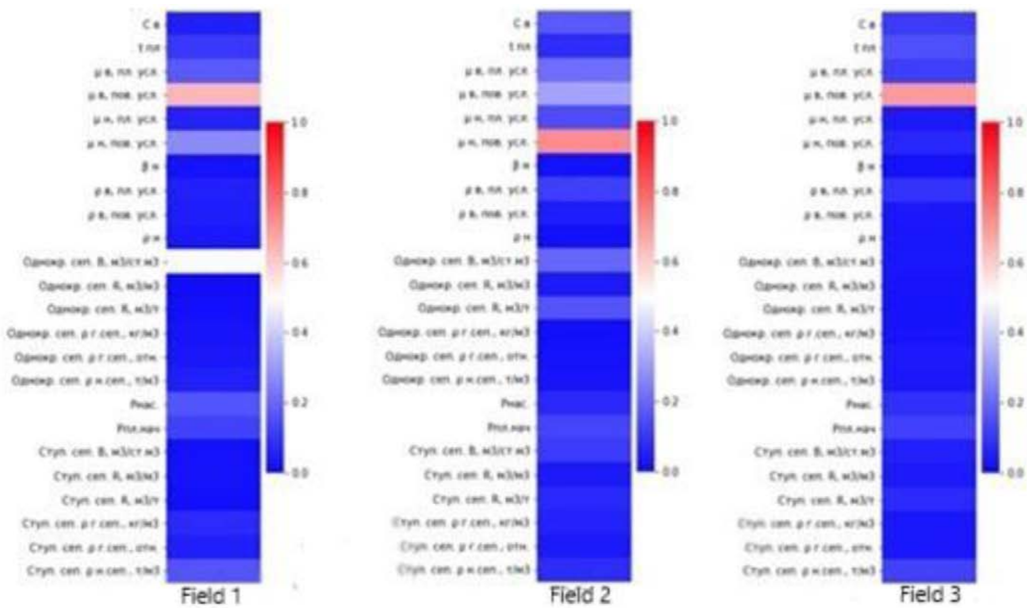

Figure 10—Visualization of the "values anomaly index" for the selected objects

For each of the objects, average anomaly index can be calculated. This value can be a measure of the quality of fluid properties studies. Figure 11 shows the first 30 objects in descending order of the average anomaly index by parameters.

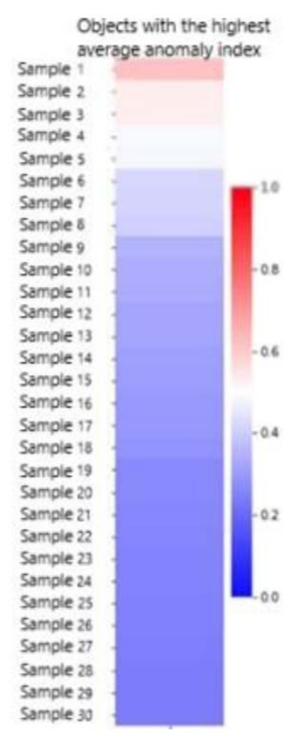

Figure 11—Objects with the largest average anomaly index in terms of parameters

There is another approach to calculate anomaly index of fluid from reservoir. As a first step atypical data are removed from train data by method IsolationForest[6]. Then projection to several last components in PCA method. Last ones altogether can only describe a small fraction of dispersion. As a result of projection typical objects are compressed to origin. Atypical instead deviate from origin. Object anomaly index (between 0 and 1) is determined as share of data from test data which after projection are lying closer to origin than observed(considered?) object.

Optimum PVT-correlation selection for the fluid sample

In practical application of the correlation approach, a problem of choice arises when selecting the type of PVT-properties correlation for restoring the missing values in the fluid sample under study. [7] showed the tables to select the most suitable correlation, the tables were built on the basis of an analysis of data available on oil samples and synthetic data.

The authors of this paper tested a statistical algorithm, which is a combination of statistical approaches, to select the best correlation based on machine learning. The studied correlations are shown below (Table 1).

Table 1—Correlation models of oil properties

Correlation	Year	Data set
Beal[8]	1946	USA
Standing[9]	1947	105 samples of California oil
Lasater[10]	1958	137 samples, Canada, USA, South America
Chew and Connally[11]	1959	Canada, USA, South America
Kouzel[12]	1965	n/a
Beggs and Robinson[13]	1975	n/a
Glaso[14]	1980	60 samples from the North Sea
Vasquez and Beggs[15]	1980	Over 600 samples from around the world
Gimatudinov[16]	1983	Udmurtia, Bashkiria, Tatarstan, Stavropol Territory, Volgograd Region, Kazakhstan, Ukraine, Belarus
Al-Marhoun[17]	1988	69 samples from the Middle East
Petrosky[18]	1993	84 samples from the Gulf of Mexico
Kartoatmodjo and Schmidt[19]	1994	740 samples, Indonesia, Middle East, North America
Velarde[20]	1997	n/a
Al-Shammasi[21]	1999	From around the world
Bergman[22]	2000	n/a
Dindoruk and Christman[23]	2001	Gulf of Mexico

The essence of the algorithm is the prediction of the most suitable correlation model by the fluid properties. To build a learning sample, the studied PVT property data sets were combined to improve the prediction quality.

The algorithm is a composite model:

1. The trained classifier predicts the type of a correlation function.
2. The algorithm uses the correlation to restore the missing value.

Implemented method allows to make the quantitative evaluations to apply preferred correlation function. These evaluations show the quality of prediction being made to determine properties of the fluid. They represent probability with which the algorithm choses PVT-correlation.

Example of algorithms' work shown below. There one can see probability distribution of PVT-correlations applied to one sample. Probability lies in range from 0 to 1.

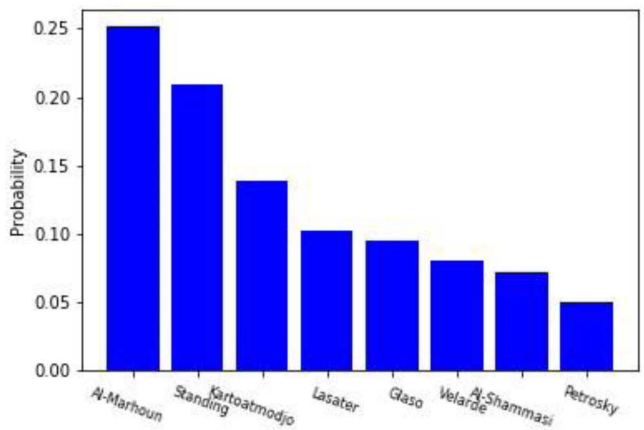

Figure 12—Probability of correlation choosing

Were set up experiment to study dependence of the parameters prediction quality from amount of the data. As a test for algorithm 20% of whole data frame were separated. The rest of the data frame were split into group by 10% of the whole data. The dependency is shown below.

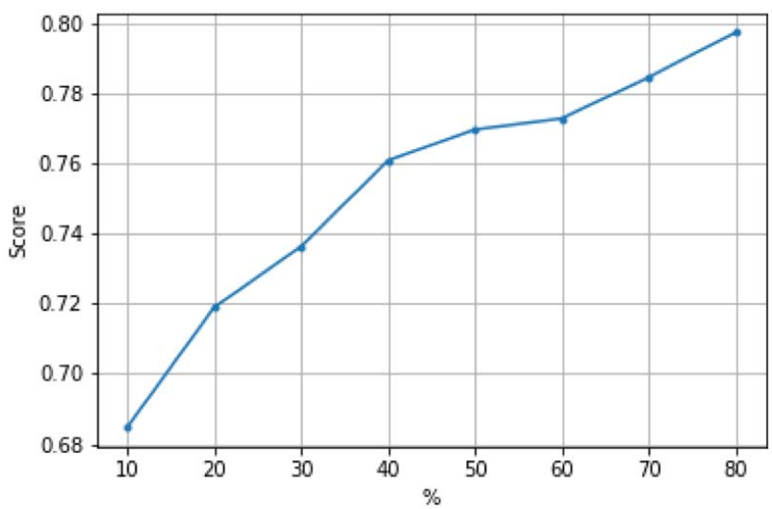

Figure 13—Score versus volume of train data

From this figure purely seen monotonous character of dependence. That fact allows to make a suggestion that larger data frame will lead to a better prediction accuracy.

In the experiment of preferred PVT-correlation prediction (classification into 8 classes), the value of the F1 metric was 0.72 after optimizing the parameters using the XGBClassifier method from the XGBoost library. The quality of the classifier is determined both by the accuracy and completeness of the prediction. In real problems, maximum accuracy and completeness are rarely achievable, therefore, to calculate the effectiveness of the algorithm, an F-measure (F1 metric) is used that provides a balance of these parameters.

The error was calculated by averaging the results obtained with sliding control on 5 breakdowns. In the second stage, the value of the dependent variable and the mean relative prediction error were calculated from the selected correlation. The results of the experiment are shown in Table 2 below.

Table 2—Relative prediction errors P_b of various methods with various data sets

Method	Error (Relative RMSE)
Standing	0.17
Lasater	0.18
Glaso	0.37
Velarde	0.23
Al-Marhoun	0.25
Al-Shammasi	0.17
Petrosky	0.25
Kartoatmodjo	0.25
According to the predicted type of correlation	**0.09**
XGBRegressor	0.37

The proposed method showed a prediction error lower than for any of the existing correlations alone. The errors listed in the table are calculated to predict the bubble-point pressure P_b for the remaining parameters.

Evaluation of the well base production

Another important issue in terms of data quality is the evaluation of the base production of a well (the expected production rate without well interventions). The base production rate of a well is the derivative information of the primary production data of the well, and the quality of this information directly determines the correct evaluation of well interventions success. The paper describes an approach to improving the quality of information on the basis of which the effectiveness of hydraulic fracturing is assessed.

The ratio of the total oil production rate of the well within half a year from the time of intervention to the frac-based baseline level was chosen as the well efficiency metric (the ratio greater than 1 indicates good performance of hydraulic fracturing). The frac-based baseline production level was estimated in two different ways - using empirical models and using machine learning methods. As a prediction quality metric, the RMSE value between the total prediction values and the actual production rate over a six-month period was considered.

The prediction of the frac-based baseline production level using machine learning models is to reduce to a regression problem. The objects in this problem are pairs (well & time point). The indicative description is built for all available characteristics of the well up to the considered time point inclusive, as well as the controlled parameters after that. The target mark is the total oil production rate of the well within half a year after the considered time point.

The machine learning objects are created for all wells and all time points. In case of a hydraulic fracturing in a well, to create the objects, only the data before the hydraulic fracturing will be used. The quality of the predictions is measured by the RMSE metric on a test data set, which is taken from the common data set of all objects corresponding to last ten months. All other objects are used to train the model. To estimate the hydraulic fracturing base, the objects corresponding to pairs (a well and a time point of a hydraulic fracturing in that well) are built.

The following characteristics are considered as an indicative description of pairs:

- All values of the oil production time series during 12 months before the fracturing
- Total oil production six months before the hydraulic fracturing, as well as with a shift by one and two months from the hydraulic fracturing
- All values of the oil production time series during 12 months before the fracturing
- All values of numerical characteristics 12 months before the fracturing and six months after (pressure, head, frequency, etc.)
- Categorical characteristics (well pad, type, etc.).

A key difference from empirical methods (in particular, the Arps model) here is predicting the total oil production within six months without using the corresponding predictions for each month. The lack of predictions values for each month is that when estimating the total value of the production rate, errors in the prediction of intermediate values are added together.

The outlier-free data was used to test various gradient boosting models, among which the CatBoost[25] model proved to be the best. The value of the RMSE metric on the test data set amounted to 14.6. Different methods of predicting the base using empirical models (using the Arps model) give the metric values of more than 80. Thus, the machine learning model allows to obtain the base prediction six times better than the empirical models, and, as a consequence, allows to more correctly determine the success of well interventions.

Figures 14 and 15 below show the oil production rates for two wells. The horizontal lines indicate the base, the level is determined by the average oil production rate over a six-month period. Note that the machine learning model gives a significantly better prediction than the empirical model.

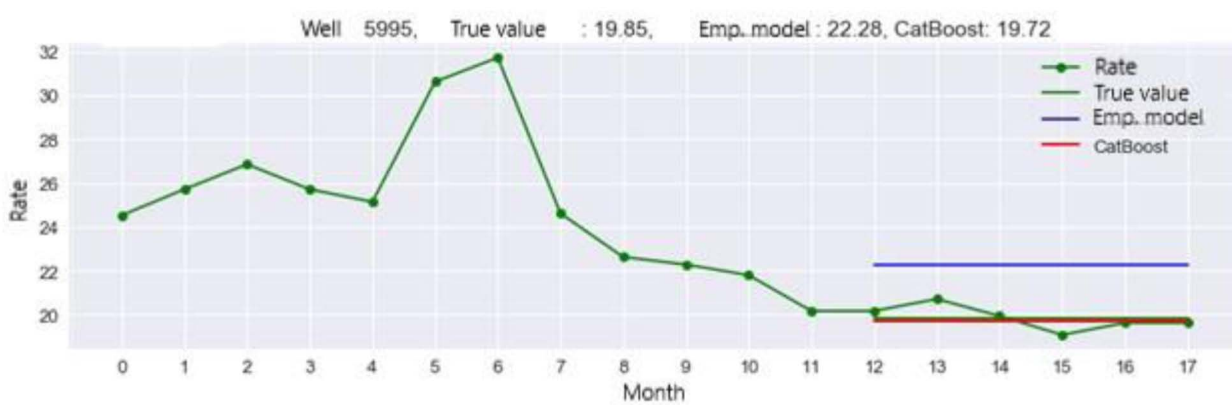

Figure 14—Oil production curve for well B and base predictions by various methods.

Figure 15—Oil production curve for well C and base predictions by various methods.

Deliverables

The real wells demonstrate the possibility of logging automation by machine and in-depth learning methods. The implemented algorithm acts as an independent expert: in the case of a strong discrepancy between the results of the expert's interpretation and the results obtained by the trained algorithm, an additional analysis of the well will be required. The quality of the machine learning algorithms is directly affected by the quality of input data, thus, the quality of reservoir presence predicted with the help of the convolutional neural network can be used to judge the quality of the log data.

To verify the data by the PVT-properties, an ‘anomaly index" was introduced, indicating potentially incorrect data that should be updated or be used with caution. Fluid analysis is an essential part of the information obtaining process on expected profitability and the field development period. A proper understanding of PVT dependencies is very important, because problems related to the quality of data on formation fluid properties can negatively affect the economic life of the project. The algorithms developed in the course of this work allow to evaluate the reliability of the available PVT data and thereby reduce project risks.

The empirical methods and machine learning models were used to analyze the effect of various characteristics of the well and the parameters of the planned hydraulic fracturing on its efficiency. To assess the efficiency of hydraulic fracturing, a method was developed for predicting the base production using machine learning models and pre-filtering of outliers. The accuracy of the base prediction according to the MSE metric is 14.6 which significantly exceeds the accuracy of the empirical methods.

Summary

The developed methods allow to more reliably (in comparison with traditional statistical approaches) restore the missing values in the field data, and also to correct the values significantly out of the total sample, which in turn makes it possible to identify errors and complications in wells operation in time, and ensures quick problem solving. In the conditions of increasing flow of information, the proposed algorithms allow to significantly reduce labor costs of specialists for preliminary data processing and to positively influence the efficiency of production decisions.

List of References

[1] V.V. Yakovlev, M.M. Khasanov, A.N. Sitnikov, D.O. Prokofiev, A.A. Pustovskikh, A.S. Margarit, M.V. Simonov, D.S. Perets «The direction of cognitive technologies development in the Upstream Division of Gazprom Neft Company» – *Neftyanoe khozyaystvo*, 12/2017

[2] A.M. Andrianova, A.S. Margarit, D.S. Perets, M.V. Simonov «Hierarchy of data verification approaches for production and development control» – *Neftyanoe khozyaystvo*, 12/2017

[3] A.I. Brusilovsky, I.O. Promselev, S.A. Fedorovsky "The problems of the lack of a single regulatory and methodological framework for justifying the properties of reservoir fluids and the experience of the company" *Gazprom Neft "//PROНЕФТЬ*. - 2017

[4] Kosyakova N.M. Parfenova D.R. Krajn,, I.M. Shafiev, V.A. Loginov, R. Yu. Narenkov «The issue of evaluation of the quality of reservoir fluids»// *Vesti gazovoy nauki*. – 2017

[5] Hoffman A.E. Equilibrium constants of a gas-condensate system: reprint // *SPE Reprint Series*. - 1981

[6] Liu Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua,. "Isolation forest." *Data Mining*, 2008. ICDM‘08. Eighth IEEE International Conference on.

[7] Khabibullin R., Khasanov, M., Brusilovsky, A., Odegov, A., Serebryakova, D., & Krasnov, V. New Approach to PVT Correlation Selection (Russian) // *Society of Petroleum Engineers*. - 2014

[8] Lake Larry W. Petroleum engeenirng handbook. - Texas, Austin - 2010

[9] Standing M. B. A Pressure-Volume-Temperature Correlation For Mixtures Of California Oils And Gases // *American Petroleum Institute*. – 1947
[10] Lasater J. A. Bubble Point Pressure Correlation // *Society of Petroleum Engineers*. - 1958
[11] Chew J. and Connally, C.A. A Viscosity Correlation for Gas-Saturated Crude Oils - 1959
[12] Kouzel B. How pressure affects liquid viscosity // *Hydrocarbon processing & petroleum refiner*. - 1965
[13] Beggs H. D. и Robinson J. R. Estimating the viscosity of crude oil systems // *Journal of petroleum technology*. - 1975
[14] Glaso O. Generalized Pressure-Volume-Temperature Correlations // *Society of Petroleum Engineers*. - 1980
[15] Vazquez M., & Beggs, H. D. Correlations for Fluid Physical Property Prediction // *Society of Petroleum Engineers*. - 1980
[16] ГиМатудиноВ Ш. К. СправВочное рукоВодстВо по проектироВанию разработки и эксплуатации нефтяных Месторождений - 1983
[17] Al-Marhoun M. A. PVT Correlations for Middle East Crude Oils // *Society of Petroleum Engineers*. - 1988
[18] Petrosky G. E., & Farshad, F. F. Pressure-Volume-Temperature Correlations for Gulf of Mexico Crude Oils // *Society of Petroleum Engineers*. - 1993
[19] Kartoatmodjo T Schmidt Z Large data bank improves crude physical property correlations // *Oil and Gas Journal*. - 1994
[20] J. Velarde W.O. McCain 48th Annual Technical Meeting of The Petroleum Society in Calgary, Alberta, Canada// Correlation of Black Oil Properties at Pressures Below Bubble Point Pressure-A New Approach. – 1997
[21] Al-Shammasi A. A. Bubble Point Pressure and Oil Formation Volume Factor Correlations // *Society of Petroleum Engineers*. - 1999
[22] David F. Bergman Robert P. Sutton A consistent andaccurate dead-oil-viscosity method // *SPE Reservoir Evaluation & Engineering*. - 2009
[23] Christman Birol Dindoruk and Peter G. PVT Properties and Viscosity Correlations for Gulf of Mexico Oils // *Society of Petroleum Engineers*. - 2001
[24] Tianqi Chen, Carlos Guestrin "XGBoost: A Scalable Tree Boosting System". arXiv:1603.02754, 2016
[25] Anna Veronika Dorogush, Vasily Ershov, Andrey Gulin "CatBoost: gradient boosting with categorical features support". *Workshop on ML Systems at NIPS* 2017.
[26] Kormaksson, M. (2015). A data driven method for sweet spot identification in shale plays using well log data. *Society of Petroleum Engineers*.
[27] Pejman Tahmase F. J., (2017). Data mining and machine learning for identifying sweet spots in shale reservoirs. *Expert Systems with Applications*, 435–447.
[28] Leila Aliouane S.-A., (2014). Sweet Spots Discrimination in Shale Gas Reservoirs Using Seismic and Well-logs Data. *A Case Study from the Worth Basin in the Barnett Shale*. Energy Procedia.
[29] Ritesh Kumar Sharma, S. C. (2015). Identification of thin sweet spots in the Duvernay Formation of north central Alberta. *SEG New Orleans Annual Meeting*.
[30] Olaf Ronneberger, P. F. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 234–241.