Forecast of water-cut at wells under design by machine learning methods

M R Enikeev[1], M F Fazlytdinov[1], L V Enikeeva[2] and I M Gubaidullin[2,3]

[1]Gazpromneft STC, Moika River emb., 75-79, liter D, St. Petersburg, Russia, 190000
[2]Ufa State Petroleum Technological University, Kosmonavtov St., 1, Ufa, Russia, 450062
[3]Institute Petrochemistry and Catalysis - Subdivision of the Ufa Federal Research Centre of RAS, Oktyabria pr., 141, Ufa, Russia, 450075

e-mail: mat-83@mail.ru, leniza.enikeeva@gmail.com

Abstract. A large amount of data is generated during the operation of oil fields. Such data can be both data already interpreted by a specialist, or "raw" data obtained directly from the devices, both structured and not structured, or locally structured (that is, allowing for local analysis, but in such form not allowing analyzing in conjunction with other types of data). To obtain from such a set of more informative data that will allow making decisions in the course of field operation, it is necessary to involve specialists from different areas of the oil industry. Therefore, it is possible and necessary to use non-deterministic methods for analyzing the data obtained. The article discusses the use of machine learning methods in the task of determining the initial water-cut based on well logging data.

1. Introduction

Initial water content (or water cut) of the well is the relative water content in the produced liquid, expressed as a percentage, at the beginning of the well operation. It allows assessing the feasibility of commissioning an oil well. One of the necessary tasks is the forecast of water cut in new wells and the allocation of the share of unproductive production/injection during the operation of wells that open water-saturated horizon in addition to the target horizon. Dynamics of water-flooding of oil wells is determined by the nature of oil reservoirs water-cut. The nature of the reservoir water-cut can significantly differ and depends on the properties of the productive layers, the initial conditions of oil occurrence in the reservoir. In addition, the nature of water-flooding and the water-flooding dynamics is influenced by layered and zonal heterogeneity. The intensity of watering depends on the permeability of the layer. The uneven watering of the layers by their thickness and strike increases with a high ratio of oil and water viscosity. Many of these factors are embedded in the methods of well logging.

Well logging is a set of exploration geophysics methods used to study the properties of rocks in the near-well and inter-well spaces and to control the technical condition of wells. Well logging is composed of two groups of methods — geophysical logging and geological logging.

Normally, well logging data represent strongly and randomly fluctuating functions (Figure 1).

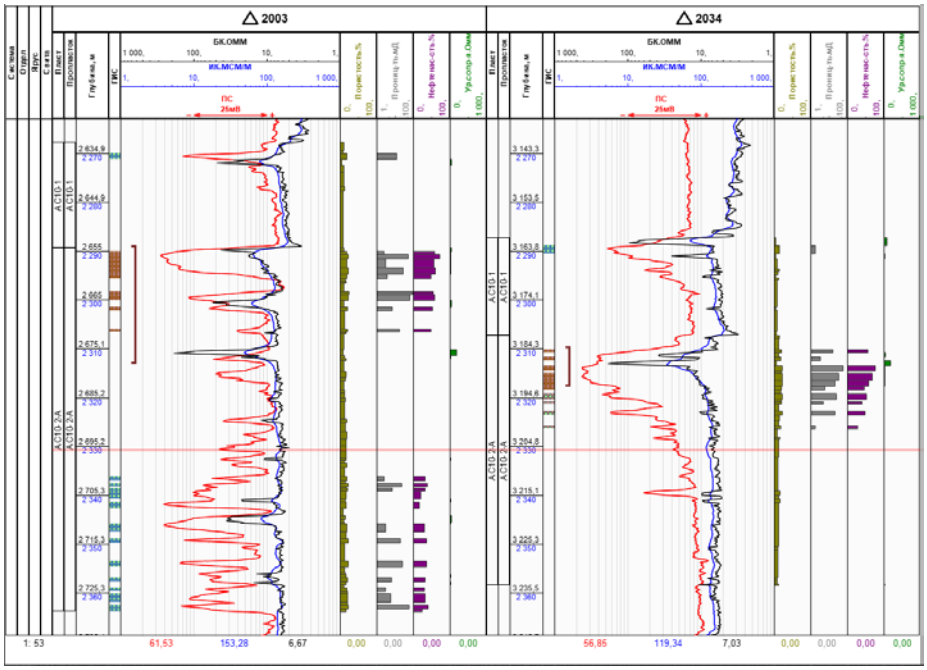


Figure 1. Example of logging data: the y-axis is usually the depth, the x-axis is the logging value.

In work [1] considers the possibility of applying analog-statistical methods in the forecast of the water-cut of production wells with the consideration of the influence of geological and technological indicators. One of the key elements of the methodology is building the modified dependences of wells water content on the degree of production of recoverable oil reserves for analog models. The conclusion about the adequacy of the forecast to the real geological object can be made by comparing the actual and forecast productivity. The use of such techniques can significantly reduce errors in the forecast of flow rates and water cut of new wells and increase the reliability of the operational production forecast. Deshenenkov [2] proposed a method for predicting the productivity and initial water content of oil deposits using: relative permeability to oil and water, field geophysical data and effective porosity. To predict water cut based on well logging data, the paper [3] uses the formulas of the Corey model, which determine the permeability for oil and water. Then, in the perforated intervals, accumulated permeabilities for water and oil are calculated. After that, taking into account the viscosities of both phases, the water-cut ratio is estimated. In these works, pre-defined model formulas are used to compare the values of water cut and logging data, but no in-depth analysis of the impact of different types of logging on the values of water cut has been made.

2. Problem statement

As mentioned earlier, well logging data are the main source of information about the reservoir at the stage of building of the geological model and the creation of the concept for oil field development. The aim of this work is to develop a method for forecasting water cut on new wells based on logging data.

Modern development of machine learning methods can effectively solve a wide range of problems in various fields. For example, when predicting the occurrence of a stroke [4], predicting the state of electromechanical systems of rolling production [5], predicting red shifts of galaxies [6], as well as in the oil industry, for example, to interpret seismic data [7] and to predict the development of corrosion of pipe steel [8]. The oil industry is a source of large amounts of structured and unstructured data. A large number of engineering (analytical / empirical) methods for studying the subject area have been developed. Machine learning methods well complement the existing set of tools for working with studied and developed oil fields.

In terms of machine learning, the problem of water cut prediction refers to the classification problem — it is necessary to divide the set of objects X (the set of logging data from the field) into M

disjoint classes Y (different water cut values). Thus, the concepts of object space and class space are defined. As it is known, there are two stages in machine learning tasks — the learning stage and the application stage. In this case, the training sample is a set of logging curves interpreted by the geophysicist, where for each element it is known what value of water cut it belongs to. An integral preparatory stage for the classification algorithm is also the selection of objects features, which will be discussed later.

3. Implementation of the approach

The problem is reduced to the feature classification problem. First, the data were pre-processed: filtering by values was carried out ('nan' values were removed from the data, obviously incorrect values were also removed (for example, values > 1 for 'aps', values < 0 for 'kint', and so on)). Then, for each parameter, its feature description was built. Features were generated based on different approaches:

- Data approximation
- Statistics
- Fourier analysis (not included in the final solution)

Further, in the report each approach is described in detail and justified. The classification was performed with regression algorithms that estimated the water-cut value from the logging data.

The capabilities of the following algorithms were studied in detail:

- Random forest
- Gradient boosting
- Neural networks

Technically, the problem was solved as follows. Using the lasio library and a python script, the data was unloaded from the LAS format into CSV (with depth parameter values). Then these data were loaded in another script, the data were filtered, the roof and floor values were determined, and the features for the training task were generated. The set of LAS features was matched with the water cut values (Matlab was used to generate features by Fourier method). The data were separated for training and control, in ratio of 70/30. These features were used to configure regressors from the skit-learn library and on tensorflow and keras neural networks.

3.1. Data loading and analysis

For the initial data analysis, it was decided to check whether there is an explicit dependence of the water cut (wc) value on the logging data. During the data analysis, it was observed that it is better to analyze the area lying between the roof and the floor. For the evaluation, it was decided to compare scaled data with the averaging values in this area for each parameter, to establish a relationship between them (Figure 2). From Figure 2 it can be concluded that the allocation of meaningful information is not possible. Dependency analysis on a logarithmic scale also gave no results. Therefore, it was decided to check how effective the various feature generation algorithms are.

3.2. Feature generation

The problem of feature classification was reduced to the classical feature classification, when each object (in this case, a curve or a set of curves) was described by a fixed set of real features. The simplest example of a feature is the average value on the curve. However, the features should be selected to describe the signal as accurately as possible, and to take into account the physics of the described processes. In addition, various heuristics can be used to generate the features.

Several approaches were used to generate features. All of them are described below. The result of such generation is a set of features, from tens to hundreds of values, depending on the method.

In order to determine more effective methods of feature generation, experiments were conducted with the setting up classifiers on different feature spaces and evaluating the classification quality on the control sample (30% of the total).

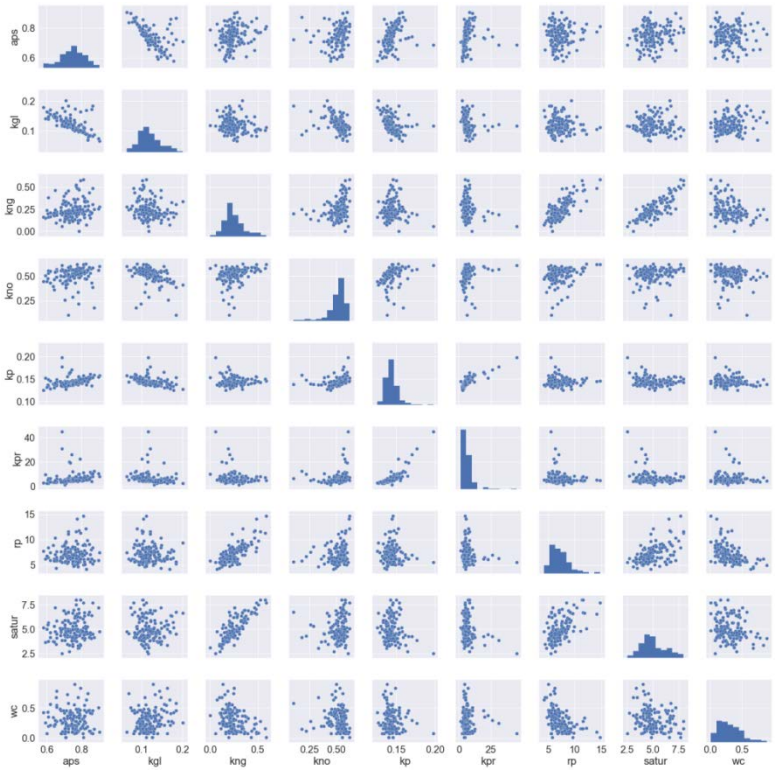


Figure 2. Correlation between various well parameters for which water cut values averaged over depth.

3.2.1. Feature evaluation

MAE (Mean absolute error) method was used to assess feature quality:

$$\text{MAE} = \frac{\sum_{i=0}^{N-1} |wc_{cor} - wc_{predict}|}{N}$$

where wc_{cor} – expected water cut value, $wc_{predict}$ – water cut value predicted by the classifier, N – number of wells in the control sample.

3.2.2. Features based on piecewise linear interpolation

The method based on piecewise linear interpolation was tested first. The data on the selected parameter (curve), after removing the incorrect values, were considered in the interval between the roof and the floor (this operation will be further denoted as preprocessing), after which the value was generated at k points, with the same depth step. Next, the values at these k points are submitted to the classifier.

3.2.3. Statistical features

First, each curve was preprocessed, then the statistical features for both the signal and its derivatives were calculated. The generation at this stage is carried out in the following order:

- Preprocessing
- For the signal $(x_1, \dots, x_n)$, derivative $(x_2 - x_1, x_n - x_{n-1})$ and modulus of the derivative $(|x_2 - x_1|, |x_n - x_{n-1}|)$ the following feature values are calculated: mean value, standard deviation, proportion of intersections with level a (a = 0, a = mean, a = mean + std)

An variant was also considered using features of percentile values (p10, p50, p90, etc.), mean value and deviation.

3.2.4. Features based on Fourier decomposition (transform)

When analyzing signals, one of the most successful methods is the use of the Fourier transform (in our case, the one-dimensional discrete Fourier transform (DFT)).

Using the DFT coefficients, it was not possible to achieve good results in the classification, so it was decided to lay out not the entire curve, but its sections. Then, these decompositions should be averaged (with logic to obtain stable features). For this, matlab was used, that has a spectrogram function, for dividing the signal into segments and calculating the DFT at each of them. Unfortunately, this method proved to be worse than the first two.

3.2.5. Selection of the optimal features and the parameter by which to conduct training

To select the optimal method of feature generation, it was decided to classify the ensemble of trees, with the search for the optimal settings for the method (changing the maximum depth of a tree, the number of trees in the ensemble of solutions, the number of selected features).

For a complete analysis of the optimal well logging curve for classification, the following curves were analyzed: 'kint', 'r05', 'r20', 'r14', 'r10', 'f07', 'f10', 'f14', 'r07', 'f20', 'f05', 'phit', 'mres', 'sg', 'kgl', 'sxwb', 'gz3', 'nphi', 'gz2', 'gz4', 'gz1', 'cild', 'prox', 'lld', 'gz7', 'aps', 'kps', 'gz5', since these curves are most densely filled with data.

The results of the comparison of the feature generation algorithms for most curves are shown in Figure 3. As a result of this test, it was decided to use an algorithm based on statistical features.

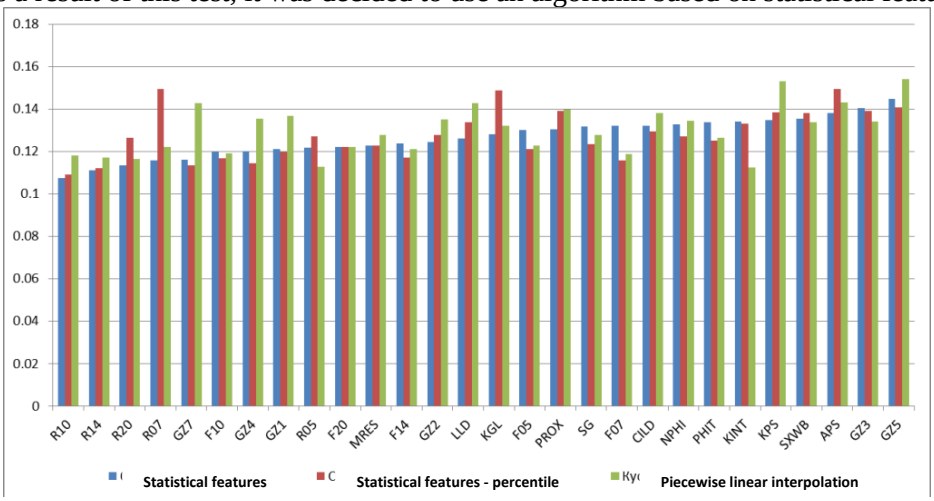


Figure 3. Comparison of feature generation methods for training.

The average MAE values for different methods are [0.127, 0.128, 0.130], and as we can see, the result is almost independent of the method of generating parameters and on the choice of the curve from the logging data.

The result on the control sample (30% of the test) on the gz5 curve is shown in Figure 4. Features were calculated by a statistical method. The real water cut values are shown in red, the predicted values are in blue.

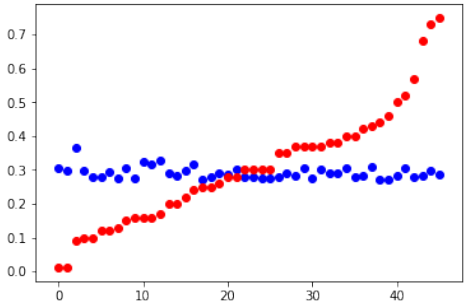
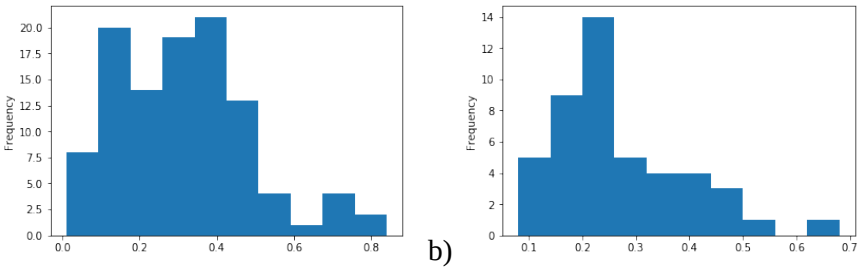


Figure 4. The result of the classifier by gz5 (red – real water cut values, blue - predicted values).

A similar pattern is observed for other logging curves. Tests were conducted on the combination of some logging curves in the generation of training parameters. There was no significant change in the prediction results.

It can be assumed that the result is a consequence of the uneven distribution of water cut values in the training sample and the classifier seeks to predict the average water cut value in the input data (Figure 5).



a) b)

Figure 5. Histograms of water-cut values: for the training sample a) and for the control sample b).

It was decided to check how the classifier would behave on a more extensive training sample.

4. Modeling/simulation of the logging data by the spectral method

One of the main requirements for the application of machine methods task is a large set of training samples. Possible solutions to the problem of expanding the training sample:

- Modeling of logging data by spectral method and interpolation (approximation) of the water cut map. Articles [9 – 11] describe the approach used to generate logging data of a field.
- Use of logging data and water cut for several "similar" fields (for example, all fields of Western Siberia).

In this paper, the first solution was used.

To generate target water cut values, a water cut map is required. It was obtained by logarithmic approximation of the initial water cut data provided.

The described scheme for obtaining logging curves data and the water-cut map is shown in Figure 6.

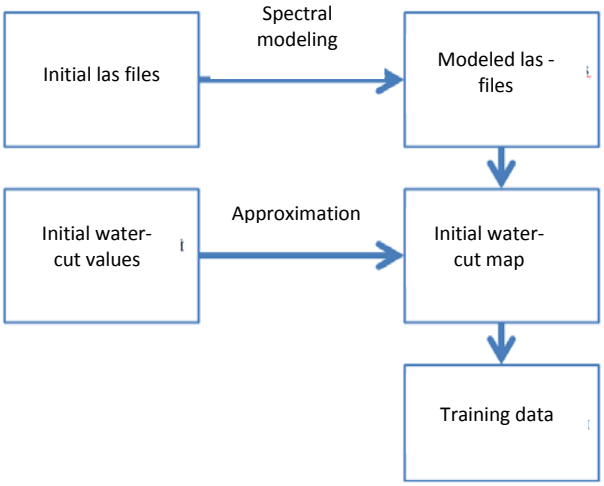


Figure 6. The generation of the extended training sample.

The results of the logging data obtained by spectral modeling are shown in Figure 7 (the original logging is shown in blue, modeled logging is shown in red) the well 105 did not participate in the modeling, and as a result was correctly predicted.

As a result of the simulation, the training sample was expanded to 5349 wells. That is, extended data were obtained for training the classifier.

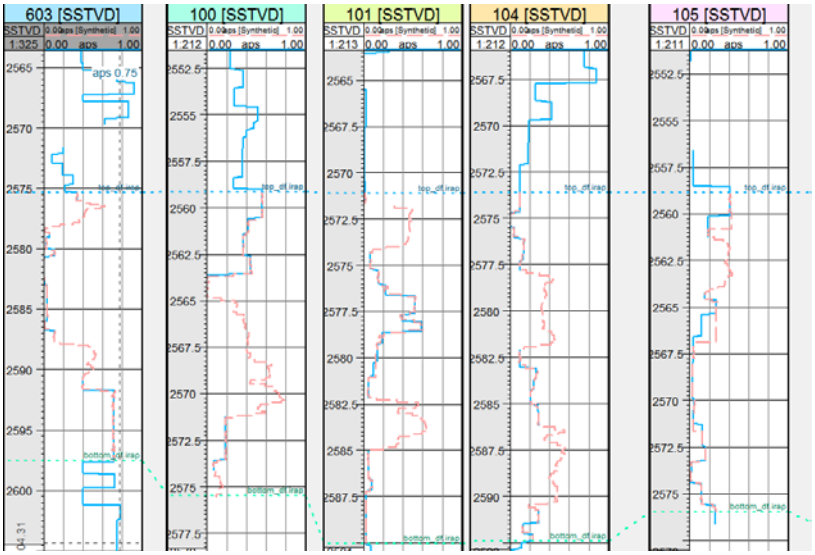


Figure 7. Results of spectral modeling on the aps curve.

4.1.1. Training on an extended training sample

In this sample, piecewise linear interpolation was used to generate features. The aps and kgl curves from well log interpretation data were considered as initial data, since it was not possible to analyze and expand the sample from well logging data within the framework of the task.

The ensemble of trees with search of optimal settings for the method and multilayer neural network were considered as a classification tool.

The structure of the considered multilayer neural network is as follows

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 250)               25250
_________________________________________________________________
dense_2 (Dense)              (None, 200)               50200
_________________________________________________________________
dense_3 (Dense)              (None, 100)               20100
_________________________________________________________________
dropout_1 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_4 (Dense)              (None, 100)               10100
_________________________________________________________________
dense_5 (Dense)              (None, 49)                4949
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 50
=================================================================
Total params: 110,649
Trainable params: 110,649
Non-trainable params: 0
```

The trained model was tested on three types of data: control sample (selected from the extended training set), *wc* values on the approximated map (for initial wells) and real *wc* values (for initial wells).

Below are the results of the analysis of the classifiers. Red colour indicates expected values, blue colour indicates predicted values. To clarify the result, the following values are additionally attached to each chart: ‘MSE’, ‘MAE’, ‘R2 score’, ‘Explained variance score’:

- MSE – mean square error.

$$\mathrm{MSE} = \frac{\sum_{i=0}^{N-1}\left(wc_{cor} - wc_{predict}\right)^2}{N}$$

- MAE - mean absolute error, described in 3.2.1.
- R2 score – coefficient of determination, is an indicator of the quality of the regression model. A value of 1 represents the ideal predictive ability, and a value of 0 represents the constant of the model that predicts the average of the responses.

$$R^2 = 1 - \frac{\sum_{i=0}^{N-1}\left(wc_{cor} - wc_{predict}\right)^2}{\sum_{i=0}^{N-1}(wc_{cor} - wc_{mean})^2}$$

- *Explained variance score* – explained variation.

$$\text{explained}_{\text{variance}} = 1 - \frac{Var\{wc_{cor} - wc_{predict}\}}{Var\{wc_{cor}\}}$$

In the above formulas: where wc_{cor} – expected water cut value, $wc_{predict}$ – water cut value predicted by the classifier, N – number of wells in the control sample, $Var\{\}$ – dispersion.

The results of classification on the aps data (ensemble of trees) are shown in Figure 8.

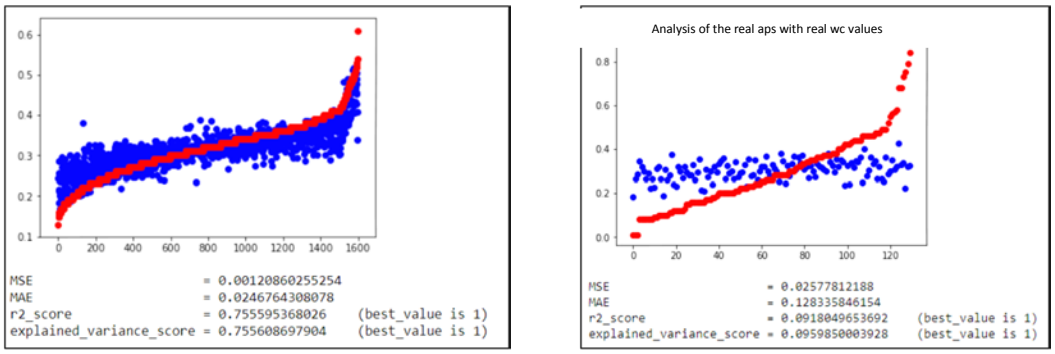


Figure 8. Results of the classification by the ensemble of trees with feature generation by piecewise interpolation method based on aps data.

The results of classification on the aps data using a neural network are shown in Figure 9.

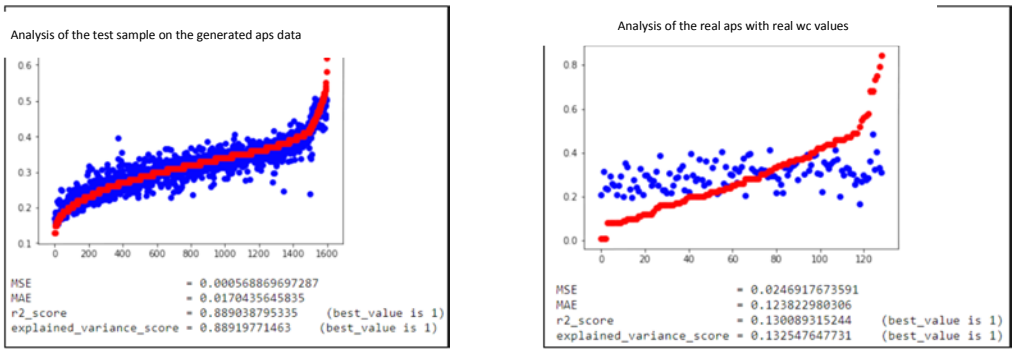


Figure 9. Results of the classification by neural network with feature generation by piecewise linear interpolation method based on aps data.

Since the results of the classifiers' work by the ensemble of trees and by neural networks differ slightly, on the example of aps, it was decided to carry out further checks by any of the classifiers. The results of classification on kgl data using a neural network are shown in Figure 10.

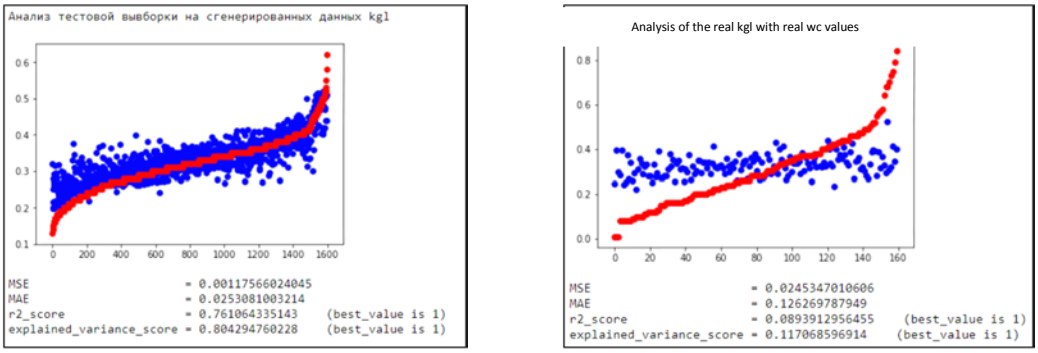


Figure 10. Results of the classification by neural network with the feature generation by piecewise linear interpolation method based on kgl data.

The results of classification on a combination of aps and kgl data were analyzed (Figure 11):

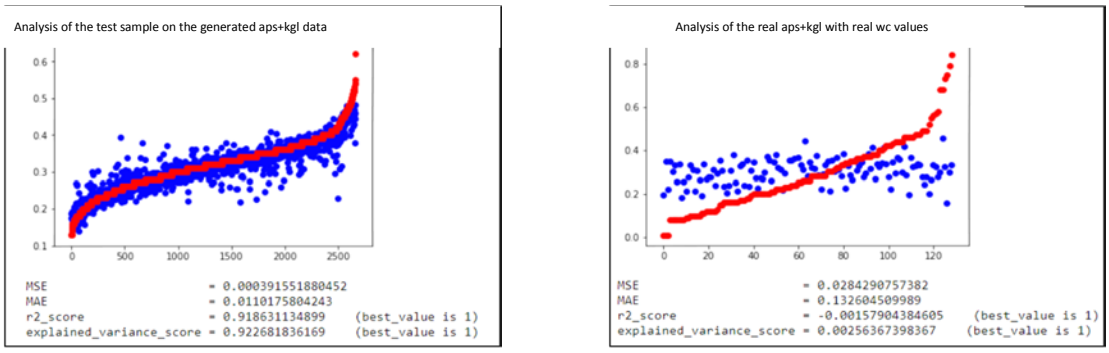
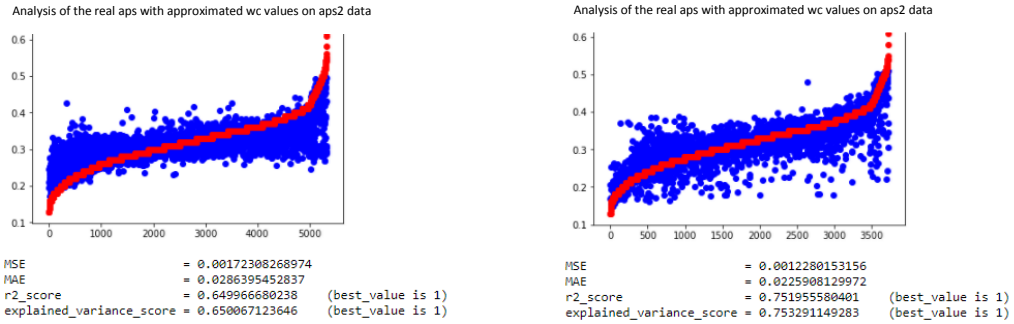


Figure 11. Results of the classification by neural network with the feature generation by piecewise linear interpolation method on combined kgl and aps data.

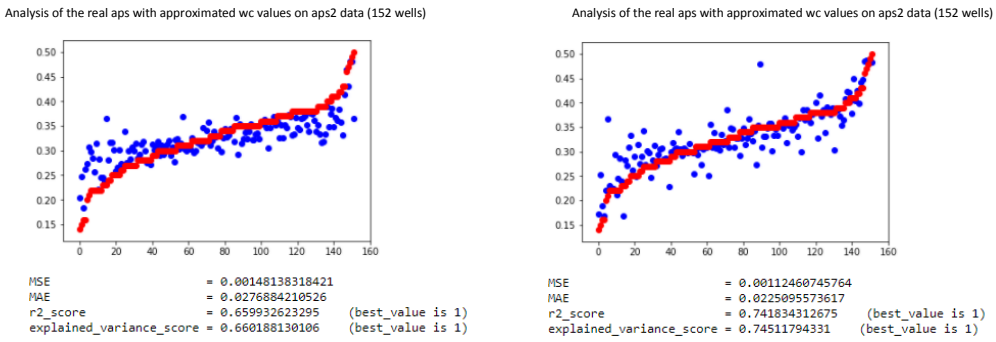
As we can see, on the real data, all models showed the result of the same order (on the extended data, the combination of kgl and aps showed the best prediction, r2_score = 0.92 on the control sample). High accuracy on the control sample, and much worse prediction on real wells. For example, r2_score = 0.92 and r2_score = -0.07, respectively. Therefore, it was decided to check how the classifier trained on extended data works for the data obtained from a different spectral experiment, but with the same initial data (Figure 12). From Figure 12 and Figures 7 – 8, it follows that on the data of the same type (modeled by the spectral method), the classifier shows similar results.

For the neural network r2_score = 0.89 and r2_score = 0.75, on the first and second spectral experiments, respectively.



a) b)

Figure 12. Results of classification (a) by an ensemble of trees, (b) by a neural network with feature generation by piecewise linear interpolation method from aps data on part of the data of the second set of spectral modeling.



a) b)

Figure 13. Results of classification (a) by an ensemble of trees, (b) by a neural network with feature generation by piecewise linear interpolation method from aps data on part of the data of the second set of spectral modeling (152 wells).

The accuracy of the r2_score prediction is determined by the amount or quality of the data, so it was decided to test the classifiers on a small amount of test data generated for the aps2 experiment. Random 152 wells were selected. As follows from Figure 13 and Figure 12, the amount of data does not affect the estimate of the forecast accuracy (the values of the accuracy metrics coincide). It is possible that the outliers lie in a small vicinity of real wells, and if they are not submitted to the classifier, the accuracy of the prediction on the extended data will be higher.

The reasons that the classifier shows high accuracy on extended data and low accuracy on real data may be:

- Incorrect or incomplete selection of curves for features generation.
- Data pre-processing practically did not take into account the physical patterns of well log interpretation data.
- Method of features generation. Perhaps, the chosen method is sensitive to the shape of the curve, and therefore it reveals patterns of data generation by the spectral modeling method.

5. Conclusions

The paper deals with the main aspects of image processing and analysis in the study of the mechanism of corrosion damage:

- The analysis and pre-processing of well logging data. Rapid analysis of the correlation between the well log interpretation data and water cut did not reveal any visible patterns. The amount of data and the distribution of water cut values (see Fig. 4) is insufficient to configure the regression and multi-class classifier.
- Four methods of features generation were studied. piecewise linear interpolation, 2 methods of generating statistical features and DFT analysis. DFT analysis turned out to be poorly applicable (perhaps, the authors could not find the optimal coefficients that should be used for training). The method based on piecewise linear interpolation and methods based on statistical features showed almost identical results.
- Four methods of classification were considered: trees, random forest, gradient boosting and multi-layer neural networks. On the generated features all methods showed close results (except for a single tree), the ensemble of trees and neural networks had a slight advantage in training time.
- The classifiers are set up on the well log interpretation data and well logging data and it is concluded that they tend to predict the average water-cut value. Therefore, in order to increase the variety of data for training, the training sample was extended to well log interpretation data, using spectral modeling for logging data and water cut approximation.
- Classifiers were trained on the well log interpretation data extended sample, which showed high prediction accuracy on the control extended sample (one set for training and one set for establishing accuracy, about 5000 in each set, and training was done on 3500). However, the accuracy of these classifiers on the source data was low.

6. References

[1] Ilyushin P I, Galkin S 2011 Forecast of production watering in producing wells of the Perm region using analog-statistical methods *Bulletin of the Perm National Research Polytechnic University. Geology, oil and gas and mining* **10(1)** 76-84

[2] Deshenenkov I S 2013 Forecast of productivity and initial water cut of oil wells of one of the fields of western Siberia according to field geophysics data *Drilling and oil* **7(8)** 32-35

[3] Alekseev A D, Aniskin A A, Volokitin Ya E, Zhitnyy M S and Khabarov A V 2011 Experience and Prospects for the Use of Modern GIS and Gdis Complexes in the Salym Group Deposits *Production engineering and technical oil and gas magazine "Engineering practice"* **11-12** 62-75

[4] Karp V P, Sayapina Yu A, Khetagurova L G and Botoeva N K 2012 Building decision rules in the study of the dynamics of cosmophysical indicators in order to predict situations that provoke the occurrence of stroke episodes *Health and education in the XXI century* **14(1)** 221-222

[5] Kozhevnikov A V, Ilatovsky I S and Solov'eva O I 2017 The use of machine learning methods in predicting the state of electromechanical systems of rolling production *Bulletin of Cherepovets State University* **1** 33-39
[6] Gerasimov S V, Meshcheryakov A V 2017 Application of the Microsoft Azure HDInsight platform for processing and analyzing large astronomical data arrays *International Journal of Open Information Technologies* **5(1)** 81-87
[7] Krasnov F V, Butorin A V and Sitnikov A N 2018 Automated detection of geological objects in seismic field images using deep learning neural networks *Business Informatics* **2** 7-16
[8] Enikeev M R, Gubaidullin I M and Maleeva M 2017 A Information-computational analytical system for the assessment and prediction of corrosion processes on the surface of steel and aluminum *Informatics systems and tools* **27(3)** 155-170
[9] Baikov V A, Bakirov N K and Yakovlev A A 2010 New approaches in the theory of geostatistical modeling *Bulletin of Ufa State Aviation Technical University* **14(2)** 209-215
[10] Baikov V A, Bochkov A S and Yakovlev A A 2011 Accounting for heterogeneity in geological and hydrodynamic modeling of the Priobskoye field *Oil industry* **5** 50-54
[11] Khasanov M M, Belozerov B V, Bochkov A S, Ushmaev O S and Fuks O M 2014 Application of spectral theory for the analysis and modeling of reservoir properties of a reservoir *Oil industry* **12** 60-64