ANALYSIS OF CONTROLLING FACTORS FOR HYDRAULIC FRACTURING PARAMETERS AND ACCUMULATED PRODUCTION USING MACHINE LEARNING

by

Zhihua ZHU[a], Maoya HSU[b], Chang LI[a], Jiacheng DAI[b], Bobo XIE[a], Zhengchao MA[b], Tianyu WANG[b*], Jie LI[a], and Shouceng TIAN[b]

[a] Research Institute of Engineering Technology,
PetroChina Xinjiang Oilfield Company, Karamay, China
[b] National Key Laboratory of Petroleum Resources and Prospecting,
China University of Petroleum, Beijing, China

Original scientific paper
https://doi.org/10.2298/TSCI230726039Z

This study, based on static data from over a thousand fracturing wells, employs data governance, data mining, and machine learning regression uncover principal controlling factors for production in the fracturing context. Preprocessing methods, including outlier identification, missing value imputation, and label encoding, address the field data challenges. Correlations among geological, engineering, and production parameters are analyzed using Pearson coefficient, grey correlation, and maximum mutual information. The AutoGluon framework and SHAP post-explanation method compute feature importance. Utilizing multiple evaluation methods, the entropy weight method comprehensively scores and ranks the principal controlling factors. A machine learning production prediction model is established for validation. Results show that DBSCAN achieves better accuracy in identifying field anomaly data.

Key words*: hydraulic fracturing design, data-driven model, machine learning, feature selection, hyperparameter optimization, production forecasting*

Introduction

Horizontal well fracturing is crucial for unconventional oil and gas development, necessitating accurate post-fracturing production prediction for optimal parameter adjustments [1-4]. However, mechanistic models are time-consuming, hindering on-site development and rapid optimization, particularly in heterogeneous reservoirs [5]. With 1226 fracturing operations in Xinjiang Oilfields, significant data is available for big data research. This study employs big data and artificial intelligence to manage on-site data, analyze production-controlling factors, and establish machine learning models for efficient fracturing design in unconventional oil reservoirs. Researchers have employed correlation algorithms and model evaluation methods to assess the relationship between well parameters and production, identifying principal controlling factors [6]. Machine learning techniques, such as dimensionality reduction and feature synthesis, capture non-linear connections between geological engineering parameters and production [7]. Various algorithms, including random forest and Locally Preserving Projections, have been applied for non-linear analysis and unconventional oil and gas production forecasting, demonstrating good applicability [8].

* Corresponding author, e-mail: wangty@cup.edu.cn

In summary, machine learning-based analysis of principal controlling factors and production forecasting has shown initial effectiveness. However, challenges such as small sample size and limited parameter ranges persist. This paper collects fracturing well data, conducts data governance and mining, comprehends principal controlling factors, and builds a machine learning production forecasting model based on field data from Xinjiang Oilfield, contributing to improved predictions and optimization.

Material and method

Data collection

This study compiled a dataset from Xinjiang Oilfields, consisting of 1226 hydraulically fractured horizontal wells with 44 feature parameters. After preliminary cleaning, wells with a single well data missing rate exceeding 40% and parameters with a data missing rate exceeding 80% were excluded. The dataset includes geological, engineering, and production data. Parameters such as horizontal section length, porosity, permeability, oil saturation, reservoir length, and reservoir type constitute geological and engineering data. Engineering data include modified section length, number of fractured sections, number of fracture clusters, liquid intensity, proppant intensity, pre-flush fluid ratio, slip water ratio, maximum construction displacement, and average sand ratio. Geological and engineering data serve as input features during model training, while production data mainly involve cumulative oil production over 330 days per unit of modified section length, serving as the output target during model training. The data were normalized to the [0, 1] interval, and kernel functions were employed to describe the data distribution, as shown in fig. 1.

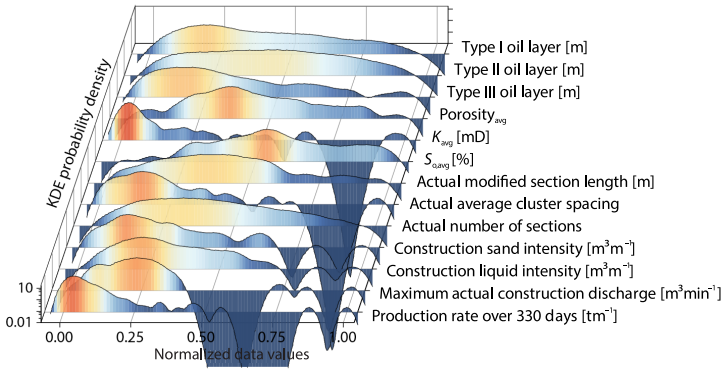


Figure 1. Waterfall plot depicting the kernel density distribution of selected data points

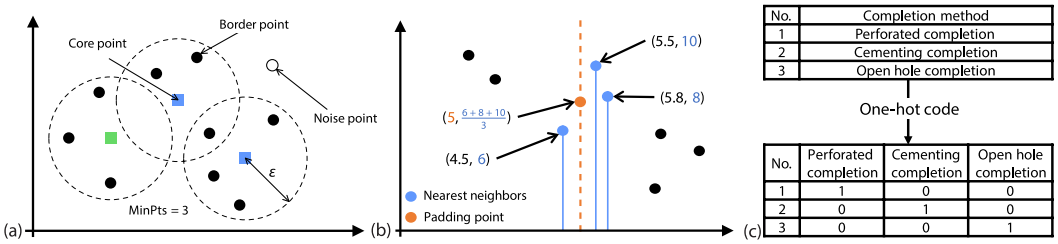
Upon observation of fig. 1, it is evident that the overall data distribution is uneven, with some data exhibiting a left-skewed pattern. Anomalies are particularly noticeable in parameters such as porosity, proppant intensity, and construction displacement, where significant outliers are present.

Data governance

Data governance encompasses outlier identification, missing value imputation, and categorical data encoding, as illustrated in fig. 2.

To handle outlier data, we used the DBSCAN unsupervised clustering method. The DBSCAN, a density-based clustering algorithm, categorizes and filters outliers by assessing

the density of neighboring data points. For missing data, various methods were employed, including zero filling, mean imputation, and KNN modelling imputation. The KNN imputation references known data from neighboring samples, improving imputation quality with different reference sample sizes. The impact of imputed data on subsequent model performance was observed through experimentation. For textual data, one-hot encoding converted textual label data into categorical labels for subsequent model predictions.



No.	Completion method
1	Perforated completion
2	Cementing completion
3	Open hole completion

No.	Perforated completion	Cementing completion	Open hole completion
1	1	0	0
2	0	1	0
3	0	0	1

Figure 2. Data governance methods; (a) DBSCAN outlier identification, (b) KNN missing value imputation, and (c) one-hot encoding of label data

Feature importance evaluation

This study employs grey correlation and maximum mutual information for correlation analysis, along with embedded feature evaluation and SHAP post-interpretation assessment for importance evaluation. Unlike typical research that often relies on a single method for analyzing main controlling factors, this study introduces the fuzzy mathematical entropy weight method. Utilizing results from various principal control analysis methods, it establishes an evaluation matrix for a comprehensive assessment of parameters through weighted calculations.

The entropy weight method is an objective weighting technique that determines the weights of indicators based on the information entropy of each indicator:

$$n_{ab} = \frac{e_{ab} - \min(E_b)}{\max(E_b) - \min(E_b)}, \quad P_{ab} = \frac{n_{ab}}{\sum_{a=1}^{n} n_{ab}}, \quad \sum P_{ab} = 1, \; r_b = \frac{-1}{\ln m} \sum_{a=1}^{n} P_{ab} \ln P_{ab} \tag{1}$$

where n_{ab} is the normalized score value of evaluation parameter a for the positive evaluation indicator b, E_b – the data vector of evaluation indicator b in the evaluation matrix, e_{ab} – the value of evaluation parameter a in evaluation indicator b within the evaluation matrix, P_{ab} – the weight of e_{ab} in evaluation indicator b, max(E_b) and min(E_b) are the max and min values of E_b, respectively, n – the number of evaluation parameters, m – the number of evaluation indicators, and r_b – the information entropy value of evaluation indicator b:

$$\omega_b = \frac{1 - r_b}{\sum_{b=1}^{m} 1 - r_b} \tag{2}$$

where ω_b is the entropy weight of evaluation indicator b.

Finally, the evaluation matrix E is multiplied by the indicator weight vector ω to obtain the comprehensive evaluation vector $S = [s_1,..., s_n]$:

$$S = E \times \omega = \begin{bmatrix} \zeta_1 & M_1 & J_1 & \varphi_1 \\ \ldots & \ldots & \ldots & \ldots \\ \zeta_n & M_n & J_n & \varphi_n \end{bmatrix} \begin{bmatrix} \omega_\zeta \\ \omega_M \\ \omega_J \\ \omega_\varphi \end{bmatrix} = \begin{bmatrix} \omega_\zeta \zeta_1 + \omega_M M_1 + \omega_J J_1 + \omega_\varphi \varphi_1 \\ \ldots \\ \omega_\zeta \zeta_n + \omega_M M_n + \omega_J J_n + \omega_\varphi \varphi_n \end{bmatrix} = \begin{bmatrix} s_1 \\ \ldots \\ s_n \end{bmatrix} \tag{3}$$

where $\zeta = [\zeta_1,..., \zeta_n]$ is the evaluation vector obtained from the grey relational analysis, $\mathbf{M} = [m_1,..., m_2]$ – the evaluation vector obtained from the maximum mutual information, $\mathbf{J} = [J_1,..., J_n]$ – the feature importance vector obtained from embedded feature evaluation, and $\boldsymbol{\omega} = [\omega 1,..., \omega_n]$ the evaluation vector obtained from the SHAP method.

This study utilized the AutoGluon machine learning framework for rapid model training andevaluation. AutoGluon achieves higher accuracy and faster predictions by integrating multiple models without the need for hyperparameter tuning.

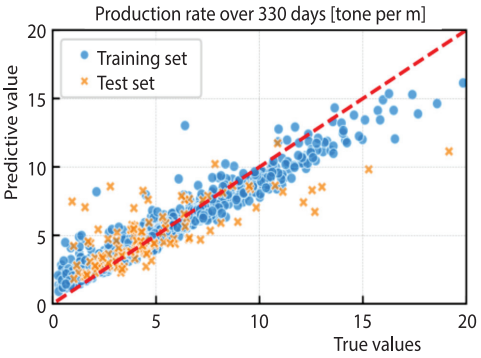


Figure 3. Comparative chart of autogluon prediction results

Results and discussion

Model performance

The training data underwent DBSCAN outlier removal, with no imputation for missing values. A test set, comprising 20% of the total dataset, was separated and excluded from training. Utilizing the AutoGluon framework, the training process involved 5-fold cross-validation bagging and a 2-layer stacking approach, using RMSE as the loss function. The training results, as depicted in fig. 3, show an average error of 0 and an R^2 of 1 for the training set. For the test set, the RMSE is 2.630, MS*E* is 6.919, MAE is 1.588, and R^2 is 0.768.

Comparison of data governance methods

Due to the limited number of training data samples, a large number of features, low data quality, and a high proportion of missing values, the data governance methods sig-nificantly affect the model results. The dataset was managed using DBSCAN, three standard deviations, and IQR methods to identify outliers, along with KNN nearest neighbor pre-diction, constant, and mean value imputation for handling missing values or not han-dling them, respectively. The model training results are compared in fig. 4.

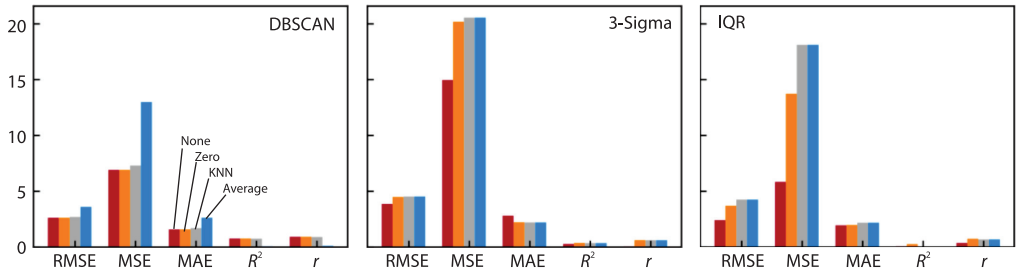


Figure 4. Test metrics for different data governance combinations

Analysis of controlling factors

The results of grey relational analysis, maximum mutual information, embedded feature evaluation, SHAP analysis, and entropy weight analysis are depicted in fig. 5.

In the experiment with varying input feature quantities, the model error is minimized when the feature quantity is set to the top four parameters, as illustrated in fig. 6(a). As the feature quantity further increases, the model error gradually increases. For different feature

evaluation methods, subsets of data were created by taking the top four parameters and inputting them into the model for training. The model accuracy variations are depicted in fig. 6(b). The results indicate that the combination of main controlling factors evaluated through entropy weight analysis exhibits lower error and better stability during model training.

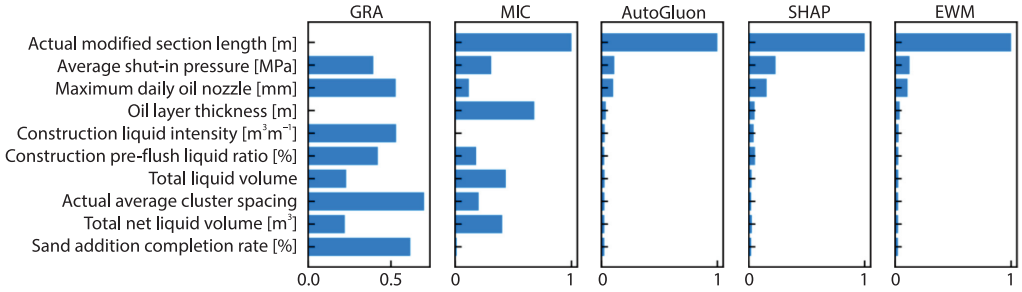


Figure 5. Evaluation of parameter importance

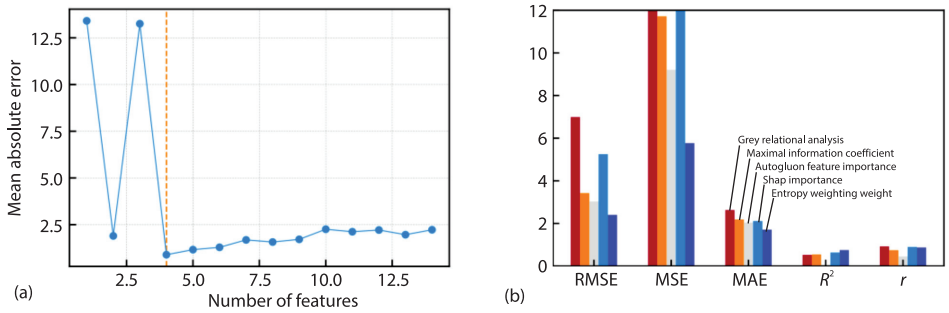


Figure 6. Impact of different principal evaluation methods on model accuracy; (a) number of features and (b) RMSE, MSE, MAE, R^2, and r

Conclusion

This research effectively utilizes large-scale field data to extract key parameters affecting production, providing technical support for the establishment of high precision prediction models and the optimization of parameters for unconventional reservoir production forecasts.

Nomenclature

a – evaluation parameter, [–]
b – evaluation indicator, [–]
r_b – the information entropy value, [–]

Greek symbol

ω_b – the entropy weight, [–]

Acknowledgment

The authors would like to thank the financial project supported by the National Natural Science Foundation of China (Grant/Award No. 5231001009).

References

[1] Liu, H., *et al.*, Application Status And Prospects of Artificial Intelligence in the Refinement of Waterflooding Development Program, *Acta Petrolei Sinica*, *44* (2023), 9, pp. 1574-1586
[2] Li, Y., *et al.*, Application Status and Prospect of Big Data And Artificial Intelligence in Oil and Gas Field Development, *Journal of China University of Petroleum (Edition of Natural Science)*, *44* (2020), 4, pp. 1-11

[3] Kuang, L., *et al*., Application and Development Trend of Artificial Intelligence in Petroleum Exploration And Development, *Petroleum Exploration and Development*, *48* (2021), 1, pp. 1-11
[4] Sheng, M., *et al.*, Research Status and Prospect of Artificial Intelligence in Reservoir Fracturing Stimulation, *Drilling and Production Technology*, *45* (2022), 4, pp. 1-8
[5] Anton, D. M., *et al*., Data-Driven Model for Hydraulic Fracturing Design Optimization: Focus on Building Digital Database And Production Forecast, *Journal of Petroleum Science and Engineering*, *194* (2020), 2, ID107504
[6] Yue, X., *et al*., Research on Main Control Factors Influencing Fracturing Effect of Jiaoshiba Area Based on Grey Relational Analysis, *Proceeding*s, ARMA-CUPB Geothermal International Conference, Beijing, China, 2019
[7] Song, X., *et al*., Productivity Forecast Based on Support Vector Machisne Optimized by Grey Wolf Optimizer, *Lithologic Reservoirs*, *32* (2020), 2, pp. 134-140
[8] Zhang, Y., *et al.*, Application of Locality Preserving Projection-Based Unsupervised Learning in Predicting the Oil Production For Low-Permeability Reservoirs, *SPE Journal*, *26* (2021), 3, pp. 1302-1313

Paper submitted: July 26, 2023
Paper revised: November 10, 2023
Paper accepted: December 11, 2023