ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

Analysis of Main Control Factors of Horizontal Well Production Capacity based on Machine Learning Algorithm

Haibiao Wang, Yancai Gao, Zheng Yuan, Kaixing Li, and Mingyue Pang

China Oilfield Services Limited, Tianjin 30000, China

Abstract

Accurately identifying the main controlling factors of the fracturing effect of horizontal wells in tight gas reservoirs, and then effectively guiding the optimisation of the fracturing scheme, is the key to enhance the fracturing capacity of horizontal wells in tight gas reservoirs. Relying on geological and engineering data of Block A of a tight gas reservoir, four algorithms, namely, Spearman coefficient, maximum mutual information coefficient, Copula entropy and grey correlation, were used to identify the main controlling factors affecting the fracturing effect of horizontal wells under different completion methods, and the weights of the four algorithms were weighted and combined in order to reduce the contingency of the evaluation results of a single weight, and the geological and engineering big data of the 126 barehole horizontal wells and 21 cased horizontal wells of this block were summed up. The study summarises the relationship between the geological and fracturing parameter characteristics of 126 openhole horizontal wells and 21 cased horizontal wells in the block and the test production of a single well, and puts forward suggestions for efficient development of horizontal wells at a later stage. The study shows that the formation pressure coefficient is the main controlling factor for determining the single-well productivity of horizontal wells in this zone, and the productivity of barehole completed horizontal wells is controlled by the fracturing fluid return rate, effective reservoir encounter rate and mud content, while the productivity of cased horizontal wells is controlled by the number of fractured sections, effective reservoir encounter rate and average section length.

Keywords

Tight Gas Reservoir; Machine Learning Algorithm; Horizontal Wells; Production Capacity.

1. Introduction

The analysis of factors affecting gas well production capacity points out the direction of how to develop the gas reservoir efficiently, clarifies the relationship between various factors and production capacity, and has guiding significance for the deployment of new wells and development adjustment in the later stage of the study area, which is conducive to proposing the optimisation technical countermeasures for horizontal well development [1]. Experts and scholars at home and abroad have also done a lot of research on the main control factors of horizontal well production capacity in different oil and gas reservoir types. Zhao Hongtao et al [2] used grey correlation method to analyze the correlation between each parameter and the specific oil recovery index (SORI), to determine the main control factors and to establish a prediction model of SORI, which provides a certain guiding basis for the preferential selection of the test layer of the exploratory wells in the regional thick oil reservoirs and the production capacity estimation before the test. Xue Ting et al [3] used grey correlation method and random forest algorithm to systematically analyze the degree of influence of geological, fracturing construction and other parameters on the production capacity, clarified the main

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

controlling factors of single-well production capacity, and optimized the deployment of geological wells and fracturing construction parameters.

However, at present, the analysis and evaluation of gas well productivity are all for the same completion method, and the change of completion method in the same block may lead to inconsistency in the main controlling factors of horizontal well productivity, resulting in the mismatch between the fracturing construction plan in the gas field site and the geological conditions of the reservoir, which restricts the gas field from realising stable production and efficient development. Therefore, there is an urgent need to clarify the indicators of gas well productivity under different completion methods, and to find out the influence of geological and engineering parameters with good correlation on productivity. In this paper, Spearman's correlation coefficient, maximum mutual information coefficient, Copula entropy and grey correlation method are used to analyze the weights of the factors influencing the production capacity of horizontal wells under different completion methods in the target block, and the weighted combination of the weights of the four algorithms is used to reduce the contingency of a single weight on the evaluation results, to sum up the laws of the influence of the geological and construction parameters of the block on the production capacity, and to choose the optimal values of the parameters. The results of the four algorithms are combined to reduce the chance of single weighting on the evaluation results, summarize the influence of geological and construction parameters on gas production, and select relatively better parameter values.

2. Methodology for Analysing the Main Control Factors Affecting Production Capacity

There are many factors affecting post-fracturing capacity, and the degree of influence often varies greatly. Clarifying the degree of influence of each factor on post-fracturing capacity is a prerequisite for the optimisation of fracturing process. In this paper, we propose to use multifactor analysis to analyse the degree of contribution of reservoir geology and construction parameters to gas production in a block. There are various methods being used to measure the correlation between sample characteristics, and common correlation metrics are as follows.

2.1 Spearman's Coefficient Method

Spearman correlation coefficient has the advantages of being independent of the data magnitude and insensitive to abnormally large numbers, etc. Spearman correlation coefficient is calculated as follows:

$$W_B(i)=\frac{\sum_{n=1}^{N}\left(X_n-\bar{X}\right)\left(Y_n-\bar{Y}\right)}{\sqrt{\sum_{n=1}^{N}\left(X_n-\bar{X}\right)^2\sum_{n=1}^{N}\left(Y_n-\bar{Y}\right)^2}} \tag{1}$$

The value of Spearman correlation coefficient is between -1 and +1; -1 means that the two variables are completely negatively correlated, +1 means that the two variables are completely positively correlated, and 0 means that the two variables are completely unrelated; the closer to 1 or -1 means that the stronger the linear correlation of the two variables.

Using this coefficient to select the best comprehensive evaluation method, it is necessary to select a group of samples first, and at the same time, establish the criteria for ranking the samples in a reasonable level of comprehensive evaluation. Then according to the different evaluation methods on the sample of different rank ordering and reasonable rank ordering of the degree of correlation between the rank of the Spearman coefficient size to select the best 错误!未找到引用源。.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

2.2 Maximum mutual information factor

Maximal Information Coefficient (MIC) is a non-parametric method proposed by Reshef as a tool for exploratory data analysis. Compared to traditional linear correlations or exponential relationships, MIC is capable of discovering a wide enough range of correlations on a sufficiently large sample set and is not limited by specific relationship types. The core idea is to encapsulate the dataset from block to block through scatterplot chunking. If there is a correlation between variables, a grid can be plotted on the scatterplot, exhausting all the possible methods of grid delineation, calculating mutual information under each delineation and using this as the information coefficient, and taking the maximum value as the maximum information coefficient. Due to its advantages of adaptivity, non-linearity and freedom from distributional assumptions, it has been widely used in data analysis in bioinformatics, medicine, finance and other fields.

For a bounded set $D \subset R^2$ and $n_x, n_y \in N^*$, definition:

$$I^*\left(D, n_x, n_y\right) = *max\left\{I\left(D\middle|_G\right)\right\} \tag{2}$$

In the formula, I^* denotes the maximum value of mutual information of grid G for column n_x and row n_y, and $I\left(D\middle|_G\right)$ denotes the value of mutual information of D under grid G partitioning.

Define the identity matrix of the bivariate data set:

$$M\left(D\right) = \frac{I^*\left(D, n_x, n_y\right)}{\log_2 * min\left\{n_x, n_y\right\}} \tag{3}$$

In the formula, I is the maximum value of M. For the bivariate data set D and the number of samples n, the maximum information coefficient can be expressed as formula (3), which is normalised to give the weights of the influencing factors.

$$MIC\left(D\right) = *max_{n_x \times n_x \le B(n)}\left\{M\left(D\right)\right\} \tag{4}$$

2.3 Copula Entropy

Copula function is a function used to describe the dependence between multidimensional random variables and is independent of the marginal distribution function, which is defined as follows:

$$C(u1, u2, ..., un) = P(X1 \le F1^{-1}(u1), ..., Xn \le Fn^{-1}(un)) \tag{5}$$

In the formula, where $X1$, ... Xn are n random variables, $F1$, ... Fn are their marginal distribution functions, and $u1$, ... un are coordinates on the unit hypercube.

Copula entropy is based on the concept of Copula function. It can be used to measure the complexity of dependencies between multidimensional random variables.Copula entropy is defined as follows:

$$\mathrm{H(C)} = -\int ... \int C(\mathrm{u1}, ..., \mathrm{un}) \log C(\mathrm{u1}, ..., \mathrm{un})\, \mathrm{du1} \ldots \mathrm{dun} \tag{6}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

In the formula, C is the Copula function and $H(C)$ is the Copula entropy.

2.4 Grey Correlation Method

Grey correlation is a method used to study the correlation between factors. The basic idea is to convert the relationship between factors and targets into similarities between factors, and then to derive the grey correlation between each factor and the target factor by comparing the similarities between the factors. Grey correlation analysis can be used to deal with various types of data, and its main advantage is that it can be analysed in the case of lack of data or incomplete data, and it does not need to carry out complex statistical processing such as hypothesis testing or parameter estimation on the data, and the specific application steps are as follows.

(1) Take the unit reservoir average daily pressure drop production as a reference series, and the influence factors as a comparison series, and carry out dimensionless processing.

(2) Calculate the grey correlation coefficient between the comparison series and the reference series using equation (7).

$$\xi_{0i}(k)=\frac{\Delta_{\min}+\rho\Delta_{\max}}{\Delta_{0i}(k)+\rho\Delta_{\max}} \tag{7}$$

In the formula, $\Delta_{\max}$ and $\Delta_{\min}$ represent the absolute difference between the largest and smallest sample data in the comparative series respectively, and represent the absolute interpolation of the corresponding sample values in the comparative series and the reference series. ρ is the resolution coefficient, ranging from 0 to 1, which is an important parameter to control the difference between the correlation coefficients, and is generally taken as 0.5. The following principles are followed in taking the value: firstly, the resolution coefficient is dynamically taken according to the actual situation of the sequence; secondly, when there is a singular value in the sequence, the resolution coefficient is taken as a small value in order to reduce the influence of the singular value; thirdly, when the sequence is relatively smooth, the resolution coefficient is taken as a large value to reflect the overall correlation of the correlation coefficient. overall nature of the correlation. According to this principle, while considering the outliers dominating the system correlation value, the mean value of the difference of all absolute values can be expressed by Δ_y.

$$\Delta_y=\frac{1}{n\cdot m}\sum_{i=1}^{m}\sum_{k=1}^{n}\left|Y_i(k)-Y_0(k)\right| \tag{8}$$

In the formula, n and m are the number of samples and the number of influences, respectively, noting that $\in_{\Delta}=\frac{\Delta_y}{\Delta_{*max}}$, when the relationship $\Delta_{\max}\le 3\Delta_y$ is satisfied, $\in_{\Lambda}\le\rho\le 1.5\in_{\Lambda}$; when $\Delta_{\max}>3\Delta_y$, $1.5\in_{\Lambda}\le\rho\le 2\in_{\Lambda}$.

(3) Calculating the grey correlation γ_{0i}:

$$\gamma_{0i}=\frac{1}{n}\sum_{k=1}^{n}\xi_{0i}(k) \tag{9}$$

(4) Calculating the weights and normalising the correlations gives the correlation weights for each comparison series $W_G(i)$:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

$$W_G(i) = \frac{\gamma_{0i}}{\sum_{i=1}^{m} \gamma_{0i}} \tag{10}$$

3. Analysis of Factors Affecting the Capacity of Horizontal Wells

3.1 Data Preprocessing

The data collected in this paper are from the actual production of the Su53 block. Due to the differences in the data records of the same block and the presence of missing values, or outliers in the actual production data, it is not possible to train directly. Therefore, before analysing, it is necessary to carry out data cleaning and other operations to obtain higher precision data collection:

(1) Missing value processing

Currently, there are two main types of missing value processing methods:

1) Directly delete sample groups or features containing missing values. If a group of samples or a feature has too much missing data, the group of samples or the feature value is deleted.

2) Fill in the sample groups or features containing missing values. Specific filling methods include: ① plurality, median and mean filling method. ② neighbouring values to fill, generally using the data before and after the missing value to fill. ③ Predictive modelling methods for filling, building corresponding machine learning models for missing value prediction filling. ④ Lagrangian and other interpolation methods for filling. ⑤ KNN algorithm for filling, by comparing the corresponding features in the complete dataset and the missing data, and calculating the distance between the missing data and each sample in the complete dataset, then the missing data value is obtained by averaging multiple samples with the smallest distance. In this case, the sample distance is calculated as follows:

$$d(\overline{p},\overline{q}) = \sqrt{\sum_{i=1}^{m} (\overline{p}_i - \overline{q}_i)^2} \tag{11}$$

In the formula, $d(\overline{p},\overline{q})$ -- the distance between two samples; $\overline{p}_i$, $\overline{q}_i$ -- the corresponding point data of different samples.

According to the above two processing methods of missing values, combined with the characteristics of gas field data and the specifics of the collected field data of fractured horizontal wells, the steps of processing missing values in this paper are as follows:

1) Since the two production indicators, gas production and water saturation, are affected by multiple factors, there is a certain degree of randomness in processing the missing values of these two production indicators purely from the perspective of data. Therefore, considering the accuracy of the results, if a group of samples is missing the two production indicators of gas production and water saturation, the group of samples will be deleted directly.

2) Features with more than half of the missing values in the original data are deleted.

3) Considering that there is not much difference in data such as layer conditions and production system between horizontal wells in the same block. Therefore, in this paper, the vacant values of features such as mud content, return rate, total porosity and total fluid volume in the collated data samples are filled by KNN. The figure reflects the principle of the KNN algorithm, whose basic idea is to find sample states in the data that are similar to the current state, and apply the sample states that match the current state to the prediction. Although the KNN method is simple and effective, the value of K also affects the final result at the same time. Therefore, the data is fitted by using different values

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

of K while utilising the Random Forest algorithm, which has a strong ability to perform in small samples.

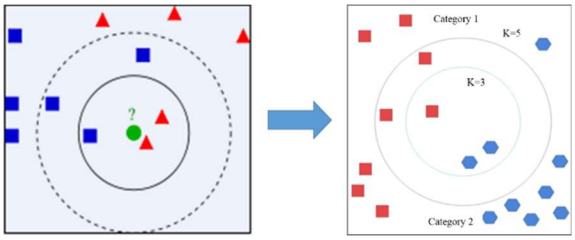

Figure 1. Principle of the KNN algorithm

(2) Outlier Handling

Before outlier processing, outlier detection should be performed first. The commonly used detection methods are as follows:

1) Lajda (3σ) criterion: assuming equal precision measurements of variables, if the residual error b of a given measurement x_b satisfies $|v_b| = |x_b - \bar{x}| > 3\sigma$, x_b is considered a bad value with a large error value and it is removed. In the formula, $\bar{x}$ is the arithmetic mean of the measured values.

2) Box plot analysis: In a box plot, an outlier is usually considered to be greater than the upper quartile + 1.5 times the interquartile spacing, or less than the lower quartile + 1.5 times the interquartile spacing. In this case, the upper quartile means that 1/4 of all values are greater than it; the lower quartile means that 1/4 of all values are less than it; and the interquartile spacing refers to the difference between the upper quartile and the lower quartile.

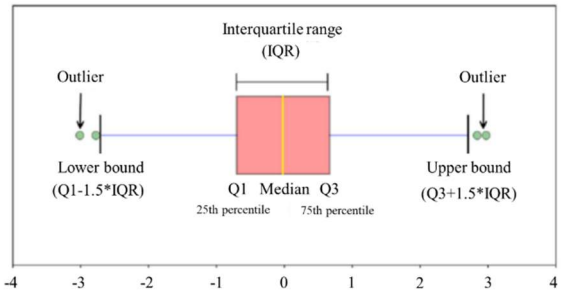

Figure 2. Principle of box-and-line diagram construction

3) According to the operating experience of field engineers, the range of all variable data is firstly sorted out, and then the outliers are identified according to the range of each variable.

Considering the specificity of data samples in the gas field field, this paper adopts the 2nd method for outlier identification. For a small number of data with excessive errors in recording, manual intervention is also required to ensure the screening accuracy of outliers. A total of 94 outlier samples are identified through this comprehensive method, and considering the small number of samples obtained from collation, the method of treating outliers as missing values and filling them in using KNN is adopted. After processing the anomalous values and missing values, the available samples of 147 wells were obtained from 241 horizontal wells in the block, of which 126 were barehole wells and 21 were cased wells, so as to establish a database for analysing the main controlling factors of the production capacity of this block.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

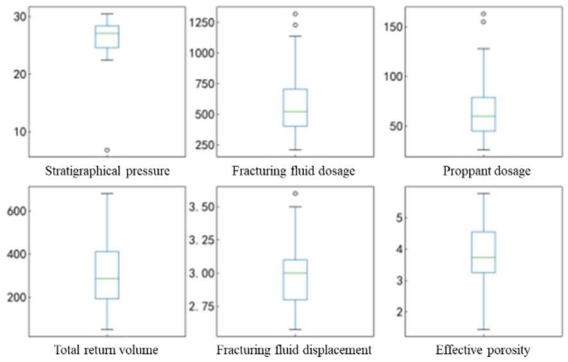

Figure 3. Identification of outliers in the dataset

3.2 Portfolio Weighting

The above four models can all solve the degree of contribution of each factor to gas production after fracturing of horizontal wells, in order to avoid the problem of chance caused by a single evaluation method, firstly, the impact weights of each parameter obtained by different evaluation methods are linearly normalised, and then each production impact indicator is combined, and the parameter weights are summed up according to formula (12) to obtain the final combined weights $W(i)$.

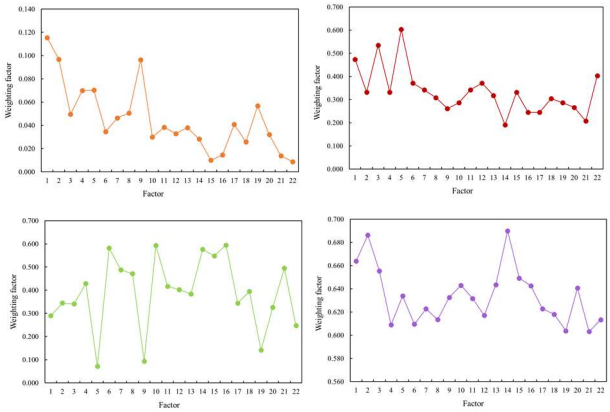

Figure 4. Weighting factors calculated by the four algorithms

$$W(i)=\frac{W_B(i)+W_D(i)+W_C(i)+W_G(i)}{\sum_{i=1}^{m}\left[W_B(i)+W_D(i)+W_C(i)+W_G(i)\right]} \tag{12}$$

3.3 Analysis of Main Control Factors

Based on the actual data of Su53 block in Ordos Basin, 126 barehole horizontal wells and 21 cased horizontal wells with relatively complete data are taken as the research objects to analyse the main influencing factors of horizontal well production capacity under different completion methods.

(1) Barehole Completion Horizontal Wells

According to the gas test data and model calculation weighting results, 10 parameters with relatively large combined weighting values (formation pressure coefficient, return rate, effective reservoir encounter rate, mud content, etc.) were selected for comparative analysis, as shown in Table I and Table II. The average weighting value of geological factors is larger than the average weighting value of engineering factors, which indicates that the geological factors have a higher degree of influence on the production than the engineering factors on the production.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

Table 1. Combined weights of different geological factors for horizontal wells with barehole completion

Factor	Spearman's coefficient	Maximum mutual information coefficient	copula entropy	Grey correlation	Combination weights
Stratigraphic pressure coefficient	0.160	0.078	0.067	0.047	0.352
Silt content	0.077	0.054	0.069	0.041	0.242
Total porosity	0.073	0.063	0.055	0.044	0.235
otal porosity	0.065	0.046	0.064	0.043	0.218

Table 2. Combined weights of different engineering factors for horizontal wells with barehole completions

Factor	Spearman's coefficient	Maximum mutual information coefficient	copula entropy	Grey correlation	Combination weights
Fracturing fluid return rate	0.126	0.053	0.034	0.037	0.249
Effective reservoir encounter rate	0.105	0.052	0.040	0.046	0.243
Fluid consumption of single section	0.053	0.056	0.038	0.046	0.193
Total sand volume	0.031	0.040	0.068	0.046	0.185
Proportion of precursor fluid	0.016	0.050	0.069	0.049	0.184
Total liquid nitrogen	0.064	0.044	0.029	0.043	0.180

(2) Casing Completion Horizontal Wells

Ten parameters (formation pressure coefficient, total porosity, number of fractured sections, effective reservoir encounter rate, etc.) with relatively large weight values are selected for comparison and analysed in Table III and Table IV, and the factors that have a greater influence on the production capacity of casing-completed horizontal wells are the formation pressure coefficient, the number of fractured sections, the effective reservoir encounter rate and the average section length.

Table 3. Combined weights of different geological factors for casing completion horizontal wells

Factor	Spearman's coefficient	Maximum mutual information coefficient	copula entropy	Grey correlation	Combination weights
Stratigraphic pressure coefficient	0.115	0.064	0.034	0.048	0.261
Total porosity	0.051	0.042	0.055	0.044	0.192

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

Table 4. Combined weights of different engineering factors for casing completion horizontal wells

Factor	Spearman's coefficient	Maximum mutual information coefficient	copula entropy	Grey correlation	Combination weights
Number of fractured sections	0.097	0.045	0.040	0.049	0.231
Effective reservoir encounter rate	0.050	0.073	0.040	0.047	0.209
Average section length	0.070	0.045	0.050	0.044	0.209
Fracturing fluid volume per stage	0.070	0.082	0.008	0.045	0.206
Total sand volume	0.035	0.051	0.068	0.044	0.197
Number of fractures	0.046	0.046	0.057	0.045	0.194
Horizontal section length	0.096	0.036	0.011	0.045	0.188
Percentage of pre-fracturing fluid	0.030	0.039	0.069	0.046	0.184

4. Conclusion and Recommendation

(1) The reliability of the weights of the factors influencing the production of tight gas wells after fracturing in the study block obtained by using the combined weighting method is better than that of the single weighting calculation method, which can effectively make up for the solution bias caused by the single method due to the limited data samples or differences in the method principles.

(2) The pattern of horizontal well production capacity under different completion methods and single fracturing influencing factors is not obvious from the analysis of field data, which indicates that the horizontal well production capacity is affected by a combination of many factors such as geological and engineering conditions.

(3) In this block, in order to obtain high production capacity, gas wells need to be geologically located in the zone of good physical properties and gas content, and at the same time, horizontal well development should be implemented according to local conditions, adopting a relatively reasonable method of reservoir modification and maximising the use of geological reserves.

References

[1] Zhao Hongbing. Research on the technology of judging the main control factors of horizontal well production capacity in volumetric fracturing of tight reservoirs[D]. Xi'an Petroleum University,2021.

[2] ZHAO Hongtao, YU Xi, YU Weiqiang, FAN Xinlei. Analysis and application of main controlling factors of thick oil reservoir capacity in Bohai Sea[J]. Complex Oil and Gas Reservoirs,2022,15(01):44-47.

[3] XUE Ting, HUANG Tianjing, CHENG Liangbian, MA Shuwei, SHI Jianchao. Optimisation of production control factors and development countermeasures for horizontal shale oil wells in Qingcheng Oilfield, Ordos Basin[J]. Natural Gas Geoscience,2021,32(12):1880-1888.

[4] LI Xianwen, WANG Lili, WANG Wenxiong, XIAO Yuanxiang, CHEN Baochun, ZHANG Yanming, ZHOU Changjing, MA Zhanguo, SHI Huaqiang. Key technological innovation and efficient development

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0027

practice of fracturing based on small borehole completion--A case study of tight gas reservoir in Surig gas field[J]. Natural Gas Industry,2022,42(09):76-83.

[5] Shang Weiping. Optimisation of comprehensive evaluation methods using Spearman's coefficient[J]. Jiangsu Statistics,1996(08):19-2.