Journal of Petroleum Exploration and Production Technology (2021) 11:911–923
https://doi.org/10.1007/s13202-020-01083-0

REVIEW PAPER-PRODUCTION ENGINEERING

Judgment method of working condition of pumping unit based on the law of polished rod load data

Chuanjun Han[1] · Yang Yue[1]

Received: 29 September 2020 / Accepted: 29 December 2020 / Published online: 8 January 2021

Abstract
At present, oil companies are committed to applying the theory and means of mathematics or data science to the research of oilfield data rules. However, for some old oil wells, aging equipment, complex environment and backward management, cause the authenticity and accuracy of the data collected by the equipment cannot be determined. According to the actual engineering demand of the old wells, this paper proposes a method based on principal component analysis, cluster analysis and regression analysis to mine and analyze the data of polished rod load of old oil wells, so as to judge the working conditions of the oil wells. Combined with the application of this study in several operation areas of some oilfields, the findings of this study can help for better understanding of the working condition information hidden in "big data" of oilfield. Meanwhile, the PCA method can reduce the complexity of the original data, the regression equation can calculate the size of the polished rod load more accurately, and the prediction model can effectively judge the working conditions of the old oil wells on site.

Keywords PCA · Cluster analysis · Regression analysis · Polished rod load · Threshold model · Working condition prediction

List of symbols

Xp	Data set of *p*-th original indexes.
Fm	Data set of m-th principal component indexes.
Emp	Data set of mp-th contribution rate.
λ_i	Data set of i-th principal component eigenvalue.
x_k	The k-th independent variable of regression equation.
y_i	The *i*-th dependent variable of regression equation.
β_k	The k-th coefficient of a linear function.
F_1	The ratio of static load sample in the regression equation of upstroke.
F_2	The ratio of dynamic load sample in the regression equation of upstroke.
F_3	The ratio of friction load sample in the regression equation of upstroke.
F_4	The ratio of static load sample in the regression equation of downstroke.
F_5	The ratio of dynamic load sample in the regression equation of downstroke.
F_6	The ratio of friction load sample in the regression equation of downstroke.
F_7	The ratio of the sample value of the downstroke static load to the predicted value of the maximum polished rod load on the upper stroke.
P_1	The confidence interval of F1.
P_2	The confidence interval of F2.
P_3	The confidence interval of F3.
P_4	The confidence interval of F4.
P_5	The confidence interval of F5.
P_6	The confidence interval of F6.
P_7	The confidence interval of F7

Introduction

The digitalization of the oil production process and the timely understanding of the operating conditions of the pumping units not only ensure the safety of oil production,

Supplementary information The online version contains supplementary material available at (https://doi.org/10.1007/s13202-020-01083-0).

✉ Chuanjun Han
hanchuanjun@126.com

Yang Yue
yueyangswpu@foxmail.com

[1] School of Mechatronic Engineering in Southwest Petroleum University, Chengdu 610500, China

but also accumulate a large number of complex original data for the oil field data information database. Therefore, how to extract useful rule information from massive data through data analysis is one of the main research directions of each oilfield (Chaodong et al. 2015). Since the 1990s, other domestic oilfields began to use the theories and methods of mathematics, economics, system theory and other disciplines to study the optimization of rational production and utilization of old oil wells and established corresponding evaluation and optimization models, such as gray system prediction model, fuzzy theory model and neural network model (Yuling and Jianping 2014; Canelon and Morles 2008).

Xinjiang oilfield has invested in digital oilfield for a long time, but the research effect of condition evaluation and efficiency optimization model is not ideal (Xiang 2007). After in-depth study of this problem, it is found that part of the reason is that the old oil wells which after decades of development, the production equipment of them has become less adaptable, resulting in the system's operating conditions are far more complex than the newly developed oil wells. As shown in Fig. 1, these old wells, which account for about 30–40 percent of the total number of wellheads in Xinjiang, are characterized by long development time, complex operation conditions, harsh environment and easy failure of digital equipment, high energy consumption and they produce a certain amount of oil but are not worthy of secondary investment. Due to the special geographical environment and limited management methods, some of the original data from these old oil wells transported to the oil field data information database are missing and inaccurate, which also makes the theoretical model of data science unable to be used well in this kind of oil wells (Gibbs and Neely 1966; Karmawijaya et al. 2009; Khakimyanov et al. 2015; Xuanyi et al. 2020).

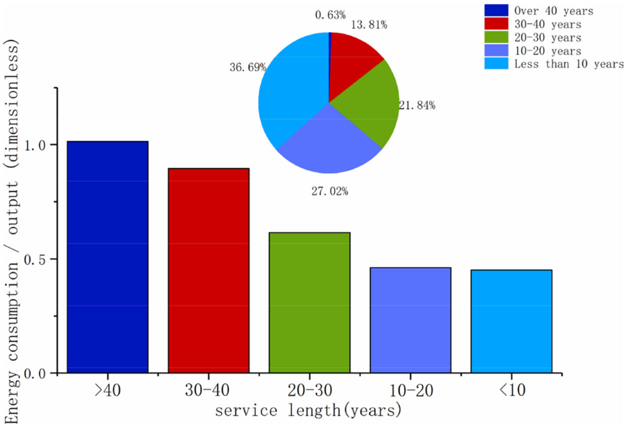


Fig. 1 Oil well exploitation length and the situation of energy consumption/output

For the newly developed oil wells, thanks to the high-quality information equipment, the measurable and perceived parameters are accurately collected and stored in the production process. These complete, reliable and valid data can easily find the underlying laws and information, and the processing methods even evolve from digital analysis to intelligent analysis. This is what researchers are more interested in. Difference from the above performed studies, it is clearly know that incomplete, unreliable data cannot be analyzed in the same way for older wells. Therefore, this study proposed a judgment method of pumping unit working condition obtained by means of statistical analysis such as PCA-cluster analysis and regression analysis, which provided a novel exploration idea for working condition evaluation and efficiency optimization model establishment of old oil wells (Jain et al. 2000; Stamatatos 2008; Wei et al. 2020).

The content of this paper includes two parts: the theoretical research process and the practical application of the research methods. First, the introduction presents the principles behind the problem under study. Then, "Principal component—cluster analysis of pumping units", "Establish the regression equation of polished rod load", "Working condition prediction model based on statistics" sections introduce the steps of the study work: 1. using the PCA- cluster analysis to cluster different working conditions to obtain the data set under normal working conditions; 2. using regression equation to get the most accurate calculation formula of polished rod load; 3. using statistical results to get the working condition prediction model. Finally, in the "Test" section, the theoretical method is applied to some operation areas of Xinjiang oilfield, and the experimental results are obtained.

Principal component—cluster analysis of pumping units

Evaluation index of dimension reduction function

Principal component analysis (PCA) is a statistical method that transforms a group of variables that may have correlation into a group of linearly uncorrelated variables through orthogonal transformation. The transformed variables are called principal components. It replaces the original index by recombining the original index into a new group of independent comprehensive indicators and takes a few less comprehensive indicators from it to reflect the information of the original index as much as possible. Through principal component analysis, the weight of each index can be determined objectively, and the deviation caused by subjective randomness can be avoided, and the multi-index variables can be scientifically evaluated (Xiaodong et al. 2011). In this part, the principal component analysis method is used to analyze the factors affecting the working conditions of pumping units, and the results of principal component analysis provide reasonable guidance information for the classification of pumping unit working conditions.

Suppose the original index variables are X_1,X_2, …, X_p, after principal component analysis, a set of principal component indexes $F_m(m<p)$ formed by linear combination of the original indexes was obtained. Among them, the first principal component F_1 is the direction corresponding to the largest data variation (contribution rate e_1) and contains the largest amount of information. If the first principal component is not enough to represent the information of the original p indicators, then consider selecting the second principal component index F_2 and so on to get the remaining principal component indicators. The expression formula of principal component analysis is as follows:

$$\begin{cases} F_1 = e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_p \\ F_2 = e_{21}X_1 + e_{22}X_2 + \ldots + e_{2p}X_p \\ \ldots\ldots \\ F_m = e_{m1}X_1 + e_{m2}X_2 + \ldots + e_{mp}X_p \end{cases} \tag{1}$$

According to the theoretical analysis, select p indexes that have influence on the working condition change of pumping unit, and m principal component indexes are obtained according to principal component analysis, and the comprehensive evaluation value F of pumping unit working condition is calculated by the weighted method of principal component eigenvalue λ_i [formula (2)].

$$F = \sum_{i=1}^{m} \lambda_i F_i / \left(\lambda_1 + \lambda_2 + \ldots + \lambda_m\right) \tag{2}$$

By using principal component analysis, many original indexes affecting working conditions can be reduced to several principal components. The factors influencing each other should be studied as a whole (principal component), rather than puzzled by the interaction between the original indicators. The higher the values of comprehensive evaluation value (*f*), the more highly correlated factors are contained in the principal component with the largest proportion of weight. Therefore, the determination of the comprehensive evaluation value (*f*) can show the complex situation of the mutual influence and weight proportion of the factors affecting the working condition of the pumping unit.

Determination method of abnormal working condition of pumping unit

In the actual production of oil field, the main basis for judging whether the working condition is normal or not is the graphic characteristics of indicator diagram. The transverse axis of the indicator diagram is the displacement of the polished rod, and the longitudinal axis is the load on the polished rod, and the lowest end of the polished rod is defined as the starting position. In a stroke of pumping unit, that is, the closed curve reflecting the load and displacement on the polished rod is the indicator diagram curve during the process of lifting the polished rod from the lowest end to the uppermost end and then descending to the lowest end.

When the oil well pump is supposed to work normally, all kinds of friction resistance and inertia force in the system are 0, and the supply capacity of oil fluid is sufficient, the indicator diagram of pumping unit can be expressed as a parallelogram. However, in actual production, due to the influence of friction, vibration and inertia in the system, the indicator diagram under different working conditions will show different trends with that under normal working conditions.

It can be seen from the above figure (Fig. 2) that the better working condition of the pumping unit is, the more the indicator diagram tends to be "parallelogram" of the theoretical indicator diagram. Therefore, in the early oil well production, the area of indicator diagram is often used to determine the abnormal condition of working condition (Zhengqin and Hongsheng 2008).

According to the theoretical derivation of the condition evaluation function, the smaller the F value is, the more stable the working condition is, and is less affected by the change of the evaluation index parameters. At the same time, the value of actual power diagram area/theoretical work diagram area is less than 1 and tends to 1, which indicates that the working condition of pumping unit is closer to the normal working condition derived from theory.

Clustering various working conditions

The above evaluation function φ_1 and indicator map area change parameter φ_2 adopt cluster analysis to optimize

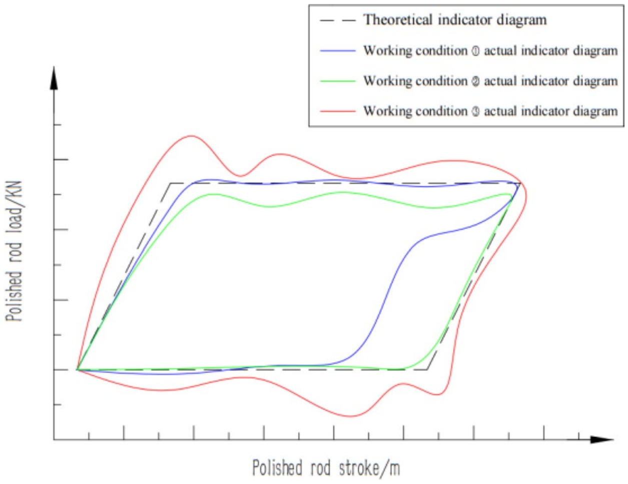


Fig. 2 Comparison between theoretical indicator diagram and actual indicator diagram

the division of different working conditions and divide all working conditions into several classes, so that each class of working conditions tends to its own characteristic attributes, so as to reduce the error of the prediction model of working conditions. The working condition cluster analysis of pumping unit is to distinguish and classify the working condition of pumping unit according to the working condition index characteristics and the situation reflected by the indicator diagram. Oriented to the engineering application of pumping unit, it mainly follows the following two principles (Zhengqin and Hongsheng 2008; Li et al. 2020):

a. The selected index characteristics must reflect the operating characteristics of the pumping unit under different working conditions and can be used to identify different working conditions;
b. Under the condition of ensuring the cumulative contribution rate of condition characteristics, the calculation amount of condition clustering should be as small as possible, which is feasible in engineering.

Let the working condition sample data set before classification be x and the corresponding characteristic matrix be W_o used the k-means clustering method to cluster the working condition samples, the condition samples are divided into k groups in advance, and the initial clustering center of each group is a matrix $m_b(b=1,2,3,\ldots K)$, then the Euclidean distance from each object x_b (characteristic matrix of corresponding working condition samples $W_{o,b}$) in different group subsets to the initial cluster center m_b.

$$d(x_b, m_b) = \left\| w_{o,b} - m_b \right\|_2 \tag{3}$$

Taking Euclidean distance as the standard, the working condition sample data are divided into the categories closest to the cluster center by iteration, and the clustering performance evaluation function, the sum of squares criterion function E, is constructed. When e reaches the minimum, it can make each class working condition sample compact and independent between each class.

$$E = \sum_{b=1}^{K} \sum_{W_K \in W_o^{(b)}} \left\| W_K - m_b \right\|_2 \tag{4}$$

In the formula, $W_o^{(b)}$ is the feature set of the class b working condition sample.

The steps of k-means clustering algorithm are as follows:

a. K initial clustering center mi is selected for all working condition sample data.
b. Euclidean distances from the clustering center of sample characteristics in different working conditions were calculated. According to the principle of minimum distance, any sample was assigned to the cluster center nearest to it.
c. The sample mean in each cluster was recalculated and used as the new cluster center.
d. Repeat steps b and c until the cluster center no longer changes.
e. K clusters are obtained. End.

Establish the regression equation of polished rod load

The second step of the judgment method of pumping unit working condition based on the law of polished rod load data is to select the theoretical calculation formula or empirical formula of polished rod load with the highest matching degree in the oilfield after obtaining the oil wells with normal working conditions through principal component clustering analysis.

Because there are some errors in the calculation formula of polished rod load, these errors make it easy to misjudge when using the polished rod load to judge the working condition. According to the composition and variation law of polished rod load of pumping unit, this paper puts forward a calculation method of polished rod load, which takes static load, dynamic load and friction load as independent variables and establishes multiple regression equation. The accuracy between the theoretical formula and regression equation and the actual polished rod load is calculated and verified.

Comparison of theoretical calculation formulas of common polished rod load

In the middle and later period of the last century, due to the low degree of automation of pumping units, the sensors are not widely used, and the polished rod load is mostly obtained by the empirical formula accumulated by foreign scholars for a long time. Later, some scholars simplified the pumping unit model as mass spring damping system and obtained the corresponding mechanical calculation formula. Due to the late start of oil exploitation in China, the calculation methods of polished rod load are introduced from abroad, including *API formula*, *Wilnofsky formula* and *Mills formula* (Yijiong 1984; Xin-fu and Yao-guang 2010).

The known data are brought into several commonly used polished rod load calculation formulas, and the calculated results of each formula are compared with the actual polished rod load. The results obtained are shown in Fig. 3, and the matching result of formula III (*Wilnovsky* formula) is better than that of other formulas. This shows that the *Wilnovsky* formula is suitable for the reservoir characteristics and production mode of Xinjiang oil field and can calculate the polished rod load more accurately.

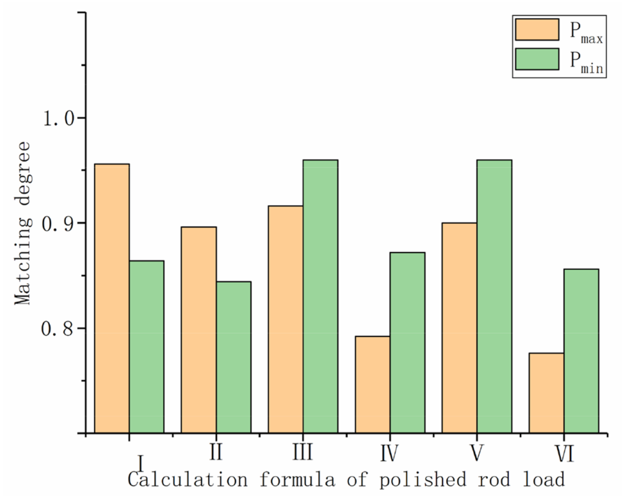


Fig. 3 Matching degree of theoretical calculation formula of common polished rod load

Wilnofsky formula avoids the appearance of empirical formula when calculating the polished rod load. It not only considers the weight of the sucker rod string, the buoyancy of the liquid column on the sucker rod string and the static load acting on the plunger area of the sucker rod, but also considers the vibration load of the system. When calculating the dynamic load, the influence of the elasticity of the sucker rod string is taken into account, as well as the influence of the dynamic load of the liquid column on the vibration are considered.

Composition of polished rod load

The polished rod load is a dynamic parameter in the operation of the pumping unit, and its maximum and minimum values occur in the up- and downstroke of the pumping unit, respectively. The load is not only composed of the weight of sucker rod string and liquid column, but also affected by inertia load, friction load (including friction force between rod string and tubing, friction force between plunger and bushing, friction force between liquid column and tubing, etc.), wellhead back pressure (increasing load) and submergence pressure (reducing load) and other factors.

Therefore, when the beam pumping unit drives the plunger of downhole oil pump, the load acting on the Horsehead suspension point is divided into the following three categories:

(1) Static load is mainly the weight of sucker rod string in liquid and the weight of oil column on plunger in tubing. The static load in the upstroke is equal to the sum of the weight of the sucker rod and the liquid column; the static load in the downstroke is only equal to the mass of the sucker rod.

(2) Dynamic load is related to the motion acceleration of the Horsehead suspension point of the pumping unit. The dynamic load includes the dynamic load generated by the movement of the sucker rod string and the oil column. In the calculation method of this study, the dynamic load of up- and downstroke is regarded as the same size.

(3) Various friction loads include friction resistance between plunger and pump barrel, friction resistance between sucker rod string and oil string, flow resistance between sucker rod string and tubing and relative movement between oil column and tubing, resistance when well fluid flows through the traveling valve of oil well pump and friction resistance between polished rod and wellhead polished rod sealer. Due to the existence of liquid column in the upstroke, the friction load in the upstroke is more than that in the downstroke.

Multiple linear regression equation

In production, there are many factors affecting the polished rod load of the pumping unit. These mechanical influencing factors are divided into static load, dynamic load and friction load. The correlation between the three types of load and the polished rod load can be calculated by regression analysis, so as to obtain the influence degree of multiple factors on the polished rod load (Xiaotao et al. 2013).

Multiple linear regression describes the linear relationship between multiple independent variables and dependent variables, such as assuming that the independent variables are $x_1, x_2, \ldots, x_k$. Then, there is a linear relationship between the independent variable and the dependent variable.

$$y_i = a_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \tag{5}$$

where y_i is the actual value of the i_{th} sample in the up- and downstroke; $x_{1i}, x_{2i}, \ldots, x_{ki}$ are the sampling data of the ith sample; β_0 is a constant; $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficient of a linear function.

Write the above expression into a matrix in the form of:

$$\hat{Y}_{n\times1} = X_{n\times k}\hat{\beta}_{k\times1} \tag{6}$$

In order to get the value of partial regression coefficient $\hat{\beta}_{k\times1}$, the least square method is introduced:

$$Q = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right) \tag{7}$$

The derivation of $\hat{\beta}$ on both sides of the above formula:

$$\hat{\beta}_{k\times1} = \left(X^T\hat{X}\right)^{-1}_{k\times k}\left(X^TY\right)_{k\times1} \tag{8}$$

Thus, the multiple linear regression equation can be calculated.

According to the theory of multiple linear regression, three kinds of loads in up- and downstroke of pumping unit are taken as independent variables, and X is used, respectively, x_1, x_2, x_3 and x_6 represent the static load, dynamic load and friction load acting on the polished rod during the up- and downstroke. Then, the regression equation of polished rod load of pumping unit is as follows:

Upstroke:

$$y_u = a_u + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} \tag{9}$$

Downstroke:

$$y_d = a_d + b_4 x_{i4} + b_5 x_{i5} + b_6 x_{i6} \tag{10}$$

where y_u and y_d are the actual polished rod load value in up- and downstroke; x_i is the ith load sample; a is a constant; $b_j(j = 1 \sim 6)$ is the partial regression coefficient.

The above regression equation can be used to calculate the polished rod load under the assumption that the working condition is normal.

Working condition prediction model based on statistics

Statistical distribution of loads in up- and downstrokes

Set the parameter $F_1, \ldots, F_6$ is the ratio of each load sample in the regression equation of up- and downstroke to the predicted value of polished rod load of that stroke, especially F_7 is the ratio of the sample value of the downstroke static load to the predicted value of the maximum polished rod load on the upper stroke. Calculate the parameter f in the well to be tested $F_1, \ldots, F_7$ and the normal distribution $N\left(\overline{X}_j, \sigma_j\right)$ of F_j is obtained by testing the normal distribution of a single sample of F_j. At the same time, through the parameter F_j.The mean value $\overline{X}$ and standard deviation σ of J data set were estimated with 95% confidence:

$$\left(\overline{X} - u_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}, \overline{X} + u_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right) \tag{11}$$

The confidence interval is $P_1, \ldots, P_7$, indicating the value of the parameter $F_1, \ldots, F_7$ has a 95% probability in this interval.

Common working conditions and corresponding threshold range of pumping units

Through the above statistical analysis, the confidence interval of each load in the upstroke and downstroke is $P_1, \ldots, P_7$. According to the following methods, the threshold limit is obtained by conversion, and the load change in different working conditions is transformed from qualitative analysis to the threshold range of quantitative solution with different thresholds(Rui et al. 2012; He Yanfeng and Xiaodong 2008; Zhiguo 2018; Zhangqi et al. 2014; Lea 1988).

Threshold limit 1: $x_{\text{line1}} = \left(P_7\right)_{\min}$;
Threshold limit 2: $x_{\text{line2}} = \left(P_1 + P_2 - P_3\right)_{\min}$;
Threshold limit 3: $x_{\text{line3}} = \left(P_1 + P_2 - P_3\right)_{\max}$;
Threshold limit 4: $y_{\text{line4}} = \left(P_4 + P_5 - P_6\right)_{\min}$;
Threshold limit 5: $y_{line5} = \left(P_4 + P_5 - P_6\right)_{\max}$;

Let the ratio between the calculated value and the measured value of the maximum polished rod load in the upstroke as the parameter ε_1. The ratio of the calculated and measured minimum polished rod load in downstroke is ε_2. By comparing and analyzing the relationship between the load of ε_1 and ε_2 and the threshold limits of the upstroke and downstroke under normal conditions, the common working conditions can be divided into the following categories.

(1) Normal condition: the change of polished rod load of oil well to be measured is similar to that under normal working condition, that is, ε_1 is between the limits of $F_1 + F_2 + F_3$,ε_2 is between the limits of $F_4 - F_5 - F_6$. Then, the expression of the threshold range under this condition is as follows: $x_{\text{line2}} < \varepsilon_1 < x_{\text{line3}}$,$y_{\text{line4}} < \varepsilon_2 < y_{\text{line5}}$.

(2) Tubing leakage: during the upstroke, the liquid is lost from the tubing, resulting in the static load between the rod weight and the sum of the rod weight and the liquid column weight, while the downstroke load is unchanged.

Threshold range:$\varepsilon_1 < x_{\text{line2}}, y_{\text{line4}} < \varepsilon_2 < y_{\text{line5}}$.

(3) Breaking off of sucker rod: the static load in the upstroke is less than the self-weight of the sucker rod, and the load of the downstroke decreases.

Threshold range: $\varepsilon_1 < x_{\text{line1}}, \varepsilon_2 < y_{\text{line4}}$.

(4) Insufficient pump charging: due to insufficient liquid supply or gas influence, the upstroke gas expands in the pump barrel, resulting in the decrease in polished rod load, and the compression of downstroke gas in the pump barrel leads to the decrease of polished rod load.

Threshold range: $x_{\text{line1}} < \varepsilon_1 < x_{\text{line2}}, \varepsilon_2 < y_{\text{line4}}$.

(5) Bump pump: at the end of upstroke, the maximum load increases due to the collision between plunger and pump body.

Threshold range: $\varepsilon_1 > x_{\text{line3}}, y_{\text{line4}} < \varepsilon_2 < y_{\text{line5}}$.

(6) At the end of downstroke, the minimum load decreases due to the collision between plunger and pump body.T.

hreshold range: $x_{\text{line2}} < \varepsilon_1 < x_{\text{line3}}, \varepsilon_2 < y_{\text{line4}}$

(7) Heavy oil: the viscosity of crude oil increases, resulting in the increase in additional resistance of

sucker rod, which increases the friction of sucker rod upstroke and downstroke, increases the upstroke load and decreases the downstroke load.

Threshold range: $\varepsilon_1 > x_{\text{line3}}, \varepsilon_2 < y_{\text{line4}}$.

(8) Friction increase: the additional resistance increases the load of polished rod in the upstroke and downstroke. The working conditions of this kind of situation include wax scaling, sand production, eccentric wear or sucker rod bending.

Threshold range: $\varepsilon_1 > x_{\text{line3}}, \varepsilon_2 > y_{\text{line5}}$.

(9) Floating valve or fixed valve leakage: the valve leakage of oil well pump, the load changes with the change of leakage in the upstroke, and the liquid column load cannot be removed in time during the downstroke.

Threshold range: $x_{\text{line2}} < \varepsilon_1 < x_{\text{line3}}, \varepsilon_2 > y_{\text{line5}}$.

(10) Unknown condition: the threshold range of this part is: $\varepsilon_1 < x_{\text{line2}}, \varepsilon_2 > y_{\text{line5}}$. Indicates that the load decreases in the upstroke and increases in the downstroke. The load variation relationship cannot find its mapping condition, and it may be the data acquisition error in the field.

Working condition prediction model

According to the threshold diagnosis range under the above working conditions, the ratio ε_1 and ε_2 of the maximum value of the measured polished rod load and the maximum value predicted by the regression equation in the up- and downstroke are used as the horizontal and vertical coordinates, and the working condition prediction diagram as shown in the figure below is drawn in combination with the threshold limit, and the obtained working condition prediction model is graphically processed (Liangyu 2019).

Then, the characteristic parameters and working condition prediction model under normal working conditions are obtained by analyzing the previous working condition data of the oil well to be tested, and then the current working condition characteristics are calculated according to the real-time data, and the working condition prediction model is brought in for comparison, so as to judge whether the working condition of the oil well to be tested is normal and point out the attribution of abnormal working conditions.

Test

Data preparation

In the process of oil production, the oil well data acquisition system stores a large amount of production parameter data, which contains the information of corresponding well conditions. Obviously, the number of parameter variables (10 production parameters) and high dimension of oil well production data are obtained in the process of oilfield production. The production parameters are analyzed to reduce the data dimension of the original data space, and then the parameters with large amount of information are extracted from the original data space, which not only reduces the analysis complexity, but also eliminates part of the noise interference.

According to the above analysis, 10 indexes which affect the pumping efficiency of pumping wells are selected as follows: the maximum load of polished rod on the upstroke (P_u), minimum load of polished rod in downstroke (P_d), maximum current of upstroke (I_{um}), maximum downstroke current (I_{dm}), stroke (S), stroke (n), oil pressure (P_t), casing pressure (P_c), submergence (H) and pump efficiency (η_p) The production and basic data of 772 wells in a month were selected randomly. Due to the large amount of data, only the statistical values of these data are given, as shown in Table 1.

Main component cluster working condition results

The whole calculation process of principal component is realized in SPSS software. The correlation coefficient and eigenvalue λ_i of the 10 indexes are shown in Figs. 4, 5, respectively.

According to the correlation coefficient matrix in Fig. 4, the original index P_u, P_d, S, n, I_{um}, I_{dm} have a strong correlation, which reflects that the operating conditions of the pumping unit have a great correlation with these indicators, among which P_u, P_d, S are the most strong correlation, which shows that the load and displacement changes at the polished rod of the pumping unit can reflect the operation condition of the pumping unit to the greatest extent, which is consistent with the theory of the working condition diagnosed by indicator diagram on the oil field. However, the correlation between submergence H and pump efficiency η_p is very poor, which reflects that submergence is basically independent of specific well conditions.

According to Fig. 5b, when the evaluation index is selected according to the variance contribution rate of principal components, the number of principal components determined by the cumulative contribution rate of eigenvalue variance is 3, and the cumulative contribution rate reaches 73.143. Although it is less than 85% specified in the principal component method, the eigenvalues of the three indicators are greater than 1. Meanwhile, the results show that KMO is 0.846 > 0.6, indicating that the data are suitable for factor score The significant p value of Bartlett's spherical test was 0.000 < 0.05, which indicated that the data were suitable for factor analysis.

At the same time, the initial factor load matrix can be directly obtained by using SPSS, and each load represents the correlation coefficient between the principal component and the corresponding variable. For the principal component

Table 1 statistical values of 10 indexes of pumping wells

Statistical	P_u(KN)	P_d(KN)	I_{um}(A)	I_{dm}(A)	S (m)
Maximum	69.2	56	1.6	2.5	101.9
Minimum	1.6	2	0.02	0	5.62
Average	16.67	16.63	0.63	0.68	28.27
Deviation	12.40	10.58	0.17	0.34	22.96
Statistical	N ($\min^{-1}$)	P_t(MPa)	P_c(MPa)	H (m)	η_p
Maximum	70.6	10	12	2310.1	349.2
Minimum	0	0.73	2.34	-532.3	0.31
Average	17.64	2.29	5.73	227.77	37.05
Deviation	16.04	1.21	1.86	325.47	33.85

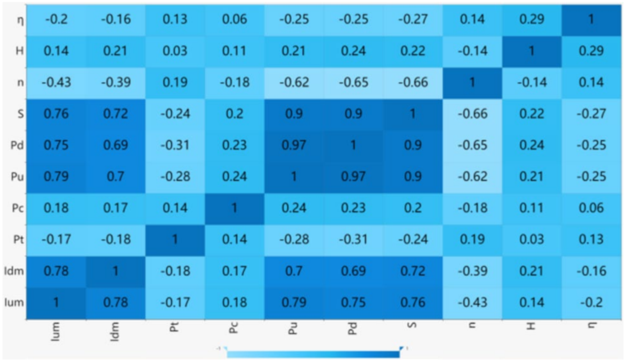


Fig. 4 Correlation coefficient matrix

load matrix, the data in the initial factor load matrix are divided by the eigenvalue corresponding to the principal component, and then the square root is calculated to obtain the corresponding coefficient of each index in the seven principal components, so as to obtain the corresponding principal component load value.

According to the principal component load, the maximum load (P_u), the minimum load of polished rod (P_d), the measured stroke (S) and the maximum current of motor upstroke ($I_{\rm um}$) and the maximum downstroke current ($I_{\rm dm}$), The absolute values of the correlation coefficients between them and the first principal component all exceed 0.8. Because these five indexes reflect the mechanical and displacement changes at the polished rod, the first principal component represents the mechanical state of the polished rod and the running state of the motor. The load on the polished rod can directly

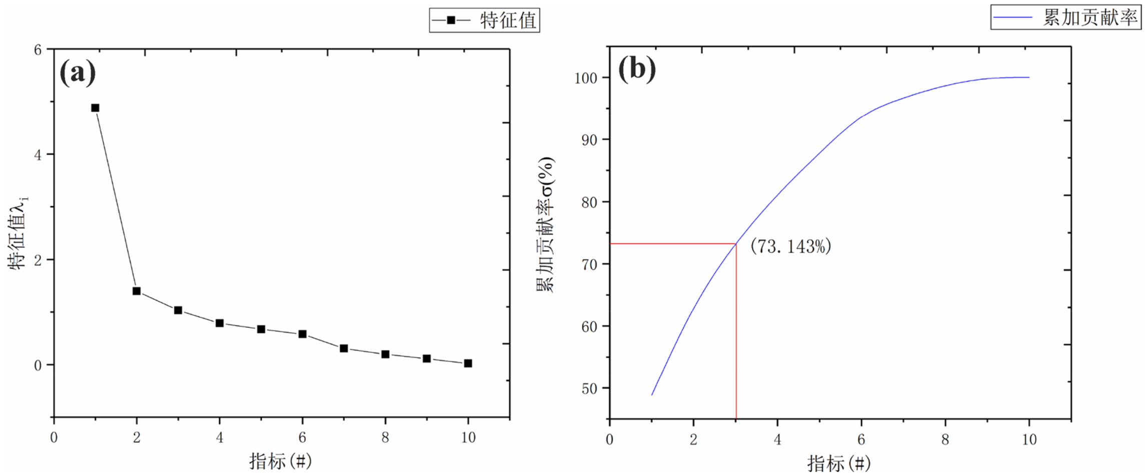


Fig. 5 Eigenvalue, eigenvalue contribution rate and cumulative contribution rate (**a**) characteristic root value attenuation broken line (#-λ_i) of oil well data (**b**) contribution rate

reflect the mechanical situation of the pumping unit in the well, and different mechanical conditions in the well correspond to different working conditions. Therefore, it is the most direct and important to determine the downhole working condition and pumping unit running state according to the load of polished rod.

Submergence and pump efficiency are closely related to the second principal component, and the correlation coefficient between them and the second principal component is greater than 0.7, which indicates that the abnormal working condition of the pumping unit is closely related to the filling degree and working efficiency of the pump.

Tubing pressure and casing pressure are closely related to the third principal component. Their correlation coefficient with the third principal component is about 0.6, which reflects the formation energy. The smaller the tubing pressure is than the casing pressure, the greater the formation capacity and the better the fluid supply capacity.

According to the load value of principal component and the standardized data of each original index, the corresponding principal component expression and comprehensive evaluation function can be obtained:

The first principal component expression: $F_1 = 0.849Z_{I_{\mathrm{um}}} + 0.806Z_{I_{\mathrm{dm}}} - 0.31Z_{P_t} + 0.262Z_{P_c}$

$+0.957Z_{P_u} + 0.956Z_{P_d} + 0.942Z_S - 0.697Z_n$.

$+0.25Z_H - 0.281Z_{\eta_p}$;

The second principal component is as follows: $F_1 = 0.001Z_{I_{\mathrm{um}}} + 0.06Z_{I_{\mathrm{dm}}} + 0.438Z_{P_t} + 0.461$

$-0.013Z_{P_c} - 0.006Z_{P_d} - 0.007Z_S - 0.02Z_n$.

$+0.7Z_H + 0.707Z_{\eta_p}$;

The third principal component is as follows: $F_1 = 0.071Z_{I_{\mathrm{um}}} + 0.002Z_{I_{\mathrm{dm}}} + 0.595Z_{P_t} + 0.612Z_{P_c}$

$+0.002Z_{P_u} - 0.037Z_{P_d} + 0.009Z_S + 0.012Z_n$ .where F_i

$-0.418Z_H - 0.354Z_{\eta_p}$.

represents the i_{th} principal component, and Z represents the value of each index after standardization.

According to the weighted method of principal component eigenvalues, the comprehensive evaluation function of pumping units can be calculated:

$$F = \sum_{i=1}^{3} \lambda_i F_i / \left(\lambda_1 + \lambda_2 + \lambda_3\right)$$

$$= \frac{4.88}{7.314}F_1 + \frac{1.399}{7.314}F_2 + \frac{1.035}{7.314}F_3$$

$$= 0.667F_1 + 0.191F_2 + 0.142F_3.$$

According to the comprehensive evaluation function of pumping unit working condition and the above judgment method for normal working condition of pumping unit, cluster analysis is carried out on the tested oil wells, and the clustering results as shown in Fig. 6 are obtained.

According to the above abnormal condition judgment method, condition III in Fig. 6 is selected as normal condition.

Comparison results of calculation methods of polished rod load

The data of 187 oil wells under normal working conditions were analyzed by multiple linear regression, and the regression equation was obtained as follows:

Upstroke:

$$y_u = 0.067 + 1.096x_{i1} + 79.48x_{i2} - 5.746x_{i3} \tag{12}$$

Downstroke:

$$y_d = -0.467 + 0.922x_{i4} + 5.945x_{i5} - 1.02x_{i6} \tag{13}$$

Figure 7 shows the comparison between the actual polished rod load value of 187 oil wells under normal working condition and the polished rod load value calculated by formula and regression equation.

It is not difficult to see from Fig. 7 that the difference between the polished rod load value calculated by regression equation and the actual polished rod load value is obviously smaller than the polished rod load value calculated by using the formula.

Working condition prediction model

SPSS was used to make descriptive statistics on the samples of each load in the up- and downstroke, and the statistical distribution results as shown in the figure below were obtained.

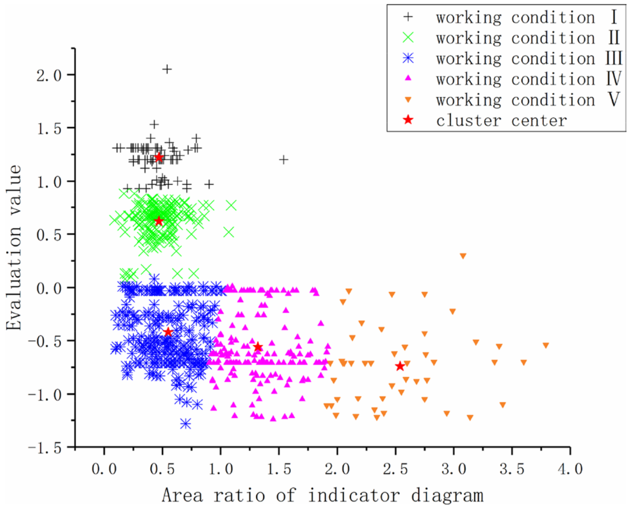
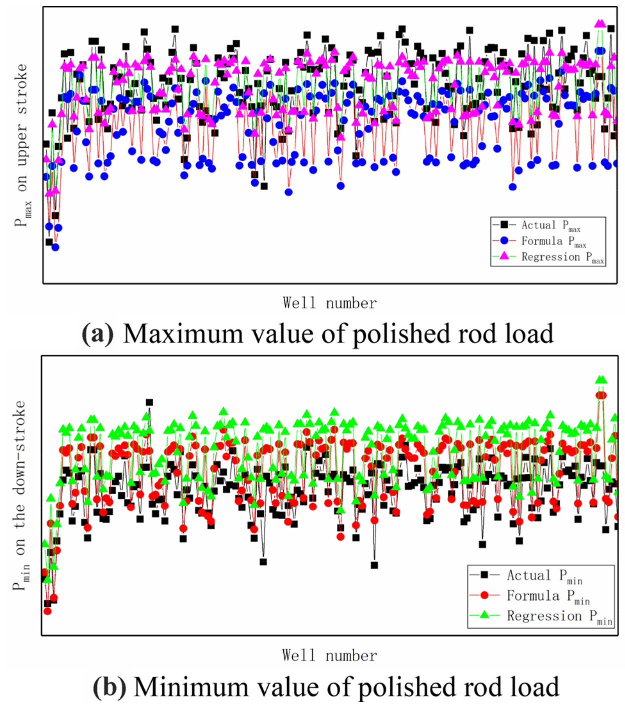


Fig. 6 Clustering results of working conditions



(a) Maximum value of polished rod load

(b) Minimum value of polished rod load

Fig. 7 Comparison of matching degree of polished rod load **a** maximum value of polished rod load **b** minimum value of polished rod load

According to the confidence interval obtained in Fig. 8, the threshold range and prediction model are calculated and transformed.

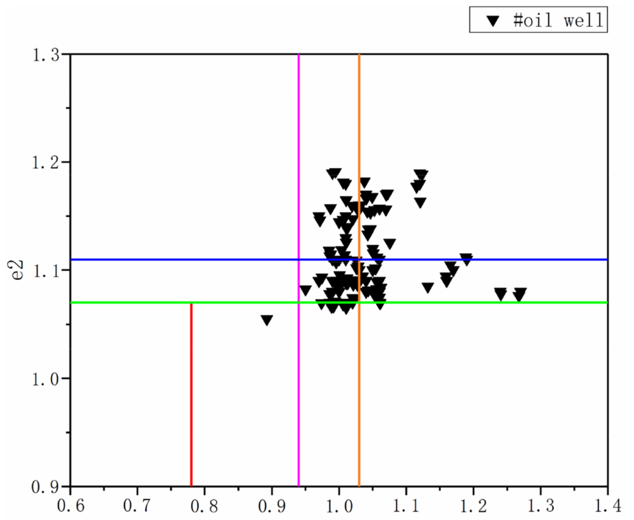
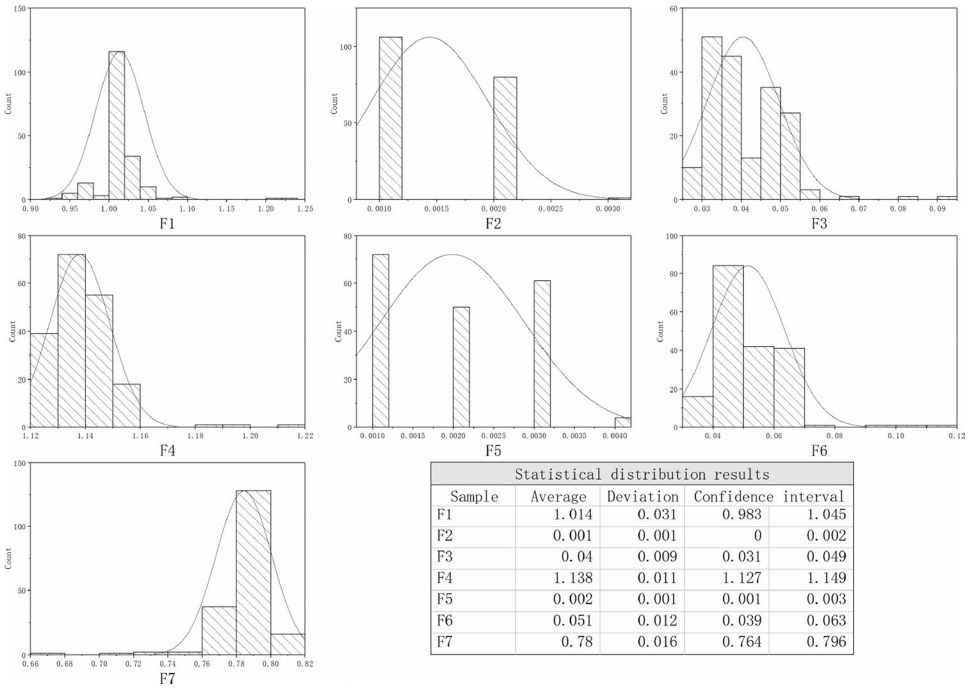


Fig. 9 Judgment results of oil well condition to be tested in an oil production area

Fig. 8 Statistical distribution results



Statistical distribution results				
Sample	Average	Deviation	Confidence interval	
F1	1.014	0.031	0.983	1.045
F2	0.001	0.001	0	0.002
F3	0.04	0.009	0.031	0.049
F4	1.138	0.011	1.127	1.149
F5	0.002	0.001	0.001	0.003
F6	0.051	0.012	0.039	0.063
F7	0.78	0.016	0.764	0.796

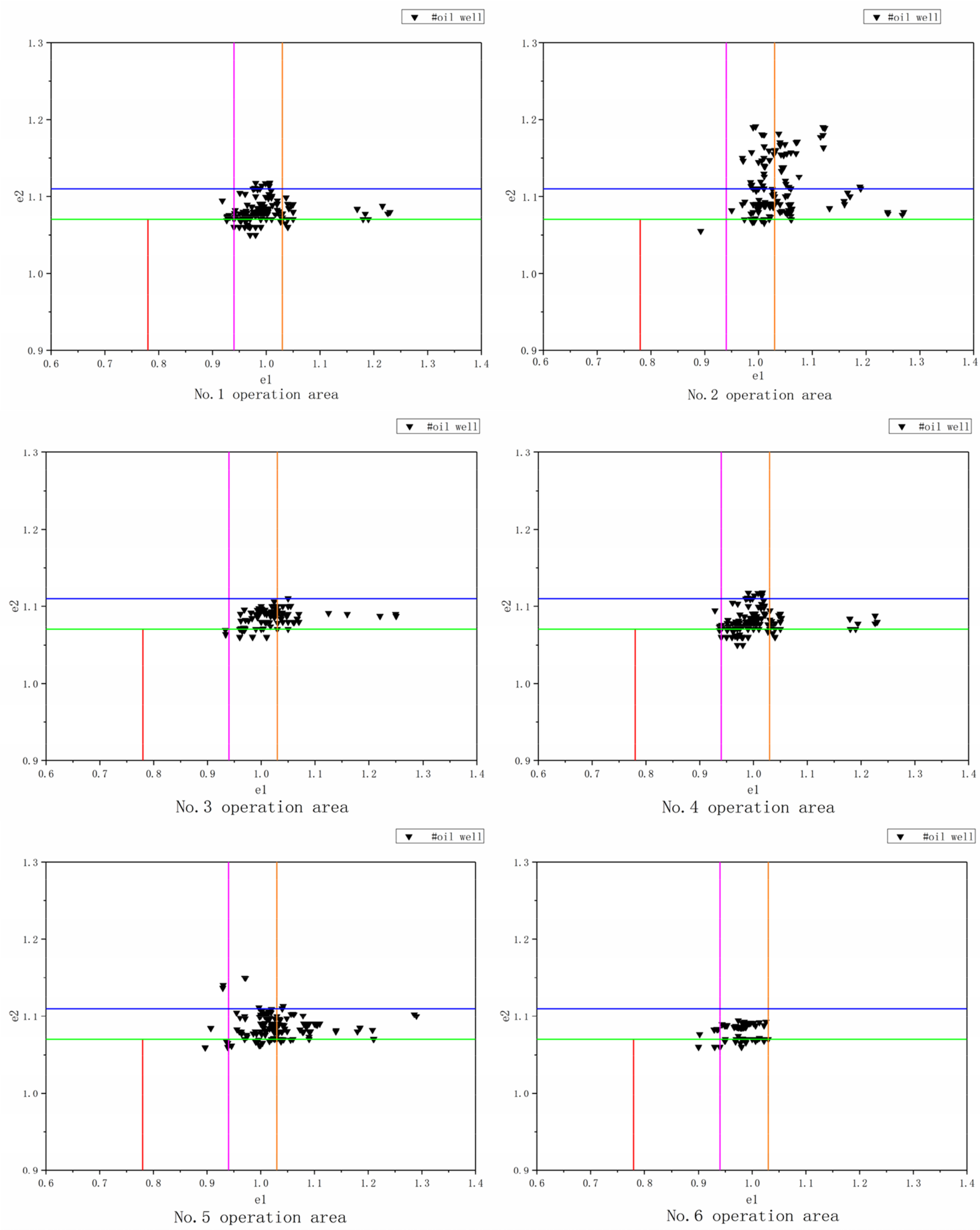


Fig. 10 Judgment results of oil well condition to be tested in some oil production area

Working condition prediction results

The production data of the old oil wells in each operation area of the oilfield are collected, and the theoretical maximum and minimum values of the polished rod load, as well as the corresponding static load, dynamic load and friction load, are calculated according to the regression optimization calculation method described in this paper.

Then, the ratio of the measured maximum value and the theoretical maximum value of the polished rod load ε_1 and the ratio of the measured minimum value and the theoretical minimum value of the polished rod load ε_2 are calculated in turn. And the values of ε_1 and ε_2 are brought into the working condition diagnosis diagram, and the following judgment results are obtained.

There are 86 old oil wells in Fig. 9, including 11 wells judged as normal working condition, 1 oil well affected by insufficient liquid supply or gas, 10 oil wells bumped by pumps and 72 wells with excessive downhole friction load. Combined with the above analysis, it can be concluded that the oil reservoir in this operation block has good liquid supply capacity and the pumping unit works normally, but the friction load of most oil wells is too large.

In Fig. 10, through the statistics of the diagnosis results of the old oil wells in the test operation area, it is found that the main low efficiency reasons of the old oil wells are the excessive friction of the sucker rod in the up- and downstroke, the leakage of the tubing and valve and the problem of the up and down hitting the pump. Therefore, in the next step of production management of old oil wells, targeted treatment measures should be adopted to save energy and improve efficiency, such as appropriately reducing the stroke to reduce the collision of oil well pumps and stopping the oil wells with serious leakage or severe friction for maintenance.

By comparing the detection results of pumping unit working condition on site with the prediction results obtained by this research method, it is found that the accuracy rate of this research method is very considerable, which can effectively reflect the working condition problems encountered by old oil wells. This test result shows that the study is more suitable than other statistical analysis models for old wells with complex operating conditions and low quality data. At the same time, this study also has the disadvantages and limitations, for example, when two or more working conditions have little difference in the influence law and influence degree on the polished rod load, the method provided in this study can only predict this kind of working condition, but cannot accurately distinguish which one is.

Summary and conclusions

(1) For research viewpoint, this research enlarges the application scope and scenario of statistical model in oilfield data processing. And in the process of research, it is not difficult to find that there is a large amount of available information in the oilfield production data, which is worthy of more researchers and explorations.

(2) The results of principal component analysis show that the load of polished rod and the running current of pumping unit motor have the highest correlation with the running state of pumping unit, which can directly reflect the working condition of pumping unit.

(3) The results show that the calculation method of polished rod load optimized by regression analysis is more accurate than various theoretical calculation formulas commonly used in oilfield.

(4) In this paper, the working condition judgment method based on the change law of polished rod load can quickly and effectively judge the operation conditions of the old oil wells that meet the conditions in a certain operation block. However, how to further judge and identify similar working conditions, such as gas influence, insufficient liquid supply and pipeline leakage, needs to be further studied.

Funding Sichuan Youth Science and Technology Foundation (19JCQN0081).

Compliance with ethical standards

Conflict of interest No conflict of interest exits in the submission of this manuscript, and all authors have approved the manuscript and agreed to submit it to" Journal of Petroleum Exploration and Production Technology ". I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in parts. All the authors listed have approved the manuscript that is enclosed. We deeply appreciate your consideration of our manuscript, and we look forward to receiving comments from the reviewers. If you have any queries, please don't hesitate to contact me at the address below.

References

Canelon MAR, Morles EC (2008) Fuzzy clustering based models applied to petroleum processes[J]. Wseas Trans Syst Control 3:159–171

Chaodong T, Jiancheng C, Zhihai L et al (2015) Prospects for the application of big data mining technology in petroleum engineering[J]. China Pet Chem Ind 01:49–51

Gibbs SG, Neely AB (1966) Computer diagnosis of down-hole conditions in sucker rod pumping wells,[J]. J Pet Tech 18:91–98

He Yanfeng Wu, Xiaodong HG et al (2008) A new method for spectrum analysis of indicator diagrams[J]. Acta Pet Sin 04:619–624

Jain A K, Duin R P W, Jianchang Mao (2000) Statistical pattern recognition: a review. Pattren analysis and machine intelligence[J]. *IEEE Transactions* 22(1): 4–37

Karmawijaya MI, Rahardja B, Hakim A (2009) Production data management and reporting system to optimize production analysis. Asia Pacific Oil & Gas Conference & Exhibition. Society of Petroleum Engineers

Khakimyanov MI, Khusainov FF, Shafikov IN (2015) Technological parameters influnce of oilwells on energy consumption sucker rod pumps [J]. Oil Gas Bus 8(10):533–540

Li R, Lin H, Bu W (2020) Ship-radiated noise evaluation method based on optimized working-conditions clustering [J]. J Huazhong Univ Sci Technol (Nat Sci Ed) 48(10):63–68, 103

Lea JF (1988) Boundary conditions used with dynamic models of beam pump performance [C]. Amoco production company, Southwestern petroleum short course-88

Liangyu X (2019) Energy consumption evaluation and working condition analysis method based on macro control chart[J]. Energy Conserv Pet Petrochem 9(03):38-40+10–11

Rui L, Sumin C, Shuzhi Z et al (2012) Research on diagnosis of pumping unit fault model using multivariate real-time data[J]. China Pet Chem Ind 01:52–54

Stamatatos E (2008) Author identification: using text sampling to handle the classimbalance problem [J]. Inf Process Manage 44(2):790–799

Wei X, Xingchang X, Yong Z, Zhiyang L, Longfei R, Chuanjun H (2020) Influence of daily liquid production on energy efficiency evaluation index of mechanical recovery system [J]. China Pet Chem Stand Qual 40(01):9–11

Xiang H (2007) Research on the intelligent integrated system for fault diagnosis of pumping wells[J]. J Oil Gas Technol (03):156–158

Xiaodong Z, Xianhua X, Zhengyao Li et al (2011) Principal component analysis of influencing factors of pumping well efficiency [J]. J Southwest Pet Univ (Nat Sci Ed) 33(05):176-180+204

Xiaotao He, Xiaomin G, Dong Z, Chen zhaona, (2013) Study on power consumption economic analysis and prediction based on principal component regression analysis [J]. Inform Syst Eng 01:140–142

Xin-fu L, Yao-guang Q (2010) Calculation of dynamic loads of the sucker rod pumping system in CBM wells[J]. J Coal Sci Eng (China) 16(02): 170-175

Xuanyi S, Yuetian L, Jing Ma, Junqiang W, Xiangming K, Xingnan R (2020) Productivity prediction based on support vector machine optimized by Gray Wolf algorithm [J]. Lithol Reserv 32(02):134–140

Yijiong Wu (1984) Calculation of suspended point load of beam pumping unit[J]. Pet Field Mach 01:15-24+64-65

Yuling W, Jianping B (2014) Research on Evaluation Technology of potential oil reservoir by reexamination of old oil wells [J]. J Chongqing Univ Sci Technol (Nat Sci Ed) 16(05):33–35

Zengguo T, Dongzhe T, Baozhu J, Jihua C (2019) Indicator diagram fault diagnosis system based on principal component analysis [J]. Intern Combust Eng Access 24:149–151

Zhengqin Li, Hongsheng Li (2008) Pumping unit fault diagnosis model based on the area change of indicator diagram[J]. Oil Gas Field Surf Eng 09:7–8

Zhiguo Z (2018) Study on the judgment method of the optimal suspension point load utilization of the pumping unit[J]. Energy Conserv Pet Petrochem 8(11):4-5+8+7

Zhong Zhangqi, Hou Dujie, Ye Xiaofeng et al. (2014) Quantitative analysis of indicator diagram load ratio and wax removal. Oil Gas Field Surf Eng 33(09) 10–11