Classification of Tight Sandstone Gas Wells Based on Time Series Similarity

Shaoyu Dong, Chunlan Zhao

***Abstract*—Scientific and effective classification of gas wells is conducive to mastering production characteristics of wells and establishing reasonable production measures. Therefore, based on the similarity of time series, and in conjunction with spectral clustering, this paper proposes a new classification model for tight sandstone gas wells: the KF-SDTW-Spectral model. The research results show that the 61 tight sandstone gas wells can be divided into clusters using different classification indicators: two clusters based on monthly gas production, three clusters based on monthly water production, and three clusters based on oil pressure. The model exhibits high clustering quality, surpassing traditional clustering methods. By using this model, the production characteristics of high-yield and low-yield wells are analyzed and corresponding production measures are proposed. This paper provides a new approach on the classification of tight sandstone gas wells, offering valuable guidance for the formulation of production measures in the gas field.**

***Index Terms*—kalman filtering, SDTW, spectral clustering, classification of gas wells**

I. Introduction

In recent years, with the rapid development of China's economy, the demand for energy has been increasing. Conventional natural gas resources are no longer sufficient to meet domestic market demands, which has led to a focus on the exploration and development of unconventional natural gas. Tight sandstone gas, also known as tight gas, is one type of unconventional natural gas[1]. China boasts rich reserves of tight sandstone gas, but these are characterized by poor reservoir properties, low production per well, and unstable yield capabilities. To achieve efficient development, large-scale well layout operations are necessary. However, within the same region, the production characteristics vary from one gas well to another. Conducting a reasonable, accurate, and efficient classification of gas wells is beneficial for enhancing the understanding of production patterns and for formulating appropriate production measures.

Currently, the commonly used methods for classifying gas wells include the unimpeded flow method, the reservoir parameter method, and the daily gas production method[2]. However, these methods have some limitations. As gas well production continues, the evaluation results of the unimpeded flow may not correspond with the actual output; the reservoir parameter method mainly classifies gas wells based on effective reservoir thickness, which cannot dynamically reflect the actual production capacity of gas wells; the daily gas production method does not take into account the impact of production time on gas production capability.

A time series is a set of records identified by time[3], widely used in various fields such as scientific research and economic analysis. In recent years, time series data mining has attracted significant attention and research. Most data mining applications for time series require similarity measure. Numerous methods have been proposed to measure the similarity of time series, broadly categorized into time-rigid measures (Euclidean distance)[4], time-flexible measures (Dynamic Time Warping) [5], feature-based measures (Fourier coefficients) [6], and model-based measures (auto-regression[7] and moving average model[8]). Euclidean distance is widely used due to its simple calculation, but it can only handle time series of the same length. Dynamic Time Warping (DTW) is a method for measuring the similarity between time series of different lengths, initially developed for speech recognition[9]. Additionally, DTW can optimally handle contractions, expansions, and shifts in time series[10]. The flexibility and effectiveness of DTW have led to its widespread application in various fields, such as gesture recognition[11], biometrics[12], and astronomy[13]. Despite DTW being successfully applied across multiple fields, it still has certain shortcomings. The "singularity" problem is one of the more serious ones[14]. Simply put, a singularity refers to a point in a time series that continuously maps to a large region of another time series, resulting in inaccurate DTW distance calculations. Singularities occur because DTW, when calculating local distances, only considers the point values of the time series (i.e., the spatial dimension), and ignores information on the time dimension. Therefore, many scholars solve the singularity problem by adding information in the time dimension, specifically derivatives. Keogh and Pazzani[14] proposed a method called Derivative Dynamic Time Warping (DDTW), which uses derivate feature instead of point value feature to calculate local distances. However, DDTW only considers the derivatives information of time series and completely ignores the point values information. Therefore, DDTW still cannot adequately measure the similarity between time series. Benedikt, Kajic, Cosker, Rosin, and Marshall[15] proposed a local distance metric that takes a weighted sum of point-to-point distance and derivative-to-derivative distance, called Weighted Derivative Dynamic Time Warping (WDDTW). However, this

Manuscript received January 19, 2024; revised July 14, 2024. This work was supported part by Natural Science Foundation of Sichuan Province (No.2022NSFSC0283).

S. Y. Dong is a postgraduate student of School of Science, Southwest Petroleum University, Sichuan, Chengdu, 610500, China. (e-mail:1257230982@qq.com).

C. L. Zhao is a professor at School of Science, Southwest Petroleum University, Key Laboratory of Energy Security and Low-carbon development, Sichuan, Chengdu, 610500, China. (Corresponding author, e-mail: 308303451@qq.com).

algorithm introduces extra weighting parameters, which increases the computational complexity and decreases the robustness of the algorithm. Addressing the shortcomings of DDTW and WDDTW, Shen, Zhu, Huang, and Liang[16] proposed Summation Dynamic Time Warping (SDTW), which considers both point values and derivative features simultaneously, and does not require additional weighted parameters. This approach can effectively solve the singularity problem and improve the accuracy of time series similarity measure.

During the production of tight sandstone gas wells, the values of each production parameter are recorded in time sequence. The production parameter data, presented in the form of time series, depict the change process of tight sandstone gas wells over time. To analyze the production characteristics during the production process of tight sandstone gas wells, this paper focuses on 61 tight sandstone gas wells in a certain gas field in Southwest China as research subjects. Based on time series similarity and incorporating spectral clustering, a new gas well classification model is proposed. Subsequently, corresponding production measures are proposed based on the classification results.

The remainder of this paper is organized as follows: Section II introduces relevant methods. Section III outlines the construction of a classification model for tight sandstone gas wells. Section IV presents the conducted experiments and results. Finally, Section V summarizes the entire paper.

II. METHODOLOGY

A. Kalman Filtering

Kalman filtering is a data processing technique used to remove noise and restore true data. It is based on the principle of minimizing the mean square error and is used for the optimal estimation of data sequences. In most cases, the Kalman filtering of discrete system is mainly used, and its mathematical model is composed of state equation and observation equation, which can be expressed as[17]

$$X_k = AX_{k-1} + W_{k-1} \tag{1}$$

$$Z_k = CX_k + V_k \tag{2}$$

Where X_k is an $n\times1$ order state vector, Z_k is an $m\times1$ order observation vector, A is an $n\times n$ order state transition matrix, W_{k-1} is an $n\times1$ order process noise, V_k is an $m\times1$ order observation noise, and C is an $m\times n$ order observation matrix.

Based on the principle of least squares, the recursive algorithm for Kalman filtering in discrete systems is illustrated in Fig.1.

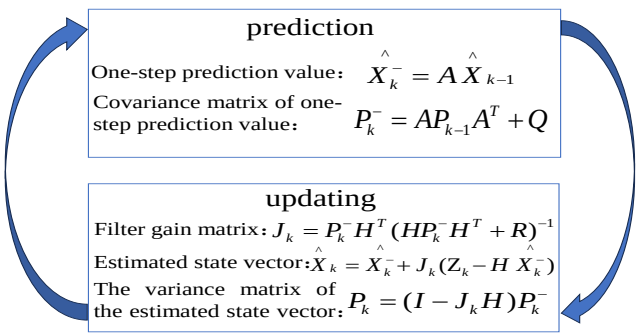

Fig .1. Recursive algorithm for the Kalman filtering

R is the observation noise covariance matrix, and Q is the process noise covariance matrix. The aforementioned recursive algorithm requires initial conditions for starting, denoted as $\hat{X}(0)$ and $P(0)$. $\hat{X}(0)$ is commonly set to zero or any other value obtainable from prior information, while $P(0)$ can be set as a factor of the identity matrix. These initial conditions are then substituted into the recursive algorithm for iteration, with continual prediction and updating until convergence is reached. This process yields the optimal estimation, denoted as the state estimation at time k, represented as $\hat{X}_k$, $(k=1,2,3,\cdots)$,achieving the filtering effect and effectively eliminating random interference noise.

B. Dynamic Time Warping(DTW)

Dynamic Time Warping (DTW) is a method used to calculate the optimal mapping between two time series through dynamic programming, thereby representing the similarity between the two series. Suppose there are two time series, $x(i)$, $i=1,2,\cdots,m$ and $y(j)$, $j=1,2,\ldots,n$. To calculate the DTW distance of these two series , we first calculate an $m\times n$ order distance matrix D, where the (i^{th}, j^{th}) element is represented by $d_{local} = (x(i)-y(j))^2$. d_{local} is called local distance, that is, the distance between two time points in two time series [16].

Define a warping path W to represent an alignment or mapping of the series x and y,

$$W = \begin{pmatrix} w_x(k) \\ w_y(k) \end{pmatrix}, k = 1,2,...,p, \tag{3}$$

Where $w_x(k)$ and $w_y(k)$ respectively represent the subscripts of elements in series x and series y, and p represents the length of the warping path W, satisfying $p \in [max(m,n), m+n-1]$. $\begin{pmatrix} w_x(k) \\ w_y(k) \end{pmatrix}$ means that the $w_x(k)$ element in series x maps to the $w_y(k)$ element in series y.

Meanwhile the warping path W must satisfy the following three constraints:

Boundary condition: The warping path W must begin at $w_1 = (1,1)$ and end at $w_k = (m,n)$. That is, the selected warping path must start from the lower left corner and end at the upper right corner.

Continuity: The adjacent elements $w_k = (a,b)$ and $w_{k-1} = (a',b')$ in the warping path W must satisfy $a-a' \le 1$ and $b-b' \le 1$.

Monotonicity: The adjacent elements $w_k = (a,b)$ and $w_{k-1} = (a',b')$ in the warping path W must satisfy $a-a' \ge 0$ and $b-b' \ge 0$.

There are many possible warping paths that satisfy the three constraints mentioned above, and DTW aims to find the optimal path among these warping paths such that the cumulative sum of the local distances along that path is

minimized. Hence, $DTW(x, y)$ is used to represent the shortest distance between time series x and y — that is, the distance corresponding to the optimal warping path among all possible warping paths W. Computing the shortest distance $DTW(x, y)$ and the optimal warping path is a dynamic programming problem that adheres to the three constraints mentioned above.

$$\begin{cases} r(i,j) = d(i,j) + \min\{r(i-1,j-1), r(i-1,j), r(i,j-1)\} \\ DTW(x,y) = \min\{r(m,n)\} \end{cases} \quad (4)$$

The $r(i, j)$ represents the cumulative distance of the local distances along the path from $(1,1)$ to (i, j) in the distance matrix D. Fig.2 illustrates the warping path for the series x and y.

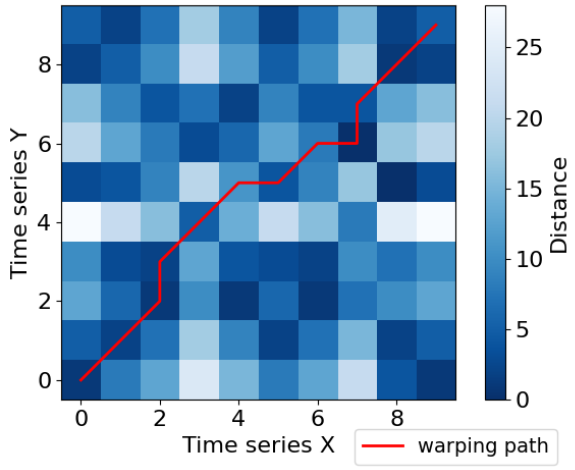

Fig .2. The warping path of DTW

C. Modified DTW-Summation Dynamic Time Warping(SDTW)

SDTW[16] first defines the feature $Fs(x(i))$ as

$$F_s(x(i)) = (1 + \frac{x(i) - x(i-1)}{\max(|\Delta x|)}) \cdot x(i), 1 < i \le m, \quad (5)$$

Where $\max(|\Delta x|)$ represents the maximum derivative of all time points in the series $x(i)$. Here, the derivative of a point is calculated by taking the difference between two adjacent points, i.e., using $x(i) - x(i-1)$ to represent the derivative of $x(i)$. The purpose of $\max(|\Delta x|)$ is to constrain the derivatives to the range $[-1, 1]$, thereby incorporating derivative information into the features in a ratio form.

Thus, the local distance (d_{local}) can be expressed as:

$$d_{local}(i,j) = (F_s(x(i)) - F_s(y(j)))^2 \quad (6)$$

From (5), it is apparent that the feature $Fs(x(i))$ contains both point values information and derivatives information, with the derivatives information being merged into the feature in the form of a ratio. Therefore, it can effectively address the problem of singularity.

D. Spectral Clustering

D.1. Algorithm Flow

Spectral clustering is an unsupervised classification method that utilizes the concept of graphs for clustering. It views all the samples in a dataset as points in space, and then connects any two points with an edge, representing the original dataset in the form of an undirected graph. Finally, it sets a criterion for cutting the undirected graph so that, under this criterion, the resulting divisions have strong connections between data points within the same cluster and weaker connections between data points in different clusters[18].

Given a set of samples $X(x_1, x_2, \cdots, x_n)$, define the weights of the edges connecting each sample as W_{ij}, calculated using the Gaussian kernel function. The adjacency matrix W is then equal to the similarity matrix S:

$$W_{ij} = S_{ij} = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}) \quad (7)$$

The sum of the weights of all edges connected to a sample point in the graph is defined as the degree d_i of that point:

$$d_i = \sum_{j=1}^{n} W_{ij} \quad (8)$$

The degree matrix is expressed as:

$$D = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix} \quad (9)$$

Define the Laplacian matrix L as:

$$L = D - S \quad (10)$$

Then the standardized Laplace matrix L' is constructed by the following formula:

$$L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \quad (11)$$

Subsequently, compute the K smallest eigenvalues $\lambda(\lambda_1, \lambda_2, \cdots, \lambda_k)$ and their corresponding eigenvectors $V(v_1, v_2, \cdots v_k)$ of the standardized Laplacian matrix L'. Finally, perform K-means clustering on the eigenvectors V to obtain classifications $C_1, C_2, \cdots, C_k$.

D.2. Determine the Parameter σ of the Gaussian Kernel Function

When computing the adjacency matrix, the Gaussian kernel function is utilized, and the parameter σ of the Gaussian kernel function can be calculated using the Adaptive Scale method [19]. This method principally sets a scale parameter σ_i for each sample point x_i based on the concept of local density to adjust the dissimilarity between two sample points. Consequently, the dissimilarity between two sample points x_i and x_j is represented as follows:

$$s_{ij} = \frac{d(x_i, x_j)}{\sigma_i} \quad (12)$$

The dissimilarity from x_j to x_i is:

$$s_{ji} = \frac{d(x_j, x_i)}{\sigma_j} \quad (13)$$

Based on the fact that $d(x_i, x_j) = d(x_j, x_i)$, we can define a similarity function with an adaptive scale parameter σ:

$$A_{ij} = \exp(-\frac{d(x_i, x_j)^2}{\sigma_i \sigma_j}) \quad (14)$$

Where $\sigma_i = d(x_i, x_k)$ represents the distance between point x_i and its $k-th$ nearest neighbor, the magnitude of σ_i reflects the density around that point: the larger the σ_i, the further the nearby points are from sample point x_i; conversely, the smaller the σ_i, the closer they are.

D.3. Determining the Optimal Number of Clusters k in Spectral Clustering

Spectral clustering is an unsupervised clustering algorithm that requires us to specify the number of clusters k. Determining the optimal number of clusters k is an important issue in spectral clustering, and the Gap Statistic can be used for this purpose.

The Gap Statistic, proposed by Tibshirani et al. [20], helps determine the optimal number of clusters by computing the Gap Statistic corresponding to different numbers of clusters. The Gap Statistic uses the output from any clustering algorithm and compares the change in within-cluster dispersion with the expected change under a null reference distribution. The optimal number is found by maximizing the Gap Statistic.

In practice, we select the first k eigenvectors of the Laplacian matrix as input, apply the K-Means clustering algorithm for clustering, and compute the Gap Statistic. Subsequently, we plot a curve graph with the number of clusters k on the horizontal axis and the Gap Statistic on the vertical axis, allowing us to observe the change in the Gap Statistic. The optimal number of clusters is thus determined by selecting the number for which the Gap Statistic is maximal.

III. Classification Model of Tight Sandstone Gas Wells

A. Model Concept

Although research on gas well classification methods such as the unimpeded flow method, the reservoir parameter method, and the daily gas production method is relatively comprehensive, these methods, though simple and feasible, have limitations. They often fail to fully consider the production status of gas wells and can be subjective. During the production period of gas wells, all production parameter values are represented as time series data. Based on the similarity of time series, a new approach for the classification of tight sandstone gas wells is proposed. DTW is a method capable of measuring the similarity between time series of different lengths. The construction ideas for a tight sandstone gas well classification model based on the similarity of time series are as follows: First, preprocess the production parameter data; then use the DTW to calculate the parameter series distance matrix between each tight sandstone gas well within each production parameter; next, use the spectral clustering to classify tight sandstone gas wells within each production parameter; and finally, select the time series curves of three representative gas wells in each class of gas wells for analysis within each production parameter. This model is named the DTW-Spectral model.

However, the problem of singularity in DTW can result in inaccurate DTW distances, thus affecting the accuracy of time series similarity measure. To tackle this singularity problem, a modified version of DTW, called SDTW, is employed to measure time series similarity, with the aim of enhancing accuracy. Consequently, the DTW-Spectral model is optimized by substituting DTW with SDTW. The optimized model is named the SDTW-Spectral model.

Taking into account that time series data often contain noise, further optimization of the SDTW-Spectral model is performed: the preprocessed production parameter data undergo Kalman filtering before SDTW and spectral clustering are applied. The optimized model is named the KF-SDTW-Spectral model.

B. Model Construction steps

To classify tight sandstone gas wells, this paper proposes a classification model based on time series similarity: the KF-SDTW-Spectral model. The construction of this model consists of five main steps, as depicted in Fig.7. The workflow of the model is illustrated in Fig. 8.

B.1. Data Preprocessing

The various time series may have missing values at different time points, which can occur due to well shut-ins or incomplete data recording. These missing values do not reflect the true characteristics of the time series and disrupt the continuity of the entire dynamic process. Filling in missing values in time series with data does not have a special effect on subsequent similarity measure between time series using SDTW; therefore, in this case, the missing values are directly eliminated from the time series. Fig.3 and Fig.4 show the original monthly gas production time series of well W1 and the monthly gas production time series after the removal of missing values, respectively. It is evident that the elimination of missing values results in a continuous monthly gas production time series. Although the length of the time series has been reduced, the overall characteristics have not changed significantly.

Due to differences in resource abundance and production parameter magnitudes across various gas wells, it is possible for two time series with very similar production patterns to exhibit a great distance from each other, resulting in a low similarity between the time series. To ensure that each pair of time series is compared for similarity on the same scale, the min-max normalization method is employed here. This method maps all the values of each production parameter's time series into a range between 0 and 1. Additionally, the normalized time series maintain the same curve shape as the original series. The data normalization process is shown in (15). Fig.5 presents the result of normalizing the monthly gas production time series for well W1.

$$X_t' = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}} \quad (15)$$

Where, X_t represents the recorded value of the time series at time t, X_t' represents the normalized value of X_t, and X_{max} and X_{min} respectively represent the maximum and minimum values of the data in the time series.

B.2. Kalman Filtering

The normalized time series is then subjected to Kalman filtering, which improves the smoothness of the time series and reduces the influence of noise. Fig.6 illustrates the monthly gas production time series for well W1 after undergoing Kalman filtering.

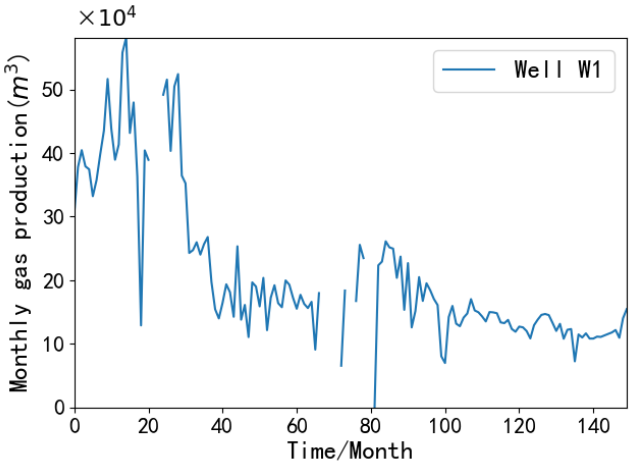

Fig. 3. Original monthly gas production time series of well W1

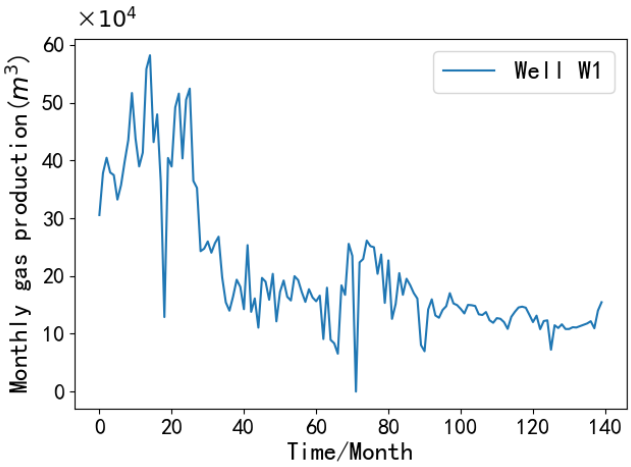

Fig .4. Monthly gas production time series of well W1 after removing missing values

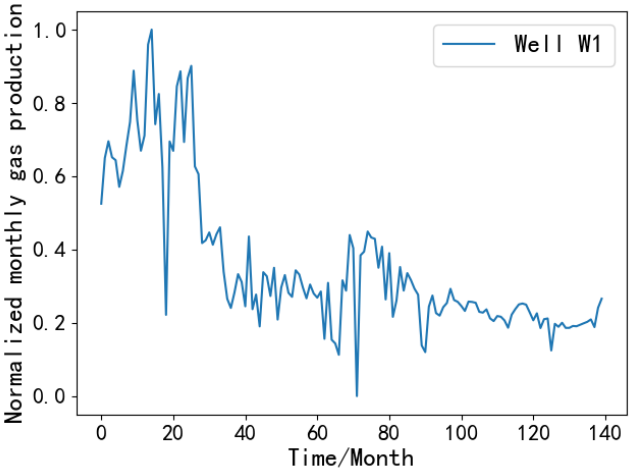

Fig .5. Normalized monthly gas production time series of well W1 after removing missing values

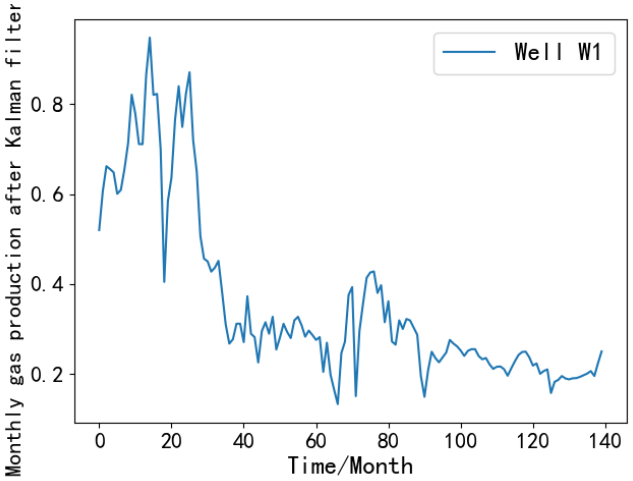

Fig.6. Monthly gas production time series of well W1 after Kalman filtering

B.3. Calculate the SDTW Distance Matrix

After removing missing values, normalization, and Kalman filter processing of each production parameter time series, the parameter series distances between tight sandstone gas wells are calculated using SDTW for each parameter. The distances between all wells under this parameter condition are then orderly stored as an inter-well series distance matrix. It is evident that this distance matrix is symmetric.

B.4. Spectral Clustering

The spectral clustering is used to transform the obtained distance matrix into an adjacency matrix. Subsequently, the Laplacian matrix is computed and normalized. The eigenvalues and eigenvectors of the normalized Laplacian matrix are then calculated. The Gap Statistic is employed to determine the optimal number of clusters in spectral clustering, which allows for the classification of tight sandstone gas wells based on each production parameter.

B.5. Time Series Curve Analysis

Since each class of gas wells for each production parameter has a large number, corresponding to numerous time series with fluctuating changes and varying lengths, displaying all time series of each class of wells on one chart would appear cluttered and chaotic, making it impossible to discern the characteristics of the time series curves. Therefore, in this paper, within each class of wells, the time series curves for three representative wells are selected to display, and a detailed analysis of these selected time series curves is conducted.

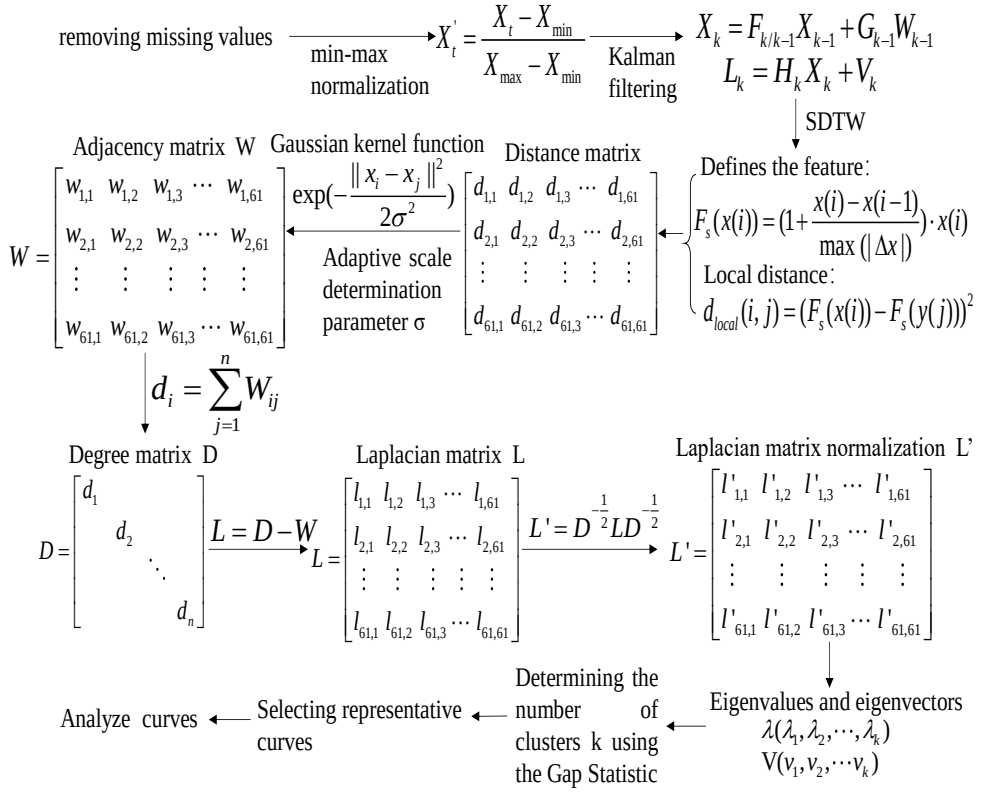

Fig.7.KF-SDTW-Spectral model structure

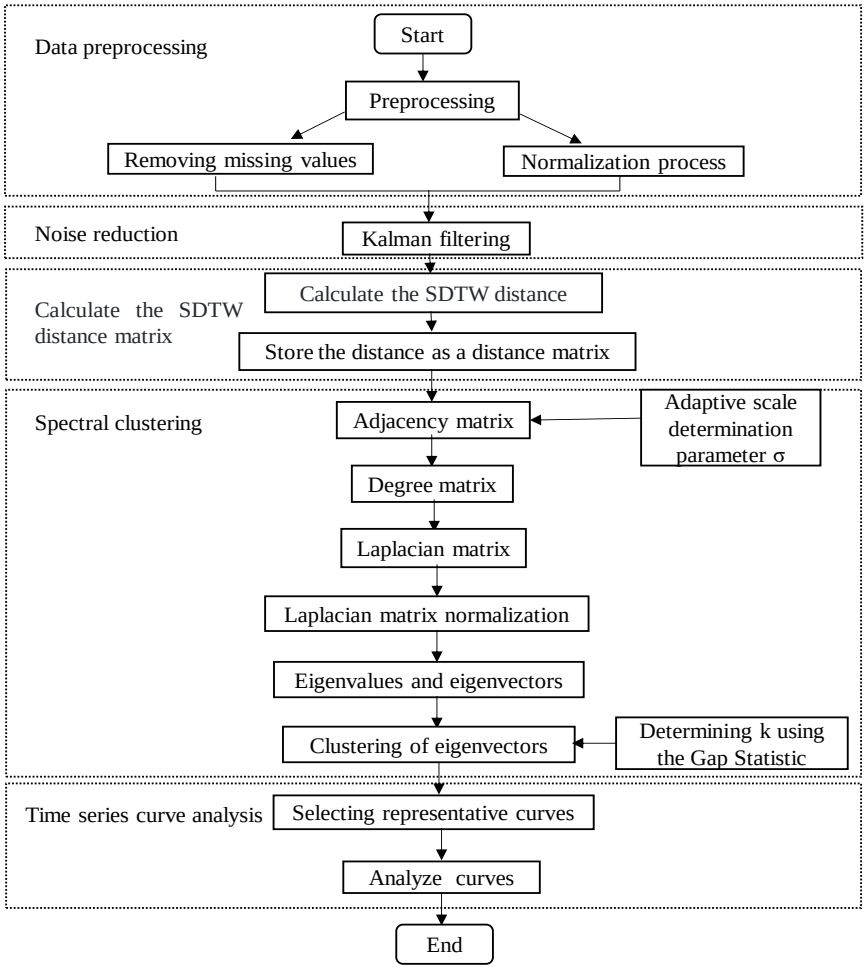

Fig.8.Flowchart of the KF-SDTW-Spectral model

IV. Case Analysis

A. Classification of Tight Sandstone Gas Wells - Monthly Gas Production

In this paper, spectral clustering is performed based on the distances between the monthly gas production series of tight sandstone gas wells, which have been calculated using SDTW after Kalman filtering. The Gap Statistic is illustrated in Fig.12. As the Gap Statistic reaches its maximum when the number of clusters is two, the 61 tight sandstone gas wells are divided into two clusters. These clusters are defined as the gradual decline-trailing pattern and the slow rise-rapid decline-trailing pattern, with 29 wells in the gradual decline-trailing pattern and 32 wells in the slow rise-rapid decline-trailing pattern. In this paper, three representative wells from each of the two patterns are selected, and their monthly gas production curves are presented in Fig.13 and Fig.14, respectively.

Fig.12.Gap Statistic

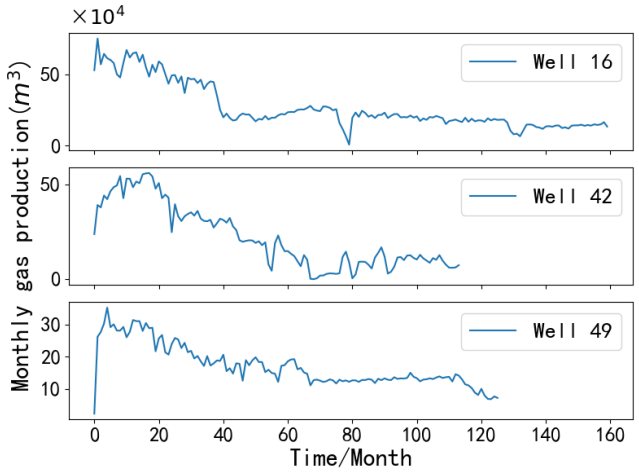

Fig.13. Monthly gas production curve of tight sandstone gas wells in the gradual decline-trailing pattern

As shown in Fig.13, the monthly gas production of tight sandstone gas wells in the gradual decline-trailing pattern rapidly increases to a peak within a short period, then fluctuates downward to lower values. Subsequently, it increases again to about one-third of the peak, followed by relatively minor fluctuations at around one-third of the peak for approximately half of the production period until production ceases. The peak values of monthly gas production for wells in this pattern typically occur within the first 20% of the entire gas production phase.

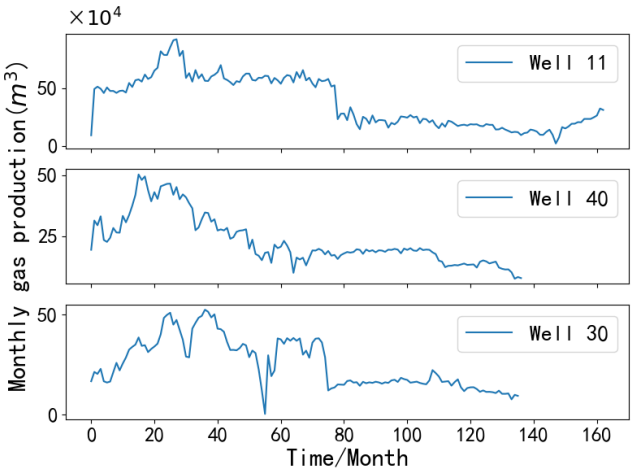

Fig.14.Monthly gas production curve of tight sandstone gas wells in the slow rise-rapid decline- trailing pattern

As shown in Fig.14, the monthly gas production of tight sandstone gas wells in the slow rise-rapid decline-trailing pattern gradually increases from lower values to a peak, then fluctuates downward to lower values. Subsequently, it fluctuates within a relatively small range around these lower values for about half of the production time until production ceases. The peak values of monthly gas production for wells in this pattern typically occur within the first 30% of the entire gas production phase.

B. Classification of Tight Sandstone Gas Wells - Monthly Water Production

In this paper, spectral clustering is performed based on the distances between the monthly water production series of tight sandstone gas wells, which have been calculated using SDTW after Kalman filtering. The Gap Statistic is illustrated in Fig.15. As the Gap Statistic reaches its maximum when the number of clusters is three, the 61 tight sandstone gas wells are divided into three clusters. These clusters are defined as the rapid decline-trailing pattern, the fluctuating multi-peak pattern, and the rapid decline-stable pattern, with 19 wells in the rapid decline-trailing pattern,36 wells in the fluctuating multi-peak pattern, and 6 wells in the rapid decline-stable pattern. In this paper, three representative wells from each of the three patterns are selected, and their monthly water production curves are presented in Fig.16, 17, and 18, respectively.

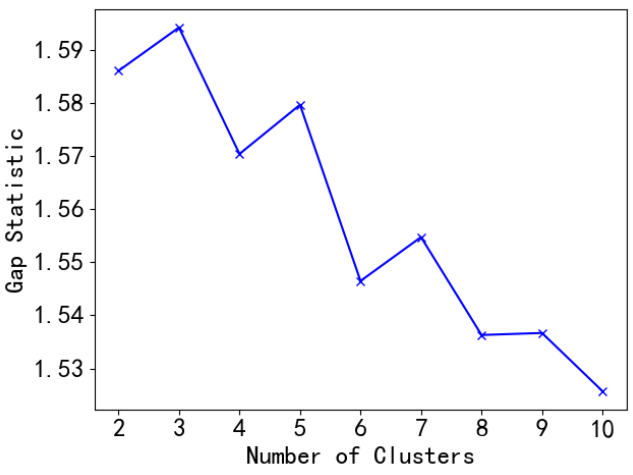

Fig.15. Gap Statistic

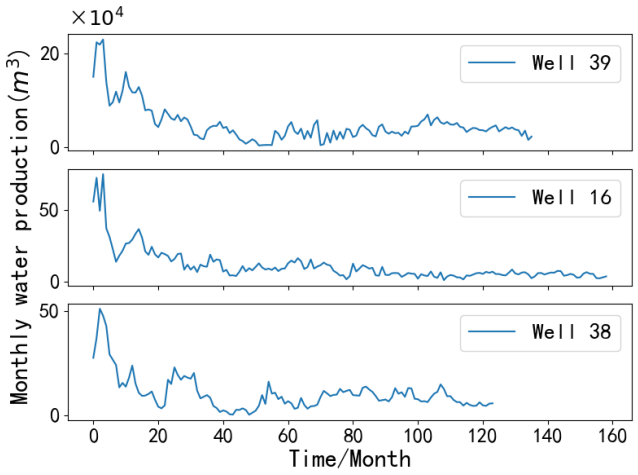

Fig.16. Monthly water production curve of tight sandstone gas wells in the rapid decline- trailing pattern

As shown in Fig.16, the monthly water production of tight sandstone gas wells in the rapid decline-trailing pattern rapidly increases from high values to a peak, then rapidly decreases to higher values. Subsequently, the monthly water production fluctuates with significant variations, followed by continuing fluctuations within a lower range for about half of the production time until production ceases, with smaller variations compared to the first half of the period. The peak values of monthly water production for wells in this pattern typically occur within the first 10% of the entire water production phase.

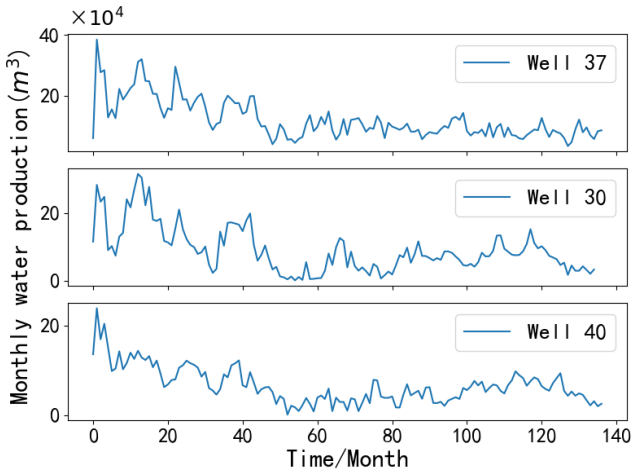

Fig.17. Monthly water production curve of tight sandstone gas wells in the fluctuating multi-peak pattern

As shown in Fig.17, the monthly water production of tight sandstone gas wells in the fluctuating multi-peak pattern rapidly increases to high values, then rapidly decreases to higher values. Subsequently, it fluctuates upwards to high values, then fluctuates downwards, and then continues to fluctuate upwards to higher values, followed by fluctuating downwards again, continuing these fluctuations until production ceases. The monthly water production undergoes frequent and large variations, with multiple local peaks throughout the entire water production phase. The peak values of monthly water production for wells in this pattern typically occur within the first 10% of the entire water production phase.

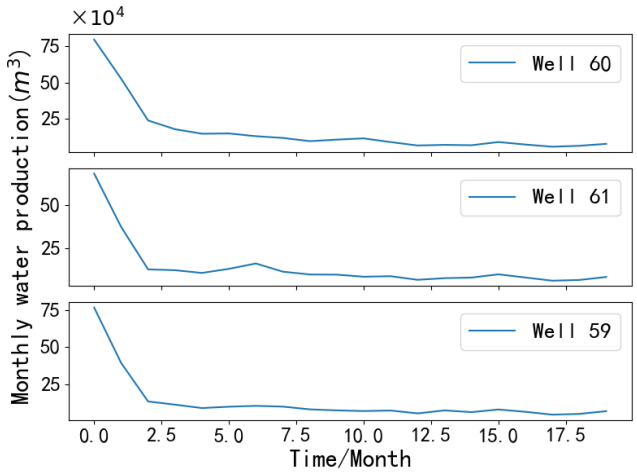

Fig.18. Monthly water production curve of tight sandstone gas wells in the rapid decline-stable pattern

As shown in Fig.18, the monthly water production of tight sandstone gas wells in the rapid decline-stable pattern decreases from peak values at a relatively fast rate to low values, after which it remains stable within this low range for approximately 6/7 of the production time until production ceases. The peak values of monthly water production for wells in this pattern occur at the beginning of the entire production phase.

C. Classification of Tight Sandstone Gas Wells - Oil Pressure

In this paper, spectral clustering is performed based on the distances between the oil pressure series of tight sandstone gas wells, which have been calculated using SDTW after Kalman filtering. The Gap Statistic is illustrated in Fig.19. As the Gap Statistic reaches its maximum when the number of clusters is three, the 61 tight sandstone gas wells are divided into three clusters. These clusters are defined as the gradual decline-trailing pattern, the gradual decline-stable-trailing pattern, and the gradual decline-fluctuating-trailing pattern, with 9 wells in the gradual decline-trailing pattern,25 wells in the gradual decline-stable-trailing pattern, and 27 wells in the gradual decline-fluctuating-trailing pattern. In this paper, three representative wells from each of the three patterns are selected, and their oil pressure curves are presented in Fig.20, 21, and 22, respectively.

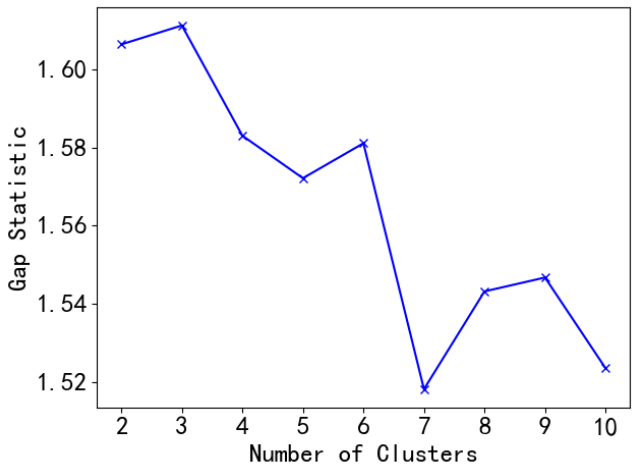

Fig.19. Gap Statistic

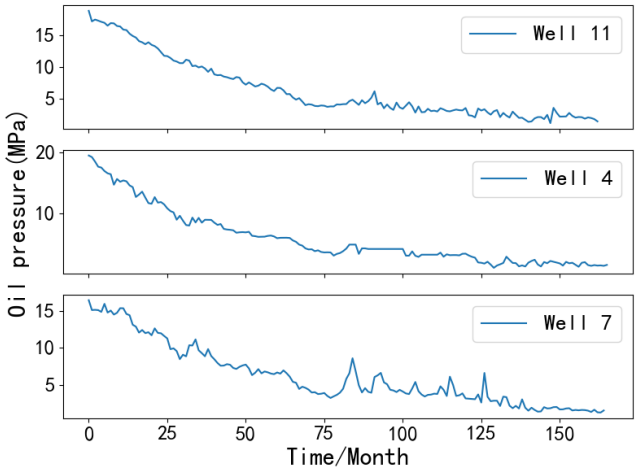

Fig.20. Oil pressure curve of tight sandstone gas wells in the gradual decline-trailing pattern

As shown in Fig. 20, the oil pressure of tight sandstone gas wells in the gradual decline-trailing pattern decreases from peak values with relatively small fluctuations to lower values, then rapidly increases to higher values, followed by a rapid decrease to lower values. Subsequently, for approximately half of the production time, the oil pressure fluctuates within this lower range with relatively small amplitudes until production ceases. The peak values of oil pressure for wells in this pattern occur at the beginning of the entire production phase.

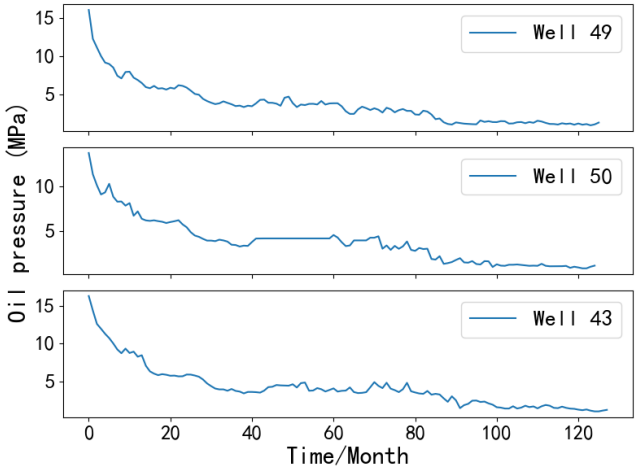

Fig.21. Oil pressure curve of tight sandstone gas wells in the gradual decline-stable-trailing pattern

As shown in Fig. 21, the oil pressure of tight sandstone gas wells in the gradual decline-stable-trailing pattern decreases gradually from peak values to lower values, then fluctuates within this lower range with relatively small amplitudes. Subsequently, the pressure fluctuates downward to a low value, and then remains relatively stable within this low range for approximately one-fourth of the production time until production ceases. The peak values of oil pressure for wells in this pattern occur at the beginning of the entire production phase.

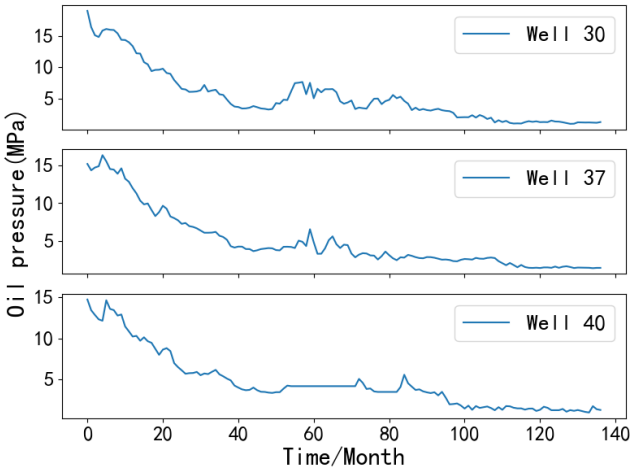

Fig.22.Oil pressure curve of tight sandstone gas wells in the gradual decline-fluctuating-trailing pattern

As shown in Fig.22, the oil pressure of tight sandstone gas wells in the gradual decline-fluctuating-trailing pattern rapidly decreases from high values to lower values, then quickly increases to high values again, followed by a fluctuation downward to lower values. Subsequently, it fluctuates within this lower range with relatively small amplitudes, then fluctuates downward again to a low value. Afterward, the pressure remains relatively stable within this low range for approximately two-sevenths of the production time until production ceases. The peak values of oil pressure for wells in this pattern typically occur within the first 10% of the entire production phase.

To summarize, based on monthly gas production as a classification indicator, the 61 tight sandstone gas wells are divided into two clusters: the gradual decline-trailing pattern and the slow rise-rapid decline-trailing pattern. The monthly gas production curves of these two patterns of wells exhibit distinct differences. The overall trend of the monthly gas production curve of the gradual decline-trailing pattern wells is rising-declining- stabilizing. The overall trend of the slow rise-rapid decline-trailing pattern wells is rising-declining-stabilizing-declining-stabilizing.

Based on monthly water production as a classification indicator, the 61 tight sandstone gas wells are divided into three clusters: the rapid decline-trailing pattern, the fluctuating multi-peak pattern, and the rapid decline-stable pattern. The monthly water production curves of these three patterns of wells exhibit distinct differences. The overall trend of the monthly water production curve of the rapid decline-trailing pattern wells is rising-declining-rising-declining-stabilizing. The overall trend of the fluctuating multi-peak pattern wells is alternating rising-declining. The overall trend of the rapid decline-stable pattern wells is declining- stabilizing.

Based on oil pressure as a classification indicator, the 61 tight sandstone gas wells are divided into three clusters: the gradual decline-trailing pattern, the gradual decline-stable-trailing pattern, and the gradual decline-fluctuating-trailing pattern. The oil pressure curves of these three patterns of wells exhibit distinct differences. The overall trend of the oil pressure curve of the gradual decline-trailing pattern wells is declining-rising-declining-stabilizing. The overall trend of the gradual decline-stable-trailing pattern wells is declining-stabilizing-

declining-stabilizing. The overall trend of the gradual decline-fluctuating-trailing pattern wells is declining-rising-declining stabilizing-declining- stabilizing.

D. Model Evaluation Index

The primary goal of clustering is to divide a set of data samples into multiple classes such that there is a high resemblance within each class and a low similarity between different classes[21]. In other words, the smaller the intra-class distance and the larger the inter-class distance, the higher the quality of the clustering. Therefore, this paper adopts the ratio of intra-class distance to inter-class distance as an indicator for evaluating the quality of clustering.

D.1. Intra-Class Distance

In this paper, the minimum value of the average distance between each object in a class and all other objects in the same class is taken as the intra-class distance for that class, while the intra-class distance of the entire sample data is the maximum value among all classes' intra-class distances. This intra-class distance of the entire sample data is denoted as $intra(k)$.

$$\text{intra(k)} = \max_{1\le i\le k}\{\min_{1\le j\le |C_i|}\{\frac{1}{|C_i|-1}\sum_{p=1,p\neq j}^{|C_i|}\| x_j - x_p \|\}\} \quad (16)$$

Where: k is the number of clusters, $intra(k)$ represents the intra-class distance of the entire sample data; $|C_i|$ is the number of objects in class C_i; x_j and x_p are objects belonging to class C_i.

The smaller the value of $intra(k)$, the more similar the data samples within the class are, and the better the clustering effect. Therefore, if the class with the maximum intra-class distance already meets the requirement for intra-class similarity, then it is certain that the other classes also meet this requirement. For this reason, this paper takes the maximum intra-class distance among all classes as the intra-class distance for the entire sample data.

D.2. Inter-Class Distance

In this paper, the distance between two classes is defined as the minimum value among the distances between objects in those two classes, and the inter-class distance of the entire sample data is the minimum value among all distances between any two classes. The inter-class distance of the entire sample data is denoted as $inter(k)$.

$$\text{inter(k)} = \min_{x_p\in C_i, x_q\in C_j, i\neq j}\| x_p - x_q \| \quad (17)$$

Where: k is the number of clusters, $i=1,2,\ldots,k$; $j=1,2,\ldots,k$; and $i\neq j$. x_p and x_q are objects within clusters C_i and C_j, respectively.

The larger the value of $inter(k)$, the less similar the data samples between different classes are, and the better the clustering effect. Therefore, if the minimum distance between two classes already meets the requirement of being sufficiently dissimilar, then it is certain that all other pairs of classes also meet this requirement. For this reason, this paper takes the minimum distance between any two classes as the inter-class distance for the entire sample data.

D.3. Cluster Validity Evaluation Index

$$\text{C(k)} = \frac{\text{inter(k)}}{\text{intra(k)}} \quad (18)$$

k is the number of clusters, and the smaller the evaluation index $C(k)$, the higher the cluster quality.

E. Comparative Analysis of Experimental Results

To assess whether the clustering quality of the KF-SDTW-Spectral model has improved, the study continues using monthly gas production, monthly water production, and oil pressure as classification indicators. The SDTW-Spectral and DTW-Spectral models were applied to classify tight sandstone gas wells. Additionally, traditional clustering methods were utilized to classify tight sandstone gas wells. This paper selected three traditional clustering methods: k-means clustering[22],hierarchical clustering[23], and DBSCAN[24].However, the k-means method is not suitable for directly processing time series data of unequal lengths because it requires calculating Euclidean distances between all time series, which in turn requires that all time series have the same length. Therefore, the study initially calculates the DTW distances between all time series to obtain a DTW distance matrix. Subsequently, the smallest k eigenvalues from the distance matrix, along with their corresponding eigenvectors, are identified. The eigenvectors are then clustered using k-means, thus classifying the tight sandstone gas wells. This model is named the DTW-k-means model. After the DTW distances between all time series are computed by DTW, hierarchical clustering is used to classify the tight sandstone gas wells. This model is named the DTW-Hierarchical model. The parameters of DBSCAN, namely the neighborhood radius and minimum number of samples, can affect the clustering results. Therefore, after computing the DTW distances between all time series, we employ grid search[25] to find the optimal parameters for DBSCAN. Subsequently, we utilize DBSCAN to classify the tight sandstone gas wells. This model is named the DTW-DBSCAN model. The cluster validity evaluation index $C(k)$ for all models is obtained and compared. The experimental results are shown in Table I.

TABLE I
COMPARISON OF MODEL EVALUATION INDEX

Models	Classification indicators		
	Monthly gas production	Monthly water production	Oil pressure
DTW- Hierarchical	7.97	8.13	9.90
DTW-K-means	7.85	7.73	11.31
DTW-DBSCAN	6.89	7.51	9.78
DTW- Spectral	5.01	7.04	9.42
SDTW-Spectral	3.33	6.37	8.31
KF-SDTW-Spectral	2.66	4.43	6.59

TABLE II
GAS WELL CLASSIFICATION RESULTS

Well number	Monthly gas production /10^4 m^3	Monthly water production /10^4 m^3	Oil pressure /MPa	Type
W1	slow rise-rapid decline-trailing pattern	rapid decline-trailing pattern	gradual decline-trailing pattern	low-yield well
W2	slow rise-rapid decline-trailing pattern	fluctuating multi-peak pattern	gradual decline-fluctuating-trailing pattern	low-yield well
W3	slow rise-rapid decline-trailing pattern	fluctuating multi-peak pattern	gradual decline-trailing pattern	high-yield well
W4	gradual decline-trailing pattern	rapid decline-trailing pattern	gradual decline-stable-trailing pattern	high-yield well
...				
W58	slow rise-rapid decline-trailing pattern	rapid decline-stable pattern	gradual decline-stable-trailing pattern	low-yield well
W59	gradual decline-trailing pattern	rapid decline-trailing pattern	gradual decline-fluctuating-trailing pattern	high-yield well
W60	slow rise-rapid decline-trailing pattern	rapid decline-stable pattern	gradual decline-trailing pattern	high-yield well
W61	gradual decline-trailing pattern	fluctuating multi-peak pattern	gradual decline-stable-trailing pattern	low-yield well

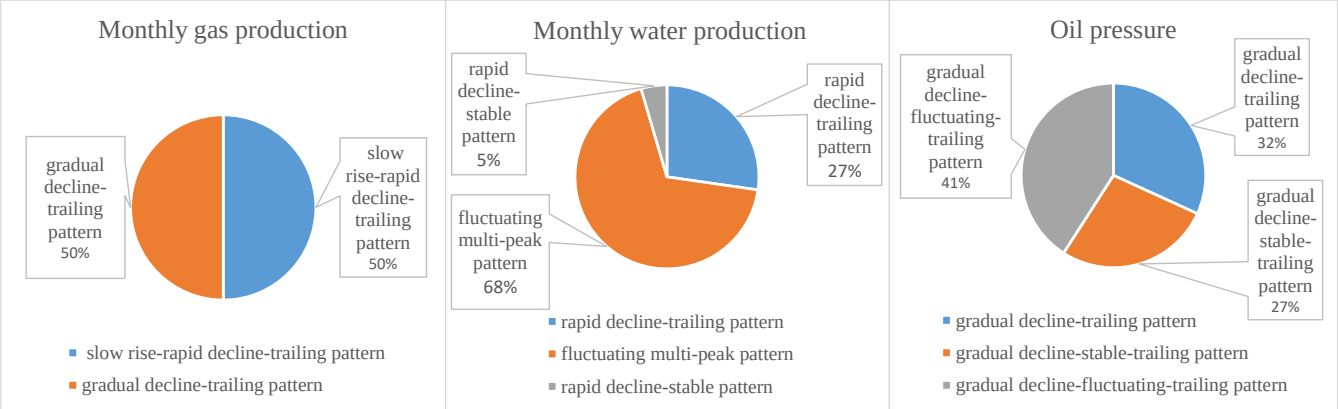

Fig.23. Production characteristics of high-yield wells

Based on monthly gas production as a classification indicator, tight sandstone gas wells are divided into two clusters. The clustering quality of DTW-Spectral, SDTW-Spectral, and KF-SDTW-Spectral models are all superior to that of traditional clustering methods. Among them, the KF-SDTW-Spectral model has the highest clustering quality.

Based on monthly water production as a classification indicator, tight sandstone gas wells are divided into three clusters. The clustering quality of DTW-Spectral, SDTW-Spectral, and KF-SDTW-Spectral models are all superior to that of traditional clustering methods. Among them, the KF-SDTW-Spectral model has the highest clustering quality.

Based on oil pressure as a classification indicator, tight sandstone gas wells are divided into three clusters. The clustering quality of DTW-Spectral, SDTW-Spectral, and KF-SDTW-Spectral models are all superior to that of traditional clustering methods. Among them, the KF-SDTW-Spectral model has the highest clustering quality.

To summarize, compared to traditional clustering methods, the DTW-Spectral, SDTW-Spectral, and KF-SDTW-Spectral models proposed in this paper have higher clustering quality. Among these, the KF-SDTW-Spectral model exhibits the highest clustering quality, making the classification of tight sandstone gas wells more scientific and rational.

F. Gas Well Production Characteristics and Production Measures

Based on the actual production situation of the gas field, this paper classifies 61 tight sandstone gas wells into high-yield and low-yield wells according to their cumulative gas production. Gas wells with a cumulative gas production of not less than 320 million cubic meters are considered high-yield wells, while those with a cumulative gas production of less than 320 million cubic meters are considered low-yield wells. Then, according to the clustering results of the KF-SDTW-Spectral model, the classification of the 61 gas wells is shown in Table II.

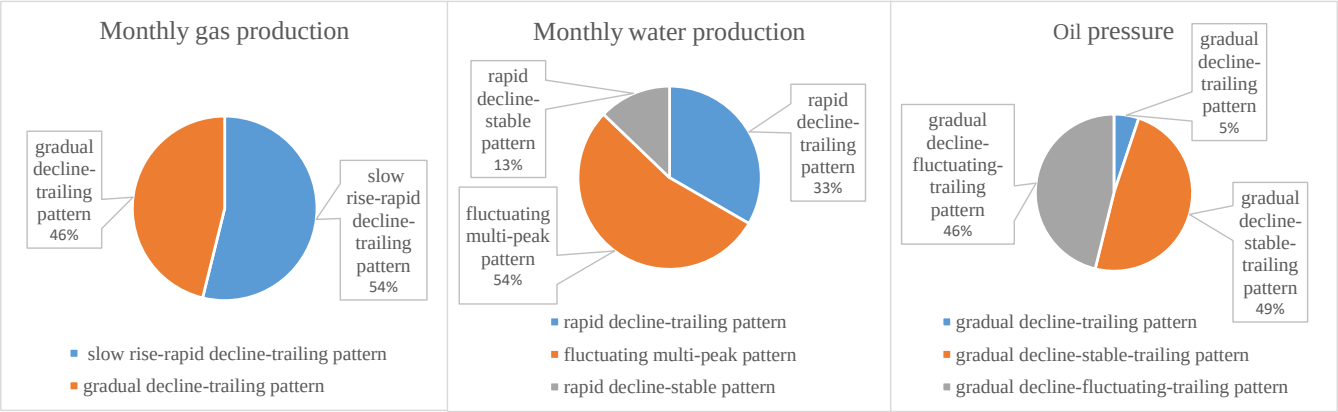

Fig.24. Production characteristics of low-yield wells

TABLE III
GAS WELL PRODUCTION MEASURES

Type	Monthly gas production /10^4 m^3	Monthly water production /10^4 m^3	Oil pressure /MPa	Production measures
High-yield wells	slow rise-rapid decline-trailing pattern	fluctuating multi-peak pattern	gradual decline-trailing pattern	I
Low-yield wells	slow rise-rapid decline-trailing pattern	fluctuating multi-peak pattern	gradual decline-stable-trailing pattern	II
Low-yield wells	gradual decline-trailing pattern	fluctuating multi-peak pattern	gradual decline-stable-trailing pattern	III

From Fig.23, it can be observed that the monthly gas production of high-yield wells exhibits two patterns, with each pattern accounting for 50% of the wells. This indicates that the monthly gas production of high-yield wells follows either the gradual decline-trailing pattern or the slow rise-rapid decline-trailing pattern. The monthly water production of high-yield wells shows three patterns, with the fluctuating multi-peak pattern being the most prevalent, at nearly 70%. It is followed by the rapid decline-trailing pattern, which accounts for close to 30%, and the rapid decline-stable pattern, making up about 5%. This indicates that monthly water production in high-yield wells is predominantly characterized by the fluctuating multi-peak pattern, followed by the rapid decline-trailing pattern. The oil pressure of high-yield wells also exhibits three patterns: the gradual decline-fluctuating-trailing pattern accounts for over 40%, the gradual decline-trailing pattern for over 30%, and the gradual decline-stable-trailing pattern for less than 30%. This indicates that oil pressure in high-yield wells is mainly characterized by the gradual decline-fluctuating-trailing pattern, followed by the gradual decline-trailing pattern and the gradual decline-stable-trailing pattern.

From Fig. 24, it can be observed that the monthly gas production of low-yield wells exhibits two patterns, with each pattern roughly accounting for 50%. This indicates that the monthly gas production of low-yield wells follows either the gradual decline-trailing pattern or the slow rise-rapid decline-trailing pattern. The monthly water production of low-yield wells shows three patterns: the fluctuating multi-peak pattern comprises more than 50%, the rapid decline-trailing pattern accounts for over 30%, and the rapid decline-stable pattern is the least common, at only 13%. This indicates that the primary pattern for monthly water production in low-yield wells is the fluctuating multi-peak pattern, followed by the rapid decline-trailing pattern. The oil pressure of low-yield wells also exhibits three patterns: the gradual decline-stable-trailing pattern and the gradual decline-fluctuating-trailing pattern each account for more than 40%, whereas the gradual decline-trailing pattern is at a mere 5%. This indicates that the oil pressure in low-yield wells is primarily characterized by either the gradual decline-stable-trailing pattern or the gradual decline-fluctuating-trailing pattern.

Based on the production characteristics of high-yield and low-yield wells, appropriate production measures can be determined. For high-yield wells, the monthly gas production follows either the gradual decline-trailing pattern or the slow rise-rapid decline-trailing pattern, with the monthly water production displaying the fluctuating multi-peak pattern, and the oil pressure exhibiting either the gradual decline-fluctuating-trailing pattern or the gradual decline-trailing pattern. For low-yield wells, the monthly gas production follows either the gradual decline-trailing pattern or the slow rise-rapid decline-trailing pattern, with the monthly water production also in the fluctuating multi-peak pattern, and the oil pressure characterized by either the gradual decline-stable-trailing pattern or the gradual decline-fluctuating-trailing pattern. The corresponding production measures are shown in Table III, contingent upon the actual situation.

From Table III, it is evident that high-yield wells correspond to production measure I, while low-yield wells correspond to production measures II and III. In other words, production measure I is intended for high-yield wells, whereas production measures II or III are intended for low-yield wells. According to Table II, there are 6 gas wells with production measure I, out of which 5 are high-yield wells, resulting in an accuracy rate of 83.3%. There are 3 gas wells with production measure II, all of which are low-yield wells, resulting in a 100% accuracy rate. There are 8 gas

wells with production measure III, out of which 6 are low-yield wells, leading to a 75% accuracy rate.

To summarize, for a certain gas field in Southwest China, if the aim is to ultimately achieve high-yield wells, it is advisable to adopt production measure I, as the probability of obtaining high-yield wells through this measure is high. Conversely, if the goal is to achieve low-yield wells, production measures II or III can be employed. There is also a high probability of achieving low-yield wells through these two measures.

V. Conclusion

This paper presents a classification model for tight sandstone gas wells based on time series similarity: the KF-SDTW-Spectral model. This model classifies the tight sandstone gas wells in a certain gas field in Southwest China scientifically and reasonably. By utilizing this model, the production characteristics of high-yield and low-yield wells are analyzed, and corresponding production measures are proposed, offering valuable references for gas well management. The research subjects of this paper include 61 tight sandstone gas wells, leading to the following conclusions.

(1) The 61 tight sandstone gas wells are divided into two clusters based on monthly gas production as a classification indicator, three clusters based on monthly water production, and three clusters based on oil pressure.

(2) According to the cluster validity assessment index, the KF-SDTW-Spectral model demonstrated the highest clustering quality and outperformed traditional clustering methods.

(3) The production characteristics of high-yield and low-yield wells have been summarized. For high-yield wells, production measure I is proposed, while for low-yield wells, production measures II and III are proposed.

To summarize, the model demonstrates high-quality clustering, offering a novel approach for the classification of tight sandstone gas wells, which holds significant reference value for guiding the development of the gas field.

REFERENCES

[1] Ailin J , Yunsheng W ,Zhi G , et al, "Development status and prospect of tight sandstone gas in China," Natural Gas Industry B,2022,9(5):467-476.

[2] Li, TT (Li, Tiantai), and Huang, X (Huang, Xing), "Classification of horizontal wells based on dynamic data and its application in ultra-low permeability gas reservoirs," Chemistry and Technology of Fuels and Oils,2017,Vol.53(1): 123-134

[3] Xiaoping Zhu, Zongmin Ma, and Qijie Tang, "UK - Means Clustering for Uncertain Time Series Based on ULDTW Distance," .Intelligent Data Engineering and Automated Learning – IDEAL 2017,2017,Vol.10585: 27-35

[4] Liao, and T.W, "Clustering of time series data-A survey(Review) ,"Pattern Recognition,2005,Vol.38(11): 1857-1874

[5] Folgado Duarte (duarte.folgado@fraunhofer.pt), Barandas Marília, Matias Ricardo, Martins Rodrigo, Carvalho Miguel, and Gamboa Hugo, "Time Alignment Measurement for Time Series," Pattern Recognition,2018,Vol.81: 268-279

[6] Yi-Leh Wu, Divyakant Agrawal, and Amr El Abbadi, " A comparison of DFT and DWT based similarity search in time-series databases,"9th International Conference on Information Knowledge Management(CIKM)2000[C],2001

[7] Joan Serrà, and Josep Ll. Arcos, "An empirical evaluation of similarity measures for time series classification ,"Knowledge-Based Systems,2014,Vol.67: 305-314

[8] Fu, and TC (Fu, Tak-chung), "A review on time series data mining ,"ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE,2011,Vol.24(1): 164-181

[9] Yuval Burstyn, Asaf Gazit, and Omri Dvir, "Hierarchical Dynamic Time Warping methodology for aggregating multiple geological time series ,"Computers & Geosciences,2021,Vol.150: 104704

[10] A. Troncoso, M. Arias, and J.C. Riquelme, "A multi-scale smoothing kernel for measuring time-series similarity ,"Neurocomputing,2015,Vol.167: 8-17

[11] Alon, J., Athitsos V., Yuan Q., and Sclaroff S., "A unified framework for gesture recognition and spatiotemporal gesture segmentation(Article) ,"IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,Vol.31(9): 1685-1699

[12] Gupta, L., Molfese D.L., Tammana R., and Simos P.G., "Nonlinear alignment and averaging for estimating the evoked potential ,"IEEE Transactions on Biomedical Engineering,1996,Vol.43(4): 348-356

[13] Eamonn Keogh[1], Li Wei[1], Xiaopeng Xi[1], Michail Vlachos[2],Sang-Hee Lee[3],and Pavlos Protopapas[4], "Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures(Article) ,"VLDB Journal,2009,Vol.18(3): 611-630

[14] E. J. Keogh, and M. J. Pazzani, "Derivative Dynamic Time Warping," in In Proceedings of the 1st SIAM international conference on data mining, 2002.

[15] L. Benedikt, V. Kajic, D. Cosker, P.L. Rosin, D. Marshall, M. Everingham, and C. Needham, "Facial Dynamics in Biometric Identification," British Machine Vision Conference 2008,2008,

[16] SHEN Jingyi, ZHU Dongyang, HUANG Weiping, and LIANG Jun, "A Novel Similarity Measure Approach for Time Series based on PLA and DTW," The 35th Chinese Control Conference [C],2016

[17] R. Kazemi, A. Farsi, M.H. Ghaed, and M. Karimi-Ghartemani, "Detection and extraction of periodic noises in audio and biomedical signals using Kalman filter," Signal Processing,2008,Vol.88(8): 2114-2121

[18] Seyoung P, and Hongyu Z, "Spectral clustering based on learning similarity matrix," Bioinformatics (Oxford, England),2018,34(12).

[19] Zelnijk-Manor, and Pietro Perona, "Self-tuning spectral clustering,"//Advances in Neural Information Processing Systems. Cambridge, USA:MIT Press,2004:1601-1608.

[20] Tibshirani R, Walther G, and Hastie T, "Estimating the Number of Clusters in a Data Set via the Gap Statistic," Journal of the Royal Statistical Society. Series B (Statistical Methodology),2001,63(2).

[21] Wu, and Edmond Hao Cun, "Independent component analysis for clustering multivariate time series data," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),2005,Vol.3584(1): 474-482

[22] MacQueen J., "Some methods for classification and analysis of multivariate observations," Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66),

[23] Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254

[24] Binbin Sun, Wentao Li, Huibin Liu, Pengwei Wang, Song Gao, and Penghang Feng, "Mathematical Method for Lidar-based Obstacle Detection of Intelligent Vehicle," IAENG Internaitonal Journal of Computer Science,vol.48,no.1, pp181-189,2021

[25] Chin Poo Lee, and Kian Ming Lim, "MFRD-80K: A Dataset and Benchmark for Masked Face Recognition," Engineering Letters, vol.29, no.4, pp1595-1600,2021