water

MDPI

Article

Application of Geologically Constrained Machine Learning Method in Characterizing Paleokarst Reservoirs of Tarim Basin, China

Wei Xin [1,*], Fei Tian [2,3,4], Xiaocai Shan [2,3,4], Yongjian Zhou [2,3,4], Huazhong Rong [1] and Changchun Yang [2,3,4]

[1] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; 2019210447@mail.buct.edu.cn
[2] Key Laboratory of Petroleum Resources Research, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China; tianfei@mail.iggcas.ac.cn (F.T.); shxc@mail.iggcas.ac.cn (X.S.); zyj@mail.iggcas.ac.cn (Y.Z.); ccy@mail.iggcas.ac.cn (C.Y.)
[3] Institutions of Earth Science, Chinese Academy of Sciences, Beijing 100029, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
[*] Correspondence: xinwei@mail.buct.edu.cn; Tel.: +86-156-0062-4834

Received: 25 May 2020; Accepted: 16 June 2020; Published: 21 June 2020

Abstract: As deep carbonate fracture-cavity paleokarst reservoirs are deeply buried and highly heterogeneous, and the responded seismic signals have weak amplitudes and low signal-to-noise ratios. Machine learning in seismic exploration provides a new perspective to solve the above problems, which is rapidly developing with compelling results. Applying machine learning algorithms directly on deep seismic signals or seismic attributes of deep carbonate fracture-cavity reservoirs without any prior knowledge constraints will result in wasted computation and reduce the accuracy. We propose a method of combining geological constraints and machine learning to describe deep carbonate fracture-cavity paleokarst reservoirs. By empirical mode decomposition, the time–frequency features of the seismic data are obtained and then a sensitive frequency is selected using geological prior constraints, which is input to fuzzy C-means cluster for characterizing the reservoir distribution. Application on Tahe oilfield data shows the potential of highlighting subtle geologic structures that might otherwise escape unnoticed by applying machine learning directly.

Keywords: karst system; machine learning; Hilbert-Huang transform; waveform cluster; seismic interpretation; paleo-channels

1. Introduction

The Ordovician carbonate paleo-water system in Tahe Oilfield is mainly formed in the Hercynian period. The Ordovician fracture-cavity carbonate reservoir is formed by dissolution, with an extremely heterogeneous internal structure and non-uniform lithological properties [1–4]. The karst paleo-channel in the karst water system area is the core of the entire karst reservoir system. The caves with different scales and forms are closely related to the ancient river channel. In general, seismic interpretation acquires subsurface information from seismic data and reveals geologic meanings. Because of the burial depth of the Ordovician carbonate strata, exceeding 5500 m [2–5], the seismic reflection signals of the unconformity surface are weak and discontinuous, with a relatively low signal-noise ratio. The weak reflection amplitude characteristics of seismic data cannot provide a better explanation for seismic attribute analysis. Phase information can detect subtle changes in the subsurface, but it is sensitive to noise, thereby masking certain stratigraphic features in full-bandwidth data. Therefore, it is difficult to characterize the sedimentary facies of carbonate reservoirs and predict their distribution through single-factor geophysical or geological methods.

Although in-depth understanding and training of geophysics is considered a prerequisite for developing effective interpretive workflows, recent advances in machine learning have provided new insights into its role in this particular domain problem, such as self-organizing diagrams [6,7], multilayer perceptron [8,9], and K-means clustering [10,11]. Machine learning models can interpret large amounts of data and understand the relationships of various types of data at the same time, providing repeatable and reliable results for seismic interpretation [12].

For the interpretation method based on machine learning, what is most important is to extract useful seismic features based on the interpreter's domain knowledge and available well logs. The sparsely available well logs have high resolution along the depth, whereas the 3-D seismic attributes have low resolution along the depth. Compared to the log curve, the smooth seismic attributes have lower lithological information content. The difference of information content between the lithological properties and the seismic attributes necessitates an information filtering scheme [13]. Therefore, using log-based curves as geological constraints and extracting features that best match the real geological information by signal processing tools can more accurately characterize the reservoir [14].

As for seismic features, time–frequency decomposition maps a 1D signal of time into a 2D image of frequency and time, which describes how the frequency content varies with time. Empirical mode decomposition (EMD), developed by Huang et al. (1998), is a powerful signal frequency–time analysis technique for nonstationary and nonlinear systems [15]. EMD is intuitive and adaptive, with basic functions derived fully from the analyzed signal without the need of the base function of the analyzed signal. EMD decomposes a seismic signal into a sum of intrinsic oscillatory components, called "intrinsic mode functions" (IMFs). Furthermore, high-resolution time–frequency analysis is possible by combining EMD with the instantaneous frequency. The resulting time–frequency resolution promises to be significantly higher than that obtained using traditional time–frequency analysis tools, such as short-time Fourier and wavelet transforms. The empirical mode decomposition has progressed from EMD to ensemble empirical mode decomposition (EEMD) [16], bivariate EMD (BEMD) [17], multivariate EMD (MEMD) [18], and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [19]. EMD related methods offer many promising features for analyzing and processing geophysical data, such as seismic attributes [20], removing cable strum noise [21], attenuate random and coherent seismic noise [22], and time-frequency (TF) analysis [20].

Different time–frequency components potentially highlight different geologic and stratigraphic information. When extract key seismic time–frequency features are based on available well logs, they need rigorous statistical-learning methods to bring out the relationship between them because the relation between the seismic data and the well logs is mostly field specific and cannot be generalized. As for integrating the high-information-carrying well logs and the low-information-carrying seismic attributes, it is firstly necessary to convert the well logs into the time domain by available velocity profiles from the well-seismic tie [13]. Cross-correlation can be used to measure the similarity between different time-signal of specific frequency and the synthetic trace by well logs, and time-domain cross correlation or the cross spectral technique can be employed to reduce the measurement error of the time differences. The use of waveform cross correlation has a long history of identifying similar waveforms [23–25]. When waveforms exhibit sufficient similarity, correlation can achieve subsample precision.

To build the final machine learning model, automatic waveform clustering, such as fuzzy C-means cluster [26], can be used to take the key seismic time–frequency signal as input and extract the natural clustering of the underlying seismic facies. Waveform cluster is a pattern recognition technique, which is often used to analyze seismic waveform [27]. Within a certain interval, the seismic facies variation gives rise to changes in the amplitude, phase, and frequency of the seismic reflection. Within a certain time-window, seismic waveforms belonging to one cluster share the same characteristics that are distinct from those in other clusters [28]. Waveform cluster can help to visualize the variation in rock properties and the possible relationships between these depositional facies and their seismic response [29].

This paper proposes a model combining geological constraints and machine learning to describe reservoirs. Firstly, the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is used to extract different constant-frequency data volumes. Then, the most sensitive constant-frequency data to the reservoirs is determined by the seismic synthesis records from well logs. Finally, the most sensitive constant-frequency data is classified by fuzzy C-means cluster to characterize the distribution of paleo-channels system reservoirs.

2. Geological Background

One-fourth of the karst area on Earth is distributed in China and one-half of the karst bare area is in northern China. Tahe Oilfield is the largest Paleozoic marine oilfield in China [2–4], covering an area of 2400 km^2. Tahe Oilfield is located on the southwestern slope of the south-central Akekule Arch (shown in Figure 1b) in the North Uplift (Shaya Uplift) of the Tarim Basin (shown in Figure 1a). Hydrocarbon reservoirs have been identified in Triassic, Carboniferous, Devonian, and Ordovician strata. The Ordovician paleokarst reservoirs account for approximately 73% of total production [30]. From the bottom to the top, the Ordovician strata are divided into Penglaiba Formation, Yingshan Formation, and Yijianfang group, Querbake Formation, Lianglitage Formation, and Santamu Formation (shown in Figure 1c).

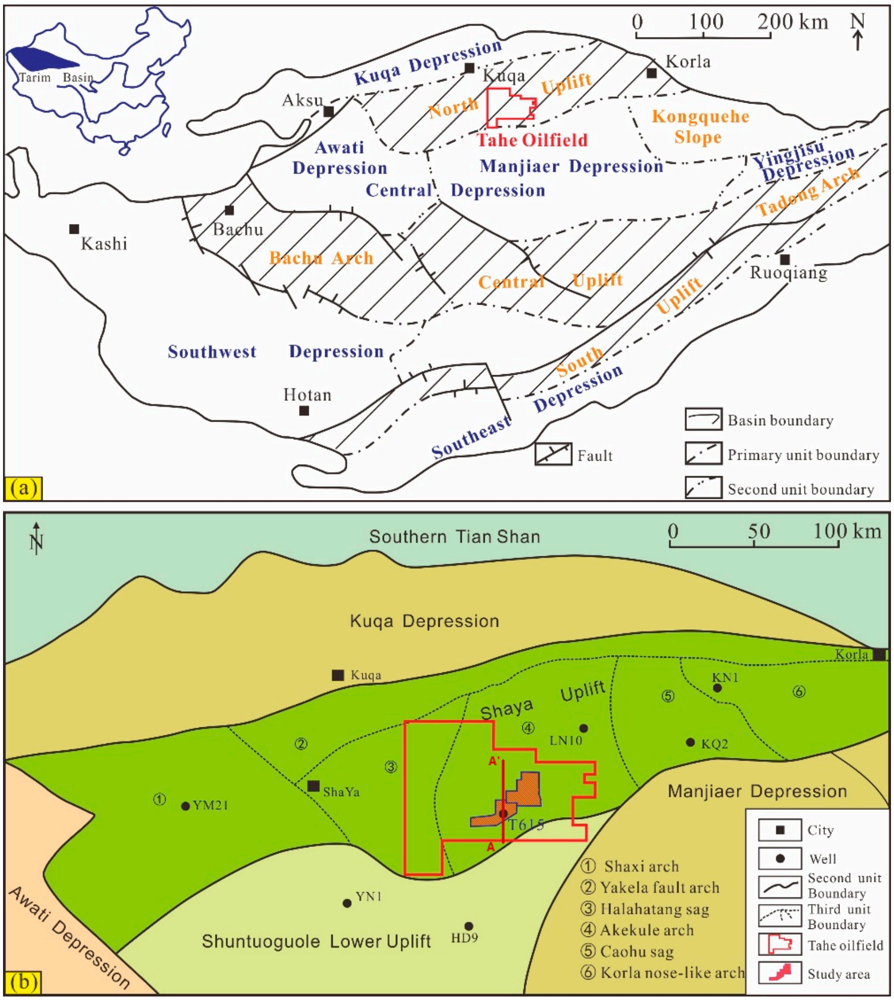


Figure 1. *Cont.*

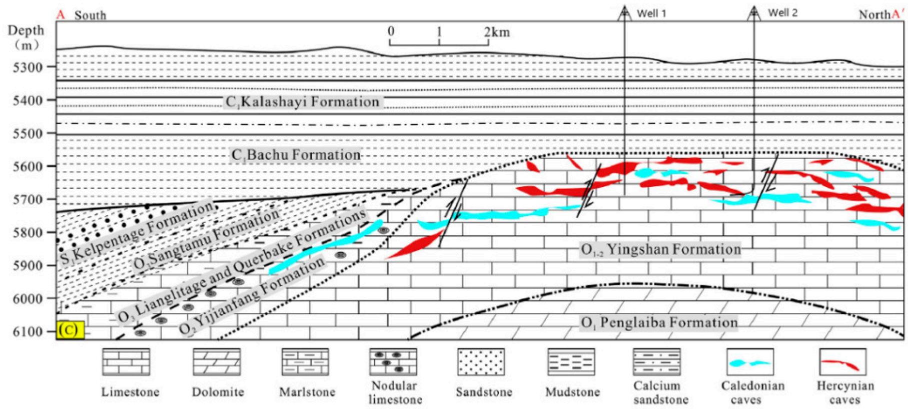


Figure 1. Location of the study area and overview map of the Tarim Basin. (**a**) Tectonic components of the Tarim Basin, including three uplifts and four depressions with east-west orientations (reference [31]). Tahe oilfield is located in the Northern Uplift of the Tarim Basin. (**b**) The study area is in the center of the Tahe oilfield, which is located in the southwestern part of the Akekule Arch in the Northern Uplift of the Tarim Basin (reference [31]). (**c**) Cross Section A-A′, showing the paleokarst system of the Tahe Oilfield. The main reservoirs are located in the northern part of the Tahe Oilfield in the Yingshan Formation (O_{1-2}ys), which is at depths exceeding 5500 m (reference [31]). See Figure 1b for the cross-section location.

Large deep faults and associated folds in Akekule Arch, forming a large karst slope, have experienced a series of erosion events from Caledonia to the Hercynian period [32]. Due to its long duration, widespread distribution, and great intensity, the early Hercynian karstification event was the most significant karstification phase in the formation of the paleokarst systems in the Tahe oilfield [3]. Groundwater dynamics in karst slopes have vertical infiltration and horizontal movement and a fracture-cavity system is thus developed. The remaining Ordovician formations are uneven and the contact relationship is complex. The strata in the northern Tahe Oilfield were denuded and the Carboniferous Bachu Formation unconformably overlies the Yingshan Formation [33]. These events identified the morphological diversity of the karst cave system, resulting in large-scale karst geomorphology and underground ancient caves, all of which represent the type of karst landforms observed in Guilin today [2].

It is this geological environment that provides a broad reservoir space for the accumulation of oil and gas, guaranteeing the high and stable production of Tahe Oilfield. Hydrocarbons were first generated in the Manjiaer depression before migrating through a series of deep-seated faults and accumulating within large karstic fault systems, which form the current deep fracture-cavity reservoirs in the north region of the Tarim Basin [33]. The ancient karst reservoir developed near the unconformity in the Middle and Lower Ordovician strata, with a burial depth exceeding 5500 m [2,3]. The rock matrix has no reservoir permeability and the accumulation of oil and gas mainly depends on the development degree of the fracture-cavity system in the strata. Previous studies have shown that paleokarst reservoirs are usually located less than 150 m below the unconformity [34]. Being the core of the entire karst system, the karst paleo-channels of multiple scales and different orientations are intertwined with various sizes of fracture-cavity reservoirs [33]. The fracture-cavity reservoirs are distributed in a random and scattered manner with great complexity [35] and have an extremely heterogeneous internal structure and non-uniform lithological properties [36].

3. Materials and Methods

3.1. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

Empirical Mode Decomposition (EMD) [15] is an adaptive method introduced to analyze non-linear and non-stationary signals. It consists of a local and fully data-driven separation of a signal in fast and slow oscillations. However, EMD experiences some problems, such as the presence of oscillations of very disparate amplitude in a mode or the presence of very similar oscillations in different modes, named as "mode mixing". To overcome these problems, a new method was proposed: Ensemble Empirical Mode Decomposition (EEMD) [16]. It performs the EMD over an ensemble of the signal plus Gaussian white noise. The addition of white Gaussian noise solves the mode mixing problem by populating the whole time-frequency space to take advantage of the dyadic filter bank behavior of the EMD [3]; however, it creates some new ones. Indeed, the reconstructed signal includes residual noise and different realizations of signal plus noise may produce different number of modes. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), a variation of the EEMD proposed by Torres et al. [19], can provide an exact reconstruction of the original signal and a better spectral separation of the modes, with a lower computational cost.

EMD decomposes a signal (t) into a small number of Intrinsic Mode Functions (IMFs). To be considered as an IMF, a signal must satisfy two conditions: (i) the number of extrema and the number of zero crossing must be equal or differ at most by one; and (ii) the mean value of the upper and lower envelope is zero everywhere. EEMD defines the "true" IMF components (here notated as *IMF* in what follows) as the mean of the corresponding IMFs obtained via EMD over an ensemble of trials, generated by adding different realizations of white noise of finite variance to the original signal $x[n]$. The EEMD algorithm can be described as [16]:

1. generate $x^i[n] = x[n] + w^i[n]$, $(i = 1, \ldots, I)$, where $w^i[n]$ are different realizations of white Gaussian noise,
2. each $x^i[n]$, $(i = 1, \ldots, I)$ is fully decomposed by EMD, getting their modes $IMF_k^i[n]$, where k = 1, . . . , K indicates the modes,
3. assign $\overline{IMF}_k$ as the k-th mode of $x[n]$, obtained as the average of the corresponding $\overline{IMF}_k[n] = \frac{1}{I}\sum_{i=1}^{I} IMF_k^i[n]$.

Observe that in EEMD, each $x^i[n]$ is decomposed independently from the other realizations and so for each one, a residue $r_k^i[n] = r_{k-1}^i[n] - IMF_k^i[n]$ is obtained.

For EEMDAN, with the decomposition modes being noted as $\widetilde{IMF}_k$, the unique first residue is:

$$r_1[n] = x[n] - \widetilde{IMF}_1[n] \tag{1}$$

where $\widetilde{IMF}_1[n]$ is obtained in the same way as in EEMD. Then, compute the first EMD mode over an ensemble of $r_1[n]$ plus different realizations of a given noise, obtaining $\widetilde{IMF}_2$ by averaging. The next residue is defined as:

$$r_2[n] = r_1[n] - \widetilde{IMF}_2[n] \tag{2}$$

This procedure continues with the rest of the modes until the stopping criterion is reached. Let us define the operator $E_j(\bullet)$ which, given a signal, produces the j-th mode obtained by EMD. Let w^i be white noise with (0, 1). If $x[n]$ is the targeted data, the CEEMDAN method can be described by the following algorithm [19]:

1. Decompose by EMD I realizations $x[n] + \varepsilon_0 w^i[n]$ to obtain their first modes and compute

$$\widetilde{IMF}_1[n] = \frac{1}{I}\sum\nolimits_{i=1}^{I} IMF_1^i[n] = \overline{IMF}_1[n] \tag{3}$$

2. At the first stage (k = 1), calculate the first residue as in

$$r_1[n] = x[n] - \widetilde{IMF}_1[n] \tag{4}$$

3. Decompose realizations $r_1[n] + \varepsilon_1 E_1\big(w^i[n]\big)$, $i = 1, \ldots, I$ until their first EMD mode and define the second mode:

$$\widetilde{IMF}_2[n] = \frac{1}{I}\sum\nolimits_{i=1}^{I} E_1\big(r_1[n] + \varepsilon_1 E_1\big(w^i[n]\big)\big) \tag{5}$$

4. For $k = 2, \ldots, K$ calculate the k-th residue:

$$r_k[n] = r_{(k-1)}[n] - \widetilde{IMF}_k[n] \tag{6}$$

5. Decompose realizations $r_k[n] + \varepsilon_k E_k\big(w^i[n]\big), i = 1, \ldots, I$ until their first EMD mode and define the $(k + 1)$-th mode as

$$\widetilde{IMF}_{(k+1)}[n] = \frac{1}{I}\sum\nolimits_{i=1}^{I} E_1\big(r_k[n] + \varepsilon_k E_k\big(w^i[n]\big)\big) \tag{7}$$

6. Go to step 4 for next k.

Steps 4 to 6 are performed until the obtained residue is no longer feasible to be decomposed (the residue does not have at least two extrema). The final residue Res[n] satisfies:

$$Res[n] = x[n] - \sum\nolimits_{k=1}^{K} \widetilde{IMF}_k[n] \tag{8}$$

with K being the total number of modes. Therefore, the given signal $x[n]$ can be expressed as:

$$x[n] = \sum\nolimits_{k=1}^{K} \widetilde{IMF}_k[n] + Res[n] \tag{9}$$

Equation (9) makes the proposed decomposition complete and provides an exact reconstruction of the original data.

Observe that the ε_i coefficients allow one to select the SNR (Signal Noise Ratio) at each stage. Concerning the amplitude of the added noise, Wu and Huang suggested to use small amplitude values for data dominated by high-frequency signals, and vice versa [16]. Following this, in this work, we used a few hundred realizations and fixed the same SNR for all the stages. This value might depend on the application.

The local symmetry property of the IMFs ensures that instantaneous frequencies are always positive, thereby rendering EMD or its variants interesting for time-frequency analysis [15]. Seismic instantaneous attributes [37] are derived from the seismic trace $x(t)$ and its Hilbert transform $y(t)$ by computing its analytic signal, given by:

$$z(t) = x(t) + iy(t) = R(t)\ exp[i\theta(t)] \tag{10}$$

where $R(t)$ and $\theta(t)$ denote the instantaneous amplitude and instantaneous phase, respectively. Instantaneous amplitude $R(t)$ is the trace envelope, also called reflection strength, defined as

$$R(t) = \sqrt{x^2(t) + y^2(t)} \tag{11}$$

Instantaneous frequency $f(t)$ is defined as the first derivative of instantaneous phase $\theta(t)$. Thus,

$$f(t) = \frac{1}{2\pi}\frac{d\theta(t)}{dt} \tag{12}$$

To prevent ambiguities due to phase unwrapping in Equation (12), the instantaneous frequency can be calculated instead from

$$f(t) = \frac{1}{2\pi}\frac{x(t)y'(t) - x'(t)y(t)}{x^2(t) + y^2(t)} \tag{13}$$

where prime denotes derivative with respect to time.

Equations (11) and (13) can be used to compute instantaneous amplitudes and frequencies for each IMF. Contrary to classical application of instantaneous attributes to the original signal, this procedure produces a multitude of instantaneous frequencies at each time sample, namely one for each IMF, allowing for a more in-depth signal analysis. The result is a time–frequency distribution that is uniformly sampled in time, but not in frequency, contrary to, for instance, the short-time Fourier transform. There are as many instantaneous frequencies as IMFs, but most applications produce up to a dozen IMFs, creating very sparse time–frequency representations.

3.2. Cross-Correlation

The similarity measure used in the time domain is the cross-correlation coefficient (CC) and the time-domain cross correlation is defined as [23]:

$$c(\tau) = n\int_{-\infty}^{+\infty} {u_1}^{*}(t)u_2(t+\tau)dt \tag{14}$$

$$n = 1/\left[\left(\int_{-\infty}^{+\infty} u_1(t)^2 dt\right)^{1/2}\left(\int_{-\infty}^{+\infty} u_2(t)^2 dt\right)^{1/2}\right] \tag{15}$$

where $u_1(t)$ and $u_2(t)$ are the signals to be measured for similarity, ${u_1}^{*}(t)$ is complex conjugate of $u_1(t)$. The τ is the delay between the two signals. $C_c = max\{c(\tau)\}$ is taken here as the final similarity value.

3.3. Waveform Cluster

The Fuzzy C-means Cluster (FCM) [26] can be used for waveform clustering, discovering the natural seismic responses of reservoir patterns without any priori information. In most of the classic cluster algorithms, each object is assigned to only one cluster, such as K-Means and KNN [26]. FCM relaxes the restriction and the object can belong to all the clusters with a certain degree of membership, which is particularly useful when the boundaries among the clusters are not well separated and ambiguous. Moreover, the memberships may help us discover more sophisticated relations between a given object and the disclosed clusters.

FCM attempts to find c fuzzy clusters for a set of data points $x_j \in \mathcal{R}^d,\ j = 1, \ldots,\ N$ while minimizing the cost function $\mathrm{J(U,M)}$ [26]:

$$\mathrm{J(U,M)} = \sum\nolimits_{i=1}^{c}\sum\nolimits_{j=1}^{N}\left(\mu_{i,j}\right)^{m} D_{ij}, \tag{16}$$

where $\mathrm{U} = \left[\mu_{i,j}\right]_{c\times N}$ is the fuzzy partition matrix and $\mu_{i,j} \in [0,1]$ is the membership coefficient of the *j*th object in the *i*th cluster; $\mathrm{M} = [m_1, m_2, \ldots,\ m_c]$ is the cluster prototype (mean or center) matrix; $m \in [1,\infty)$ is the fuzzification parameter and usually is set to 2; $D_{ij} = D\left(x_j,\ m_i\right)$ is the distance measure between x_j and m_i, for instance, the Euclidean L_2 norm distance function. The basic flow of the fuzzy C-means cluster for the waveform is as follows:

1. Select the time window to extract the waveforms, $x_j \in \mathcal{R}^d,\ j = 1, \ldots,\ N$, where x_j is the *j*th waveform in the time window, d is the time sampling number, and N is the number of waveforms;
2. Select appropriate values of m, c, and a small positive number ε. Initialize the prototype matrix M randomly and set step variable $t = 0$.
3. Calculate (at $t = 0$) or update (at $t > 0$) the membership matrix U by:

$$\mu_{ij}{}^{(t+1)} = 1/(\sum_{l=1}^{c}\left(D_{lj}/D_{ij}\right)^{1/(1-m)}),\ for\ i = 1,\ldots,c\ and\ j = 1,\ldots,\ N. \tag{17}$$

4. Update the prototype matrix M by:

$$m_i{}^{(t+1)} = \left(\sum_{j=1}^{N} \left(\mu_{ij}{}^{(t+1)}\right)^m x_j) / (\sum_{j=1}^{N} \left(\mu_{ij}{}^{(t+1)}\right)^m \right), \ for\ i = 1, \ldots, c. \tag{18}$$

5. Repeat steps 2–3 until $||M^{(t+1)} - M^{(t)}|| < \varepsilon$. The jth waveform is assigned to the lth cluster if $\mu_{l,j}$ is the maximum of all $\mu_{i,j}$, $i = 1, \ldots, c$.

4. Results and Discussions

4.1. Dataset

The study area is located in Block 7 of the Tahe Oilfield, centered on Well T615 (shown in Figure 1). The 3-D seismic data used is a post-stack dataset with a grid size of 15 × 15 m covering 15.5 km^2 (4.5 km long and 4.5 km wide), as shown in Figure 2. Well logging data from eight vertical wells can provide sufficient information for characterizing reservoirs in the target strata. Our proposed method supposes to analyze the spatial distribution of the fracture-cavity paleo-channels reservoir.

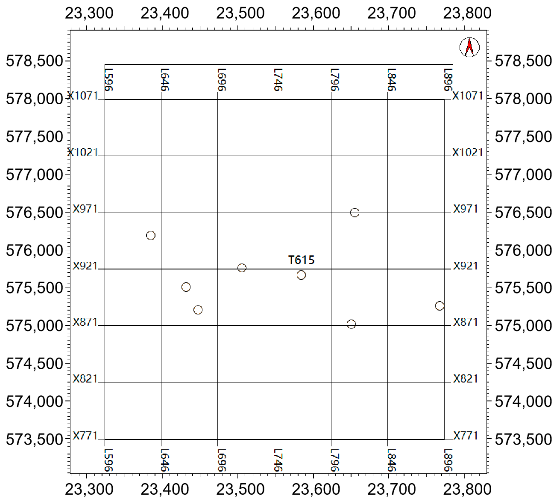


Figure 2. Distribution of the study area along two dimensions, i.e., inlines and xlines, where eight wells are marked by the dark circles.

One 2-D slice of seismic data from Figure 2 is show as an example in Figure 3, a full-band seismic profile, within which the target reservoir is between about 3.5 s and 3.7 s in time with strong heterogeneity. The geomorphological features around the target reservoir show that the reflection of the paleo-channel has a "pull-down" feature, the channel exhibits lateral strong heterogeneity, and the reflection characteristics of the "beads" develop longitudinally. Although controlled by the diving datum and hydrodynamic changes, the paleo-channel morphological structure and scale formed along the layers and fracture zones are very different, but the common feature is the layered reflection structure of the cave (shown between the T74 Top and Bottom). All of the seismic profiles that have been cut vertically along the river direction have the above characteristics. The statistical results show that the average depth of the paleo-channel is 63 m and the width is between 45 and 500 m. This indicates that during the development of the paleo-channel system, the hydrodynamic conditions are strong, resulting in deep undercuts and a large width.

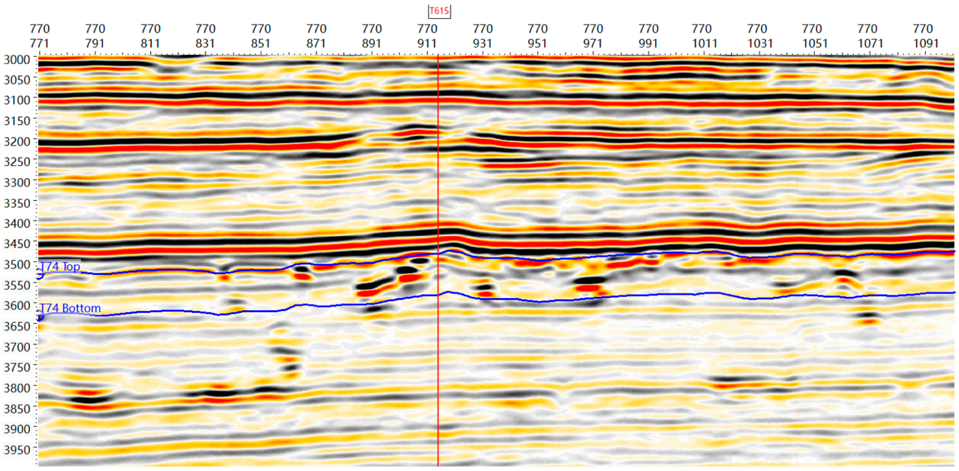


Figure 3. 2-D seismic data cross T615. The area between T74 Top and Bottom is the target reservoir.

4.2. *Time–Frequency Features*

In this section, we compare the EMD and CEEMDAN methods using seismic signal (shown in Figure 4) to demonstrate the advantages of CEEMD.

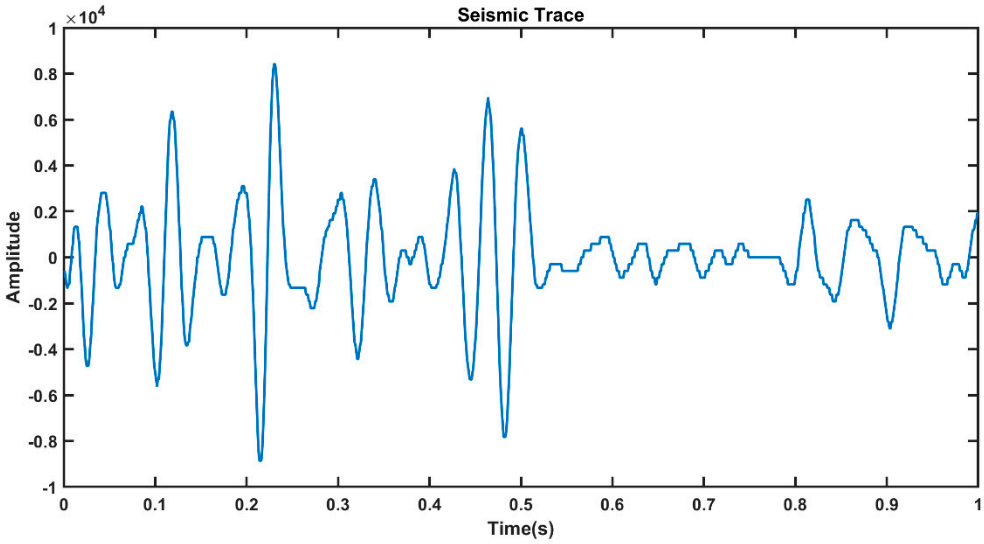
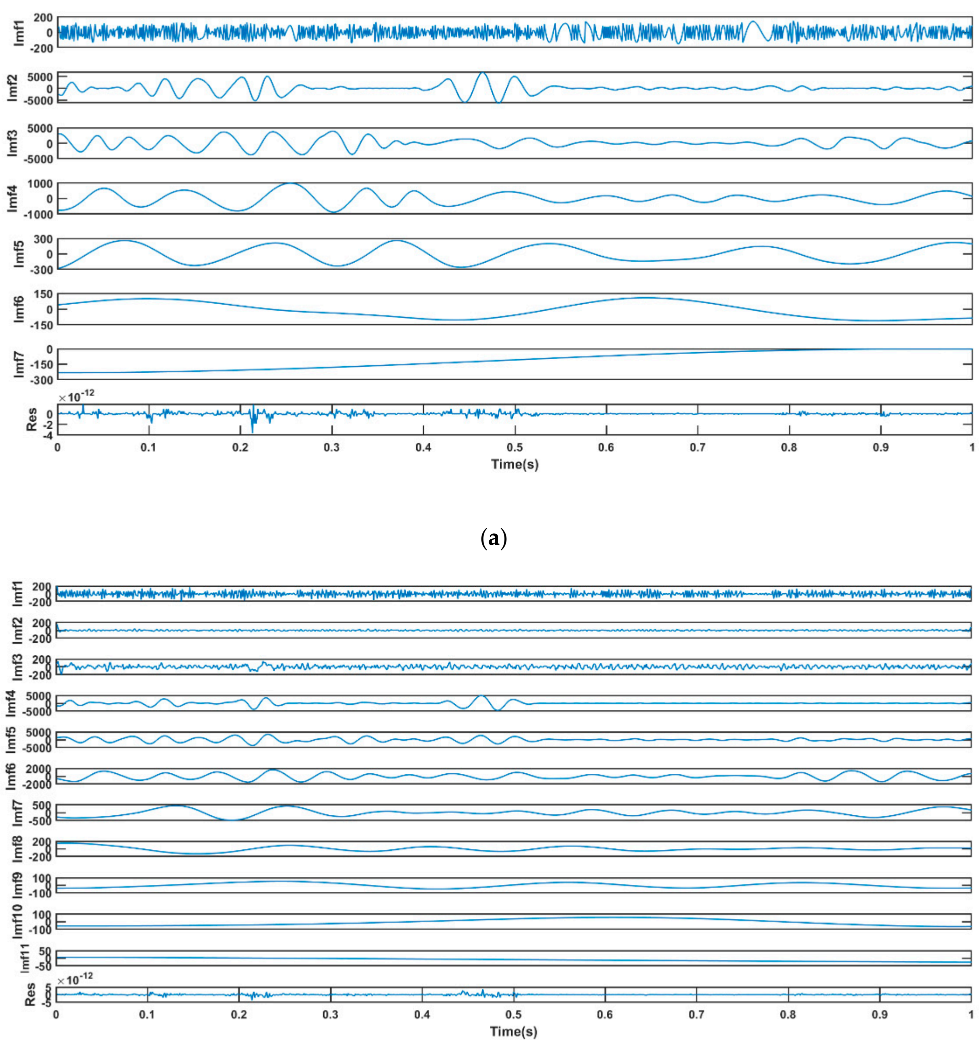


Figure 4. The seismic trace signal.

EMD decomposes the synthetic data into seven IMFs (Figure 5a). The IMFs in Figure 5a show mode-mixing deficiencies. IMF1 does not solely extract the high-frequency components, as it is polluted with low-frequency components. Likewise, IMF2 and IMF3 mix low- and high-frequency components from a variety of signal components. This makes it difficult to recognize the individual contributions of each component to various IMFs, thereby complicating signal analysis.

The CEEMDAN result using 20% Gaussian white noise and 500 realizations is shown in Figure 5b. After the CEEMDAN decomposition, each IMF is locally symmetric. The resulting IMF1, IMF2, and IMF3 are similar to the IMF1 obtained by EMD, but without low-frequency components. The frequency component of each IMF4-IMF10 is also stable and single. CEEMDAN is less affected by mode mixing of all EMD variants. For both EMD and CEEMDAN decomposition, the IMFs can perfectly reproduce the original signal with a reconstruction error of negligible percentage of the total energy.

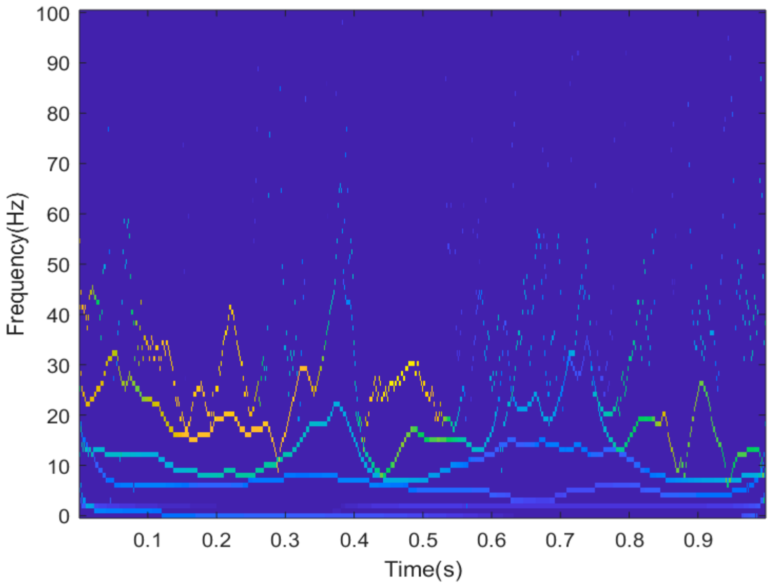


(**a**)

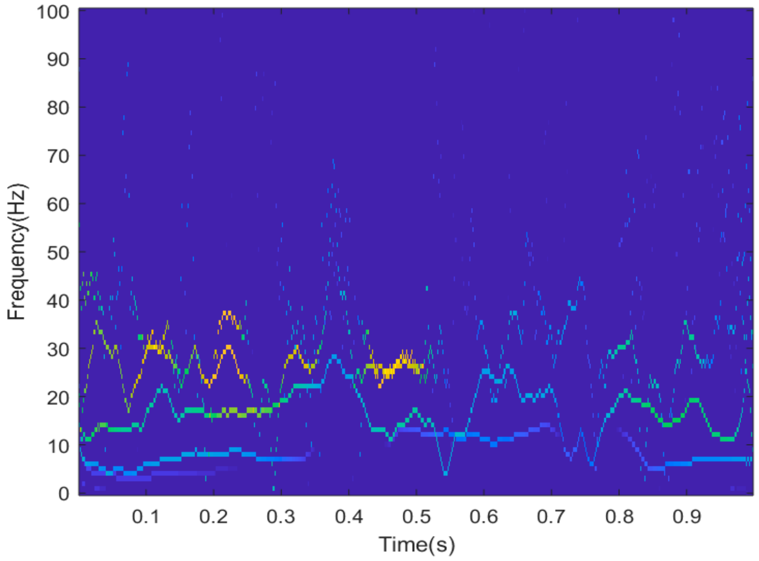
(**b**)

Figure 5. Intrinsic mode functions (IMFs) and residual of empirical mode decomposition (EMD) (**a**) and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) (**b**).

We compute the instantaneous frequency of each IMF using Equation (13) and associated instantaneous amplitude with Equation (11). The resulting time–frequency image is smoothed by a convolution with a 2D Gaussian-weighted filter of prespecified 30 × 30. We compare the resulting instantaneous spectrum with EMD (Figure 6a) and CEEMDAN (Figure 6b) for the same seismic trace signal shown in Figure 4. CEEMDAN analysis, which shows a better compromise between time and frequency resolution than the EMD as the frequency components, is more dispersed on the time and frequency axis. The time–frequency image of better resolution is helpful for more accurately locating spectral anomalies and thus facilitating further interpretation.



(**a**)

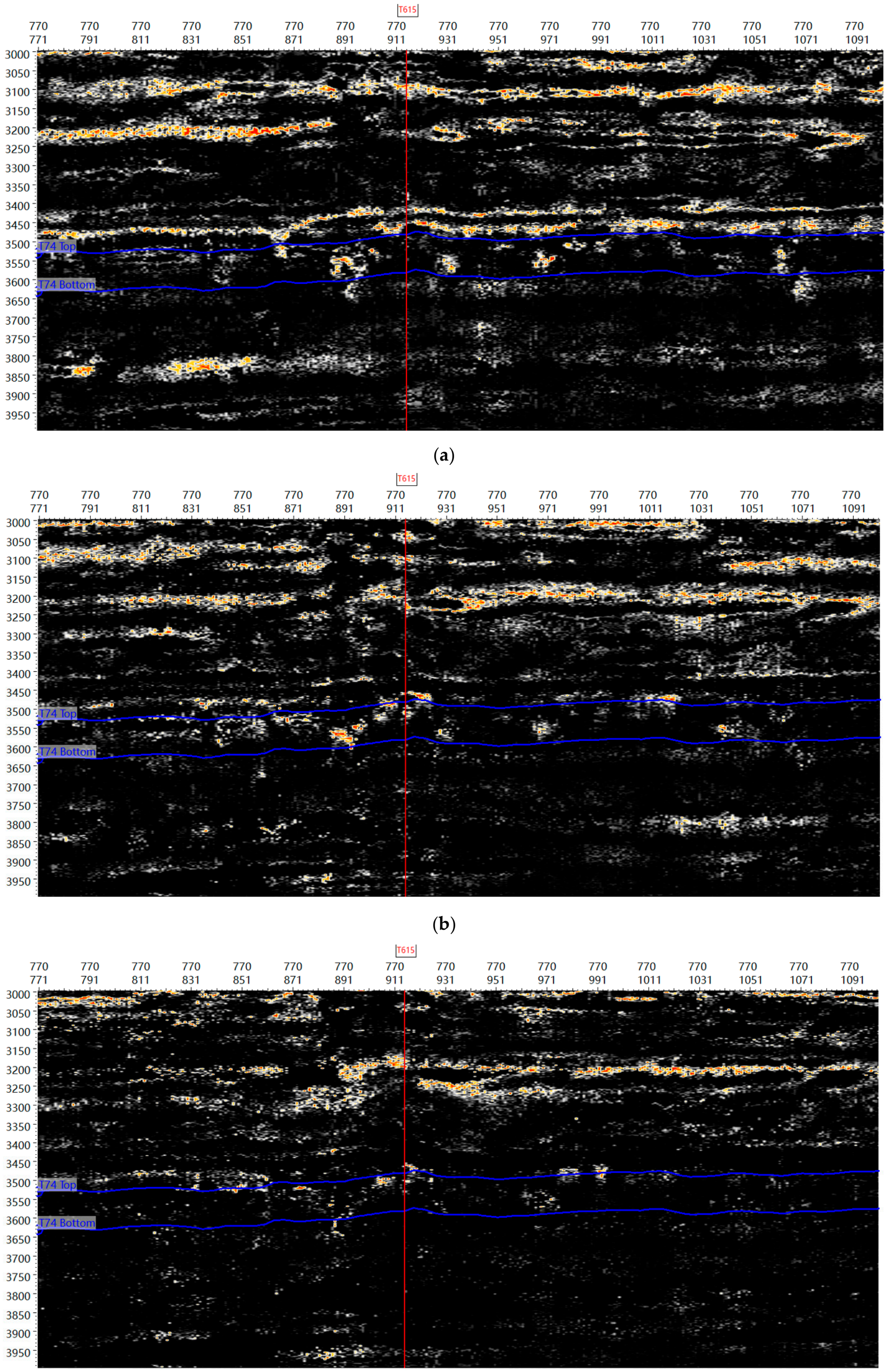


(**b**)

Figure 6. Time-frequency maps by EMD (**a**) and CEEMDAN (**b**) of the seismic signal (Figure 4).

Next, we extract the 30, 35, and 40 Hz frequency slices after CEEMDAN and calculate the instantaneous frequencies (Figure 7) to illustrate that the instantaneous spectrum of different frequencies show different reservoir information. The instantaneous spectrum shows much sparser outputs and resolves the spectral characteristics of the various reflections more clearly than the original seismic data (shown in Figure 3). Single-frequency sections have two main advantages over the full-band data: (1) The signal-to-noise ratio is significantly improved; (2) The recognition of small-scale caves and fractures is obviously improved, the energy is more concentrated, and the cave edge is much clearer.

The section with 30 Hz (Figure 7a) mainly reflects large sets of formation information, with relatively low resolution. The profile with 40 Hz (Figure 7c) seems to have higher resolution, but too many details of the caves are lost, which is due to the attenuation of high-frequency energy and the reduction of the signal-to-noise ratio. Profiles between the 30 Hz–40 Hz, such as 35 Hz (Figure 7b), reveals that the caves (“bead” shape) in the target layer are distinguishable. As a result, we chose 30 Hz–40 Hz to be the sensitive frequency range for the following analysis.



(**a**)

(**b**)

(**c**)

Figure 7. Constant-frequency slices, 30-Hz (**a**), 35-Hz (**b**), and 40-Hz (**c**) of the time–frequency map of the 2D seismic data cross Well T615 (Figure 3). The instantaneous spectrum of different frequencies shows different reservoir information.

4.3. Sensitivity of Time–Frequency Features

In the signal preprocessing of applying machine learning on seismic data, the input features are generally chosen by experience, lacking geological constraints. In terms of priority illumination, seismic events of the reservoir are preferentially illuminated by certain frequency components [38]. The reliability of the suitable single–frequency data has a complex relationship with the surrounding environment of the target reservoir, the strength of the reflection events, and the size of the target reservoir. As for geomorphology, the fracture-caves vary in size, distribute in a random, dispersed, and discontinuous manner, and intertwine with paleo-channels of different scales and different strikes [36]. Most importantly, the Ordovician carbonate strata are deeply buried, the seismic response of the unconformity surface is weak and discontinuous, and the main frequency and SNR are relatively low [2–4].

The distributions of lithological properties in the subsurface are more accurate in borehole locations due to the direct high-resolution measurements along with the wells. On the other hand, the seismic attributes are recorded spatially with a low vertical resolution in the time domain. Therefore, it is imperative to use densely available 3-D seismic attributes along with the existing well logs as a guide to generate the target lithological properties over the study area [14].

What we tried to investigate here is which of the selected single-frequency data profiles in the previous part is consistent with the real strata. We selected the logging data (AC, DEN) of eight wells to synthesize seismic traces by GEOSCOPE software, Beijing RockStar Tech, Inc. (Beijing, China), which were compared to the single-frequency traces around the wells. Ricker wavelet was selected as the wavelet with 30 Hz main frequency, which was the main frequency of the seismic data around the well. For example, synthesize seismic traces around T615 are shown in Figure 8.

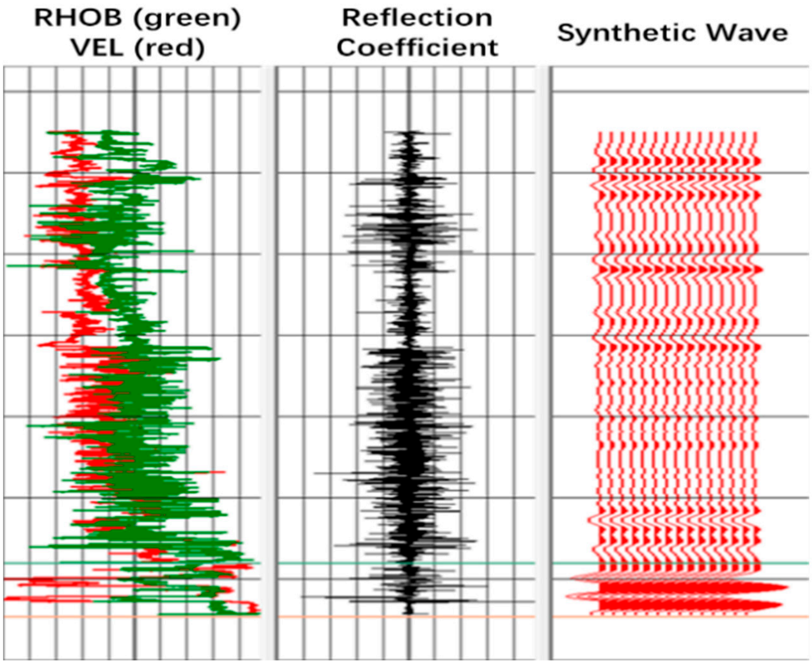


Figure 8. The synthetic seismic trace around well T615 as an example.

By Equation (14), we calculated the maximum of the cross-correlation value between the synthetic seismic signals and the related 11 single-frequency signals (30 Hz–40 Hz) of the seismic trace surrounding each well. Through statistics, we found that the frequency with the highest occurrence (six wells of all eight wells) was 35 Hz.

4.4. Waveform Cluster

The clusters of seismic facies provide interpreters with meaningful insights into the structure underlying the seismic data. The natural cluster structure can be related to stratigraphic features or reservoir properties by extracting information about changes in strata, faults, fluid, fractures, lithology, and lithofacies. The connectivity and geological similarity of the ancient river channel can help to understand the trapping of the reservoir.

Before applying the waveform cluster method mentioned in Section 3.3, the prepicked horizons were used as the geological constraint for selecting the time slice window. We selected the seismic data slice between the T74 top and T74 bottom shown in Figure 7, which is the location of a fracture-cavity paleo-channel reservoir. When selecting the suitable number of typical waveforms, prior geological understanding of the target layer and existing well-logging data should be combined to make the best decision. In this project, we chose three clusters representing paleo-channel, fractured caves, and the rock basis, respectively.

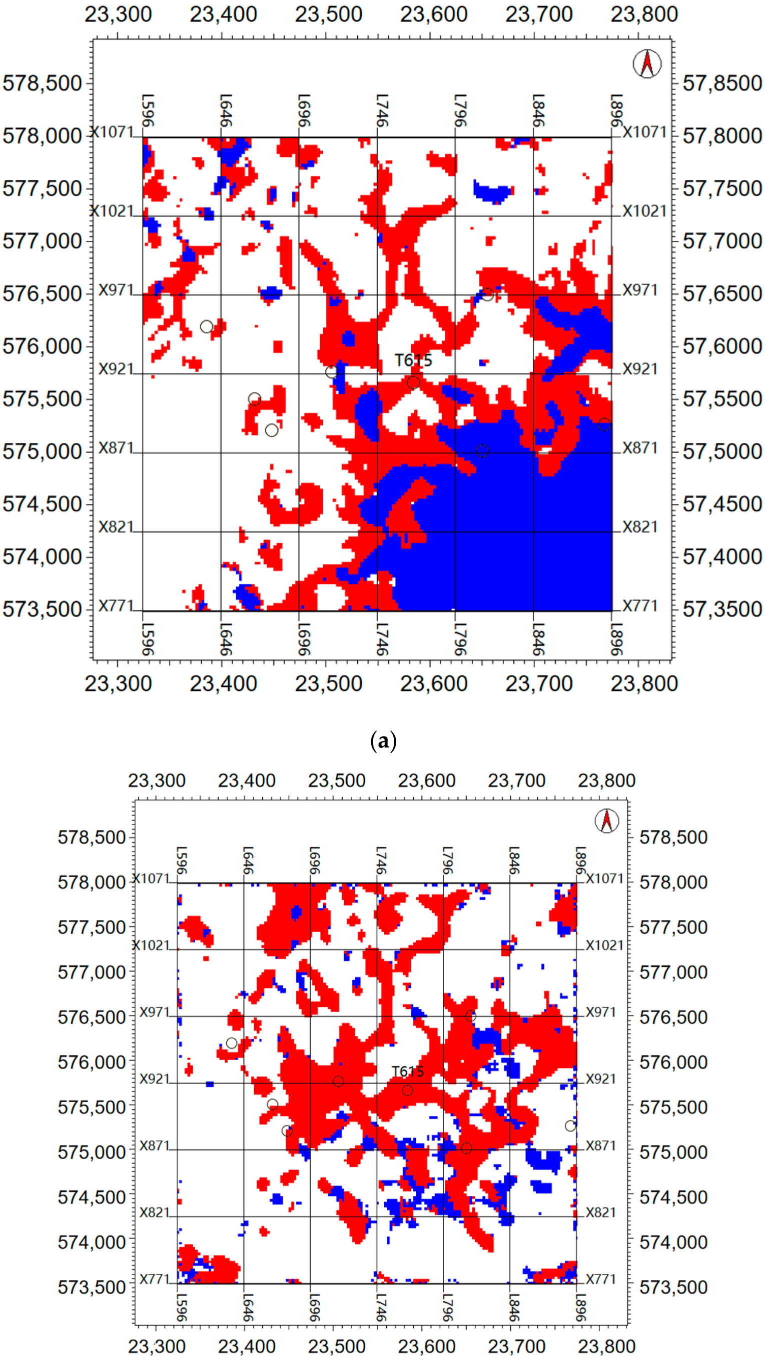
The waveform clustering results (Figure 9) show that the karst paleo-channels of multiple scales and different directions are intertwined with fractured caves of various sizes in the middle-upper depth of the Ordovician. The distribution of the surface paleokarst drainage system at the top of the Ordovician is similar to that of modern karst-drainage systems [39]. The red part represents the main trunk and branches of the ancient karst paleo-river; the blue represents the karst caves developed along the river; and the white represents the surrounding rock. In the middle north-south direction, a network-like paleo-channel (shown in red) with good connectivity is developed. The main paleo-river channel and branch paleo-river channel are clearly distributed. In the middle part of the study area, a channel, with a length of about 3 km and a width of 100 to 300 m, is developed. In the other area where channels are not developed, "bead" shaped caves (shown in blue) are scatter-likely distributed. It can be inferred that the target reservoir has good inner diameter flow conditions with a continuous distribution of soluble rock masses. The obvious trend of the branch channel in both Figure 9a and Figure 9b indicates that the "beads" (fractured caves) near the channels or the edge of the channels may be favorable Carbonate fracture-cavity reservoirs.

Figure 9a,b are results of waveform clustering on two kinds of data, original data, and 35-frequency. Clustering result for 35 Hz data is significantly better than the full band data:

(1) The connectivity between wells is much better, such as the connection of the six wells between X871 and X921, which is consistent with the previous study of the karst development [2,3,33,40] and the actual drilling process. The near-shore karst platform and karst gentle slopes of the ancient channel are formed by strong hydrodynamic erosion, which is easy to form a pipeline system of "crossing the mountain". The pipeline system is less damaged by the filling in the later stage and the formed oil and gas reservoirs are larger. An increase in connectivity of paleo-channel may indicate a corresponding increase in reservoir connectivity.

(2) The channels trend portrayed is clearer with fewer messy details. This provides a new perspective for understanding the crack caves in the Tahe Oilfield. From Figure 9b, the main channel accounts for the largest part of the total volume of the region, while the blue part, which represents the caves, accounts for almost all the area in the south-east corner in Figure 9a, which implies that the full-band data cannot capture the reservoir features very well.

As the frequency of the seismic signal increases, the resolution of the strata it depicts will increase accordingly. The main frequency of the seismic signal is 25 Hz and, as shown in Figure 7, 30 Hz (Figure 7a) mainly reflects a large amount of formation information and the resolution is lower, while 40 Hz (Figure 7c) seems to have a higher resolution. But too much cave detail is lost and this is due to the attenuation of high frequency energy and the reduction of the signal-to-noise ratio. Therefore, we decided to filter the time–frequency information from 30 Hz to 40 Hz to better match the real formation response. We calculated the maximum value of the cross-correlation between the synthesized seismic signal and the 11 single-frequency signals (30 Hz to 40 Hz) associated with the seismic traces around each well. Through statistics, we found that the highest frequency (6 of all 8 wells) is 35 Hz. In the end, the sensitive time–frequency data we chose was 35 Hz.



(**a**)

(**b**)

Figure 9. Waveform cluster results. (**a**) is obtained by the full-bands data; (**b**) is obtained by key time–frequency data.

Our selection of reservoir-sensitive single-frequency data may be related to thin-layer tuning effects. Carbonate fractures usually contain fluids (oil, gas, or water), causing a decrease in seismic reflection frequency. But sometimes the reflection frequency may increase when the reservoir consists of a series of small holes and cracks. This is mainly because the scattering or diffracted waves generated by a series of small cracks and caves are superimposed on each other, causing the main frequency to rise. In fact, the amplitude component of a certain frequency associated with a cracked cave is still weakened [39]. Fractured-cavity reservoirs, on the other hand, are typically relatively thin. It is known from the tuning characteristics of the thin layer [41] and the characteristics of the low-pass frequency filtering that the reflection of the fracture-caves not only have strong amplitude, but also the amplitude corresponding to a certain high-frequency component varies. Therefore, the time–frequency characteristics corresponding to a specific high frequency can be used to predict carbonate reservoirs. The sharpness of geological phenomena in different frequency data is significantly different. Thin-layer tuning effects are not elaborated in detail for the limited length of the article and related articles include [38,41,42].

In summary, 35 Hz-frequency seismic data show us with a better view of seismic data analysis a more reliable geological description to understand the history of reservoir formation and development and more accuracy to predict potential well locations in the area. The reservoir connectivity information can be mined from logging data and seismic data can provide a new understanding of reservoir prediction. By using high-resolution time-frequency analysis methods and using available logging constraints to select input features for machine learning, the effectiveness of machine learning models to mine geological reservoir information can be greatly improved.

5. Conclusions

The Ordovician carbonate strata in the Tahe oilfield are deeply buried. The irregular geometry, the associated extreme heterogeneity, combined with the weak quality of the seismic data make it difficult to accurately predict the distribution of reservoirs using the traditional full-band seismic data. We combine geological constraints and machine learning to extract the reservoir-sensitive single-frequency data and characterize the distribution of fracture-cavity paleo-channel reservoirs. By empirical mode decomposition, the time-frequency features of the seismic data are obtained and then a sensitive frequency is selected using geological prior constraints, which is finally input to fuzzy C-means cluster for characterizing reservoir. The method is applied to a work area in Tahe Oilfield. The results show that the method shows the potential of highlighting subtle geologic structures, which might be unnoticed by applying machine learning directly to the raw seismic data. This indicates that the method is promising and can effectively guide oil and gas exploration in Tahe Oilfield and other similar karstic groundwater and karst hydrogeological.

Author Contributions: Conceptualization, W.X., F.T., and X.S.; methodology, W.X., F.T., and X.S.; software, X.S. and W.X.; validation, F.T., X.S., Y.Z., and C.Y.; formal analysis, C.Y.; investigation, W.X., Y.Z., and H.R.; data curation, F.T.; writing—original draft preparation, W.X., X.S., Y.Z., and H.R.; writing—review and editing, W.X. and F.T.; visualization, W.X. and F.T.; supervision, C.Y.; project administration, F.T. and W.X.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Chinese National Natural Science Foundation (Grant Nos. 41504142, 41502149 and U1663204), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA14050101 and XDA14040201), the Chinese National Major Fundamental Research Developing Project (Grant No. 2017ZX05008-004), the Fundamental Research Funds for the Central Universities (Grant No. ZY1927) of China, and the China Postdoctoral Foundation Funded Project (Grant No. 2015M570148). Horizontal Cooperation Project (Grant No. H2020009).

Acknowledgments: We thank the two anonymous reviewers and academic editor. Their thorough and critical reviews and suggestions have significantly improved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Milad, B.; Ghosh, S.; Slatt, R.M. Comparison of Rock and Natural Fracture Attributes in Karsted and Non-Karsted Hunton Group Limestone: Ada and Fittstown Area, Oklahoma. *Okla. City Geol. Soc.* **2018**, *69*, 70–86.

2. Tian, F.; Jin, Q.; Lu, X.; Lei, Y.; Zhang, L.; Zheng, S.; Zhang, H.; Rong, Y.; Liu, N. Multi-layered ordovician paleokarst reservoir detection and spatial delineation: A case study in the Tahe Oilfield, Tarim Basin, Western China. *Mar. Pet. Geol.* **2016**, *69*, 53–73. [CrossRef]
3. Tian, F.; Lu, X.; Zheng, S.; Zhang, H.; Rong, Y.; Yang, D.; Liu, N. Structure and Filling Characteristics of Paleokarst Reservoirs in the Northern Tarim Basin, Revealed by Outcrop, Core and Borehole Images. *Open Geosci.* **2017**, *9*. [CrossRef]
4. Zhu, G.; Zhang, B.; Yang, H.; Su, J.; Liu, K.; Zhu, Y. Secondary alteration to ancient oil reservoirs by late gas filling in the Tazhong area, Tarim Basin. *J. Pet. Sci. Eng.* **2014**, *122*, 240–256. [CrossRef]
5. Zhang, B.M.; Liu, J.J. Classification and characteristics of karst reservoirs in China and related theories. *Pet. Explor. Dev.* **2009**, *36*, 12–29. [CrossRef]
6. Milad, B.; Slatt, R. Impact of lithofacies variations and structural changes on natural fracture distributions. *Interpretation* **2018**, *6*, T873–T887. [CrossRef]
7. Strecker, U.; Uden, R. Data mining of 3D poststack seismic attribute volumes using Kohonen self-organizing maps. *Lead. Edge* **2002**, *21*, 1032–1037. [CrossRef]
8. Sandham, W.; Leggett, M. *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*; Kluwer Academic Publishers: Dordrecht, The Netherlands; Boston, MA, USA, 2003; ISBN 978-1-4020-1729-2.
9. Zheng, Z.H.; Kavousi, P.; Di, H.B. Multi-Attributes and Neural Network-Based Fault Detection in 3D Seismic Interpretation. *Adv. Mater. Res.* **2013**, *838–841*, 1497–1502. [CrossRef]
10. Milad, B.; Ghosh, S.; Slatt, R.; Marfurt, K.; Fahes, M. Practical Aspects of Upscaling Geocellular Geological Models for Reservoir Fluid Flow Simulations: A Case Study in Integrating Geology, Geophysics, and Petroleum Engineering Multiscale Data from the Hunton Group. *Energies* **2020**, *13*, 1604. [CrossRef]
11. Di, H.; AlRegib, G. Seismic Multi-Attribute Classification for Salt Boundary Detection—A Comparison. In *79th EAGE Conference and Exhibition 2017*; European Association of Geoscientists & Engineers: Paris, France, 2017; Volume 2017, No. 1.
12. Wang, Z.; Di, H.; Shafiq, M.A.; Alaudah, Y.; AlRegib, G. Successful leveraging of image processing and machine learning in seismic structural interpretation: A review. *Lead. Edge* **2018**, *37*, 451–461. [CrossRef]
13. Chaki, S.; Routray, A.; Mohanty, W.K. A Novel Preprocessing Scheme to Improve the Prediction of Sand Fraction From Seismic Attributes Using Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1808–1820. [CrossRef]
14. Chaki, S.; Routray, A.; Mohanty, W.K. Well-Log and Seismic Data Integration for Reservoir Characterization: A Signal Processing and Machine-Learning Perspective. *IEEE Signal Process. Mag.* **2018**, *35*, 72–81. [CrossRef]
15. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.-C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A* **1998**, *454*, 903–995. [CrossRef]
16. Wu, Z.; Huang, N.E. ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD. *Adv. Adapt. Data Anal.* **2009**, *1*, 1–41. [CrossRef]
17. Rilling, G.; Flandrin, P.; Goncalves, P.; Lilly, J.M. Bivariate Empirical Mode Decomposition. *IEEE Signal Process. Lett.* **2007**, *14*, 936–939. [CrossRef]
18. Rehman, N.; Mandic, D.P. Multivariate empirical mode decomposition. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2010**, *466*, 1291–1302. [CrossRef]
19. Torres, M.E.; Colominas, M.A.; Schlotthauer, G.; Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; IEEE: Prague, Czech Republic, 2011; pp. 4144–4147.
20. Han, J.; Van der Baan, M. Empirical mode decomposition for seismic time-frequency analysis. *Geophysics* **2013**, *78*, O9–O19. [CrossRef]
21. Battista, B.M.; Knapp, C.; McGee, T.; Goebel, V. Application of the empirical mode decomposition and Hilbert-Huang transform to seismic reflection data. *Geophysics* **2007**, *72*, H29–H37. [CrossRef]
22. Bekara, M.; Van der Baan, M. Random and coherent noise attenuation by empirical mode decomposition. *Geophysics* **2009**, *74*, V89–V98. [CrossRef]
23. Schaff, D.P. Optimizing Correlation Techniques for Improved Earthquake Location. *Bull. Seismol. Soc. Am.* **2004**, *94*, 705–721. [CrossRef]

24. Schaff, D.P. Waveform Cross-Correlation-Based Differential Travel-Time Measurements at the Northern California Seismic Network. *Bull. Seismol. Soc. Am.* **2005**, *95*, 2446–2461. [CrossRef]
25. Moriya, H. Multiplet-Clustering Analysis Reveals Structural Details within the Seismic Cloud at the Soultz Geothermal Field, France. *Bull. Seismol. Soc. Am.* **2003**, *93*, 1606–1620. [CrossRef]
26. Xu, R.; WunschII, D. Survey of Clustering Algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [CrossRef]
27. Arshin, B.M.; Ghazali, A.R.; Amin, Y.K.; Barnes, A.E. Hybrid Waveform Classification Applied to Delineate Compartments in a Complex Reservoir in the Malay Basin. In Proceedings of the International Petroleum Technology Conference, Kuala Lumpur, Malaysia, 10–12 December 2014; International Petroleum Technology Conference: Kuala Lumpur, Malaysia, 2014.
28. John, A.K.; Lake, L.W.; Torres-Verdin, C.; Srinivasan, S. Seismic Facies Identification and Classification Using Simple Statistics. *SPE Reserv. Eval. Eng.* **2008**, *11*, 984–990. [CrossRef]
29. Zeng, H. Seismic geomorphology-based facies classification. *Lead. Edge* **2004**, *23*, 644–688. [CrossRef]
30. Tian, F.; Di, Q.; Jin, Q.; Cheng, F.; Zhang, W.; Lin, L.; Wang, Y.; Yang, D.; Niu, C.; Li, Y. Multiscale geological-geophysical characterization of the epigenic origin and deeply buried paleokarst system in Tahe Oilfield, Tarim Basin. *Mar. Pet. Geol.* **2019**, *102*, 16–32. [CrossRef]
31. Shan, X.; Tian, F.; Cheng, F.; Yang, C.; Xin, W. Spectral Decomposition and a Waveform Cluster to Characterize Strongly Heterogeneous Paleokarst Reservoirs in the Tarim Basin, China. *Water* **2019**, *11*, 256. [CrossRef]
32. Li, C.; Wang, X.; Li, B.; He, D. Paleozoic fault systems of the Tazhong Uplift, Tarim Basin, China. *Mar. Pet. Geol.* **2013**, *39*, 48–58. [CrossRef]
33. Lu, X.; Wang, Y.; Tian, F.; Li, X.; Yang, D.; Li, T.; Lv, Y.; He, X. New insights into the carbonate karstic fault system and reservoir formation in the Southern Tahe area of the Tarim Basin. *Mar. Pet. Geol.* **2017**, *86*, 587–605. [CrossRef]
34. Sun, S.Z.; Zhou, X.; Yang, H.; Wang, Y.; Wang, D.; Liu, Z. Fractured reservoir modeling by discrete fracture network and seismic modeling in the Tarim Basin, China. *Pet. Sci.* **2011**, *8*, 433–445. [CrossRef]
35. Wilson, J.P.; Grotzinger, J.P.; Fischer, W.W.; Hand, K.P.; Jensen, S.; Knoll, A.H.; Abelson, J.; Metz, J.M.; Mcloughlin, N.; Cohen, P.A.; et al. Deep-water incised valley deposits at the ediacaran-cambrian boundary in southern Namibia contain abundant treptichnus pedum. *Palaios* **2012**, *27*, 252–273. [CrossRef]
36. Liu, C.Y.; Lin, C.S.; Yi, W.; Wu, M.B. Burial Dissolution of Ordovician Granule Limestone in the Tahe Oilfield of the Tarim Basin, NW China, and Its Geological Significance. *Acta Geol. Sin. Engl. Ed.* **2010**, *82*, 520–529. [CrossRef]
37. Taner, M.T.; Koehler, F.; Sheriff, R.E. Complex seismic trace analysis. *Geophysics* **1979**, *44*, 1041–1063. [CrossRef]
38. Castagna, J.P.; Sun, S.; Siegfried, R.W. Instantaneous spectral analysis: Detection of low-frequency shadows associated with hydrocarbons. *Lead. Edge* **2003**, *22*, 120–127. [CrossRef]
39. Zeng, H.; Loucks, R.; Janson, X.; Wang, G.; Xia, Y.; Yuan, B.; Xu, L. Three-dimensional seismic geomorphology and analysis of the Ordovician paleokarst drainage system in the central Tabei Uplift, northern Tarim Basin, western China. *Bulletin* **2011**, *95*, 2061–2083. [CrossRef]
40. Tian, F.; Luo, X.; Zhang, W. Integrated geological-geophysical characterizations of deeply buried fractured-vuggy carbonate reservoirs in Ordovician strata, Tarim Basin. *Mar. Pet. Geol.* **2019**, *99*, 292–309. [CrossRef]
41. Partyka, G.; Gridley, J.; Lopez, J. Interpretational applications of spectral decomposition in reservoir characterization. *Lead. Edge* **1999**, *18*, 353–360. [CrossRef]
42. Marfurt, K.J.; Kirlin, R.L. Narrow-band spectral analysis and thin-bed tuning. *Geophysics* **2001**, *66*, 1274–1283. [CrossRef]