IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2020, accepted August 24, 2020, date of publication September 14, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022625

Two-Phase Multivariate Time Series Clustering to Classify Urban Rail Transit Stations

LIYING ZHANG[1,2], TAO PEI[2,3], BIN MENG[4], YUANFENG LIAN[1], AND ZHOU JIN[1]

[1]College of Information Science and Engineering, China University of Petroleum, Beijing 102249, China
[2]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China
[3]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[4]College of Applied Arts and Sciences, Beijing Union University, Beijing 100191, China

Corresponding author: Tao Pei (peit@lreis.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41525004, Grant 41421001, Grant 41877523, and Grant 41671165; in part by the State Key Laboratory of Resources and Environmental Information System; and in part by the Science Foundation of China University of Petroleum, Beijing, under Grant ZX20200100.

ABSTRACT Consider the problem of clustering objects with temporally changing multivariate variables, for instance, in the classification of cities with several changing socioeconomic indices in geographical research. If the changing multivariate can be recorded simultaneously as a multivariate time series, in which the length of each subseries is equal and the subseries can be correlated, the problem is transformed into a multivariate time series clustering problem. The available methods consider the correlations between distinct time series but overlook the shape of each time series, which causes multivariate time series with similar correlations and opposite shapes to be clustered into the same class. To overcome this problem, this paper proposes a two-phase multivariate time series clustering algorithm that considers both correlation and shape. In Phase I, the discrete wavelet transform is applied to capture the wavelet variances and the correlation coefficients between each pair of variables to realize the initial clustering of multivariate time series, where time series with a similar correlation but opposite shape may be assigned to the same cluster. In Phase II, multivariate time series are clustered based on shape via the symbolic aggregate approximation (SAX) method. In this phase, time series with similar correlations but opposite morphologies are differentiated. The method is evaluated using multivariate time series of incoming and outgoing passenger volumes from Beijing IC card data; these volume data were collected between March 4, 2013 and March 17, 2013. Based on the silhouette coefficient, our approach outperforms two popular multivariate time series clustering methods: a wavelet-based method and the SAX method.

INDEX TERMS Multivariate time series, cluster, maximum overlap discrete wavelet transform, symbolic aggregate approximation (SAX), urban rail transit stations.

I. INTRODUCTION

The aggregation of objects with many time-dependent variables has been considered in research on data mining, such as the classification of cities with multiple changing socioeconomic indices, the identification of crop type from various remotely sensed image series, and the categorization of the point of interest (POI) social function based on incoming and outgoing passenger flow series. This problem can be regarded as a cluster determination problem in a multivariate time series that is composed of single-variable time series of equal length. In the multivariate time series, the single-variable time series may be correlated. It may not be possible to adapt the available methods that are designed for univariate time series to multivariate time series, even if the multivariate time series can be treated as a single time series, because the correlations between single-variable time series may be disregarded by this adaptation. Hence, it is necessary to develop clustering methods for multivariate time series.

Univariate time series clustering has been well studied in the literature [1]–[6], whereas multivariate time series clustering, for which univariate methods are not suitable, has been less extensively addressed. The available approaches for multivariate time series clustering can be divided into three main categories: model-based methods, feature-based methods and shaped-based methods [7]. Model-based approaches

The associate editor coordinating the review of this manuscript and approving it for publication was Keli Xiao.

assume that each time series can be represented by a known model [8]. For example, Maharaj employed the vector autoregressive model to fit multivariate time series and utilized an algorithm that was based on the p-values of hypothesis tests to cluster time series [9]. However, it was necessary to estimate the model parameters of each time series; this approach failed to capture the correlations between the variables of the multivariate time series. In a feature-based method, a raw multivariate time series is represented by a lower dimensional feature vector. The extracted feature vectors are then clustered via a conventional clustering algorithm [10]–[13]. To extract feature vectors that represent a raw time series, Ye *et al.* performed a generalized principal component analysis (GPCA) [10], and Guo *et al.* [11] and Wu and Philip [12] applied an independent component analysis (ICA). However, they disregarded the correlations between the variables of the multivariate time series. To overcome this problem, D'Urso and Maharaj [13] proposed decomposing each variable of the multivariate time series into the wavelet series on various scales and calculating the wavelet variance at each scale. The wavelet correlations at each scale for the multivariate time series of every pair of variables are calculated, and the wavelet variance and correlation coefficient are concatenated into a single vector to represent the multivariate time series. This approach has the advantage of constructing the time series of wavelet features while considering the correlations. This approach can be applied to average nonstationary time series. However, time series with similar variances and correlation coefficients but opposite morphologies may be clustered into the same category. In shape-based methods, the time series are similar in terms of time and shape. For time series with high dimensionality, the dimension reduction technique is typically employed to reduce the noise and simplify the variables. Symbolic Aggregate approXimation (SAX) [14], which was proposed by Lin *et al.* as a transformation function in a time series similarity measure study, boasts the advantages of a high compression ratio, retention of data locality details and effective dimensionality reduction. However, SAX is unable to capture the correlations between the variables of multivariate time series.

Although a variety of approaches have been proposed [9]–[12], [14]–[16] for multivariate time series clustering, they either disregard the correlations between the variables of the multivariate time series or the cluster time series with opposite morphology in the same category. To overcome this problem, we propose two-phase clustering of multivariate time series based on wavelet transform and SAX (WSAX), which has the following characteristics: 1) in Phase I, the inherent correlations between variables of multivariate time series are considered by using a wavelet transform to represent the wavelet features of the original time series, and 2) in Phase II, shape-based clustering can effectively distinguish multivariate time series with opposite morphologies. Thus, multivariate time series with opposite shapes but similar variances and correlation coefficients can be effectively identified and clustered into different classes.

The remainder of the paper is organized as follows: Section 2 briefly describes the background of our proposed method. Section 3 details the two-phase multivariate time series clustering algorithm, namely, WSAX. The experimental results and analysis are presented in Section 4. The conclusions of this work are presented in Section 5.

II. BACKGROUND

In this section, we briefly describe the notations of multivariate time series, the maximal overlap discrete wavelet transform, SAX and the similarity measure, from which our proposed method is extended.

A. NOTATIONS AND PROBLEM

Let S represent a set of multivariate time series S_i as

$$S = \{S_i : i = 1, \ldots, I\} \tag{1}$$

$$S_i = \{s_{iqt} : q = 1, \ldots, Q; t = 1, \ldots, T\} = \begin{pmatrix} s_{i11} & \cdots & s_{iq1} & \cdots & s_{iQ1} \\ \vdots & & \vdots & & \vdots \\ s_{i1t} & \cdots & s_{iqt} & \cdots & s_{iQt} \\ \vdots & & \vdots & & \vdots \\ s_{i1T} & \cdots & s_{iqT} & \cdots & s_{iQT} \end{pmatrix} \tag{2}$$

where S_i represents the multivariate time series of the i^{th} object, with q ($q = 1, \ldots, Q$) as the variable, where t ($t = 1, \ldots, T$) denotes the observation time interval. Thus, s_{iqt} represents the observation of the i^{th} object's q^{th} variable at time t. S_i is a univariate time series if q is equal to 1.

Multivariate time series clustering is defined as a specified set of the multivariate time series S clustered into K clusters, $C_1, C_2, \cdots, C_K$, where C_i ($i = 1, \cdots, K$) is the i^{th} cluster, which contains at least 3 objects in S.

B. WAVELET-BASED CLUSTERING OF MULTIVARIATE TIME SERIES

The maximal overlap discrete wavelet transform (MODWT) [13], [17], [18] is a modified version of the discrete wavelet transform. An example of using the MODWT to analyze time series is available in the literature [18]. Assume that $g_{jl}, l = 0, \cdots, L_j$, is a j-level wavelet filter of length L_j that is associated with the scale $\tau_j \equiv 2^{j-1}$ and the univariate time series $S_{iq} = (s_{iq1}, s_{iq2}, \ldots, s_{iqT})$. An unbiased estimator of the time-independent variance at scale τ_j is

$$\widehat{v}^2_{S_{iq}}(\tau_j) \equiv \frac{1}{N_j} \sum\nolimits_{t=L_j}^{T-1} \widehat{H}^2_{S_{iq},jt} \tag{3}$$

where $\widehat{H}_{S_{iq},jt}$ denotes the MODWT coefficients of the time series S_{iq} and $N_j = T - L_j + 1$ is the number of wavelet coefficients.

The wavelet covariance of two specified univariate time series S_{iq} and S_{ik} with the MODWT coefficients $\widehat{H}_{S_{iq},jt}$ and $\widehat{H}_{S_{ik},jt}$, respectively, is defined as $\widehat{v}_{S_{iq}S_{ik}}(\tau_j) \equiv cov(\widehat{H}_{S_{iq},jt}, \widehat{H}_{S_{ik},jt})$. $\sum_{j=1}^{\infty} \widehat{v}_{S_{iq}S_{ik}}(\tau_j) = cov(S_{iq}, S_{ik})$ if j is infinite. Thus, we obtain the unbiased estimator of the correlation

coefficient between S_{iq} and S_{ik}:

$$\widehat{\rho}_{S_{iq}S_{ik}}(\tau_j) \equiv \frac{\widehat{v}_{S_{iq}S_{ik}}(\tau_j)}{\widehat{v}^2_{S_{iq}}(\tau_j)\widehat{v}^2_{S_{ik}}(\tau_j)} \tag{4}$$

C. SYMBOLIC AGGREGATE APPROXIMATION (SAX)

SAX [14], [19], which was proposed by Keogh E, is an effective discrete dimensionality reduction method of a time series that is based on piecewise aggregate approximation (PAA) [20]. SAX represents the conversion of a time series of length T into a symbol string of length N ($N \ll T$), where N is the number of subseries. Given the univariate time series $S_{iq} = (s_{iq1}, s_{iq2}, \ldots, s_{iqT})$, SAX can be conducted via the following three steps:

Step 1: Normalization. Each original univariate time series S_{iq} is normalized into $S'_{iq} = (s'_{iq1}, s'_{iq2}, \ldots, s'_{iqT})$ with mean 0 and variance 1 via equation (5). The normalization does not affect the shape or scale of the original univariate time series S_{iq}. [21].

$$s'_{iqt} = \frac{s_{iqt} - u_{S_{iq}}}{\sigma_{S_{iq}}} \tag{5}$$

where s_{iqt} is the observed value at time t in the univariate time series S_{iq}, $u_{S_{iq}}$ is the mean of all observed values in the univariate time series S_{iq} and $\sigma_{S_{iq}}$ is the standard deviation of all observed values in the time series S_{iq}.

Step 2: PAA dimensionality reduction. PAA is used to divide the univariate time series S_{iq} of length T into the time series $\bar{S}_{iq} = (\bar{s}_{iq1}, \bar{s}_{iq2}, \ldots, \bar{s}_{iqN})$ of length N ($N \ll T$), according to the subseries length T/N. The mean of each subseries is calculated via equation (6).

$$\bar{s}_{iqn} = \frac{N}{T} \sum_{t=\frac{T}{N}(n-1)+1}^{\frac{T}{N}n} s'_{iqt} \tag{6}$$

Step 3: Symbolic representation. The time series $\bar{S}_{iq}$ of the approximate Gaussian distribution can be divided into α equiprobable intervals, and the breakpoints β_i can be obtained as specified in the literature [19]. The sequence values in the same interval are represented by the same symbol, and the original univariate time series S_{iq} is symbolized as $\widetilde{S}_{iq} = (\widetilde{s}_{iq1}, \widetilde{s}_{iq2}, \ldots, \widetilde{s}_{iqN})$.

For two univariate time series of length T, namely, $S_{iq} = (s_{iq1}, s_{iq2}, \ldots, s_{iqT})$ and $S_{ik} = (s_{ik1}, s_{ik2}, \ldots, s_{ikT})$, SAX is utilized to obtain two symbolic representations of length N: $\widetilde{S}_{iq} = (\widetilde{s}_{iq1}, \widetilde{s}_{iq2}, \ldots, \widetilde{s}_{iqN})$ and $\widetilde{S}_{ik} = (\widetilde{s}_{ik1}, \widetilde{s}_{ik2}, \ldots, \widetilde{s}_{ikN})$. Here, equation (7). Reference [19] is employed to calculate the distance between the two symbolic representations of $\widetilde{S}_{iq}$ and $\widetilde{S}_{ik}$ to determine their similarity.

$$MINDIST(\widetilde{S}_{iq}, \widetilde{S}_{ik}) = \sqrt{\frac{T}{N}} \sqrt{\sum_{n=1}^{N} (dist(\widetilde{s}_{iqn} - \widetilde{s}_{ikn}))^2} \tag{7}$$

where $dist(\widetilde{s}_{iqn} - \widetilde{s}_{ikn})$ represents the distance between the two symbols; the calculation is demonstrated in the literature [19].

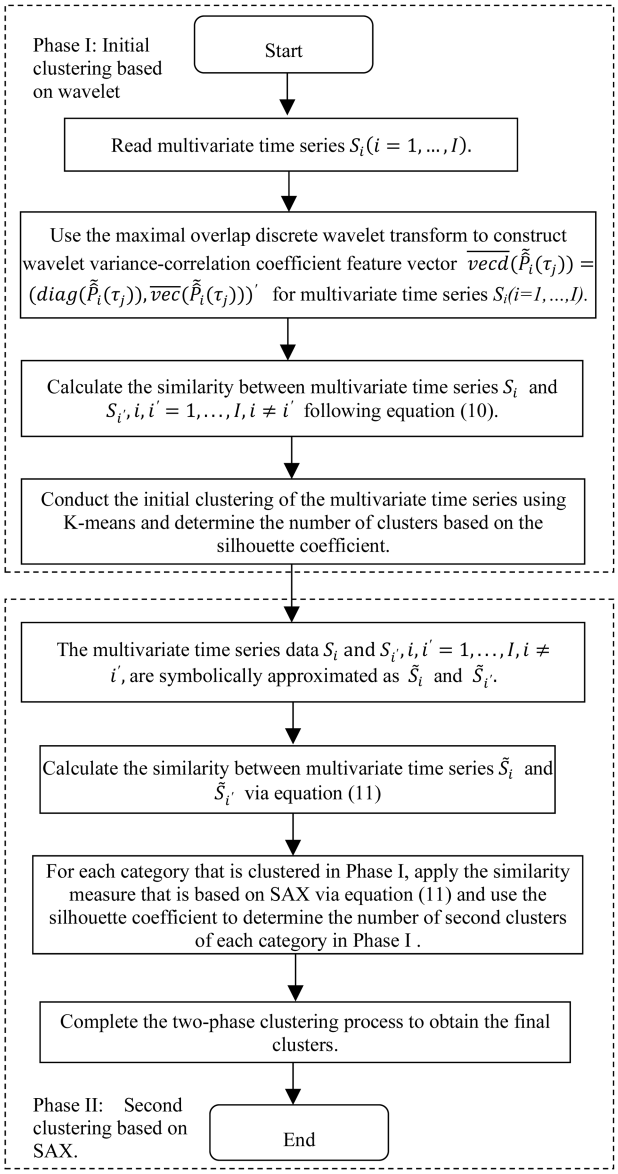

FIGURE 1. Flow chart of WSAX.

III. TWO-PHASE CLUSTERING OF MULTIVARIATE TIME SERIES

The algorithm for two-phase clustering of multivariate time series based on wavelet analysis and SAX is denoted as WSAX and consists of two phases. A flow chart of WSAX is shown in figure 1.

In Phase I, the wavelet variance of each variable and the correlation coefficient between each pair of variables of multivariate time series are calculated. The wavelet variances and correlations are concatenated into a single vector to represent the multivariate time series. The wavelet variance-correlation coefficient feature vector is applied for clustering to yield the Phase I clustering result. The advantage of wavelet analysis is that the inherent correlations between variables of multivariate time series are considered. However, only using wavelet features for multivariate time series clustering will

create the problem that time series with similar variances and covariances but opposite morphologies are clustered into the same category. Therefore, the shape-based method SAX is utilized to perform secondary clustering with the results of the Phase I. In Phase II, SAX is utilized to reduce the dimensionality of each variable of the multivariate time series, and the similarity measure is applied. For each category that was clustered in Phase I, K-means is applied to perform a second clustering of the multivariate time series based on shape.

A. PHASE I: WAVELET-BASED CLUSTERING OF MULTIVARIATE TIME SERIES

For the multivariate time series of each object, first, the wavelet variance of each variable and the correlation coefficient between each pair of variables can be obtained by using the MODWT. This wavelet information is represented as a wavelet variance-correlation coefficient matrix (equation (8)) using equation (3) and equation (4). Wavelet variances and correlations are concatenated into a single vector to represent the multivariate time series, as expressed in equation (9).

$$\widetilde{\widetilde{P}}_i(\tau_j) = \begin{pmatrix} \widehat{v}^2_{s_{i1}}(\tau_j) & \widehat{\rho}_{s_{i1}s_{i2}}(\tau_j)\cdots\widehat{\rho}_{s_{i1}s_{iq}}(\tau_j)\cdots\widehat{\rho}_{s_{i1}s_{iQ}}(\tau_j) \\ \widehat{\rho}_{s_{i2}s_{i1}}(\tau_j) & \widehat{v}^2_{s_{i2}}(\tau_j)\cdots\widehat{\rho}_{s_{i2}s_{iq}}(\tau_j)\cdots\widehat{\rho}_{s_{i2}s_{iQ}}(\tau_j) \\ \vdots & \vdots \quad \vdots \quad \vdots \\ \widehat{\rho}_{s_{iq}s_{i1}}(\tau_j) & \widehat{\rho}_{s_{iq}s_{i2}}(\tau_j)\cdots\widehat{v}^2_{s_{iq}}(\tau_j)\cdots\widehat{\rho}_{s_{iq}s_{iQ}}(\tau_j) \\ \vdots & \vdots \quad \vdots \quad \vdots \\ \widehat{\rho}_{s_{iQ}s_{i1}}(\tau_j) & \widehat{\rho}_{s_{iQ}s_{i2}}(\tau_j)\cdots\widehat{\rho}_{s_{iQ}s_{iq}}(\tau_j)\cdots\widehat{v}^2_{s_{iQ}}(\tau_j) \end{pmatrix} \quad (8)$$

$$\overline{vecd}(\widetilde{\widetilde{P}}_i(\tau_j)) = (diag(\widetilde{\widetilde{P}}_i(\tau_j)), \overline{vec}(\widetilde{\widetilde{P}}_i(\tau_j)))' \quad (9)$$

where $diag(\widetilde{\widetilde{P}}_i(\tau_j)) \equiv (\widehat{v}^2_{s_{i1}}(\tau_j), \widehat{v}^2_{s_{i2}}(\tau_j), \ldots, \widehat{v}^2_{s_{iq}}(\tau_j), \ldots, \widehat{v}^2_{s_{iQ}}(\tau_j))'$; the operator $diag(.)$ returns a vector of the diagonal elements of the matrix $\widetilde{\widetilde{P}}_i(\tau_j)$, namely, $\overline{vec}(\widetilde{\widetilde{P}}_i(\tau_j)) \equiv (\widehat{\rho}_{s_{i1}s_{i2}}(\tau_j), \ldots, \widehat{\rho}_{s_{i1}s_{iq}}(\tau_j), \cdots, \widehat{\rho}_{s_{i1}s_{iQ}}(\tau_j), \ldots, \widehat{\rho}_{s_{i(Q-1)}s_{iQ}}(\tau_j))'$; the operator $\overline{vec}(.)$ creates a $([Q \times (Q+1)/2] \times 1$ column vector from the $Q \times Q$ matrix $\widetilde{\widetilde{P}}_i(\tau_j)$ by stacking the elements of this upper-triangular matrix; and $\overline{vecd}\left(\widetilde{\widetilde{P}}_i(\tau_j)\right)$ creates a new vector by stacking the vectors from the operators $diag(.)$ and $\overline{vec}(.)$.

For two multivariate time series S_i and $S_{i'}, i, i' = 1, \ldots, I, i \neq i'$, after their respective wavelet features have been obtained, equation (10), which is based on the wavelet variance-correlation coefficients, is applied to calculate the Euclidean distance between S_i and $S_{i'}$. K-means is applied to cluster the multivariate time series to yield the Phase I clustering result.

$$\boldsymbol{d_{wvc}(S_i, S_{i'}) = \sum_{j=1}^{J} \overline{vecd}(\widetilde{\widetilde{P}}_i(\tau_j)) - \overline{vecd}(\widetilde{\widetilde{P}}_{i'}(\tau_j))} \quad (10)$$

B. PHASE II: SAX-BASED CLUSTERING OF MULTIVARIATE TIME SERIES

For any object *i*, the data representation is a multivariate time series S_i, as expressed in equation (2), where the q^{th} column is a univariate time series of length T, which is denoted as $S_{iq} = (S_{iq1}, S_{iq2}, \ldots, S_{iqT})'$, and its symbolic representation is a sequence $\widetilde{S}_{iq} = (\widetilde{S}_{iq1}, \widetilde{S}_{iq2}, \ldots, \widetilde{S}_{iqN})'$ of reduced length N (N$\ll$T). The symbolic representation of the multivariate time series S_i is approximated as $\widetilde{S}_i = (\widetilde{S}_{i1}, \widetilde{S}_{i2}, \ldots, \widetilde{S}_{iQ})$.

The symbolic representations of any two multivariate time series S_i and $S_{i'}$ are approximated as $\widetilde{S}_i = (\widetilde{S}_{i1}, \widetilde{S}_{i2}, \ldots, \widetilde{S}_{iQ})$ and $\widetilde{S}_{i'} = (\widetilde{S}_{i'1}, \widetilde{S}_{i'2}, \ldots, \widetilde{S}_{i'Q})$. The similarity between $\widetilde{S}_i$ and $\widetilde{S}_{i'}$ is calculated via equation (11):

$$MDIST(\widetilde{S}_i, \widetilde{S}_{i'}) = \sum_{q=1}^{Q} w_q MINDIST(\widetilde{S}_{iq}, \widetilde{S}_{i'q}), \quad 0 < w_q < 1, \sum_{q=1}^{Q} w_q = 1 \quad (11)$$

To overcome the problem of clustering time series with similar variances and correlation coefficients but with opposite morphology into the same category in Phase I, for each category that is clustered in Phase I, we apply a similarity measure that is based on SAX, as expressed in equation (11), and use the silhouette coefficient to determine the number of second clusters of each category in Phase I.

C. CLUSTERING VALIDITY

The dataset size, classification objective and clustering validity must be considered in determining the number of clusters. Compared with other clustering validity indices [22], the silhouette coefficient is a comprehensive index that is highly suitable for clustering validity analysis of univariate or multivariate time series. The silhouette coefficient [23] considers both the discreteness of all clusters of samples and the coherence of each cluster of samples in the clustering structure. The larger is the silhouette coefficient, the higher is the clustering performance. Assume a total of H ($H \geq 2$) clusters and that the K^{th} cluster has N_K objects. The silhouette coefficient of each object that belongs to cluster $L (1 \leq L \leq H)$ is calculated via equation (12).

$$Sil_i^L = \frac{b_i^L - a_i^L}{\max(a_i^L, b_i^L)} \quad i = 1, \cdots, \sum_{k=1}^{H} N_k \quad (12)$$

where $Sil_i^L (-1 \leq Sil_i^L \leq 1)$ is the silhouette coefficient index of the i^{th} object that belongs to cluster L; a_i^L (equation (13)) is the mean distance between the i^{th} object and the other objects that belong to the same cluster L; and b_i^L (equation (14)) is the minimum of the mean distances between the i^{th} object and all objects that do not belong to cluster L.

$$a_i^L = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{N_L - 1} d_{ij}^{L,L}}{N_L - 1} \quad i = 1, \cdots, N_L \quad (13)$$

$$b_i^L = min_{\substack{K=1 \\ K \neq L}}^{H} (d_{iK}), d_{iK} = \frac{\sum_{j=1}^{N_K} d_{ij}^{L,K}}{N_L - 1} \quad i = 1, \cdots, N_L \quad (14)$$

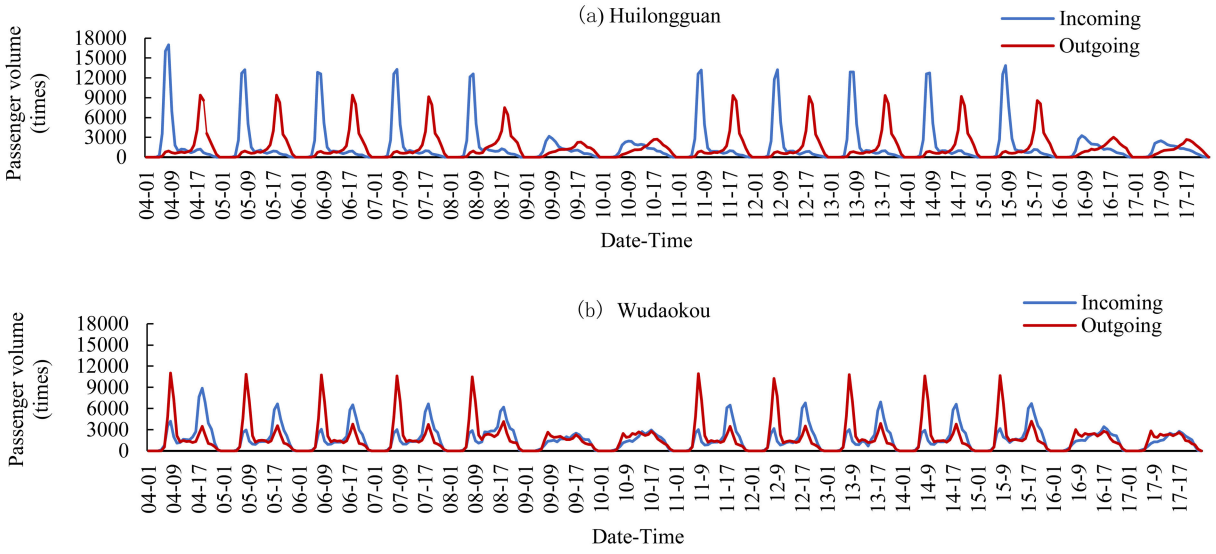

FIGURE 2. Two stations' distributions of the hourly volumes of incoming and outgoing passengers over a two-week period.

where $d_{ij}^{L,K}$ is the Euclidean distance between the i^{th} object that belongs to cluster L and the j^{th} object that belongs to cluster K. In this paper, the mean silhouette coefficient index over all objects (referred to as the mean silhouette coefficient index) is used to evaluate the clustering performance.

IV. EXPERIMENTS AND ANALYSIS

A. DATA DESCRIPTION

We employ card touch in/out data from various rail transit stations in Beijing during a two-week period from March 4 to March 17, 2013. We cleaned the data and selected 195 rail transit stations with complete card touch in/out records for the study.

We aggregate these data in hours according to the incoming and outgoing directions and obtain multivariate time series of incoming and outgoing passenger volumes. Figure 2 plots the data from two stations (Huilongguan and Wudaokou) for two weeks. In Figure 2(a), Huilongguan station shows a single peak each day for each of the incoming and outgoing directions, and the daily peaks are scattered among time intervals. On weekdays, the peaks are higher in the incoming direction than in the outgoing direction, whereas on weekends, they are approximately the same, except that they are substantially lower on weekends than on weekdays. In Figure 2(b), Wudaokou station shows dual peaks for both the incoming direction and the outgoing direction, and the daily peaks are scattered among the time intervals. The peaks in the incoming direction are substantially lower in the morning hours than in the evening hours; the opposite is observed for the outgoing direction. No peaks are readily observed on weekends and the volume of passengers is steady throughout the day.

Data preprocessing normalizes the original incoming and outgoing time series to the time series with a mean of 0 and a variance of 1. The incoming and outgoing time series data for the entire period include T = 336 time points. We use the augmented Dickey-Fuller (ADF) test to perform statistical tests on the stationarity of the incoming and outgoing time series data. The ADF test results of 195 inbound and outbound time series are less than the critical statistical values t the 5% confidence level, and all p-values are close to zero, which shows that the data are stationary time series. The set of maximum and minimum ADF test result values for the incoming and outgoing time series are shown in Table 1.

TABLE 1. Maximum and minimum ADF test result values.

Type	ADF Statistic	p-value	5%
maximum for incoming	-3.08	0.03	-2.87
minimum for incoming	-12.29	0.00	-2.87
maximum for outgoing	-5.24	0.00	-2.87
minimum for outgoing	-13.79	0.00	-2.87

B. EXPERIMENTAL RESULTS

The multivariate time series of incoming and outgoing passenger volumes for each of the 195 rail transit stations are clustered via our method. In the experiment, the wavelet filter is LA(8), the scale is $\tau_j = 2^{j-1}$, for $j = 2, 3, \ldots, 5$, where j is the number of wavelet filter layers, and the number of clusters is $K = 2, 3, \ldots, 10$. To avoid the local optimal solution, for each K, we repeat K-means 100 times, calculate the average of its silhouette coefficient, and select the maximum of the silhouette coefficient that corresponds to K equal to 2 as the optimal number of clusters in Phase I. The average silhouette coefficients are listed in Table 2.

Based on the Phase I clustering results, we conduct a second clustering according to shape for the stations of each category. We apply SAX to the incoming/outgoing passenger volume time series for a symbolic representation with a window interval of 2 and apply equation (12) to calculate the similarity between the rail stations. Next, we conduct

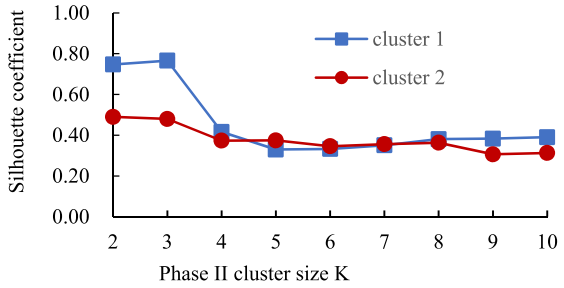

FIGURE 3. Silhouette coefficient vs. K value.

TABLE 2. Average silhouette coefficients.

	Number of clusters K								
j	2	3	4	5	6	7	8	9	10
5	**0.69**	0.56	0.49	0.54	0.49	0.50	0.46	0.45	0.46
4	**0.67**	0.52	0.58	0.52	0.49	0.46	0.45	0.45	0.44
3	**0.67**	0.56	0.47	0.47	0.42	0.41	0.39	0.39	0.40
2	**0.71**	0.60	0.52	0.46	0.45	0.45	0.43	0.42	0.42

TABLE 3. Average wavelet variances and correlation coefficients for the first category and for each cluster of Phase II clustering.

Index	Category 1 in Phase I	Cluster No. in Phase II		
		1	2	3
Incoming variance	0.14	0.07	0.14	0.16
Outgoing variance	0.12	0.08	0.19	0.10
Correlation coefficient	0.21	0.09	0.24	0.21
Percentage of stations	61.54%	6.67%	17.44%	37.44%

TABLE 4. Average wavelet variances and correlation coefficients for the second category and for each cluster of Phase II clustering.

Index	Category 2 in Phase I	Cluster No. in Phase II	
		4	5
Incoming variance	0.14	0.15	0.13
Outgoing variance	0.15	0.13	0.17
Correlation coefficient	0.62	0.62	0.61
Percentage of stations	38.46%	15.38	23.08

K-means clustering to yield the results of Phase II clustering for Categories 3 and 2, according to the clustering validity index-silhouette coefficient (Figure 3) and the presence of at least 3 stations for each cluster. The average wavelet variance, correlation coefficient and station number for the first category and the 3 clusters after Phase II clustering are listed in Table 3 and those for the second category and the 2 clusters after Phase II clustering are listed in Table 4. After the two-phase clustering has been completed, the rail transit stations are divided into 5 clusters.

C. CONTROLLED EXPERIMENT

To evaluate and compare the clustering performances on the multivariate time series of incoming and outgoing passenger volumes of 195 rail transit stations, where the number of clusters K ranges between 4 and 10, we conduct experiments using the wavelet transform (WVC), which is proposed by D'Urso and Maharaj [13]; SAX, which was proposed by Lin *et al.* [31]; and WSAX, which was proposed in this paper. Figure 4 plots the silhouette coefficients versus the K values of the three clustering methods, according to which the silhouette coefficients are larger for WSAX than for WVC and SAX. Hence, the WSAX clustering method is more effective in clustering rail transit stations. Combined with the curve characteristics of Figure 5(a)-(e), our proposed clustering method can accurately differentiate the time series with similar variances but opposite morphologies that were obtained in the initial clustering, which demonstrates the satisfactory performance of WSAX.

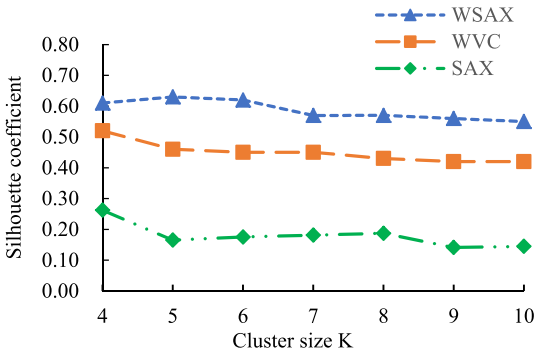

FIGURE 4. Silhouette coefficients vs. K values of the three clustering methods.

The larger is the silhouette coefficients that correspond to the cluster size, the better is the clustering performance, which can be applied to determine the cluster size. In Figure 4, the blue broken line depicts the change in the silhouette coefficients of the WSAX method with the cluster size K, which shows that the silhouette coefficient is the largest when k is equal to 5, that is, the best clustering performance is achieved.

D. ANALYSIS OF THE CLUSTERING RESULTS

Figure 5 (a)-(e) shows the curves for 5 clusters of incoming and outgoing passenger volumes over one week after normalization (March 11-17, 2013). The 5 clusters of 195 rail transit stations that are identified via the WSAX method are listed in the Appendix. In Phase I, clusters 1∼3 in Figure 5(a)-(e) are grouped into the same category, namely, Category 1, and clusters 4 and 5 are grouped into another category, namely, Category 2. According to Tables 3 and 4, the multivariate time series with weak correlations (0.09-0.24) and strong correlations (0.61-0.62) between incoming passenger volumes and outgoing passenger volumes are divided into Category 1 and Category 2, respectively. However, the multivariate time series of Category 1 in Figure 5(a) and Figure 5(b) have opposite morphologies. The same problem is observed in Category 2. In Phase II, this problem is effectively overcome by applying SAX. Category 1 is divided into 3 clusters, namely, cluster 1, cluster 2, and cluster 3, as plotted in Figure 5(a)-(c), and Category 2 is divided into 2 clusters, namely, cluster 4 and cluster 5, as plotted in Figure 5(d)-(e).

The multivariate time series with weak correlations between incoming passenger volumes and outgoing

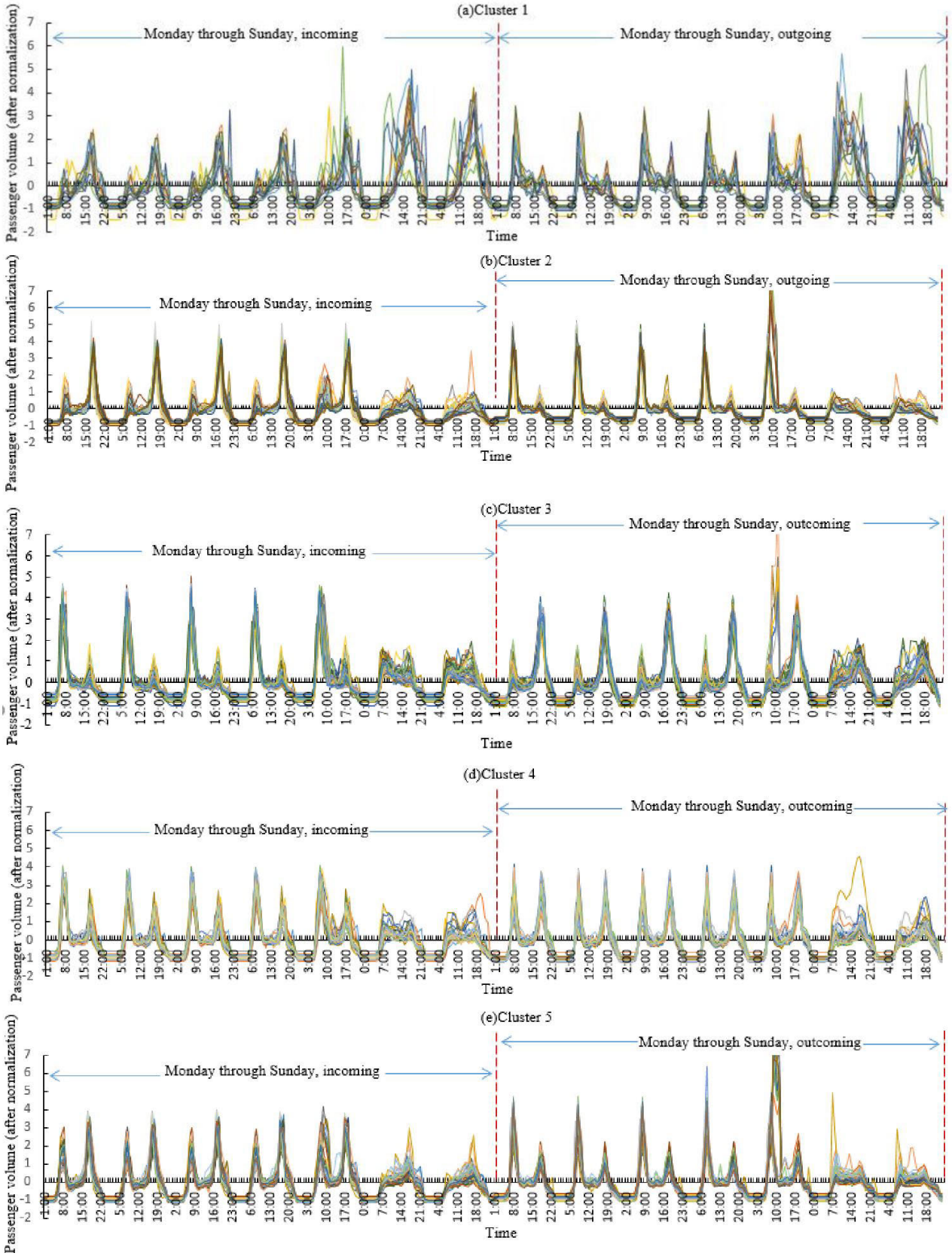

FIGURE 5. Time vs. incoming and outgoing passenger volumes of five categories of stations during one week.

passenger volumes are shown in Figure 5(a)-(c). The curves of the incoming and outgoing passenger volumes exhibit single peaks, except during peak periods. The five clusters, cluster 1, which is plotted in Figure 5(a), show single peaks on weekdays and weekends. The weekday incoming peaks last from 3 p.m. to 9 p.m., and the weekday outgoing peaks last from 7 a.m. to 5 p.m. The weekend incoming/outgoing peaks last longer and are more readily observed. This finding demonstrates that these stations are in integrated service functional areas, namely, they are comprehensive stations (such as stations near scenic locations and large shopping centers). For example, Zoo Station is located near a large attraction, the Beijing Zoo, and a transportation hub; and Olympic Green Station and the Olympic Sports Center Station are located near the Beijing Olympic Park and large shopping malls, such as the XinAo Shopping Mall and Rainbow Shopping Mall. These stations are either located in the same urban functional area or have similar POIs. Cluster 2, which is plotted in Figure 5(b), shows a weekday incoming passenger volume peak between 5 p.m. and 7 p.m. and a weekday outgoing passenger volume peak between 8 a.m. and 10 a.m. Weekends are similar to weekdays but with substantially lower passenger volumes. Hence, these stations are located in business service functional areas, namely, these stations are business-related stations. Zhongguancun Station and Xi'erqi Station are located in typical business areas. Cluster 3, which is plotted in Figure 5(c), shows a weekday incoming passenger volume peak between 8 a.m. and 10 a.m. and a weekday outgoing passenger volume peak between 5 p.m. and 7 p.m. Weekends are similar to weekdays but with substantially lower passenger volumes and a morphology that is exactly the opposite that in Figure 5(b). Hence, these stations are located in residential service functional areas, namely, these stations are residential-related stations. For example, Tiantongyuan Station and Huilongguan Station are located in typical residential communities.

The multivariate time series with strong correlations between incoming passenger volumes and outgoing passenger volumes, which are depicted in Figure 5(d)-(e), show strong dual peaks during peak hours. Cluster 4, which is plotted in Figure 5(d), shows slightly higher morning peaks of weekday incoming passenger volumes, with morphologies that are opposite those of the weekday outgoing passenger volumes. Hence, these stations are located in residential service functional areas and business service functional areas, and the residential function outweighs the business function, namely, these stations are primarily residential- and secondarily business-related stations. Taoranting, Beiyuan and Sihuidong stations, for example, are located near residential areas that also contain business areas. In cluster 5, which is plotted in Figure 5(e), the shape is opposite that in Figure 5(d). Hence, these stations are located in both residential service functional areas and business service functional areas, and the business function outweighs the residential function, namely, these stations are primarily business- and secondarily residential-related stations. Zhichunlu, Xizhimen and Wangjing stations, for instance, are located near business areas that are mixed with residential areas, and the business area functions outweigh the residential area functions.

V. CONCLUSION

The available methods for clustering multivariate time series, which contain multiple variables, are insufficient. In this paper, we propose a two-phase multivariate time series clustering algorithm, namely, WSAX, which combines the advantages of feature-based and shape-based clustering methods. WSAX not only explores the correlations between variables but also considers the similarity in terms of the time series morphology. Phase I obtains the wavelet variance of each variable and the correlation coefficients between the variables via the MODWT. Subsequently, Phase I uses the wavelet variance-correlation coefficient feature vector for clustering. Phase II uses SAX to reduce the dimensionality of each variable of the multivariate time series and applies the similarity measure to realize the second clustering of multivariate time series based on shape. In the experiment, in which real rail transit station data from Beijing IC cards are employed, the WSAX method outperformed the WVC and SAX methods. The rail transit stations were divided into 5 clusters: comprehensive, business-related, residential-related, primarily residential- and secondarily business-related, primarily business- and secondarily residential-related. The clustering validity index silhouette coefficient is employed to compare the WSAX, WVC and SAX methods, which in combination with the clustering results that are presented in figure 5(a)-(e), demonstrates the satisfactory performance and rationality of the WSAX algorithm.

The experimental results demonstrated two main advantages of the proposed WSAX method: 1) This method considers the inherent correlations between the variables of the multivariate time series in Phase I and re-clusters the initial clustering results into a cluster of weak correlations between incoming and outgoing passenger volumes and a cluster of strong correlations. 2) This method considers the morphological similarity of the multivariate time series in Phase II and overcomes the problem of clustering time series with similar variances and correlation coefficients but with opposite morphologies into the same category in Phase I. In addition, the results provide a scientific basis of reference for studying urban functions, planning rail transit stations, and managing related services.

APPENDIX

Five clusters of 195 rail transit stations that were obtained via the WSAX method are listed as follows:

Thirteen rail transit stations in cluster 1: Olympic Green Station, Olympic Sports Center Station, Beihai North Station, Beijing Railway Station, Beijing Zoo Station, Liangxiang University Town Station, Liangxiang University Town North Station, Nanluoguxiang Station, Olympic Green South Gate

Station, Tian'anmen East Station, Tian'anmen West Station, Wangfujing Station, and Xidan Station.

Thirty-four rail transit stations in cluster 2: Baishiqiao South Station, Beiyuanlu North Station, Chaoyangmen Station, Dabaotai Station, Dawanglu Station, Dengshikou Station, Dongdaqiao Station, Dongdan Station, Dongsi Station, Dongsishitiao Station, Fuchengmen Station, Fuxingmen Station, Gaobeidian Station, National Library Station, Guomao Station, China International Exhibition Center Station, Haidian Huangzhuang Station, Hujialou Station, Jianguomen Station, Jintaixizhao Station, Jinghailu Station, Liangmaqiao Station, Lingjinghutong Station, Liufang Station, Agricultural Exhibition Center Station, Rongchangdongjie Station, Rongjingdongjie Station, Biomedical Base Station, Suzhoujie Station, Wanyuanjie Station, Xi'erqi Station, Yonghegong Lama Temple Station, Yong'anli Station, and Zhongguancun Station.

Seventy-three rail transit stations in cluster 3: Anheqiao North Station, Babaoshan Station, Bajiao Amusement Park Station, Baliqiao Station, Beigongmen Station, Caofang Station, Changying Station, Communication University of China Station, Cishousi Station, Ciqu Station, Ciqu South Station, Dalianpo Station, Daotian Station, Fengbo Station, Gaomidian North Station, Gongyixiqiao Station, Gonghuacheng Station, Guanzhuang Station, Guangyangcheng Station, Guoyuan Station, Houshayu Station, Huangcun Railway Station, Huangcunxidajie Station, Huangqu Station, Huilongguan Station, Huilongguan Dongdajie Station, Huoying Station, Jiaomen West Station, Jiukeshu Station, Jiugong Station, Liyuan Station, Libafang Station, Lishuiqiao Station, Lishuiqiao South Station, Liangxiang University Town West Station, Liangxiangnanguan Station, Linheli Station, Liujiayao Station, Longze Station, Majiapu Station, Maquanying Station, Nanfaxin Station, Nanshao Station, Pingguoyuan Station, Puhuangyu Station, Qingnianlu Station, Qingyuanlu Station, Shahe Station, Shahe University Park Station, Life Science Park Station, Shilipu Station, Shimen Station, Shuangqiao Station, Shunyi Station, Songjiazhuang Station, Suzhuang Station, Tiantongyuan Station, Tiantongyuan North Station, Tiantongyuan South Station, Tongzhou Beiyuan Station, Tuqiao Station, Xihongmen Station, Xiyuan Station, Xiaohongmen Station, Xingong Station, Yihezhuang Station, Yizhuangqiao Station, Yongtaizhuang Station, Yuxin Station, Yuanmingyuan Park Station, Zaoyuan Station, Changyang Station, and Zhuxinzhuang Station.

Thirty rail transit stations in cluster 4: Bagou Station, Beijing South Railway Station, Beiyuan Station, Cuigezhuang Station, Gaomidian South Station, Guchenglu Station, Guangximen Station, Haidian Wuluju Station, Hepingmen Station, Hualikan Station, Jishuitan Station, Jintailu Station, Jinsong Station, Lincuiqiao Station, Qianmen Station, Shangdi Station, Shaoyaoju Station, Sihui East Station, Sunhe Station, Taiyanggong Station, Taoranting Station, Tiangongyuan Station, Tongjinanlu Station, Wanshoulu Station, Wukesong Station, Xixiaokou Station, Xiaocun Station, Yizhuang Culture Park Station, Yuquanlu Station, and Changchunjie Station.

Forty-five rail transit stations in cluster 5: Beixinqiao Station, Chongwenmen Station, Ciqikou Station, Datunlu East Station, Hepingli Beijie Station, Hepingxiqiao Station, Huixinxijie Beikou Station, Huixinxijie Nankou Station, Tiantandongmen Station, Zhangzizhonglu Station, Andingmen Station, Anhuaqiao Station, Anzhenmen Station, Peking University East Gate Station, Beitucheng Station, Caishikou Station, Chegongzhuang Station, Chegongzhuang West Station, Dazhongsi Station, Dongzhimen Station, Guloudajie Station, Guogongzhuang Station, Huayuanqiao Station, Jiandemen Station, Military Museum Station, Mudanyuan Station, Muxidi Station, Nanlishilu Station, Ping'anli Station, Renmin University Station, Sanyuanqiao Station, Shuangjing Station, Sihui Station, Tuanjiehu Station, Wangjing Station, Wangjing West Station, Weigongcun Station, Wudaokou Station, Xisi Station, Xitucheng Station, Xizhimen Station, Xinjiekou Station, Xuanwumen Station, Zhichunli Station, and Zhichunlu Station.

REFERENCES

[1] W. Meesrikamolkul, V. Niennattrakul, and C. A. Ratanamahatana, "Shape-based clustering for time series data," in *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science). Kuala Lumpur, Malaysia: Springer-Verlag, 2012, pp. 530–541.

[2] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: A survey and empirical demonstration," *Data Mining Knowl. Discovery*, vol. 7, no. 4, pp. 349–371, Oct. 2003.

[3] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Min. Knowl. Discov.*, vol. 13, no. 3, pp. 335–364, Nov. 2006.

[4] P. P. Rodrigues, J. Gama, and J. P. Pedroso, "Hierarchical clustering of time-series data streams," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 615–627, May 2008.

[5] J. Paparrizos and L. Gravano, "K-shape: Efficient and accurate clustering of time series," *Sigmod Rec.*, vol. 45, no. 1, pp. 69–76, Mar. 2016.

[6] J. Paparrizos and L. Gravano, "Fast and accurate time-series clustering," *ACM Trans. Database Syst.*, vol. 42, no. 2, pp. 1–49, Jun. 2017.

[7] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering–a decade review," *Inf. Syst.*, vol. 53, pp. 16–38, Oct. 2015.

[8] T. W. Liao, "Clustering of time series data-a survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[9] E. A. Maharaj, "Comparison and classification of stationary multivariate time series," *Pattern Recognit.*, vol. 32, no. 7, pp. 1129–1138, Jul. 1999.

[10] J. Ye, R. Janardan, and Q. Li, "GPCA: An efficient dimension reduction scheme for image compression and retrieval," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Seattle, WA, USA, 2004, pp. 354–363.

[11] C. Guo, H. Jia, and N. Zhang, "Time series clustering based on ICA for stock data analysis," in *Proc. 4th Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Dalian, China, Oct. 2008, pp. 1–4.

[12] E. H. Wu and L. Philip, "Independent component analysis for clustering multivariate time series data," in *Advanced Data Mining and Applications* (Lecture Notes in Computer Science). Wuhan, China: Springer-Verlag, 2005, pp. 474–482.

[13] P. D'Urso and E. A. Maharaj, "Wavelets-based clustering of multivariate time series," *Fuzzy Sets Syst.*, vol. 193, pp. 33–61, Apr. 2012.

[14] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining Knowl. Discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[15] H. Li, "Multivariate time series clustering based on common principal component analysis," *Neurocomputing*, vol. 349, pp. 239–247, Jul. 2019.

[16] G. Du, L. Zhou, L. Wang, and H. Chen, "Multivariate time series clustering via multi-relational community detection in networks," in *Web and Big Data* (Lecture Notes in Computer Science). Macau, China: Springer-Verlag, 2018, pp. 138–145.

[17] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2006, pp. 159–205.
[18] K. Roushangar and F. Alizadeh, "Using multi-temporal analysis to classify monthly precipitation based on maximal overlap discrete wavelet transform," *J. Hydroinform.*, vol. 21, no. 4, pp. 541–557, Jul. 2019.
[19] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proc. 8th ACM SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery (DMKD)*, San Diego, CA, USA, 2003, pp. 2–11.
[20] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM Trans. Database Syst.*, vol. 27, no. 2, pp. 188–228, Jun. 2002.
[21] D. Q. Goldin and P. C. Kanellakis, "On similarity queries for time-series data: Constraint specification and implementation," in *Principles and Practice of Constraint Programming—CP'95* (Lecture Notes in Computer Science). Cassis, France: Springer-Verlag, 1995, pp. 137–153.
[22] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013.
[23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

LIYING ZHANG received the B.S. and M.S. degrees in computer science and technology from the China University of Petroleum, Beijing, in 2001 and 2004, respectively, and the Ph.D. degree in cartography and geographic information engineering from the China University of Mining and Technology, Beijing, in 2019. She is currently a Lecturer with the College of Information Science and Engineering, China University of Petroleum. Her current research interests include spatio-temporal data mining and machine learning.

TAO PEI received the Ph.D. degree from the China University of Geosciences, in 1998. He is currently a Professor with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research. His research interests include spatial big data mining and Geostatistics.

BIN MENG received the Ph.D. degree from the Institute of Geographic Sciences and Natural Resources Research, in 2005. He is currently a Professor with Beijing Union University. His main research interests include urban geography and geographic information science.

YUANFENG LIAN received the Ph.D. degree from Beihang University, China, in 2013. He is currently an Associate Professor with the China University of Petroleum, Beijing, China. His research interests include computer graphics and visualization, image analysis, and informatics.

ZHOU JIN was born in Jiangsu, China. She received the B.E. degree in computer science and technology from Nanjing University, Nanjing, China, in 2010, and the M.E. and Ph.D. degrees with the Graduate School of Information, Production and Systems, Waseda University, Kitakyushu, Japan, in 2012 and 2015, respectively. From 2015 to 2017, she was a Postdoctoral Researcher with the Research Center, Waseda University. She is currently a Lecturer with the College of Information Science and Engineering, China University of Petroleum, Beijing. Her research interests include verification technologies for nonlinear circuits and systems, LSI simulation technologies, and intelligence computation technologies.