A Climate Classification of China through k -Nearest-Neighbor and Sparse Subspace Representation

JIARONG SHI

*State Key Laboratory of Green Building in Western China, and School of Architecture, and School of Science,
Xi'an University of Architecture and Technology, Xi'an, China*

LIU YANG

*State Key Laboratory of Green Building in Western China, and School of Architecture, Xi'an University of Architecture
and Technology, Xi'an, China*

(Manuscript received 29 November 2018, in final form 19 July 2019)

ABSTRACT

Climate classification aims to divide a given region into distinct groups on the basis of meteorological variables, and it has important applications in fields such as agriculture and buildings. In this paper, we propose a novel spectral clustering-based method to partition 661 meteorological stations in China into several zones. First, the correlations are analyzed among five meteorological variables: daily average temperature, average relative humidity, sunshine hours, diurnal temperature range, and atmospheric pressure. The resulting weak linear correlation supports the classification under multiple views. Next, a similarity matrix/graph is constructed by combining the advantages of k -nearest-neighbor and sparse subspace representation. The blocking effect of the similarity matrix confirms the rationality of the classification. Then, we consider respectively the climate classification under a single view and multiple views. For the single view, atmospheric pressure has the highest imbalance degree, and sunshine hours and diurnal temperature range have the strongest classification consistency. The consistency rates are improved evidently in a multiple-view situation. Afterward, we propose a determining method for the number of clusters and make a sensitivity analysis on various parameters. Finally, we provide a further statistical analysis on the classification results and compare the consistency with another climate dataset. All experimental results show that the proposed classification method is feasible and effective in the climate zoning of China.

1. Introduction

There are many different climate types across the world and they have distinctive influences on several fields such as agriculture and energy usage of most residential and commercial buildings (Briggs et al. 2003a). Given sufficient climate data, climate classification will provide valuable guidelines to agriculture and building design. There are many climate classification methods based on different meteorological variables and indices. A proper and reasonable classification method can be chosen according to the purpose and motivation of classification.

Considering the impact of climate on building design principles and architectural similarities, Olgyay (2015) proposed four main climate types around the world,

namely cool, temperate, hot and arid, and hot and humid. To improve the building energy codes and standards in the United States, a new climate classification technique has been developed in which 17 climate zones were designed (Briggs et al. 2003b). Recently, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standard determined nine thermal climate zones using the heating and cooling degree-days of all stations under study (ASHRAE 2013). In China, a cumulative temperature above 10°C is an important index for defining temperature zones. The latest climate regionalization was established by using the daily climate observations from 609 stations (Zheng et al. 2010). For the thermal design of buildings, five climate types were presented, namely severe cold, cold, hot summer and cold winter, mild, and hot summer and warm winter (Ministry of Construction of China 1993).

Corresponding author: Dr. L. Yang, 626224056@qq.com

DOI: 10.1175/JCLI-D-18-0718.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

The climate classification criteria usually involve the mean temperature of the coldest and warmest month, annual cooling degree-days, annual heating degree-days, precipitation, humidity, and solar irradiation. When classifying climate zones or regions, we should obey an important principle that two points or samples from the same zone should have similar climate characteristics and those from different zones have dissimilar characteristics. In a detailed classification, the observation data are typically from a set of meteorological stations or grid points. Meanwhile, climate data of one or more variables over many years are required (Mimmack et al. 2001). Mathematically, cluster analysis, also named data clustering (Jain 2010), is just a statistical classification technique and it is commonly used in climate classification.

The goal of cluster analysis is to group samples according to similarity or dissimilarity, where the samples from the same cluster have similar features and the features of samples from different clusters are as dissimilar as possible (Filippone et al. 2008). Clustering methods can be broadly divided into two categories: hierarchical and partitioning. The first category is used to seek nested clusters in an agglomerative manner. The second category finds all clusters simultaneously by solving a minimization problem whose objective function measures the dissimilarities between each sample and its corresponding potential cluster center.

The following cites several representative climate classification examples based on cluster analysis. Unal et al. (2003) used hierarchical cluster analysis to redefine climate zones of Turkey based on temperatures (mean, maximum, and minimum) and total precipitation. Subsequently, hierarchical clustering was also employed to develop five solar climate zones according to the monthly average clearness index from 123 stations (Lau et al. 2007). Lund and Li (2009) considered three agglomerative methods and classified the state of Colorado into six zones based on monthly maximum and minimum temperatures from 292 meteorological stations. Wan et al. (2010) designed a hierarchical clustering on annual cumulative heat and cold stress derived by monthly minimum dry-bulb temperature (DBT), maximum DBT, and vapor pressure, and classified 3973 grids in China into five bioclimate zones. Gerstengarbe et al. (1999) discussed the cluster number and cluster separation of nonhierarchical cluster analysis, and investigated the classification of European climate using monthly precipitation, temperature, and temperature range at 228 meteorological stations. A *K*-means clustering, a simple and popular nonhierarchical clustering method, was used to classify China into different climate types based on monthly temperature and

precipitation from 753 meteorological stations (Zhang and Yan 2014). The expectation maximization clustering was utilized to find climate regions based on the grid climate data over Indian regions (Mallenahalli 2015). Netzel and Stepinski (2016) considered monthly precipitation, temperature, and temperature range from grid points obtained by interpolation, and adopted two clustering methods: hierarchical clustering and partitioning around medoids. Three climate zoning methods (degree-days, cluster analysis, and administration divisions) were compared for building energy efficiency applications in Nicaragua (Walsh et al. 2017). Recently, a two-tier climate zoning method was applied to the hot summer and cold winter zone of China, where the second tier adopted hierarchical agglomerative clustering on climate data (degree-days, relative humidity, solar radiation, and wind speed) (Xiong et al. 2019). The *K*-means algorithm was used for the climate zoning in the extreme southern part of Chile (Verichev et al. 2019).

Regardless of the clustering method that is used, we have to face two critical problems, that is, how to determine the number of clusters and how to measure the dissimilarity or similarity between two samples. The paper will focus on the latter problem. Dissimilarity is equivalent to distance and similarity is the opposite of dissimilarity. The distance measurements are mainly composed of the widely used Euclidean distance of original data, Mahalanobis distance (Mimmack et al. 2001), and Euclidean distance of normalized data (Lund and Li 2009). The similarity measurements include the cross-correlation function (Steinhaeuser et al. 2010) and dynamic time warping (Netzel and Stepinski 2016). However, the aforementioned distance or similarity may not be the most appropriate measurement for high-dimensional and noisy data.

Among all clustering methods, *K*-means clustering attracts wide attention for its simplicity, ease of implementation, efficiency, and empirical success (Jain 2010). However, this clustering technique is not suitable for high-dimensional data. As a recently emerged alternative of *K*-means clustering, spectral clustering has many fundamental advantages over traditional clustering algorithm such as *K*-means clustering and hierarchical clustering (Von Luxburg 2007). Spectral clustering originates from the spectral graph theory and it configures the clustering problem as a graph cut problem (Filippone et al. 2008). In spectral clustering, we first construct a similarity graph/matrix (or affinity matrix) and then calculate the graph Laplacian matrix. Finally, *K*-means clustering is employed to cluster on the eigenvectors of the graph Laplacian matrix. The most difficult problem in spectral clustering is still how to

build a good similarity matrix. To construct a better similarity matrix using global information, several spectral clustering-based methods were proposed successively. For instance, spectral curvature clustering constructed multiway similarities taking the curvature into account (Chen and Lerman 2009). Low-rank recovery (Liu et al. 2013) and low-rank subspace clustering (Vidal and Favaro 2014) obtained the similarity graph by seeking a low-rank representation of the data. Sparse subspace clustering built the similarity matrix by sparse subspace representation of all data (Elhamifar and Vidal 2013). One prominent advantage of these spectral clustering methods is that they are robust to noise and outliers. Among them, sparse subspace clustering is competitive with other state-of-the-art subspace clustering algorithms.

This paper proposes a novel spectral clustering-based method to perform climate classification. This method combines the advantages of k -nearest-neighbor and sparse subspace representation, and possesses strong stability and robustness. Furthermore, it can also deal with high-dimensional cases. We employ the proposed method to divide China into several zones according to different meteorological variables. These include daily average temperature, average relative humidity, sunshine hours, diurnal temperature range, and atmospheric pressure.

2. Data and methods

This section will introduce the climate data used in this paper, provide two manners of building similarity matrix, and present a new spectral clustering method.

a. Climate data

China has a vast and diverse landscape and its climate is dominated mainly by dry seasons and wet monsoons. There exists a significant temperature difference between winter and summer. We choose 661 national meteorological stations in China and perform climate classification. Figure 1 lists the spatial distribution of all stations, where the color of each station is used to indicate the altitude whose unit is 1 m. It can be seen from Fig. 1 that the east of China has a denser station distribution and lower altitude than the west.

The surface-based observation data of 661 stations in the period 2004–13 were obtained from the National Climate Center of China. Five main meteorological variables are considered as the classification indexes: daily average temperature (0.1°C), average relative humidity (%), sunshine hours (0.1 h), diurnal temperature range (0.1°C), and atmospheric pressure (10 Pa). To suppress adverse effects of stochastic fluctuation, we

FIG. 1. Spatial distribution of 661 stations in China.

split a single time series into 365 sub-time series according to 10 consecutive days, and average the observation values of each sub-time series. Therefore, for a fixed meteorological variable, the observation values of each station are represented by a d -dimensional column vector, where $d = 365$. The above dimensionality reduction method can also preserve the interannual variability.

b. Construction of similarity matrix

1) k -NEAREST-NEIGHBOR

For each meteorological variable, we investigate the data from N meteorological stations. Obviously, the value of N is set to 661 for given data. Let $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ be the set of climate data, where $\mathbf{y}_i \in \mathbb{R}^{d \times 1}$ is the observation vector in the average of 10-day periods for the i th station over many years. In traditional data analysis, a common assumption is that the dataset Y lies in a linear subspace. Under this circumstance, the distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ is measured by the l_2 norm: $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|_2$. The above distance is also named as the Euclidean distance.

The first step in spectral clustering is to build a weighted graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of N nodes and $\mathcal{E}$ is a set of edges between two nodes. The i th node is used to denote the observation vector $\mathbf{y}_i$ and the graph G is also called the similarity graph. Constructing the similarity graph is the most crucial step for spectral clustering. There are several popular construction manners of the similarity graph such as the ϵ -neighborhood graph, the k -nearest-neighbor graph, and the fully connected graph. Among these manners, the k -nearest-neighbor graph is generally recommended as the first choice (Von Luxburg 2007). In the following, we introduce the construction of similarity matrix based on the k -nearest-neighbor graph.

For each $\mathbf{y}_i$, we first calculate $N - 1$ distances $\{d_{i1}, d_{i2}, \dots, d_{i,i-1}, d_{i,i+1}, \dots, d_{iN}\}$ and then sort these distances by the increasing order. As a result, k nearest neighbors for $\mathbf{y}_i$ can be found by k smallest distances. Based on this, we build a similarity graph G : if $\mathbf{y}_j$ is one of k nearest neighbors for $\mathbf{y}_i$ or $\mathbf{y}_i$ is one of k nearest neighbors for $\mathbf{y}_j$, then we add an edge between the i th and the j th nodes; otherwise, there is no edge between the two nodes. Finally, the similarity matrix $\mathbf{W} = (w_{ij})_{N \times N}$ is computed by the resulting similarity graph, where w_{ij} is the similarity between data points $\mathbf{y}_i$ and $\mathbf{y}_j$. If there is an edge between the i th and the j th nodes, then we set $w_{ij} = 1$; otherwise, $w_{ij} = 0$. The constructed similarity matrix $\mathbf{W}$ is symmetrical and binary, and $w_{ij} = 1$ means $\mathbf{y}_i$ and $\mathbf{y}_j$ have large similarity. In addition, $\mathbf{W}$ is also sparse in the case of $k \ll N$.

2) SPARSE SUBSPACE REPRESENTATION

In the aforementioned subspace clustering, we make an additional assumption that the dataset Y lies in the union of c linear subspaces. The aim of clustering analysis is to identify and separate these c subspaces. If Y is noise-free and there are sufficient samples or points in each subspace, then each $\mathbf{y}_i$ can be represented by the linear combination of other samples:

$$\mathbf{y}_i = c_{1i}\mathbf{y}_1 + c_{2i}\mathbf{y}_2 + \dots + c_{i-1,i}\mathbf{y}_{i-1} + c_{i+1,i}\mathbf{y}_{i+1} + \dots + c_{Ni}\mathbf{y}_N, \quad (1)$$

where $c_{ji} (j \neq i)$ is the linear representation coefficient of the j th sample. Large $|c_{ji}|$ means that the i th and the j th samples might have strong similarity.

Denote $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$, $\mathbf{c}_i = (c_{1i}, c_{2i}, \dots, c_{Ni})^T$, where $c_{ii} = 0, i = 1, 2, \dots, N$. Then Eq. (1) is rewritten as $\mathbf{y}_i = \mathbf{Y}\mathbf{c}_i$. Generally, the linear system $\mathbf{y}_i = \mathbf{Y}\mathbf{c}_i$ with respect to $\mathbf{c}_i$ has infinite solutions when $d < N$. We usually desire that the solution vector $\mathbf{c}_i$ has some special structure such as lower model complexity. In practical implementation, the optimal $\mathbf{c}_i$ can be determined by solving the following l_p norm minimization problem:

$$\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_p, \text{ s.t. } \mathbf{y}_i = \mathbf{Y}\mathbf{c}_i, c_{ii} = 0. \quad (2)$$

The value of p is commonly set to 1 or 2. The case $p = 1$ corresponds to the sparse representation, and it has made exciting breakthroughs in the field of pattern recognition (Wright et al. 2009). If samples $\mathbf{y}_i$ and $\mathbf{y}_j$ come from different subspaces, then we have $c_{ij} = 0$. When the sample numbers of c subspaces are roughly equivalent, the proportion of zero elements in optimal vector $\mathbf{c}_i$ is approximately $(c - 1)/c$, which indicates $\mathbf{c}_i$ is very sparse for large c .

The real observation data are usually contaminated by noise or even large sparse noise. Under this situation, vector $\mathbf{y}_i$ is decomposed as the sum of three terms:

$$\mathbf{y}_i = \mathbf{Y}\mathbf{c}_i + \mathbf{e}_i + \mathbf{z}_i, \quad (3)$$

where $\mathbf{z}_i$ is a dense noise vector and $\mathbf{e}_i$ is a large sparse noise vector. We suppose that $\mathbf{z}_i$ is a Gaussian white noise vector, and both $\mathbf{c}_i$ and $\mathbf{e}_i$ obey two different multivariate Laplacian distributions. According to the maximum likelihood estimation, we can establish a minimization problem to obtain the optimal linear representation:

$$\begin{aligned} \min_{\mathbf{c}_i, \mathbf{e}_i, \mathbf{z}_i} \|\mathbf{c}_i\|_1 + \mu_1 \|\mathbf{e}_i\|_1 + \frac{1}{2} \mu_2 \|\mathbf{z}_i\|_2^2, \\ \text{s.t. } \mathbf{y}_i = \mathbf{Y}\mathbf{c}_i + \mathbf{e}_i + \mathbf{z}_i, c_{ii} = 0, \end{aligned} \quad (4)$$

where μ_1 and μ_2 are two positive regularization parameters.

The above optimization problem is more robust to dense or large sparse noise than problem (2). We call the representation based on problem (4) a sparse subspace representation. In practice, we further consider an assumption that each subspace is reduced to an affine space. Consequently, the constraint $\mathbf{c}_i^T \mathbf{1} = 1$ is imposed on the constraints of problem (4), where $\mathbf{1}$ is an N -dimensional column vector whose entries are all ones. In the following, we take the temperature variable of Huashan station (34.48°N, 110.08°E) located at Shaanxi Province of China to illustrate the superiority of sparse subspace representation. The observation vector of the chosen station is represented as the linear combination of vectors from the other 660 stations. The linear representation coefficients obtained by solving problems (2) and (4) are compared in Fig. 2.

We can draw two conclusions from Fig. 2. Compared with the l_2 norm minimization model, more coefficients in the l_1 norm minimization are close to zero, which means linear representation based on l_1 norm is more sparse than the l_2 norm. Furthermore, most coefficients in sparse subspace representation are zeros and only a few coefficients are larger than zero. In other words, the sparse subspace representation yields the most sparse representation. This conclusion is consistent with the linear subspace assumption with noise. In summary, the sparse subspace representation model has better noise robustness than the other two models in this example.

All linear representation coefficient vectors $\{\mathbf{c}_i\}_{i=1}^N$ can be integrated into a comprehensive model. Let $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_N)$, $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N)$, and $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)$. Then we have $\mathbf{Y} = \mathbf{Y}\mathbf{C} + \mathbf{E} + \mathbf{Z}$. The sparse subspace

FIG. 2. Coefficients under different representations. (a) Representation based on l_2 norm. (b) Representation based on l_1 norm. (c) Sparse subspace representation.

representations of all stations can be obtained simultaneously by solving the minimization problem:

$$\begin{aligned} \min_{\mathbf{C}, \mathbf{E}, \mathbf{Z}} & \|\mathbf{C}\|_1 + \mu_1 \|\mathbf{E}\|_1 + \frac{1}{2} \mu_2 \|\mathbf{Z}\|_F^2, \\ \text{s.t. } & \mathbf{Y} = \mathbf{Y}\mathbf{C} + \mathbf{E} + \mathbf{Z}, \mathbf{C}^T \mathbf{1} = \mathbf{1}, \text{diag}(\mathbf{C}) = \mathbf{0}, \end{aligned} \quad (5)$$

where $\|\cdot\|_1$ and $\|\cdot\|_F$ indicate respectively the l_1 and the Frobenius norms of a matrix, and $\text{diag}(\cdot)$ is the vector composed by the main diagonal of a matrix. By eliminating the variable matrix $\mathbf{Z}$, problem (5) is reformulated as

$$\begin{aligned} \min_{\mathbf{C}, \mathbf{E}} & \|\mathbf{C}\|_1 + \mu_1 \|\mathbf{E}\|_1 + \frac{1}{2} \mu_2 \|\mathbf{Y} - \mathbf{Y}\mathbf{C} - \mathbf{E}\|_F^2, \\ \text{s.t. } & \mathbf{C}^T \mathbf{1} = \mathbf{1}, \text{diag}(\mathbf{C}) = \mathbf{0}. \end{aligned} \quad (6)$$

The optimal $\mathbf{C}$ in the above optimization problem measures the similarities among all stations or nodes.

Although problem (6) is convex and continuous, it is not smooth. The alternating direction method of multipliers (ADMM) is a very efficient method for solving the distributed optimization problem with linear equality constraints (Boyd et al. 2011). ADMM adopts the alternating update strategy; that is, we minimize or maximize the augmented Lagrangian function of problem (6) with respect to each block variable at each iteration. Once the optimal solution of the above problem is obtained, we can calculate the weights on the edge set $\mathcal{E}$ based on matrix $\mathbf{C}$. For this purpose, each column of $\mathbf{C}$ is first normalized as $\mathbf{c}_i \leftarrow \mathbf{c}_i / \|\mathbf{c}_i\|_\infty$, where $\|\cdot\|_\infty$ is the infinite norm of a vector. Then we set the weights matrix $\mathbf{W} = (|\mathbf{C}| + |\mathbf{C}|^T)/2$, where $|\cdot|$ denotes the absolute value for each element of a matrix. It is obvious that $\mathbf{W}$ is nonnegative and the maximum value of each column is within the interval $[0.5, 1]$. Furthermore, $\mathbf{W}$ is also sparse due to the fact that $\mathbf{C}$ is sparse.

c. Spectral clustering through k -nearest-neighbor and sparse subspace representation

Let $\mathbf{W}_{\text{knn}}$ be the similarity matrix constructed by the k -nearest-neighbor graph and $\mathbf{W}_{\text{ssr}}$ be the similarity matrix obtained by sparse subspace representation. Both $\mathbf{W}_{\text{knn}}$ and $\mathbf{W}_{\text{ssr}}$ are sparse and symmetric in view of their construction principles. The k -nearest-neighbor is a commonly used means for graph construction, and it is based on the assumption that the data space is locally linear. The main disadvantage of k -nearest-neighbor method is that it does not utilize the global information, which may result in the defect that this method is not robust to data noise. In contrast, the sparse subspace representation method adopts the overall contextual information. The sparsity-based method can convey valuable information in the task of classification, offer the datum-adaptive neighborhoods, and possess greater robustness to noise. In essence, the similarity graph constructed by sparse subspace representation is a robust version of the l_1 graph (Cheng et al. 2010). However, a critical deficiency of the sparse subspace method is that two samples with very large Euclidean distance may be classified as the same cluster, which is generally contrary to our intuition. To sum up, these two graph construction approaches complement each other.

To overcome the shortcomings of k -nearest-neighbor and sparse subspace representation, we propose a weighted similarity matrix defined by the compromise between $\mathbf{W}_{\text{knn}}$ and $\mathbf{W}_{\text{ssr}}$:

$$\mathbf{W} = \beta \mathbf{W}_{\text{knn}} + (1 - \beta) \mathbf{W}_{\text{ssr}}, \quad (7)$$

where the trade-off parameter $\beta \in [0, 1]$. If $\beta = 0$, the similarity matrix is constructed by sparse subspace representation. If $\beta = 1$, the similarity matrix is constructed by the k -nearest-neighbor graph. Large β indicates great importance of $\mathbf{W}_{\text{knn}}$. Based on the resulting matrix $\mathbf{W}$, we can divide the dataset Y into c clusters through the spectral

clustering method. The following lists the implementation procedure of spectral clustering (Ng et al. 2002):

- Step 1. Construct a diagonal matrix $\mathbf{D}$ whose (i, i) element is the sum of the i th row of $\mathbf{W}$.
- Step 2. Calculate the Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$.
- Step 3. Carry out eigen-decomposition on $\mathbf{L}$ and obtain its c mutually orthogonal unit eigenvectors $\{\mathbf{x}_i \in \mathbb{R}^{N \times 1}\}_{i=1}^c$ corresponding to c largest eigenvalues.
- Step 4. Form the matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_c)$ and normalize its rows to have unit length in the sense of the l_2 norm.
- Step 5. Treat each row of normalized $\mathbf{X}$ as a point and cluster these N points into c clusters via K -means clustering.
- Step 6. The observation vector $\mathbf{y}_i$ has the same class label with the i th row of normalized $\mathbf{X}$.

If $c < d$, spectral clustering means that we first perform dimensionality reduction on the original data and then cluster them in fewer dimensions by K -means clustering. Compared with directly using K -means clustering on the original data, the spectral clustering can preserve the nonlinear manifold structure to some extent.

3. Results

Because China is roughly partitioned into five zones or clusters in the thermal design of buildings, the default value of c is 5 in this section. The experiments consist of six parts: correlation analysis, visualization of similarity matrices for five meteorological variables, single-view climate classification, multiple-view climate classification, determination of the number of clusters, and sensitivity analysis.

a. Correlation analysis

The exploration of correlation structure among variables is necessary for climate classification. For two arbitrary meteorological variables, we will investigate the correlation coefficients between two stations. Daily average temperature is an extremely important and special meteorological variable; the correlation coefficients between two stations are larger than 0.67 and 94.65% of coefficients are larger than 0.9. The above results show that two stations generally have a strong positive linear relationship, which can be explained as the fact that the temperature has a similar trend everywhere; namely, it reaches the highest value in summer and the lowest value in winter. The correlation matrices among five variables are displayed as 15 images shown in

Fig. 3. In this figure, the x axis and y axis denote the number of stations. Moreover, Fig. 3 adopts the following abbreviations:

- Tem = daily average temperature,
- Hum = average relative humidity,
- SunHour = sunshine hours,
- TemRange = diurnal temperature range, and
- Pressure = atmospheric pressure.

Two conclusions are drawn by observing Fig. 3. For identical variables, each correlation coefficient matrix shows an apparent blocking phenomenon (see Figs. 3a,f,j,m,o), which will be conducive to classification. As for two different variables, the correlation coefficients vary in a relatively large range, which indicates that different variables do not have strong linear relationships. In summary, it is reasonable to carry out climate classification based on the five meteorological variables we selected.

b. Visualization of similarity matrix

This subsection sets the neighbor number $k = 10$ in constructing the k -nearest-neighbor graph. In minimization problem (6), we denote $\mu_1 = \alpha_1 / \min_i \max_{j \neq i} \|\mathbf{y}_j\|_1$, $\mu_2 = \alpha_2 / \min_i \max_{j \neq i} |\mathbf{y}_i^T \mathbf{y}_j|$. The setting of α_1 and α_2 must satisfy the conditions $\alpha_1 > 1$ and $\alpha_2 > 1$. We carry out respectively experiments on five meteorological variables. For the sake of brevity, both α_1 and α_2 are set to 50, and the compromise parameter β is chosen as 0.5.

For the constructed similarity graph, we expect that the similarity matrix $\mathbf{W}$ is sparse. The index of sparsity rate is employed to measure the sparsity of a matrix and its definition is given by

$$\text{SR}(\mathbf{W}) = 1 - \|\mathbf{W}\|_0 / N^2, \quad (8)$$

where $\|\cdot\|_0$ is the l_0 norm of a matrix (i.e., the number of nonzero elements). The value of $\text{SR}(\mathbf{W})$ lies in the interval $[0, 1]$. If $\mathbf{W}$ is the zero matrix, then $\text{SR}(\mathbf{W}) = 1$. If all elements of $\mathbf{W}$ are nonzero, then $\text{SR}(\mathbf{W}) = 0$. Large $\text{SR}(\mathbf{W})$ means large sparsity.

The sparsity rates of five similarity matrices are daily average temperature (0.9713), average relative humidity (0.9656), sunshine hours (0.9634), diurnal temperature range (0.9663), and atmospheric pressure (0.7630). We can see that the first four meteorological variables have very sparse similarity matrices and the last variable (i.e., atmospheric pressure) has the worst sparsity. Figure 4 visualizes the similarity matrices of five meteorological variables, where the white pixel represents 0 and the dark red indicates 1. The first four subfigures in Fig. 4 indicate that most nonzero elements of the similarity

FIG. 3. Visualization of correlation coefficients matrices.

matrices are concentrated near the subdiagonal, but the last subfigure shows a significant blocking effect in the subdiagonal. This observation can be interpreted as follows. The data matrix of all stations is stacked by the increasing order of station number. In general, the adjacent stations have similar meteorological characteristics owing to their relatively small geographic distances. In addition, the first four similarity matrices are sparser than the last matrix, which may be caused by small change of atmospheric pressure. By comparing Figs. 3 and 4, we find out that the similarity matrices are more sparse than the correlation coefficients matrices. This observation shows that it is unwise to measure the similarities by correlation coefficients.

In the subsequent experiments, spectral clustering is used to classify all stations according to the constructed

similarity graph or matrix under single view or multiple views. As an important component of spectral clustering, *K*-means clustering is sensitive to the centroids initialization. To reduce the uncertainty of clustering results, we repeat the *K*-means clustering 20 times and seek the solution with the lowest sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. We will plot cluster locations, the neighbors graph, and the mean of each zone. To simplify the neighbor relationship, we only draw the edges whose weights or similarities are larger than 0.2 in the neighbors graph. Moreover, the edge from the same zone is plotted by a solid line whose color is the same as its nodes, and the edge from different zones is drawn by a dotted line whose color is black.

FIG. 4. Visualization of similarity matrices for five meteorological variables: (a) daily average temperature, (b) average relative humidity, (c) sunshine hours, (d) diurnal temperature range, and (e) daily atmospheric pressure.

c. Single-view climate classification

1) DAILY AVERAGE TEMPERATURE

All stations are divided into five zones according to the data of daily average temperature. We visualize the neighbors graph to display the similarity relationship among all stations. The mean of each zone is utilized to evaluate the difference between different zones. Figure 5 shows the classification results of daily average temperature.

In Fig. 5a, different colors or markers represent different climate zones or clusters. It can be seen from this subfigure that the adjacent stations are partitioned into the same cluster in most cases. By observing Figs. 5a and 5c, we find out the classification results are roughly consistent with five thermal climate clusters. Concretely speaking, most stations of Zone 1 are in the hot summer and warm winter region, and fewer stations are in the mild cluster. Zone 2 belongs to the cluster of hot summer and cold winter. Zone 3 involves two thermal climate clusters, the cold and the severe cold region. Zone 4 is part of the severe cold cluster. Zone 5 includes three thermal climate clusters, the severe cold, the cold, and the mild regions. Among these five zones, Zone 1 has the

highest average temperature and Zone 2 has the second highest temperature. Most stations of Zone 5 are located in the Qinghai–Tibet Plateau and the Yunnan–Guizhou Plateau; compared with Zone 4, they have lower temperature in summer because of their higher altitudes. In addition, the latitudes of Zone 3 lie between Zone 5 and Zone 4. Figure 5b constructs the neighbors graph of daily average temperature. There are 4074 solid lines and 290 dotted lines in total, and each node has 13.2 neighbors in the average sense.

2) AVERAGE RELATIVE HUMIDITY

The classification results of average relative humidity is shown in Fig. 6. It can be seen from this figure that all stations are approximately divided into five disjoint regions. Compared with the distribution of dry and wet areas in China, we have the following observations from Figs. 6a and 6c. Zone 1 is the wet region with the highest relative humidity. Zone 2 includes two regions, the semiwet and the semiarid. Zone 3 has the lowest relative humidity and it consists of two regions, the arid and the semiarid. Zone 4 mainly belongs to the wet and the semiwet. Zone 5 is mainly located at plateau and it has

FIG. 5. Classification of daily average temperature. (a) Cluster locations of five zones. (b) Neighbors graph. (c) Mean of each zone.

three regions, the wet, the semiwet, and the semiarid. In conclusion, the trend of average relative humidity in China is decreasing from the southeast to the northwest. Figure 6b shows the neighbors graph of average relative humidity. In this subfigure, there are totally 4017 solid lines from the same clusters and 336 dotted lines from different clusters. On the average, each station has 13.17 neighbors.

3) SUNSHINE HOURS

The sunshine hours of a location is a general indicator of cloudiness. Generally, the cloudiness has a strongly positive relationship with the relative humidity within a narrow latitude sector. Figure 7 gives the classification results of sunshine hours. We can see from Figs. 7a and 7c that the classification of sunshine hours is almost identical with that of average relative humidity except for Zone 1.

Zone 1 in Fig. 7 has more stations than that in Fig. 6. Similarly, Zone 1 has the shortest sunshine hours and Zone 2 has the second shortest sunshine hours; the stations of Zones 3 and 4 are located at high altitude and they have the longest sunshine hours. Figure 7b plots the neighbors graph of sunshine hours. In this subfigure, 3887 solid lines and 291 dotted lines are drawn to describe the neighbor relationships and each station has 12.64 neighbors on average.

4) DIURNAL TEMPERATURE RANGE

Diurnal temperature range is the difference between the maximum and the minimum temperature. Generally speaking, the lower relative humidity or the higher the altitude, the greater the diurnal temperature range. Based on this meteorological variable, the classification results

FIG. 6. Classification of average relative humidity. (a) Cluster locations of five zones. (b) Neighbors graph. (c) Mean of each zone.

are shown in Fig. 8. By observing Figs. 8a and 8c, we have the following conclusions. Zone 1 has the smallest temperature range, which is consistent with this zone also having the highest relative humidity. Zone 2 has the second smallest temperature range. Both Zone 3 and Zone 5 have large temperature ranges, which makes sense given that Zone 3 is the driest region and Zone 5 has high altitude. Figure 8b displays the neighbors graph of diurnal temperature range. There are 4128 solid lines and 322 dotted lines, and each station has 13.46 neighbors.

5) ATMOSPHERIC PRESSURE

Atmospheric pressure decreases with increasing altitude. Figure 9 gives the classification results of atmospheric pressure. According to Figs. 9a and 9c, we draw several

observations. Zone 5 has the lowest atmospheric pressure due to the fact the corresponding stations are located in the Yunnan–Guizhou Plateau and the Qinghai–Tibet Plateau. Zone 3 has the second lowest atmospheric pressure because of high altitude. The other three zones have low altitude, and they are distributed from south to north in turn. The mean atmospheric pressure of Zone 1 lies between Zone 2 and Zone 4. Furthermore, the mean atmospheric pressure of Zone 5 changes slightly and the other four zones have same change tendency, namely, the mean atmospheric pressure in winter is higher than in summer. The final classification result is in line with three steps of China's terrain: Zone 5 belongs to the first step, Zone 3 lies in the second step, and the third step includes the other three zones.

FIG. 7. Classification of sunshine hours. (a) Cluster locations of five zones. (b) Neighbors graph. (c) Mean of each zone.

The sparsity rate of the revised $\mathbf{W}$ (i.e., the element less than 0.2 is replaced by 0) is reduced to 0.9390 and the neighbor relationships are shown in Fig. 9b. The subfigure shows 11 667 solid lines and 1976 dotted lines, and each station has 40.91 neighbors on the average. The large neighbor number may result from the fact that the atmospheric pressure has the worst sparsity among five similarity matrices. Figure 9c also illustrates the separability of the mean atmospheric pressure.

d. Multiple-view climate classification

In section 3c, we study the climate classification with only one meteorological variable. However, in many practical applications, we need to provide a comprehensive climate classification by integrating all

interested meteorological variables. This subsection will present a composite classification method via multiple views.

We utilize Q meteorological variables to perform classification and obtain the similarity matrices by the proposed method in this paper. Let $\mathbf{W}_i$ be the similarity matrix of the i th meteorological variable, where $i = 1, 2, \dots, Q$. The integrated similarity matrix is defined by

$$\mathbf{W} = \sum_{i=1}^Q \gamma_i \mathbf{W}_i, \quad (9)$$

where $\gamma_i \geq 0$ is the weight of the i th meteorological variable and it holds that $\gamma_1 + \gamma_2 + \dots + \gamma_Q = 1$. Large weight means great importance, and the case $\gamma_i = 0$

FIG. 8. Classification of diurnal temperature range. (a) Cluster locations of five zones. (b) Neighbors graph. (c) Mean of each zone.

indicates that the i th meteorological variable is not considered in the classification procedure. The weights $\{\gamma_i\}_{i=1}^Q$ can be determined by the purpose of classification.

The following provides a method to determine automatically all weights. We first perform classification on the i th meteorological variable and all stations are partitioned into c clusters. Let $n_j^{(i)}$ be the number of stations from the j th cluster. For the i th meteorological variable, the imbalance degree of the above classification procedure is defined as

$$b^{(i)} = \left[\max_j n_j^{(i)} - \min_j n_j^{(i)} \right] / \max_j n_j^{(i)}. \quad (10)$$

A larger imbalance degree indicates a more imbalanced grouping. If there is no prior information, the case

that $n_1^{(i)} = n_2^{(i)} = \dots = n_c^{(i)}$ is expected, that is, $b^{(i)} = 0$. Large $b^{(i)}$ is propitious to classification. Hence, we set $\gamma_i = b^{(i)} / [b^{(1)} + b^{(2)} + \dots + b^{(Q)}]$.

According to the experimental results of section 3c, we list the number of stations in each zone for all meteorological variables, as shown in Table 1. The imbalance degree of each variable is also displayed in the last column of Table 1. From this table, we can see that the daily atmospheric pressure has the largest imbalance degree, the relative humidity has the second largest imbalance degree and the sunshine hours has the smallest imbalance degree. The above observations mean that both daily atmospheric pressure and relative humidity are of primary importance.

Spectral clustering is carried out on the synthetic similarity matrix. Figure 10 shows the integrated classification results, where all stations are divided into five

FIG. 9. Classification of atmospheric pressure. (a) Cluster locations of five zones. (b) Neighbors graph. (c) Mean of each zone.

zones. It can be seen from this figure that each zone has a reasonable layout and the classification result is the closest to that of the atmosphere pressure. The comprehensive zoning in Fig. 10a has larger difference with that of other alone meteorological variable. Figure 10b plots the neighbors graph, and there are 5089 lines within the same zones and 517 lines from different

zones. The numbers of stations from five zones are 112, 135, 199, 89, and 126 respectively. The imbalance degree of the above classification is 0.5528, which lies between the smallest and the largest values of five individual meteorological variables.

The consistency rate is employed to evaluate the compatibility of two classification methods. All stations are

TABLE 1. Number of stations in each zone and imbalance degree. The bold font emphasizes the largest value.

Meteorological variable	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Imbalance degree
Tem	97	164	177	108	115	0.4520
Hum	217	133	110	88	113	0.5945
SunHour	114	141	142	113	151	0.2517
TemRange	160	142	152	102	105	0.3625
Pressure	101	133	243	87	97	0.6420

FIG. 10. Climate classification by multiple views. (a) Cluster locations of five zones. (b) Neighbors graph.

divided into c zones $\{Z_1^{(i)}, Z_2^{(i)}, \dots, Z_c^{(i)}\}$ according to the i th classification method, where these zones are rearranged by some order. For a given set Z , let $\text{Card}(Z)$ be the cardinality of Z . The consistency rate (CR) between the i th and the j th classification methods is defined by

$$\text{CR}_{ij} = \frac{1}{N} \sum_{l=1}^c \text{Card}[Z_l^{(i)} \cap Z_l^{(j)}]. \quad (11)$$

The values of CR_{ij} lie in the interval $[0, 1]$. Large CR_{ij} means large consistency and $\text{CR}_{ij} = 1$ when the i th and the j th classification methods have identical classification results.

Table 2 shows the consistency rates between different classification methods. It can be seen from this table that the sunshine hours variable has large classification consistency with diurnal temperature range and average relative humidity among the first five classification variables. This conclusion can be explained as follows. Both diurnal temperature range and average relative humidity have large impact on the sunshine hours. A small diurnal temperature range and large average relative humidity possibly shorten the sunshine hours. In addition, the classification under multiple views improves generally the consistency rate and it has the largest consistency with classification result of the atmosphere pressure.

e. Determination of the number of clusters

Choosing the number c of clusters is a fundamental problem for all clustering tasks. In spectral clustering, the eigengap heuristic is a recommended tool designed for determining the number c (Von Luxburg 2007). Let $\{\lambda_1, \lambda_2, \dots, \lambda_N\}$ be N eigenvalues of the Laplacian matrix $\mathbf{L}$ and they are sorted in descending order (only the real part is considered for the complex eigenvalue). It is obvious that the largest eigenvalue $\lambda_1 = 1$. Inspired by matrix perturbation theory, we hope that the first c eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_c$ are relatively large, but λ_{c+1} is relatively small. For determining c , the i th eigengap is defined as $\delta_i = \lambda_i - \lambda_{i+1}$, $i = 1, 2, \dots, N - 1$. The relatively large eigengap provides us a reference for choosing the number c .

For simplicity, we synthesize a similarity matrix $\mathbf{W}$ through averaging the similarity matrices of five meteorological variables used in this paper. In the process of constructing similarity matrices, three numbers of neighbors are considered, namely, $k \in \{0, 10, 30\}$. For the Laplacian matrix corresponding to $\mathbf{W}$, we compute its first 30 eigengaps and plot them in Fig. 11. Without considering the first eigengap, δ_5 reaches the largest value, which means that the number c of clusters should be greater than or equal to 5. We suggest that the

TABLE 2. Consistency rates between different classification methods. The bold font emphasizes the largest and second largest values.

Consistency rate	Hum	SunHour	TemRange	Pressure	Multiple views
Tem	0.6021	0.6172	0.6717	0.6868	0.7231
Hum		0.7519	0.7927	0.6581	0.7156
SunHour			0.8366	0.6989	0.7352
TemRange				0.7126	0.7640
Pressure					0.9032

FIG. 11. Eigengaps for integrated Laplacian matrix.

number c of clusters should satisfy the following two conditions: $\delta_c \geq \delta_0$ and δ_c is a local maximum, where the threshold δ_0 is a given small positive value. If δ_0 is set to 0.02, then we can take $c \in \{5, 8, 11, 13, 16\}$ for $k = 0$, $c \in \{5, 7, 9, 11, 13, 15\}$ for $k = 10$, and $c \in \{5, 10, 15\}$ for $k = 30$. On the whole, the number c of clusters can be chosen within the range from 5 to 16.

f. Sensitivity analysis

1) REGULARIZATION PARAMETERS

This part will discuss the influence of regularization parameters on the experimental results. For the sake of simplicity, we might as well take daily average temperature as an example. When both α_1 and α_2 are relatively small, the optimal coefficient matrix $\mathbf{C}^*$ in problem (6) trends to the zero matrix. Conversely, the optimal noise matrix $\mathbf{E}^*$ or $\mathbf{Z}^*$ in problem (5) also trends to the zero matrix. Based on numerical experiments, we have the following observations. If $\alpha_1 = \alpha_2$ and $1 \leq \alpha_1 \leq 4$, some columns of $\mathbf{C}^*$ are zero vectors, which is in conflict with our motivation. If $\alpha_1 = \alpha_2 = 50$, each column of $\mathbf{C}^*$ has 8.07 nonzero elements in the average sense. If $\alpha_1 = \alpha_2 = 500$, each column of $\mathbf{C}^*$ has 38.72 nonzero elements on the average. Generally, the value of $\text{SR}(\mathbf{C}^*)$ decreases with the increasing of α_1 and α_2 . In summary, $\alpha_1 = \alpha_2 = 50$ is an appropriate choice.

Because the true classification result is unknown, we take the classification result for the case $\alpha_1 = \alpha_2 = 50$ as a comparison object and further study the consistency rates under different values of parameters α_1 and α_2 . Set $\alpha_1 = 10 + 20i$ and $\alpha_2 = 10 + 20j$, where $i, j = 0, 1, \dots, 25$. For all possible combinations of α_1 and α_2 , the corresponding consistency rates are shown in Fig. 12. From this figure, we have the following observations. For fixed α_1 , the consistency rates with different α_2 are relatively stable and the corresponding standard deviation is less

FIG. 12. Consistency rates under different α_1 and α_2 .

than 0.032. While the consistency rates show a decreasing trend with the increasing of α_1 for fixed α_2 . Furthermore, if $\alpha_1 \in [30, 330]$ and $\alpha_2 \in [10, 510]$, the smallest consistency rate is 0.9213, which means the classification results have good consistency. The above observations demonstrate that the classification results are not very sensitive to the choices of α_1 and α_2 to some extent.

2) TRADE-OFF PARAMETER

The trade-off parameter β controls the importance of balancing sparse subspace representation and the k nearest neighbor. For five given meteorological variables, we will discuss the impact of different β on classification results. Similarly, we take $\beta = 1/2$ as a reference object and explore the consistency rates for other values of β . Let $\{i/10\}_{i=0}^{10}$ be the set of values of β . The consistency rates are compared in Fig. 13. If β lies in the interval $[0.2, 0.8]$, the smallest consistency rates are daily average temperature (0.8835), average relative humidity (0.9198), sunshine hours (0.8729), diurnal temperature range (0.9592), and atmospheric pressure (0.8926). This observation means that the classification results have good consistency for the case $\beta \in [0.2, 0.8]$. Moreover, the consistency rate of $\beta = 0$ has large difference with that of $\beta = 1$, which shows that it is valuable to choose an appropriate parameter β .

3) NUMBER OF NEIGHBORS

The choice of neighbor numbers directly affects the similarity matrix $\mathbf{W}_{\text{knn}}$, which has a strong influence on the compromise matrix $\mathbf{W}$. This part adopts the same computation formulation for integrated similarity matrix as section 3d. In the following, we will compare the classification results for different values of k . Let $k = 5i$, where $i = 0, 1, 2, \dots, 10$. The climate classification is performed for each k and the consistency rate is employed to evaluate the consistency of classification results under two values of k . According to experimental

FIG. 13. Consistency rates under different β .

results, the consistency rate lies in the interval $[0.8215, 0.9955]$, which indicates that the value of k has a certain impact on classification. The consistency rate comparison is displayed in Fig. 14 as an image.

Let k_1 and k_2 be two numbers of neighbors. Three observations can be seen from Fig. 14. In general, the smaller the value of $|k_1 - k_2|$ is, the larger the consistency rate is. The case that $k_1 = 0$ and $k_2 = 40$ yields the worst consistency rate. When $|k_1 - k_2| = 5$, the difference between two corresponding consistency rates is less than or equal to 0.0439 in the absolute sense. These observations mean that the climate classification results are stable for minor changes in k .

4) NUMBER OF CLUSTERS

In section 3e, we discuss a determination method for the number of clusters. For the climate data used in this paper, it is proposed that the number c of clusters ranges from 5 to 16. This part will consider several other values of c , that is, $c \in \{7, 9, 11, 13\}$. For the above four values of c , spectral clustering is performed on the similarity matrix used in section 3d. The detailed classification results are shown in Fig. 15.

By observing Figs. 10a and 15a, we can see that Zones 1 and 4 in Fig. 10a have slight changes, but the remaining three zones are approximately divided into five zones when $c = 7$. We also have the following observations by further comparing four subfigures of Fig. 15. Zones 1 and 3 in Fig. 15a are roughly partitioned into two zones respectively when $c = 9$. Zones 1 and 7 in Fig. 15b are respectively split into two zones when $c = 11$. Zone 11 in Fig. 15c are segmented into Zone 1 and Zone 10 in Fig. 15d, and Zone 3 and Zone 6 are changed into three zones. Based on the above observations, we draw three conclusions: all stations are regrouped regularly with the increasing of the number of clusters, there is no serious imbalance in the number of each zone, and all stations

FIG. 14. Consistency rates under different k .

are divided into several disjoint regions except a few outliers.

4. Discussion

This section first conducts a further statistical analysis on the climate classification on the basis of daily average temperature, and then it discusses the classification consistency by providing another meteorological dataset.

a. Further statistics of classification results

Section 3c has compared the average results of each zone for five meteorological variables. This section will make a statistical analysis on other indicators. These include the mean temperature of the coldest month (January), the mean temperature of the warmest month (July), cooling degree-days based on 26°C (CDD26), and heating degree-days based on 18°C (HDD18). They are important classification elements in climate zoning for building thermal design.

For brevity, only the classification result of daily average temperature in section 3c is utilized. To count the aforementioned four indicators, the daily average temperatures for each station are averaged over 10 years and thus an annual average temperature time series is obtained. The boxplot is used to graphically depict groups of numerical data through their quartiles. The experimental results are displayed in Fig. 16. By observing the first two subfigures in Fig. 16, we find that Zone 1 has the highest average temperature in January and July, Zone 4 has the lowest average temperature in January, and Zone 5 has the lowest average temperature in July. There are in total three outliers in January and

FIG. 15. Classification under different numbers of clusters: (a) $c = 7$, (b) $c = 9$, (c) $c = 11$, and (d) $c = 13$.

21 outliers in July. It can be seen from the last two subfigures that the average degree-days in HDD18 gradually increase from Zone 1 to Zone 4, and CDD26 has the opposite changes. Zone 5 has a large change range than Zone 4 for HDD18, and the degree-days of these two zones are almost zeros for CDD26. Coincidentally, HDD18 and CDD26 have 3 outliers and 21 outliers respectively. To sum it up, the statistical analysis on these four indicators demonstrates the rationality and effectiveness of the classification result.

b. Comparison with another dataset

Due to data limitations, the previous work only considers the climate dataset covering a 10-yr period. In this subsection, we provide another dataset during the period from 1 January 2014 to 31 May 2017. This dataset is also from the National Climate Center of China. As for the aforementioned meteorological variables, atmospheric pressure is not included in the new dataset.

Moreover, this dataset involves 706 stations shown in Fig. 17 and there are 552 stations common to the two datasets. Compared with Fig. 1, the stations located at the regions of Zhejiang, Jiangxi, Guizhou, and Tibet are not taken into account. By adopting a 5-day period, each meteorological variable is represented by a 249-dimensional vector.

For fair comparison, we set the same parameters as in the previous section, namely, $c = 5$, $k = 10$, and $\alpha_1 = \alpha_2 = 50$. Under single view, the resulting classifications are shown in Fig. 18. For each meteorological variable in both datasets, we compare their classification consistency on 552 common stations, and the consistency rates are 0.6757, 0.7572, 0.8514, and 0.9275 respectively. The best consistency occurs at diurnal temperature range, which indicates that the two datasets have good consistency on classification. For other three variables, station distribution in Zone 5 is obviously influenced by the lack of stations in Tibet, which leads to a direct impact on the

FIG. 16. Comparison of four indicators on five zones: (a) January, (b) July, (c) HDD18, and (d) CDD26.

stations in Zone 1 and Zone 3. Especially, Zone 5 and Zone 2 are severely disrupted in terms of daily average temperature. In summary, the proposed classification method has good stability and obtains acceptable consistency between two consecutive periods except daily average temperature.

5. Conclusions

In this paper, we investigate the climate classifications by using a spectral clustering-based method. Daily climate observations of five variables are employed and they are from 661 stations in China during the period of 2004–13. To obtain a stable and reliable climate classification, the daily observations are averaged through 10-day period. For classification, a novel construction method of the similarity graph or matrix is proposed. The construction method combines the strengths of k -nearest-neighbor and sparse subspace representation.

In the implementation process of experiments, weak linear correlation is first validated by correlation analysis among five meteorological variables. Next,

classification of each individual meteorological variable is studied based on the constructed similarity matrix. Then the climate classification is performed by integrating all meteorological variables. Regardless of single-view or multiple-view classification, the proposed method has good consistency. Afterward, the

FIG. 17. Spatial distribution of 706 stations in China.

FIG. 18. Classification under four meteorological variables for the second dataset: (a) daily average temperature, (b) average relative humidity, (c) sunshine hours, and (d) diurnal temperature range.

classification of each individual meteorological variable is studied based on the constructed similarity matrix. Then the climate classification is performed by integrating all meteorological variables. Regardless of single-view or multiple-view classification, the proposed method has good consistency. Afterwards, the determination method of the number of clusters is explored. Finally, we research the influence of various parameters on classification and the results show the proposed method is not very sensitive to all parameters.

In the discussion section, we make a further statistical analysis on classification results and compare the classification consistency by supplementing another dataset. All experimental results validate the feasibility and effectiveness of the proposed climate classification method. This paper provides a mathematical model to climate zones, and also presents a valuable reference and verification on climate zones. In the future, we will

consider other meteorological variables in climate classification, such as monthly precipitation, solar radiation, and wind. Furthermore, it is worth studying the construction manner of similarity matrix by integrating all interested meteorological variables. We will also explore the application of the climate classification to the field of building.

Acknowledgments. This study was supported by “the 13th Five-Year” National Science and Technology Major Project of China (2018YFC0704500), the State Key Program of National Natural Science of China (51838011), and the China Postdoctoral Science Foundation (2017M613087).

REFERENCES

- ASHRAE, 2013: Climatic Data for Building Design Standards. ASHRAE Standard 169-2013, American Society of Heating, Refrigerating and Air-Conditioning Engineers, 98 pp.

- Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein, 2011: *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Foundations and Trends, 128 pp., <https://doi.org/10.1561/22000000016>.
- Briggs, R. S., R. G. Lucas, and Z. T. Taylor, 2003a: Climate classification for building energy codes and standards: Part 1—Development process. *ASHRAE Trans.*, **109**, 109–121.
- , —, and —, 2003b: Climate classification for building energy codes and standards: Part 2—Zone definitions, maps, and comparisons. *ASHRAE Trans.*, **109**, 122–130.
- Chen, G., and G. Lerman, 2009: Spectral curvature clustering (SCC). *Int. J. Comput. Vis.*, **81**, 317–330, <https://doi.org/10.1007/s11263-008-0178-9>.
- Cheng, B., J. Yang, S. Yan, Y. Fu, and T. S. Huang, 2010: Learning with ℓ^1 -graph for image analysis. *IEEE Trans. Image Process.*, **19**, 858–866, <https://doi.org/10.1109/TIP.2009.2038764>.
- Elhamifar, E., and R. Vidal, 2013: Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 2765–2781, <https://doi.org/10.1109/TPAMI.2013.57>.
- Filippone, M., F. Camastra, F. Masulli, and S. Rovetta, 2008: A survey of kernel and spectral methods for clustering. *Pattern Recognit.*, **41**, 176–190, <https://doi.org/10.1016/j.patcog.2007.05.018>.
- Gerstengarbe, F. W., P. C. Werner, and K. Fraedrich, 1999: Applying non-hierarchical cluster analysis algorithms to climate classification: Some problems and their solution. *Theor. Appl. Climatol.*, **64**, 143–150, <https://doi.org/10.1007/s007040050118>.
- Jain, A. K., 2010: Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.*, **31**, 651–666, <https://doi.org/10.1016/j.patrec.2009.09.011>.
- Lau, C. C., J. C. Lam, and L. Yang, 2007: Climate classification and passive solar design implications in China. *Energy Convers. Manage.*, **48**, 2006–2015, <https://doi.org/10.1016/j.enconman.2007.01.004>.
- Liu, G., Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, 2013: Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**, 171–184, <https://doi.org/10.1109/TPAMI.2012.88>.
- Lund, R., and B. Li, 2009: Revisiting climate region definitions via clustering. *J. Climate*, **22**, 1787–1800, <https://doi.org/10.1175/2008JCLI2455.1>.
- Mallenahalli, N. K., 2015: Predicting climate variability over the Indian region using data mining strategies. 8 pp., <https://arxiv.org/abs/1509.06920>.
- Mimmack, G. M., S. J. Mason, and J. S. Galpin, 2001: Choice of distance matrices in cluster analysis: Defining regions. *J. Climate*, **14**, 2790–2797, [https://doi.org/10.1175/1520-0442\(2001\)014<2790:CODMIC>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<2790:CODMIC>2.0.CO;2).
- Ministry of Construction of China, 1993: Thermal design code for civil building (GB50176-93) (in Chinese). China Architecture and Building Press.
- Netzel, P., and T. Stepinski, 2016: On using a clustering approach for global climate classification. *J. Climate*, **29**, 3387–3401, <https://doi.org/10.1175/JCLI-D-15-0640.1>.
- Ng, A. Y., M. I. Jordan, and Y. Weiss, 2002: On spectral clustering: Analysis and an algorithm. *NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, MIT Press, 849–856.
- Olgyay, V., 2015: *Design with Climate: Bioclimatic Approach to Architectural Regionalism*. Princeton University Press, 224 pp.
- Steinhaeuser, K., N. V. Chawla, and A. R. Ganguly, 2010: An exploration of climate data using complex networks. *ACM SIGKDD Explorations Newsletter*, No. 12, Association for Computing Machinery, New York, NY, 25–32, <https://doi.org/10.1145/1882471.1882476>.
- Unal, Y., T. Kindap, and M. Karaca, 2003: Redefining the climate zones of Turkey using cluster analysis. *Int. J. Climatol.*, **23**, 1045–1055, <https://doi.org/10.1002/joc.910>.
- Verichev, K., M. Zamorano, and M. Carpio, 2019: Assessing the applicability of various climatic zoning methods for building construction: Case study from the extreme southern part of Chile. *Build. Environ.*, **160**, 106165, <https://doi.org/10.1016/j.buildenv.2019.106165>.
- Vidal, R., and P. Favaro, 2014: Low rank subspace clustering (LRSC). *Pattern Recognit. Lett.*, **43**, 47–61, <https://doi.org/10.1016/j.patrec.2013.08.006>.
- Von Luxburg, U., 2007: A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416, <https://doi.org/10.1007/s11222-007-9033-z>.
- Walsh, A., D. Cóstola, and L. C. Labaki, 2017: Comparison of three climatic zoning methodologies for building energy efficiency applications. *Energy Build.*, **146**, 111–121, <https://doi.org/10.1016/j.enbuild.2017.04.044>.
- Wan, K. K., D. H. Li, L. Yang, and J. C. Lam, 2010: Climate classifications and building energy use implications in China. *Energy Build.*, **42**, 1463–1471, <https://doi.org/10.1016/j.enbuild.2010.03.016>.
- Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, 2009: Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **31**, 210–227, <https://doi.org/10.1109/TPAMI.2008.79>.
- Xiong, J., R. Yao, S. Grimmond, Q. Zhang, and B. Li, 2019: A hierarchical climatic zoning method for energy efficient building design applied in the region with diverse climate characteristics. *Energy Build.*, **186**, 355–367, <https://doi.org/10.1016/j.enbuild.2019.01.005>.
- Zhang, X., and X. Yan, 2014: Temporal change of climate zones in China in the context of climate warming. *Theor. Appl. Climatol.*, **115**, 167–175, <https://doi.org/10.1007/s00704-013-0887-z>.
- Zheng, J., Y. Yin, and B. Li, 2010: A new scheme for climate regionalization in China (in Chinese). *Acta Geogr. Sin.*, **65**, 3–13.