Journal of Hydrology 597 (2021) 126146

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Using *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data

Honghua Liu [a], Jing Yang [b,c,*], Ming Ye [b,*], Scott C. James [d], Zhonghua Tang [c], Jie Dong [a], Tongju Xing [a]

[a] Qingdao Geo-Engineering Surveying Institute (Qingdao Geological Exploration and Development Bureau) and Key Laboratory of Urban Geology and Underground Space Resources, Shandong Provincial Bureau of Geology and Mineral Resources, Qingdao 266100, China
[b] Department of Earth, Ocean, and Atmosphere Science, Florida State University, Tallahassee, FL 32306, USA
[c] School of Environmental Studies, China University of Geosciences, Wuhan 430074, China
[d] Departments of Geosciences and Mechanical Engineering, Baylor University, Waco, TX 76798, USA

ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Saket Pande, Associate Editor

Keywords:
Statistical analysis
Geochemical zones
Dimension reduction
Cluster validity
Principal component analysis

ABSTRACT

Cluster analysis is a valuable tool for understanding spatial and temporal patterns (e.g., spatial zones) of groundwater geochemistry. To determine cluster numbers and cluster memberships that are unknown in real-world problems, a number of methods have been used to assist cluster analysis, among which graphic approaches are popular and intuitive. This study introduced, for the first time, the *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) method as a graphic approach to assist cluster analysis for groundwater geochemistry data. The hierarchical cluster analysis (HCA) was applied to original groundwater geochemistry data, and *t*-SNE was used to help determine the number of cluster and cluster memberships. Afterward, *t*-SNE was used to help delineate spatial zones of groundwater geochemistry. The *t*-SNE-based cluster visualization was compared to the visualization based on principal component analysis (PCA). By applying HCA, PCA, and *t*-SNE to three geochemical datasets (Oslo transect, Taiyuan karst water, and Jianghan Plain groundwater datasets, which are characterized by different number of samples and features collected across different space and time scales), we found that *t*-SNE outperformed PCA to assist HCA as a promising tool for helping determine the number of HCA clusters and delineate spatial zones of groundwater geochemistry. It should be noted that *t*-SNE alone cannot be used for cluster analyses, partly because *t*-SNE visualization depends on a hyperparameter called perplexity that is *a priori* unknown for real-world problems. The perplexity values used in this study were determined empirically, and a small value of 0.1 was used for the Taiyuan karst water dataset with 14 samples. For the other two datasets with hundreds of samples, the corresponding perplexity values were 20 and 30, within the range of 5 – 50 commonly used in *t*-SNE.

1. Introduction

Groundwater resources may be effectively and efficiently protected if spatial and temporal patterns of groundwater geochemistry are understood, especially in light of variable future conditions that will change according to climate, population expansion, and decreasing freshwater availability (Cloutier et al., 2008; Fendorf et al., 2010; Gorelick and Zheng, 2015; Green et al., 2011; Zhu et al., 2020). Groundwater geochemistry is affected by both natural processes and anthropogenic activities with broad variability in space and time especially at the regional scale (Güler and Thyne, 2004b; Haile and Fryar, 2017; Tóth, 1999, 2009). Discovering spatio-temporal patterns and further delineating spatial zones of groundwater geochemistry is a primary goal of hydrogeochemical studies (Güler et al., 2012; Nguyen et al., 2015; Yang et al., 2020). This goal is complicated because groundwater geochemical datasets are multivariate in nature; one groundwater sample has multiple physical, chemical, and biological features while multiple samples are collected at various locations and times. To understand the spatio-temporal patterns embedded in a groundwater geochemistry dataset, cluster analyses have been used to separate the dataset into a number of clusters, each of which has similar groundwater geochemistry that reflects the controlling processes of groundwater quality (Gan et al., 2018;

* Corresponding authors.
E-mail addresses: jingyang@cug.edu.cn (J. Yang), mye@fsu.edu (M. Ye), sc_james@baylor.edu (S.C. James), zhhtang@cug.edu.cn (Z. Tang).

https://doi.org/10.1016/j.jhydrol.2021.126146
Received 24 December 2020; Received in revised form 23 February 2021; Accepted 24 February 2021
Available online 2 March 2021

Güler and Thyne, 2004a,b; Pacheco Castro et al., 2018; Pant et al., 2018; Templ et al., 2008). Cluster analysis addresses two fundamental questions: (1) how many clusters are appropriate (i.e., cluster numbers)? and (2) how are individual groundwater samples assigned to a cluster (i.e., cluster memberships)? Because neither the number of clusters nor cluster memberships are known *a priori*, cluster analyses must be evaluated. Evaluation can be as simple as mapping the locations of clusters or by applying statistical techniques such as principal component analysis (PCA). For the convenience of visualizing PCA results in two-dimensional graphs, high-dimensional geochemical data are always projected onto two principal components (e.g., PC1 vs. PC2 where PC standing for principal component). It has been found that PCA plots are not always successful at distinguishing clustered data, and overlaps between clusters are constantly observed, making clear interpretation challenging. In the review article by Templ et al. (2008) wherein cluster analysis was applied to regional geochemical data, a question arose: "*Is there a graphical way to evaluate the stability or validity of clusters*?" This study was motivated to answer this question.

This study introduced the *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) method as a new graphical technique to support cluster analysis. The *t*-SNE method, developed by van der Maaten and Hinton (2008), is a state-of-the-art machine learning technique for dimensionality reduction to visualize high-dimensional data. The method has been used in many research fields, such as gene expression (Aizarani et al., 2019; Dobie et al., 2019; Kobak and Berens, 2019), tumor classification (Abdelmoula et al., 2016; Roche et al., 2018), hyperspectral imaging analysis (Melit Devassy and George, 2020; Pouyet et al., 2018), and fault diagnosis (Zheng et al., 2018; Zheng and Zhao, 2020). Recently, there have been several studies using *t*-SNE for visualization of geochemical and hydrological data. Balamurali and Melkumyan (2016) used *t*-SNE for dimension reduction, and found that *t*-SNE outperformed other dimension reduction methods (e.g., PCA, kernel PCA, and locally linear embedding) for visualizing patterns of shape/trend embedded in large geological assay datasets. Balamurali et al. (2019) further applied different clustering algorithms to reduced dimensions produced by *t*-SNE, such as self-organizing map and spanning-tree progression analysis of density-normalized events. Leung et al. (2019) conducted a similar research to apply the spectral clustering algorithm to *t*-SNE-reduced dimensions, and they further linked clustering results to a consensus matrix obtained from multiple *t*-SNE runs and spectral clustering classifications for detecting outliers in a geochemical dataset. Horrocks et al. (2019) used random forest methods to select 11 out of 31 elements for separating unaltered and altered host rock specimens, and then used *t*-SNE to validate the selection of the 11 elements. Mazher (2020) applied *t*-SNE to a large dataset generated by a hydrologic model that produces 9 variables over 273 time steps. The high dimensions of $2{,}457 = 9 \times 273$ was reduced to two dimensions using four methods, i.e., PCA, generative topographic mapping, *t*-SNE, and uniform manifold approximation and projection. Mazher (2020) concluded that the latter two methods outperformed the former two methods for visualizing spatial patterns of the modeling results. To the best of our knowledge, *t*-SNE has not been used to assist cluster analysis for groundwater geochemistry data.

This study tackled the following two questions that have not been attempted by the groundwater hydrology community:

(1) Can *t*-SNE be used as a graphical method to assist cluster analyses with respect to determining the number of clusters?
(2) Can *t*-SNE be used as a tool to delineate spatial zones of groundwater geochemistry based on clustering results?

In response to the first question, we used *t*-SNE in conjunction with hierarchical cluster analyses (HCA). In HCA, one way to determine the number of clusters is to subjectively place the so-called phenon line at a linkage distance, which is discussed in Section 2. Adjusting the phenon line upward or downward changes the number of clusters. Determining the appropriate number of clusters requires additional lines of evidence such as geochemical analysis using Piper or Stiff plots (Appelo and Postma, 2005; Yang et al., 2020). The appropriate number of clusters can also be determined by using various statistical methods, and the Elbow method, average Silhouette method, and the Gap statistic method (Kassambara, 2017) explored in this study. This study focused on the PCA method, which produces two-dimensional visualization of the clustered data. PCA visualizations, however, are often not ideal, because data of one cluster may significantly overlap data of another cluster. In this work, we illustrated that *t*-SNE outperformed PCA when visualizing clustered geochemical data in the two-dimensional *t*-SNE visualization.

To answer the second question about delineating spatial zones of groundwater geochemistry based on clustering results, the conventional approach is to first plot cluster data on a map and then examine the resulting spatial distribution of the clusters to determine spatial zones while considering the geological, hydrogeological, and geochemical information relevant to the site of interest. A drawback of this approach is that, when a series of samples are collected over time at a sampling location, it is impossible to plot the data series on the map, and one has to either incorporate the statistics of the data or discard certain data samples. For example, in the study of Yang et al. (2020), because groundwater samples collected at one well belonged to different clusters in different years, the clusters of the majority samples were used to delineate spatial zones. This problem is intrinsically resolved when using *t*-SNE, because it reduces high dimensional data to low dimensions and can use all data for delineating spatial zones (without discarding any data). An example of doing so was given by Mazher (2020), who used *t*-SNE to visualize a dataset comprising nine variables simulated at 273 time steps on a two-dimensional plot. More importantly, when *t*-SNE visualizes high-dimensional data on a reduced set of dimensions, pairwise distances and structures in the high-dimensional data space are maintained to the extent possible in the low-dimensional *t*-SNE space. Therefore, *t*-SNE, by default, is suitable for delineating spatial zones of groundwater geochemistry, a feature that has not received adequate attention.

To explore the two questions discussed above, *t*-SNE was applied to three geochemical datasets (the Oslo transect, the Taiyuan karst water, and the Jianghan Plain groundwater datasets) with different sample sizes and dimensions (i.e., the number of geochemical features). The Oslo transect dataset included geochemical data of nine plant materials (e.g., different species or leaves, wood, bark of birch and spruce) collected at 40 sites (Reimann et al., 2007). Because it is theoretically known that the dataset can be divided into nine clusters (each for one plant material), this dataset was used to benchmark *t*-SNE's capability of assisting HCA. It should be noted that this dataset is not suitable for spatial zone delineation, because the sampling sites were along a transect and the nine samples corresponding to nine plant materials were collected from the same sites. The Taiyuan karst water dataset of Ma et al. (2011) was used not only to assist HCA but also to evaluate *t*-SNE's ability to delineate spatial zones of groundwater geochemistry. The evaluation was possible because Ma et al. (2011) divided the groundwater system into three sub-systems and further delineated three groundwater geochemistry zones (e.g., a recharge and flow-through zone, a cold-water discharge zone, and a thermal-water discharge zone) for each subsystem. The Jianghan Plain groundwater dataset of Yang et al. (2020) was used in the same manner as the Taiyuan karst water dataset, except that the former dataset is substantially more complicated than the latter dataset in terms of dimensionality and spatial and temporal scales over which the data were collected. The Jianghan Plain groundwater dataset can better evaluate *t*-SNE's capability of both assisting cluster analysis and spatial zone delineation for regional aquifers.

In this study, HCA, PCA, and *t*-SNE analyses were applied to the three geochemical datasets. HCA was used as the basis for the cluster analysis, and PCA and *t*-SNE were used as graphic means of evaluating the HCA-determined number of clusters and cluster memberships. For this purpose, *t*-SNE outperformed PCA for all three datasets because *t*-SNE

provides better visualizations of clustered data. We demonstrated that *t*-SNE is an effective and efficient graphical way to assist cluster analysis with respect to determining the number of clusters and cluster memberships; it is also a promising tool for delineating spatial zones of groundwater geochemistry based on clustering results. However, since *t*-SNE was not designed for cluster analysis, we suggest using *t*-SNE only as a graphical way to assist HCA-based cluster analyses. *t*-SNE visualization strongly depends on a hyperparameter called perplexity that is *a priori* unknowable for real-world problems.

2. Geochemistry data and statistical methodologies

2.1. Three geochemical datasets

The Oslo transect dataset includes 360 samples of nine different plant materials collected at 40 sites along a 120-km transect crossing Oslo, Norway (Templ et al., 2008). The nine plant materials are terrestrial moss (MOS), fern (FER), European mountain ash leaves (ROG), birch leaves (BIL), bark (BBA) and wood (BWO) and spruce needles (SNE), twigs (TWI), and wood (STW). Details of the samples, element concentrations of the sample, and quality control of the concentrations are available, see Reimann et al. (2007). The dataset used in this analysis was downloaded from the *R* package "`rrcov`" developed by Todorov and Filzmoser (2010), available at https://cran.r-project.org/web/packages/rrcov/index.html (accessed 2/20/2021). The dataset includes concentrations of 24 elements (Ag, As, B, Ba, Ca, Cd, Co, Cr, Cu, Fe, Hg, K, La, Mg, Mn, Mo, Ni, P, Pb, S, Sb, Sr, Ti, and Zn) and loss on ignition for 350 samples (concentrations of ten samples are missing in this dataset). The Oslo transect dataset comprises 350 × 25 = 8,750 measurements. It is known theoretically that these data can be separated into nine clusters, one for each plant materials, because different plants uptake nutrients in different ways and variously partition those elements between wood, leaves, and bark (Templ et al., 2008). This "logical result" of nine clusters makes the dataset suitable as a benchmark to evaluate the potential of using *t*-SNE to evaluate cluster analyses.

The Taiyuan karst water dataset (Ma et al., 2001) consists of 37 samples of cold (water temperature < 30℃) and thermal karst groundwater from Taiyuan City, China, and the locations of the samples are shown in Fig. 1. For each sample, a total of 31 geochemical parameters were analyzed, including 3 physiochemistry variables (temperature, pH, and EC), nine major elements (CO_3^{2-}, HCO_3^-, F^-, Cl^-, NO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, and Na^+), 4 minor elements (K^+, Fe, Si, and Sr), and 15 trace elements (As, Ag, Al, B, Ba, Cd, Co, Cu, Hg, Li, Mn, Mo, Ni, Sb, and Zn). Therefore, the dataset has a total of 37 × 31 = 1,147 measurements. Based on site-specific information related to structural geology, hydrogeology, and hydrogeochemistry, the karst groundwater system was divided into three sub-systems as follows: the Dongshan Mountain karst groundwater subsystem (DMK), the Beishan Mountain karst groundwater subsystem (BMK), and the Xishan Mountain karst groundwater subsystem (XMK). Ma et al. (2011) further grouped samples of DMK into three zones (recharge and flow-through zone, cold-water discharge zone, and thermal-water discharge zone), samples of BMK into three zones (recharge and flow-through zone, cold-water discharge zone at the margin of the mountain, and cold-water discharge zone in buried karst zone), and samples from XMK into three zones (recharge and flow-through zone, cold-water discharge zone, and thermal-water discharge zone). Generally speaking, the groundwater geochemistry evolved from the recharge and flow-through zone toward the cold-water discharge zones and further to the thermal-water discharge zones. In Fig. 1, the samples collected from the DMK, BMK, and XMK subsystems are denoted by squares, circles, and triangles, respectively. Within each subsystem, the samples in the recharge

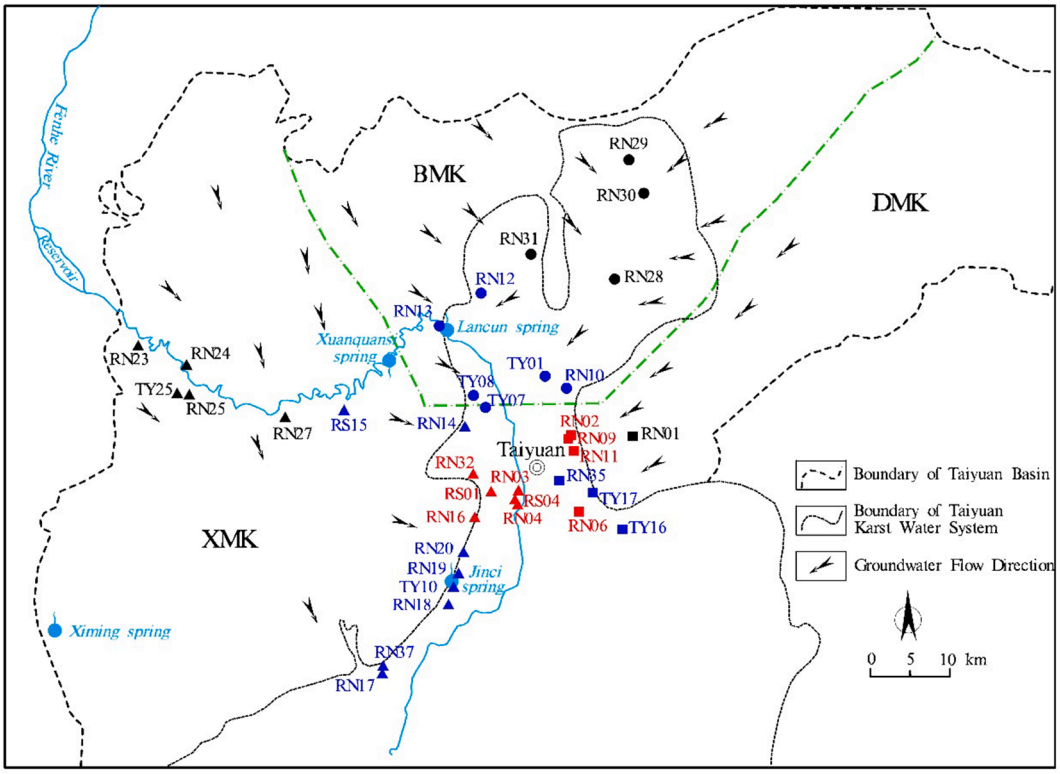

Fig. 1. Taiyuan karst water dataset sampling locations. The green dashed segments are the boundaries between the Dongshan Mountain (DMK), Beishan Maintain (BMK), and Xishan Maintain (XMK) karst water subsystems. Squares denote samples collected from DMK, circles from BMK, and triangles from XMK. In each subsystem, samples from the recharge and flow-through zone, cold-water discharge zone, and thermal-water discharge zone are indicated with black, blue, and red labels, respectively. This figure was modified from Ma et al. (2011). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and flow-through zone, cold-water discharge zone, and thermal-water discharge zone are highlighted in black, blue, and red colors, respectively. The dataset and geochemical analysis of Ma et al. (2011) facilitate an evaluation of the potential for *t*-SNE to assist cluster analysis for hundreds of groundwater geochemical data and to delineate groundwater geochemical zones at the scale of tens of kilometers.

The Jianghan Plain groundwater dataset, used in our previous study of Yang et al. (2020), includes 1,184 groundwater samples collected over 23 years (1992–2014) from 29 monitoring wells drilled into the middle-confined aquifer of Jianghan Plain, China (Fig. 2). Each sample has 11 geochemical parameters (pH, Ca^{2+}, Mg^{2+}, K^{+}, Na^{+}, Cl^{-}, SO_4^{2-}, HCO_3^{-}, NH_4^{+}, F^{-}, and Fe), and the dataset has a total of 1,184 × 11 = 13,024 measurements. Yang et al. (2020) conducted cluster and hydrogeochemical analyses for this large dataset, and delineated seven clusters and four groundwater geochemical zones of the regional aquifer, i.e., recharge (Zone I), transition (Zone II), flow-through (Zone III), and discharge-mixing zones (Zone IV) (Fig. 2). This large dataset and the four delineated zones enable us to evaluate the effectives of *t*-SNE in assisting cluster analysis for tens of thousands of groundwater geochemical data and in delineating groundwater geochemical zones at the scale of hundreds of kilometers.

2.2. Data preprocessing

Preprocessing geochemical data before conducting cluster analysis is always necessary (Templ et al., 2008; Ellefsen et al., 2014). The three datasets were pre-processed by: (1) geochemical feature selection, (2) substitution of censored data, (3) screening missing values, (4) data transformation, and (5) standardization. The first step defines appropriate parameters for cluster analysis. A geochemical parameter may be excluded if one or more of the following conditions is met: (1) the parameter is not continuously measured over time, i.e., the parameter values are not reported for many sampling campaigns, resulting in a large number of missing values, (2) the parameter is closely related to other parameters (e.g., alkalinity can be computed by using pH and HCO_3^{-} according to Appelo and Postma (2005)), and (3) >40% of measurements are censored data that are either "less than" or "greater than" a detection limit (Sanford et al., 1993). If censored data exist for a selected geochemical parameter, further data preprocessing is required. This study used a simple approach to replace the censored values with 3/4 of the detection limit for less-than conditions and 4/3 of the upper limit for greater-than conditions (Sanford et al., 1993). For the issue of missing values, samples that contain missing values are removed. Afterward, the natural-log transformation was applied to the datasets, and the *z*-score standardization (subtracting the data mean and dividing the residuals by data standard deviation) was applied to the transformed data to remove the impacts of data units and scales (Reimann and Filzmoser, 2000; Templ et al., 2008). It however should be noted that the use of substitution method and log transformation is simple but may not be ideal, as they may skew the distribution of the data and hamper the statistical analysis (Reimann and Filzmoser, 2000; Sanford et al., 1993).

2.3. HCA and PCA statistical analyses

The HCA and PCA statistical methods are described briefly here, and a more thorough description is available in statistical books (e.g., Rencher, 2003). HCA with the Ward method (Ward, 1963), as an agglomerative approach, starts by treating each sample as its own cluster, and merges the clusters stepwise to generate larger clusters, ending with one cluster containing all samples. At each successive step, clusters are merged according to the Ward criterion with the smallest increase of Sum of Squared Errors (*SSE*). For cluster A_l with n_l observations:

$$SSE_{A_l} = \sum_{i=1}^{n_l} \left\| O_i - \overline{O} \right\|^2,$$

where O_i ($i = 1, 2 \ldots, n_l$) is the i^{th} observation in the cluster, $\overline{O}$ is the mean of all observations in the cluster, and $||O_i - \overline{O}||^2$ is the squared Euclidean distance between O_i and $\overline{O}$. At the beginning of HCA, there are n clusters for n samples, and *SSE* of each cluster is zero. In the next step, all possible cluster combinations are considered, and *SSE* is calculated for each

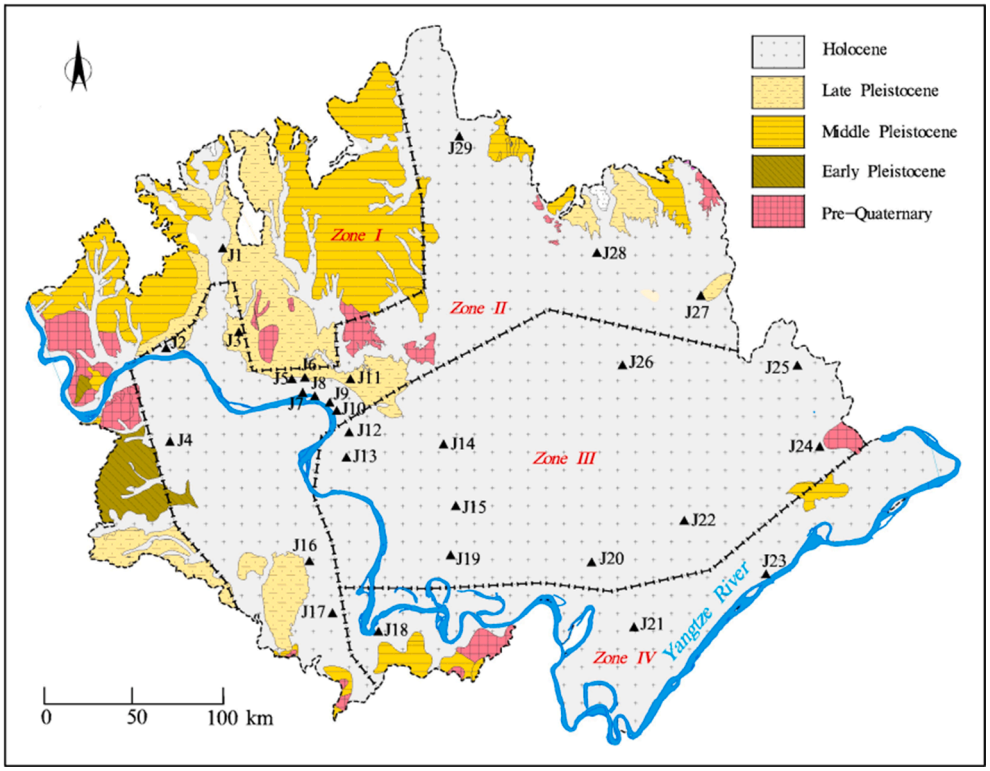

Fig. 2. Locations of 29 monitoring wells and four geochemical zones in the regional aquifer of the Jianghan Plain. The black dashed segments are the approximate boundaries of the Zones I – IV: the recharge, transition, flow-through, and discharge-mixing zones, respectively. The background shows the spatial distribution of outcrops of Holocene, Late Pleistocene, Middle Pleistocene, Early Pleistocene, and Pre-Quaternary sediments at the land surface. This figure was modified from Yang et al. (2020).

combination. For the case that clusters A and B are merged into cluster C, the SSE changes to:

$$\Delta SSE = SSE_C - SSE_A - SSE_B.$$

In the Ward method, the two clusters that yield the smallest ΔSSE are merged. The merging continues until only one cluster remains, when the variance of clusters is minimized in such a way. The merging history is recorded in a dendrogram, whose horizontal axis is for all the samples and the vertical axis shows the linkage distances, defined as $\sqrt{2\Delta SSE}$, between the merged clusters. The number of clusters is determined by placing the phenon line at a linkage distance, and one needs to carefully evaluate whether the number of clusters is appropriate by using various approaches, among which PCA is a popular one.

PCA is a dimension-reduction technique that performs a linear mapping of high-dimensional space to a lower-dimensional space, with the variance of the low-dimensional data maximized. PCA reduces data dimensionality, and allows focusing on a few new combinatorial components that describe a large portion of the variance in the data (Ouyang, 2005). In PCA, the covariance matrix of the high-dimensional variables is evaluated, and then the eigenvectors of the covariance matrix corresponding to the largest eigenvalues (the principal components) are used to reconstruct a significant fraction of the variance of the high-dimensional data. In the context of using PCA to assist cluster analysis, the first two principal components (or any two selected principal components) can be used to generate a two-dimensional graph to visualize the high-dimensional data and to examine whether the assigned cluster number is reasonable. Generally speaking, data in one cluster should be close to one another, but are maximally separated from data in any other clusters on the two-dimensional plot.

2.4. SNE and t-SNE methods

The t-SNE algorithm is an improved variation of stochastic neighbor embedding (SNE) developed by Hinton and Roweis (2002). The first task of SNE is to convert the distance between two points in a high-dimensional space to a conditional probability that represents the similarity of the two points in the high-dimensional space, and then to match the conditional probability between two points (data points) in the high-dimensional space to the conditional probability between two points (map points) in a low-dimensional space. The conditional probability, $p_{j|i}$, between data points, x_i and x_j, is the probability that x_i would pick x_j as its neighbor, and $p_{i|i} = 0$ by definition. The conditional probability is defined using a Gaussian kernel:

$$p_{j|i} = \frac{exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k\neq i} exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)},$$

where $\|x_i - x_j\|$ is the Euclidean distance between data points x_i and x_j. The variance, σ_i^2, of the Gaussian kernel is calculated using a binary search such that the entropy, $H(P_i) = -\sum_j p_{j|i} log_2 p_{j|i}$, of the probability distribution over all the data points is equal to $\log_2(Perp)$, where $Perp$ is the user-specified perplexity.

In the low-dimensional space, SNE computes a similar conditional probability for map points y_i and y_j (corresponding to the two data points, x_i and x_j) using another Gaussian kernel with variance equaling 1/2:

$$q_{j|i} = \frac{exp\left(-\|y_i - y_j\|^2\right)}{\sum_{i\neq k} exp\left(-\|y_i - y_k\|^2\right)},$$

and $q_{i|i} = 0$ by definition. The conditional probabilities, $p_{j|i}$ and $q_{j|i}$, should be equal, if the map points, y_i and y_j, exactly represent the similarity between the high-dimensional data points, x_i and x_j. SNE thus arranges map points in the low-dimensional space to minimize the discrepancy between $p_{j|i}$ and $q_{j|i}$ measured by the Kullback-Leibler divergence considering all data points. The cost function to be minimized is:

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}},$$

where P_i is the conditional probability distribution of data point x_i over all other data points, and Q_i is the conditional probability distribution of map point y_i over all other map points. The cost function is minimized through various optimization methods such as gradient descent.

SNE was further developed into t-SNE by van der Maaten and Hinton (2008) with two major improvements. One was to use a symmetric version of SNE to estimate pairwise similarities in the both high- and low-dimensional spaces. For data points x_i and x_j, t-SNE introduces:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n},$$

as the probability that x_i would pick x_j as its neighbor such that $p_{ij} = p_{ji}$ (the symmetric property), where n is the number of data points. The other improvement was to use a Student's t-distribution rather than a Gaussian kernel to compute the similarity between map points, so that the map points are more scattered in low-dimensional space. Strictly speaking, for map points y_i and y_j in low-dimensional space, t-SNE uses a heavy tailed t-distribution with one degree of freedom to compute:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1}},$$

which is the probability that y_i would pick y_j as its neighbor. Again, $q_{ij} = q_{ji}$ as a symmetric property. Correspondingly, the cost function becomes:

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}},$$

where P and Q are the joint probability distributions in high- and low-dimensional spaces, respectively.

Perplexity is the most important hyperparameter in the t-SNE method (there are other important hyperparameters such as learning rate and number of iterations used in gradient descent optimization). A small perplexity corresponds to a small σ_i^2 used in Eq. (3), which results in selecting a data pair, x_i and x_j, with small distance. A large perplexity corresponds to a large σ_i^2, which can pair data at large distances. In the extreme case of $Perp = +\infty$, x_i would pick any point as its neighbor with the equal possibility of $1/(n-1)$, with n being the number of data points. A commonly used perplexity is 30, and a typical range is 5–50. van der Maaten and Hinton (2008) argued that "*performance of SNE is fairly robust to changes in the perplexity*." This study explored the effects of perplexity on t-SNE performance for the three geochemical datasets, and found that a smaller perplexity should be used for small sample sizes.

With unknown perplexity values, t-SNE was used in this study as a visualization tool to assist HCA, and this is empirically described below:

Step 1: Conduct HCA to determine the number of clusters and cluster memberships based on the best understanding of the problem of interest and/or appropriate statistical methods.
Step 2: Assign different colors to different clusters (e.g., red for cluster 1 and blue for cluster 2); the colors will be used in Step 4 below.
Step 3: Conduct t-SNE with different perplexity values. The t-SNE runs are independent of the HCA in Step 1.
Step 4: For each perplexity value, plot the samples in a two-dimensional t-SNE space with samples highlighted in the colors determined in Step 1. For example, if a sample belongs to cluster 1, it will be plotted in red.

Step 5: Examine the spatial patterns of the colored samples in the *t*-SNE plots to determine which perplexity value yields the best results according to the following two criteria: (a) whether samples of the same cluster are close to each other, and (b) whether the samples of different clusters are maximally separated.

Step 6: If the visual examination in Step 5 yields satisfactory results for one perplexity value (likely for multiple perplexity values), it confirms that the cluster number and membership determined in Step 1 are reasonable. Otherwise, it is necessary to adjust the cluster number and memberships in Step 1 and to repeat Steps 2, 4, and 5. There is no need to repeat the *t*-SNE runs in Step 3.

This procedure indicates that *t*-SNE is not use as a stand-alone algorithm for cluster analysis but as a visualization tool to assist cluster analysis for determining the number of clusters and cluster memberships.

All the statistical analyses were performed using Python 3.7, and HCA, PCA, and *t*-SNE were implemented using the Python libraries Scikit-Learn (Pedregosa et al., 2011, https://scikit-learn.org/stable/index.html) and SciPy (Jones et al., 2001, https://www.scipy.org/). The codes and the datasets used in this study are available at https://github.com/jyangfsu/geochemical-t-SNE.

3. Results and discussion

3.1. Cluster analysis for Oslo transect dataset

Because it is known theoretically that the Oslo transect dataset should be separated into nine clusters (one for each plant material), this information was used to evaluate the potential of using *t*-SNE to assist cluster analysis. Delineation of spatial zones of geochemistry was not attempted for this dataset, because the sampling sites were along a transect and the nine samples corresponding to nine plant materials were collected from the same sites. Given that data preprocessing was completed by Reimann et al. (2007) and Todorov and Filzmoser (2010), the dataset used by Todorov and Filzmoser (2010) was used directly in this study.

Fig. 3(a) illustrates the two-dimensional *t*-SNE visualization of the Oslo transect dataset with $Perp = 20$, which yields the best visualization. The *t*-SNE plot shows that the nine plant material clusters are distinct and without overlap. This separation is consistent with the HCA results. Fig. 3(c) plots the dendrogram of the HCA using the Ward method with Euclidian distance. Placing the phenon line at a linkage distance of 18 results in nine clusters corresponding almost completely to the nine plant materials, indicating that *t*-SNE provides a graphic way to evaluate validity of clusters. PCA does not yield such a consistent graphic validation as shown in Fig. 3(b), which is the two-dimensional PCA visualization using the same color scheme of Fig. 3(a). Fig. 3(b) shows substantial overlaps between the clusters, indicating that PCA is not as adept as *t*-SNE for graphically validating the nine HCA-based clusters.

It is noted that, while HCA misclassified one sample (with an index of X.ID = 134 in the Oslo transect dataset) of ROG into the FER cluster (Fig. 3(c)), this did not occur to *t*-SNE. Because discrepancy between *t*-SNE and HCA results is not uncommon, it is worth analyzing the reasons. Fig. 4(a) plots the HCA dendrogram for clusters ROG (C5) and FER (C6). Cluster ROG is further separated into two clusters, B_1 and B_2, and FER into four clusters $B_3 - B_6$. The misclassified sample (X.ID = 134) of ROG is denoted as cluster *A*. In HCA, cluster *A* was merged with cluster B_4, not with B_1 or B_2. As explained in Section 2.3, HCA initially treats each sample as an individual cluster, and proceeds to merge the two closest clusters until only a single cluster remains. If cluster *A* is merged with cluster B_1 or B_2, the increase in *SSE* is 7.17 or 6.06, larger than the

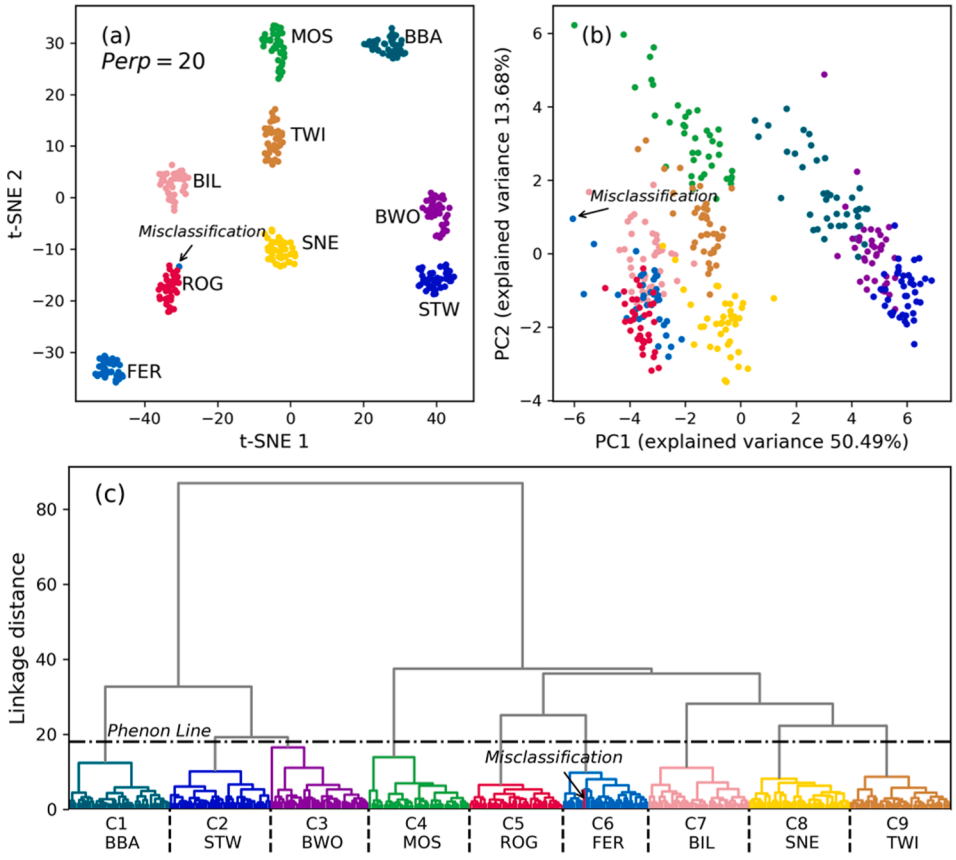

Fig. 3. Two-dimensional (a) *t*-SNE and (b) PCA visualizations of the Oslo transect dataset and (c) the HCA dendrogram using the Ward method with Euclidian distance. Each dot in the PCA and *t*-SNE visualizations represents one sample. Dot colors correspond to those used for the HCA clusters.

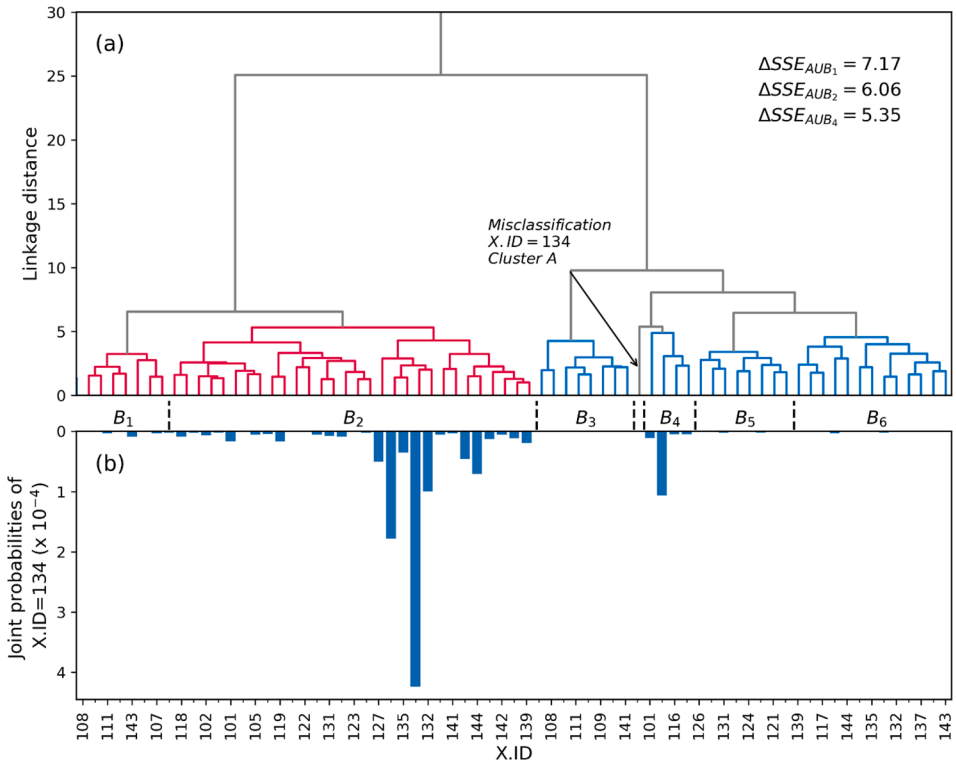

Fig. 4. (a) HCA dendrogram of Clusters C5 (ROG) and C6 (FER), which is part of the dendrogram of Fig. 3(c), and (b) joint probability that the sample (X.ID = 134) would pick another sample as its neighbor according to Equation (6). The *x*-axis of (b) is the index of the samples of Clusters C5 and C6.

increase of 5.35 upon merging clusters *A* and B_4. As a result, the sample (X.ID = 134) of ROG was assigned to cluster FER because of the smaller increase in *SSE*. This is consistent with HCA theory because *SSE* does not consider distances between samples, instead *SSE* considers distances between the sample and centers of other clusters. This differs from *t*-SNE, which is based on distance between paired data points or map points. Fig. 4(b) shows the probability of sample X.ID = 134 picking another sample in clusters $B_1 - B_6$ as its neighbor according to Eq. (6). The probability is substantially higher for samples in cluster B_2 of ROG, which explains why sample X.ID = 134 was assigned to ROG by *t*-SNE (Fig. 3a). Because *t*-SNE and HCA are based on different theories, it is not surprising that the two methods yield different results for certain samples. Although the nine groups of samples shown in the *t*-SNE visualization fully agree with the theoretically known number of sample groups and group memberships, it should be noted that *t*-SNE alone cannot be used for cluster analysis, because the visualizations strongly depend on perplexity (see Section 4). Because both perplexity and the number of clusters are unknown in practice (the Oslo transect dataset is a special case with known number of clusters), we only recommend using *t*-SNE in conjunction with HCA.

Because the theoretical number of clusters is known for the Oslo transect dataset, it is an opportunity to evaluate the effects of alternative transformations. We applied the Box-Cox transformation as an alternative to the log transformation to the 24 elements and the loss on ignition of the Oslo transect dataset. The histograms of the log-transformed data are plotted in Fig. S1, and the histograms of the Box-Cox-transformed data are plotted in Fig. S2 in the supplementary information file. The Box-Cox transformation was conducted by using the box-cox function of SciPy, and the resulted λ values are shown in Fig. S2. The Lilliefors test was used to examine normality of the transformed data (Lilliefors, 1967), and the *p* values are shown in Figs. S1 and S2. The results of the Lilliefors test show that neither of the two transformations yield data following normal distributions, although the results of the Box-Cox transformation are slightly closer to the normal distribution than the results of the log transformation. PCA analysis was applied to the log-transformed and Box-Cox-transformed data, and the results are plotted in Fig. S3. The similar visualization patterns in this figure indicate that the two transformations have negligible effects on the PCA results. The HCA and *t*-SNE were also applied to the Box-Cox-transformed data, and the results are plotted in Fig. S4. The figure shows that the performance of the Box-Cox transformation is worse than that of the log transformation, in that there are more samples misclassified as shown in Fig. S4(a) for the Box-Cox-transformed data.

We also evaluated other statistical methods that have been used to assist cluster analysis for determining the appropriate number of clusters. They are the Elbow, Silhouette, and Gap statistic methods, and a description of the methods is given in Text S1 of the supplementary information file. The results of the three methods are shown in Fig. 5. While the Elbow and Silhouette methods indicated that optimal cluster number was 2, the Gap statistic values monotonically increased and the optimal number of clusters was 16 by using the rule of one-standard-error. It was surprising that all three methods failed to yield the correct number of clusters. This is also the case for the Taiyuan Karst water dataset and the Jianghan Plain groundwater dataset. The results for these two datasets are shown in Figs. S5 and S6. Exploring the reasons is beyond the scope of this study, but it is warranted in future studies.

3.2. *Cluster analysis and spatial zone delineation for Taiyuan karst water dataset*

The Taiyuan karst water dataset was subject to a *t*-SNE analysis to support HCA clustering results and to delineate spatial zones of groundwater geochemistry. The results of Ma et al. (2011) were used a reference for the evaluation. By following the preprocessing procedure described in Section 2.2, parameter CO_3^{2-} was excluded from the dataset, because all its measurements were less than the detection limit. In addition, ten samples were removed because of missing measurements. Ultimately, fourteen samples in XMK, seven in BMK, and six in DMK were used in this study. Due to the small sample size for the BMK and DMK subsystems, only the results of XMK are presented and discussed

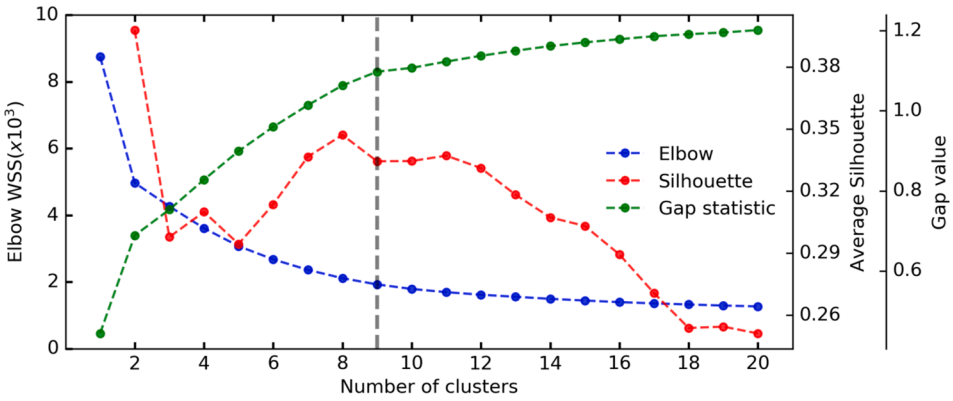

Fig. 5. Results of the Elbow method, Silhouette coefficient analysis, and Gap statistic method for determining the optimal number of clusters for the Oslo transect dataset. The vertical dashed line denotes the nine plant materials of Templ et al. (2008).

below. The results of HCA, PCA, and *t*-SNE for BMK and DMK data are shown in Fig. S7 of the supplementary information file.

Fig. 6(a) is the two-dimensional *t*-SNE visualization (using *Perp* = 0.1) of the 14 geochemical samples collected from the XMK sub-system. The three clusters shown in Fig. 6(a) correspond exactly to the three clusters shown in Fig. 6(c) by placing the phenon line at a linkage distance of 11. Comparing the *t*-SNE visualization to the PCA visualization shown in Fig. 6(b) revealed that *t*-SNE outperformed PCA when graphically validating the clustering results. For the samples collected from the XML sub-system (denoted by triangles in Fig. 1), the three clusters shown in Fig. 6(a) for *t*-SNE visualization agreed with the three groundwater geochemistry zones identified by Ma et al. (2011) and shown in Fig. 1. To facilitate a visual comparison, the three clusters and the three spatial zones were distinguished using the same color scheme, i.e., black for the recharge and flow-through zone, blue for the cold-water discharge zone, and red for the thermal-water discharge zone. The only disagreement between the clusters and the spatial zones was for sample RN32. Because the sample's temperature was 30°C, it should have belonged to the thermal-water discharge zone (Fig. 1). However, both HCA and *t*-SNE misclassified the sample into the cold-water discharge zone, because the sample's SO_4^{2-} concentration is 172 mg/L, about one order of magnitude less than the concentrations of other samples in the thermal-water discharge zone but about the same order of magnitude as the concentrations of the samples in the cold-water discharge zone.

3.3. Cluster analysis and spatial zone delineation for Jianghan Plain groundwater dataset

The Jianghan Plain groundwater dataset has 1,184 samples collected over 23 years (1992–2014) from 29 monitoring wells distributed throughout a regional aquifer at the scale of hundreds of kilometers. The 13,024 geochemical measurements pre-processed by Yang et al. (2020) were used in this study. Fig. 7(a) shows that, by drawing the phenon line at a linkage distance of 25, the 1,184 samples were grouped into seven clusters, denoted as C1–C7. Fig. 7(b) is the two-dimensional PCA visualization of the seven clusters. While the seven clusters were generally separated, there were substantial overlaps, especially for C5–C7. The overlaps were substantially decreased in the two-dimensional *t*-SNE visualization shown in Fig. 7(c). For example, the dots of C5–C7 were mostly separated, demonstrating that *t*-SNE outperformed PCA.

Fig. 7(c) shows that *t*-SNE divided C4 and C7 each into two sub-clusters. The same results were obtained by moving the phenon line down from the linkage distance of 24 (Fig. 7(a)) to that of 22 (Fig. 7(d)). The two sub-clusters of C4 were denoted as C4S1 and C4S2 and the two sub-cluster of C7 as C7S1 and C7S2. Fig. 7(e) and 7(f) are the corresponding PCA and *t*-SNE visualizations, respectively, for the nine clusters. The overlap shown in Fig. 7(b) is also observed in Fig. 7(e). Comparing the PCA-based visualization and the *t*-SNE-based visualization indicates that *t*-SNE was a better visualization tool for the reduced dimensions.

The investigation on whether it is reasonable to have seven or nine clusters started with C7S1 and C7S2. Of the 50 samples comprising

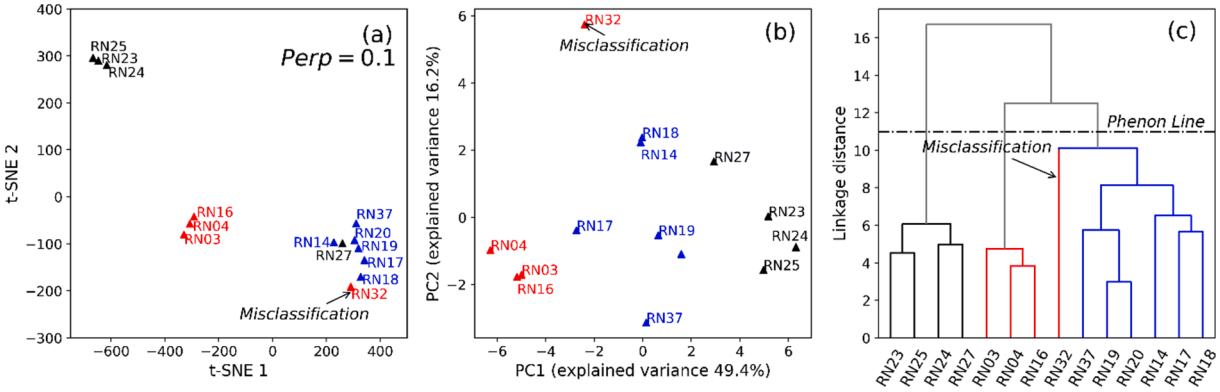

Fig. 6. (a) Two-dimensional *t*-SNE visualization of the geochemical data of the XMK sub-system of the Taiyuan karst water dataset, (b) two-dimensional PCA visualization of the geochemical data, and (c) HCA dendrogram using the Ward method with Euclidian distance. Each triangle in the PCA and *t*-SNE visualizations represents one sample. The colors correspond to those used in Fig. 1, i.e., black for the recharge and flow-through zones, blue for the cold-water discharge zone, and red for the thermal-water discharge zone identified by Ma et al. (2011). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

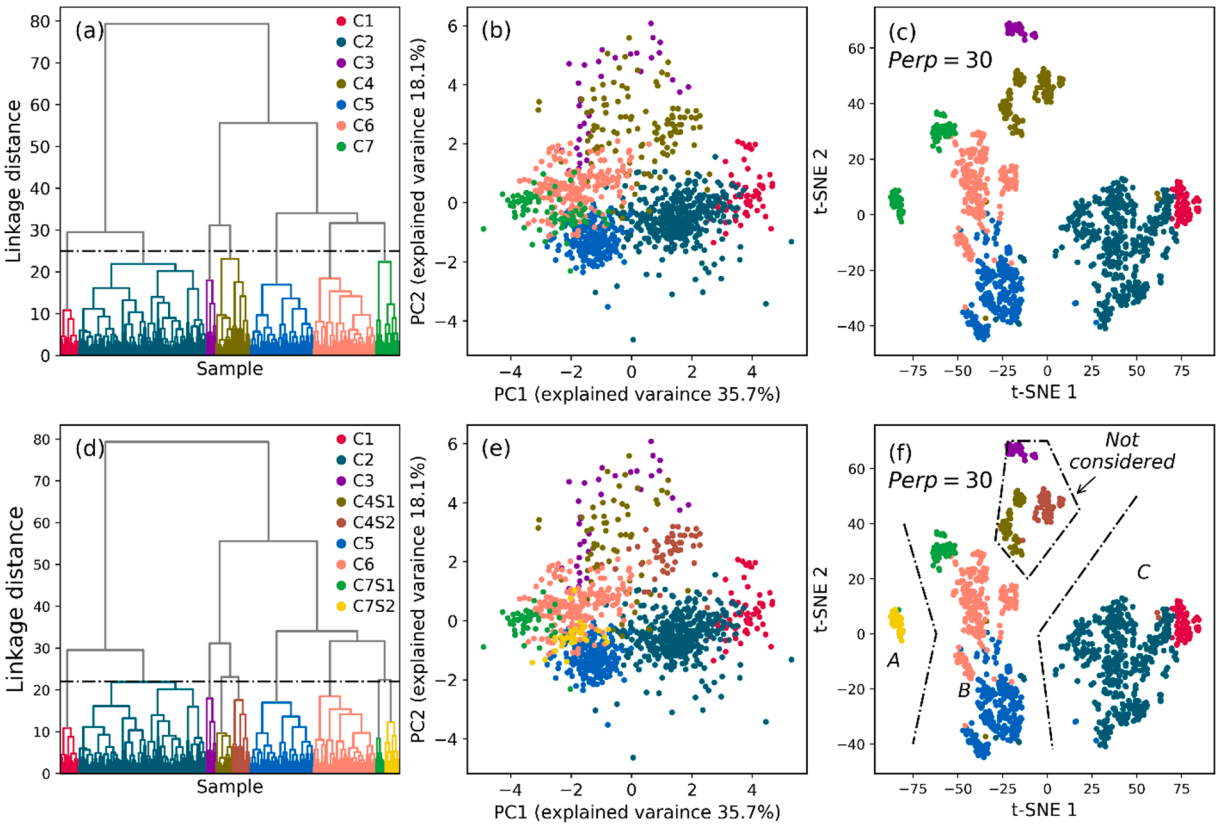

Fig. 7. (a) HCA dendrogram using the Ward method with Euclidian distance with a phenon line location at a linkage distance of 25, which yielded seven clusters, (b) two-dimensional PCA visualization of the seven clusters, and (c) two-dimensional *t*-SNE visualization of the seven clusters. (d)–(f) are equivalent to (a)–(c), respectively, except that there were nine clusters with the phenon line at 22. Each point in the PCA and *t*-SNE visualizations represents one sample.

C7S1, 39 were from well J1 (Fig. S8(a)); of the 33 samples in C7S2, 31 were from well J8 (Fig. S8(b)). As shown in Fig. 2, well J1 was located in the recharge zone (Zone I) with limited water–rock interaction, while well J8 was located in the transition zone (Zone II). It appears reasonable to separate the data from the two wells into two clusters based on their locations. This was also supported by examining the groundwater geochemistry of the two clusters. The box plots of 11 groundwater geochemical parameters in Fig. 8(a) indicate that the concentrations of K^+, NH_4^+, and Fe were substantially lower in C7S1 than in C7S2, and that the concentrations of Mg^{2+}, Na^+, Cl^-, and HCO_3^- were substantially higher in C7S1 than in C7S2. The differences between the two clusters may justify separation of C7 into C7S1 and C7S2. Part of the difference may be explained by the impacts of the Yangtz River on the groundwater samples in C7S2 wherein 31 out of 33 samples were from well J8, which was only 200 m from the river. Yang et al. (2020) found that river water recharged groundwater in the area where well J8 was located, and Li et al. (2014) reported that the mean concentrations of Ca^{2+}, Mg^{2+}, and HCO_3^- of the river water were 37.16, 6.77, and 126.89 mg/L, respectively, which were lower than the groundwater concentrations. Groundwater geochemistry of well J8 reflected the mixing of river water and groundwater, and this may explain why the concentrations of Ca^{2+}, Mg^{2+}, and HCO_3^- were substantially lower in C7S2 than in C7S1. Separation of C7 into C7S1 and C7S2 inspired by the *t*-SNE visualization helped better understand groundwater geochemistry subject to groundwater and surface water interactions.

The geochemical reasons that C4 was separated into C4S1 and C4S2 were not fully understood. Yang et al. (2020) found that C4 had high SO_4^{2-} concentrations caused by anthropogenic activities, and this was true for both C4S1 and C4S2, as shown in Fig. 8(b). However, the 62 samples of C4S1 were from 16 wells (Fig. S8(c)), while the 57 samples of C4S2 were from only nine wells with 40 out of the 57 samples from well J9 (Fig. S8(d)). As a result, the concentration ranges of the geochemical parameters (except pH, NH_4^+, and Fe) were larger in C4S1 than in C4S2, as shown in Fig. 8(b). This finding however was not useful for better understanding groundwater geochemistry in the Jianghan Plain aquifer.

t-SNE visualization is a promising tool for delineating spatial zones of groundwater geochemistry at the regional scale. The two-dimensional *t*-SNE visualization of Fig. 7(f) shows three zones, denoted as A, B, and C, that correspond to Zones I – IV delineated by Yang et al. (2020) and shown in Fig. 2. Zone A of cluster C7S2 corresponds to Zone I, the recharge zone. Zone B of clusters C5, C6, and C7S1 corresponds to Zone II, the transition zone. Zone C of clusters C1 and C2 corresponds to Zones III and IV, the flow-through zone and discharge-mixing zone, respectively. Zones III and IV were separated by Yang et al. (2020) to reflect the impacts of the Three Gorges Dam on the groundwater geochemistry observed at well J21. This separation was supported by the *t*-SNE visualization because clusters C1 and C2 were broadly separated. Although Yang et al. (2020) delineated the four zones of groundwater geochemistry without using *t*-SNE visualization, the delineation would have been more obvious if *t*-SNE visualization had been used. For example, Yang et al. (2020) first determined the clusters that the majority of the samples belonged to, and then mapped the locations of the wells and their corresponding clusters, based on which zone delineation was made. The two steps could be simplified into a single step by examining the zone patterns revealed by the *t*-SNE visualization.

In Fig. 7(f), the samples from C3, C4S1, and C4S2 were not used for zone delineation because they were impacted by anthropogenic actives. For example, the samples in C3 were from well J10 located at one of the largest pesticide factories in China, and the samples were characterized by extremely high concentrations of Na^+ and Cl^- due to wastewater infiltration from the factory to the aquifer. Of the 62 samples in C4S1, 15 were from well J25 located in a water-supply plant while 10 samples

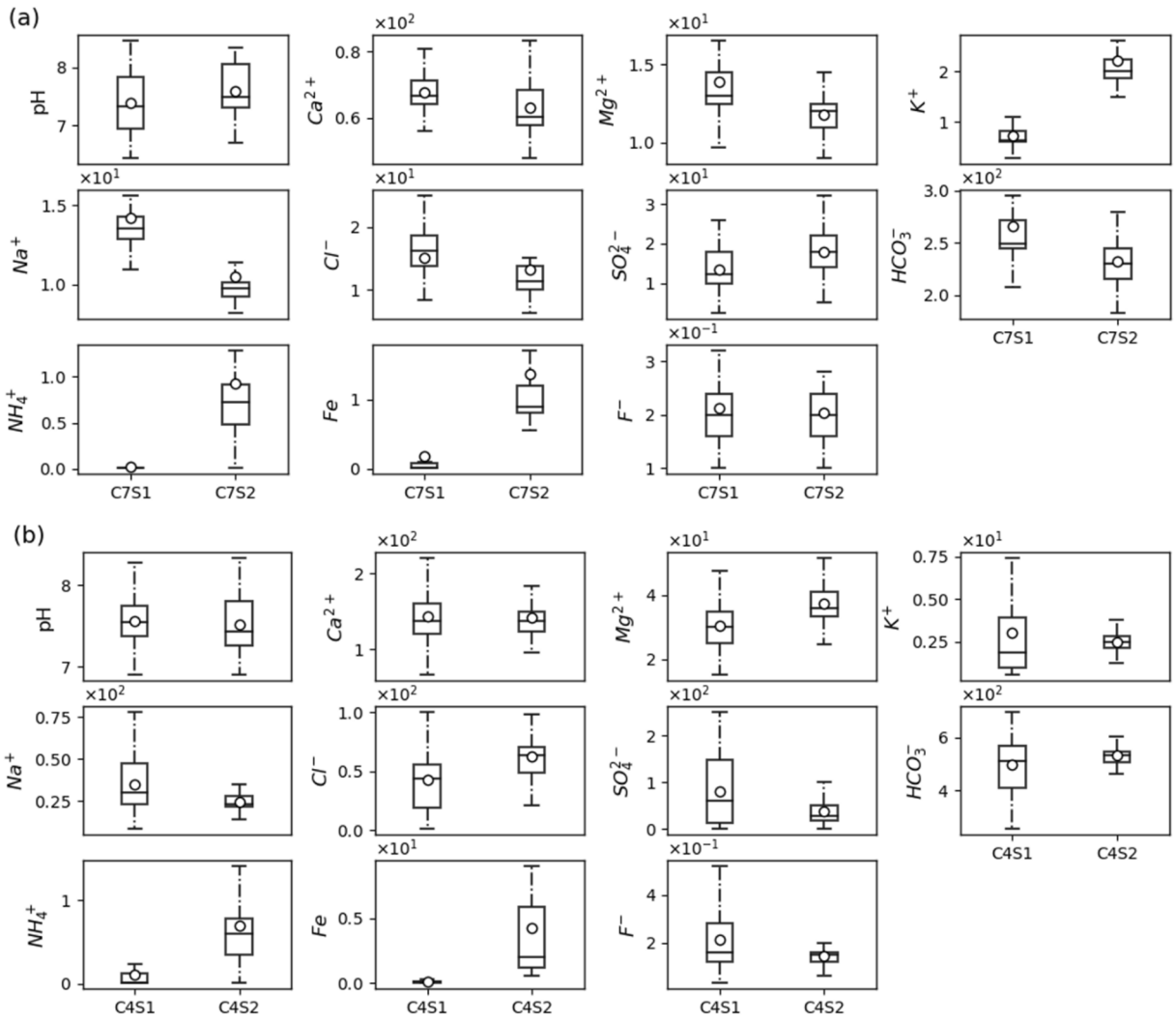

Fig. 8. Box plots for the concentrations of the 11 parameters of the two sub-Clusters (a) C7S1 and C7S2 and (b) C4S1 and C4S2. The units of all geochemical parameters (except pH) are mg/L.

were from well J8 about 300 m away from a cotton mill (Fig. S2(c)). Of the 57 samples in C4S2, 40 were from well J9 located near another cotton mill (Fig. S2(d)). It is interesting that the impacts of anthropogenic activities on groundwater quality are unambiguously revealed on the *t*-SNE visualization.

4. Discussion

4.1. Running t-SNE with different perplexity values

Using different perplexity values can substantially change the *t*-SNE visualization patterns (Wattenberget al., 2016). In the original *t*-SNE paper, van der Maaten and Hinton (2008) used five datasets to compare the performance of *t*-SNE, and the size of the datasets ranged from 400 to 10,000. They stated that "*the performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50*". Our literature review indicates that 30 and 50 are two commonly used perplexity values for datasets with hundreds to thousands of samples (Table 1). To explore the impacts of perplexity on clustering analyses of groundwater geochemistry, *t*-SNE was conducted with different perplexity values for all three datasets. Fig. 9 illustrates results from the Jianghan Plain groundwater dataset while the results from the Oslo transect and Taiyuan karst water datasets are shown in Figs. S9 and S10, respectively, in the supplementary information file.

Table 1
Sample size and perplexity values used in the literature.

	Sample size	Perplexity
Balamurali and Melkumyan (2016)	239 and 14,906	30
Aizarani et al. (2019)	10,372	Not mentioned
Roche et al. (2018)	2,016	Not mentioned
Pouyet et al. (2018)	Not mentioned	50
Balamurali et al. (2019)	66,344	30
Horrocks et al. (2019)	16,165	30
Mazher (2020)	5,688	Not mentioned

van der Maaten and Hinton (2008) reported stable *t*-SNE results for $5 < Perp < 50$. This was true for the Oslo transect dataset with 350 samples as well as for the Jianghan Plain groundwater data with 1,184 samples. Fig. 9 shows that with the perplexity values in the range from 20 to 50, similar patterns were observed in the *t*-SNE visualizations. For $Perp < 10$, local variations within each cluster dominated, and too many clusters were returned. For $Perp > 300$, the samples could not be separated into distinct clusters. Because perplexity is *a priori* unknown for most real-world problems, it must be determined in conjunction with HCA. A small perplexity value should be assigned when sample size is small. For the 14 samples of XMK in the Taiyuan karst water dataset, as shown in Fig. S4, $Perp < 1$ yielded three groups; for $Perp > 1$, the three groups started mixing into one. Based on our experience of applying *t*-SNE to

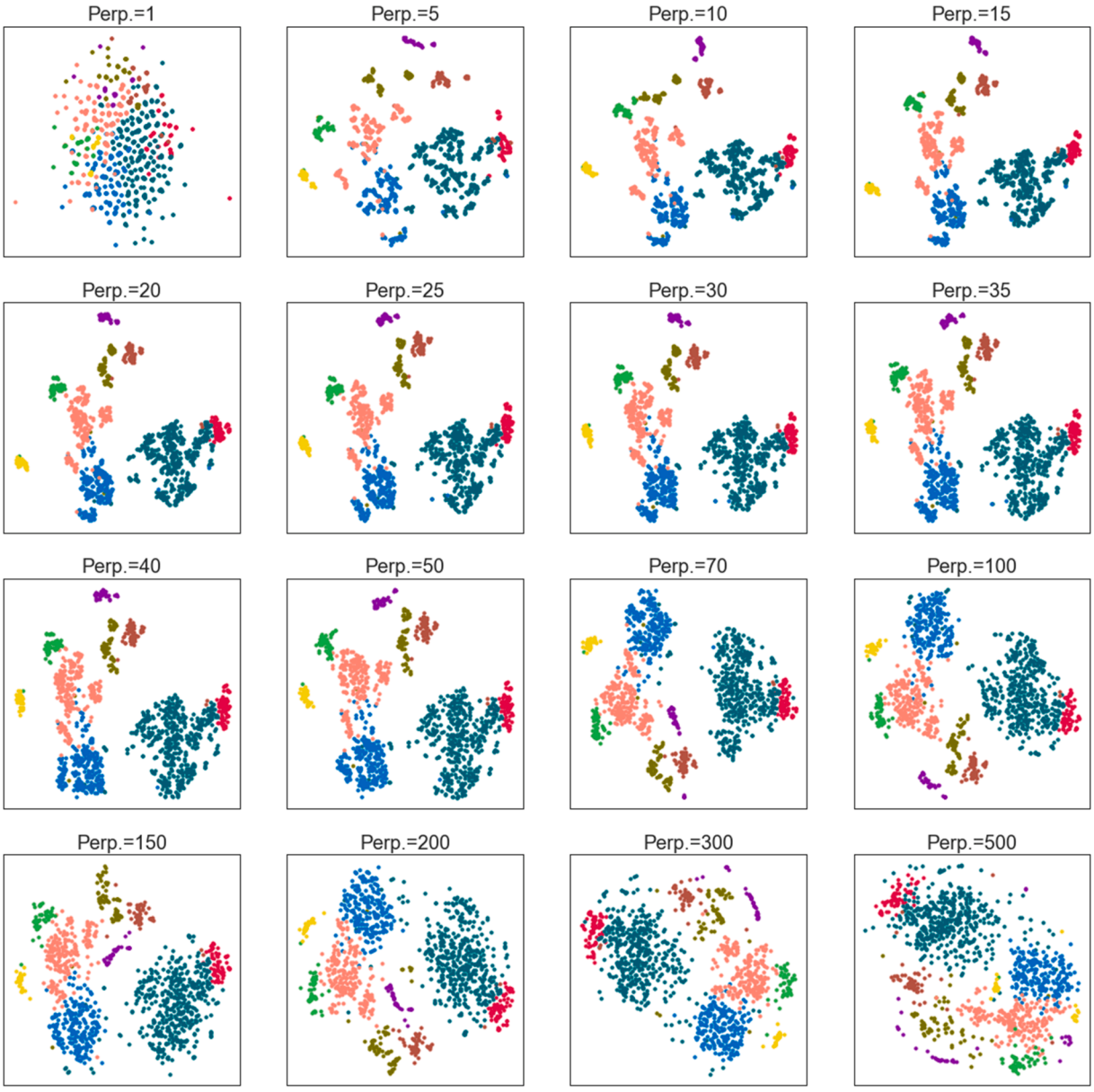

Fig. 9. *t*-SNE visualization of the Jianghan Plain groundwater dataset with various perplexities. Dot colors correspond to those used for the nine HCA clusters shown in Fig. 5(d).

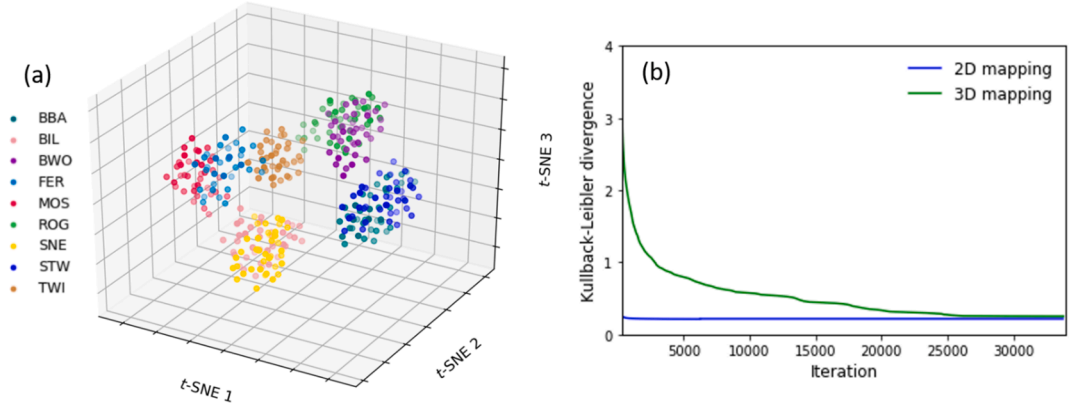

Fig. 10. (a) Three-dimensional *t*-SNE visualization of the Oslo transect dataset and (b) convergence of the Kullback-Leibler divergence for the two-dimensional (2D) mapping and three-dimensional (3D) mapping. The convergence line of the 2D mapping was extended for a visual comparison with the converge of the 3D mapping.

the three datasets, rather than using the range of $5 < Perp < 50$, the range should be $n/200 < Perp < n/10$, where n is number of samples.

4.2. Three-dimensional t-SNE visualization

t-SNE can map high dimensional data to any number (e.g., two or three) of lower dimensions, and a three-dimensional *t*-SNE visualization may be more useful than a two-dimensional *t*-SNE visualization for visually determining the number of clusters and cluster memberships. We explored this for the Oslo transect dataset, for which the theoretical number of clusters and cluster memberships are known. Fig. 10(a) shows the three-dimensional *t*-SNE visualization based on a perplexity value of 20 that was used to generate the two-dimensional *t*-SNE visualization shown in Fig. 3(a). Although cluster overlaps are observed in Fig. 10(a), these were simply caused by plotting the three-dimensional figure on a two-dimensional paper. Rotating the three-dimensional figure shows that the nine clusters for the nine plant materials were well separated in the three-dimensional *t*-SNE visualization.

In comparison with the two-dimensional visualization shown in Fig. 3(a), the three-dimensional visualization in Fig. 10(a) does not add value to cluster analysis, because the nine clusters were equally well separated in the two figures (after rotating the three-dimensional figure). This can be explained by examining the Kullback-Leibler divergence, the cost function (Eq. 8) minimized during a *t*-SNE run. The Kullback-Leibler divergence measures discrepancy between the joint probability distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ in high- and low-dimensional spaces, respectively. A smaller value of the Kullback-Leibler divergence indicates that $\boldsymbol{P}$ and $\boldsymbol{Q}$ are closer to each other. Fig. 10(b) shows convergence of the Kullback-Leibler divergence for generating the two- and three-dimensional *t*-SNE visualizations of the Oslo transect dataset. The figure shows that the Kullback-Leibler divergence converged to a similar value, but the two-dimensional *t*-SNE run converged substantially faster than the three-dimensional *t*-SNE run. This case study for the Oslo transect dataset suggested that it was unnecessary to generate the three-dimensional *t*-SNE visualization. It was also the case for the other two datasets considered in this study, because their corresponding two-dimensional *t*-SNE visualizations sufficiently separated the HCA clusters. Three-dimensional (or higher) *t*-SNE visualizations may be useful when two-dimensional *t*-SNE visualizations cannot separate HCA clusters, and the Kullback-Leibler divergence may be an indicator for when three-dimensional *t*-SNE visualizations are needed.

5. Conclusions

This study for the first time applied *t*-SNE as a graphic approach to assist HCA cluster analysis and delineation of spatial zones of groundwater geochemistry based on clustering results. The application of HCA, PCA, and *t*-SNE to the three geochemical datasets leads to the following major conclusions:

(1) In comparison with PCA, *t*-SNE was a better graphical way to assist HCA cluster analysis with respect to determining the number of clusters and cluster memberships. The *t*-SNE visualization outperformed the PCA visualization to separate the clusters determined by HCA.
(2) In comparison with the conventional way of delineating spatial zones based on clustering results, *t*-SNE is a promising tool for effectively and efficiently delineating spatial zones of groundwater geochemistry, because *t*-SNE maintains pair-wise distances and structures in the high-dimensional data space to the extent possible in the low-dimensional *t*-SNE space.
(3) HCA clustering results may differ from *t*-SNE grouping results, because the two methods are based on different theories. It is recommended to use *t*-SNE as a graphic way to support HCA cluster analysis, not to use *t*-SNE alone for cluster analysis, because *t*-SNE visualization depends substantially on perplexity that is unknown for real-world problems.
(4) Although *t*-SNE developers recommended constraining perplexity to within 5 – 50, a substantially smaller perplexity (i.e., <1) was appropriate for this small geochemical dataset with only tens of samples. Based on our experience, we suggest setting perplexity within the range of $n/200$ – $n/10$, where n is number of samples.

t-SNE cannot be used as a standalone algorithm for cluster analysis, and can only be used as a graphic approach to assist clustering methods such as HCA and *k*-means. There remains room to improve *t*-SNE for general applications. For example, measures other than Euclidian distance or kernels other than Gaussian could be used to estimate similarities between geochemical samples. These are warranted in a future study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by FSU Research Foundation project RF04063, National Science Foundation grant EAR-1828827, and National Natural Science Foundation of China grant U1911205. We are grateful to Associate Editor Saket Pande and two anonymous reviewers for their constructive comments that substantially improved the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2021.126146.

References

Abdelmoula, W.M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M.J., Walch, A., McDonnell, L.A., Lelieveldt, B.P., 2016. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proc. Natl. Acad. Sci. U. S. A. 113 (43), 12244–12249.
Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Herman, J.S., Pessaux, P., Baumert, T.F., Grun, D., 2019. A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature 572 (7768), 199−204.
Appelo, C.A.J., Postma, D., 2005. Geochemistry. Groundwater and Pollution, Taylor and Francis, Great Britain.
Balamurali, M., Melkumyan, A., 2016. t-SNE based visualisation and clustering of geological domain, International Conference on Neural Information Processing. Springer International Publishing, Cham, pp. 565−572.
Balamurali, M., Silversides, K.L., Melkumyan, A., 2019. A comparison of t-SNE, SOM and SPADE for identifying material type domains in geological data. Comput. Geosci. 125, 78–89.
Cloutier, V., Lefebvre, R., Therrien, R., Savard, M.M., 2008. Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. J. Hydrol. 353 (3–4), 294–313.
Dobie, R., Wilson-Kanamori, J.R., Henderson, B.E.P., Smith, J.R., Matchett, K.P., Portman, J.R., Wallenborg, K., Picelli, S., Zagorska, A., Pendem, S.V., Hudson, T.E., Wu, M.M., Budas, G.R., Breckenridge, D.G., Harrison, E.M., Mole, D.J., Wigmore, S. J., Ramachandran, P., Ponting, C.P., Teichmann, S.A., Marioni, J.C., Henderson, N. C., 2019. Single-Cell Transcriptomics Uncovers Zonation of Function in the Mesenchyme during Liver Fibrosis. Cell Rep. 29 (7), 1832−1847.
Ellefsen, K.J., Smith, D.B., Horton, J.D., 2014. A modified procedure for mixture-model clustering of regional geochemical data. Appl. Geochem. 51, 315–326.
Fendorf, S., Michael, H.A., van Geen, A., 2010. Spatial and temporal variations of groundwater arsenic in south and southeast Asia. Science 328 (5982), 1123–1127.
Gan, Y., Zhao, K., Deng, Y., Liang, X., Ma, T., Wang, Y., 2018. Groundwater flow and hydrogeochemical evolution in the Jianghan Plain, central China. Hydrogeol. J. 26 (5), 1609–1623.
Gorelick, S.M., Zheng, C., 2015. Global change and the groundwater management challenge. Water Resour. Res. 51 (5), 3031–3051.
Green, T.R., Taniguchi, M., Kooi, H., Gurdak, J.J., Allen, D.M., Hiscock, K.M., Treidel, H., Aureli, A., 2011. Beneath the surface of global change: Impacts of climate change on groundwater. J. Hydrol. 405 (3–4), 532–560.

Güler, C., Kurt, M.A., Alpaslan, M., Akbulut, C., 2012. Assessment of the impact of anthropogenic activities on the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) using fuzzy clustering, multivariate statistics and GIS techniques. J. Hydrol. 414–415, 435–451.
Güler, C., Thyne, G.D., 2004. Delineation of hydrochemical facies distribution in a regional groundwater system by means of fuzzy c-means clustering. Water Resour. Res. 40 (12).
Güler, C., Thyne, G.D., 2004. Hydrologic and geologic factors controlling surface and groundwater chemistry in Indian Wells-Owens Valley area, southeastern California. USA. J. Hydrol. 285 (1–4), 177–198.
Haile, E., Fryar, A.E., 2017. Chemical evolution of groundwater in the Wilcox aquifer of the northern Gulf Coastal Plain. USA. Hydrogeol. J. 25 (8), 2403–2418.
Hinton, G.E., Roweis, S.T., 2002. Stochastic Neighbor Embedding, Advances in Neural Information Processing Systems. The MIT Press, Cambridge, MA, USA, 833−40.
Horrocks, T., Holden, E.-J., Wedge, D., Wijns, C., Fiorentini, M., 2019. Geochemical characterisation of rock hydration processes using t-SNE. Comput. Geosci. 124, 46–57.
Jones, E., Oliphant, T., Peterson, P., 2001. In. Open Source Scientific Tools for Python, SciPy.
Kassambara, A., 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA, Poland.
Kobak, D., Berens, P., 2019. The art of using t-SNE for single-cell transcriptomics. Nature Comm. 10 (1), 5416.
Leung, R., Balamurali, M., Melkumyan, A., 2019. Sample truncation strategies for outlier removal in geochemical data: the MCD robust distance approach versus t-SNE ensemble clustering. Mathemat. Geosci.
Li, X., Liu, Y., Zhou, A., Zhang, B., 2014. Sulfur and oxygen isotope compositions of dissolved sulfate in the Yangtze River during high water period and its sulfate source tracing. Earth Sci. J. China Univ. Geosci. 39 (11), 1647–1654 (in Chinese).
Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 62 (318), 399–402.
Ma, R., Wang, Y., Sun, Z., Zheng, C., Ma, T., Prommer, H., 2011. Geochemical evolution of groundwater in carbonate aquifers in Taiyuan, northern China. Appl. Geochem. 26 (5), 884–897.
Mazher, A., 2020. Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques. Water 12 (2), 590–604.
Melit Devassy, B., George, S., 2020. Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. Forensic Sci. Int. 311, 110194–110203.
Nguyen, T.T., Kawamura, A., Tong, T.N., Nakagawa, N., Amaguchi, H., Gilbuena, R., 2015. Clustering spatio–seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta. Vietnam. J. Hydrol. 522, 661–673.
Ouyang, Y., 2005. Evaluation of river water quality monitoring stations by principal component analysis. Water Res. 39 (12), 2621–2635.
Pacheco Castro, R., Pacheco Avila, J., Ye, M., Cabrera Sansores, A., 2018. Groundwater quality: analysis of its temporal and spatial variability in a karst aquifer. Groundwater 56 (1), 62–72.
Pant, R.R., Zhang, F., Rehman, F.U., Wang, G., Ye, M., Zeng, C., Tang, H., 2018. Spatiotemporal variations of hydrogeochemistry and its controlling factors in the Gandaki River Basin, Central Himalaya Nepal. Sci. Total Environ. 622, 770–782.
Pedregosa, F., et al., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pouyet, E., Rohani, N., Katsaggelos, A.K., Cossairt, O., Walton, M., 2018. Innovative data reduction and visualization strategy for hyperspectral imaging datasets using t-SNE approach. Pure Appl. Chem. 90 (3), 493–506.
Reimann, C., et al., 2007. Element contents in mountain birch leaves, bark and wood under differ ent anthropogenic and geogenic conditions. Appl. Geochem. 22 (7), 1549–1566.
Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ. Geol. 39 (9), 1001–1014.
Rencher, A.C., 2003. Methods of Multivariate Analysis, 2nd Edition. John Wiley & Sons Inc, New York.
Roche, K.E., Weinstein, M., Dunwoodie, L.J., Poehlman, W.L., Feltus, F.A., 2018. Sorting five human tumor types reveals specific biomarkers and background classification genes. Sci. Rep. 8 (1), 8180–8191.
Sanford, R.F., Pierson, C.T., Crovelli, R.A., 1993. An objective replacement method for censored geochemical data. Math. Geol. 25 (1), 59–80.
Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: Problems and possibilities. Appl. Geochem. 23 (8), 2198–2213.
Todorov, V., Filzmoser, P., 2010. Robust statistic for the one-way MANOVA. Comput. Stat. Data Anal. 54 (1), 37–48.
Tóth, J., 1999. Groundwater as a geologic agent: An overview of the causes, processes, and manifestations. Hydrogeol. J. 7 (1), 1–14.
Tóth, J., 2009. Gravitational Systems of Groundwater Flow: Theory, Evaluation, Utilization. Cambridge University Press, Cambridge.
van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11), 2579–2605.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244.
Wattenberg, M., Viégas, F., Johnson, I., 2016. How to use t-SNE effectively. Distill 1 (10), Article e2.
Yang, J., Ye, M., Tang, Z., Jiao, T., Song, X., Pei, Y., Liu, H., 2020. Using cluster analysis for understanding spatial and temporal patterns and controlling factors of groundwater geochemistry in a regional aquifer. J. Hydrol. 583, 124594–125604.
Zheng, J., Jiang, Z., Pan, H., 2018. Sigmoid-based refined composite multiscale fuzzy entropy and t-SNE based fault diagnosis approach for rolling bearing. Measurement 129, 332–342.
Zheng, S.D., Zhao, J.S., 2020. A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis. Comput. Chem. Eng. 135, 106755–106872.
Zhu, J., Nolte, A.M., Jacobs, N., Ye, M., 2020. Using machine learning to identify karst sinkholes from LiDAR-derived topographic depressions in the Bluegrass Region of Kentucky. J. Hydrol. 588.