Science of the Total Environment 905 (2023) 167176

Contents lists available at ScienceDirect

Science of the Total Environment

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

Unveiling elemental fingerprints: A comparative study of clustering methods for multi-element nanoparticle data

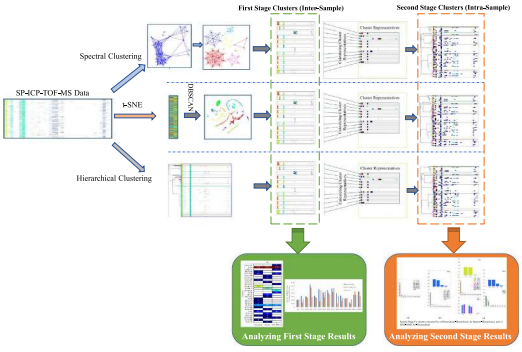
Mahdi Erfani [a], Mohammed Baalousha [b,*], Erfan Goharian [a,*]

[a] Department of Civil and Environmental Engineering, University of South Carolina, SC 29208, USA
[b] Center for Environmental Nanoscience and Risk, Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29201, USA

HIGHLIGHTS

- Single particle elemental compositions were acquired using SP-ICP-TOF-MS.
- Elemental composition data were transformed using tSNE and Spectral clustering.
- Data transformation allowed extracting subclusters not extractable by hierarchical clustering.
- Hierarchical, tSNE-DBSCAN, and spectral clustering approaches were used to extract nanoparticle clusters.
- Spectral clustering is more robust and flexible compared to tSNE-DBSCAN and hierarchical clustering.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Kevin V. Thomas

Keywords:
Engineered nanoparticles
Multi-element single nanoparticle
Mass spectrometry
High dimensional data
Nonlinear clustering
tSNE
Spectral clustering

ABSTRACT

Single particle-inductively coupled plasma-time of flight-mass spectrometers (SP-ICP-TOF-MS) generates large datasets of the multi-elemental composition of nanoparticles. However, extracting useful information from such datasets is challenging. Hierarchical clustering (HC) has been successfully applied to extract elemental fingerprints from multi-element nanoparticle data obtained by SP-ICP-TOF-MS. However, many other clustering approaches can be applied to analyze SP-ICP-TOF-MS data that have not yet been evaluated. This study fills this knowledge gap by comparing the performance of three clustering approaches: HC, spectral clustering, and t-distributed Stochastic Neighbor Embedding coupled with Density-Based Spatial Clustering of Applications with Noise (tSNE-DBSCAN) for analyzing SP-ICP-TOF-MS data. The performance of these clustering techniques was evaluated by comparing the size of the extracted clusters and the similarity of the elemental composition of nanoparticles within each cluster. Hierarchical clustering often failed to achieve an optimal clustering solution for SP-ICP-TOF-MS data because HC is sensitive to the presence of outliers. Spectral clustering and tSNE-DBSCAN extracted clusters that were not identified by HC. This is because spectral clustering, a method developed based on graph theory, reveals the global and local structure in the data. tSNE reduces and maps the data into a lower-dimensional space, enabling clustering algorithms such as DBSCAN to identify subclusters with subtle differences in their elemental composition. However, tSNE-DBSCAN can lead to unsatisfactory clustering solutions because tuning the perplexity hyperparameter of tSNE is a difficult and a time-consuming task, and the relative distance

* Corresponding authors.
E-mail addresses: mbaalous@mailbox.sc.edu (M. Baalousha), goharian@cec.sc.edu (E. Goharian).

https://doi.org/10.1016/j.scitotenv.2023.167176
Received 2 July 2023; Received in revised form 3 September 2023; Accepted 16 September 2023
Available online 18 September 2023

between datapoints is not maintained. Although the three clustering approaches successfully extract useful information from SP-ICP-TOF-MS data, spectral clustering outperforms HC and tSNE-DBSCAN by generating clusters of a large number of nanoparticles with similar elemental compositions.

1. Introduction

Engineered nanoparticles (ENPs) are increasingly used in many consumer products and industries due to the novel properties and functionalities of ENPs compared to their larger-sized counterparts (Leitch et al., 2012; Majestic et al., 2010; Bundschuh et al., 2018). Increasing usage of ENPs leads to their release into environmental systems and subsequently increased environmental exposure to ENPs. To determine environmental exposures to ENPs, different methods have been used to detect and quantify the concentration of ENPs in the environment (Montaño et al., 2014; Zhao et al., 2018; Peters et al., 2018; Cheng and Compton, 2014; Benoit et al., 2013). However, the quantification of ENP concentrations in environmental systems remains challenging due to the high background concentrations of natural nanoparticles (NNPs) with similar properties to ENPs and the limited methodologies developed for the quantification of of ENP concentrations in environmental systems (Wiedensohler et al., 2000; Gottschalk et al., 2009; Gottschalk et al., 2013). Single particle-inductively coupled plasma-time of the flight-mass spectrometer (SP-ICP-TOF-MS) is a promising method to differentiate ENPs from NNPs in the environment at the single particle level and to determine their number concentration (Hendriks et al., 2019; Praetorius et al., 2017). However, analyzing the large datasets generated by SP-ICP-TO-MS remains a challenging task due to the large amount of the produced data as well as the high dimensionality of the data, which calls for taking advantage of big-data analytics, such as clustering and dimensionality reduction approaches.

Unsupervised clustering methods such as HC or K-means clustering have been recently implemented to cluster SP-ICP-TOF-MS data (Tokalıoğlu et al., 2018; Song et al., 2018; Mehrabi et al., 2021; Baalousha et al., 2021; Maione et al., 2017; Bi et al., 2014; Tokalioğlu, 2012; Fathinezhad et al., 2020; Wang et al., 2022). However, clustering methods, such as K-means and Density-based spatial clustering of applications with noise (DBSCAN), do not perform well when clustering high dimensional and sparse data such as those generated by SP-ICP-TOF-MS. This is because the optimization search and solutions are often trapped in local optima (Ding et al., 2010). Therefore, novel and complex clustering methods should be considered for the analysis of the SP-ICP-TOF-MS high dimensional data in order to overcome the shortcomings of clustering algorithms such as K-means and DBSCAN (Kriegel et al., 2009; Weber and Robinson, 2016; Esmin et al., 2015).

Spectral Clustering (SC) is one of the most powerful clustering methods for the analysis of high-dimensional data (Ng et al., 2001). Spectral clustering denotes a family of graph-based clustering algorithms which are based on spectral graph theory (Von Luxburg, 2007). Thus, the SC technique can find arbitrarily shaped clusters and optimal solutions (Wang et al., 2015a). The SC technique has been implemented in other fields that require high dimensional data analysis such as single-cell RNA-sequencing analysis data (Park and Zhao, 2018), computer vision (Ochs and Brox, 2012), Lagrangian vortex detection (Hadjighasem et al., 2016), text document clustering (Janani and Vijayarani, 2019), and transportation problems (Banisch and Koltai, 2017), but has not yet been explored for the analysis of SP-ICP-TOF-MS data (Wu et al., 2014; Wang et al., 2015b; Borges et al., 2019). For instance, SC has been applied to cluster single-cell RNA-sequencing data, which are high dimensional and sparse, similar to the SP-ICP-TOF-MS data, to cluster different cell types based on gene expression patterns, and has been shown to outperform other clustering methods such as K-means, Greedy, and FINCH (Qi et al., 2021). In computer vision, SC has been used for super-pixel image segmentation has been shown to outperform five other super-pixel segmentation algorithms (Li and Chen, 2015).

Another approach to improve the quality of high dimensional data clustering is to reduce the data dimensionality through feature reduction techniques such as redundancy maximum relevance (mRMR), principle component analysis (PCA) (Ziasabounchi and Askerzade, 2014), and t-distributed stochastic neighbor embedding (tSNE). Reducing the number of dimensions means finding a subset of determining dimensions or transforming the data to a lower dimension form without reducing the information present in the original data (Chan and Hall, 2010; Yang and Sinaga, 2019; Nguyen et al., 2019; Chen et al., 2014). The mRMR algorithm is a feature selection method that finds a subset of features that are most correlated with the target variable and the least correlated with each other (Ding and Peng, 2005). The PCA is a linear feature transformation algorithm that finds a small set of uncorrelated lines (axes) in a way that when the original data are projected on those lines, most of the variation in the data is captured (Jolliffe, 1990). The mRMR approach has been applied as a feature selection method to create a model for the prediction of active sites of enzymes (Gao et al., 2013). Pinciple component analysis has been used to reduce the dimensions of a heart disease dataset (Ziasabounchi and Askerzade, 2014). While PCA and other linear techniques have been applied successfully, they are typically unable to preserve the local patterns in the high dimensional data, an area in which nonlinear methods excel (Nguyen and Holmes, 2019). tSNE is a nonlinear feature transformation method that maps the high dimensional data into a lower dimensional space, minimizing the differences between the probability distributions of proximity (the probability of a datapoint being close to another datapoint) in the low dimensional, and the same probability distributions in high dimensional (original) space (Van der Maaten and Hinton, 2008). The tSNE has been shown to be able to better identify local data structures and minimize the effect of outliers compared to PCA (Li et al., 2017). The tSNE method has been used as a preprocessing step to visualize and cluster high dimensional data (Li et al., 2017; Alibert, 2019; Kobak and Berens, 2019). The tSNE has also been applied as a preprocessing technique before performing SC on geological data and tSNE outperformed PCA, Kernel PCA, and Locally Linear Embedding (LLE) in extracting more distinct data clusters (Balamurali and Melkumyan, 2016).

Therefore, tSNE is potentially an excellent data reduction technique and SC is potentially an excellent clustering approach for the analysis of SP-ICP-TOF-MS data. This is because tSNE can reveal local data structures and reduce the effect of outliers, and SC can extract arbitrarily shaped clusters and because it has been implemented successfully to extract useful information from similarly complex and sparse data. Therefore, this study aims to evaluate the performance of HC, tSNE-DBSCAN, and SC techniques for the classification of multi-element nanoparticle (mmNP) composition data obtained by SP-ICP-TOF-MS. Clustering performance is defined as identifying clusters with a large number of members of similar/narrow elemental compositions. For detailed information on the interpretation of the extracted mmNP clusters, the reader is referred to our previous publication (Wang et al., 2022).

2. Materials and methods

2.1. Data

A detailed description of the data used in this study, including the sampling sites, sample collection, and preparation for SP-ICP-TOF-MS analysis, can be found elsewhere (Wang et al., 2022). In summary, the elemental composition data of nanoparticles (<1 μm particles) extracted from urban rain collected near the Blossom Street Bridge and urban

runoff samples obtained from the drains of Blossom Street Bridge were acquired using an ICP-TOF-MS instrument (TOFWERK, Thun, Switzerland). This instrument allows the simultaneous quantification of all isotopes within a single nanoparticle (Hendriks et al., 2017). Each replicate was acquired for a duration of 200 s, and the data from three replicates were combined to enable a comprehensive analysis due to limited detection events for certain elements. Data processing, including particle/baseline signal separation and elemental mass calculation, was conducted using Tofware software, as described in previous studies (Loosli et al., 2019; Tanner, 2010). The dataset comprises 15 samples, each containing several thousands of particles.

2.2. Methods

This study employed a two-stage clustering approach to analyze the elemental composition of nanoparticles. The first stage is an intra-sample clustering that aim to seperate mmNPs within each sample into clusters of particles with similar elemental composition using three distinct clustering methods: Hierarchical Clustering (HC), Spectral Clustering (SC), and tSNE-DBSCAN (Fig. S1). These methods differ in their approach to determining the number of clusters. The second stage is an inter-sample clustering that aims to compare the elemental compositions of the clusters generated in the intra-sample clustering across all samples, using HC.

HC relies on a distance cutoff value, determined by visually examining the cluster dendrogram, to identify major clusters with similar elemental compositions in each sample. SC requires the number of clusters to be predefined, and for this study, it was set to match the number of clusters obtained by HC. Relying alone on performance metrics during clustering task, without domain knowledge, may be misleading. For this purpose, first, we used experts' opinion to visually inspect the dendrograms and to select the cutoff value based on expert judgment for identifying the number of clusters. After selecting the cutoff values based on visual inspection, performance metrics has been calculated to test the effect of choosing various cutoff values on HC clustering performance. Silhouette score has been used here as the performance metric for clustering tasks and to examine if the cut off value selected by experts' opinion was actually an optimized value to deliver the best clustering performance. The effect of the number of clusters in SC was also evaluated similarly. In the case of tSNE-DBSCAN, the number of clusters is determined by visually inspecting the tSNE plots of the samples to identify meaningful clusters.

Subsequently, the mean elemental composition of the clusters identified by HC, tSNE-DBSCAN, and SC methods were compared among the different samples. The following sections provide a detailed description of the first stage and the second stage of the analysis.

2.2.1. First stage clustering

2.2.1.1. Hierarchical clustering (HC). Hierarchical clustering was performed using a method described elsewhere (Baalousha et al., 2021). In brief, pairwise correlation distances were calculated between the mmNPs, and each datapoint (*e.g.*, mmNP) was paired with its closest counterpart. This process resulted in the formation of numerous small clusters (consisting of a few members) of mmNPs with similar elemental compositions. Subsequently, these small clusters were grouped together to form larger clusters based on the average correlation distance between them. This iterative process continued until all datapoints were interconnected, forming a hierarchical tree. The final step involved cutting the hierarchical tree using a threshold to identify the major clusters of mmNPs.

Two methods are used in HC to determine the threshold for clustering: prespecifying a distance cutoff value, or setting the maximum number of clusters. In this study, a cutoff value of 0.5 was selected. The algorithm determined the number of clusters based on the specified distance cutoff. Fig. S1-a illustrates the hierarchical clustering process.

2.2.1.2. Spectral clustering (SC). Spectral clustering is a graph-based method that utilizes a similarity matrix calculated from the graph representation of the data (Jia et al., 2014). The stepwise algorithm used to perform spectral clustering is described below:

– The first step involves constructing a similarity matrix. If datapoints are represented as a set of nodes $X = \{x_1, x_2, \ldots, x_n\}$, and the set of edges as E, then the similarity matrix W will be a $N \times N$ matrix, and each element, w_{ij}, is calculated as:

$$w_{i,j} = exp\left(\frac{-d(i,j)^2}{\sigma^2}\right), (i,j) \in E \quad (1)$$

The distance between datapoints j and j, $d(i,j)$ is calculated based on a specified distance and σ is the scaling factor set to 1. Correlation distance is used here to estimate d.

There are three different approaches to form the set of edges: K-nearest neighbor ($K-$NN), maximum radius search, and full graph. In the full graph method, it is assumed that all nodes are connected to each other. In the K-NN method, each node is only connected to its K nearest neighbors, based on a pre-specified parameter K. In the maximum radius search, each node is only connected to other nodes within a specified distance. The similarities and differences between these approaches will be discussed in more detail in the results section.

– In the next step, Laplacian matrix L should be calculated using the similarity matrix W.
– Next, a $n \times k$ matrix V should be calculated, where columns of V are the *kth* smallest eigenvalues of the Laplacian matrix. Here, n denotes the number of datapoints, and k is the predefined number of clusters.
– Finally, using a clustering method, like k-means, the matrix V is clustered into a pre-specified number of clusters where each of the rows in V is considered as a datapoint and the original datapoints are assigned to the same clusters as their corresponding rows in the V matrix. Fig. S1-b shows the flowchart of the SC method.

2.2.1.3. tSNE-DBSCAN. In this method, the number of data dimensions is reduced to two using the t-SNE algorithm and the correlation distance metric. By plotting the transformed data using t-SNE and inspecting the results, natural clusters in the t-SNE space become apparent. To extract these clusters, DBSCAN with the Euclidean distance metric was employed. Once the transformed data were clustered, the corresponding clusters in the original data were extracted. Therefore, this method combines a nonlinear feature reduction technique (t-SNE) with DBSCAN for clustering purposes (Fig. S1-c). The steps involved in this approach are described as follows:

2.2.1.3.1. t-distributed Stochastic Neighbor Embedding (tSNE). t-distributed Stochastic Neighbor Embedding, a variation of Stochastic Neighbor Embedding (SNE) (Hinton and Roweis, 2002), is a statistical method for reducing high dimensional data. The first step in tSNE is to convert the distance, such as correlation distance, between datapoints into conditional probabilities. The probability $p_{j|i}$, the measure of similarity between datapoints, represents the probability of x_i being the neighbor of x_j if the neighbors were selected based on their probability density under a Gaussian distribution. $p_{j|i}$ is calculated as:

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

where σ_i is standard deviation of the Gaussian distribution centered at x_i and is found by a binary search in way that produces a fixed perplexity value predetermined by the user:

$$Perp(P_i) = 2^{-\sum_j p_{j|i} log_2 p_{j|i}} \quad (3)$$

After mapping the datapoints into lower dimensions, the conditional probability can also be calculated for these datapoints in low dimensional space, y_i. However, in tSNE, instead of using a Gaussian distribution a t-Student distribution is used for the low dimensional space and the probability $q_{j|i}$ is calculated as:

$$q_{j|i} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1}} \quad (4)$$

Then, the algorithm tries to minimize the difference between the conditional probabilities $q_{j|i}$ and $p_{j|i}$, by minimizing the Kullback-Leibler divergence (KL divergence) between the two distributions.

2.2.1.3.2. Density-based spatial clustering of applications with noise (DBSCAN). DBSCAN is a density-based clustering algorithm that defines local density as the total number of datapoints in the neighborhood of a given datapoint, denoted as x_i. The general steps of the DBSCAN algorithm are as follows:

- Calculation of the distance between datapoints.
- The algorithm begins by selecting one of the datapoints, x_i, and assigns it to cluster 1.
- The neighboring datapoints should be found. This is done using a predetermined epsilon value which controls the radius of search around the datapoint x_i. Thus, the datapoints that are located within the epsilon distance of x_i become the new neighbors.
- The process continues by iteratively searching for neighbors within the specified radius until no new neighbors can be found.
- Once the first cluster is formed, an unlabeled datapoint is selected as a new point belonging to the next cluster. The same iterative neighbor search steps are then repeated to identify the neighbors of the new cluster, and the process continues until no additional neighbors can be found.
- The entire process is repeated, creating new clusters and searching for neighbors, until there are no more unlabeled points remaining.

2.2.2. Second stage clustering

The objective of the second stage clustering is to conduct cluster analysis on the clusters obtained from the first stage (intra-sample clusters). To perform this analysis, it is essential to identify a suitable representative for each cluster. In this study, the representative of each first stage cluster was defined as the mean mass fraction of the most frequently occurring element within that cluster, considering only elements present in >5 % of the particles in the cluster.

For the three first stage clustering approaches, the HC method was employed to perform the inter-sample clustering, as depicted in Fig. S2. A cutoff value of 0.2 was utilized for all three clusters. The general procedure of the second stage (inter-sample) clustering is illustrated in Fig. S2, highlighting the overall process.

3. Results and discussion

3.1. First stage clustering

Fig. S3 shows the effect of various cutoff values for HC and number of clusters in SC on average Silhouette score using sensitivity analysis. The average silhouette scores for HC and SC tend to be low for low and high cutoff values (Fig. S3a) and for low and high number of clusters (Fig. S3b). For HC, the average silhouette score is highest for a distance cutoff of 0.5 in good agreement with the expert decision based on visual inspection of the dendrograms. Similarly, for SC, the average silhouette score is highest for the number of clusters that matches the number of clusters identified in HC. Therefore, both expert choice and sensitivity analysis of the average silhouette scores can be used to identify clusters for HC and SC.

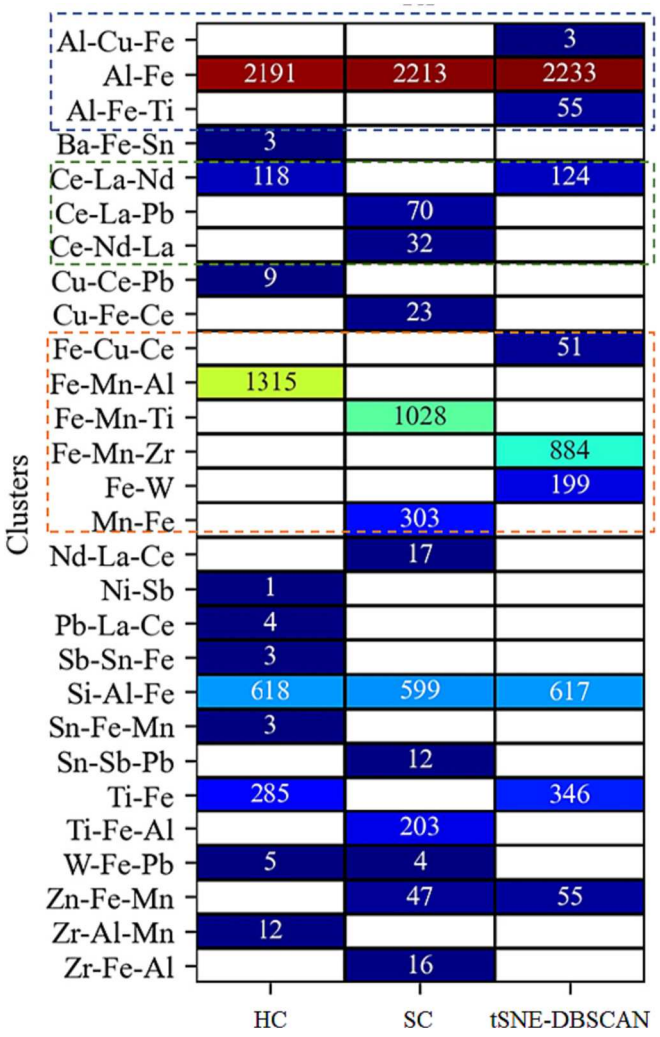


Fig. 1. Number of mmNP particles identified in each mmNP cluster in sample R1 following first stage clustering using the three clustering approaches. HC refers to hierarchical clustering, SC refers to spectral clustering, and tSNE-DBSCAN refers to t-distributed Stochastic Neighbor Embedding coupled with Density Based Spatial Clustering of Applications with Noise.

Fig. S4 presents a heatmap displaying the elemental composition of each member within the first stage clusters, using sample R1 as an example. Hierarchical clustering (HC) yielded clusters with a wide range of member counts, including clusters with a large number of mmNP members as well as clusters with very few members (considered outliers). In contrast, spectral clustering (SC) and tSNE-DBSCAN resulted in clusters with a predominantly large number of mmNP members and avoided the formation of clusters with very few members or outliers. Additionally, SC and tSNE-DBSCAN further divided some of the large clusters identified by HC into subclusters.

Fig. 1 provides a summary of the data presented in Fig. S4, presenting the elemental composition (*e.g.,* mean mass fraction) and the number of mmNP members for the clusters extracted by the three clustering methods in sample R1. Overall, HC extracted five large clusters, each containing over 100 mmNPs (*e.g.,* 118 to 2191), as well as eight clusters with a small number of mmNPs (*e.g.,* 1 to 12). On the other hand, SC identified six clusters with >50 members (*e.g.,* 70 to 2213) and seven clusters with fewer than 50 members (*e.g.,* 4 to 47). In contrast, tSNE-DBSCAN revealed nine clusters with over 50 mmNPs (*e.g.,* 51 to 2233) and only one cluster with three mmNPs. This trend, with HC extracting a few large clusters and other small clusters, tSNE-DBSCAN extracting

medium to large clusters, and SC falling somewhere in between the two methods, was observed consistently across all other samples as well (data not presented).

Further details regarding the differences observed between the clusters extracted by the different methods will be discussed in the following sections for a selected set of representative clusters.

3.1.1. Al-rich cluster

All clustering methods extracted one large Al—Fe cluster, with a mean mass fraction of $Al_{79}Fe_{20}$ by the HC and SC and $Al_{77}Fe_{22}$ by the tSNE-DBSCAN (Fig. 1). The median Al/Fe in these three clusters are 4.58, 4.64, and 4.48, respectively. However, tSNE-DBSCAN extracted two other smaller (*e.g.,* 3 and 55 members) Al—Fe clusters as well, with the mean mass fractions of $Al_{77}Cu_{16}Fe_9$ and $Al_{65}Fe_{20}Ti_{15}$ (Fig. 1). The large $Al_{77}Fe_{22}$ cluster has a mean Ti mass fraction of only 0.0033, while the $Al_{77}Cu_{16}Fe_9$ cluster shows no trace of Ti. Comparatively, in HC and SC, the mean mass fractions of Ti in the large Al—Fe cluster are 0.0069 and 0.0065, respectively. The median Al/Fe in all three Al—Fe clusters extracted by tSNE are 4.48, 7.44, and 3.21.

3.1.2. Fe-rich cluster

HC identified only one large cluster with a mean mass fraction of $Fe_{69}Mn_{18}Al_3$ (1315) (Fig. 1). In contrast, SC identified two clusters with mean mass fractions of $Mn_{61}Fe_{37}$, and $Fe_{79}Mn_{16}Ti_4$ (1028 and 303). The tSNE-DBSCAN method identified three clusters with mean mass fractions of $Fe_{85}W_{15}$, $Fe_{65}Mn_{27}Zr_2$, and $Fe_{49}Cu_{47}Ce_2$ (199, 884, and 51 members). One notable difference between HC and the other clustering methods is that SC and tSNE-DBSCAN were able to extract sub-clusters such as Fe—W and Mn—Fe, which were not identified in HC, suggesting that SC and tSNE-DBSCAN are better than HC in extracting clusters with subtle differences in their elemental composition. This is because tSNE transforms the original data to a lower dimension in which it is easier to extract subclusters with subtle differences in their elemental compositions. Similarly, SC transforms the data to a graph similarity matrix which is less affected by outliers and therefore, identifies sub-clusters with small differences. Decreasing the cutoff values in HC would create a large number of very small clusters since HC first extracts the outliers and then place them in separate clusters instead of extracting sub-clusters that exist within the large clusters.

3.1.3. Ce- and Ti-rich cluster

Both HC and tSNE-DBSCAN methods extracted Ce—La clusters with similar elemental compositions and nanoparticle counts (*e.g.,* 118 $Ce_{46}La_{36}Nd_{14}$ for HC and 124 $Ce_{45}La_{35}Nd_{13}$ for tSNE-DBSCAN, Fig. 1), while SC extracted two different Ce—La clusters with different mass fractions of Nd (70 $Ce_{52}La_{46}Pb_2$ and 32 $Ce_{47}Nd_{29}La_{13}$). This is because tSNE transformation alters the relative distance between datapoints in comparison to the distance between the original data, which can disrupt the global data structure. Therefore, while tSNE makes it easier to differentiate between datapoints within some clusters, it could sometimes make it more difficult to differentiate between other clusters. However, SC is sensitive to small differences between subclusters as well, while maintaining the global data structure intact. The three clustering methods extracted Ti—Fe clusters with similar nanoparticle counts but slightly different elemental compositions (285 $Ti_{66}Fe_{31}$ for HAC, 203 $Ti_{77}Fe_{19}$ for SC, and 346 $Ti_{58}Fe_{40}$ for tSNE-DBSCAN). The slight differences in the mean mass fraction composition of the different clusters are attributed to the differences in the way nanoparticles are clustered (members of each cluster are identified) by the different clustering methods.

To further understand the similarities/differences between mmNPs extracted within each cluster extracted by the different clustering methods, the distributions of Al, Fe, and Ti mass fraction are presented in Fig. S5 for the extracted clusters. The SC and tSNE-DBSCAN differentiated between mmNPs with low and high mass fractions of Al and Fe, whereas HC places them all in one cluster. Nonetheless, tSNE-DBSCAN mixed some of the mmNPs with very low traces of Al in the $Al_{77}Fe_{22}$ cluster (Fig. S5-a), which makes this cluster less consistent, compared to the similar cluster extracted by SC and HC. Fig. S5-b show that HC identified only one Fe cluster with a wide Fe mass fraction distribution ($Fe_{69}Mn_{18}Al_3$). In contrast, tSNE-DBSCAN identified an additional cluster with low traces of W ($Fe_{85}W_{15}$) and SC differentiated between the particles with higher Fe/Mn mass fraction and those with low Fe/Mn mass fraction, leading to the formation of $Mn_{61}Fe_{37}$ cluster. Fig. S5-c shows the distribution of Ti mass fraction in Ti-rich clusters. The Ti-rich cluster extracted by SC has a narrower mass fraction distribution with higher Ti mass fractions compared to those extracted by HC and tSNE-DBSCAN, indicating that the Ti-rich nanoparticle cluster extracted by SC consists of nanoparticles within tighter (more similar) elemental compositions.

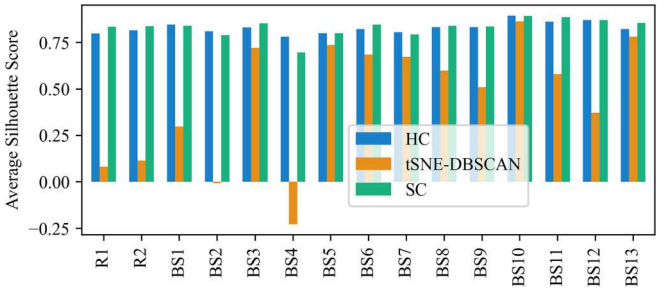


Fig. 2. Average Silhouette score calcualted for the three clustering approaches used in this study.

Fig. 2 shows the average Silhouette score calculated for each clustering method and each sample. In most cases, SC performance is better or similar to that of HC in terms of Silhouette score, and they both outperform (*i.e.,* display a higher Silhouette score) the tSNE-DBSCAN significantly. Additionally, two measures—namely, the coefficient of variation of the mass fractions within each cluster and the mean distance between mmNPs within each cluster—were used to determine the quality/performance of the clustering approaches with respect to the extraction of clusters with consistent (more similar) elemental compositions. The similarity of the elemental compositions of mmNPs in the clusters extracted by different methods is determined by calculating the coefficient of variation of mass fraction for the main elements of each cluster (Fig. 3-a). The coefficient of variation of the mass fractions is generally lower for the clusters extracted by SC than those extracted by HC and tSNE-DBSCAN. This indicates that SC outperforms the other clustering methods in generating clusters of mmNPs with similar elemental compositions. These differences are attributed to how HC, SC, and tSNE-DBSCAN separates clusters and identify cluster members. The tSNE-DBSCAN, while highlighting the local structure of data by separating some clusters with very narrow differences, sometimes groups a number of the mmNPs with very different elemental compositions into one cluster. On the other hand, HC is sensitive to outliers, and it separates outliers into separate clusters. This leads to the placement of the rest of the mmNPs in large clusters that sometimes possess relatively diverse elemental compositions. However, SC using a full graph, accounts for both local and global structures of the data. Thus, SC maintains a balance between separating clusters with small differences and not disrupting the global structure of the data by mixing mmNPs with very different elemental compositions in the same cluster. This is why SC usually extracts clusters with closely tight mmNP elemental compositions.

As another way of comparing the three clustering methods, Al—Fe and Ti—Fe clusters for each sample were tracked and the mean distance between cluster members was calculated (Fig. 3-b and -c). In most cases, the average "within cluster distance" for Al—Fe and Ti—Fe clusters was the smallest for the clusters extracted by SC or tSNE-DBSCAN. However, there were a few samples in which the Ti—Fe clusters extracted by HC

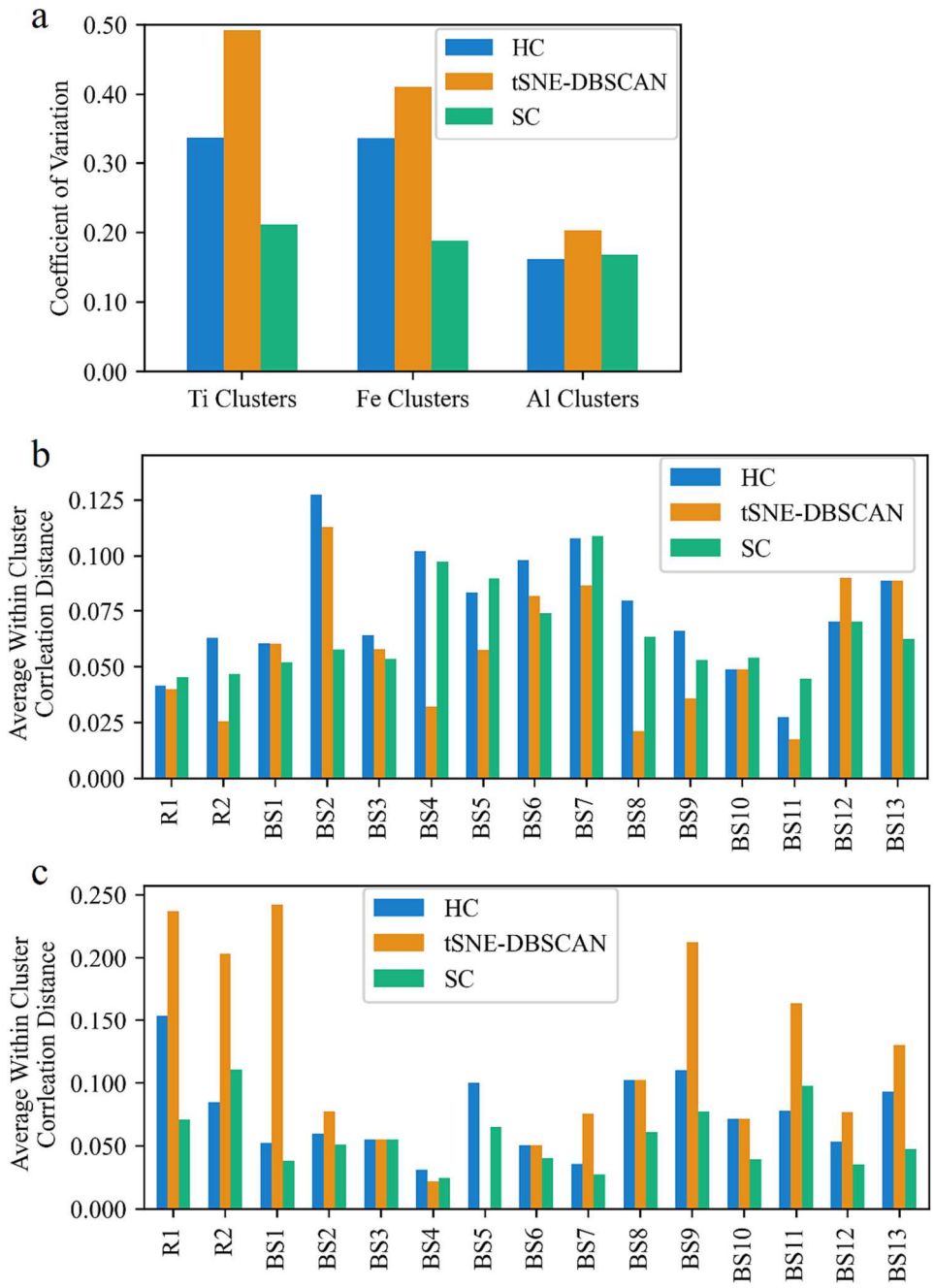


Fig. 3. Measures of the performance of the three clustering approaches used in this study: (a) coefficient of variation of Ti, Fe, and Al mass fraction in their respective main clusters in sample R1, (b) average within cluster correlation distance for Al—Fe clusters, and (c) average within cluster correlation distance of Ti—Fe clusters.

have a smaller average "within cluster distance" compared to the other clustering methods. This is because SC and tSNE-DBSCAN move additional nanoparticles that were not identified by HC into Ti—Fe clusters. In general, SC generated clusters of mmNPs with tighter elemental compositions (less variability in the elemental composition within each cluster) than HC and tSNE.

3.2. Second stage clustering

Merging all the sample data into a single dataset and performing clustering on that merged dataset is typically not feasible due to memory constraints on most computers. Consequently, to compare the elemental composition of different samples, the clustering process was conducted in two stages. In the first stage, clustering was performed individually on each sample. Then, in the second stage, HC was employed to cluster and compare the representatives of the first stage clusters. Fig. S6 illustrates the dendrogram of the cluster representatives generated by each of the three clustering methods in the first stage.

HC classified the first stage clusters obtained from HC, SC, and tSNE-DBSCAN into 34, 37, and 18 major clusters, respectively. Similar to the first stage, the second stage clusters were tracked, and the number of nanoparticles in each cluster is depicted in Fig. 4. Notably, Fig. 4 demonstrates that HC-HC yielded a higher number of very small outlier clusters, which was expected as HC tends to separate outliers and assign them to their own clusters in the first stage. Moreover, differences between the SC-HC and HC-HC methods are primarily observed among the smaller clusters, specifically the identification of Mn-rich clusters in samples R1 and R2 (Fig. 4).

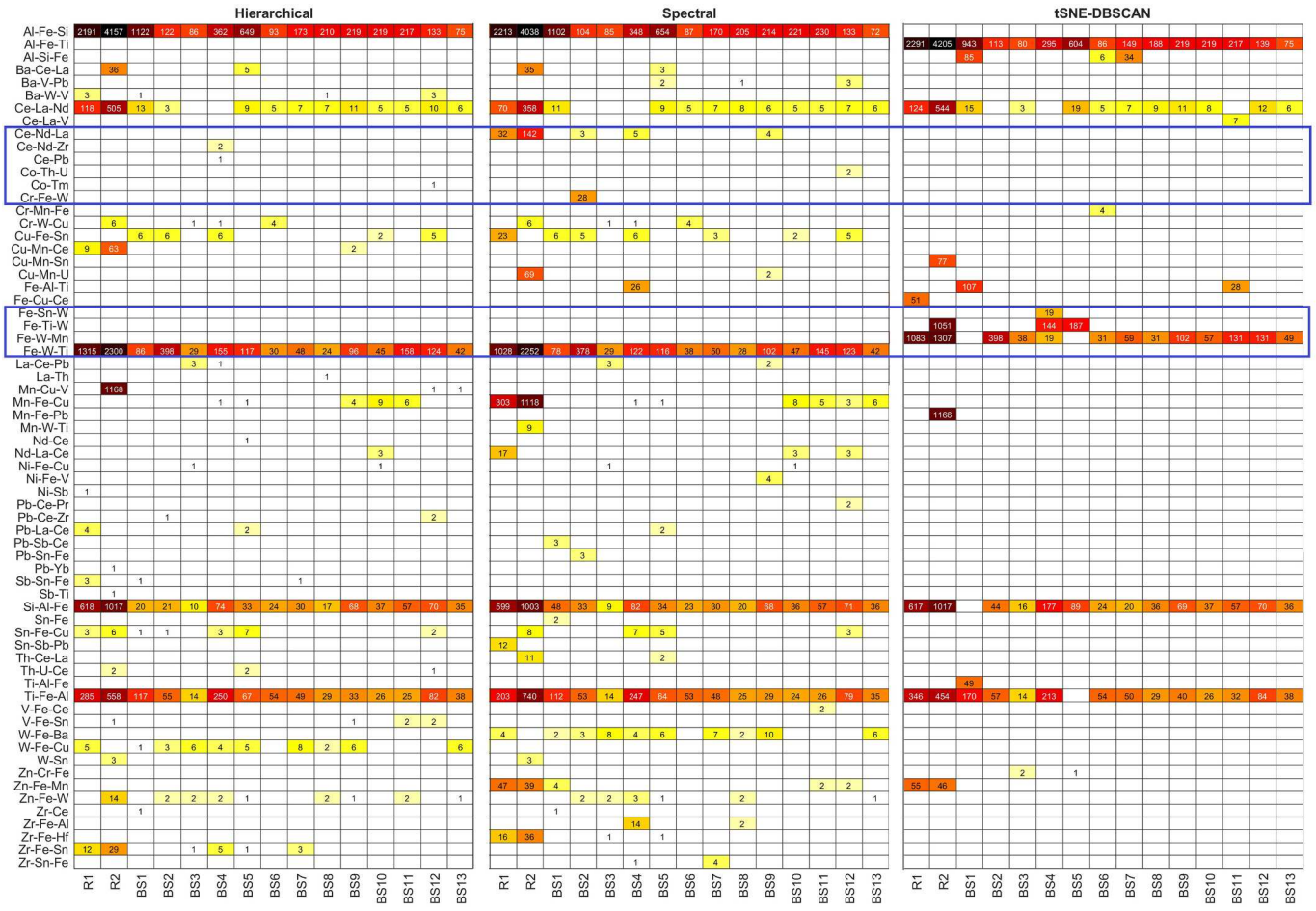


Fig. 4. Number of mmNPs in the clusters identified in all samples following the two stage clustering of mmNPs (left) hierarchical- hierarchical, (middle) spectral-hierarchical, and (right) tSNE-DBSCAN-hierarchical. The second stage hierarchical clustering distance cutoff value was set to 0.2. The corresponding number of mmNPs in clusters identified with a second stage distance cutoff value of 0.05 is presented in Fig. S7.

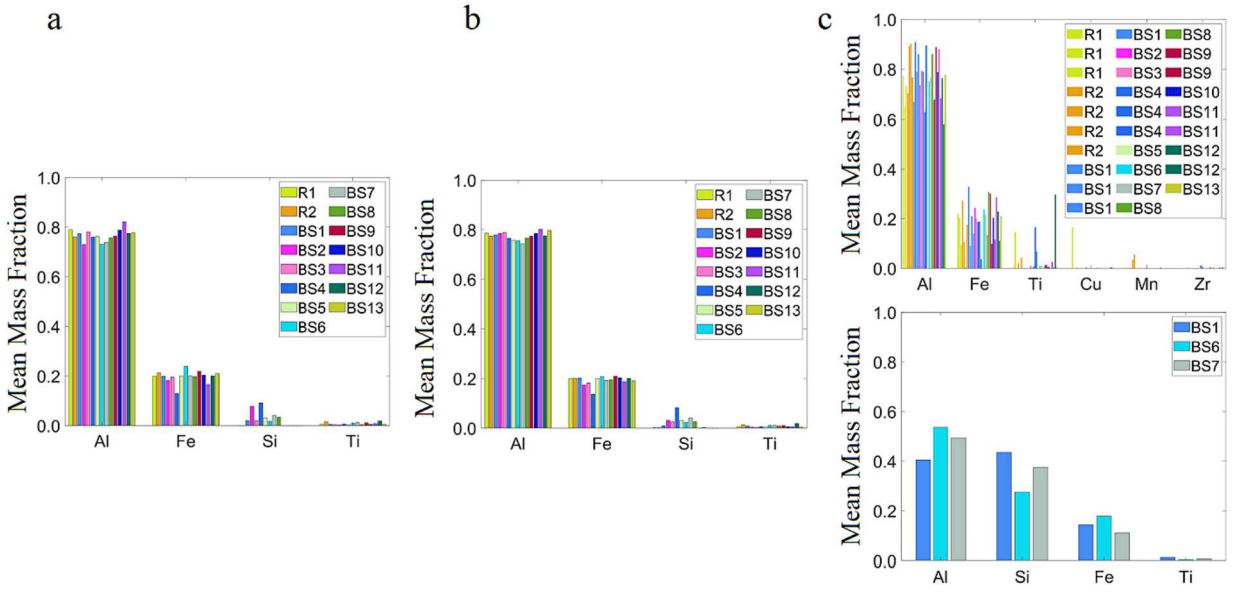


Fig. 5. Mean mass fraction of the major elements in mmNPs in Al-rich cluster extracted by the different clustering methods in all samples: (a) hierarchical-hierarchical, (b) spectral-hierarchical, and (c) tSNE-DBSCAN-hierarchical.

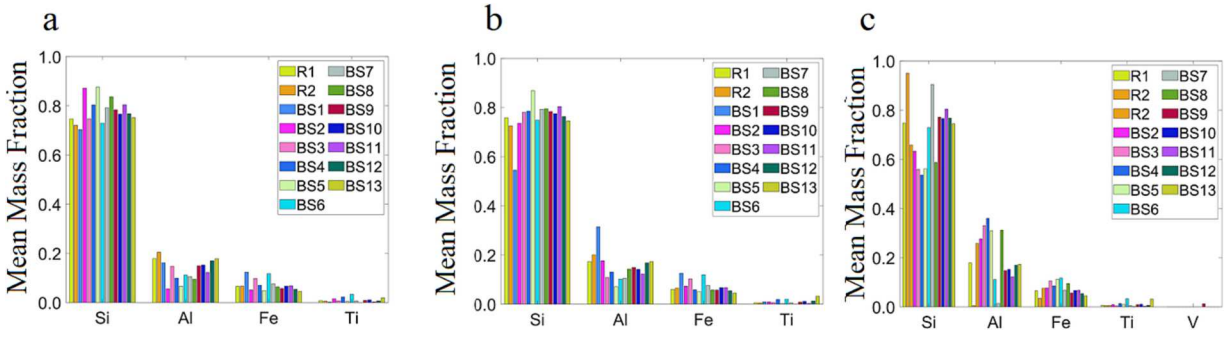


Fig. 6. Mean mass fraction of the major elements in mmNPs in Si-rich cluster extracted by the different clustering methods in all samples: (a) hierarchical-hierarchical, (b) spectral-hierarchical, and (c) tSNE-DBSCAN-hierarchical.

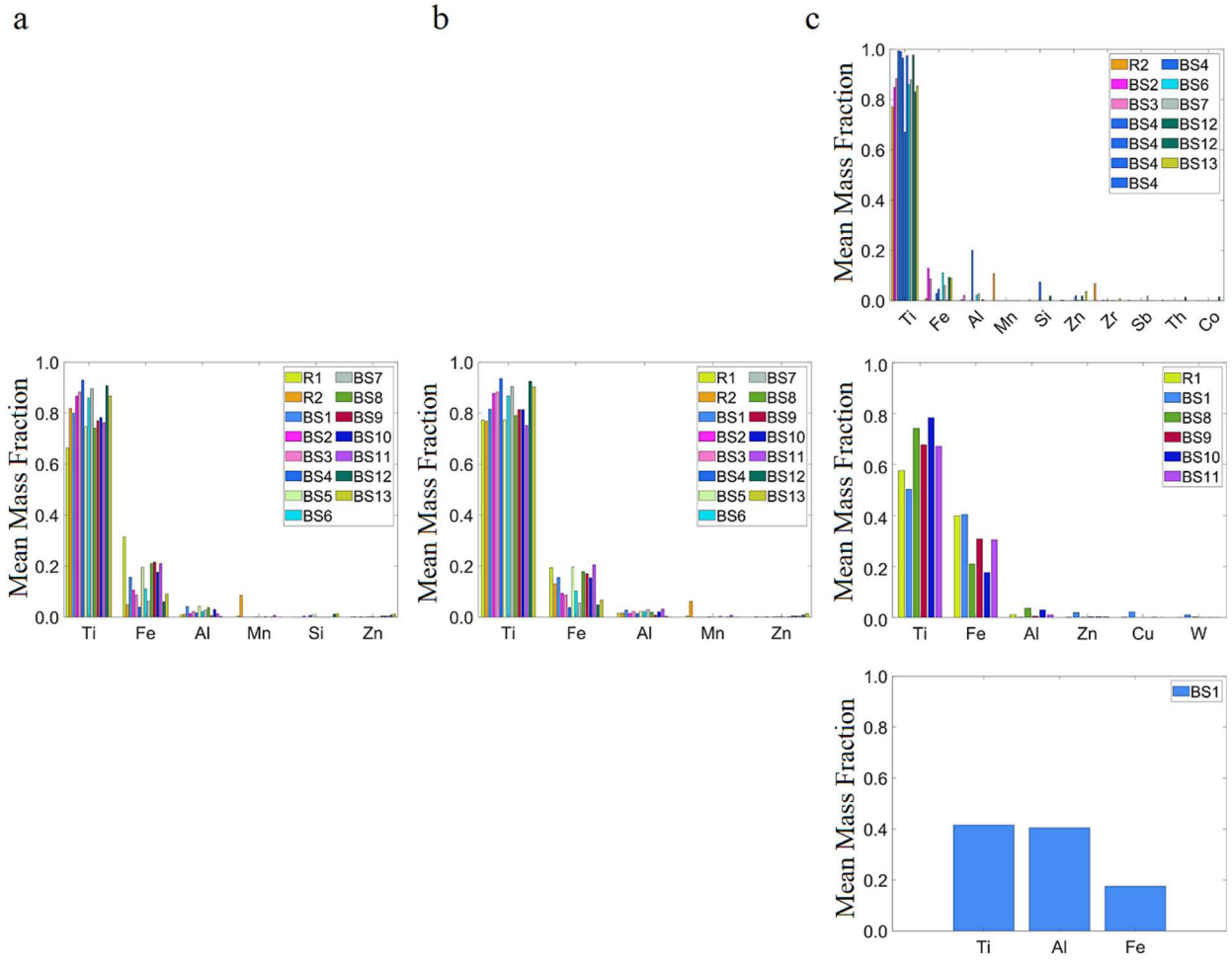


Fig. 7. Mean mass fraction of the major elements in mmNPs in Ti-rich cluster extracted by the different clustering methods in all samples: (a) hierarchical-hierarchical, (b) spectral-hierarchical, and (c) tSNE-DBSCAN-hierarchical.

Most of the mmNPs in all samples are clustered into Al—Fe, Fe—W, Si—Al, and Ti—Fe clusters. Another noticeable mmNP cluster is Ce—La cluster which is almost similar regardless of the first stage method. After the second stage, the final outcome of clustering based on all the methods is relatively similar. This indicates that evaluating the performance of clustering methods and identifying unique clusters in a given sample is more useful after the first stage of clustering on individual samples.

Figs. 5-8 show the mean mass fraction of the four most frequent second stage clusters. The mass fraction plots are very similar for Al, Si, Ti, and Fe clusters which are extracted by HC-HC and SC-HC methods. tSNE-DBSCAN-HC method, however, has extracted a more diverse set of frequent clusters (Al, Fe, Ti, and Si), especially for Al and Fe-rich nanoparticles. tSNE-DBSCAN-HC divided the Al and Fe-rich clusters into a number of smaller clusters compared to HC-HC and SC-HC. This is consistent with the analysis of the first stage results since tSNE extracted clusters such as Fe—W from the Fe-rich clusters. It is worth noting that in the second stage, many of these clusters were merged with the rest of the Fe-rich clusters since their cluster representatives were relatively similar. However, lowering the second stage cut-off (0.05 instead of 0.2) may differentiate some of the mmNP clusters identified in the data as shown in Fig. S7.

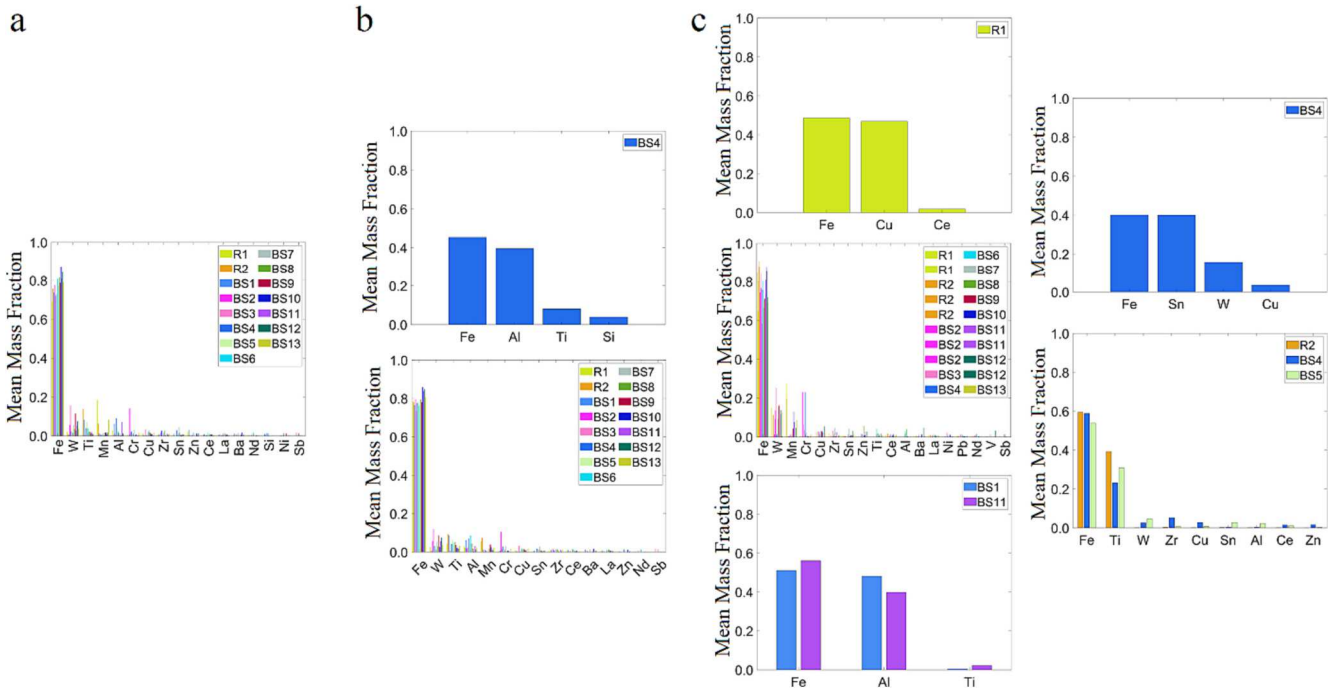


Fig. 8. Mean mass fraction of the major elements in mmNPs in Fe-rich cluster extracted by the different clustering methods in all samples: (a) hierarchical-hierarchical, (b) spectral-hierarchical, and (c) tSNE-DBSCAN-hierarchical.

Even though tSNE-DBSCAN extracted additional clusters from the Al and Fe-rich mmNPs in the first stage, they were merged with other clusters in the second stage. HC-HC extracted similar major clusters as SC-HC, but the rest of the clusters extracted by HC-HC are mostly very small (below 10 nanoparticles) (Fig. 4).

3.3. Overall comparison between the clustering methods

In order to assess the overall performance of the three different clustering approaches used in this study, we focus more on the analysis of the first stage clustering phase. Even though HC-HC and SC-HC methods generated generally similar clusters after the second stage of clustering, SC extracted more consistent large clusters and a higher number of relatively large clusters, while HC extracted many outlier clusters with very few nanoparticles in them. tSNE-DBSCAN identified clusters that existed within the large clusters, which were not detected by the other methods. However, tSNE-DBSCAN also generated several inconsistent clusters (clusters with large variations in elemental composition). It's worth mentioning that after transforming the data using tSNE, the clusters that emerge for the majority of samples have arbitrary shapes (Fig. S8). This fact indicates that clustering methods, such as K-means, which are suitable for identifying spherical-shaped clusters are not appropriate for clustering SP-ICP-TOF-MS data. Instead, an algorithm that can identify arbitrary-shaped clusters is needed. That is the reason DBSCAN was used as the clustering method after tSNE transformation.

While the difference between the three clustering methods is most significant for samples with a higher number of mmNPs, *i.e.*, R1 and R2, SC generally outperformed HC and tSNE-DBSCAN across all samples. The large clusters extracted by SC in most cases possess a much more consistent elemental composition compared to tSNE-DBSCAN. On the other hand, since SC is not as sensitive to outliers as HC, it extracted informative clusters (relatively large and having a consistent elemental composition) not identified by HC. Another important feature of SC is its flexibility. Based on the method used for forming the similarity matrix, *i.e.*, full graph (the method used in this study), maximum radius search, or K-NN, the clustering outcomes may change considerably. Based on several tests we performed, it was found that if a maximum radius search method is used, considering the specified maximum radius of search, the clustering result may become similar to HC. The preliminary tests showed that a full graph generates homogeneous clusters and extracts more information from the data. Also, the K-NN based SC, similar to tSNE-DBSCAN, extracted clusters not extractable by the other methods but generated inconsistent clusters as well. In general, K-NN connects outlier datapoints to their nearest neighbors, and therefore, it increases the probability of separating them from the datapoints which are very similar to them. Considering these issues, we conclude that K-NN is not appropriate for clustering datasets that may contain outliers such as the SP-ICP-TOF-MS data used in this study.

4. Conclusions

Spectral clustering demonstrates higher sensitivity towards the multi-element associations within nanoparticles, allowing the identification of clusters of mmNPs with closer similirities in their elemental compositions to those identified by HC. This enables the differentiation of mmNPs with very similar elemental compositions, even with subtle differences. Transforming the data using tSNE can sometimes lead to inconsistent clusters, as tSNE does not preserve the relative distances between datapoints (mmNPs). Additionally, determining appropriate values for the perplexity hyperparameter of tSNE and the epsilon hyperparameter of DBSCAN can be challenging, as they need to be individually determined for each sample. Incorrect parameter selection can significantly impact the quality of clustering. Similarly, for SC with K-NN or maximum search radius, the selection of hyperparameters greatly affects the clustering results. However, SC using a full graph method is more robust compared to tSNE-DBSCAN and less sensitive to outliers compared to HC. It is worth mentioning that the differences between clustering methods are more pronounced when applied to larger samples. For smaller samples, the extracted clusters tend to be relatively similar across all methods.

Another characteristic of SC is its flexibility in clustering mmNPs based on the chosen method for constructing the similarity matrix. Depending on the application and desired outcomes, the appropriate

method can be selected. In our case, the full graph method was chosen as it provides more informative results by extracting larger clusters with tighter elemental compositions, rather than isolating outlier mmNPs into separate clusters as HC does. However, one limitation of SC is that the number of clusters needs to be predetermined. This can be addressed by initially evaluating the data using HC and determining the number of clusters based on the results or through a trial-and-error process. The SC framework utilized in this study can be extended to other datasets with similar characteristics, as it provides additional insights into the mmNP datasets compared to HC. For detailed information on the interpretation of the extracted mmNP clusters, the reader is referred to our previous publication (Wang et al., 2022).

CRediT authorship contribution statement

Mr. Mahdi Erfani proposed the overall idea of the study, performed all analysis, and wrote the first draft. All authors contributed to the study plan and the manuscript writing and editing.

Declaration of competing interest

The authors declare no competing interest.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by NSF grants (1553909, 1828055, 2101983) from the United States National Science Foundation (NSF).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.167176.

References

Alibert, Y., 2019. New metric to quantify the similarity between planetary systems: application to dimensionality reduction using T-SNE. Astron. Astrophys. 624, 1–10.
Baalousha, M., Wang, J., Erfani, M., Goharian, E., 2021. Elemental fingerprints in natural nanomaterials determined using SP-ICP-TOF-MS and clustering analysis. Sci. Total Environ. 792, 148426.
Balamurali, M., Melkumyan, A., 2016. t-SNE based visualisation and clustering of geological domain. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 565–572, 9950 LNCS.
Banisch, R., Koltai, P., 2017. Understanding the geometry of transport: diffusion maps for lagrangian trajectory data unravel coherent sets. Chaos 27, 1–36.
Benoit, R., Wilkinson, K.J., Sauvé, S., 2013. Partitioning of silver and chemical speciation of free Ag in soils amended with nanoparticles. Chem. Cent. J. 7, 1–7.
Bi, X., et al., 2014. Quantitative resolution of nanoparticle sizes using single particle inductively coupled plasma mass spectrometry with the K-means clustering algorithm. J. Anal. At. Spectrom 29, 1630–1639.
Borges, H., Guibert, R., Permiakova, O., Burger, T., 2019. Distinguishing between spectral clustering and cluster analysis of mass spectra. J. Proteome Res. 18, 571–573.
Bundschuh, M., et al., 2018. Nanoparticles in the environment: where do we come from, where do we go to? Environ. Sci. Eur. 30.
Chan, Y.B., Hall, P., 2010. Using evidence of mixed populations to select variables for clustering very high-dimensional data. J. Am. Stat. Assoc. 105, 798–809.
Chen, H., Tan, C., Lin, Z., Wu, T., 2014. The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. Comput. Biol. Med. 50, 70–75.
Cheng, W., Compton, R.G., 2014. Electrochemical detection of nanoparticles by 'nano-impact' methods. TrAC - Trends Anal. Chem. 58, 79–89.
Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3, 185–205.
Ding, S., Zhang, L., Zhang, Y., 2010. Research on spectral clustering algorithms and prospects. ICCET 2010–2010. Int. Conf. Comput. Eng. Technol. Proc. 6, 149–153.
Esmin, A.A.A., Coelho, R.A., Matwin, S., 2015. A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artif. Intell. Rev. 44, 23–45.
Fathinezhad, M., AbbasiTarighat, M., Dastan, D., 2020. Chemometrics heavy metal content clusters using electrochemical data of modified carbon paste electrode. Environ. Nanotechnology, Monit. Manag. 14, 100307.
Gao, Y.-F., et al., 2013. Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. Mol. Biosyst. 9, 61–69.
Gottschalk, F., Sonderer, T., Scholz, R.W., Nowack, B., 2009. Modeled environmental concentrations of engineered nanomaterials (TiO2, ZnO, ag, CNT, fullerenes) for different regions. Environ. Sci. Technol. https://doi.org/10.1021/es9015553.
Gottschalk, F., Sun, T., Nowack, B., 2013. Environmental concentrations of engineered nanomaterials: review of modeling and analytical studies. Environ. Pollut. https://doi.org/10.1016/j.envpol.2013.06.003.
Hadjighasem, A., Karrasch, D., Teramoto, H., Haller, G., 2016. Spectral-clustering approach to Lagrangian vortex detection. Phys. Rev. E 93, 1–17.
Hendriks, L., Gundlach-Graham, A., Hattendorf, B., Günther, D., 2017. Characterization of a new ICP-TOFMS instrument with continuous and discrete introduction of solutions. J. Anal. At. Spectrom 32, 548–561.
Hendriks, L., Gundlach-Graham, A., Günther, D., 2019. Performance of sp-ICP-TOFMS with signal distributions fitted to a compound Poisson model. J. Anal. At. Spectrom 34, 1900–1909.
Hinton, G.E., Roweis, S., 2002. Stochastic neighbor embedding. Adv. Neural Inf. Process. Syst. 15.
Janani, R., Vijayarani, S., 2019. Text document clustering using spectral clustering algorithm with particle swarm optimization. Expert Syst. Appl. 134, 192–200.
Jia, H., Ding, S., Xu, X., Nie, R., 2014. The latest research progress on spectral clustering. Neural Comput. Appl. 24, 1477–1486.
Jolliffe, I.T., 1990. Principal component analysis: a beginner's guide—I. Introduction and application. Weather 45, 375–382.
Kobak, D., Berens, P., 2019. The art of using t-SNE for single-cell transcriptomics. Nat. Commun. 10.
Kriegel, H.P., Kröger, P., Zimek, A., 2009. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data 3.
Leitch, M.E., Casman, E., Lowry, G.V., 2012. Nanotechnology patenting trends through an environmental lens: analysis of materials and applications. J. Nanopart. Res. 14.
Li, Z., Chen, J., 2015. Superpixel segmentation using linear spectral clustering. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 1356–1363.
Li, W., Cerise, J.E., Yang, Y., Han, H., 2017. Application of t-SNE to human genetic data. J. Bioinform. Comput. Biol. 15, 1–14.
Loosli, F., et al., 2019. Sewage spills are a major source of titanium dioxide engineered (nano)-particle release into the environment. Environ. Sci. Nano 6, 763–777.
Maione, C., et al., 2017. Using cluster analysis and ICP-MS to identify groups of ecstasy tablets in Sao Paulo state, Brazil. J. Forensic Sci. 62, 1479–1486.
Majestic, B.J., et al., 2010. A review of selected engineered nanoparticles in the atmosphere: sources, transformations, and techniques for sampling and analysis. Int. J. Occup. Environ. Health 16, 488–507.
Mehrabi, K., Kaegi, R., Günther, D., Gundlach-Graham, A., 2021. Quantification and clustering of inorganic nanoparticles in wastewater treatment plants across Switzerland. Chimia (Aarau). 75 (642–646).
Montaño, M.D., Badiei, H.R., Bazargan, S., Ranville, J.F., 2014. Improvements in the detection and characterization of engineered nanoparticles using spICP-MS with microsecond dwell times. Environ. Sci. Nano 1, 338–346.
Ng, A., Jordan, M., Weiss, Y., 2001. On spectral clustering: analysis and an algorithm. Adv. Neural Inf. Process. Syst. 14.
Nguyen, L.H., Holmes, S., 2019. Ten quick tips for effective dimensionality reduction. PLoS Comput. Biol. 15, 1–19.
Nguyen, T.T., Krishnakumari, P., Calvert, S.C., Vu, H.L., van Lint, H., 2019. Feature extraction and clustering analysis of highway congestion. Transp. Res. Part C Emerg. Technol. 100, 238–258.
Ochs, P., Brox, T., 2012. Higher order motion models and spectral clustering. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 614–621. https://doi.org/10.1109/CVPR.2012.6247728.
Park, S., Zhao, H., 2018. Spectral clustering based on learning similarity matrix. Bioinformatics 34, 2069–2076.
Peters, R.J.B., et al., 2018. Detection of nanoparticles in Dutch surface waters. Sci. Total Environ. 621, 210–218.
Praetorius, A., et al., 2017. Single-particle multi-element fingerprinting (spMEF) using inductively-coupled plasma time-of-flight mass spectrometry (ICP-TOFMS) to identify engineered nanoparticles against the elevated natural background in soils. Environ. Sci. Nano 4, 307–314.
Qi, R., Wu, J., Guo, F., Xu, L., Zou, Q., 2021. A spectral clustering with self-weighted multiple kernel learning method for single-cell RNA-seq data. Brief. Bioinform. 22 bbaa216.
Song, X., et al., 2018. Multi-element analysis of baijiu (Chinese liquors) by ICP-MS and their classification according to geographical origin. Food Qual. Saf. 2, 43–49.
Tanner, M., 2010. Shorter signals for improved signal to noise ratio, the influence of Poisson distribution. J. Anal. At. Spectrom 25, 405–407.
Tokalioğlu, Ş., 2012. Determination of trace elements in commonly consumed medicinal herbs by ICP-MS and multivariate analysis. Food Chem. 134, 2504–2508.
Tokalıoğlu, Ş., Çiçek, B., İnanç, N., Zararsız, G., Öztürk, A., 2018. Multivariate statistical analysis of data and ICP-MS determination of heavy metals in different Brands of Spices Consumed in Kayseri, Turkey. Food Anal. Methods 11, 2407–2418.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE Laurens. J. Mach. Learn. Res. 9.
Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17, 395–416.
Wang, S., Gu, J., Chen, F., 2015a. Clustering high-dimensional data via spectral clustering using collaborative representation coefficients. Lect. Notes Comput. Sci.

(including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 9226, 248–258.

Wang, W. Chuan, Chau, K. Wing, Xu, D. Mei, Chen, X.Y., 2015b. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resour. Manag. https://doi.org/10.1007/s11269-015-0962-6.

Wang, J., Nabi, M.M., Erfani, M., Goharian, E., Baalousha, M., 2022. Identification and quantification of anthropogenic nanomaterials in urban rain and runoff using single particle-inductively coupled plasma-time of flight-mass spectrometry. Environ. Sci. Nano. https://doi.org/10.1039/d1en00850a.

Weber, L.M., Robinson, M.D., 2016. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytom. Part A 89, 1084–1096.

Wiedensohler, A., Stratmann, F., Tegen, I., 2000. Environmental particles. Particle-Lung Interactions 1.

Wu, S., Feng, X., Zhou, W., 2014. Spectral clustering of high-dimensional data exploiting sparse representation vectors. Neurocomputing 135, 229–239.

Yang, M.S., Sinaga, K.P., 2019. A feature-reduction multi-view k-means clustering algorithm. IEEE Access 7, 114472–114486.

Zhao, B., Yang, T., Zhang, Z., Hickey, M.E., He, L., 2018. A triple functional approach to simultaneously determine the type, concentration, and size of titanium dioxide particles. Environ. Sci. Technol. 52, 2863–2869.

Ziasabounchi, N., Askerzade, I.N., 2014. A comparative study of heart disease prediction based on principal component analysis and clustering methods. Turkish J. Math. Comput. Sci. 16, 18.