Available online at www.sciencedirect.com
ScienceDirect
Atmospheric Environment 42 (2008) 881–892
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

Comparison of two cluster analysis methods using single particle mass spectra

Weixiang Zhao[a], Philip K. Hopke[b,*], Kimberly A. Prather[c]

[a]*Department of Mechanical and Aeronautical Engineering, University of California, Davis, CA 95618, USA*
[b]*Department of Chemical and Biomolecular Engineering, Center for Air Resources Engineering and Science, Clarkson University, P.O. Box 5708, Potsdam, NY 13699, USA*
[c]*Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093-0314, USA*

Received 16 June 2007; received in revised form 14 October 2007; accepted 16 October 2007

Abstract

Cluster analysis of aerosol time-of-flight mass spectrometry (ATOFMS) data has been an effective tool for the identification of possible sources of ambient aerosols. In this study, the clustering results of two typical methods, adaptive resonance theory-based neural networks-2a (ART-2a) and density-based clustering of application with noise (DBSCAN), on ATOFMS data were investigated by employing a set of benchmark ATOFMS data. The advantages and disadvantages of these two methods are discussed and some feasible remedies proposed for problems encountered in the clustering process. The results of this study will provide promising directions for future work on ambient aerosol cluster analysis, suggesting a more effective and feasible clustering strategy based on the integration of ART-2a and DBSCAN.

Keywords: Aerosol time-of-flight mass spectrometer; Single particle; Density-based cluster analysis; Adaptive resonance theory neural networks; Source identification

1. Introduction

Ambient particles have been shown to adversely impact both environmental quality and human health (Dockery et al., 1993), so it is becoming more and more urgent to correctly identify the sources of ambient particles. Aerosol time-of-flight mass spectrometry (ATOFMS) is a novel and effective aerosol analysis technique and its single aerosol particle mass spectrometry data have been widely employed for aerosol source identification via cluster analysis (Song et al., 1999). Among various cluster analysis methods, a type of adaptive resonance theory-based neural networks, called ART-2a, seems the most popular tool for cluster analysis of single particles' mass spectrometry data, because of its adaptive cluster nucleation and expansion process (Fergenson et al., 2001; Zhao et al., 2005). Recently, a novel cluster analysis method based on a different grouping principle, called density-based spatial clustering of application with noise (DBSCAN), was introduced to the cluster analysis of ATOFMS data and compared with ART-2a (Ester et al., 1996; Zhou et al., 2006).

*Corresponding author. Tel.: +1 315 268 3861; fax: +1 315 268 4410.
E-mail address: hopkepk@clarkson.edu (P.K. Hopke).

1352-2310/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2007.10.024

However, without any information on particle origins, the explanation of the clustering results was mainly based on the physical interpretability of each cluster center. Therefore, it was difficult to make a full comparison of these two methods, particularly because of the similarity of the mass spectrometry data of the particles from similar sources and the ambiguity of the categorization of these particles.

In this study, a set of benchmark ATOFMS data (i.e., with a priori known source indexes) of the ambient aerosols from six sources (gasoline emission, diesel emission, biomass burning, coal combustion, sea salt and soil dust) are employed for a complete and convincible comparison of the clustering by these two methods. The goal of this study is to find advantages and disadvantages of these methods, investigate feasible remedies for potential problems of these methods, and provide some suggestions and a feasible starting point for future ATOFMS cluster analyses. However, it cannot be guaranteed that this approach will always perform better than any other method. Specific data sets may require fine tuning of these procedures.

2. Methods

2.1. ART-2a

There are a number of reports involving the use of ART-2a for the cluster analysis of single particle mass spectrometry data (e.g., Song et al., 1999; Bhave et al., 2001; Phares et al., 2001). For ATOFMS data, the inputs of ART-2a usually are the mass spectral data for each particle and the output is the index of the class each particle belongs to. Compared with most clustering methods, the significant advantage of ART-2a is the ability to add a new cluster without disturbing the existing clusters, and thus, it has the potential to be used for on-line data analyses.

The training algorithm for ART-2a is briefly described below. The details are provided in the literature (Carpenter et al., 1991; Zhao et al., 2005).

1. Randomly select an input vector and scale it into unit length.
2. Compare the resonances between the input vector and the cluster vectors of all existing output neurons and determine the neuron with the largest resonance as "winner". The resonance is represented as the dot product of the input vector and the existing cluster vector.
3. If the resonance of the winner neuron is larger than a predefined vigilance factor (VF), ρ_{vig}, modify the cluster vector of the winner neuron toward the input sample vector. Otherwise, create a new cluster for this sample. The modification process of the winner cluster center is

$$\mathbf{v}_{ij} = \begin{cases} \mathbf{r}_{ij} & \text{if } w^{\text{old}}_{(\text{win}),j} > \theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\mathbf{u}_i = \frac{\mathbf{v}_i}{||\mathbf{v}_i||} \tag{2}$$

$$\mathbf{t}_i = \mathbf{w}^{\text{old}}_{\text{win}} + \mu(\mathbf{u}_i - \mathbf{w}^{\text{old}}_{\text{win}}) \tag{3}$$

$$\mathbf{w}^{\text{new}}_{\text{win}} = \frac{\mathbf{t}_i}{||\mathbf{t}_i||} \tag{4}$$

where $\mathbf{r}_{ij}$ is the jth element of the ith normalized training sample vector, $\mathbf{w}^{\text{old}}_{(\text{win}),j}$ is the jth element of the winner cluster center vector, θ is a predefined threshold value, μ is the learning rate, and $\mathbf{w}^{\text{new}}_{\text{win}}$ is the modified winner cluster center vector.
4. Repeat the above steps for all the input vectors, which is defined as one cycle, till reach a stopping criterion.

In ideal cases, the criterion for stopping the training of ART-2a is that the change between the cluster vectors of two consecutive cycles is zero or smaller than a pre-defined threshold value. However, in ATOFMS studies, it is almost impossible to reach the above ideal criteria within a foreseeable time period, so usually a limit is set on the maximal number of cycles. The vigilance factor is a key parameter to control the cluster number. An overly large vigilance would result in an "overly fine" clustering result (the extreme case is one cluster for one aerosol mass spectral sample) by generating many homogeneous small clusters, while an overly small vigilance would result in an "overly coarse" result. There is no general rule to determine a "correct" vigilance value. The initial cluster vectors are randomly selected from the sample set and scaled.

2.2. DBSCAN

Different from many cluster analysis methods including ART-2a, DBSCAN performs a cluster territory expansion process based on the density and continuity of sample distribution. It is a one-step cluster process employing a recursion procedure.

Two parameters, neighbor number (k) and neighborhood radius (ε), control the entire clustering process by examining if a current cluster territory can be further expanded. If not, a new cluster will be generated. The DBSCAN clustering process (Ester et al., 1996) can be briefly described as follows. The details of this algorithm can also be found in Daszykowski et al. (2001, 2002).

1. Initialize the status of all samples to be unprocessed.
2. Randomly choose an unprocessed sample as a current sample, mark it as processed. If the current sample is a "core" (a sample is a core object, if in its neighborhood of radius ε, there are more than k samples), create a new cluster and assign the current sample to it and go to '3'. If not, move to the next unprocessed sample. When all the samples are processed, terminate the algorithm.
3. Find neighbors of the current sample within the distance ε, assign them to the cluster created in step '2', mark as processed, and transfer them to 'seeds' (a set of samples that have a potential of further expansion).
4. For each sample in 'seeds', find its neighbors within the neighbor radius. If they are not processed, add them to 'seeds' and assign them to the same cluster. Continue this expansion process based on the sample density and continuity.
5. When all samples in 'seeds' are processed, go back to '2'.

In ATOFMS cluster analysis, the similarity of two particles is defined as the their dot product rather than the Euclidean distance, so in this study, unless specifically indicated, neighborhood radius (ε) denotes dot product of particle mass spectral vectors. Higher ε values produce smaller neighborhood areas.

2.3. Comparison of ART-2a and DBSCAN

It is clear that these two algorithms represent two different clustering processes. The major difference between them is that DBSCAN not only considers the distance (similarity) between samples but also takes into account the continuity of sample locations in the measured variable space. As a schematic illustration, Fig. 1 shows an extreme case to show the advantages of DBSCAN. Clearly, DBSCAN is able to cluster the samples that have a continuous distribution into one group, while ART-2a divides each group into three small groups based on the sample similarity (distance). The DBSCAN algorithm is a deterministic one-step process. For ART-2a, the initialization, in terms of the order in which the objects are presented to the program, may have some effect on final cluster centers, since convergence may be achieved at slightly different locations depending on the order in which the particles are analyzed.

Because of the possible mixing and aging of the particles during the transportation and the similarity among the particles from different origins, the distributions of the particle data from the same origin could be deformed and the distributions for the particles from various sources could be quite much entangled. The complexity of the ATOFMS data distribution encourages the exploration of various methods for cluster analysis. Zhou et al. (2006) found that DBSCAN is better able to define clusters with varying shapes and sizes that may be

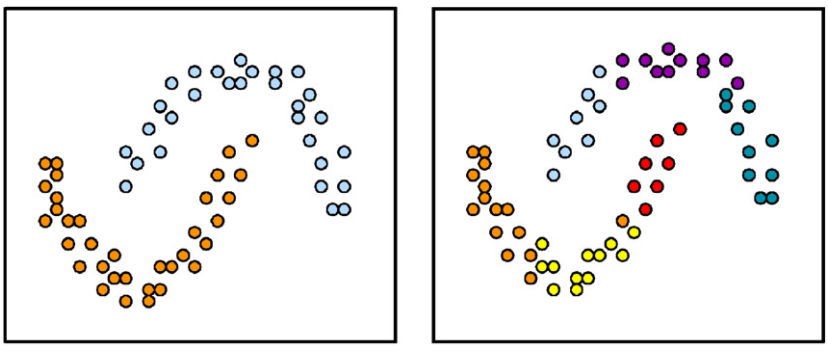

Fig. 1. Schematic illustration of the difference between the clustering principles of ART-2a and DBSCAN.

more representative of real world aerosols. In this study, ART-2a and DBSCAN are applied to a set of benchmark ATOFMS data to see which method is better able to separate the particles into groups that can be associated with specific sources, if there are any remedies for possible problems of these methods, and if these two methods could supplement each other.

3. Sample description

The data employed for this study were a set of benchmark ATOFMS data (i.e., with a priori known source indexes) from six sources (gasoline, diesel, biomass burning, coal combustion, sea salt and soil dust). Each source provided 1600 particle samples and each sample was composed of both positive ions with the charge from +1 to +350 and negative ions with the charge from −1 to −350. The positive and negative ions were concatenated to form a 700 dimensional feature vector.

The gasoline and diesel emission particle samples were randomly selected from the ultrafine particles (aerodynamic diameter (D_a) <100 nm) produced in the dilute exhaust from vehicles operating on a chassis dynamometer. The collection and measurement processes are described by Sodeman et al. (2005) and Toner et al. (2006).

The particles from the other sources were in the 500–1000 nm size range. Biomass and coal burning particles were collected from laboratory tests, whereas dust and sea salt particles were taken from clusters made using ART-2a run on ambient data taken in Trinidad Head, CA, in April 2004.

To ensure that there is comparability in the particle characterization in different studies, rigorous QA procedures are employed in the ATOFMS measurements. Size calibrations are performed on a daily basis with multiple known sizes of polystyrene latex (PSL) particles into the ATOFMS. During ambient measurements, hourly scaling curves are derived to measure particle detection efficiencies by comparing ATOFMS counts to those from external size distribution measurements (APS, SMPS). These comparisons with other measurements test whether the instruments remain stable since the curves are not expected to change over time. In addition, pressures and flow rates in the ATOFMS are monitored continuously. To ensure that the mass spectral sensitivities remain the same from study to study, PSL particle standards of known size are used to generate mass spectra. The intensities are optimized by adjusting the voltage settings on the mass spectrometer so the same absolute ion response is obtained for each study to allow direct comparison of the spectra between studies. These checks insure stable sizes, mass spectral peak intensities, and constant ion transmission as a function of size into the ATOFMS instruments, critical checks to allow comparisons to be made between studies.

4. Results

4.1. Criterion for cluster comparison

The objective of this study is to compare the clustering results of two methods by employing a set of benchmark data. In order to have a quantitative comparison, a reasonable criterion needs to be designed. Table 1 presents a clustering result as an example to demonstrate the criterion used in this study.

It is almost unavoidable to generate some small clusters that contain only a few particles because of the complex and entangled distribution of single particle mass spectra. However, these minor clusters may not be very helpful in characterizing the sources even if they have a very high accuracy (an extreme case for a minor cluster is one particle in one cluster). Therefore, in this study, only the clusters containing at least 80 particles (i.e., 5% of the 1600 sample particles from each source) were included in the comparisons and called "significant" clusters while those containing <80 particles were called "minor" clusters and excluded from the evaluation of clustering result. Then each significant cluster was attributed to one of the six sources according to the categorization of the dominant part of samples in this cluster. For example, the first cluster in Table 1 was attributed to gasoline emission, because 750 of 1198 particles were from gasoline emissions.

Assuming there are n significant clusters that can be attributed to one source, the criterion to evaluate the clustering result on this source can be expressed as

$$S = \sum_{i=1}^{n} \left(\frac{c_i}{L_i} \frac{c_i}{1600} \right) \tag{5}$$

where S is the score of the clustering result on this source, c_i is the number of matched samples of the

Table 1
Significant clusters obtained by ART-2a (vigilance factor = 0.5)

Cluster index	No. of samples	No. of samples matched to each source					
		Gas	Diesel	Biomass	Sea salt	Coal	Soil dust
1	1198	750	143	105	0	200	0
2	932	0	12	3	870	47	0
3	843	41	6	4	3	55	734
4	735	129	2	24	0	578	2
5	631	1	0	1	0	0	629
6	523	1	1	499	1	18	3
7	501	0	3	426	0	71	1
8	473	62	405	3	2	1	0
9	391	3	2	294	0	89	3
10	389	2	0	1	386	0	0
11	299	30	17	43	1	205	3
12	297	6	291	0	0	0	0
13	290	161	21	1	0	107	0
14	288	25	263	0	0	0	0
15	170	0	1	1	167	1	0
16	143	1	0	0	1	16	125
17	137	28	37	4	1	65	2
18	106	105	1	0	0	0	0
19	97	0	1	92	0	4	0
20	96	21	52	14	0	9	0
21	85	18	67	0	0	0	0
22	83	0	1	2	1	79	0

dominant source for cluster i, L_i is the number of all the samples in cluster i, 1600 is the sample number of each benchmark source. c_i/L_i is the direct accuracy of cluster i, and $c_i/1600$ is the weight for this direct accuracy. Thus, a cluster containing more matched dominant particles makes a greater contribution to the accuracy for this observed source. Clearly, the ideal highest score for a source is "1" that corresponds to a single cluster or a number of significant clusters that cover all the particles from one source and do not contain any particles from any of the other sources. The objective of this paper is to find a method to make the clustering accuracy as close to "1" as possible. As an example, the clustering accuracy of the soil dust source in Table 1 can be calculated by summing up $734/843 \times 734/1600$ (cluster 3), $629/631 \times 629/1600$ (cluster 5) and $125/143 \times 125/1600$ (cluster 16), which equals 0.86. Similarly, applying this criterion to the soil dust source in Table 2 (VF = 0.9) can give us a clustering accuracy of 0.17 by summing up $101/101 \times 101/1600$ (cluster 5), $84/84 \times 84/1600$ (cluster 12), and $81/81 \times 81/1600$ (cluster 13). The reason for the sharp decrease of the soil dust will be discussed later and this type of poor result will be termed a "crashed" cluster result. The cluster result for the whole system (i.e., for the six sources) is the mean value of the scores for the six sources.

Because it is almost unavoidable that the particles from one source will group into multiple clusters, the above criterion does not pay particular attention to the potential difference caused by the number of significant clusters. For example, the result in which 1600 samples are grouped into two clusters and that in which 1600 sample are grouped in three clusters are considered to represent the same quality of result.

4.2. Clustering results for ART-2a

Fig. 2 shows the clustering results of ART-2a. The corresponding system average accuracies for different vigilance factors (from 0.4 to 0.9) in Fig. 2 are 0.63, 0.65, 0.65, 0.61, 0.47, and 0.14, respectively. It can be seen that ART-2a with the vigilance factors being 0.5 and 0.6 yield the best clusters. The sources of soil dust, coal combustion, and biomass burning show the highest accuracies. One possible reason could be that these three sources have some dominant and stable inorganic or metal ions that

Table 2
Significant clusters obtained by ART-2a (vigilance factor = 0.9)

Cluster index	No. of samples	No. of samples matched to each source					
		Gas	Diesel	Biomass	Sea salt	Coal	Soil dust
1	135	26	0	1	0	108	0
2	120	0	0	120	0	0	0
3	104	0	0	104	0	0	0
4	102	0	0	99	0	3	0
5	101	0	0	0	0	0	101
6	99	1	0	0	98	0	0
7	98	0	0	0	98	0	0
8	92	0	0	0	92	0	0
9	92	0	0	0	92	0	0
10	88	0	1	86	0	1	0
11	86	0	0	0	86	0	0
12	84	0	0	0	0	0	84
13	81	0	0	0	0	0	81
14	80	0	0	0	80	0	0

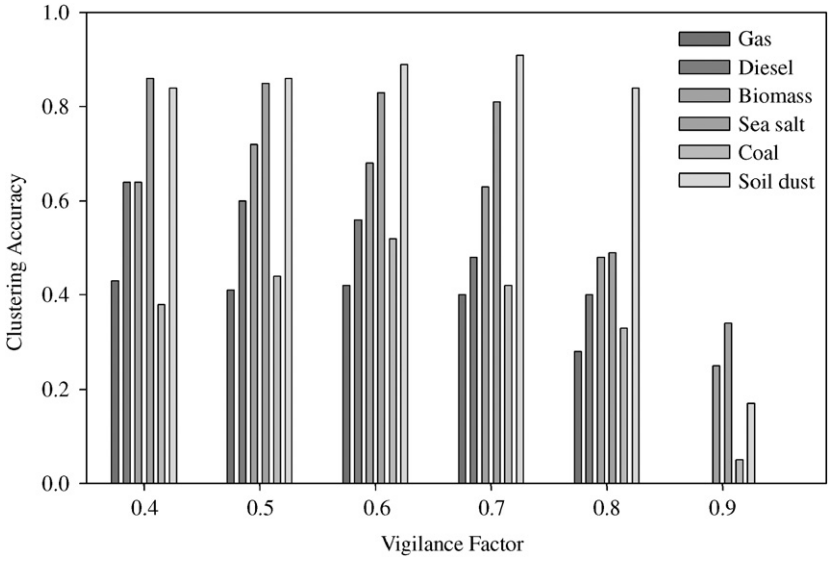

Fig. 2. Clustering efficiency of ART-2a (the system average accuracies for different vigilance factors (from 0.4 to 0.9) are 0.63, 0.65, 0.65, 0.61, 0.47, and 0.14, respectively).

can relatively easily distinguish these sources from the others. For other sources like coal combustion, as shown in Table 1, these clusters only cover about half of the total 1600 samples although some of their significant clusters have good direct accuracies. Therefore, the accuracies for these sources are just ~50% for the case of the vigilance factor being 0.5 or 0.6.

The clustering results decreased significantly with an increase of the vigilance factor (such as 0.8 and 0.9). Table 2 shows the results of ART-2a (VF = 0.9). It can be seen from this table that the significant clusters only cover a small fraction of the particles and most particles are grouped into minor clusters that are not qualified to be evaluated. The sources of gasoline and diesel emission do not have any significant clusters, so their accuracies are zero. Clearly, an overly high vigilance factor could result in a "crashed" clustering result that does not represent a good cluster result since all of the particles from a single source are expected to fall into a limited number of clusters.

4.3. Regrouping analysis for ART-2a

In ART-2a, a sample that does not belong to any existing cluster is classified into a new cluster. During the following training process, this new cluster will only expand but never be merged with its neighbor cluster, even if the gradually modified center of this new cluster is very similar to its neighbor cluster (i.e., the dot product of these two cluster centers is larger than the vigilance factor). Therefore, some clusters generated by ART-2a could have a significant overlap between their sample distribution spaces, and this problem could be more severe in the cases with high vigilance factors.

A possible remedy for this problem is to regroup ART-2a clusters with the same vigilance factor. In regrouping analysis, any two cluster centers whose similarity (dot product) is larger than the vigilance factor are merged to form a new cluster center using the number of samples in each cluster as the weights. The newly formed cluster center is compared with other centers and the whole regrouping process continued until there are no overly similar (i.e., dot product > vigilance factor) cluster centers. Then, all the particles are re-matched to all the new cluster centers based on the same vigilance factor. The results of the regrouping analysis in this study are shown in Fig. 3 and the corresponding average system clustering accuracies are 0.59 (VF = 0.6), 0.59 (VF = 0.7), 0.62 (VF = 0.8), and 0.49 (VF = 0.9), respectively.

Comparison with Figs. 2 and 3 shows that the regrouping analysis significantly improved the clustering results for ART-2a with relatively high vigilance factors (like 0.8 and 0.9) by regrouping small and minor clusters. As an example, Table 3 shows the regrouping analysis results of VF = 0.9. It can be seen from the comparison with Table 2 that the regrouping analysis reunited the small and minor clusters into bigger and significant clusters. This reunion resulted in an increase in the clustering accuracy.

However, the regrouping analysis seems to have an adverse effect on the results of lower vigilance factors. For the vigilance factor (VF) = 0.6 (Fig. 3), the regrouped cluster pattern for coal combustion particles sharply decreased from the cluster result of the ART-2a alone. One possible reason is that the regrouping analysis combined some coal combustion clusters with some clusters of other similar sources because of their similar distribution spaces. The regrouping analysis provides a feasible improvement for the ATOFMS cluster analysis and shows the ability to recover from "crashed" cluster results caused by a high vigilance factor.

4.4. Clustering results for DBSCAN

DBSCAN uses two key parameters to control its clustering behavior. Neighborhood radius (ε) performs similarly to the vigilance factor for ART-2a

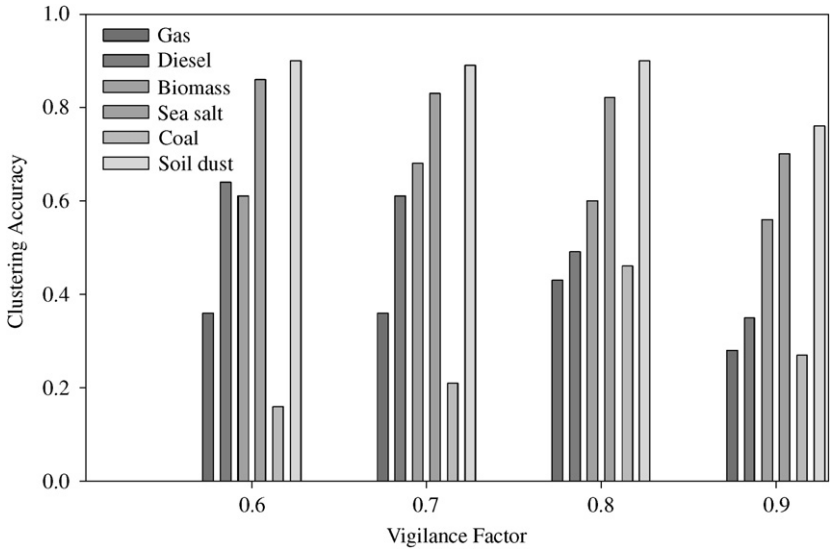

Fig. 3. Clustering efficiency of ART-2a integrated with regrouping analysis (the system average accuracies for different vigilance factors (from 0.6 to 0.9) are 0.59, 0.59, 0.62, and 0.49, respectively).

Table 3
Significant clusters obtained by ART-2a integrated with regrouping analysis (vigilance factor = 0.9)

Cluster index	No. of samples	No. of samples matched to each source					
		Gas	Diesel	Biomass	Sea salt	Coal	Soil dust
1	632	471	79	49	0	33	0
2	536	4	0	1	0	0	531
3	433	75	0	14	0	344	0
4	382	0	0	0	382	0	0
5	288	0	0	0	0	0	288
6	281	0	0	269	0	12	0
7	274	29	244	0	1	0	0
8	250	0	1	225	0	24	0
9	230	0	0	229	0	0	1
10	227	151	7	1	0	68	0
11	199	0	0	0	199	0	0
12	190	0	190	0	0	0	0
13	182	0	0	0	182	0	0
14	165	0	0	0	0	0	165
15	155	0	0	0	0	0	155
16	148	0	148	0	0	0	0
17	143	1	0	0	142	0	0
18	133	0	0	0	133	0	0
19	123	0	0	120	0	3	0
20	116	1	0	1	0	114	0
21	112	14	2	28	1	67	0
22	88	0	0	0	88	0	0
23	83	0	0	83	0	0	0
24	80	0	0	0	0	0	80

Table 4
Clustering effect of DBSCAN

ε	k	Clustering accuracy						
		Gas	Diesel	Biomass	Sea salt	Coal	Soil dust	Average
0.7	400	0.05	0.52	0	0.83	0.05	0.24	0.28
0.7	600	0.44	0.58	0.43	0.83	0	0.83	0.52
0.7	800	0.39	0.58	0.46	0.83	0	0.83	0.51
0.8	100	0.27	0.17	0.69	0.82	0	0.9	0.48
0.8	300	0.37	0.46	0.65	0.8	0.38	0.85	0.58
0.8	500	0.36	0.44	0.61	0.76	0.25	0.85	0.55
0.9	20	0.05	0.35	0.63	0.83	0.2	0.8	0.48
0.9	30	0.3	0.29	0.62	0.81	0.31	0.87	0.53
0.9	50	0.3	0.23	0.53	0.8	0.27	0.87	0.50

and neighbor number (k) helps check the density and continuity of sample distribution space.

The clustering accuracies over a range of ε and k values are shown in Table 4. For each given neighbor radius (ε), the clustering accuracies for most sources and the entire system increased and then they decreased as the neighbor number (k) increased further. A small value of neighbor number (k) could result in a mixture of clusters from different sources while a large value of k could lead to an overly fine or even "crashed" clustering result because of an overly strict criterion. Both cases

could influence the clustering results. In addition, the examination of the best system clustering result for each given ε shows that the highest system accuracy is obtained for $\varepsilon = 0.8$ that suggests that ε appears to have a more important role in controlling the clustering than the neighbor number (k).

The best clustering result of DBSCAN in this figure is obtained when ε and k equal 0.8 and 300, respectively, but it is still poorer than the best results of the ART-2a-based methods. Therefore, it could be inferred that the distribution of ATOFMS objects does not resemble the schematic case in Fig. 1. Thus, DBSCAN does not provide as substantial advantages for ATOFMS data as had been suggested by Zhou et al. (2006).

4.5. Clustering results for ART-2a integrated with DBSCAN

Fig. 3 indicates that the "crashed" clustering in the ART-2a analysis with a high vigilance factor can be recovered by employing a regrouping process that suggests investigating if DBSCAN has a similar (even better) function to recover the "crashed" clustering results for ART-2a. The assumption for this experiment is if the centers of some clusters can be grouped based on the density and continuity by DBSCAN, the samples in these clusters should also be grouped. The results obtained by ART-2a with the vigilance factors being 0.7, 0.8, and 0.9 were used to check the effect of the integration of ART-2a with DBSCAN.

Table 5 shows the results of this integration strategy. The neighbor radius (ε) values were set to be the same as the vigilance factor for ART-2a, as they have a similar function. The clustering efficiencies in Table 5 are all significantly better than the corresponding results of the sole ART-2a, demonstrating the result of DBSCAN on reuniting the small (overly fine, even "crashed") clusters generated by ART-2a with high vigilance factors. Moreover, the best clustering result for each vigilance factor (or neighbor radius) in this table are better than the corresponding result of the ART-2a integrated with regrouping analysis.

In addition, for each neighbor radius (ε), the clustering accuracies of most sources and the entire system first are increased along with the increase on the neighbor number (k) but are not significantly changed after k reaching a level. (For example, for the case of VF = 0.8 and $\varepsilon = 0.8$, there is no salient change of the clustering accuracy while k gradually increasing from 20 to 80.) This changing trend could indicate that (1) a small k value could increase the clustering result through cluster re-union, but it could result in excessively reunited clusters, and (2) an increase of k value could gradually solve the reuniting problem ultimately yielding a better clustering result. This table also shows that various ε values have the different best clustering accuracies. Assuming that the current search ranges for k cover the most of the areas in which clustering accuracy would significantly change, this result further supports the conclusion that ε has a more dominant role in controlling cluster effect than k.

All the results in Table 5 indicate that DBSCAN may be a more effective and robust post-processing strategy for ART-2a with high vigilance factors than regrouping analysis.

Table 5
Clustering effect of ART-2a integrated with DBSCAN

VF	ε	k	Clustering accuracy						
			Gas	Diesel	Biomass	Sea salt	Coal	Soil dust	Average
0.7	0.7	3	0.07	0.24	0.7	0.84	0.25	0.84	0.49
0.7	0.7	10	0.41	0.55	0.71	0.84	0.43	0.89	0.64
0.7	0.7	50	0.41	0.55	0.71	0.84	0.43	0.89	0.64
0.8	0.8	5	0.14	0.25	0.7	0.83	0.32	0.88	0.52
0.8	0.8	20	0.42	0.52	0.72	0.83	0.46	0.91	0.64
0.8	0.8	80	0.42	0.52	0.72	0.83	0.46	0.91	0.64
0.9	0.9	5	0.31	0.41	0.61	0.72	0.12	0.85	0.50
0.9	0.9	10	0.39	0.37	0.55	0.72	0.36	0.76	0.52
0.9	0.9	150	0.39	0.36	0.55	0.72	0.36	0.73	0.52

5. Discussion

Four methods have been employed to test the clustering results on the ATOFMS data. In order to review and evaluate the results of these methods, a principal component analysis was performed on the ATOFMS data employed in this study. The first three principal components are shown pairwise in Fig. 4 although they cannot completely and directly reflect the distribution of the ATOFMS samples. It can be seen that soil dust and sea salt are relatively well separated while the distribution of the other four sources overlap and are mutually quite much entangled.

The distributions shown in Fig. 4 support the following results of this study: (1) all the methods give the best clustering results for the soil dust and sea salt sources while the effects on the other sources are just about 50% or even lower; (2) DBSCAN does not show its specific advantages, probably because the highly overlapped distributions of the different sources could lead DBSCAN in the wrong direction (i.e., the space of a different source); and (3) none of the methods shows a significantly superior clustering result on the entire system composed of six sources.

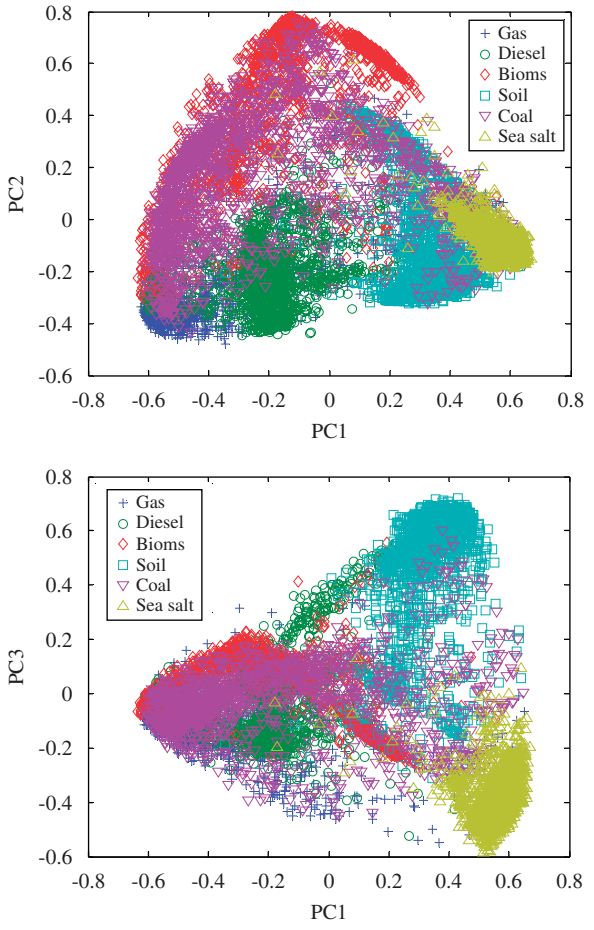

Fig. 4. Distribution of the first three principal components of the particle samples from the six sources (*top*: the first component vs. the second component; *bottom*: the first component vs. the third component).

DBSCAN is a one-step territory expansion process while ART-2a is an iterated process in which all the samples need to be compared with the cluster centers generated in each cycle. Therefore, DBSCAN could use (much) less computation time than ART-2a. Regrouping analysis is a straightforward cluster merging process without any iteration steps, so there is no significant difference between the regrouping analysis and DBSCAN in terms of computation time when they are used to post-process ART-2a preliminary results.

Although it is almost impossible to find a set of fixed parameters or rules that could be directly generalized to other studies, some suggestions are summarized below for a wide spectrum of applications of the methods investigated in this study. No matter whatever methods are used, one important key for a cluster analysis of ATOFMS data without known origins is the interpretability of the cluster centers. Interpretability is unfortunately subjective. For example, depending on the users, either a result in which gasoline and diesel emissions are clustered into two separated groups or a result in which they are grouped into a single class called mobile emissions could be acceptable.

For ART-2a, both a relatively small vigilance factor (like VF = 0.5 or 0.6 in this study) and a relatively high vigilance factor (like VF = 0.8 in this study) followed by a regrouping analysis show the almost equivalent clustering. A similar result in which small vigilance factors (VF = 0.4 and 0.5) yielded an interpretable clustering result for 50,000 or so ATOFMS samples collected in NYC was reported in Zhou et al. (2006). However, as discussed before, this current study also shows there is still some similarity (over the predefined vigilance factor) between the cluster centers obtained with the small vigilance factor that may indicate that solely regrouping analysis could seem to be necessary for both high and low vigilance factors to overcome possible over-similarity among the generated cluster centers. Therefore, considering the regrouping analysis could make the result obtained

with a small vigilance factor worse, it is suggested to use a relatively high vigilance factor followed by regrouping.

For DBSCAN, the neighbor radius (ε) has a similar function to the vigilance factor in ART-2a, so it is suggested to begin with a relatively high value like 0.7 or 0.8. The clustering of DBSCAN is based on the co-action of the neighbor radius (ε) and the neighbor number (k). As shown in Table 4, a larger ε value makes a sample (object) have fewer neighbors, so a relatively smaller k value may be required for good clustering.

Since DBSCAN did not show its advantages when it was directly applied to the ATOFMS data in this study, feasible suggestions will be provided to use DBSCAN as a post-processing approach for the ART-2a results. First, continue using the vigilance factor for ART-2a as the ε value for DBSCAN. As discussed in Section 4 of the ART-2a and DBSCAN-integrated method, after k reaches some level, further increases in k do not significantly increase the clustering accuracy. Therefore, it seems to be a good choice to select a relatively large value (like one-tenth of the cluster number of ART-2a or even higher) as the k value for post-processing.

Various evaluation criteria could show slightly different comparison results of clustering results. The criterion used in this study fairly evaluated the clustering of each method, considering both the direct accuracy of each cluster and the modification weight based on the percentage of the matched dominant samples over the all 1600 samples of that matched source.

6. Conclusions

The benchmark samples employed in this study cover some major sources of ambient aerosols and reflect the real spatial distribution of most particles from these sources, so the results of this study could be applied to other real ATOFMS cluster analyses. This study examined the clustering results of various methods by employing a set of benchmark ATOFMS data. The ART-2a-based approaches show better initial clustering results than the usage of DBSCAN alone. The territory expansion principle of DBSCAN largely prevented the overlap of generated cluster spaces, but it does not show its specific advantages in this study because of the complex and entangled distribution spaces of the six sources. A proper vigilance factor can produce a reasonable ART-2a clustering result, but an overly fine or "crashed" clustering result for ART-2a with a high vigilance factor can be recovered by a post-processing strategy. DBSCAN seems to be more effective and robust in post-processing than the regrouping analysis. Thus, a recommended approach is to begin the analysis with ART-2a with a relatively high vigilance factor (>0.7) and regroup the clusters using DBSCAN.

Acknowledgments

This work was supported by the US Environmental Protection Agency under Science to Achieve Results (STAR) grant number R831083. Although the research described in this article has been funded by the US EPA, the views expressed herein are solely those of the authors and do not represent the official policies or positions of the US EPA.

References

Bhave, P.V., Fergenson, D.P., Prather, K.A., Cass, G.R., 2001. Source apportionment of fine particulate matter by clustering single-particle data: tests of receptor model accuracy. Environmental Science and Technology 35, 2060–2072.

Carpenter, G.A., Grossberg, S., Rosen, D.B., 1991. ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition. Neural Networks 4, 493–504.

Daszykowski, M., Walczak, B., Massart, D.L., 2001. Looking for natural patterns in data. Part 1. Density-based approach. Chemometrics and Intelligent Laboratory Systems 56, 83–92.

Daszykowski, M., Walczak, B., Massart, D.L., 2002. Representative subset selection. Analytica Chimica Acta 468, 91–103.

Dockery, D.W., Pope, C.A., Xu, X.P., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B., Speizer, F.E., 1993. An association between air-pollution and mortality in 6 United-States cities. New England Journal of Medicine 329, 1753–1759.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

Fergenson, D.P., Song, X., Ramadan, Z., Allen, J.O., Hughes, L., Cass, G.R., Hopke, P.K., Prather, K.A., 2001. Quantitation of ATOFMS data using multivariate methods. Analytical Chemistry 73 (15), 3535–3541.

Phares, D.J., Rhodes, K.P., Wexlere, A.S., Kane, D.B., Johnston, M.V., 2001. Application of the ART-2a algorithm to laser ablation aerosol mass spectrometry of particle standards. Analytical Chemistry 73, 2338–2344.

Sodeman, D.A., Toner, S.M., Prather, K.A., 2005. Determination of single particle mass spectral signatures from light-duty vehicle emissions. Environmental Science and Technology 39 (12), 4569–4580.

Song, X., Hopke, P.K., Fergenson, D.P., Prather, K.A., 1999. Classification of single particles analyzed by ATOFMS using

an artificial neural network, ART-2A. Analytical Chemistry 71 (4), 860–865.

Toner, S.M., Sodeman, D.A., Prather, K.A., 2006. Single particle characterization of ultrafine and accumulation mode particles from heavy duty diesel vehicles using aerosol time-of-flight mass spectrometry. Environmental Science and Technology 40 (12), 3912–3921.

Zhao, W., Hopke, P.K., Qin, X., Prather, K.A., 2005. Predicting bulk ambient aerosol compositions from ATOFMS data with ART-2a and multivariate analysis. Analytica Chimica Acta 549, 179–187.

Zhou, L., Hopke, P.K., Venkatachari, P., 2006. Cluster analysis of single particle mass spectra measured at Flushing, NY. Analytica Chimica Acta 555, 47–56.