International Journal of Mass Spectrometry 275 (2008) 1–10

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

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

ClusterSculptor: Software for expert-steered classification of single particle mass spectra

Alla Zelenyuk[a,*], Dan Imre[b], Eun Ju Nam[c], Yiping Han[c], Klaus Mueller[c]

[a] *Pacific Northwest National Laboratory, Richland, WA 99354, USA*
[b] *Imre Consulting, Richland, WA 99352, USA*
[c] *State University of New York at Stony Brook, Stony Brook, NY 11794, USA*

ARTICLE INFO

Article history:
Received 26 December 2007
Received in revised form 29 April 2008
Accepted 30 April 2008
Available online 18 May 2008

Keywords:
Single particle mass spectrometry
Data classification
Data visualization

ABSTRACT

To take full advantage of the vast amount of highly detailed data acquired by single particle mass spectrometers requires that the data be organized according to some rules that have the potential to be insightful. Most commonly cluster analysis methods are used to classify the individual particle mass spectra on the basis of their similarity. Cluster analysis is a powerful strategy for the exploration of high-dimensional data in the absence of a-priori hypotheses or data classification models. However, more often than not, the examination of the data clustering results reveals that many clusters contain particles of different types and that many particles of one type end up in a number of separate clusters. Our experience with cluster analysis shows that we have a vast amount of non-compiled knowledge and intuition that if brought to bear in this effort has the potential to greatly improve it. ClusterSculptor is software package designed to provide a comprehensive and intuitive visual framework to aid scientists introduce their vast knowledge into the data classification process. ClusterSculptor offers a wide variety of tools that are necessary for a high-dimensional, expert-driven activity we call cluster sculpting. ClusterSculptor is designed to be coupled to SpectraMiner, our data mining and visualization software package. The data are first visualized with SpectraMiner and identified problems are exported to ClusterSculptor, where the user steers the reclassification and recombination of clusters of tens of thousands of particle mass spectra in real-time. The resulting sculpted clusters can be then imported back into SpectraMiner.

Here we present the results of a study, in which ClusterSculptor is used to classify a complex dataset that includes single particle mass spectra of a variety of particle types. The compositions of these laboratory generated particles were carefully chosen to test some of the more difficult aspects of single particle mass spectroscopy. We demonstrate the use of ClusterSculptor to greatly improve chemical speciation of single particles by introducing expert input into data classification process.

1. Introduction

Single particle mass spectrometers are sophisticated instruments designed to measure the sizes and compositions of individual aerosol particles in-situ and in real-time. These instruments characterize hundreds of thousands or millions of particles, generating vast amounts of rich and complex data. To make efficient use of these data calls for the development of software that is designed to transform the process of mining hundreds of gigabytes of data into a tractable task.

Several data clustering approaches have been applied to the analysis of data produced by single particle mass spectrometers [1–10]. These methods typically treat the individual particle mass spectra as multidimensional vectors and calculate their proximity in *N*-D space. The distances between these spectra can be expressed by a number of metrics such as: Euclidean distances, dot products, correlation coefficients, Mahanalobis distances, etc. On the basis of these distances a variety of clustering procedures like: hierarchical cluster analysis [4,8,9], artificial neural networks [3,7], principal components analysis [6], k-means clustering [5,9], and fuzzy *c*-means clustering [8] are then used to organize particle mass spectra into classes.

The hierarchical methods initially yield a large number of clusters, which are then combined until stopping conditions are met, resulting in a limited or, "manageable" number of clusters. It is important to note that in this approach there is a simple option for a scientist to determine the final clustering outcome and with it select the classification's final level of detail. These decisions are

* Corresponding author. Tel.: +1 509 3767696.
E-mail address: alla.zelenyuk@pnl.gov (A. Zelenyuk).

1387-3806/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2008.04.033

made by visually examining the clusters' content and on the basis of "expert" knowledge. Other clustering approaches classify particles into classes on the basis of a set distance metric [1,3,7]. In all cases, at the end of the classification process each class is represented by an average/representative mass spectrum. It is important to keep in mind that once the data are organized and reduced into classes, there is no convenient path to a higher level of details.

Trimborn et al. [2,8] developed and applied a fuzzy classification algorithm in an attempt to account for the complex internal mixtures of individual atmospheric aerosol particles. In this approach internally mixed particles belong to more than one class, with varying degree of membership that is intended to represent their compositions.

In the past few years we have developed and applied a unique data mining and visualization software package we call SpectraMiner that makes it possible to handle hundreds of clusters, limiting loss of information and thus overcoming the boundaries set by traditional data cluster analysis approaches [9]. SpectraMiner organizes the data in a circular interactive hierarchical tree, or dendrogram, and provides the user with a visually driven, intuitive interface to easily access the data at all levels, and to mine the data in *real-time*. The user can view the mass spectral data at any point on the hierarchical tree, from mass spectra of individual particles to the average mass spectrum of the entire dataset, and any intermediate point. SpectraMiner makes it possible to account for the complex internal mixtures of atmospheric aerosol particles [11] without resorting to particle partial membership in several different classes as used in a fuzzy classification algorithm. With SpectraMiner the user can easily identify classes that contain very few particles, can create animations that illustrate the temporal evolution of the data, or the behavior of the particle data as a function of any other variable of interest. These are but a few of the features of this software that was specifically developed to handle data generated by single particle mass spectrometers.

While the methods, metrics and threshold distances that are used as criteria for particles to belong to the same class can be different for the different data classification methods, the vast majority of these approaches rely on unsupervised cluster analysis. The examination of the data clustering results reveals that more often than not, users are not completely satisfied with the results of this type of data classification. The two most common problems with unsupervised data classification are that they often fail to properly separate different particle types and that identical particle types are spread over a number of clusters. The first problem makes it almost impossible to properly follow particle behavior and the second results in unnecessary complexity that hinders data analysis.

Past experience with different clustering approaches [1,3,4,9,10,12] reveals that classification of single particle mass spectra could be greatly improved if the users aid the clustering process by utilizing their scientific knowledge to steer the classification.

It is clear that the vast amount of expert knowledge that scientist accumulated by working with his/her instrument in the field and in the laboratory should be harnessed to help guide the data classification. This approach would allow the clustering process to take into account the peculiarities of the instruments used to acquire the data and their impact on the data they generate. It may even include information about the properties of the specific particles that are the subject of the study.

A few different approaches were used to refine data classification by inserting scientific expert knowledge. The most common is for the user to manually combine clusters on the basis of a visual inspection of their average mass spectra [4,12]. Other approaches rely on using different clustering algorithms [10] or classification parameters [9,10]. Some have defined discriminant chemical markers or characteristic peaks [1,9] that are used in particle assignment. In other cases data are even modified prior to classification [4,9,10].

While each of these approaches demonstrated some improvements, their actual efficacy is difficult to establish because they were mostly applied to ambient data for which the correct final result is unknown. To date there are very few papers that describe the results of classification of individual particle mass spectra of particles whose compositions are known. Phares et al. [13] provided results of an application of the ART-2a algorithm to data classification of seven particle types. The final outcome of this exercise revealed that many particle types were poorly classified. The authors point to impurities, contaminations, and shortfalls of laser ablation to explain their findings. Murphy et al. [4] conducted a study in order to test their hierarchical clustering algorithm and to compare it with ART-2a. This study reports the results of classifying nine types of laboratory generated particles, a dataset of particle mass spectra that were acquired in the ambient atmosphere and the very same dataset of seven laboratory generated particles that was used by Phares et al. [13]. In the Murphy et al. [4] study "expert" knowledge about particle chemistry, the atmosphere, and the instrument response was used in a number of steps to achieve "satisfactory" results of data classification. In this case "expert" input was used to: guide the determination of stopping conditions, choose the type of data scaling to be used, and decide which clusters should be "manually" combined. An examination of the final classification results revealed a number of problems with particles that were composed of: ammonium sulfate, ammonium nitrate, organics and their mixtures. Again, the authors attribute the difficulty to achieve proper classification to the shortcomings of ablation and the presence of trace contaminants. Moffet and Prather [12] described results of a study of two particle types—one composed of polystyrene latex (PSL) and the other of dioctyl sebacate (DOS), whose mass spectra were classified, using the ART-2a algorithm, into 35 clusters. These clusters were then hand sorted on the basis of the intensity of the Na^+ mass spectral peak intensity, with DOS particles exhibiting higher Na^+ intensity. The authors justify this process by claiming that, in their system, DOS contains higher impurity of sodium than PSL. Rebotier and Prather [10] used a mixture of presorted ambient particles to test the accuracy and speed of a number of data clustering approaches. This study shows that it is possible to improve data classification by altering the mass spectra prior to the classification process. In this case the authors used the square root of peak intensities [10] to reduce the impact of some of the dominant peaks. In a publication that presented the use of SpectraMiner to classify the data of 12 types of laboratory generated particles [9], the compositions of which were chosen to test some of the more difficult aspects of single particle mass spectroscopy, we discussed the limitations inherent to unsupervised data clustering approaches. We demonstrated that more stringent criteria for particles to belong to the same cluster can in some cases improve the separation of different particle types, but we also found that an unavoidable outcome of this approach is an increase in the number of clusters and overall complexity. We also showed that SpectraMiner can be used to identify and diagnose specific causes for the failure to properly classify particle types and illustrated how the mass spectral data can be manually sculpted prior to classification to achieve improved data clustering.

It is important to note that all past cases, in which "expert" input was used to influence the classification, were implemented in the absence of any intuitive tools that have the potential to turn this type of activity into an iterative process, in which the scientist can steer the data classification process with ease, and offer the user the framework to test and verify a proper outcome.

In some respects, as we advanced our understanding of the data we acquired and the analysis tools we deployed, we begun to view

data classification and exploration with SpectraMiner as an important first step in data analysis. SpectraMiner presents the results of data classification in easy to explore formats that allow the user to identify points that need to be, or could be improved by including expert knowledge in the clustering process.

It is important to note that there are two elements that have the capacity to limit how much detail can be extracted from the data generated by single particle mass spectrometers: the first being determined by the quality of the signal that is generated by the instrument and the second relates to the data analysis process. Our goal here is to demonstrate how our recently developed software, we call ClusterSculptor [14], makes it possible to push data analysis to the point where it offers the option to reach the limits that are set by the quality of the data. ClusterSculptor is interactive software that is designed to work either in conjunction with SpectraMiner by importing any subset of the data from it, or be used as a standalone real-time visual data classification program. ClusterSculptor is designed to provide the researcher with the tools to input their expert knowledge by sculpting the data in a manner that steers the clustering process.

The software is built around an intuitive visually driven interface, through which the user can mold the data with a variety of tools, visually inspect the transformation using a number of representations, modify the process and re-cluster the data. An essential aspect of this software is that the process of data shaping, clustering and result visualization are all carried out with ease, and in real-time, which assures that the user is able to attempt a number of approaches until a satisfactory result is achieved.

In the sections below we will demonstrate how some of the numerous features of ClusterSculptor can be used to refine classification of single particle mass spectra and will test the software by applying it to several of the particle types that are of interest to atmospheric science and which were presented in the previous publication as a laboratory test case of SpectraMiner [9].

2. Methods

2.1. Particles, instrumentation, and mass spectra generation

All particle types, except soot, were generated by aerosolizing them from solutions using an atomizer (TSI Inc., Model 3076) as described in detail elsewhere [9,15]. Soot particles were sampled from a diesel engine (Mercedes 1.7L A-Class) at the National Transportation Research Center (Oak Ridge, TN) [9,16].

Individual particle mass spectra and their vacuum aerodynamic diameters were acquired by our single particle mass spectrometer, SPLAT. The instrument has previously been described in details elsewhere [9,16]. The data presented here were originally processed, classified, visualized and analyzed using SpectraMiner as described in detail by Zelenyuk et al. [9]. Here we focus only on the use of ClusterSculptor to extend this process. It is important to keep in mind that SpectraMiner is not an essential part of the process and that the data can be directly and completely analyzed by ClusterSculptor.

2.2. ClusterSculptor basics

Fig. 1 is a screen capture of ClusterSculptor displaying 7076 individual particle mass spectra of particles containing ammonium sulfate and organics. Below we provide a brief description of the elements that make-up ClusterSculptor, their functionalities and applicability to the classification of single particle mass spectra.

At the center, the program displays all of the individual particle mass spectra in a color map that is constructed from horizontal lines of colored pixels. Each of the horizontal lines represents a normalized, individual particle mass spectrum in which peak intensities are indicated by colored pixels using the rainbow color scheme, with red being high intensity and blue being low. Here we use a $\log(\log(I))$ scale, where I is peak intensity, to enhance the presence

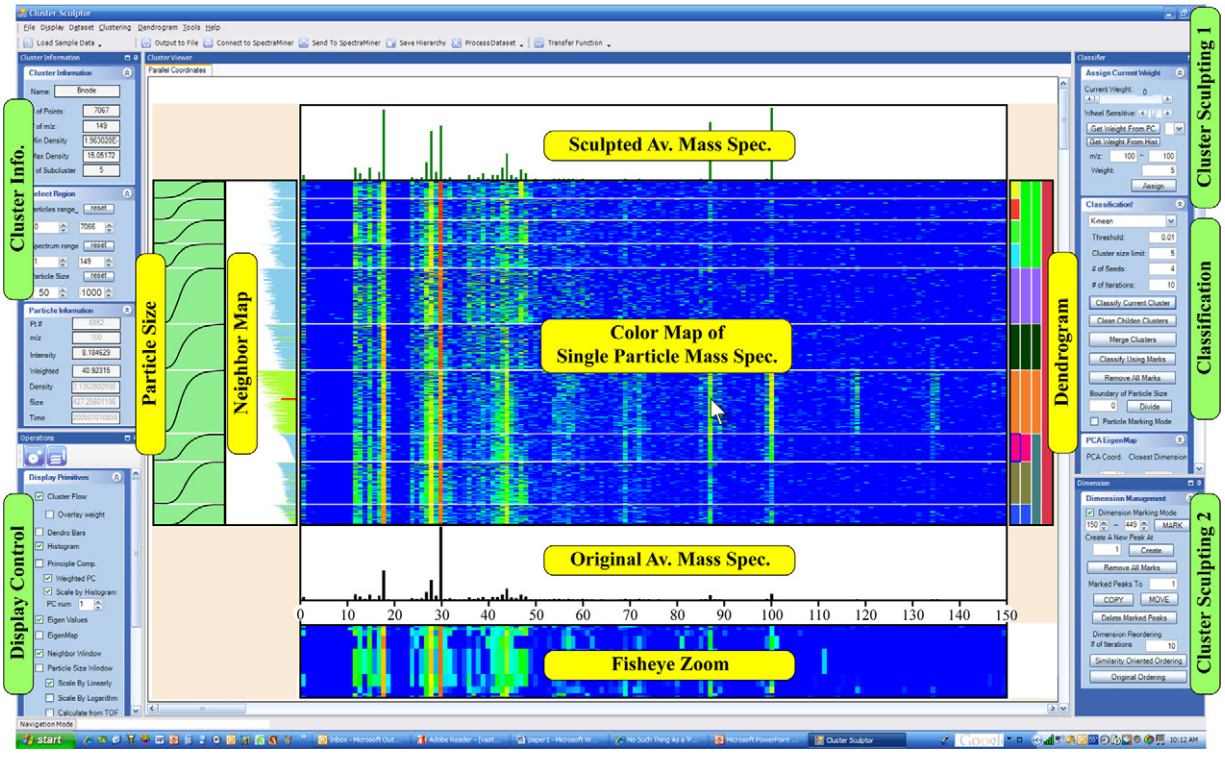

Fig. 1. Screen capture of ClusterSculptor displaying 7076 individual particle mass spectra. See text for a detailed description.

of very low intensity peaks, making it easy to observe variations between spectra. The user can select the m/z range to be displayed. The mass spectra are grouped into the clusters they belong to at the current stage of the data organization. The clusters are separated by white lines. By default the mass spectra within each cluster are organized according to their distance to the center of the cluster, but users can also sort the data by mass spectral intensity of any m/z value, by particle size or acquisition time (timestamp). In the lower panel, the program displays the average mass spectrum, reflecting unaltered relative intensities. At the top, the program displays the average sculpted mass spectrum in which the relative intensities reflect the scientists' input. A comparison between the average mass spectra in the top bottom panels reveals the differences.

The color map of the individual particle mass spectra remains unchanged when the mass spectral intensities of existing peaks are modified. However, when "new" peaks are created, they appear in the color map as well as in other clusters' views. What we mean here by "new" peaks and how they are created will become clear as we go through the examples later in the paper.

The cluster information window displays: (1) the cluster name, (2) the number of particles in the active cluster, (3) information relating to the size of the particles in the cluster, and (4) the mass spectral peak intensities and timestamps of the individual particle mass spectra. Double clicking on any of the clusters makes it active, at which point the displayed information, including the average mass spectra, number of particles and neighborhood map, refer to that cluster only. The active cluster can be further clustered.

In addition to the ability to sculpt the data, the users may click on any particle to set it as the initial cluster center for k-means clustering, making it possible to identify a particle type of interest and have the program search for other particles that are similar to it.

The green panel on the left displays the particle sizes. User can control the displayed size range and time-of-flight to aerodynamic size calibration parameters. The individual particle mass spectra can also be sorted by particle size, as is the case here. Particle size can be used to classify the data as well.

The dendrogram, shown on the right, represents the cluster hierarchy. Clicking on a dendrogram leaf turns that cluster into the active one and displays the information about the particles it contains. The user can name any or all of the leaves and select their colors. Most importantly, based on visual examination of the classification results, the user can subdivide or merge any of the clusters on any level of the hierarchical tree until satisfactory results are achieved.

The neighbor map on the left is designed to show the relations between a mass spectrum or a cluster and the rest of the dataset. When the mouse points to any particle mass spectrum on the color map, the neighbor cluster that is closest to that mass spectrum is indicated by a blue frame that appears in the dendrogram. In addition, the distances from each of the individual particle mass spectra to the center of the cluster that is being indicated are shown by the lengths of the bars in the neighbor map. The selected active mass spectrum is rendered in red, while the green bars indicate mass spectra that have the same closest neighbor. Distances of other mass spectra to the center of the emphasized cluster are rendered in blue. This information is helpful to identify similar clusters and to assist with the cluster merging process as described in detail by Nam et al. [14].

In the lower window on the left, the program provides the users with various display options, which we will illustrate in the next section.

In the windows on the right, the program displays the different control parameters. The upper frame is used to allow the user to sculpt the data by applying different weights to any of the mass spectral peaks, or to a range of m/zs, thereby increasing or decreasing peak intensities or a range of m/zs. The bottom right window is used to sculpt the data by creating "new" peaks on the basis of existing ones. Most commonly we use this window to create a "new" peak by first summing the intensities of a number of related peaks and then placing the value of the calculated sum at an m/z that typically has zero intensity. This option provides the means to minimize the, always-present, noise in the mass spectral relative peak intensities that is due to variations in fragmentation patterns.

The classification control window is used to set the clustering criteria, such as distance metrics, minimum number of particles per cluster, minimum (threshold) distances, the number of clusters to be formed (for K-mean clustering) and number of iterations.

The program also provides the user with the option to classify the data on the basis of particle size or on the basis of visually identifiable mass spectral features, where the later is achieved by inserting demarcation lines on the color map to serve as break points between clusters.

ClusterSculptor also performs a Principle Component Analysis (PCA) of the data and offers the user the option to view a projection of the dataset onto a 3D Eigen map, where the 3 dimensions are the largest Eigen values of the principal components. Since the principal components represent the axes with the most variant N-D space, the presence of independent clusters might be more perceptible in the *Eigen Space*. The utility of these views for data sculpting and clustering is described in detail by Nam et al. [14]. At any point the user can easily switch between the 3D Eigen map and the color map views.

Once data sculpting, classification, and merging are complete, the data can be exported back into SpectraMiner, where the new clusters are placed at the same position of the circular dendrogram they were originally extracted from. From this point on, all the data visualization and mining tools offered by SpectraMiner apply to the re-clustered data, which are merged into the overall dataset.

Below we demonstrate the application of ClusterSculptor to a number of particle types that are of interest to atmospheric science and which were presented in the previous publication as a laboratory test case of SpectraMiner [9].

3. Results and discussion

3.1. The case of particles containing alkali metals

In laser ablation single particle mass spectrometry it is common to find that the mass spectra of particles containing substances that produce positive ions with high efficiency, like alkali metals, are dominated by those ions [17–20]. Since the intensities of the other peaks, which are often used to properly distinguish different particle types and their sources, are significantly lower, it is very difficult to properly classify these data when using unsupervised data clustering. For example, urban dust and biomass burning particles are two common, but very different atmospheric particle types that exhibit high K^+ mass spectral peak intensity.

Another example is sodium-containing particles, like freshly emitted sea-salt particles containing NaCl and processed sea-salt particles containing $NaNO_3$. To better understand atmospheric processes, it is clearly important to be able to distinguish between these two particle types.

In our previous publication [9] we showed that the high intensity of the sodium peak in the mass spectra of NaCl and $NaNO_3$ particles made it impossible to properly separate the two particle types when a standard, unsupervised clustering approach was used. We showed that it is possible to improve the classification

to some degree by decreasing the distance threshold parameter, but it results in the creation of a large number of clusters. We also illustrated that great improvement could be achieved simply by reducing the intensity of the sodium peak prior to data clustering. Here we demonstrate the application of ClusterSculptor to the very same dataset.

In Fig. 2a we present the color map of 3000 NaCl particles and 3000 $NaNO_3$ particles. The data clearly indicate that all the particles contain Na, but in this format the color map provides no clear indication of the fact that two distinct particle types are present. To explore the data content, we reorder particles on the color map according to the intensity of a mass spectral peak that carries significant intensity, but is not the $m/z = 23$ peak. For example, the results of reordering the data on the basis of the intensity at $m/z = 62$ is shown in Fig. 2b. An inspection of the color map in this, reordered form clearly reveals the presence of two particle types.

In order to improve the clustering process, the next step involves increasing the weight of all the mass spectral peaks with $m/z > 23$ from a default value of 1–10, i.e., increasing their intensity by a factor of 10. The sculpted data are classified in this case into five clusters using K-mean clustering and the results of this classification are presented in Fig. 2c.

A comparison between the mass spectra of particles in these five clusters suggests that the 3rd (green) and 5th (blue) clusters contain $NaNO_3$ particles and the other three contain NaCl particles. Also shown in Fig. 2c in the green panel on the left are the particle size distributions. An inspection of the particle size distributions, on the left, provides support for these assignments. Particles in clusters three and five exhibit three distinct and relatively narrow sizes—these are consistent with particles that were selected classified with Differential Mobility Analyzer (DMA).

To complete the data organization, the clusters are combined to produce the two clusters shown in Fig. 2d, where the data are sorted by particle size, as is evident from the pattern on the particle size panel. The classified data can also be viewed in the 3D Eigen map view (not shown). The 3D Eigen map and principal component window can also be used to guide data sculpting and clustering as described by Nam et al. [14].

The two new clusters shown in Fig. 2d are now ready to be exported back to SpectraMiner to replace the original Na-containing clusters. It is important to note that the entire process of exporting the data from SpectraMiner, data visualization, sculpting, clustering, merging, and exporting the new clusters back to SpectraMiner is carried out in real-time and requires only a few minutes to complete.

3.2. Distinguishing between soot and organics

Understanding the role that organics play in determining the properties of atmospheric particles is at present one of the most important challenges of atmospheric aerosol research. Progress in this area clearly depends on our ability to identify and quantify the various organic compounds that are commonly present in aerosols. At the very least, it is important to be able to confidently distinguish between elemental carbon and organic carbon. Unfortunately, laser ablation can result in extensive fragmentation of organics compounds, often to the point that their mass spectral intensities are dominated by carbon progressions similar to those of soot. As a result, unsupervised classification has difficulty separating soot particles from particles that are composed of organic carbon and consequently a significant fraction of the particles that are composed of organic compounds can end up being classified erroneously as soot.

Here we present the use of ClusterSculptor to refine the clustering of three particle types that were classified into a single

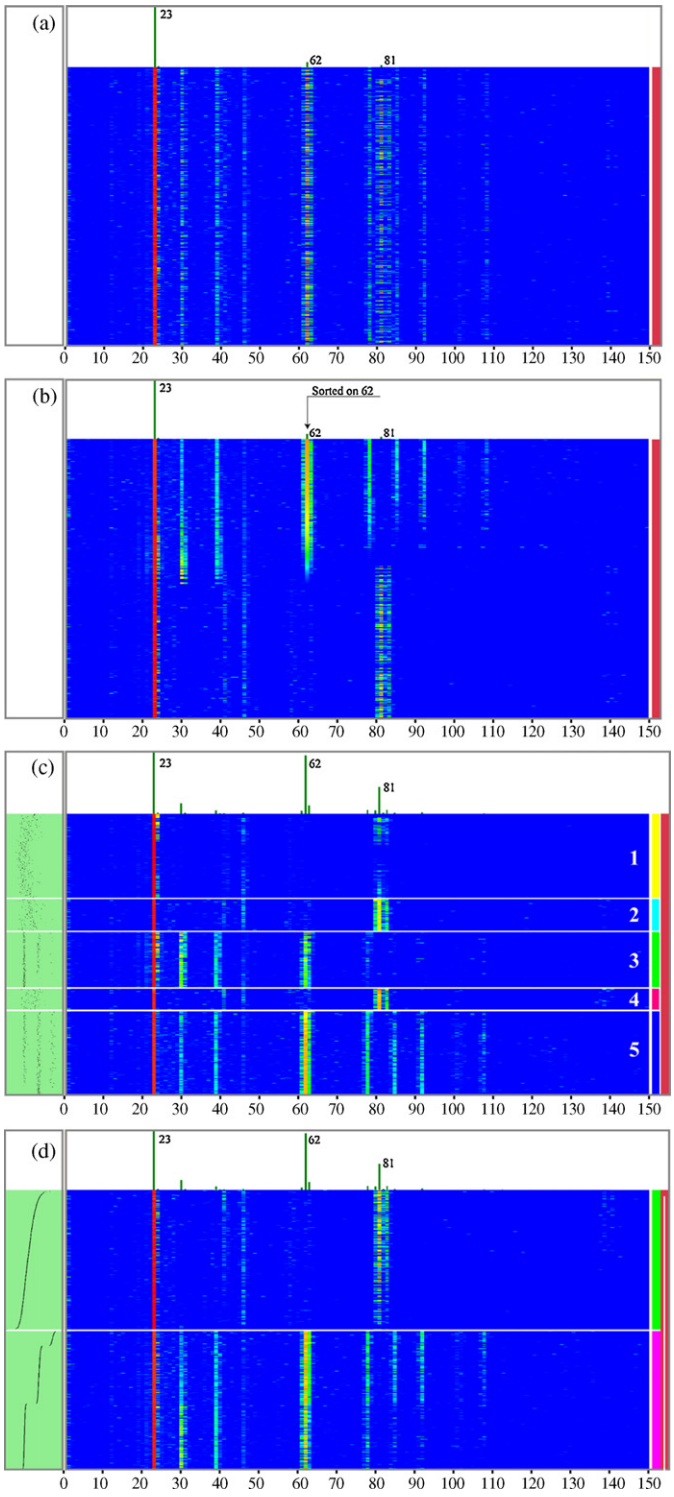

Fig. 2. (a) A color map of 3000 NaCl and 3000 $NaNO_3$ particles. (b) The same color map but sorted according to the intensity of the peak at $m/z = 62$. (c) The average mass spectrum of the sculpted data (top). The color map indicating the five clusters the data were classified into and the resulting dendrogram (right). The particle vacuum aerodynamic sizes (left green panel). (d) The final two clusters produced by recombining the five clusters in (c). Here the data are sorted by particle size.

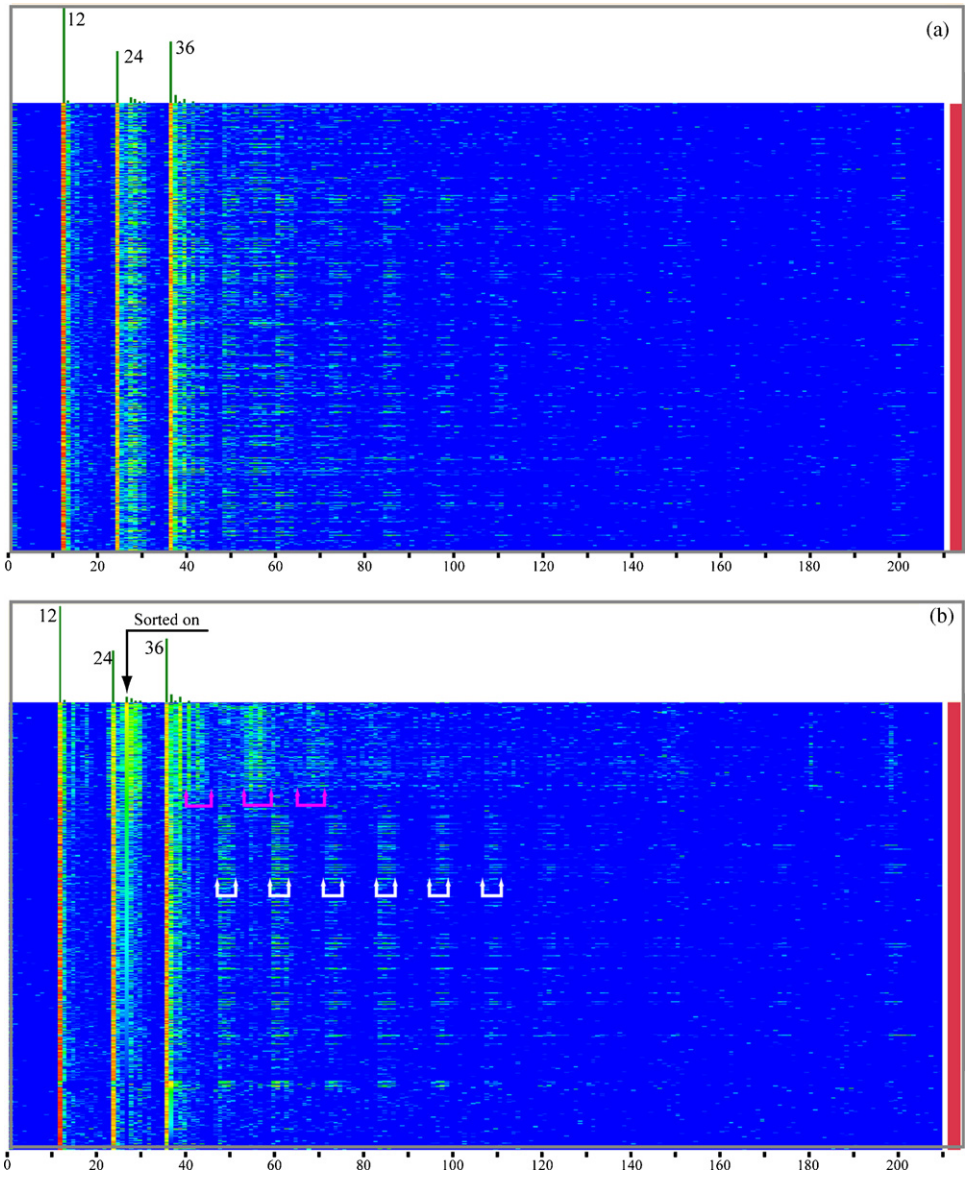

Fig. 3. (a) A color map of 5547 individual particle mass spectra of particles composed of pyrene, lauric acid and fresh diesel soot. (b) The same data as in (a), but sorted by the intensity of the peak at $m/z = 27$.

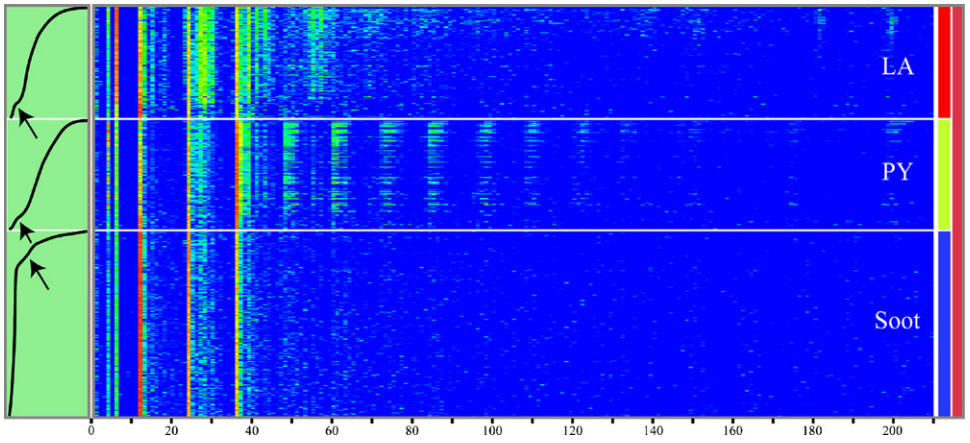

Fig. 4. A color map of the mass spectra of 5547 particles separated into three clusters and marked according to their composition. The data were sorted according to the particle vacuum aerodynamic sizes, which are shown in the left panel. The arrows mark discontinuities in the particle size distributions that indicate misclassified mass spectra.

branching point of the original SpectraMiner hierarchical tree [9]. The branching point was identifiable by soot-like progression. This dataset includes 5547 individual particle mass spectra: 2915 spectra of freshly emitted soot particles sampled directly from diesel exhaust; 1451 spectra of particles composed of pyrene, which is a polyaromatic hydrocarbon; and 1181 lauric acid particles, a 10-carbon organic molecule.

Fig. 3a shows the color map of all 5547 individual particle mass spectra and the average mass spectrum of all these particles, which is clearly dominated by the progression of carbon peaks at $m/z = 12$, 24, 36 common to soot particles. A careful inspection of the color map reveals the presence of other, much weaker peaks, which are consistent with organic compounds. A second indication that organics are present in addition to soot is provided by the relative intensities of the three carbon peaks. Our experience has shown that in the mass spectra of soot particles, the relative intensities steadily decrease when proceeding from C_1^+ to C_2^+ and then to C_3^+. In contrast, Fig. 3a shows that the intensity of the C_3^+ mass spectral peak is higher than that of C_2^+, suggesting that organic molecules are present. The same data are shown in Fig. 3b except that here the data have been sorted by the intensity of the peak at $m/z = 27$, which is commonly observed in the mass spectra of organic particles, but is not present in soot. In this view of the data, two types of organic particles become apparent, one exhibiting a short progression that is marked by magenta arrows and a second, longer progression that is indicated with white arrows.

Our experience with mass spectra of organic compounds generated by laser ablation shows that, in addition to a high degree of fragmentation, they often exhibit significant particle-to-particle variations in mass spectral peak intensities. This clearly presents significant obstacles to simple classification. In coarse classification, the lower intensity peaks play a minor role and many of the organics are classified together with soot. Under finer classification, with more stringent requirements on mass spectral similarity, particle-to-particle fluctuations result in a large number of clusters, making data analysis more difficult. Our experience

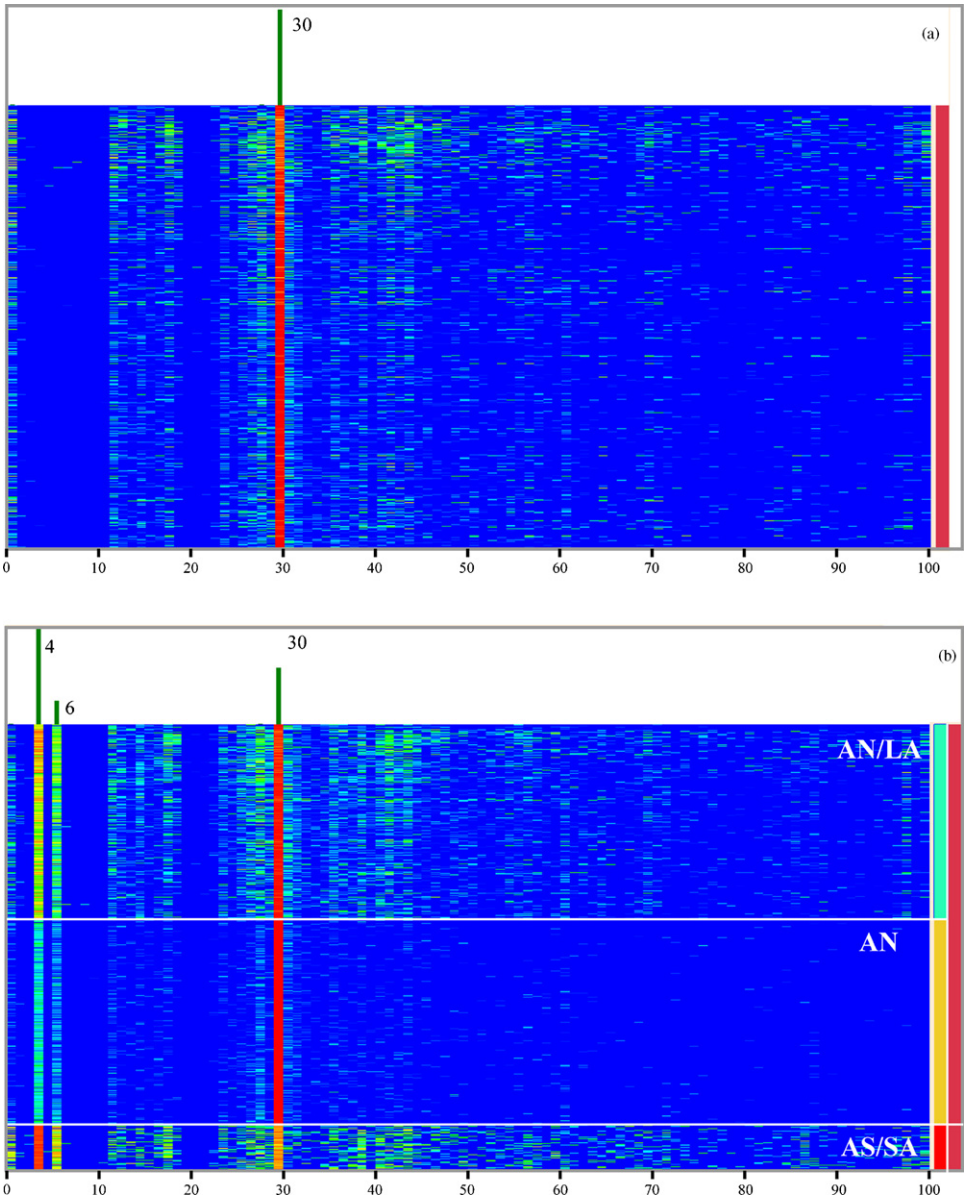

Fig. 5. (a) A color map of 6352 individual particle mass spectra of particles composed of ammonium nitrate, ammonium nitrate/lauric acid and ammonium sulfate/succinic acid. (b) The sculpted data showing the newly created peaks at $m/z = 4$ and 6 and the 3 clusters that the data were classified into.

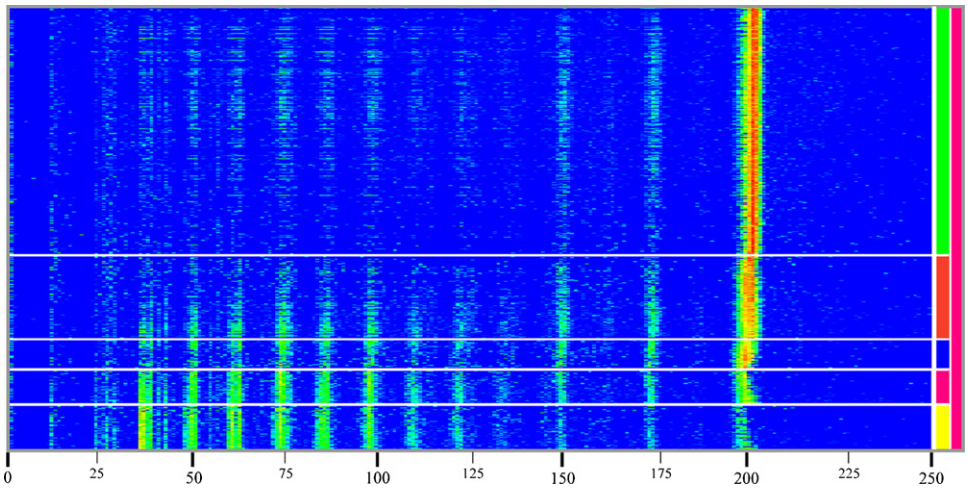

Fig. 6. A color map of 1552 individual mass spectra of pyrene particle. The data were originally classified into the five indicated clusters.

shows that in many cases, when progressions of related mass spectral peaks can be identified, it is possible to significantly reduce the particle-to-particle fluctuations by summing the related peaks.

In the present case we sum all the intensities of the peaks that are marked by the white arrows and create a new peak at $m/z = 4$. Similarly, we sum the intensities of the peaks marked by the magenta arrows and place that sum in the $m/z = 6$ position. We then assign a weight of 5 to these two new peaks to enhance their impact on the data classification. Once sculpted, the data are first classified into eight clusters, which are subsequently recombined to produce the three clusters shown in Fig. 4, where the individual particles are sorted by their vacuum aerodynamic diameters that are shown on the left. An examination of the size distribution in Fig. 4 reveals discontinuities we indicated with arrows. These discontinuities signify the presence of a small fraction of particles that remain misclassified. Since the fractal nature of fresh soot results in particles with vacuum aerodynamic size distribution that is relatively narrow, peaks at 100 nm and is independent of their physical diameters, we can rely on this property to complete the classification process by transferring these particle into the soot class.

3.3. Ammonium nitrate and ammonium nitrate with organics

The mass spectra of nitrate particles characteristically have very intense NO^+ signals. When organics are added to nitrate particles, the mass spectra of these internally mixed particles are still dominated by the NO^+ peak and the organics are almost invisible and hence difficult to properly identify. Here we present the analysis of 6352 individual particle mass spectra with a number of different compositions: 2999 pure ammonium nitrate (AN) particles, 2665 internally mixed AN/lauric acid (AN/LA) particles with 1:1 weight fraction ratio, and 688 ammonium sulfate particles mixed with succinic acid (AS/SA) at a number of different weight fraction ratios. Since the mass spectra of all of these particles are dominated by an intense NO^+ peak, they were originally classified together into the same branching point of the SpectraMiner dendrogram [9]. It is common for us to find that approximately 10% of the ammonium sulfate particles are classified together with AN particles. In the present case, the 688 AS/SA particles, which were included in the same branching point, represent less than 5% of AS/SA particles in the overall study.

Fig. 5a is a color map of the 6352 individual particle mass spectra. It reveals the fact that the mass spectra of all the particles in

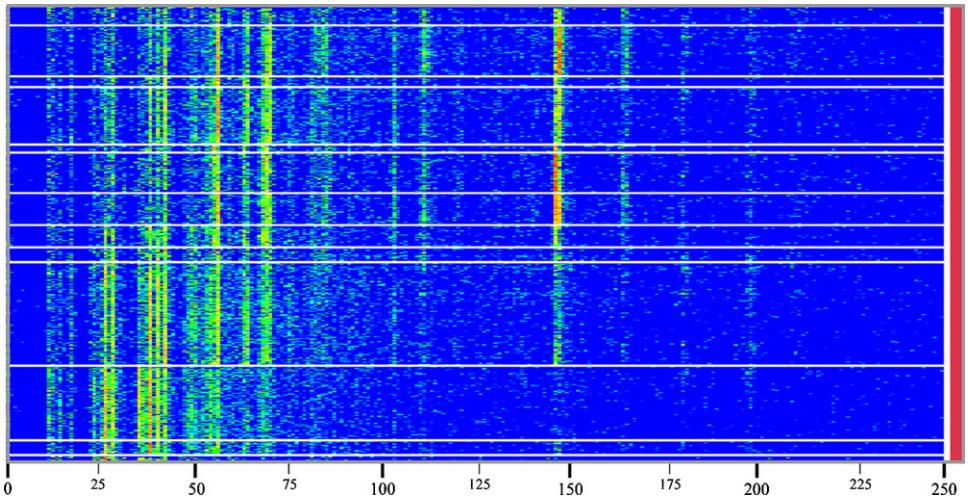

Fig. 7. A color map of 1852 mass spectra of lauric acid particles. The data were originally classified into the 14 clusters.

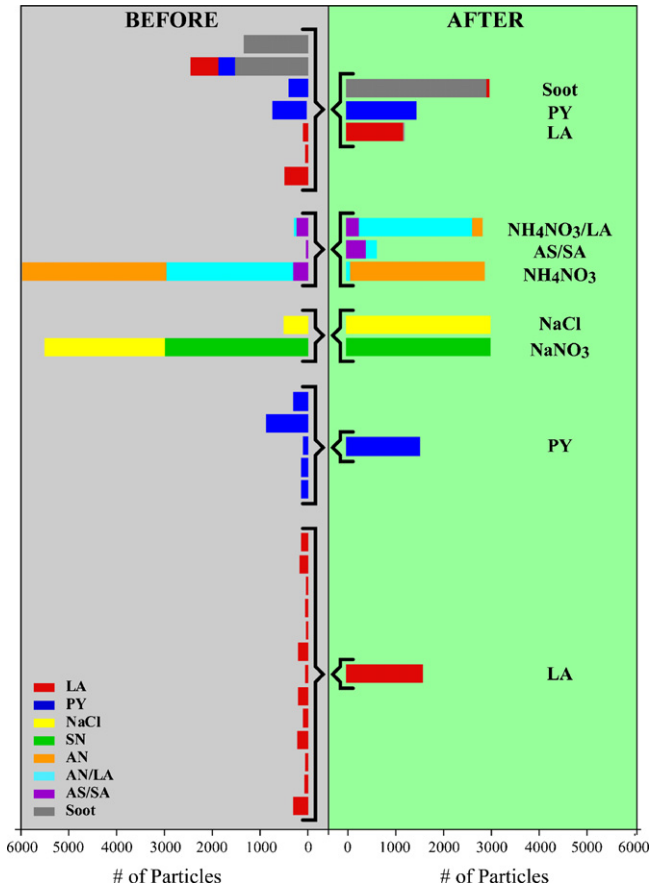

Fig. 8. A bar graph representation of results of two clustering approaches: unsupervised data classification (left) and expert-steered classification (right).

this class are clearly dominated by the NO^+ mass spectral peak at $m/z = 30$. The existence of organics in some of these particles is indicated by presence of several other peaks, even though no clear intensity patterns are observed that could be used for particle identification. To enhance the differences between particles containing organics and those that do not, we created a new peak at $m/z = 4$, which is the sum of all peak intensities except peaks at $m/z = 30$, 18 and 28. To help identify the internally mixed AS/SA particles, we created a second peak at $m/z = 6$, which is the sum of the two peaks at $m/z = 18$ and 28. Finally, a weight of 10 was applied to the newly created peaks, after which the data were classified and sequentially recombined to yield the three clusters shown in Fig. 5b. An examination of the classification results shows that ~90% of the AN/LA and AN particles were properly classified into the 1st and 2nd clusters, respectively, and that 60% of the mixed AS/SA particles were classified into the 3rd cluster, while the rest were classified into the 1st.

3.4. Fragmentation patterns of organic molecules and data clustering

So far, we have demonstrated the application of ClusterSculptor to cases where the original classification did not properly separate different particle types. In this section we address the opposite state, in which data clustering produces a number of clusters all populated by identical particle type. Since organic molecules often exhibit a wide range of fragmentation patterns, this problem is most commonly encountered during the classification of organic particles. To deal with this issue, we have already built into SpectraMiner the option to explore the data at all the hierarchal branches and even collapse the hierarchical tree at any branching point. However, SpectraMiner does not provide the means to conveniently visualize and inspect thousands of mass spectra and combine or separate different branches on different levels of the hierarchical tree.

To demonstrate this, we will use the example of two very different organic particle types composed of lauric acid and pyrene, both of which were previously encountered when re-clustering the node containing soot particles. In this section we examine the individual mass spectra of these particle types that were not classified with soot particles. Fig. 6 shows the color map of 1552 mass spectra of pyrene particles, which like most other polyaromatic hydrocarbons produce relatively simple and easy to identify fragmentation patterns. The five clusters that are shown in Fig. 6 were produced by the original data classification in SpectraMiner. A close examination of the color map of these mass spectra and the 3D Eigen map (not shown) strongly suggests that they all represent the same organic compound and should be combined together.

Fig. 7 provides a similar example of 1851 single particle mass spectra of particles that are composed of lauric acid. Again, the clusters represent the results of the previous classification in which the data were subdivided into 14 clusters. The extensive fragmentation pattern that is clearly visible in Fig. 7 is common to laser ablation generated mass spectra of long chain organics, like lauric acid. A high degree of particle-to-particle variability in the mass spectral relative peak intensities results in the large number of clusters that are indicated in the figure. Fig. 7 shows a simple trend of increasing degree of fragmentation as we move from top to bottom. As previously noted this trend does not end in this figure but continues to the point where some of the lauric acid particles are classified together with soot particles. ClusterSculptor is used here to combine all of these clusters and reduce the complexity of the data analysis.

4. Conclusions

We presented the results of an application of ClusterSculptor to the classification of individual particle mass spectra of laboratory generated particles of different types. We showed that unsupervised data clustering produces results that an expert can easily recognize as lacking. We demonstrated that the vast amount of scientific knowledge accumulated in the field of single particle mass spectrometry could, and should be used to steer the process of data clustering. To this end, we have developed new software called ClusterSculptor that provides the user interactive and intuitive tools to visualize and sculpt the data in a way that would guide data clustering process to yield the proper results.

We applied this software to address the problems previously identified by SpectraMiner – our dedicated data visualization and mining software package. Our goal was to put ClusterSculptor to the test and to gauge its performance in identifying and separating different particle types that were wrongly classified together and combining particles of the same particle type that were clustered into many smaller classes. Our ultimate goal is to produce an interactive visual display of the data in its simplest possible form, but with minimum loss of information. In Fig. 8 we provide bar graphs illustration what has been accomplished by applying ClusterSculptor to the data. On the left we indicate the state of classification results prior to the application of ClusterSculptor and on the right we present the end results of this study. We have been able to drastically reduce the number of clusters that represent the 8 particle types used in this study and most importantly, we have shown that, with the exception of the AS/SA particles, the new clusters are nearly pure.

Acknowledgments

This work was supported by the US Department of Energy Office of Basic Energy Sciences, Chemical Sciences Division. Part of this research was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research at Pacific Northwest National Laboratory (PNNL). PNNL is operated by the US Department of Energy by Battelle Memorial Institute under contract No. DE-AC06-76RL0 1830.

References

[1] P.V. Tan, O. Malpica, G.J. Evans, S. Owega, M.S. Fila, J. Am. Soc. Mass Spectrom. 13 (2002) 826.
[2] A. Trimborn, K.P. Hinz, B. Spengler, Aerosol Sci. Technol. 33 (2000) 191.
[3] D.J. Phares, K.P. Rhoads, A.S. Wexler, D.B. Kane, M.V. Johnston, Anal. Chem. 73 (2001) 2338.
[4] D.M. Murphy, A.M. Middlebrook, M. Warshawsky, Aerosol Sci. Technol. 37 (2003) 382.
[5] N. Erdmann, A. Dell'Acqua, P. Cavalli, C. Gruning, N. Omenetto, J.P. Putaud, F. Raes, R. Van Dingenen, Aerosol Sci. Technol. 39 (2005) 377.
[6] K.P. Hinz, R. Kaufmann, B. Spengler, Aerosol Sci. Technol. 24 (1996) 233.
[7] X.H. Song, P.K. Hopke, D.P. Fergenson, K.A. Prather, Anal. Chem. 71 (1999) 860.
[8] K.P. Hinz, M. Greweling, F. Drews, B. Spengler, J. Am. Soc. Mass Spectrom. 10 (1999) 648.
[9] A. Zelenyuk, D. Imre, Y. Cai, K. Mueller, Y.P. Han, P. Imrich, Int. J. Mass Spectrom. 258 (2006) 58.
[10] T.P. Rebotier, K.A. Prather, Anal. Chim. Acta 585 (2007) 38.
[11] A. Zelenyuk, D. Imre, J.H. Han, S. Oatis, Anal. Chem. 80 (2008) 1401.
[12] R.C. Moffet, K.A. Prather, Anal. Chem. 77 (2005) 6535.
[13] D.J. Phares, K.P. Rhoads, M.V. Johnston, A.S. Wexler, J. Geophys. Res.-Atmospheres (2003) 108.
[14] E.J. Nam, Y. Han, K. Mueller, A. Zelenyuk, D. Imre, IEEE Symposium on Visual Analytics Science and Technology, VAST 2007, 2007, p. 75.
[15] A. Zelenyuk, Y. Cai, D. Imre, Aerosol Sci. Technol. 40 (2006) 197.
[16] A. Zelenyuk, D. Imre, Aerosol Sci. Technol. 39 (2005) 554.
[17] D.S. Gross, M.E. Galli, P.J. Silva, K.A. Prather, Anal. Chem. 72 (2000) 416.
[18] Y. Cai, A. Zelenyuk, D. Imre, Aerosol Sci. Technol. 40 (2006) 1111.
[19] P.T.A. Reilly, A.C. Lazar, R.A. Gieray, W.B. Whitten, J.M. Ramsey, Aerosol Sci. Technol. 33 (2000) 135.
[20] A.N. Zelenyuk, J. Yang, C. Song, R. Zaveri, D. Imre, J. Phys. Chem. 112 (2008) 669.