Atmos. Meas. Tech., 10, 1323–1334, 2017
www.atmos-meas-tech.net/10/1323/2017/
doi:10.5194/amt-10-1323-2017

Atmospheric Measurement Techniques
Open Access
EGU

FATES: a flexible analysis toolkit for the exploration of single-particle mass spectrometer data

Camille M. Sultana[1], **Gavin C. Cornwell**[1], **Paul Rodriguez**[2], **and Kimberly A. Prather**[1,3]

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA
[2]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[3]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093, USA

Correspondence to: Kimberly A. Prather (kprather@ucsd.edu)

Received: 4 September 2016 – Discussion started: 6 October 2016
Revised: 14 February 2017 – Accepted: 14 March 2017 – Published: 4 April 2017

Abstract. Single-particle mass spectrometer (SPMS) analysis of aerosols has become increasingly popular since its invention in the 1990s. Today many iterations of commercial and lab-built SPMSs are in use worldwide. However, supporting analysis toolkits for these powerful instruments are outdated, have limited functionality, or are versions that are not available to the scientific community at large. In an effort to advance this field and allow better communication and collaboration between scientists, we have developed FATES (Flexible Analysis Toolkit for the Exploration of SPMS data), a MATLAB toolkit easily extensible to an array of SPMS designs and data formats. FATES was developed to minimize the computational demands of working with large data sets while still allowing easy maintenance, modification, and utilization by novice programmers. FATES permits scientists to explore, without constraint, complex SPMS data with simple scripts in a language popular for scientific numerical analysis. In addition FATES contains an array of data visualization graphic user interfaces (GUIs) which can aid both novice and expert users in calibration of raw data; exploration of the dependence of mass spectral characteristics on size, time, and peak intensity; and investigations of clustered data sets.

1 Introduction

Single-particle mass spectrometers (SPMSs) yield the size and chemical composition of individual aerosol particles in real time. SPMSs can generate tens of single-particle mass spectra per second, utilizing laser desorption–ionization (LDI). However mass spectra generated by LDI exhibit ion signals only qualitatively dependent on particle chemical composition (e.g., Ge et al., 1998; Gross et al., 2000; Hinz and Spengler, 2007) and also can exhibit large particle-to-particle variation even for chemically uniform particles (e.g., Steele et al., 2005; Wenzel and Prather, 2004; Zelenyuk et al., 2008a, b). Thus SPMSs generate both large and highly complex data sets, requiring sophisticated data analysis techniques for exploration and distillation of information.

As Table 1 illustrates, individual laboratories have independently developed a variety of SPMSs, and two commercial versions have also been produced. Due to the many iterations of SPMSs that exist and the lack of a standard data format, individual laboratories have had to build their own data analysis software, though these toolkits are often not reported in the literature (Table 1). Only two of these data analysis toolkits have been made publicly available, YAADA (www.yaada.org) and ENCHILADA (www.cs.carleton.edu/enchilada). YAADA is specific to the lab-built and commercial versions of the aerosol time-of-flight mass spectrometer (ATOFMS), a version of SPMS (Allen, 2005). ENCHILADA is reported to be compatible with three SPMS designs: SPASS, PALMS, and TSI ATOFMS (Gross et al., 2010). However, the authors could only find reported use of the ENCHILADA toolkit for TSI ATOFMS and SPASS data sets. Despite their age these toolkits are still utilized, with YAADA being the toolkit of choice for the burgeoning SPMS community in China. The differences and limitations between these two software tools have been extensively described previously (Gross et al., 2010), but a brief summary is given here. YAADA is an object-oriented frame-

Table 1. Summary of SPMSs developed and data analysis packages used.

SPMS version	Analysis toolkit utilized
Lab-developed instruments	
ALABAMA[a]	CRISP (IGOR toolkit)[b]
ATOFMS[c] (UF-ATOFMS[d])	YAADA (MATLAB toolkit)[e]
LAMPAS[f] (LAMPAS 2[g], LAMPAS 3[h])	Not reported
PALMS[i]	Not reported
RSMS[j] (RSMS-II[k], RSMS-III[l])	Not reported
SPASS[m]	ENCHILADA[n,o]
SPLAT[p] (SPLAT II[q], mini-SPLAT[r])	SpectraMiner[s], ClusterSculptor[t]
Commercial instruments	
Guangzhou-Hexin ATOFMS/SPAMS (currently manufactured)[u]	YAADA[v]
TSI ATOFMS (discontinued)[w]	YAADA[x], ENCHILADA[y]

[a] Brands et al. (2011). [b] Klimach (2012). [c] Gard et al. (1997). [d] Su et al. (2004). [e] Allen (2005). [f] Hinz et al. (1994). [g] Trimborn et al. (2000). [h] Hinz et al. (2011). [i] Thomson et al. (2000). [j] Carson et al. (1995). [k] Phares et al. (2002). [l] Lake et al. (2003). [m] Erdmann et al. (2005). [n] Healy et al. (2010). [o] Gross et al. (2010). [p] Zelenyuk and Imre (2005). [q] Zelenyuk et al. (2009). [r] Zelenyuk et al. (2015). [s] Zelenyuk et al. (2006). [t] Zelenyuk et al. (2008). [u] www.tofms.net/content.aspx?info_lb=387&flag=103. [v] Zhang et al. (2015). [w] www.tsi.com/aerosol-time-of-flight-mass-spectrometers-series-3800. [x] Dall'Osto et al. (2012). [y] Sierau et al. (2014).

work implemented in MATLAB that allows user-developed script-based data exploration and can also leverage the extensive set of built-in functions within MATLAB. This allows a degree of flexibility in creating graphical outputs and exploring ATOFMS data in tandem with other data types. However, the extensive amount of code required at the time of development to create the object-oriented framework for YAADA has made the toolkit highly susceptible to updates and changes in MATLAB. Thus continued use of YAADA requires either using outdated MATLAB versions or extensive maintenance of the scripts underlying the toolkit. Also considerable knowledge of YAADA-specific data classes and framework in addition to general MATLAB understanding is required to be able to manipulate the data. Additionally, YAADA's accessibility is limited for novice users as there are no graphic user interfaces (GUIs) for data exploration. In comparison ENCHILADA is a software package with a graphical user interface. Therefore data analysis functions and workflows built into ENCHILADA are leveraged by interacting with the GUI, without the need to create scripts or interact in a command line interface. However any addition of functionality requires modifying the underlying source code and rebuilding the software. ENCHILADA relies primarily on SQL for accessing and storing the mass spectral database and Java for implementation of the GUI, though a number of other drivers, toolkits, and C++ are also integrated into its implementation. Thus modifications are a significant programming task and likely infeasible for scientists not highly experienced in programming and computer science.

Motivated by the continued use of SPMS and the limitations of the currently available software, we have developed a new flexible analysis toolkit for the exploration of single-particle mass spectrometer data (FATES). To encourage the widespread adoption of this toolkit, it was purposely designed in an extensible manner to adapt to the ever-evolving and varied implementations of SPMS. It is clear that building open-source tools in a standard, well-known platform and creating a work flow with user-defined parameters for data analysis would be beneficial to the SPMS community, increasing the rate of knowledge discovery and enabling collaboration between researchers. For example, maintenance and alterations of the software should be easily accessible to chemists and aerosol scientists without extensive training in computer science. In addition, any new toolkit should not be explicitly limited to expected common analyses, which may be built into GUIs, but should give the user complete freedom to access, explore, and utilize SPMS data and also integrate with other temporally and spatially resolved data sets. Finally any framework needs to make careful consideration of both memory and speed constraints imposed by the possible large size of SPMS data sets. Given these constraints, the FATES toolkit (Sultana et al., 2017) was developed completely in the MATLAB environment, and an extensive manual was written and is provided in the Supplement. MATLAB is a popular language for numerical data analysis by scientists because it has an extensive library of well-documented built-in functions, utilizes libraries optimized for speed in matrix manipulation, and can support both graphical and script-based exploration of data. By taking advantage of native MATLAB data types, FATES is easier to maintain and computationally more efficient than YAADA, the previous publicly available MATLAB toolkit for SPMS analysis. The FATES framework allows users to creatively explore their data without previous assumptions or constraints with simple scripts and by leveraging built-in MATLAB functions. Additionally FATES of-

fers a suite of GUIs for interactive visualizations which can aid both novice and expert users in calibration of raw data; exploration of data sets using temporal, size, and mass spectral filters; and investigations of clustered data sets. FATES is the first publicly available SPMS toolkit to allow creative, efficient script-based data mining along with GUI-based visual data exploration and calibration all within a single programming environment.

2 FATES software description

FATES is implemented completely in MATLAB. No other languages, drivers, or software are needed to utilize FATES. In addition FATES was purposely developed in a manner that demands few presumptions about the instrument, particle, and spectral variables collected by the SPMS. For example one SPMS may only record the speed and time of detection for each particle, while another SPMS may also record the power of the desorption–ionization laser pulse. These differences are handled easily as FATES allows users to specify, define, and change the instrument, particle, and spectral variables they would like imported into and saved to a study. To make these alterations, users only need modify simple scripts where the desired variables are listed, and then these changes are carried over throughout the entirety of the source code. This flexible but simple design gives high utility for the SPMS community because it prevents users from needing expert knowledge of any language and having to search for and make line-by-line or structural changes within the source code. Detailed instructions for making these simple modifications are included in the FATES manual (Supplement M-5) and commented within the code. As distributed, the FATES source code already contains the necessary modifications to read in data sets from three SPMS designs: ATOFMS, ALABAMA, and TSI ATOFMS. In addition FATES avoids the explicit creation of new class objects, which minimizes the lines of source code and number of scripts by over an order of magnitude when compared to YAADA. This greatly minimizes the maintenance needed to keep FATES compatible with future versions of MATLAB. FATES has been tested for compatibility with MATLAB versions 2014b through 2016b.

2.1 FATES data architecture

SPMS data imported within FATES is stored within separate variables for the experiment description, the particle data, and the spectral data. A SPMS data set imported into MATLAB via FATES is referred to as a FATES study, the data architecture of which is comprehensively detailed in the FATES manual (Supplement M-4). Logically, the data mostly consist of one-to-many relationships from study to experiment, experiment to particle, and particle to spectral peaks. The data are most typically loaded once and then accessed and filtered in bulk. Therefore, it is more efficient to organize the observed measurements into denormalized matrices for particle and spectral data, where key information is duplicated in each matrix.

Each FATES study stores a data structure that contains a number of user-defined fields (e.g., instrument name, operator, location) to describe the experiment in which the data within the study were collected. Each row of the structure describes a unique experiment, which pertains to a unique experiment identifier (ID). All particle data (e.g., speed, power of desorption–ionization laser pulse) are stored in a MATLAB matrix. More specifically, each particle within a FATES study has a unique two-column particle ID. The first column of the particle ID is the experiment ID, previously described, to which the particle belongs. This framework allows users to easily select for particle or spectral data collected during a specific experiment within a FATES study that contains data from multiple experiments. The mass spectral data for all particles in the FATES study are held in an external binary file. Users can easily and quickly retrieve spectral peak data (e.g., m/z, area, height) for user-selected particles using functions provided by the FATES toolkit (Supplement M-6). The spectral data when imported are then stored in MATLAB cell arrays or matrices. Each peak for all of the spectra within a FATES study has a unique four-column peak ID. The first three columns of the peak ID are the experiment ID, particle ID, and polarity indicator of the spectrum to which the peak belongs. Note that each FATES study contains auxiliary data structures that list the name of the variable (e.g., particle speed, peak area, peak ID) that each column in a data matrix holds. Thus all data within a FATES study are self-contained and self-described, from experimental conditions to peak information. Therefore despite the flexibility of the FATES framework, users can still share FATES studies without confusion or need for external README files to determine the source and identify of the data.

2.2 FATES optimization

Considerable work has been completed to optimize the FATES framework for memory demands, speed, and ease of use. An ATOFMS data set collected at Bodega Bay, CA, in February and March of 2016 is used throughout this paper to illustrate the speed of data analysis within the FATES toolkit. This data set contains 1 386 042 dual-polarity single-particle mass spectra as well as particle data for an additional 11 454 356 particles that were detected in the light-scattering region but did not generate spectra. All FATES analysis is performed in MATLAB 2014b with an Intel Core i7-4930K CPU running at 3.4 GHz with 16.0 GB of RAM. Run time comparisons, summarized in Table 2, are made using the same computer utilizing a version of YAADA, which had been maintained by Kim Prather's research group to be compatible with MATLAB 2013a.

To begin working with a SPMS data set, a new FATES study has to be created (Supplement M-2). This process only

Table 2. Comparison of run times for various operations in YAADA and FATES.

	YAADA	FATES
Study creation	20.8 min (ATOFMS) 127 077 hit particles 1 050 174 missed particles	28.4 min (ATOFMS) 1 386 042 hit particles 11 454 356 missed particles 24.8 s (TSI ATOFMS) 68 400 hit particles 639 145 missed particles 3.2 s (ALABAMA) 1000 hit particles 86 744 missed particles
Mass spectra retrieval	42.5 s 127 077 mass spectra 17.3 s 50 000 mass spectra	3.3 min 1 386 042 mass spectra 26 s 400 000 mass spectra 2.7 s 50 000 mass spectra
Retrieval of particle IDs for hit submicron particles	0.6 s 127 077 hit particles	0.01 s 1 386 042 hit particles
ART-2a clustering	70 min 100 000 mass spectra	2.1 min 100 000 mass spectra

needs to occur once for any data set, but the source code was still designed to minimize the time for study initialization. Despite the large size of the Bodega Bay data set, the creation of the FATES study only took 28.4 min. Even initiating a subset of the Bodega Bay study roughly one-tenth of the FATES study (127 077 dual-polarity mass spectra) in YAADA still required 20.8 min. Small ALABAMA and TSI ATOFMS data sets were also initiated expediently in FATES (Table 2). Note the version of YAADA maintained by Kim Prather's research group is not able to import these data sets into MATLAB for comparison. FATES has also been designed so that additional data can be added to an existing study without having to re-initialize the entire data set (Supplement M-A). This is especially useful for field studies, where daily examination of the data is required, but initialization of increasingly large data sets can become onerous and time consuming.

Once a FATES study is initiated, it is crucial to efficiently handle the spectral data. Users may desire to examine data sets with millions of mass spectra, and each spectrum can contain hundreds of peaks. SPMS spectra data formats usually contain mass-to-charge (m/z) ratio and area for each peak, but they may also specify peak width, peak height, and other values. This amounts to many gigabytes of data, and therefore the trade-off between making all the spectral data available and managing memory requirements had to be taken into consideration. MATLAB facilities for tables were considered, but they are more appropriate for heterogeneous data, whereas in our case all the spectral data are numeric or binary indicators. We also found MATLAB memory mapped files to have unpredictable performance, and it was difficult to append data rows because matrices are stored in column order. We determined the best way to build up and maintain a large matrix of spectral data, without keeping it in memory, was to create a single external binary file, append to it as needed, and provide a lightweight interface so that FATES programs, or other users, could easily execute functions against the file. Essentially, this interface is an API (application programming interface), which takes a regular MATLAB command or script, shuffles data in/out of memory in blocks of rows, executes the commands against the data in memory, and gathers results. The block sizes are set to default values that are reasonable for current workstation capacities but can also be changed as appropriate in the future. The possible commands are unconstrained, but summaries and filtering operations are most appropriate and most likely to be called for.

In addition, the binary format minimizes both the time required to write and retrieve spectral data and the storage requirements for the file. Retrieving all 1 386 042 dual-polarity mass spectra in a single call from the external binary file created for the Bodega Bay study and loading it into a MATLAB array only took 3.3 min. It is important to note that this example is used for benchmarking purposes, but rarely would users need or choose to load into and hold all spectra information for entire large data sets within memory at the same time. The FATES framework automatically employs data pointers so that the whole binary file does not need to be read if the user is only attempting to retrieve spectra from particles which make up a subset of all the data in the FATES study.

Run times for retrieving all and contiguous subsets (i.e., the raw data files from which the study was created were contiguous) of the dual-polarity mass spectra from the FATES and YAADA studies are summarized in Table 2. Retrieving a subset of 50 000 mass spectra from the FATES study (2.7 s) was over 6 times faster than in the YAADA study (17.3 s). Searching through and sorting data by particle information is also quickly performed in the FATES framework. By holding all hit particle data in memory, any operation querying the particle data does not require any data input/output calls and therefore is nearly instantaneous in MATLAB. For example retrieving the particle IDs for all submicron particles from the Bodega Bay study only took 0.01 s, while performing a similar analysis on the much smaller YAADA study required 0.6 s.

The quickness of the FATES framework depends partially upon minimizing retrieval calls to external files outside of the MATLAB workspace. Thus formatting of the data held within the MATLAB workspace has been carefully considered to minimize the memory demands of the FATES framework. Because spectral data are held in an external binary file, users can choose to store spectra data in the study at a high resolution without increasing the study's working memory. When retrieving spectra from the external binary file, users may specify the resolution to hold the data in the workspace. This feature allows users to tailor the resolution of the spectra in the workspace to its application and therefore the memory requirements. Mass spectral data loaded into the MATLAB workspace are stored in a single-precision floating-point format, saving memory compared to the standard MATLAB double-precision format, which requires twice the space. Particle data stored within a FATES study have also been formatted to minimize memory demands. If the user loads data into a FATES study for both detected particles that generated mass spectra (hit) and detected particles that did not generate spectra (missed), only hit particle data are stored in the particle matrices in MATLAB. Most data analyses utilize spectra, and therefore only hit particle information is necessary, but hit particles usually make up a small fraction of total particles detected by the light-scattering region of the SPMS. Therefore storing missed particle data in MATLAB memory would take up large amounts of space needlessly. All missed particle data are written to an external binary file and can be loaded by the user into MATLAB using a script provided in the FATES toolkit. Furthermore particle data stored in MATLAB memory are split between a single-precision and double-precision matrix. It is not necessary to store most data collected for particles (e.g., speed, laser power) in a double-precision format, so this choice further relieves the space required to store all particle data in memory. Therefore storing data for 1 million hit particles in memory where three variables require double-precision format (particle ID, time) and three variables only need single-precision format (speed, size, laser power) only requires 0.036 GB, which is very feasible for most modern desktop computers. Finally because all SPMS data when loaded into a FATES study are held in native MATLAB data types, interacting with the data requires very few FATES-specific functions. Almost all common analyses can be patterned off a basic script, provided with demonstration data in the FATES toolkit and relying on a handful of MATLAB built-in functions and matrix indexing, which makes the FATES framework accessible and powerful for both expert and novice users.

3 Data analysis within FATES

In this section we provide a brief overview of common analyses that can be performed on SPMS data within a FATES study. However it should be mentioned that it is impossible to describe or predict all data analyses and plotting options easily available to FATES users due to the extensive library of built-in and user-developed MATLAB functions. A large array of analyses can be performed using concise code (Supplement M-6), with only a few examples quickly discussed here. By utilizing logical indexing, particles and spectra can be filtered using any single or combination of particle and mass spectral characteristics (e.g., particle size, peak area at a certain m/z, etc.). Binning of particles and spectra by these characteristics, such as binning data based on time, can be accomplished in a single line with the built-in function histc. Additionally lists of particles can be compared with the built-in function intersect. Grouping data based on algorithmic clustering of the spectra is also easily performed. Clustering methods commonly used by the SPMS community such as k means, hierarchical clustering, and k medoids are built in to MATLAB, and ART-2a, a fast adaptive resonance algorithm popular among ATOFMS users, is supplied in the FATES toolkit. Clustering data, which necessitates a large number of matrix operations, can be performed quickly even with naïve user scripts because MATLAB utilizes BLAS, LAPACK, and proprietary libraries which speed up common linear algebra computations. Clustering 100 000 particles from the Bodega Bay study with ART-2a (vigilance factor = 0.80, learning rate = 0.05) in the YAADA study required 70 min; however improvements in the ART-2a scripts in FATES allow the same analysis to be completed in only 2.1 min. With the built-in MATLAB k means function, the same data was grouped into 15 clusters in 2.9 min (77 iterations) in FATES. Finally other types of data can be easily loaded into MATLAB and examined along with the SPMS data.

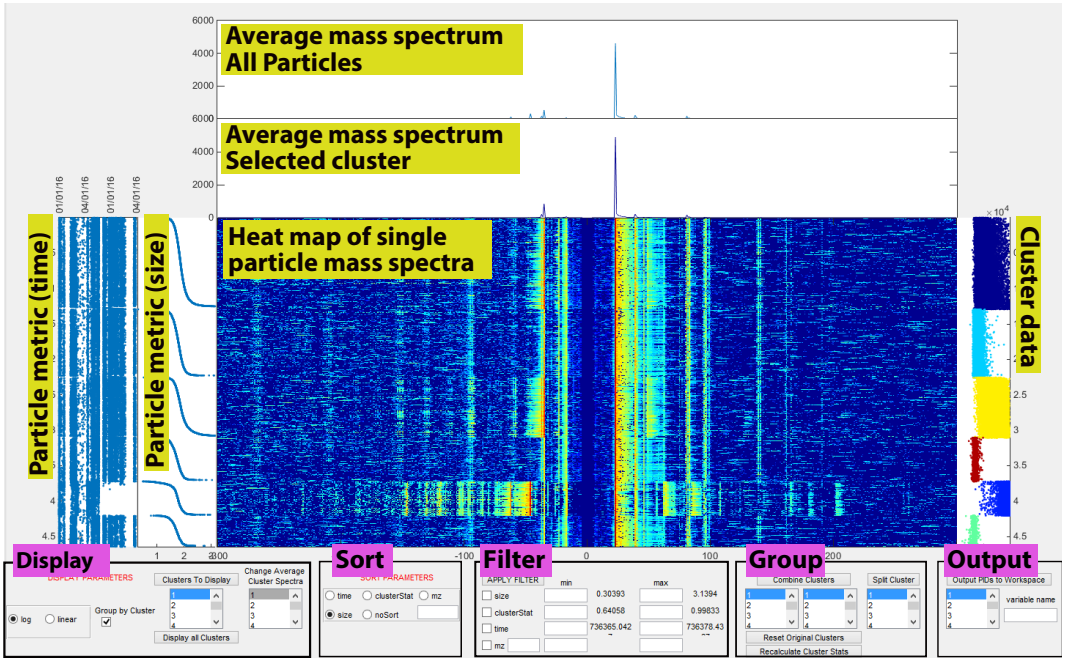

Figure 1. Screen capture of a guiFATES window with data from 46 432 individual particles.

4 Exploration of data utilizing FATES GUIs

4.1 guiFATES: spectra visualization, grouping, and exploration

While the FATES toolkit allows flexibility in script-based SPMS data analysis, graphical tools can also be an effective way to explore the data and quickly identify trends and patterns. To this end the FATES toolkit includes GUIs, built within MATLAB, which allow users to easily examine trends in spectra based on particle metrics such as size and time, and cluster and spectral characteristics. Figure 1 is a screen capture of the FATES spectra explorer guiFATES, displaying data for 46 432 particles. This spectra explorer has been modeled after ClusterSculptor, a SPMS data analysis GUI developed by Zelenyuk et al. (2008a) that has not been made publicly available. To initiate guiFATES, the user provides the function with the mass spectra, two user-selected particle metrics, and cluster data for a set of particles. A description of the functionality and abilities of guiFATES is given below.

The main panel of the guiFATES display is the heat map of the individual particle mass spectra. Each row is an individual mass spectrum with peak intensity indicated by color. The user can choose to display the provided mass spectra peak intensity utilizing a linear or log10 scale. The logarithmic scale makes it easier to visually detect relatively small peak intensities in the spectra, while the linear scale helps users visualize absolute differences between peak intensities. In Fig. 1 the logarithmic scale has been selected. Users can choose to provide any two characteristic particle metrics, such as particle size, time of detection, laser pulse energy, or total ion intensity, which are displayed in the left panels. In Fig. 1 particle time and size have been provided. Clustering information is displayed in the right panel. The cluster or group assigned to each particle is indicated by the color of the points on the right, while the location on the x axis is a user-provided clustering statistic for each particle. The clustering statistic provided for display in Fig. 1 is the dot product of each normalized particle spectrum with the normalized representative spectrum of the cluster to which the particle had been assigned. However, the user can provide any clustering or neighbor statistic they feel is effective for exploring their data set. The top plot in guiFATES is the average of all the provided spectra, and immediately below is plotted a select average cluster spectra, specified by the user in the display parameters. The line color in the average cluster spectra plot matches the colors used to indicate the assigned cluster for each particle in the right vertical plot. The bottom of the guiFATES windows contains all the display, sorting, filtering, and grouping parameters that the user may select and change.

guiFATES provides the user with many options for displaying and exploring the data, and all functionalities are thoroughly detailed in the manual (Supplement M-7). A check box allows the user to display all data with or without grouping by cluster. In addition the user can select to sort the data by any of the particle metrics in the vertical side panels or by a m/z value in the spectra. In Fig. 1 the data are displayed by cluster and sorted by size. Figure S1a in the Supplement is a screen capture where the same data are not

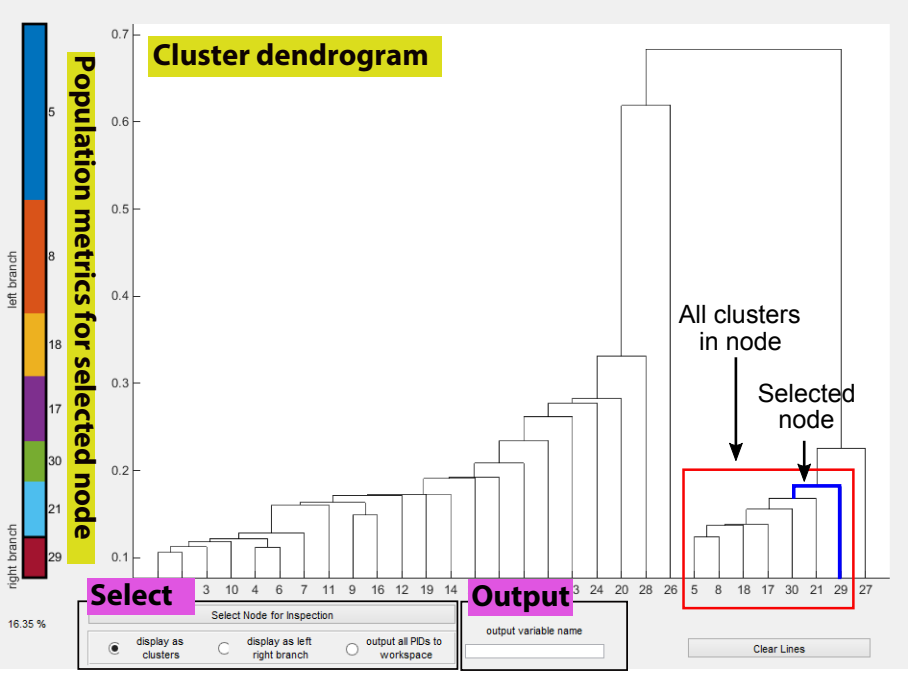

Figure 2. Screen capture of a dendroFATES window showing the cluster tree or dendrogram for 30 input clusters. The cluster contributions to the user-selected node are shown in the plot on the left. The particle data for the selected node are automatically plotted in a guiFATES window (Fig. S1).

grouped by cluster and have been sorted by peak intensity of m/z −35. While users may initially provide guiFATES with a large amount of data, they will likely desire to display smaller selections at a time to enable better visual exploration. This can be accomplished in a number of ways within guiFATES. Users can use mouse clicks to quickly zoom in and out of a single plot using MATLAB's native figure handling capabilities. guiFATES is designed so that when this occurs all plot axes within the GUI are scaled appropriately and instantaneously. Figure S1b is a screen capture where the user utilized this functionality to select the bottom half of the particles in Fig. S1a and also decreased the range of the m/z values displayed. For more complex selections users can enter in filtering parameters so that displayed particles only fall within a desired range of particle metrics, peak intensity of a certain m/z value, or any combination thereof. Figure S1c is a screen capture where the data, sorted by cluster, have been filtered by size (1–2 µm), m/z −35 peak area (0–3000), and clustering statistic (0.8–1). Lastly users can also choose to only display select clusters. Figure S1d is a screen capture utilizing the same filters as in Fig. S1c albeit limiting the display to only clusters 2 and 5.

These visual sorting and filtering methods enable users to efficiently inspect data sets and visually discover mass spectral trends, differences, and similarities both between distinct particle types and within populations of chemically similar particles. Due to the high variability and qualitative nature of single-particle mass spectra generated by laser desorption–ionization techniques, clustering algorithms utilized to group SPMS mass spectra within a data set often do not generate a one-to-one relationship between the number of chemical particle types in the population and spectra clusters generated (e.g., Giorio et al., 2012; Murphy et al., 2003; Rebotier and Prather, 2007; Wenzel and Prather, 2004; Zelenyuk et al., 2006, 2008a). Therefore it is necessary to leverage expert knowledge either to combine multiple spectra clusters, generated algorithmically, into a single chemical particle type or to further split clusters into smaller groups as has been noted in many SPMS studies of unconstrained aerosol populations (e.g., Dall'Osto and Harrison, 2006; Pratt et al., 2009; Qin et al., 2012). The authors emphasize that there is not a consensus on the most suitable algorithms and thresholds for SPMS analysis and suggest users investigate the previously listed references before embarking on mass-spectral-based algorithmic analysis. However, despite the conditions of initial clustering, guiFATES aids this process by allowing users to visualize all clustered particles at once and combine any number of clusters or split any cluster in any location during the data exploration process. Users can choose to output the particle identifiers of any cluster in the guiFATES window to the MATLAB workspace. All plotting, sorting, filtering, and grouping applications of guiFATES have been tested on a set of 100 000 particles with dual-polarity mass spectra, and at this size all updates to the displayed plots occurred nearly instantaneously, making guiFATES an appropriate and efficient tool for the large data sets common to SPMS analysis.

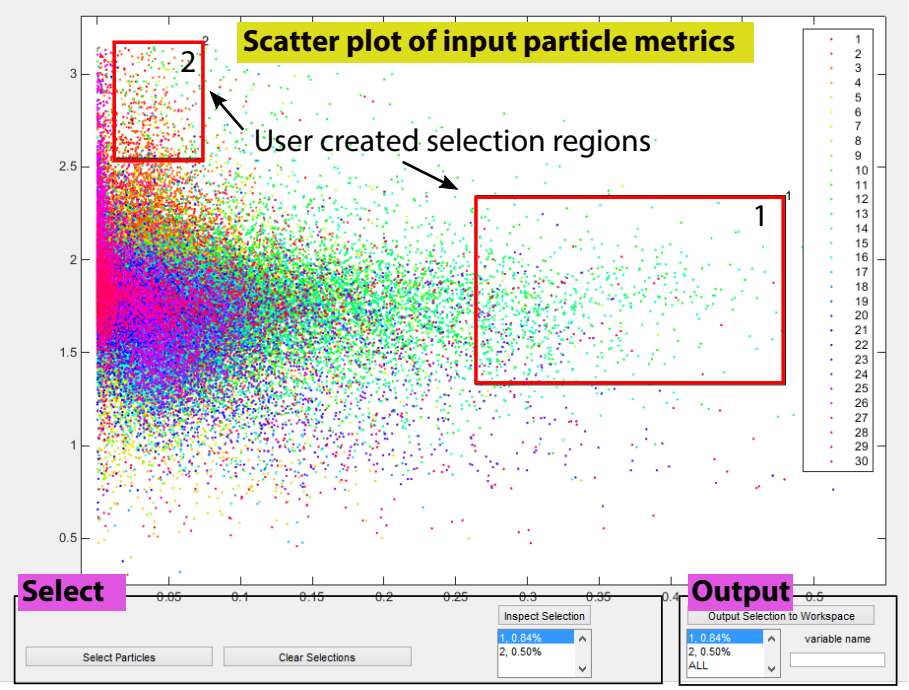

Figure 3. Screen capture of a scatterFATES window showing the -35 to $-93\,m/z$ ratio plotted against particle size for 166 666 particles. Any two particle metrics can be input into scatterFATES. Two regions have also been created by the user for further inspection in guiFATES.

The advantages and benefits of this general method of data visualization and exploration for refining particle clusters have been discussed at length previously (Zelenyuk et al., 2008) and with the publication of FATES will be available to the SPMS community at large. A specific detail of note is that Zelenyuk et al. (2008) demonstrate that discontinuities in the particle cluster size distributions were characteristic of misclassifications of their mass spectra. Because this technique is not dependent on specific ion markers, it has the potential to be effective for a broad range of particle types but is yet to be extensively explored. guiFATES also enables future investigations of the extension of this cluster-discriminating technique to other common particle metrics, such as total ion intensity. Finally many studies have examined the influences of particle and experimental characteristics on the mass spectra generated from particles of uniform composition (e.g., Neubauer et al., 1998; Reinard and Johnston, 2008; Steele et al., 2003; Zelenyuk et al., 2008b). guiFATES can also be utilized in the exploration of these data sets consisting of a single particle type, where algorithmic grouping of particles utilizing mass spectra is unnecessary or even inappropriate.

4.2 dendroFATES: hierarchical cluster relations

FATES also includes two supplementary GUIs which allow the users to graphically select the particles to feed into the guiFATES spectra explorer. dendroFATES is a GUI where the user supplies the clusters and representative cluster mass spectra output from any clustering algorithm of the user's choice. The clusters are then automatically grouped into a cluster tree by a hierarchical analysis performed within MATLAB which is displayed in the dendroFATES GUI window. Hierarchical analyses have been utilized previously with SPMS data sets (Giorio et al., 2012; Hinz et al., 2006; Murphy et al., 2003; Rebotier and Prather, 2007; Zelenyuk et al., 2006), but a brief description is given here. The dendrogram links clusters in a binary fashion, creating new groups which are then further linked. Lower linkage heights indicate a higher degree of similarity between groups, and large distances between levels in the dendrogram are indicative of natural divisions in the data set. Figure 2 is a screenshot of the dendroFATES window with a dendrogram generated from the 30 most populous clusters generated using the ART-2a algorithm to cluster a subset of 166 666 particles from the Bodega Bay data set. Zooming in and out of the dendrogram is handled by MATLAB's native graphics functionality and makes it possible to supply dendroFATES with hundreds of clusters and still explore the cluster tree quickly and intuitively. Because the dendrogram allows the user to easily visualize similarities and natural groupings of clusters generated, it is an excellent tool to select clusters for further exploration of the particle and spectral data using the guiFATES tool. Clicking linkages in dendroFATES automatically opens a guiFATES window displaying all particles belonging to the selected node. When a linkage is selected, the fractional cluster contribution to the selected node is dis-

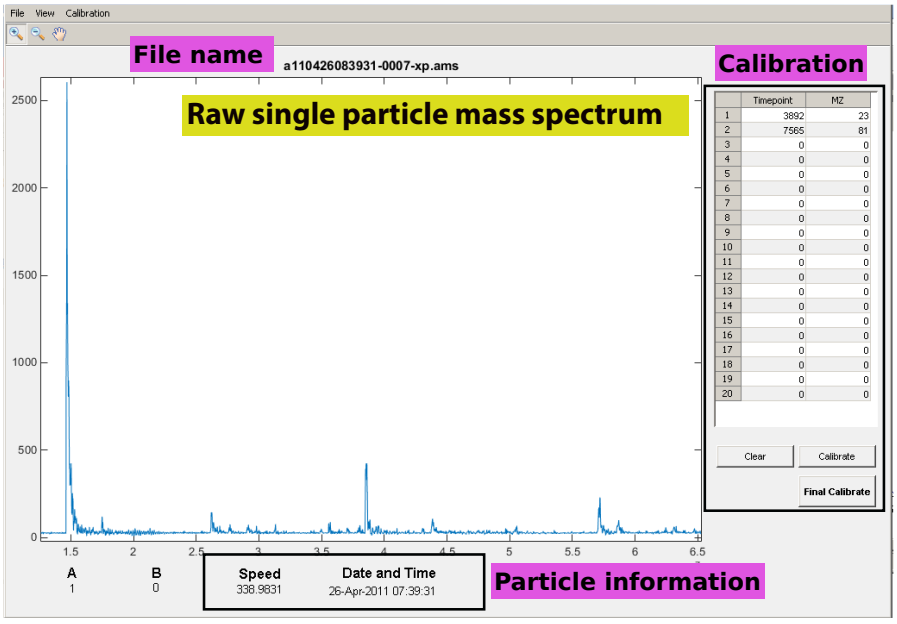

Figure 4. Screenshot of a calibFATES window displaying a single-particle uncalibrated mass spectrum. Calibration data are input and displayed on the right, and particle size and time are displayed on the bottom.

played on the right in the dendroFATES window, and the fraction of the selected node to the total population is also displayed in text. Figure S2 illustrates the guiFATES window generated with the node selection made in Fig. 2 when the user chooses to display particles by their cluster label (Fig. S2a) or grouped by the left and right branch (Fig. S2b). As illustrated in Fig. S2a, when guiFATES is populated by dendroFATES, the clusters are displayed in the same order as displayed in the dendogram. Therefore very similar clusters are adjacent in the guiFATES window, assisting intuitive visual comparisons and combinations of data. Because all FATES GUIs are in MATLAB and the user can also access the data programmatically, it is straightforward and fast for the user to iteratively select clusters from the dendrogram in dendroFATES, refine them in guiFATES, output new clusters to the workspace, and feed the new cluster results back into dendroFATES until the user is satisfied with the grouping of the data set.

4.3 scatterFATES: user-defined particle relations

The complexity of SPMS data sets means there are numerous relationships that could be explored, and predicting all desired comparisons is impossible. scatterFATES is another GUI used to populate guiFATES with user-selected particles. However, rather than grouping particles via clusters as in dendroFATES, scatterFATES creates a scatterplot of particles using any two particle data metrics the user supplies as the axes. The points are then color-coded by cluster or group. Figure 3 is an example scatterFATES window, where the -35 to $-93\,m/z$ ratio is plotted against particle size for the 166 666 particles that had been previously clustered. Once a scatterplot is created in scatterFATES, the user can click on the figure to draw regions within the scatterplot as shown in Fig. 1. All particle data within a created region can then be selected and automatically populated into guiFATES for spectra visualization and exploration.

4.4 calibFATES: raw spectra calibration

FATES has been designed so that all aspects and functionalities of SPMS data analysis and exploration are contained within a single programming environment and language. To this end we developed calibFATES, a GUI to quickly scan through raw spectra data files before importation into FATES and generate calibrations to convert raw time-of-flight spectra to mass-to-charge spectra. calibFATES allows SPMS users to quickly visually examine generated spectra on the fly without any time-consuming processing, even during data acquisition, to ensure the quality and consistency of the data being acquired. While calibFATES is currently written to be able to read the raw spectra files generated by the ATOFMS and TSI ATOFMS, it could be easily modified to read in any raw spectra file (Supplement M-B). Figure 4 is a screenshot of a calibFATES window displaying a single uncalibrated raw spectrum. Users can scan through and display spectra contained in any raw spectra files within the folder. A calibration can be generated by setting selected times to entered m/z values. To generate as accurate a calibration as possible, it is suggested that users choose peaks with a di-

verse set of m/z values that span the SPMS mass spectral range and utilize multiple raw spectra to generate a single calibration. Generating calibration parameters from 20 peaks selected from five spectra has been found to produce generally satisfactory results for ATOFMS data sets. Calibration parameters can be output to a text file for future reference, and any calibration file generated can be loaded into and applied to the raw spectra in calibFATES so that the spectra are displayed as calibrated mass spectra rather than time-of-flight spectra.

5 Conclusions

FATES is the first software package for SPMS data sets to include flexible script-based data analysis and graphical user interfaces for data exploration integrated within a single programming language. Because FATES is designed to be easily extensible to diverse input data formats and implemented completely in MATLAB, a highly documented language popular among scientists, it should be accessible and employable across the SPMS community despite the many independent instrumental designs. SPMS data importation and programmatic and graphical data analyses can be performed quickly in FATES even for large data sets thanks to both speed and memory optimizations and utilization of native MATLAB data types and built-in functions. Within a FATES study data are structured so that complex analyses can be performed using concise code with little reliance on FATES-specific functions. In addition a set of GUIs with many display, sorting, filtering, and grouping functionalities have been developed to assist both expert and novice users to intuitively visualize a complex SPMS data set and create robust particle groupings. For these reasons we believe FATES will greatly improve the efficiency of data processing and knowledge discovery from SPMS data sets.

Code and data availability. The FATES software package (v1.0.0), an extensive manual, and an example data set are available online at doi:10.5281/zenodo.398847 (Sultana et al., 2017), and all future releases will be available at www.github.com/CMSultana/FATESmatlabToolKit. This site is a forum where updates to the code and new functions can be shared amongst the SPMS community.

The Supplement related to this article is available online at doi:10.5194/amt-10-1323-2017-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was funded by the National Science Foundation through the Center for Aerosol Impacts on Climate and the Environment (CHE 1305427). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Edited by: G. Phillips
Reviewed by: two anonymous referees

References

Allen, J. O.: YAADA – Software Toolkit to Analyze Single-Particle Mass Spectral Data: Reference Manual Versions 1.3 and 2.0, Tempe, 2005.

Brands, M., Kamphus, M., Böttger, T., Schneider, J., Drewnick, F., Roth, A., Curtius, J., Voigt, C., Borbon, A., Beekmann, M., Bourdon, A., Perrin, T., and Borrmann, S.: Characterization of a Newly Developed Aircraft-Based Laser Ablation Aerosol Mass Spectrometer (ALABAMA) and First Field Deployment in Urban Pollution Plumes over Paris During MEGAPOLI 2009, Aerosol Sci. Tech., 45, 46–64, doi:10.1080/02786826.2010.517813, 2011.

Carson, P. G., Neubauer, K. R., Johnston, M. V., and Wexler, A. S.: On-line chemical analysis of aerosols by rapid single-particle mass spectrometry Peter, J. Aerosol Sci., 26, 535–545, doi:10.1016/0168-1176(95)04312-8, 1995.

Dall'Osto, M. and Harrison, R.: Chemical characterisation of single airborne particles in Athens (Greece) by ATOFMS, Atmos. Environ., 40, 7614–7631, doi:10.1016/j.atmosenv.2006.06.053, 2006.

Dall'Osto, M., Ceburnis, D., Monahan, C., Worsnop, D. R., Bialek, J., Kulmala, M., Kurtén, T., Ehn, M., Wenger, J., Sodeau, J., Healy, R., and O'Dowd, C.: Nitrogenated and aliphatic organic vapors as possible drivers for marine secondary organic aerosol growth, J. Geophys. Res., 117, D12311, doi:10.1029/2012JD017522, 2012.

Erdmann, N., Dell'Acqua, A., Cavalli, P., Grüning, C., Omenetto, N., Putaud, J.-P., Raes, F., and Dingenen, R. Van: Instrument Characterization and First Application of the Single Particle Analysis and Sizing System (SPASS) for Atmospheric Aerosols, Aerosol Sci. Tech., 39, 377–393, doi:10.1080/027868290935696, 2005.

Gard, E., Mayer, J. E., Morrical, B. D., Dienes, T., Fergenson, D. P., and Prather, K. A.: Real-Time Analysis of Individual Atmospheric Aerosol Particles: Design and Performance of a Portable ATOFMS, Anal. Chem., 69, 4083–4091, doi:10.1021/ac970540n, 1997.

Ge, Z., Wexler, A. S., and Johnston, M. V.: Laser Desorption/Ionization of Single Ultrafine Multicomponent Aerosols, Environ. Sci. Technol., 32, 3218–3223, doi:10.1021/es980104y, 1998.

Giorio, C., Tapparo, A., Dall'Osto, M., Harrison, R. M., Beddows, D. C. S., Di Marco, C., and Nemitz, E.: Comparison of three techniques for analysis of data from an Aerosol Time-of-Flight Mass Spectrometer, Atmos. Environ., 61, 316–326, doi:10.1016/j.atmosenv.2012.07.054, 2012.

Gross, D. S., Gälli, M. E., Silva, P. J., and Prather, K. a: Relative sensitivity factors for alkali metal and ammonium cations in single-particle aerosol time-of-flight mass spectra, Anal. Chem., 72, 416–22, 2000.

Gross, D. S., Atlas, R., Rzeszotarski, J., Turetsky, E., Christensen, J., Benzaid, S., Olson, J., Smith, T., Steinberg, L., Sulman, J., Ritz, A., Anderson, B., Nelson, C., Musicant, D., Chen, L., Snyder, D., and Schauer, J.: Environmental chemistry through intelligent atmospheric data analysis, Environ. Model. Softw., 25, 760–769, doi:10.1016/j.envsoft.2009.12.001, 2010.

Healy, R. M., Hellebust, S., Kourtchev, I., Allanic, A., O'Connor, I. P., Bell, J. M., Healy, D. A., Sodeau, J. R., and Wenger, J. C.: Source apportionment of $PM_{2.5}$ in Cork Harbour, Ireland using a combination of single particle mass spectrometry and quantitative semi-continuous measurements, Atmos. Chem. Phys., 10, 9593–9613, doi:10.5194/acp-10-9593-2010, 2010.

Hinz, K. and Spengler, B.: Instrumentation, data evaluation and quantification in on-line aerosol mass spectrometry, J. Mass Spectrom., 42, 843–860, doi:10.1002/jms.1262TS7, 2007.

Hinz, K., Kaufmann, R., and Spengler, B.: Laser-Induced Mass Analysis of Single Particles in the Airborne State, Anal. Chem., 66, 2071–2076, 1994.

Hinz, K. P., Erdmann, N., Grüning, C., and Spengler, B.: Comparative parallel characterization of particle populations with two mass spectrometric systems LAMPAS 2 and SPASS, Int. J. Mass Spectrom., 258, 151–166, doi:10.1016/j.ijms.2006.09.008, 2006.

Hinz, K. P., Gelhausen, E., Schäfer, K. C., Takats, Z., and Spengler, B.: Characterization of surgical aerosols by the compact single-particle mass spectrometer LAMPAS 3, Anal. Bioanal. Chem., 401, 3165–3172, doi:10.1007/s00216-011-5465-6, 2011.

Klimach, T.: Chemische Zusammensetzung der Aerosole- Design und Datenauswertung eines Einzelpartikel- Laserablationsmassenspektrometers, University of Mainz, 2012.

Lake, D. A., Tolocka, M. P., Johnston, M. V., and Wexler, A. S.: Mass spectrometry of individual particles between 50 and 750 nm in diameter at the Baltimore supersite, Environ. Sci. Technol., 37, 3268–3274, doi:10.1021/es026270u, 2003.

Murphy, D. M., Middlebrook, A. M., and Warshawsky, M.: Cluster Analysis of Data from the Particle Analysis by Laser Mass Spectrometry (PALMS) Instrument, Aerosol Sci. Tech., 37, 382–391, doi:10.1080/02786820300971, 2003.

Neubauer, K. R., Johnston, M. V., and Wexler, A. S.: Humidity effects on the mass spectra of single aerosol particles, Atmos. Environ., 32, 2521–2529, doi:10.1016/S1352-2310(98)00005-3, 1998.

Phares, D. J., Rhoads, K. P., and Wexler, A. S.: Performance of a Single Ultrafine Particle Mass Spectrometer, Aerosol Sci. Tech., 36, 583–592, doi:10.1080/02786820252883829, 2002.

Pratt, K. A., Hatch, L. E., and Prather, K. A.: Seasonal volatility dependence of ambient particle phase amines., Environ. Sci. Technol., 43, 5276–81, 2009.

Qin, X., Pratt, K. A., Shields, L. G., Toner, S. M., and Prather, K. A.: Seasonal comparisons of single-particle chemical mixing state in Riverside, CA, Atmos. Environ., 59, 587–596, doi:10.1016/j.atmosenv.2012.05.032, 2012.

Rebotier, T. P. and Prather, K. A.: Aerosol time-of-flight mass spectrometry data analysis: a benchmark of clustering algorithms, Anal. Chim. Acta, 585, 38–54, doi:10.1016/j.aca.2006.12.009, 2007.

Reinard, M. S. and Johnston, M. V: Ion Formation Mechanism in Laser Desorption Ionization of Individual Nanoparticles, J. Am. Soc. Mass Spectrom., 19, 389–399, doi:10.1016/j.jasms.2007.11.017, 2008.

Sierau, B., Chang, R. Y.-W., Leck, C., Paatero, J., and Lohmann, U.: Single-particle characterization of the high-Arctic summertime aerosol, Atmos. Chem. Phys., 14, 7409–7430, doi:10.5194/acp-14-7409-2014, 2014.

Steele, P. T., Tobias, H. J., Fergenson, D. P., Pitesky, M. E., Horn, J. M., Czerwieniec, G. A., Russell, S. C., Lebrilla, C. B., Gard, E. E., and Frank, M.: Laser Power Dependence of Mass Spectral Signatures from Individual Bacterial Spores in Bioaerosol Mass Spectrometry, Anal. Chem., 75, 5480–5487, 2003.

Steele, P. T., Srivastava, A., Pitesky, M. E., Fergenson, D. P., Tobias, H. J., Gard, E. E., and Frank, M.: Desorption/Ionization Fluence Thresholds and Improved Mass Spectral Consistency Measured Using a Flattop Laser Profile in the Bioaerosol Mass Spectrometry of Single Bacillus Endospores, Anal. Chem., 77, 7448–7454, 2005.

Su, Y., Sipin, M. F., Furutani, H., and Prather, K. A.: Development and Characterization of an Aerosol Time-of-Flight Mass Spectrometer with Increased Detection Efficiency, Anal. Chem., 76, 712–719, doi:10.1021/ac034797z, 2004.

Sultana, C., Cornwell, G., and Rodriguez, P.: KPrather-Lab/FATESmatlabToolKit: Version 1 of FATES (v1.0.0), Data set, Zenodo, doi:10.5281/zenodo.398847, 2017.

Thomson, D. S., Schein, M. E., and Murphy, D. M.: Particle Analysis by Laser Mass Spectrometry WB-57F Instrument Overview, Aerosol Sci. Tech., 33, 153–169, doi:10.1080/027868200410903, 2000.

Trimborn, A., Hinz, K.-P., and Spengler, B.: Online Analysis of Atmospheric Particles with a Transportable Laser Mass Spectrometer, Aerosol Sci. Tech., 33, 191–201, doi:10.1080/027868200410921, 2000.

Wenzel, R. J. and Prather, K. A.: Improvements in ion signal reproducibility obtained using a homogeneous laser beam for on-line laser desorption/ionization of single particles, Rapid Commun. Mass Spectrom., 18, 1525–1533, doi:10.1002/rcm.1509, 2004.

Zelenyuk, A. and Imre, D.: Single Particle Laser Ablation Time-of-Flight Mass Spectrometer: An Introduction to SPLAT, Aerosol Sci. Tech., 39, 554–568, doi:10.1080/027868291009242, 2005.

Zelenyuk, A., Imre, D., Cai, Y., Mueller, K., Han, Y., and Imrich, P.: SpectraMiner, an interactive data mining and visualization software for single particle mass spectroscopy: A laboratory test case, Int. J. Mass Spectrom., 258, 58–73, doi:10.1016/j.ijms.2006.06.015, 2006.

Zelenyuk, A., Imre, D., Nam, E. J., Han, Y., and Mueller, K.: ClusterSculptor: Software for expert-steered classification of single particle mass spectra, Int. J. Mass Spectrom., 275, 1–10, doi:10.1016/j.ijms.2008.04.033, 2008a.

Zelenyuk, A., Juan, Y., Chen, S., Zaveri, R. A., and Imre, D.: "Depth-profiling" and quantitative characterization of the size, composition, shape, density, and morphology of fine particles with SPLAT, a single-particle mass spectrometer, J. Phys. Chem. A, 112, 669–671, doi:10.1021/jp077308y, 2008b.

Zelenyuk, A., Yang, J., Choi, E., and Imre, D.: SPLAT II: An Aircraft Compatible, Ultra-Sensitive, High Precision Instrument for In-Situ Characterization of the Size and Composition of Fine and Ultrafine Particles, Aerosol Sci. Tech., 43, 411–424, doi:10.1080/02786820802709243, 2009.

Zelenyuk, A., Imre, D., Wilson, J., Zhang, Z., Wang, J., and Mueller, K.: Airborne single particle mass spectrometers (SPLAT II & miniSPLAT) and new software for data visualization and analysis in a geo-spatial context, J. Am. Soc. Mass Spectrom., 26, 257–270, doi:10.1007/s13361-014-1043-4, 2015.

Zhang, G., Han, B., Bi, X., Dai, S., Huang, W., Chen, D., Wang, X., Sheng, G., Fu, J., and Zhou, Z.: Characteristics of individual particles in the atmosphere of Guangzhou by single particle mass spectrometry, Atmos. Res., 153, 286–295, doi:10.1016/j.atmosres.2014.08.016, 2015.