A Machine-Learning Approach to Aerosol Classification for Single-Particle Mass Spectrometry

by

Costa Christopoulos

Submitted to the Department of Earth, Atmospheric and Planetary Sciences

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Earth, Atmospheric and Planetary Sciences

at the Massachusetts Institute of Technology

May 10, 2017 [June 2017]

Author___ Signature redacted

Department of Earth, Atmospheric and Planetary Sciences
May 16, 2017

Certified by_ Signature redacted

[Supervisor's Name]
Thesis Supervisor

Accepted by___ Signature redacted

Richard P. Binzel
Chair, Committee on Undergraduate Program

Abstract

Compositional analysis of atmospheric and laboratory aerosols is often conducted via single-particle mass spectrometry (SPMS), an *in situ* and real-time analytical technique that produces mass spectra on a single particle basis. In this study, machine learning classification algorithms are created using a dataset of SPMS spectra to automatically differentiate particles on the basis of chemistry and size. While clustering methods have been used to group aerosols into broad categories based on similarity, these models do not incorporate known aerosols labels and are not explicitly formulated for classification. Furthermore, traditional methods often rely on a smaller set of well-known, important variables whereas the proposed method is more general and flexible, allowing researchers to automatically quantify and select important variables from any aerosol subset. In this work, machine learning algorithms build a predictive model from a training set in which the aerosol type associated with each mass spectrum is known. Several such classification models were created to differentiate aerosol types in four broad categories: fertile soils, mineral/metallic particles, biological, and all other aerosols using ~40 common positive and negative spectral features. For this broad categorization, machine learning resulted in a classification accuracy of ~93%. More complex models were developed to classify aerosols into specific categories which resulted in a classification accuracy of ~87%. The trained model was then applied to a 'blind' mixture of aerosols with model agreement on the presence of secondary organic aerosol, coated and uncoated mineral dust and fertile soil. Additionally, the model is used to characterize an ambient atmospheric dataset collected from the free troposphere.

1. Introduction

The interaction of atmospheric aerosols with clouds and radiation contributes to the uncertainty in determinations of both anthropogenic and natural climate forcing [Boucher et al., 2013; Lohmann and Feichter, 2005]. Aerosols directly affect atmospheric radiation by scattering and absorption of radiation from both solar and terrestrial sources. The radiative forcing from particulates in the atmosphere depends on optical properties that vary significantly among different aerosol types [Lesins et al., 2002]. Aerosols also indirectly affect climate via their role in the development and maintenance of clouds [Vogelmann et al., 2012; Lubin et al., 2006]. Ultimately, the formation, appearance, and lifetime of clouds are sensitive to aerosol properties like shape, chemistry, and morphology [Lohmann and Feichter, 2008]. Characterization of aerosol properties, therefore, plays a vital role in understanding weather and climate.

The chemical composition and size of aerosols has been analyzed on a single particle basis *in situ* and in real-time using single particle mass spectrometry (SPMS; Murphy [2007]). First developed ~2 decades ago, SPMS permits the analysis of aerosol particles in the 150 – 3000 nm size range, while differentiating internal and external aerosol mixtures and characterizing both volatile (e.g. organics and sulfates) and refractory (e.g. crystalline salts, elemental carbon and mineral dusts) particle components. Particles are typically desorbed and ionized with a UV laser and resultant ions are detected using time-of-flight mass spectrometry [Murphy, 2007]. A complete mass

spectrum of chemical components is produced from each analyzed aerosol particle [Coe et al., 2006]. Despite universal detection of components found in atmospheric aerosols, SPMS is not normally considered quantitative without specific laboratory calibration [Cziczo et al., 2001].

Aerosols with different properties can appear similar in the context of SPMS. For example, fly ash spectra and mineral dust contain peaks corresponding to sulfates, phosphates, metals, and metal oxides despite different origins and emission sources [Garimella et al., 2017]. This complicates analysis of aerosol populations because their properties need to be well-defined in order to increase agreement between models and observations [Niemand et al., 2012; Hoose and Möhler, 2012; Welti et al., 2009]. Even minor compositional changes can be atmospherically important. As one example, mineral dusts are known to be effective at nucleating ice clouds [Cziczo et al., 2013]. Particles in the atmosphere undergo chemical and morphological changes as they age and eventually contain material from several sources [Boucher et at. 2013]. Despite minor addition of mass, aged mineral dust is less suitable for ice formation [Cziczo et al., 2013], but these particles then act as cloud condensation nuclei and participate in warm cloud formation [Andreae et al., 2008]. As a second example, ice nucleation in mixed-phased clouds has been suggested to be predominantly influenced by feldspar, a single component among the diverse mineralogy of atmospheric dust [Atkinson et al., 2013].

Here we show that supervised training and a rule-based probabilistic classification of a decision tree ensemble can be used for differentiation of SPMS spectra. Various clustering methods have been used to group aerosol types [Murphy et al., 2003; Gross et al., 2008] but these algorithms are known to struggle with chemically-similar aerosols as

they do not incorporate known particle labels in the training process. Such 'unsupervised' clustering algorithms automatically group unlabeled data points on the basis of a specified distance metric in feature space, in this case mass spectral features. For the purposes of setting broad aerosol categories, which are chemically similar and easily separable in feature space, clustering is the simpler tool and the data easier to interpret. For identifying new or potentially unexpected atmospheric aerosols, such properties are desirable; however, the advantages of clustering greatly diminish when considering similar particles types that overlap in feature space. Fertile soils, for instance, are often grouped into a single category despite different sources and atmospheric histories. Clustering algorithms should therefore be considered as a tool to use alongside supervised classification. The latter may be used to further explore unique aerosol types or verify manually labeled clusters with higher precision. Furthermore, the ensemble approach presented here also produces variable rankings and probabilistic predictions that help address measurement uncertainty.

In this study, we demonstrate the capabilities of machine learning to automatically differentiate particles on the basis of chemistry and size. The resulting model can capture minor compositional differences between aerosol mass spectra. By testing predictions using independent, or 'blind', datasets, we illustrate the feasibility of combining on-line analysis techniques such as SPMS with machine learning to infer the behavior and origin of aerosols in the laboratory and atmosphere.

2. Methodologies

2.1 PALMS

The Particle Analysis by Laser Mass Spectrometry (PALMS) instrument was employed for these studies. PALMS has been described in detail previously [Cziczo et al. 2006]. Briefly, the instrument samples aerosol particles in the size range from ~200 to ~3000 nm using an aerodynamic lens inlet into a differentially-pumped vacuum region. Particle aerodynamic size (hereafter "deltayagtime") is acquired by measuring particle transit time between two 532 nm continuous wave neodymium-doped yttrium aluminum garnet (Nd:YAG) laser beams. A pulsed UV 193 nm excimer laser is used to desorb and ionize the particles and the resulting ions are extracted using a unipolar time-of-flight mass spectrometer. The resulting mass spectra correspond to single particles. The UV ionization extracts both refractory and volatile components and allows analysis of all chemical components present in atmospheric aerosol particles [Cziczo et al. 2013]. SPMS is not normally considered quantitative since the ion abundance measured with the mass spectrometer depends on ionization efficiency by the UV laser rather than an absolute concentration of chemical species [Cziczo et al. 2013].

2.2 Dataset

A set of 'training data' was acquired by sampling atmospherically-relevant aerosols. The majority of the dataset was acquired at the Karlsruhe Institute of Technology (KIT) Aerosol Interactions and Dynamics in the Atmosphere (AIDA) facility during Fifth Ice Nucleation workshop—part 1 (FIN01) and the remainder at the Aerosol and Cloud Laboratory at MIT. The FIN01 workshop was an intercomparison effort of ~10 SPMS instruments, including PALMS. The training data correspond to spectra of known particle types that were aerosolized into KIT's AIDA and an auxiliary chamber

(NAUA) for sampling by PALMS and the other SPMSs (Table 1). The number of training spectra acquired varied by particle type, ranging from ~250 for secondary organic aerosol (SOA) to ~1500 for potassium-rich feldspar ("K-feldspar"). In total, ~50,000 spectra are considered with each spectrum containing 512 possible mass peaks and a deltayagtime (Table 2). Additionally, the FIN01 workshop included a blind sampling period, where the NAUA chamber was filled with 3 - 4 types of unknown aerosol at unknown concentrations for those previously sampled (i.e., for which spectra had already been acquired).

Figure 1 illustrates the simplistic differentiation of particles using only two mass peaks in one (negative) polarity. Mass peaks represent fractional ion abundance, measured as a normalized total ion current. In this example, the normalized areas of negative mass peaks 24 (C_2^-) and 16 (O^-) are plotted. Distinct aerosol types are differentiated by color with clusters forming in two-dimensional space. Note that spectra of the same aerosol type form distinct clusters (e.g. Arizona Test Dust, ATD), as do similar aerosol classes (e.g., soil dusts). Co-plotted in Figure 1 are data from the blind experiment. Distinct clusters of spectra from the blind experiment are noticeable in Figure 1. Described in the next section, machine learning algorithms draw "decision boundaries" that best separate different groups of data points based on set rules. Machine learning is not bound by the simplistic two dimensional space shown in Figure 1 and can instead use all 512 mass peaks and deltayagtime.

2.2 Automatic Aerosol Classification

Machine-learned aerosol classification models map a continuous input vector X to a discreet output value using a set of parameters 'learned' from the data. Figure 2

illustrates the mapping of a mass spectrum to vector X space. In contrast to traditional, hard-coded, rule-based classification methods, machine learning automatically determines parameters that partition the data set. To form X, mass spectra are converted to dimensional vectors normalized to the total ion current (i.e., the total of all mass peaks sum to 1 in each spectra). While 512 mass peaks are available the treatment here uses the first ~210 (i.e., hydrogen through lead). The elements of this vectorized mass spectrum hold information about the ionization efficiency and relative abundance of chemical species in each aerosol and serve as the variables for the machine learning model.

Machine learning is conducted in two phases: training and testing. During training, a model is constructed and iteratively updated based on data (i.e., mass spectra) from the training set. For this work, a set of known aerosol types sampled by PALMS was converted to dimensional vectors. These data form the basis set for defining each aerosol type. An ensemble of decision trees was used to then generate predictions of aerosol type. A single decision tree is a statistical decision model that performs classification based on a series of comparisons relating a variable X_i (in this case a normalized mass peak in X) to a learned threshold value [Breiman, 2001]. Represented as an algorithmic tree, a binary decision tree consists of a hierarchy of nodes where each node connects via branches to two other nodes deeper in the tree. At each node, one of the two branches is taken based on whether a normalized peak X_i is greater or less than a threshold value. Each branch leads to another node where a different test is performed. After a series of tests, one at each node, a class is assigned to a given sample; this is a so-called 'leaf'. Figure 2 illustrates the classification model for a single decision tree.

Each test in the tree narrows the set of reachable output leaves and thus the

sample space of possible aerosol labels. After h tests in this study, where h ranges from 10 to 3000, the set of reachable leaves and possible labels is 1 and the decision tree outputs that prediction. Because PALMS is unipolar – either a positive or negative mass spectrum is produced – simultaneous generation of positive and negative spectra on a particle-by-particle basis is not possible. Two separate classification models, one for each polarity, are therefore generated to classify aerosols. These are hereafter referred to as the 'positive' and 'negative classification algorithms'.

2.3 Decision Tree Ensembles

An ensemble consists of a collection of classifiers where each independently labels a spectrum vector X. To make a final prediction of aerosol type, decision trees within an ensemble 'vote' on a classification label. Each vote has equal weight and the spectrum is assigned to the majority choice. Each tree within an ensemble is independently grown on a subset of the training data so that a commonly voted label implies a higher certainty. Adding members to an ensemble increases the robustness of a classification model by providing alternative hypotheses and is therefore preferable to single classifiers.

Before an ensemble method is implemented for classification, trees are independently grown during training. A total of k trees, with $k = 1000$, were grown using a bootstrap sample from the training set. In bootstrap sampling, each tree sees an independent sample set of equal size drawn from the full training set by sampling spectra with replacement. On average, each tree is built with ~63% of the data. The unsampled

data, known as 'out-of-bag' observations, provide a means to assess classification error for each tree during the training process.

Given a bootstrap sample, a binary decision tree is grown by sequentially creating tests that maximize the separation between classes in parameter space. A test is created by defining a comparison that minimizes the information entropy of a possible split, thus minimizing the randomness of prediction labels [Breiman, 1996]. To generate variability in the model, a best split is chosen among a random set of possible splits at each node on the basis of entropy [Breiman, 2001]. After iteratively defining thresholds for each new node, the tree grows in size until a series of tests ending at some node S_q uniquely characterizes an aerosol as a particle type. A leaf is then appended to node S_q with the corresponding label. In classification mode, an aerosol spectrum that passes the same tree will undergo the same series of tests and will end in the same leaf, thus being labeled in the same way. For the purposes of this study, each tree had ~3,300 nodes.

2.3 Dimensionality Reduction and Chemical Feature Selection

Dimensionality reduction involves representing data with fewer variables than initially present in the dataset, in this case less than the original 512 mass peaks. In addition to facilitating data visualization and limiting overfitting [Mjolsnes, 2001], dimensionality reduction in the context of aerosol mass spectra points to the most important chemical makers for differentiation. Because a majority of the 512 mass peaks were not significant for classification, it was important to first identify variables with predictive accuracy. Variable ranking was algorithmically determined by comparing the performance of trees before and after removing information about peak X_i. The method is

that the values of variable X_i is permuted for tree k in the out-of-bag set so that the variable is irrelevant to the final label. The change in misclassification before and after the permutation is calculated and then repeated for all trees so that a variable ranking is obtained [Breimann, 2001]. Table 2 ranks mass peaks by polarity in importance using this method. The columns at left list variable rankings (i.e., most to least important for correct classification) for the entire set of aerosol types (labels). The columns at right list rankings when aerosol types are grouped into broader, chemically similar, categories. Both the specific aerosol label and broad aerosol category models were retrained using the subset of the initial variables. The final dimensionality was determined by sequentially adding variables in order of decreasing rank and observing classification performance response. All variables preceding two e-foldings in classification error were maintained in the final model.

3. Results

3.1 Confusion Matrices and Probabilistic Model Performance

A confusion matrix captures misclassification tendencies by pair-wise matching model prediction with true aerosol labels [Powers, 2007]. Confusion matrices represent model predictions as columns i and true aerosol labels as rows j, where class names are mapped to integers $i, j \in \{1,2, \ldots, y\}$. In this study, matrices have been normalized along each column to show the fraction of aerosols labeled as j that actually belong to i (Figure 3). For aerosol classification, these matrices can also be interpreted as similarity measures between particle types. Since the basis of decision tree classification is mathematical separation of physical quantities, misclassifications result from similarity in

mass peaks and their ion abundance between aerosol types. This is most easily visualized as overlapping clusters in the simple two dimensional space in Figure 1.

Because the size of the set is large (~22,300), the general classification behavior can be quantified in term of conditional probability. If $\hat{Y}_i$ is the set of predicted aerosol spectra with aerosol label i and Y_j is the corresponding set of true spectrum-label pairs for label j, then the conditional probability of assigning an aerosol to label i given a predicted label j is given by:

$$p(i \mid j) = \frac{|Yj \cap \hat{Y}i|}{|Yj|} \qquad (1)$$

C is the raw confusion matrix of spectrum counts and $p(i \mid j)$ is the conditional probability distribution over all true aerosol labels i, conditioned on some model-generated label j. To obtain matrix P, which encodes $p(i \mid j)$ for all possible labeling, columns of C are normalized with respect to the total aerosol counts for each label with Eq. 1.

Model performance for each aerosol is summarized in the diagonal elements of P, which represent the fraction of aerosol in column j labeled correctly. The classification accuracy (a) is given by averaging diagonal elements of P. A perfect classification model produces the identity matrix, as all data points are classified correctly 100% if the time. For example, in the positive confusion matrix, SOA and Agar growth medium are correctly labeled in the test set 100% of the time. Barring element truncation, all columns of P add to 1.

Figures 3 and 4 display confusion matrices as heat maps for the full set of particle labels and broad grouped particle categories, respectively. Broad grouped categories are delineated in Figure 3 as fertile soil (Argentinian, Chinese, Ethiopian, Moroccan and two

German soils), pure mineral dust and fly ash (ATD, illite NX, fly ash, Na-feldspar, K-feldspar), other (K-feldspar with sulfuric acid (SA) and SOA coatings, soot, and SOA) and biological (Agar growth medium, *P. syringae* bacteria, cellulose, Snomax, and hazelnut pollen). Model confusion exists between fertile soils and coated/uncoated feldspars which can be explained since soils are mineral dust mixed with organic and other materials.

Positive mass spectra appear to hold more information with respect to differentiating aerosols than negative. Label-wise classification accuracy for the negative algorithm ranges from 3-5% lower. A large part of this performance discrepancy is due to greater ability of positive spectra to differentiate coated particles within the 'other' category.

In addition to quantifying misclassification tendencies between classes, the confusion matrix can be redefined to show confusion for aerosols within classes themselves. Intraclass misclassification analysis is accomplished by considering smaller portions of C and using the same probabilistic assumptions highlighted for the full confusion matrices to form modified probability distributions. The full confusion matrix is partitioned into submatrices representing confusion in a specific aerosol category and renormalized with respect to matrix columns. L is the subset of particle labels of a broader set of aerosols. Integrating the full conditional probability distribution over labels that are impossible to observe gives the probability distribution over members of L:

$$P_l(i,j) = p(i \in L \mid j \in L) = \frac{C(i \in L, j \in L)}{\sum_{i' \in L} C(i', j \in L)} \qquad (2)$$

For example, to determine $P_l(i|j)$ for fertile soils, a submatrix is formed by collecting spectrum counts in the first 6 rows and columns of the full confusion matrix (Figure 3). Column normalization is then applied to derive a probability distribution over labels in the fertile soil category, conditioned on the aerosol actually being a fertile soil. This analysis is repeated over all categories in both models. Finally, the relative performance of both models is isolated and considered with respect to each specific aerosol category.

The precision score [Powers, 2007] captures the classification behavior for some subset of aerosol L by averaging fractions of correctly classified aerosols for labels within that category:

$$\text{Precision Score(L)} = \frac{1}{|L|}\sum_{i=j}^{|L|} P(i \in L, j \in L) \qquad (3)$$

When applied to P_l, the precision score captures classification performance on a population with only aerosol labels contained in L. The algorithm is expected to correctly label an aerosol in such a population with a probability equal to the precision score. The precision score is valuable when using the classification model as a particle screener, producing probability distributions over a subset of aerosol labels of interest. The confusion characteristics are shown in Table 3 for each category in terms of the precision score and the mean and standard deviation of misclassification within each category. Although both models perform similarly for biological spectra, discrepancies of 2-5% appear in the remaining categories. For regimes consisting of only mineral/metallic or other particles, the positive algorithm shows intraclass performance advantages in terms of the precision score, but most notably in terms of fewer mislabeling of mineral/metallic particles. The largest precision discrepancy is observed for fertile soils, where the

positive ion algorithm has a 5% advantage in precision with approximately half the false labeling rate.

3.2 Characterization of Blind Data

As part of the FIN01 workshop, 3 - 4 aerosol types from Table 1 were aerosolized into the NAUA chamber. PALMS, one member of this blind intercomparison effort, collected ~25,000 spectra.

The presence or absence of particle types in the blind set was initially diagnosed by choosing particles predicted at or above the 1% level. We note here that this step was based on the knowledge that (1) a distinct set of particles would be placed in the chamber and (2) particles present at or below the 1% level were most likely contamination. We further note that this step is unique to a blind study and would not be applicable to the atmosphere. Normalized confusion matrices were redefined for the aerosols in the population (i.e., those above the 1% level), which forms the labels of set L in Eq. 2. Finally, particle counts are re-computed by reassigning particle labels based on the modified confusion matrix. For each particle label j, a fraction $n' = P(i \mid j)$ of particles labeled as j are reassigned to i. This probabilistic correction accounts for aerosol mislabeling tendencies observed during testing, producing statistics that better represent the underlying aerosol population. The expected fraction of particles belonging to label i (denoted $\hat{n}i$) is given by:

$$< \hat{n}i > = \frac{< n_i >}{|n|} = \frac{1}{|n|} \sum_j P(i \mid j) \; |n_j| \qquad (4)$$

where n is a set containing all blind spectra and n_j is the set of particles labeled as j.

Figure 5 illustrates the results after this step, where the bottom charts show corrected fractional percentages for each aerosol category. Because SOA was nearly always labeled correctly (Figure 3), the remaining aerosols are considered separately using the full set of candidate aerosol labels. Both positive and negative models arrived at similar results, with inconsistencies primarily associated with the presence of trace fertile soils and mineral dust / fly ash particles. The positive algorithm identifies ~2-4% of the AIDA population as each Argentinean soil, German soil, ATD, and cellulose whereas the frequency of these aerosols was too low to consider in the negative. Alternatively, the negative model estimates Na-Feldspar at ~8% of the total population, a label not identified by the positive algorithm. This discrepancy can be explained by the 1% selection criterion for aerosols present in the population. Fertile soils, ATD, and cellulose frequently accumulate error along rows in the full positive confusion matrix, indicating frequent confusion with other categories (Figure 3). Furthermore, with the observed misclassification rates ranging ~1-4%, it is expected that these aerosol labels are false positives. The negative model offers an alternative hypothesis, suggesting these miscellaneous aerosols are Na-feldspar. Since there is significant model agreement on the percentages of SOA, K-Feldspar, and coated feldspars, this part of the blind mixture population (~90%) can be characterized with most certainty. For the disputed aerosol labels, more credence is lent to the negative classification algorithm on the basis of improved precision for fertile soils.

The aerosols reported in the blind mixture were soot, K-Feldspar, and SOA. Because the soot aerosols were below the cutoff diameter for PALMS, they were not measured by the instrument above the 1% level and therefore were not considered by the

model. Both algorithms robustly labeled SOA with large agreement, consistent with the 100% accuracy observed in the test set. Furthermore, SOA coated K-Feldspar (from coagulation) was correctly identified. While both models incorrectly identified fertile soils, these results are largely consistent with the known uncertainties highlighted by the confusion matrices discussed previously. Given the presence of K-Feldspar, some confusion with fertile soils, SA coated Feldspar, and Na-Feldspar is expected (Figure 3). As discussed previously [Gallavardin et al., 2008], AIDA and NAUA backgrounds are not completely particle-free. During FIN01 study, contamination particles from previous test aerosol were frequently observed as background and they could be the origin of some particles matching fertile soil chemistry.

3.2 Characterization of FIN03 Data

FIN03 -the latest phase of the FIN campaign - establishes an intercomparison of sampling methods for ice nucleating particles collected from the atmosphere. The FIN03 dataset used for this study consists of ~26,000 negative spectra of aerosols sampled directly from the ambient troposphere at Storm Peak Observatory in Steamboat, Co during September, 2015. Unlike the FIN01 experiment, FIN03 contains numerous particles that are absent from the training set, so care is taken to select for ambient particles that fall into regions of the subspace that were trained on. The training set should contain roughly 10% of the atmospheric aerosol types in FIN 03, and any aerosol type not trained on will just be labeled as a known aerosol.

To account for this, a two-dimensional probability density is estimated for the training data in a space defined by area 16 (O-) and area 24 (C2-). This is done using Kernel density estimation - a machine learning method that approximates a

multidimensional probability density function based on data in a non-parametric manner. The bandwidth hyperparameter, which sets the scale for smoothing and regularizes the final distribution, does not significantly affect predictions around the values chosen. For the purposes of this study, this was chosen to be 0.079 for Peak 16 and 0.1210 for Peak 24. Once defined, the probability density of belonging to the training set is defined as a non-linear function of area 16 and area 24. To filter FIN03 data, a fixed probability density threshold of 0.3 is selected, defining a 2D contour that may be used for screening. This contour contains ~80% of the original training data and filters out 84% of the original FIN 03 data. Figure 6 shows several probability density contours along with the training data used for fitting the function

Because of the uncertainties involved, the negative model is used with broad, four-category classification. FIN03 samples that have a probability density > 0.3 are selected for, classified, and shown in figure 7 as a pie chart. While the model identifies significant amounts of fertile soil, mineral/metallic, and other particles, biologicals made up less than 1% of the identified aerosols. It is here noted that because of the subspace selection problem described previously, percentage calculations are expected to be less robust and more qualitative than for FIN 01. Unidentified FIN03 particles are likely organics, sulfates, and nitrates, which are among the aerosol types not considered in this study.

4. Conclusions and Future Work

The machine learning approach described here allows for differentiation of aerosols within a SPMS dataset, augmenting existing tools and reducing the need for a qualitative comparison between mass spectra. This study lays out a framework for training and implementing an ensemble classification model and interpreting results in the context of laboratory and atmospheric aerosol populations. Across a representative sample of possible aerosol types, the behavior of each algorithm predictably allows users to infer the presence or absence of specific aerosols and quantify aerosol abundance. Machine learning is automated and the output of the model must then be informed by human knowledge of aerosol chemistry. Machine learning should therefore be considered as an additional tool to interpret mass spectra to better distinguish aerosols with unique properties in terms of atmospheric chemistry, biogenic cycles, and population health.

The ensemble decision tree classification framework described here may be generalized to any instrument, or set of instruments, capable of collecting physical and chemical information that distinguishes particles. Although the method described here is applied to a stand-alone SPMS and tested with a set of 'blind' data, ancillary laboratory or field data can be integrated to expand the data set. The success of these algorithms is data-dependent, where better performance is expected for instruments that provide more, and more quantitative, analysis of the aerosol properties. Although the algorithms implemented in this study were primarily used to categorize SOA, mineral dust, fertile soil and biological aerosols, these models can adopt an arbitrary large set of aerosol data.

4. Acknowledgements

I would like to greatly acknowledge Maria Zawadowicz for providing instrument data as well as information about PALMS and particle chemistry, Dan Cziczo for years of guidance, resources, and support, and Sarvesh Garimella for programming help and support.

References

Andreae, M. & Rosenfeld, D.: Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols, Earth-Sci. Rev., 89, 13-41, doi:10.1016/j.earscirev.2008.03.001, 2008.

Atkinson, J., Murray, B., Woodhouse, M., Whale, T., Baustian, K., & Carslaw, K., Dobbie, S., O'Sullivan, D., and Malkn, T. L: The importance of feldspar for ice nucleation by mineral dust in mixed-phase clouds, Nature, 498, 355-358, doi:10.1038/nature12278, 2013.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M. , Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S.K., Sherwood, S., Stevens B., and Zhang, X. Y.,: Clouds and Aerosols, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 5, 571-657, 2013.

Breiman L.: Bagging Predictors. Machine Learning, 24, 123-140, 1996.

Breiman L.: Random Forests. Machine Learning, 45, 5-32, 2001.

Coe, H., Allan, J. D.: In Analytical Techniques for Atmospheric Measurement; Heard, D. E., Ed., Blackwell Publishing, 265–311, 2006.

Cziczo, D., Thomson, D., Thompson, T., DeMott, P., and Murphy, D.: Particle analysis by laser mass spectrometry (PALMS) studies of ice nuclei and other low number density particles, Int. J. Mass. Spectrom., 258, 21-29, 2006.

Cziczo, D. J., Froyd, K., Hoose, C., Jensen, E., Diao, M., Zondlo, M., Smith, J. B., Twohy, C. H., and Murphy, D. M.: Clarifying the Dominant Sources and Mechanisms of Cirrus Cloud Formation, Science, 340, 1320-1324, doi:10.1126/science.1234145, 2013.

Cziczo, D. J., Thomson, D. S., and Murphy, D. M.: Ablation, flux, and atmospheric implications of meteors inferred from stratospheric aerosol, Science, 291 (5509), 1772–1775, 2001.

Gallavardin, S., Lohmann, U., and Cziczo, D.: Analysis and differentiation of mineral dust by single particle laser mass spectrometry, Int. J. Mass. Spectrom., 274, 56-63, doi:10.1016/j.ijms.2008.04.031, 2008.

Gallavardin, S. J., Froyd, K. D., Lohmann, U., Moehler, O., Murphy, D. M., Cziczo, D. J.: Single Particle Laser Mass Spectrometry Applied to Differential Ice Nucleation Experiments at the AIDA Chamber, Aerosol Sci. Tech., 42, 773-791, doi: 10.1080/02786820802339538, 2008.

Garimella, S., Wolf, M. J., Christopoulos, C. D., Zawadowicz, M. A., and Cziczo, D. J.: Measuring the cloud formation potential of fly ash particle, Atmos. Chem. Phys. (in prep)

Gross, D., Atlas, R., Rzeszotarski, J., Turetsky, E., Christensen, J., Benzaid, S., Olson, J., Smith, T., Steinberg, L., and Sulman, J.: Environmental chemistry through intelligent atmospheric data analysis, Environ. Modell. Softw., 25, 760-769, 2008.

Henning, S., Ziese, M., Kiselev, A., Saathoff, H., Möhler, O., Mentel, T. F., Buchholz, A., Spindler, C., Michaud, V., Monier, M., Sellegri, K. and

Stratmann, F.: Hygroscopic growth and droplet activation of soot particles: uncoated, succinct or sulfuric acid coated, Atmos. Chem. Phys., 12(10), 4525–4537, doi:10.5194/acp-12-4525-2012, 2012.

Hoose, C. and Möhler, O.: Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments, Atmos. Chem. Phys., 12, 9817-9858, doi:10.5194/acpd-12-12531-2012, 2012.

Hiranuma, N., Augustin-Bauditz, S., Bingemer, H., Budke, C., Curtius, J., Danielczok, A., Diehl, K., Dreischmeier, K., Ebert, M., Frank, F., Hoffmann, N., Kandler, K., Kiselev, A., Koop, T., Leisner, T., Möhler, O., Nillius, B., Peckhaus, A., Rose, D., Weinbruch, S., Wex, H., Boose, Y., Demott, P. J., Hader, J. D., Hill, T. C. J., Kanji, Z. A., Kulkarni, G., Levin, E. J. T., McCluskey, C. S., Murakami, M., Murray, B. J., Niedermeier, D., Petters, M. D., O'Sullivan, D., Saito, A., Schill, G. P., Tajiri, T., Tolbert, M. A., Welti, A., Whale, T. F., Wright, T. P. and Yamashita, K.: A comprehensive laboratory study on the immersion freezing behavior of illite NX particles: A comparison of 17 ice nucleation measurement techniques, Atmos. Chem. Phys., 15(5), doi:10.5194/acp-15-2489-2015, 2015a.

Hiranuma, N., Möhler, O., Yamashita, K., Tajiri, T., Saito, A., Kiselev, A., Hoffmann, N., Hoose, C., Jantsch, E., Koop, T. and Murakami, M.: Ice nucleation by cellulose and its potential contribution to ice formation in clouds, Nat. Geosci., 8(4), 273–277, doi:10.1038/ngeo2374, 2015b.

Lesins, G., Chylek, P., & Lohmann, U.: A study of internal and external mixing scenarios and its effect on aerosol optical properties and direct radiative forcing, J. Geophys. Res.-Atmos., 107, 1-12, doi:10.1029/2001jd000973, 2002.

Lohmann, U., and Feichter, J.: Global indirect aerosol effects: a review, Atmos. Chem. Phys., 5, 715-737, doi:10.5194/acp-5-715-2005, 2005.

Lubin, D., and Vogelmann, A.: A climatologically significant aerosol longwave indirect effect in the Arctic. Nature, 439, 453-456, doi:10.1038/nature04449, 2006.

Mjolsness, E.: Machine Learning for Science: State of the Art and Future Prospects, Science, 293, 2051-2055, doi:10.1126/science.293.5537.2051, 2001.

Murphy, D. M.: The design of single particle laser mass spectrometers, Mass Spectrom. Rev., 26 (2), 150–165, 2007.

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., 37:4, 382-391, doi:10.1080/02786820300971, 2003.

Niemand, M., Möhler, O., Vogel, B., Vogel, H., Hoose, C., Connolly, P., Klein, H., Bingemer, H., DeMott, P., Skrotzki, J. and Leisner, T.: A Particle-Surface-Area-Based Parameterization of Immersion Freezing on Desert Dust Particles, J. Atmos. Sci., 69, 3077-3092, 2012.

Peckhaus, A., Kiselev, A., Hiron, T., Ebert, M. and Leisner, T.: A comparative study of K-rich and Na/Ca-rich feldspar ice-nucleating particles in a nanoliter droplet freezing assay, Atmos. Chem. Phys., 16(18), 11477–11496, doi:10.5194/acp-16-11477-2016, 2016.

Powers D. W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies, 7, 1-24, 2007.

Saathoff, H., Naumann, K.-H., Schnaiter, M., Schöck, W., Möhler, O., Schurath, U., Weingartner, E., Gysel, M. and Baltensperger, U.: Coating of soot and (NH4)2SO4 particles by ozonolysis products of α-pinene, J. Aerosol Sci., 34(10), 1297–1321, doi:10.1016/S0021-8502(03)00364-1, 2003.

Steinke, I., Funk, R., Busse, J., Iturri, A., Kirchen, S., Leue, M., Möhler, O., Schwartz,T., Schnaiter, M., Sierau, B., Toprak, E., Ullrich, R., Ulrich, A., Hoose, C. and Leisner, T.: Ice nucleation activity of agricultural soil dust aerosols from Mongolia, Argentina, and Germany, J. Geophys. Res. Atmos., doi:10.1002/2016JD025160, 2016.

Vogelmann, A., McFarquhar, G., Ogren, J., Turner, D., Comstock, J., Feingold, G., Long, C., Jonsson, H., Bucholtz, A., Collins, D., Diskin, G., Gerber, H., Lawson, R., Woods, R., Andrews, E., Yang, H., Chiu, J., Hartsock, D., Hubbe, J., Lo, C.,Marshak, A., Monroe, J., McFarlane, S., Schmid, B., Tomlinson, J. and Toto, T.: Racoro Extended-Term Aircraft Observations of Boundary Layer Clouds, Bull. Amer. Meteor. Soc., 93, 861-878, 2012.

Welti, A., Lüönd, F., Stetzer, O., and Lohmann, U.: Influence of particle size on the ice nucleating ability of mineral dusts, Atmos. Chem. Phys., 9, 6929-6955, doi:10.5194/acpd-9-6929-2009, 2009.

Zawadowicz, M. A., Froyd, K. D., Murphy, D. M. and Cziczo, D. J.: Improved identification of primary biological aerosol particles using single particle

mass spectrometry, Atmos. Chem. Phys., doi: 10.5194/acp-2016-1119, 2016.

5. Table and Figure Captions

Table 1: Characterization of aerosol training set.

Table 2: Ensemble chemical features rankings for all particle labels and between broad aerosol categories in positive and negative ion modes.

Table 3. Intraclass model performance metrics for each aerosol category and ion mode. Rows characterize classification on a population consisting entirely of aerosols within that category. Left: Average classification accuracy within each category, where 1.0 = 100% precision (Powers, 2007) Right: mean and standard deviations of misclassification within each category.

Figure 1: Aerosol training data plotted as factor area 16 (O-) verses area 24 (C2-). Axes represent normalized mass-to-charge peak areas obtained from PALMS. Note closing of aerosol types. Co-plotted are ~500 randomly drawn spectra from the AIDA blind experiment, which was known to contain a subset of the training data types.

Figure 2. Schematic of decision tree classification for a single aerosol spectrum. From left to right, mass spectra are normalized with respect to total ion current, forming the elements of a normalized vector X. A trained decision tree then applies a series of tests to

a discreet number of peaks in order to arrive at a categorical aerosol prediction.

Figure 3. Column-normalized confusion matrices (P) showing fraction of aerosols labeled as j that belong to i, where i, j are row and column indices, respectively. Confusion matrices are computed with out-of-bag test examples, and for the purpose of this study are used to compute conditional probability distributions given model generated labels j. Aerosol labels are grouped into the follow broad categories: fertile soil, mineral/metallic, biological, and other. For all labels, A+ = 88% and A- = 86%.

Figure 4. Column-normalized confusion matrices for broad categorization of aerosols (Pl). Classification accuracy, or the average probability of a correct aerosol prediction across all labels, is computed by averaging diagonal elements of Pl. For all categories,

A+ = 94% and A- = 92%.

Figure 5. Maximum likelihood model predictions of ~5000 aerosols sampled from the AIDA FIN01 blind mixture. Bottom: broad aerosol categories. Top: breakout of non-SOA labels above the 1% level.

Figure 6. Aerosol training data plotted as factor area 16 (O-) verses area 24 (C2-) with aerosols grouped by category. Also shown are several contours from a probability density function fit on the training data using kernel density estimation.

Figure 7. Pie chart based on classification of negative FIN 03 data with a probability density > 0.3. The model identified mostly mineral/metallic particles and fertile soils, with smaller amounts of the other category and trace amounts of biologicals (< 1%). Unidentified FIN03 particles are likely organics, sulfates, and nitrates, which are among

the common, atmospherically-relevant cloud condensation nuclei not included in the model.

Aerosol type	Description and/or supplier	Generation method	Location of sampling	Reference
Argentina	Soil dust collected in La Pampa province, Argentina	Dry-dispersed	KIT	(Steinke et al., 2016)
China	Soil collected from Xilingele steppe, China/Inner Mongolia	Dry-dispersed	KIT	(Steinke et al., 2016)
Ethiopian	Soil collected in Lake Shala National Park, Ethiopia (collection coordinates: 7.5 N, 38.7 E)	Dry-dispersed	KIT	N/A
German	Arable soil collected near Karlsruhe, Germany	Dry-dispersed	KIT	(Steinke et al., 2016)
Moroccan	Soil collected in a rock desert in Morocco (collection coordinates: 33.2 N, 2.0 W)	Dry-dispersed	KIT	N/A
Paulinenaue	Arable soil collected in Northern Germany (Brandenburg)	Dry-dispersed	KIT	(Steinke et al., 2016)
ATD	Arizona Test Dust, Powder Technology, Inc. (Arden Hills, MN)	Dry-dispersed	MIT	N/A
Illite	Illite NX (Arginotec, Germany)	Dry-dispersed	KIT	(Hiranuma et al., 2015a)
Fly ash	Four samples of fly ash from U.S. power plants: J. Robert Welsh Power Plant (Mount Pleasant, TX), Joppa Power Station (Joppa, IL), Clifty Creek Power Plant (Madison, IN) and Miami Fort Generating Station (Miami Fort, OH) (Fly Ash Direct, Cincinnati, OH)	Dry-dispersed	MIT	(Garimella, 2016; Zawadowicz et al., 2016)
Na-Feldspar	Sodium and calcium-rich feldspar, samples provided by Institute of Applied Geosciences, Technical University of Darmstadt (Germany) and University of Leeds (UK)	Dry-dispersed	KIT	(Peckhaus et al., 2016)
K-Feldspar	Potassium-rich feldspar, samples provided by Institute of Applied Geosciences, Technical University of Darmstadt (Germany) and University of Leeds (UK)	Dry-dispersed	KIT	(Peckhaus et al., 2016)
Agar	Agar growth medium for bacteria, Pseudomonas Agar Base (CM0559, Oxoid Microbiology Products, Hampshire, UK)	Wet-generated	KIT	N/A
Bacteria	Two different cultures of *Pseudomonas syringae*.	Cultures grown on the agar growth medium (as above), suspended in nanopure water and wet-generated	KIT	(Zawadowicz et al., 2016)

Cellulose	Microcrystalline and fibrous cellulose (Sigma Aldrich, St. Louis, MO)	Wet-generated	KIT	(Hiranuma et al., 2015b)
Hazelnut	Natural hazelnut pollen (GREER, Lenoir, NC) wash water	Wet-generated	KIT	(Zawadowicz et al., 2016)
Snomax	Snomax, (Snomax International, Denver, CO) irradiated, desiccated and ground *Pseudomonas syringae*	Wet-generated	KIT	(Zawadowicz et al., 2016)
PSL	Polystyrene latex spheres (Polysciences, Inc. Warrington, PA), various sizes	Wet-generated	MIT	N/A
Soot	CAST soot	miniCAST flame soot generator	KIT	(Henning et al., 2012)
SOA	Secondary organic aerosol	Ozonolysis of α-pinene	KIT	(Saathoff et al., 2003)
K-Feldspar cSA	Potassium-rich feldspar (as above) coated with sulfuric acid (SA).	Small amounts of sulfuric acid were incrementally added to the chamber filled with K-feldspar to achieve thin coatings, as judged from PALMS spectra	KIT	(Saathoff et al., 2003)
K-Feldspar cSOA	Potassium-rich feldspar (as above) coated with secondary organic aerosol (SOA, as above).	Small amounts of α-pinene were incrementally added to the chamber filled with K-feldspar to achieve thin coatings, as judged from PALMS spectra	KIT	(Saathoff et al., 2003)

Table 1.

Between labels				Between categories			
Negative		Positive		Negative		Positive	
ion	label	ion	label	ion	label	ion	label
35	$^{35}Cl^-$	23	Na^+	35	$^{35}Cl^-$	23	Na^+
25	C_2H^-	59	$Co^{+(1)}/CaF^+/C_2H_2OOH^+$	26	$CN^-/C_2H_2^-$	59	$Co^{+(1)}/CaF^+/C_2H_2OOH^+$
24	C_2^-	39	$^{39}K^+$	46	NO_2^-	44	$SiO^+/COO^+/^{44}Ca^+/AlOH^+$
57	C_2OOH^-	12	C^+	1	H^-	39	$^{39}K^+$
59	$C_2H_2OOH^-/AlO_2^-$	24	C_2^+	57	C_2OOH^-	28	Si^+/CO^+
43	HCN^-/AlO^-	41	$^{41}K^+/C_3H_5^+$	59	$C_2H_2OOH^-/AlO_2^-$	41	$^{41}K^+/C_3H_5^+$
1	H^-	204-208	Pb region (^{204}Pb, ^{206}Pb, ^{207}Pb and ^{208}Pb)	45	$COOH^-$	54	$^{54}Fe^+$
26	$CN^-/C_2H_2^-$	27	$Al^+/C_2H_3^+$	42	$CNO^-/C_2H_2O^-$	56	Fe^+/CaO^+
46	NO_2^-	44	$SiO^+/COO^+/^{44}Ca^+/AlOH^+$	43	HCN^-/AlO^-	27	$Al^+/C_2H_3^+$
16	O^-	57	$^{57}Fe^+/CaOH^+/C_3H_4OH^+$	16	O^-	45	$SiOH^+/COOH^+$
17	OH^-	N/A	aerodynamic diameter	73	$C_2O_3H^-/C_3H_2OOH_3^-$	66	Zn^+
61	$SiO_2H^-/^{29}SiO_2^-/C_5H^-/CHO_3^-$	83	$H_3SO_3^+/C_4H_2OOH^+$	63	PO_2^-	57	$^{57}Fe^+/CaOH^+/C_3H_4OH^+$
63	PO_2^-	87	$^{87}Rb^+/CaPO^+$	60	$SiO_2^-/C_5^-/CO_3^-/AlO_2H^-$	87	$^{87}Rb^+/CaPO^+$
19	F^-/H_3O^-	13	CH^+	15	NH^-/CH_3^-	85	$^{85}Rb^+$
76	SiO_3^-	66	Zn^+	24	C_2^-	83	$H_3SO_3^+/C_4H_2OOH^+$
77	$SiO_3H^-/^{29}SiO_3^-$	28	Si^+/CO^+	76	SiO_3^-	24	C_2^+
79	PO_3^-	85	$^{85}Rb^+$	32	O_2^-	204-208	Pb region (^{204}Pb, ^{206}Pb, ^{207}Pb and ^{208}Pb)
60	$SiO_2^-/C_5^-/CO_3^-/AlO_2H^-$	72	FeO^+/CaO_2^+	N/A	aerodynamic diameter	40	Ca^+
45	$COOH^-$	54	$^{54}Fe^+$	71	$C_3H_2OOH^-$	153	$^{137}BaO^+$
N/A	aerodynamic diameter	82	ZnO^+	50	$C_4H_2^-$	N/A	aerodynamic diameter

(1) Contamination

Table 2.

Category	Negative	Postive
Fertile Soil	0.89	0.84
Mineral/Metallic	0.93	0.97
Biological	1.00	1.00
Other	0.93	0.96

Category	Negative	Postive
Fertile Soil	0.022 ±0.021	0.032 ±0.031
Mineral/Metallic	0.017 ± 0.031	0.006 ±0.013
Biological	0.000	0.001 ± 0.003
Other	0.025 ±0.075	0.010 ±0.029

Table 3.

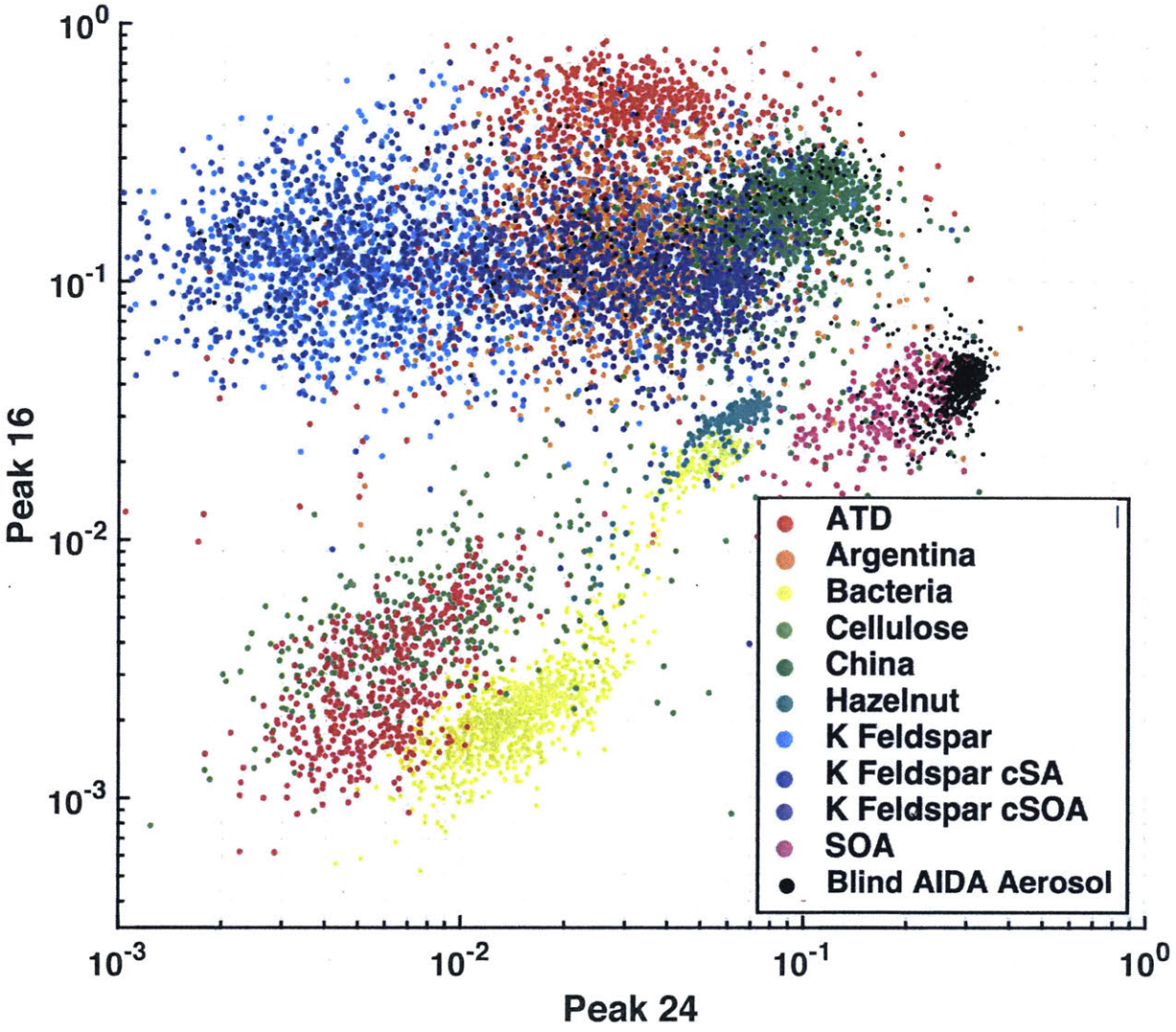

Peak 16
Peak 24
10^0
10^{-1}
10^{-2}
10^{-3}
ATD
Argentina
Bacteria
Cellulose
China
Hazelnut
K Feldspar
K Feldspar cSA
K Feldspar cSOA
SOA
Blind AIDA Aerosol

Figure 1.

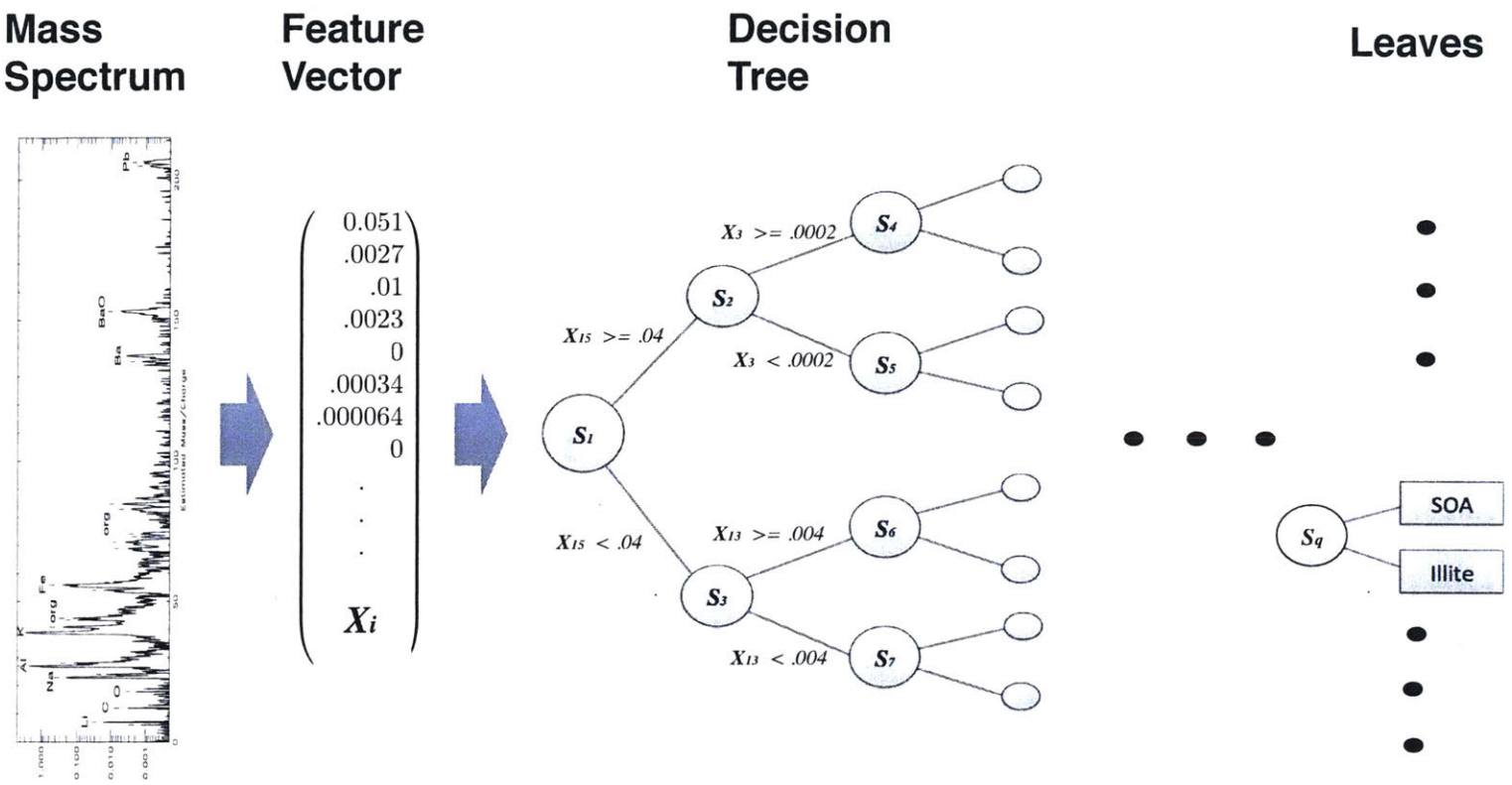
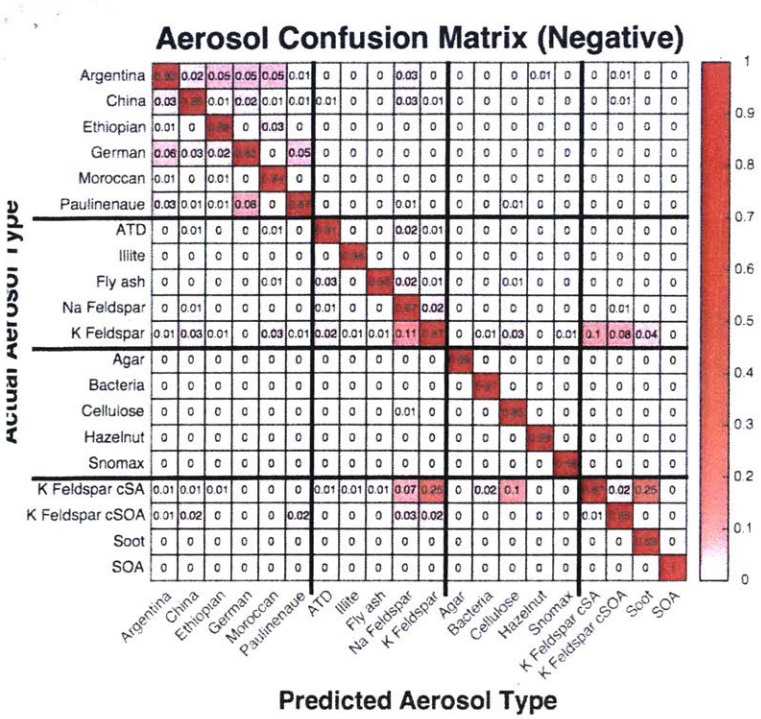

Mass Spectrum
Feature Vector
Decision Tree
Leaves
0.051
.0027
.01
.0023
0
.00034
.000064
0
Xi
X15 >= .04
X15 < .04
X3 >= .0002
X3 < .0002
X13 >= .004
X13 < .004
S1
S2
S3
S4
S5
S6
S7
Sq
SOA
Illite

Figure 2.


Aerosol Confusion Matrix (Negative)
Actual Aerosol Type
Predicted Aerosol Type
Argentina
China
Ethiopian
German
Moroccan
Paulinenaue
ATD
Illite
Fly ash
Na Feldspar
K Feldspar
Agar
Bacteria
Cellulose
Hazelnut
Snomax
K Feldspar cSA
K Feldspar cSOA
Soot
SOA

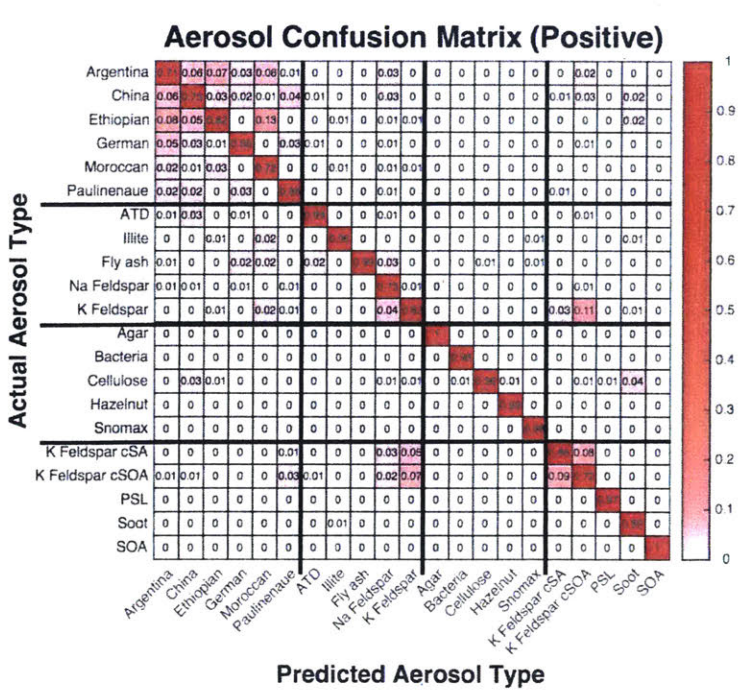

Aerosol Confusion Matrix (Positive)
Actual Aerosol Type
Predicted Aerosol Type
Argentina
China
Ethiopian
German
Moroccan
Paulinenaue
ATD
Illite
Fly ash
Na Feldspar
K Feldspar
Agar
Bacteria
Cellulose
Hazelnut
Snomax
K Feldspar cSA
K Feldspar cSOA
PSL
Soot
SOA

Figure 3.

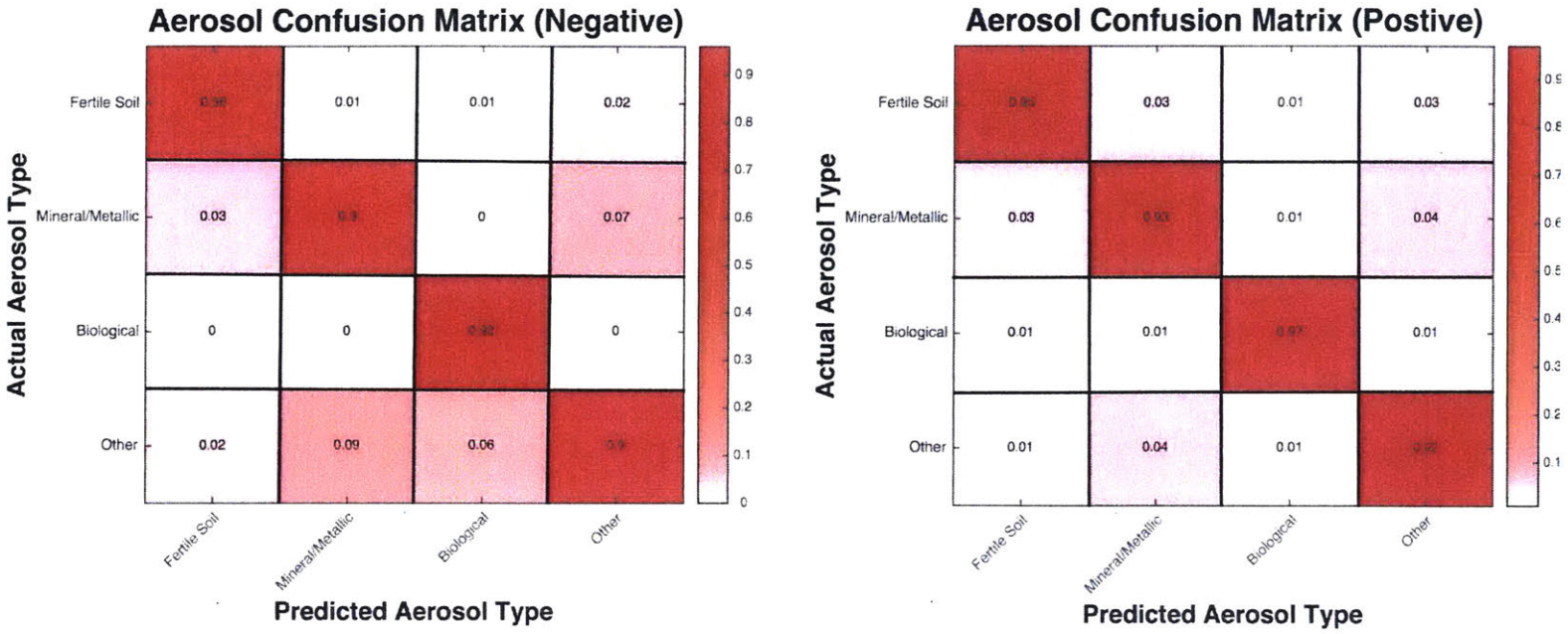

Aerosol Confusion Matrix (Negative)
Actual Aerosol Type
Fertile Soil
Mineral/Metallic
Biological
Other
0.01
0.01
0.02
0.03
0
0.07
0
0
0
0.02
0.09
0.06
Fertile Soil
Mineral/Metallic
Biological
Other
Predicted Aerosol Type
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Aerosol Confusion Matrix (Postive)
Actual Aerosol Type
Fertile Soil
Mineral/Metallic
Biological
Other
0.03
0.01
0.03
0.03
0.01
0.04
0.01
0.01
0.01
0.04
0.01
Fertile Soil
Mineral/Metallic
Biological
Other
Predicted Aerosol Type

Figure 4

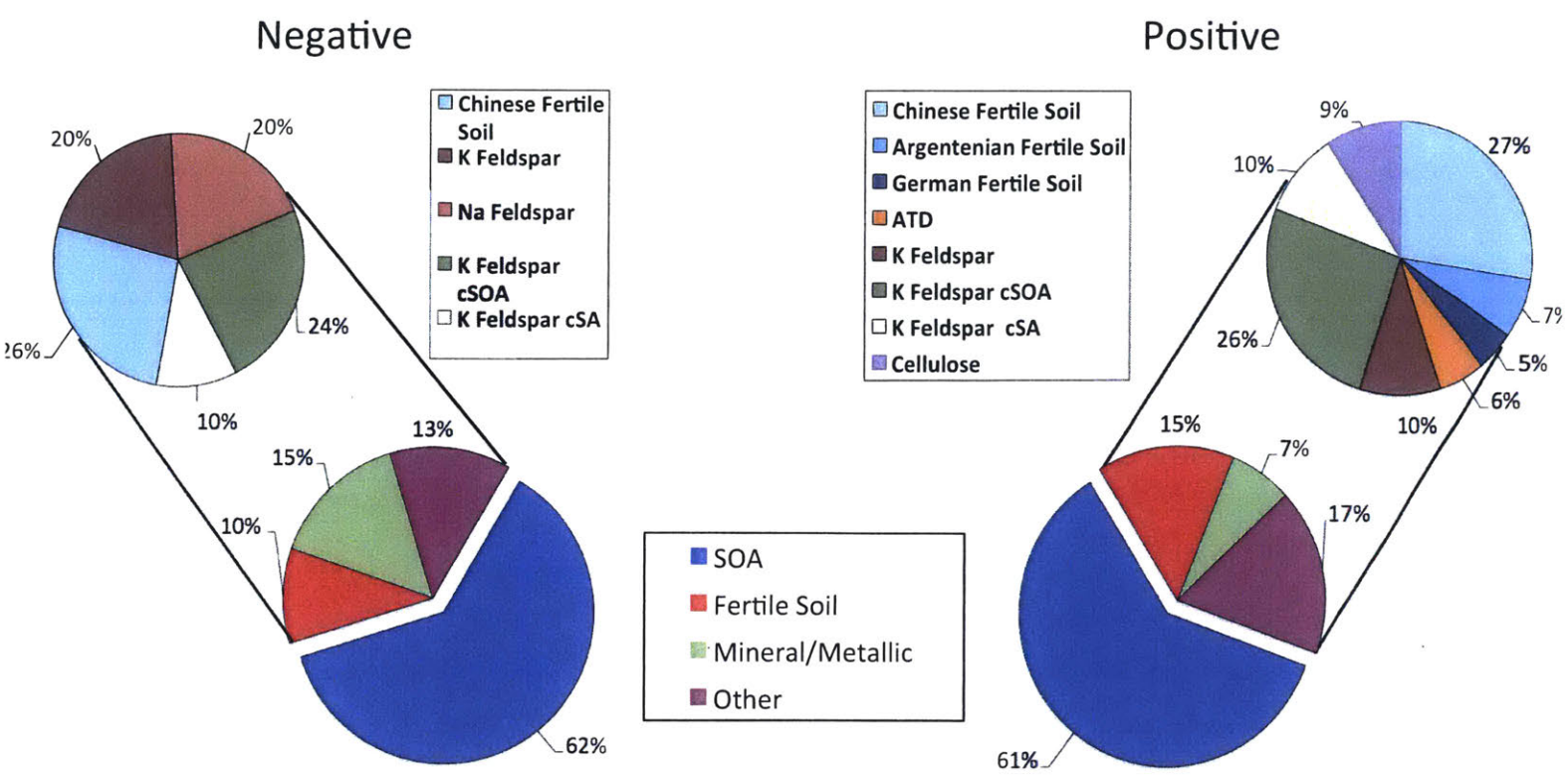

FIN01 Aerosol Characterization
Negative
Positive
Chinese Fertile Soil
K Feldspar
Na Feldspar
K Feldspar cSOA
K Feldspar cSA
20%
20%
24%
10%
13%
15%
10%
62%
Chinese Fertile Soil
Argentenian Fertile Soil
German Fertile Soil
ATD
K Feldspar
K Feldspar cSOA
K Feldspar cSA
Cellulose
9%
27%
10%
26%
5%
6%
10%
15%
7%
17%
61%
SOA
Fertile Soil
Mineral/Metallic
Other

Figure 5.

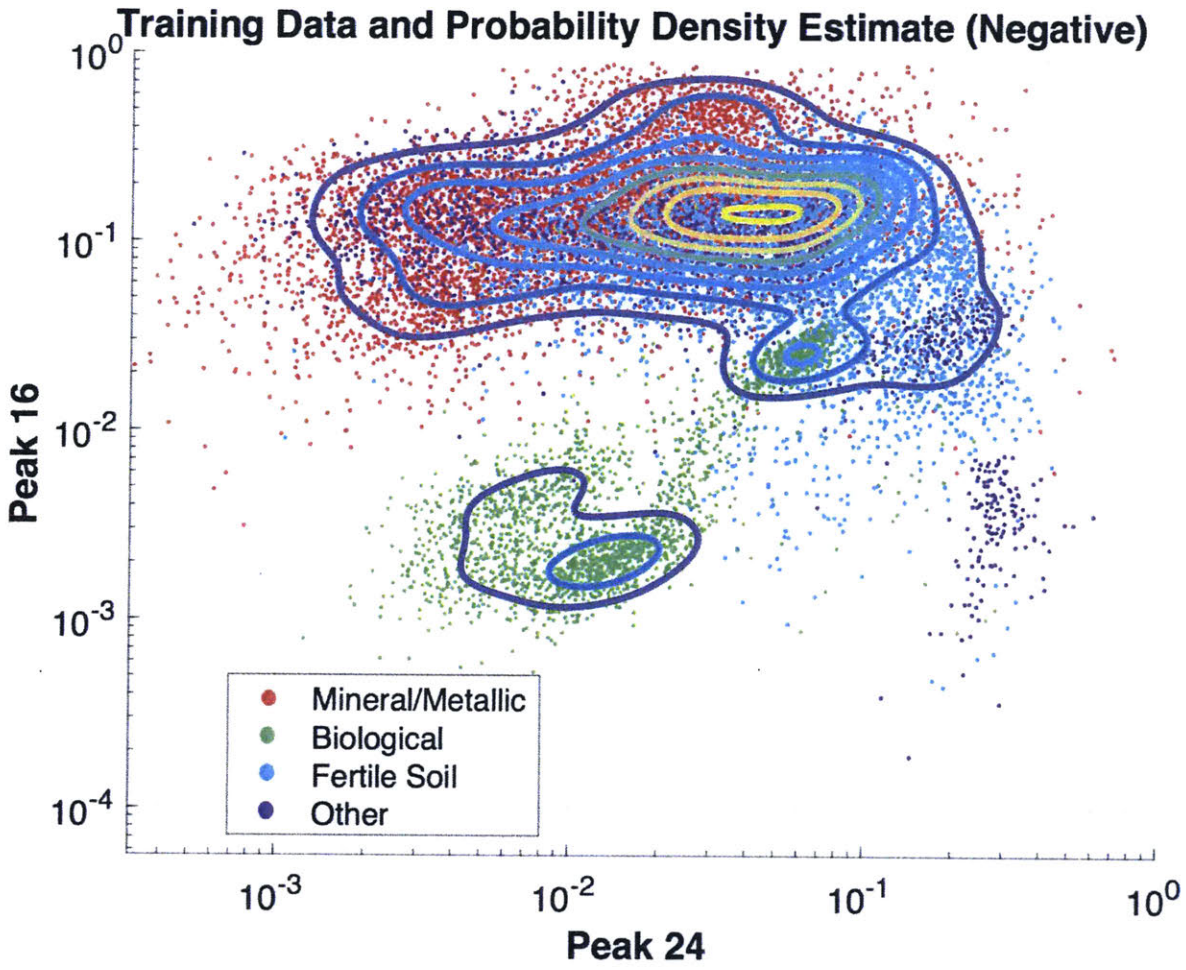

Training Data and Probability Density Estimate (Negative)
Peak 16
Peak 24
Mineral/Metallic
Biological
Fertile Soil
Other

Figure 6

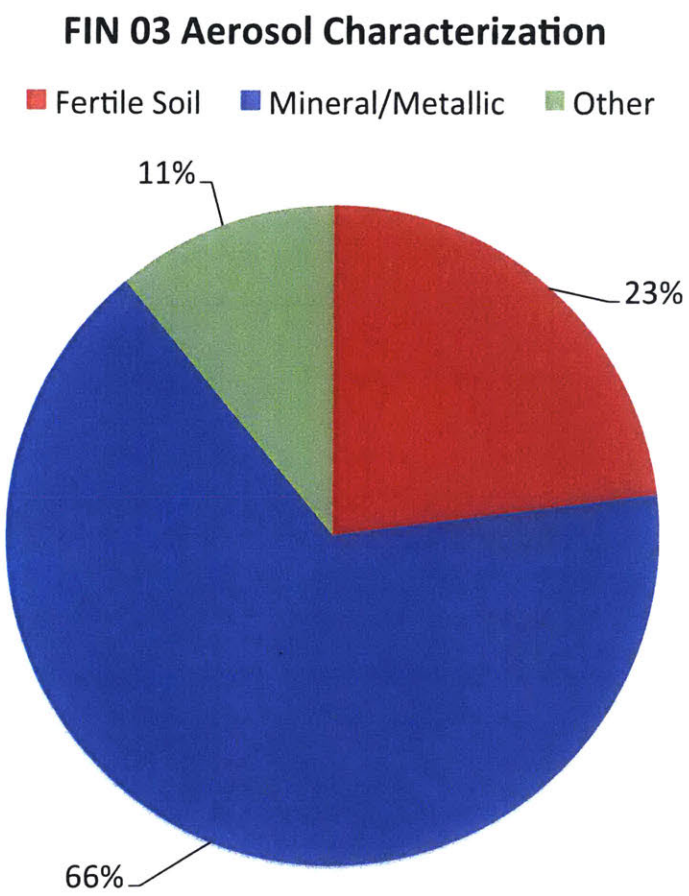

FIN 03 Aerosol Characterization
Fertile Soil
Mineral/Metallic
Other
11%
23%
66%

Figure 7