Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

Machine learning for improved data analysis of biological aerosol using the WIBS

Simon Ruske[1], David O. Topping[1], Virginia E. Foot[2], Andrew P. Morse[3], and Martin W. Gallagher[1]

[1]Centre of Atmospheric Science, SEES, University of Manchester, Manchester, UK
[2]Defence, Science and Technology Laboratory, Porton Down, Salisbury, Wiltshire, SP4 0JQ, UK
[3]Department of Geography and Planning, University of Liverpool, Liverpool, UK

Correspondence to: simon.ruske@postgrad.manchester.ac.uk

Abstract.

Primary biological aerosol including bacteria, fungal spores and pollen have important implications for public health and the environment. Such particles may have different concentrations of chemical fluorophores and will provide different responses in the presence of ultraviolet light which potentially could be used to discriminate between different types of biological aerosol. Development of ultraviolet light induced fluorescence (UV-LIF) instruments such as the Wideband Integrated Bioaerosol Sensor (WIBS) has made is possible to collect size, morphology and fluorescence measurements in real-time. However, it is unclear without studying responses from the instrument in the laboratory, the extent to which we can discriminate between different types of particles. Collection of laboratory data is vital to validate any approach used to analyse the data and to ensure that the data available is utilised as effectively as possible.

In this manuscript we test a variety of methodologies on traditional reference particles and a range of laboratory generated aerosols. Hierarchical Agglomerative Clustering (HAC) has been previously applied to UV-LIF data in a number of studies and is tested alongside other algorithms that could be used to solve the classification problem: Density Based Spectral Clustering and Noise (DBSCAN), k-means and gradient boosting.

Whilst HAC was able to effectively discriminate between the reference particles, yielding a classification error of only 1.8%, similar results were not obtained when testing on laboratory generated aerosol where the classification error was found to be between 11.5% and 24.2%. Furthermore, there is a worryingly large uncertainty in this approach in terms of the data preparation and the cluster index used, and we were unable attain consistent results across the different sets of laboratory generated aerosol tested.

The best results were obtained using gradient boosting, where the misclassification rate was between 4.38% and 5.42%. The largest contribution to this error was the pollen samples where 28.5% of the samples were misclassified as fungal spores. The technique was also robust to changes in data preparation provided a fluorescent threshold was applied to the data.

Where laboratory training data is unavailable, DBSCAN was found to be an potential alternative to HAC. In the case of one of the data sets where 22.9% of the data was left unclassified we were able to produce three distinct clusters obtaining a classification error of only 1.42% on the classified data. These results could not be replicated however for the other data set where 26.8% of the data was not classified and a classification error of 13.8% was obtained. This method, like HAC, also

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

appeared to be heavily dependent on data preparation, requiring different selection of parameters dependent on the preparation used. Further analysis will also be required to confirm our selection of parameters when using this method on ambient data.

There is a clear need for the collection of additional laboratory generated aerosol to improve interpretation of current databases and to aid in the analysis of data collected from an ambient environment. New instruments with a greater resolution are likely improve on current discrimination between pollen, bacteria and fungal spores and even between their different types, however the need for extensive laboratory training data sets will grow as a result.

1 Introduction

Biological aerosol, such as bacteria, fungal spores and pollen have important implications for public health and the environment (Després et al., 2012). They have been linked to the formation of cloud condensation nuclei and ice nuclei which in turn may have important influence on the weather (Crawford et al., 2012; Cziczo et al., 2013; Gurian-Sherman and Lindow, 1993; Hader et al., 2014; Hoose and Möhler, 2012; Möhler et al., 2007). The particles have impacts on health (Kennedy and Smith, 2012), particularly for those who suffer from asthma and allergic rhinitis (D'Amato et al., 2001).

It is therefore of paramount importance that we continue to develop methods of detecting these particles, to quantify them, determine seasonal trends and to compare different environments. One such method for detecting biological aerosol is to use an ultraviolet light induced fluorescence (UV-LIF) spectrometer such as the Wideband Integrated Bioaerosol Spectrometer (WIBS). Particles with different concentrations of the chemical fluorophores tryptophan and NADH will provide different responses when excited. In addition to the fluorescence measurements collected, a measurement of size and shape for each particle is taken to further aid in discrimination.

These measurements have limited application in isolation. However, data analysis techniques, such as those available within the field of machine learning, are potentially able transform these measurements into quantities of pollen, bacteria and fungal spores. There are a variety of machine learning algorithms that are applicable to solving this classification problem, and they can be divided broadly into two groups: supervised and unsupervised.

It is not clear whether the supervised or unsupervised approach is to be preferred as both approaches have their advantages and disadvantages. Supervised machine learning uses data, usually collected within laboratory settings, where the correct classification is known. Subsequently, this data is split into training data and testing data. The training data is used to fit a model which is then validated using the test set. Once the model is fitted and validated it may then be applied to ambient data.

During unsupervised analysis, ambient data is classified without using training data from the laboratory. Instead, an attempt is made to split the data into groups using natural differences in the data. Ideally the data would be naturally split into broad biological classes, but this may not necessarily be the case. Instead, for example, two sets of similar bacteria and fungal spores could be grouped together.

The supervised methods, including gradient boosting, have the disadvantage that the training data collected may not include the entirety of what may be collected during an ambient campaign. Particularly in an urban environment, the instrument will collect a large quantity of non-biological material that will still need to be either classified as such or removed from the analysis.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

We would expect most of this non-biological material to either be non-fluorescent or weakly fluorescent and therefore it should be removed prior to analysis by applying a justifiable threshold to the fluorescent measurements (see Section 2.2). Nonetheless, a few weakly fluorescent non-biological particles may remain and could be overlooked if the training data is incomplete.

Clearly there are issues to be explored with either approach, therefore it seems unlikely that we will be able to abandon either supervised or unsupervised techniques at this point in time.

In an ambient setting, determining the number of clusters is difficult so Hierarchical Agglomerative Clustering (HAC) has been the preferred method over other methods such as k-means since the method naturally presents a clustering for all possible number of clusters (Robinson et al., 2013). A suggestion of the number of clusters can then be provided using indices such as the Caliński Harabasz Index (CH Index) (Caliński and Harabasz, 1974) by maximising a statistic which yields a peak for clusterings which contain clusters that are compact and far apart. HAC has previously been used on data collected using the WIBS to discriminate between different Polystyrene Latex Spheres (PSLs) and has been applied to ambient measurements collected as part of the BEACHON RoMBAS experiment (Crawford et al., 2015; Gallagher et al., 2012; Robinson et al., 2013).

Nonetheless, little has been done to demonstrate the effectiveness of HAC on laboratory generated aerosol. Evaluating the effectiveness of HAC on generated aerosol is crucial to support or repudiate conclusions made using HAC on ambient data, especially since the fluorescence response from the laboratory generated aerosol will much better reflect fluorescence responses from the environment, when compared with PSLs.

During the process of HAC there are also a number of vital choices that have to be made, that could have a substantial implication on the effectiveness of the method (these are discussed in detail in Section 2.2). For the PSLs previously analysed (Crawford et al., 2015), we determined standardising using the z-score, with removal of non-fluorescent particles, taking logarithms of shape and size was most effective. The CH index was selected to determine the number of clusters as it was demonstrated to perform best in the literature (Milligan and Cooper, 1985). It is however, not clear whether these choices will remain the most effective for laboratory generated aerosol nor ambient data. See Section 2.3 for further details on data preparation for HAC.

Furthermore, data analysis using HAC can take a matter of hours, if not days depending on the number of particles. The time requirements for HAC are between N^2 and N^3 meaning that a doubling of the number of particles will require between four and eight times as much time. This means that not only is the method already quite slow, but will get increasingly slower as more data is collected. This may limit the real time effectiveness of the method.

Within the Python programming language, a package called Scikit-learn (Pedregosa et al., 2011) offers implementations of several unsupervised methods. Some of these methods i.e. Affinity Propagation, Mean-shift, Spectral Clustering and Gaussian mixtures are not explored as they will scale poorly as the number of particles increases (Pedregosa et al., 2011). Instead, our analysis is focused on K-means, HAC and DBSCAN which can be used on larger data-sets.

For HAC we continue to use the fastcluster package (described in Section 2.3). Sklearn does have a HAC implementation but it is not as fast or memory efficient. We do use sklearn for DBSCAN and kmeans, although if one was to use DBSCAN for ambient data we would suggest exploring alternatives such as ELKI (Schubert et al., 2015) as the sklearn implementation

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

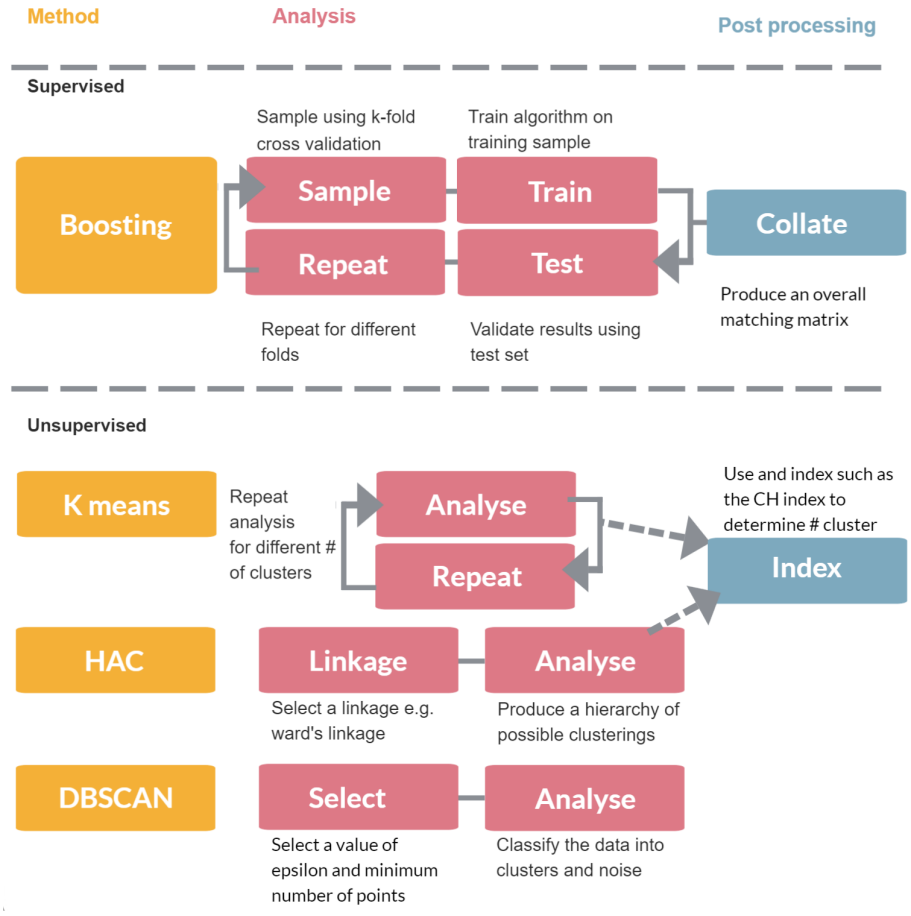

Figure 1. Overview of different analysis approaches

of DBSCAN by default is not memory efficient making it difficult to utilise for more than $30,000$ particles. Sklearn has a fast implementation for Gradient Boosting, so this is used.

2 Methods

In this section we discuss the variety of approaches that could be used to classify particles such as bacteria, fungal spores or pollen. In Section 2.1 we provide an overview of the instrument used to collect the data. In Section 2.2 we discuss the variety of decisions that need to be made prior to passing the data to the machine learning algorithms which are discussed in Sections 2.3 - 2.6. An overview of the different methods is given in Figure 1.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

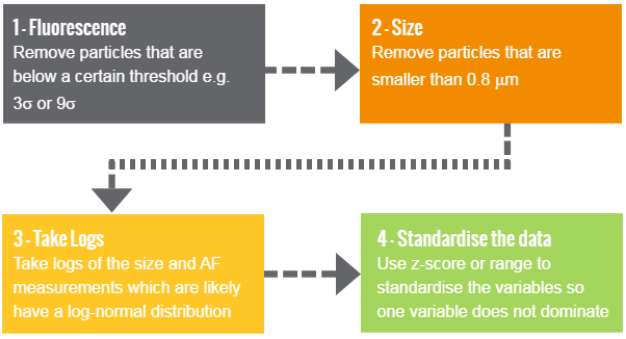

Figure 2. Overview of preprocessing steps for WIBS data

2.1 Instrumentation

The Wideband Integrated Bioaerosol Sensor (WIBS) collects size, shape and fluorescence measurements (Kaye et al., 2005). The size is a single measurement; the shape measurement consists of four measurements (one for each quadrant) which are combined to produce a single asymmetry factor measurement. Four fluorescence measurements are collected by firing a flash lamp at 280nm and 370nm and detecting the resultant fluorescence on two fluorescence detectors. The measurement collected using the second detector from the excitation at 280nm is ignored as it saturates the instrument. After removal of this fluorescent measurement, there are three remaining fluorescence measurements which are combined with the size and asymmetry factor measurements. A more detailed description of the instrument can be found in previous publications (e.g. Gabey et al., 2010; Healy et al., 2012a)

2.2 Data preparation

Prior to analysis using the machine learning algorithm we may choose to make a variety of decisions to pre-process the data with the aim to improve performance (see Figure 2). An overview for the decisions often made are outlined below.

First we may elect to remove particles which are non-fluorescent. Forced trigger data is collected which is a measurement of the instrument response when particles are not present. We then set a threshold, for which if a particle fails to exceed this threshold in at least one of the fluorescent channels we conclude that the particle is non-fluorescent. Usually we set the threshold to be three standard deviations above the average forced trigger measurement although a recent laboratory study has suggested that nine standard deviations may be more appropriate (Savage et al., 2017).

Another threshold is usually then applied to the size. A size threshold of 0.8μm is usually applied as detection efficiency of the instrument drops below 50% at this point. (Gabey, 2011; Gabey et al., 2011; Healy et al., 2012b).

Natural logarithms of the size and the asymmetry factor are often taken as these measurements are often log normally distributed and it is postulated that this will increase performance in the case of hierarchical agglomerative clustering.

It is also widely regarded that standardising the data prior to analysis is utmost importance (Milligan and Cooper, 1988). We often subtract the average measurement in each of the five variables and divide by the standard deviation, often referred to as

'standardising using the z-score'. Standardisation is used to prevent variables with larger magnitude, such as the fluorescent measurements, from dominating the analysis. An alternative approach to standardising is to divide each of the five variables by the range.

2.3 Hierarchical Agglomerative Clustering

In order for particles to be clustered, we need to define a measurement of how similar two clusters are. These similarity measures are often referred to as linkages. We use the Python package fastcluster (Müllner, 2013) which provides modern implementations of single, complete, average, weighted, Ward, centroid and median linkages (Müllner, 2011). A thorough detailing of the definitions of the different linkages can be found in the fastcluster manual (Müllner, 2013). For the memory efficient mode, which is essential when using the algorithm for large data sets, only Ward, centroid, median and single linkages are available.

Initially each particle is placed into an individual cluster. Next, using the linkage selected, the two most similar clusters are merged. The merging process is repeated until all the particles are placed in a single cluster, which provides a clustering from $k = 1, \cdots, N$, where k is the number of clusters and N is the number of particles being analysed. A cluster validation index such as the Calinski-Harabasz index (Caliński and Harabasz, 1974) is then used to identify an appropriate number of clusters. The index is maximised for clusterings that contain compact clusters that are far apart.

2.4 K-Means Clustering

K-Means clustering is designed to place particles into k clusters. However we can repeat the method multiple times e.g. for $k = 1, 2, \cdots, 10$, where k is the number of clusters. Similar to HAC we can then use a cluster validation index to determine which choice of k gives the most effective results.

The method works as follows. Initially k cluster centroids are set by selecting k particles at random. The rest of the particles are then placed into these k clusters depending on which of the centroids the particle is closest to. At this point a new centroid is calculated for each cluster. The process is then repeated many times until convergence occurs and the centroids do not change significantly from one iteration to the next.

2.5 DBSCAN

For DBSCAN we set two parameters, the radius for a neighbourhood ϵ, and the number of particles required for a neighbourhood to be identified as dense.

Initially a random point, say A, is selected. If there are sufficient number of points in the neighbourhood of A then all the points in A's neighbourhood are also checked and so on, until the cluster has fully expanded and there are no points left to check. Should the point not have a sufficient number of other points in its neighbourhood then it is left unclassified. Further points are then selected and the above process is repeated until all points have been considered.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

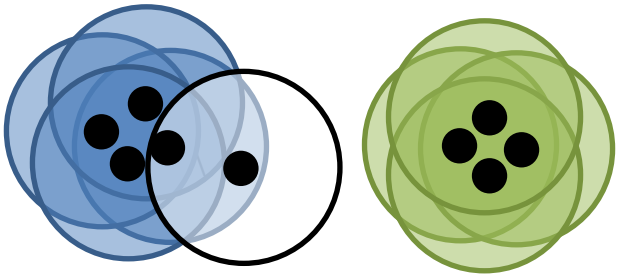

Figure 3. Visual representation of DBSCAN. Here each point is represented as a black dot and it's neighbourhood is represented by a circle. Here ϵ is the radius of the circle and the minimum number of points is 3. Four points have each been placed into the blue cluster and green cluster, all of which having at least 3 other points in their neighbourhood. One point is classified as noise as it has only 1 other point in it's neighbourhood.

We give an example of DBSCAN in Figure 3. Note that cluster validation indices are *not* required for DBSCAN, since the number of clusters is intrinsically calculated within the algorithm.

2.6 Gradient Boosting

A basic decision tree is constructed by considering each possible split across all variables and evaluating which split best divides the data. For example, we may consider the third fluorescence channel and split the data on the basis of whether the measurement is more or less than 10 arbitrary units (AU). This process is then repeated many times until a tree is built.

There are two ways in which trees can be combined into an ensemble. The first is by averaging multiple trees in the hope to produce a more accurate classification. This is known as a Random Forest. Here the data-set is sampled with replacement, meaning that the same particle could be selected more than once or not at all. Sampling in this way enables the algorithm to produce a subtly different version of the data from which to build each tree. In addition, instead of considering all possible variables to use to split the data, only a random subset is used.

Alternatively we can fit a single decision tree to the data, evaluate where the tree is performing well and then fit a second tree to the particles in the data for which the current model is performing poorly. This process can be repeated many times, each time adding a new tree to the model in the hope of making an improvement. This approach is known as AdaBoost. Gradient Boosting is an extension of AdaBoost to allow for other loss functions.

For the current study we elect to use Gradient Boosting to indicate the performance of the supervised approach since it was the best performer for the Multiparameter Bioaerosol Spectrometer, a similar UV-LIF spectrometer similar to the WIBS but with single waveband fluorescence, 8 fluorescence detection channels and very high shape analysis capability (Ruske et al., 2017).

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

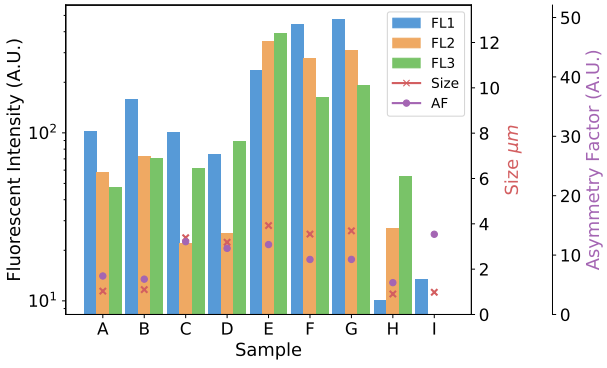

Figure 4. Average fluorescent characteristics for the different aerosol samples collected in 2008

3 Data

The efficacy of the different data analysis approaches was evaluated using three different data sets. The first of which comprised several industry standard polystyrene latex spheres of various different sizes and colours. This data set was first analysed in Crawford et al. (2015), where Hierarchical Agglomerative Clustering was successfully applied to the data yielding a classification accuracy of 98.2%. This data set presents a simple challenge for which we would expect any reasonable algorithm to be able to discriminate between the different sizes and colours of particles.

To further extend the previous analysis in Crawford et al. (2015) we include two previously unpublished data sets from 2008 and 2014 which are similar to data previously published using the Multiparameter Bioaerosol Spectrometer (Ruske et al., 2017). These data sets consist of various different pollen, fungal, bacterial and non-biological samples, and should present a much more difficult challenge for the algorithms.

The samples of laboratory generated aerosol were collected as follows. Material was aerosolised into a large, clean HEPA filtered chamber, which incorporated a recirculation fan. The *Bacillus atrophaeus* and *Escherichia coli* (*E.coli*) bacteria were aerosolised into the chamber using a mini-nebuliser (e.g. Hudson RCI Micro-Mist nebuliser) as were the salt and phosphate buffered saline samples. The dry samples, which included the pollen, and fungal samples were aerosolised directly into the chamber from small quantities of powder utilising a filtered compressed air jet. The diesel smoke and grass smoke samples were generated by burning a small amount within a fume cupboard using a smoker (piece of bespoke equipment).

We present a summary of the number of particles for each sample in total as well as when using a fluorescent threshold of 3σ and 9σ in Tables 1 and 2. Plots of the average fluorescent characteristics and size and shape for each sample are provided in Figures 4 and 5. Plots and tables for the polystyrene spheres previously published in Crawford et al. (2015) are omitted.

For most of the interferent particles we see that a fluorescent threshold of either 3σ or 9σ will remove the vast majority of these particles. The exception to this is in the case of the 2008 data we are unable to remove a significant number of the

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

Table 1. Counts of different aerosols collected in 2008 before and after a fluorescent threshold is applied.

ID	Sample	#	# (3σ)	# (9σ)	Classification
A	*Bacillus atrophaeus*	30946	12631	3239	bacteria
B	*E.coli*	15237	8332	3681	bacteria
C	Bermuda grass smut	5220	2681	423	fungal
D	Johnson grass smut	7248	3882	637	fungal
E	Paper-mulberry	1030	630	312	pollen
F	Ragweed pollen	569	332	151	pollen
G	Birch pollen	164	111	56	pollen
H	Grass smoke	14457	3357	299	interferent
I	Diesel smoke	7900	11	5	interferent

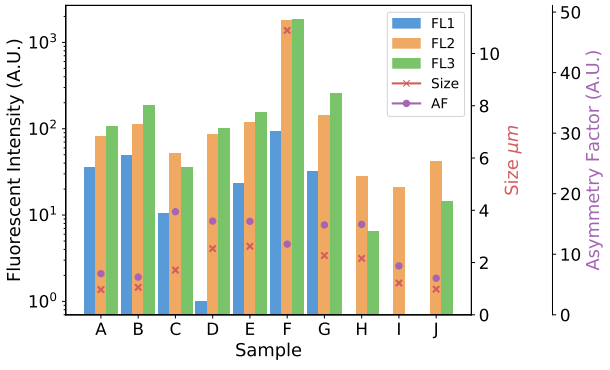

Figure 5. Average fluorescent characteristics for the different aerosol samples collected in 2014

grass smoke samples even using a fluorescent threshold of 9σ, providing an example of an interferent that does fluoresce in the instrument.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

Table 2. Counts of different aerosols collected in 2014 before and after a fluorescent threshold is applied.

ID	Sample	#	#(3σ)	#(9σ)	Classification
A	*Bacillus atrophaeus*	6217	3050	1292	bacteria
B	*E.coli*	2534	1290	632	bacteria
C	Puffballs	3919	555	252	fungal
D	Aspen pollen	398	74	31	pollen
E	Poplar pollen	375	104	50	pollen
F	Paper-Mulberry	565	541	537	pollen
G	Ryegrass	47	21	15	pollen
H	Fullers' Earth	3226	35	3	interferent
I	NaCl	2197	3	0	interferent
J	Phosphate Buffered Saline	3064	61	20	interferent

Table 3. Outline of the different approaches tested when using Hierarchical Agglomerative Clustering

Consideration	Option
Take Logs	True or False
Size Threshold	None or 0.8
Fluorescent Threshold	None, 3σ or 9σ
Standardisation	Z-score or Range
Linkage	Ward, Centroid, Median or Single

4 Results

4.1 Hierarchical Agglomerative Clustering

Usually when using Hierarchical Agglomerative Clustering we use the following data preparation strategy: take logs of the size and the asymmetry factor, apply a size threshold of 0.8 microns, apply a fluorescent threshold of 3 or more recently 9 standard

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

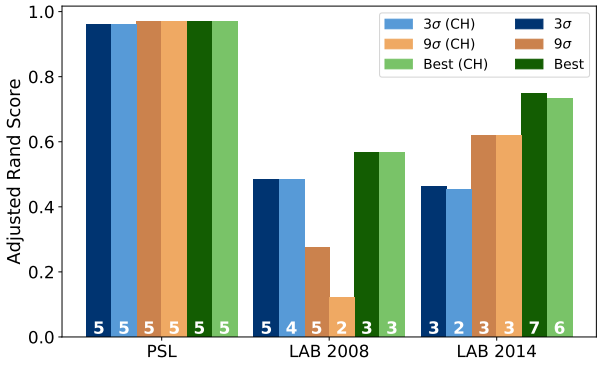

Figure 6. Performance of Hierarchical Agglomerative Clustering using the adjusted rand score for the data sets tested across different data preparation strategies. The number of clusters concluded in each case is indicated at the bottom of each bar.

deviations and standardise using the z-score. This approach is used as it has been demonstrated to be the most effective for the PSL data previously tested. We varied this approach by using a variety of different data preparation methods outlined in Table 3.

In Figure 6 we outline how well Hierarchical Agglomerative Clustering performed when using the standard strategy varying between 3σ and 9σ, and how well the algorithm worked with the best data preparation strategy across all 96 possible combinations of options for each data set.

First, we see that the high performance achieved for the PSLs (AR = 0.958), previously studied in Crawford et al. (2015), could not be fully extended to the laboratory generated aerosol studied where the highest adjusted rand score attained was 0.567 and 0.747 in 2008 and 2014 respectively. This is to be expected as the laboratory generated aerosol particles are much more complex, and therefore more difficult to differentiate.

We note the best performing data strategy for the PSL's previously studied (Crawford et al., 2015) was not the best performing for the laboratory generated aerosol. For the data set collected in 2008, the best strategy was found to be: taking logs; using a size threshold of 0.8 microns; using 3 standard deviations as a fluorescent threshold; standardising using the range; and using Ward linkage. In 2014, the best results were obtained by not taking logs, not applying a size threshold, using a fluorescent threshold of 9 standard deviations and using the centroid linkage.

In addition, there was a substantial difference between the quality of results attained for 3 standard deviations vs. 9 standard deviations. In 2008, we see a decrease in the adjusted rand score from 0.482 to 0.277 when using 3 and 9σ respectively. In 2014, we see an increase in the adjusted rand score from 0.462 to 0.625 when using 3 and 9σ respectively. So not only is there a substantial difference between the quality of results dependent on the data preparation technique used, but the difference is inconsistent across different data sets.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

Table 4. Matching matrix for the best case scenario when using the current data preparation strategy with 9σ on the data collected in 2014

	bacteria	fungal spores	pollen	non-biological
CL1	4	80	13	5
CL2	85	4	550	3
CL3	1835	168	70	15

Table 5. Matching matrix for the worst case scenario when using the current data preparation strategy with 9σ on the data collected in 2008

	bacteria	fungal spores	pollen	non-biological
CL1	547	69	298	0
CL2	6373	991	221	304

It is indeed the case that the data preparation approach currently used could be improved upon for the laboratory generated aerosol. However, due to inconsistencies in results across different data-sets it becomes difficult to provide an accurate recommendation as to what data preparation strategy should be used for hierarchical agglomerative clustering.

The adjusted rand score is often quite difficult to interpret, so we provide matching matrices for the best and worst case scenario using the current data preparation strategy in Tables 4 and 5. In the best case scenario we are able to discriminate between the pollen and the rest of the data placing 86.8% of the pollen into Cluster 2. Most of the bacteria is also placed into Cluster 3 with 66.6% of the fungal spores. A third of the fungal spores are differentiated from the rest of the data and placed into Cluster 1. In the worst case scenario two clusters are provided both primarily containing bacteria. In this case we can conclude that algorithm has failed to differentiate between any of the biological classes.

From Figure 6, it is clear that data preparation strategy can have a substantial impact upon the quality of clustering results. From Tables 4 and 5 we demonstrate that for a particular data preparation approach the quality of the clustering results could vary substantially across the different data sets. Therefore, it is important that in future analysis one should demonstrate that a particular data preparation performs consistently across a variety of different types of samples and performance is repeatable.

4.2 DBSCAN

One of the main difficulties of using DBSCAN is selecting the minimum number of points to form a neighbourhood and the radius of the neighbourhood (Khan et al., 2014). For 9σ and 3σ using z-score, taking logs of the size and asymmetry factor and removing particles smaller than 0.8 microns we repeat the DBSCAN algorithm for a variety of ϵ (neighbourhood radii) and minimum number of points values. The range of values of ϵ we test is $0.1, 0.2, \cdots, 1.0$. The range of minimum number of points is set using the following range relative to the number of particles collected $0.1\%, 0.2\%, \cdots, 1.0\%, 2.0\%, \cdots, 10.0\%$.

We found wide variety of performance across the different parameters. Often high accuracy could be obtained when using a high value of the minimum number of points but this resulted in removing a substantial portion of the data. In Figure 7 we

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

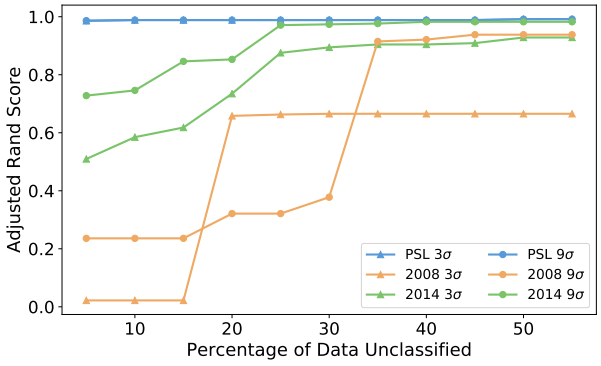

Figure 7. Adjusted rand score using different thresholds of percentage of points we allow to be left in the analysis for DBSCAN.

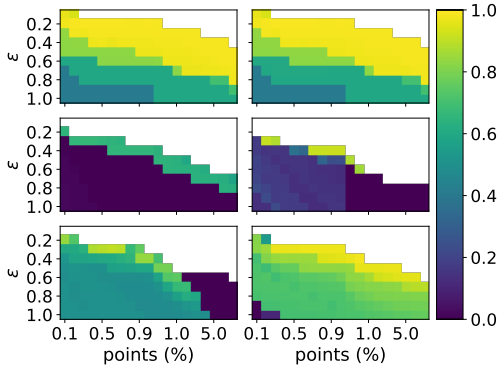

Figure 8. Adjusted rand score for DBSCAN, over a range different values of ϵ and minimum number of points required to form a neighborhood. The minimum number of points is expressed relative to the total number of points. The columns correspond to 3 and 9σ respectively. The rows correspond to the PSL, 2008 and 2014 data respectively.

filter our results using a range of thresholds for the maximum number of points that can be left unclassified $(5\%, 10\%, \cdots 60\%)$ and plot the corresponding best performance under this filter. In all the data sets there was a point of diminishing returns where no further benefit could be attained by removing any more of the data. In the case of the PSL data, this point happened after removing around 5% of the particles. For the laboratory data sets between 25 and 40% of the data was left unclassified before a peak in performance was attained. Nonetheless, we note in the case of the laboratory data collected in 2014 and using a 9σ fluorescent threshold, we can attain performance similar to that which we attain for the PSL data.

In order to investigate further a choice of ϵ and the minimum number of points which would maximise performance in terms of the adjusted rand score we plot the adjusted rand score for each test across all of the data sets. In Figure 8 we see that there is a large window of different values for which a higher value of the adjusted rand score can be achieved on the PSLs. Contrary to

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

Table 6. Matching matrix for the best case scenario when using DBSCAN with 9σ, $\epsilon = 0.4$ and a minimum number of points of 0.7% on 2014 data.

	bacteria	fungal spores	pollen	non-biological
Unclassified	329	169	134	16
CL1	0	0	490	0
CL2	12	80	4	0
CL3	1583	3	5	7

Table 7. Matching matrix for the worst case scenario when using DBSCAN with 3σ, $\epsilon = 0.3$ and a minimum number of points of 0.4% on 2008 data

	bacteria	fungal spores	pollen	non-biological
Unclassified	5858	1893	636	752
CL1	15025	15	44	2616
CL2	80	4655	393	0

this, in 2008 when using 9σ there is a very narrow window for which higher values of the adjusted rand score could be attained. It can also be seen that as ϵ increases the number of points required to create a cluster needs to be increased to compensate.

Overall our results indicate setting $\epsilon = 0.3$ and $\epsilon = 0.4$ when using 3σ and 9σ respectively. Best results can then be obtained by setting the number of points between 0.4% and 0.7% of the data when using an ϵ of 0.3 and 0.7% and 1.0% when using an ϵ of 0.4. However, future research will be required to demonstrate these conclusions are applicable when studying ambient data.

We provide matching matrices for the worst and best case scenarios in Tables 6 and 7. We see that in the best case scenario, leaving a decent proportion of data left unclassified we are able to produce three distinct clusters containing predominantly one broad class of biological aerosol. In the worst case scenario we manage only to distinguish between the bacteria from the fungal spores combined with the pollen.

In the worst case scenario i.e. using 3σ, on the 2008 data we fail to remove a sizable fraction of the non-biological particles, which was also the case when using HAC, however we would have expected that the algorithm would leave the particles unclassified. There is some argument that this worst case scenario could be circumvented by simply using the 9σ threshold instead. But further research needs to be conducted on the handling of non-biological material that appears fluorescent in the instrument.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

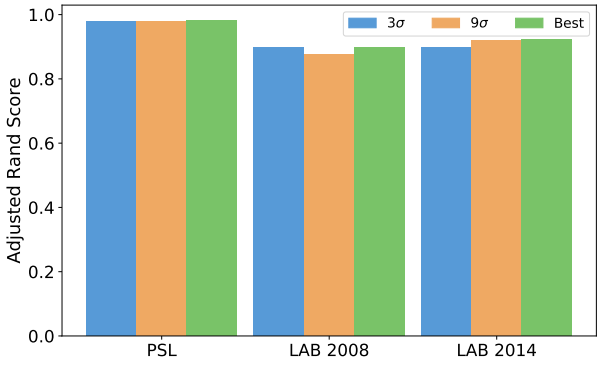

Figure 9. Performance of Gradient Boosting for the different data sets when using 3σ and 9σ.

4.3 Gradient Boosting

We conducted a similar analysis varying data preparation approaches as in Section 4.1. We found data preparation to have a very small impact upon performance when using Gradient Boosting as long as some kind of fluorescence threshold is applied where a high value of the adjusted rand score was obtained regardless of whether we took logs, what standardization was used or the size threshold imposed.

Figure 9 shows the performance using 3σ and 9σ using z-score, taking logs and applying a size threshold of 0.8 microns. High performance was attained across both laboratory generated aerosol data sets and for the PSLs. As we did in the previous sections we provide matching matrices of the worst-case scenario and best case scenario when using Gradient Boosting using the current data preparation in Tables 8 and 9. In the best case scenario we provide a very good classification with very small errors (AR=0.919). The algorithm does a poor job with the remaining non-biological material but there are only 13 non biological particles left for this data set, so the algorithm has very little to train on, but these few particles have very little impact on the quality of the result.

In the worst case scenario a similar performance is achieved (AR = 0.877). Nonetheless, a few particles are incorrectly classified within the fungal spore and pollen classes. The classification for the bacteria is still very strong and most of the remaining non-biological particles are correctly classified.

5 Conclusions

We evaluated a variety of different methods that could be used for classification of biological aerosol. Gradient Boosting offered by far the best performance consistently across different data preparation strategies and the different data sets tested. That being said it is unclear at this point how this will translate to ambient data and whether or not the training data currently collected will be sufficient to outline the variety of environments that could potentially be studied.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-126
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 18 June 2018

Atmospheric Measurement Techniques Discussions

Table 8. Matching matrix for the best case scenario when using Gradient Boosting. This is when using 9σ on 2014 data.

	bacteria	fungal spores	pollen	non-biological
bacteria	1908	8	18	10
fungal spores	6	216	31	7
pollen	6	23	584	5
non-biological	4	5	0	1

Table 9. Matching matrix for the worst case scenario when using Gradient Boosting. This is when using 9σ on 2008 data.

	bacteria	fungal spores	pollen	non-biological
bacteria	6852	89	79	7
fungal spores	51	892	148	3
pollen	9	75	288	1
non-biological	8	4	4	293

Should there not be sufficient training data available we will have to use an unsupervised approach. In this case, a possible alternative to Hierarchical Agglomerative Clustering is found. In the best case scenario DBSCAN, despite leaving a decent proportion of the data unclassified, was able to produce three distinct clusters containing predominantly one biological class each.

To the best of our knowledge this is the first manuscript using DBSCAN to classify biological aerosol using the WIBS. So we will need to continue to evaluate the performance of this algorithm in the context of the ambient setting. In particular, we have provided details of what we believe to be sensible selections of epsilon and the minimum number of points on the basic of the laboratory data collected. However, it is unclear at this point how effective these selections will be when analysing ambient data.

It is clear that Hierarchical Agglomerative Clustering certainly has it drawbacks. When applied the laboratory generated aerosol tested, we found that performance was in general much lower than what could be achieved for the PSLs. Performance was heavily dependent on the data preparation strategy used and often results could vary substantially between different strategies and data sets. Caution will therefore be required when applying the algorithm to ambient data.

In the future, more laboratory generated aerosol particles will need to be collected to continue to evaluate the performance of the algorithms which we use. In addition, even when Gradient Boosting was used we failed to classify the some of the pollen and fungal spores analysed. It is possible that higher spectral instruments will be required to provide a more accurate classification.

Code and data availability. The code used produce the above manuscript is part of an ongoing development of a software suite for analysis of various UV-LIF instruments, which will be made public at https://github.com/simonruske/UVLIF upon publication. Other code not currently included within the software package i.e. code files which are used to produce the plots and figures specific to this paper will be made available on the same site but placed into a different repository.

The data used is available upon request by contacting the lead author.

Competing interests. There are no competing interests that the authors are aware of

Acknowledgements. Simon Ruske is funded by NERC (NERC Grant number: NE/L002469/1) and the University of Manchester.

References

Caliński, T. and Harabasz, J.: A dendrite method for cluster analysis, Communications in Statistics-theory and Methods, 3, 1–27, 1974.

Crawford, I., Bower, K., Choularton, T., Dearden, C., Crosier, J., Westbrook, C., Capes, G., Coe, H., Connolly, P., Dorsey, J., et al.: Ice formation and development in aged, wintertime cumulus over the UK: observations and modelling, Atmospheric Chemistry and Physics, 12, 4963–4985, 2012.

Crawford, I., Ruske, S., Topping, D., and Gallagher, M.: Evaluation of hierarchical agglomerative cluster analysis methods for discrimination of primary biological aerosol, Atmospheric Measurement Techniques, 8, 4979–4991, 2015.

Cziczo, D. J., Froyd, K. D., Hoose, C., Jensen, E. J., Diao, M., Zondlo, M. A., Smith, J. B., Twohy, C. H., and Murphy, D. M.: Clarifying the dominant sources and mechanisms of cirrus cloud formation, Science, 340, 1320–1324, 2013.

D'Amato, G., Liccardi, G., D'amato, M., and Cazzola, M.: The role of outdoor air pollution and climatic changes on the rising trends in respiratory allergy, Respiratory medicine, 95, 606–611, 2001.

Després, V., Huffman, J. A., Burrows, S. M., Hoose, C., Safatov, A., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M., Pöschl, U., et al.: Primary biological aerosol particles in the atmosphere: a review, Tellus B: Chemical and Physical Meteorology, 64, 15 598, 2012.

Gabey, A., Gallagher, M., Whitehead, J., Dorsey, J., Kaye, P. H., and Stanley, W.: Measurements and comparison of primary biological aerosol above and below a tropical forest canopy using a dual channel fluorescence spectrometer, Atmospheric Chemistry and Physics, 10, 4453–4466, 2010.

Gabey, A., Stanley, W., Gallagher, M., and Kaye, P. H.: The fluorescence properties of aerosol larger than 0.8 μm in urban and tropical rainforest locations, Atmospheric Chemistry and Physics, 11, 5491–5504, 2011.

Gabey, A. M.: Laboratory and field characterisation of fluorescent and primary biological aerosol particles, Ph.D. thesis, The University of Manchester, Manchester, UK, 2011.

Gallagher, M., Robinson, N., Kaye, P. H., and Foot, V.: Hierarchical Agglomerative Cluster Analysis Applied to WIBS 5-Dimensional Bioaerosol Data Sets, Atmospheric Chemistry and Physics, 10, 4453–4466, 2012.

Gurian-Sherman, D. and Lindow, S. E.: Bacterial ice nucleation: significance and molecular basis., The FASEB journal, 7, 1338–1343, 1993.

Hader, J., Wright, T., and Petters, M.: Contribution of pollen to atmospheric ice nuclei concentrations, Atmospheric Chemistry and Physics, 14, 5433–5449, 2014.

Healy, D. A., O'Connor, D. J., Burke, A. M., and Sodeau, J. R.: A laboratory assessment of the Waveband Integrated Bioaerosol Sensor (WIBS-4) using individual samples of pollen and fungal spore material, Atmospheric environment, 60, 534–543, 2012a.

Healy, D. A., O'Connor, D. J., and Sodeau, J. R.: Measurement of the particle counting efficiency of the "Waveband Integrated Bioaerosol Sensor" model number 4 (WIBS-4), Journal of Aerosol Science, 47, 94–99, 2012b.

Hoose, C. and Möhler, O.: Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments, Atmospheric Chemistry and Physics, 12, 9817–9854, http://www.atmos-chem-phys.net/12/9817/2012/, 2012.

Kaye, P. H., Stanley, W., Hirst, E., Foot, E., Baxter, K., and Barrington, S.: Single particle multichannel bio-aerosol fluorescence sensor, Optics express, 13, 3583–3593, 2005.

Kennedy and Smith: Health Effects of Climate Change in the UK 2012 : Effects of aeroallergens on human health under climate change, 2012.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., and Sarasvady, S.: DBSCAN: Past, present and future, in: Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the, pp. 232–238, IEEE, 2014.

Milligan, G. W. and Cooper, M. C.: An examination of procedures for determining the number of clusters in a data set, Psychometrika, 50, 159–179, 1985.

Milligan, G. W. and Cooper, M. C.: A study of standardization of variables in cluster analysis, Journal of classification, 5, 181–204, 1988.

Möhler, O., DeMott, P., Vali, G., and Levin, Z.: Microbiology and atmospheric processes: the role of biological particles in cloud physics, Biogeosciences, 4, 1059–1071, 2007.

Müllner, D.: Modern hierarchical, agglomerative clustering algorithms, arXiv preprint arXiv:1109.2378, 2011.

Müllner, D.: fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python, Journal of Statistical Software, 53, 1–18, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Robinson, N. H., Allan, J., Huffman, J., Kaye, P. H., Foot, V., and Gallagher, M.: Cluster analysis of WIBS single-particle bioaerosol data, Atmospheric Measurement Techniques, 6, 337, 2013.

Ruske, S., Topping, D. O., Foot, V. E., Kaye, P. H., Stanley, W. R., Crawford, I., Morse, A. P., and Gallagher, M. W.: Evaluation of machine learning algorithms for classification of primary biological aerosol using a new UV-LIF spectrometer, Atmospheric Measurement Techniques, 10, 695, 2017.

Savage, N. J., Krentz, C. E., Könemann, T., Han, T. T., Mainelis, G., Pöhlker, C., and Huffman, J. A.: Systematic characterization and fluorescence threshold strategies for the wideband integrated bioaerosol sensor (WIBS) using size-resolved biological and interfering particles, Atmospheric Measurement Techniques, 10, 4279, 2017.

Schubert, E., Koos, A., Emrich, T., Züfle, A., Schmid, K. A., and Zimek, A.: A Framework for Clustering Uncertain Data, PVLDB, 8, 1976–1979, http://www.vldb.org/pvldb/vol8/p1976-schubert.pdf, 2015.