Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

Evaluation of Machine Learning Algorithms for Classification of Primary Biological Aerosol using a new UV-LIF spectrometer

Simon Ruske[1], David O. Topping[1, 2], Virginia E. Foot[4], Paul H. Kaye[5], Warren R. Stanley[5], Ian Crawford[1], Andrew P. Morse[3], and Martin W. Gallagher[1]

[1]Centre for Atmospheric Science, SEAES, University of Manchester, Manchester, UK
[2]NCAS, National Centre for Atmospheric Science, University of Manchester, Manchester, UK
[3]Department of Geography and Planning, University of Liverpool, Liverpool, United Kingdom
[4]Defence Science and Technology Lab. (United Kingdom)
[5]Particle Instruments Research Group, University of Hertfordshire (United Kingdom)

Correspondence to: Simon Ruske
(simon.ruske@postgrad.manchester.ac.uk)

Abstract.

Characterisation of bio-aerosols has important implications within Environment and Public Health sectors. Recent developments in Ultra-Violet Light Induced Fluorescence (UV-LIF) detectors such as the Wideband Integrated bio-aerosol Spectrometer (WIBS) and the newly introduced Multiparameter bio-aerosol Spectrometer (MBS) has allowed for the real time collection of fluorescence, size and morphology measurements for the purpose of discriminating between bacteria, fungal Spores and pollen.

This new generation of instruments has enabled ever larger data sets to be compiled with the aim of studying more complex environments. In real world data sets, particularly those from an urban environment, the population may be dominated by non-biological fluorescent interferents bringing into question the accuracy of measurements of quantities such as concentrations. It is therefore imperative that we validate the performance of different algorithms which can be used for the task of classification.

For unsupervised learning we test Hierarchical Agglomerative Clustering with various different linkages. For supervised learning, ten methods were tested; including decision trees, ensemble methods: Random Forests, Gradient Boosting and AdaBoost; two implementations for support vector machines: `libsvm` and `liblinear`; Gaussian methods: Gaussian naïve Bayesian, quadratic and linear discriminant analysis and finally the k-nearest neighbours algorithm.

The methods were applied to two different data sets measured using a new Multiparameter bio-aerosol Spectrometer which provides multichannel UV-LIF fluorescence signatures for single airborne biological particles.

Clustering, in general performs slightly worse than the supervised learning methods correctly classifying, at best, only 72.7 and 91.1 percent for the two data sets respectively. For supervised learning the gradient boosting algorithm was found to be the most effective, on average correctly classifying 88.1 and 97.8 percent of the testing data respectively across the two data sets.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

1 Introduction

Primary biological aerosol particles (PBAP) such as fungal spores, bacteria and pollen have been linked to global atmospheric processes but their impact remains uncertain. In particular, cloud and precipitation feedback mechanisms are dependent on the concentration of the particles as well as the surface area. Quantification of the bio-geography and seasonal variability of such quantities is vital for better understanding the impacts of atmospheric aerosol on the environment.

It is thought that bacteria, pollen and fungal spores can act as cloud condensation nuclei (CCN) and heterogeneous ice nuclei (IN) (Möhler et al., 2007; Hoose and Möhler, 2012). For example bacterial species such as *Pseudomonas syringae* and *Erwina herbicola*, have been shown to be catalysts for the formation of ice at temperatures as warm as -2°C (Gurian-Sherman and Lindow, 1993). Furthermore, ice nucleation active (INA) bacteria have been recovered from cloud water (Joly et al., 2013) demonstrating that bio-aerosols, acting as IN, can be found in the atmosphere, at least where these clouds are present and therefore may be influencing various atmospheric processes.

Only a few bacterial and fungal species have been shown to be INA at the higher range of sub-zero temperatures and even in these cases only a small amount of cells nucleate at these temperatures, leading some to question the significance of bio-aerosols as ice nucleators (Cziczo et al., 2013). However, since on-going research has led to the discovery of new biological ice nucleators (Huffman et al., 2013), there are likely to be more INA species to be found and under certain conditions, such as during rainfall especially at warmer temperatures, these particles may be having a much more profound impact than previously thought (Huffman et al., 2013; Hader et al., 2014; Prenni et al., 2013; Tobo et al., 2013).

The above recent research has led to the development of the hypothesis of a bio-precipitation feedback cycle, whereby plants release aerosol containing microorganisms and spores that then act as ice catalysts at warmer temperatures than other more common ice nucleators, such as mineral dusts. This in turn facilitates precipitation which is beneficial for the growth of plants and microorganisms (Morris et al., 2014). Within such a cycle it may be the case that biological particles initiate secondary ice nucleation processes, also at warmer temperatures (Crawford et al., 2012) leading to more rapid cloud glaciation which may also impact the development of precipitation. Emissions of certain bio-aerosols are also predicted to increase in a warming climate (Jacobson and Streets, 2009), resulting in changing patterns of plant and animal disease spread (Kennedy and Smith, 2012).

Whilst the technology for identification and quantification of specific airborne bio-aerosols exists, measurements of their concentrations and surface properties remain some way off. Nonetheless, the practicality of long-term, continuous, real-time monitoring and discrimination of at least some of these properties for the more common types has already been demonstrated, e.g. at rural and semi-rural background sites in Germany, Ireland and Finland, (Healy et al., 2014; Toprak and Schnaiter, 2013; Schumacher et al., 2013).

Despite the limited observations of the concentrations of bio-aerosols, their effects on the outcomes of global and regional aerosol models have been investigated (Spracklen and Heald, 2014; Hummel et al., 2015). In Spracklen and Heald (2014), simulated concentrations of fungal spores and bacteria are used in a global aerosol model from which they conclude that whilst PBAP contribute very little to average global immersion freezing ice nucleating rates, PBAB dominates ice nucleation

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

at warmer temperatures at certain altitudes. In Hummel et al. (2015), measurements from a number of field sites have also been used to test high-resolution bio-aerosol emission models on European regional scale, from which it is suggested that simulated Fluorescent Biological Aerosol Particle (FBAP) concentrations based on literature emission parametrisation are lower than the corresponded measured concentrations. As well as further field research, evaluation of the algorithms discussed in this manuscript could allow for more certainty in the measurements of the concentrations which would allow for better validation of the above models.

Furthermore, there are other uncertainties which arise from the potential misclassification from interferents, particularly in complex urban environments. Potential non-biological fluorescent aerosol interferents may include black carbon aerosols from seasonally varying solid fuel sources (Herich et al., 2014). Addition of organic films via deposition of Polycyclic aromatic hydrocarbons (PAHs) emitted by vehicle exhausts is another potential interferent, as are common mineral dusts containing fluorescent rare-earth metals. In addition to the compilation of larger data catalogues to help address the issue of interferents (e.g. Hernandez et al., 2016), there also needs to be a focus on testing the effectiveness of approaches to distinguish between particles reliably in real-time.

Hierarchical cluster analysis (HCA), an unsupervised learning technique has been used previously to discriminate between bio-aerosol (Gallagher et al., 2012; Robinson et al., 2013; Crawford et al., 2014, 2015). This technique has been shown to be successful in discriminating between various Polystyrene Latex Spheres (PSLs) and has been applied to ambient data where correct classification is unknown. In this manuscript we extend this research to encompass laboratory samples where correct classification is known, in an attempt to evaluate the performance of such algorithms with data that is more similar to that which could be produced during an ambient campaign.

To enhance our study, we also conduct analysis using a range of supervised methods. There are many advantages and disadvantages of supervised methods versus unsupervised methods. Firstly supervised techniques allow one to choose training data and groupings that better reflect the research problem at hand. For example, for discriminating between bacteria, fungal spores and pollen with the aim of studying how they interact with the atmosphere, one could collect various different samples of the different groups and use this to train supervised methods to identify the particles in ambient data. Conversely the results from the unsupervised methods are dependent on natural differences in the data and cannot be tailored towards a particular application.

On the other hand when faced with a previously unseen particle, the supervised methods may be dependent on the data with which they were trained. Unsupervised methods may offer an advantage in these cases since they are not reliant on training data. Another factor that needs to be considered which is not mentioned in detail in this manuscript, is the time cost of the different methods. Supervised methods such as decision trees and linear discriminant analysis offer much faster alternatives to hierarchical cluster analysis which would be important when considering real-time applications in the future.

Clearly, supervised methods may offer additional benefits making their study worthwhile, but the laboratory data collected prior to ambient studies will be of paramount importance. Specifically we test ten methods available in the scikit-learn package (Pedregosa et al., 2011) including decision trees, ensemble methods: Random Forests, Gradient Boosting and AdaBoost;

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

two implementations for support vector machines: `libsvm` and `liblinear`; Gaussian methods: Gaussian naïve Bayesian, quadratic and linear discriminant analysis (QDA and LDA) and finally the k-nearest neighbours algorithm.

2 Methods

In the classification of biological aerosol the primary aim is to attribute a label to each of the particles. Unsupervised learning requires no prior knowledge and splits the particles into different groups using natural differences in the data. Supervised learning takes a subset of the data, which we will call the *training set*, and uses this 'learn' differences between groups. A testing stage on the remaining data, which we will call the *testing set*, is then conducted. The percentage of the testing set correctly classified is then recorded to evaluate how well the method has 'learnt' how to distinguish between the groups.

For Hierarchical Cluster Analysis (HCA) we varied whether we (a) included both saturated and non fluorescent data (b) included saturated data but not non-fluorescent (c) included only non-fluorescent data but removed saturated data and (d) removed both. We concluded that a particle was non-fluorescent if it's eight fluorescence measurements lay within three standard deviations of the mean measurements when the instrument was empty. Such filtering is common for previous studies using hierarchical clustering but filtering was not considered for the supervised learning methods since the methods should be able to incorporate some kind of filtering within their own classification schemes. For example when using decision trees, removal of non fluorescent data would be replicated using branches that split the data based on the fluorescence above and below a certain threshold. We therefore conclude in the case of supervised methods it is beneficial to allow the method to have full control of how the data is grouped for classification rather than to filter any of the data ourselves.

The structure of Sect.2 is as follows: In Sect. 2.1 we discuss the only unsupervised method we tested - Hierarchical Cluster Analysis (HCA). In Sect. 2.2 we highlight decision trees and ensemble methods encompassing everything from a single decision tree to any method which can be used to combine multiple decision trees in an attempt to create a better classifier (AdaBoost, Gradient Boosting and Random Forests). Gaussian methods are introduced in Sect. 2.3, these include any method that fits a Gaussian model to the data for classification, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Gaussian Naive Bayesian. In Sect. 2.4 we highlight the k-nearest neighbour classifier and finally in Sect. 2.5 we discuss the main differences between the two implementations of support vector machines.

2.1 Hierarchical Cluster Analysis (HCA)

Hierarchical Cluster Analysis (HCA) is the only unsupervised method that we tested. Other unsupervised methods such as k-means clustering and are not considered here since they rely on the user to input the number of clusters, which in an ambient situation this is unlikely to be known prior to the analysis. There are seven available linkages in the Fastcluster package (Müllner, 2013, 2011): single (closest point), complete (furthest point), average (average distance), weighted (weighted average distance), ward (minimisation of variance), centroid (difference between means) and median (differences between medians). Prior to this analysis we used the z-score to standardise the data.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

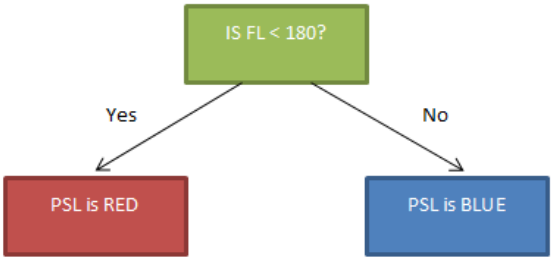

Figure 1. An example of a small decision tree

2.2 Decision Trees and Ensemble Methods

When using Decision Trees, data is split by sorting the data by each variable and using a metric to find the best place to split. An example of a Decision Tree is given in Figure 1. In our example there are two groups: blue particles and red particles, and the variable we use to split them is a measurement of fluorescent intensity of the particle. In reality the tree will be much more complicated with many more branches.

To construct the decision tree we consider all possible splits within the data. For example if we had three particles with fluorescent (FL) intensity of 180, 300 and 1400 absolute units (au), to determine the best split for the first branch we would consider all possible splits. For three particles there would be three possible cases for the first branch: FL $>$ 180, FL $>$ 300, FL$>$ 1400. Each split then will be evaluated using a criterion to determine how effective the split is to distinguish between the different groups. All of the other variables are then considered in the same fashion and the most effective split for the first branch in the data is selected. The process is then repeated to split the data multiple times creating a larger tree with many branches. In the case of our example we would have a tree with two splits. When classifying a new particle we simply start at the top of the tree evaluating the criteria until a conclusion about the particle is made.

Multiple Decision Trees can be combined to create ensemble classifiers. These classifiers often achieve improvements in one of two ways. Firstly, classifiers such as bagging and random forests take samples of the data and the variables which are used to produce different decision trees each capable of classifying a particle. Averaging the classifications made by each tree is then thought to give an overall better result. An alternative approach used by the AdaBoost Classifier and the Gradient Boosting Classifier is to begin by weighting all the data equally and over several iterations have decision trees focus on the parts of the data that are being misclassified most often. This can yield an improvement over the single decision tree as the classifier is modified to correct the mistakes that it is making. These ensemble methods could be theoretically used with other classifiers but the simplicity and speed of the decision trees mean that they are most often used. We give further details of the ensemble methods below.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

Bagging (Breiman, 1996) is where multiple samples of the data are taken and a different tree is fitted to each of the samples. The samples taken are bootstrap samples, a common statistical technique used to create multiple data sets from one set of data. This can be thought as putting all the samples into a bag, taking out one sample at a time and putting it back into the bag. This is repeated until a new data set which is the same size as the original is obtained. Some samples will have been selected more than once from the bag and others may not get selected at all. This gives a subtly different data set. This can be repeated multiple times in order to create multiple versions of the data set. From each of the samples a decision tree is constructed and the results from the different trees are then averaged to give an overall result. The rationale behind the method is that slight differences in the different versions of the data set will produce different trees and in averaging the results we will get a better estimation of which group the particle belongs to.

Bagging is extended to *Random Forests* in Breiman (2001). Instead of selecting the best split when constructing any particular tree, a random subset of variables is chosen to build the tree. It is hypothesised that using only a subset of the variables will produce trees that are more independent and thus the improvement from averaging can be larger. Random forests are generally considered to perform better than bagging hence we do not consider bagging in our analysis.

An alternative method for combining decision trees into an ensemble classifier is AdaBoost (Freund and Schapire, 1995). Here weights are assigned to each of the particles and very small decision trees are fitted to the data. Performance is evaluated using a loss function (exponential loss function) and the data is re-weighted to focus on particles that are being misclassified most often. Gradient Boosting is a generalisation of the AdaBoost algorithm to allow for different loss functions.

2.3 Gaussian Based Methods

An alternative approach to solve the classification problem is to fit multivariate normal distributions to each of the groups within the training data. This distribution is a generalisation of the normal distribution for one variable and depends on the means and covariance of the different variables.

Different assumptions of how the covariance between the variables behaves leads to different classifiers. If one places no requirements on the covariance this results in *Quadratic Discriminant Analysis* (QDA). The assumption that each of the groups vary equally results in *Linear Discriminant analysis* (LDA) and finally making the assumption that each of the variables are independent of each other results in *Gaussian Naïve Bayesian*.

Once the normal distributions are fitted we can classify new particles by calculating the probability that the particle comes from each of the groups and assigning it to the group that it is most likely to have come from.

As an example, in Figure 2, we have plotted normal distributions for three groups. If we were to observe a value of $x = -2$ then the particle would be assigned to the green group as the probability that the particle comes from the green group is much higher than that for the red and for the blue.

2.4 K-nearest neighbour classification (KNN)

This method does not require a training step, instead to classify a previously unseen particle the particle is compared to each of the particles in the training set and the k nearest neighbours in the training set to the previously unseen particle are recorded.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

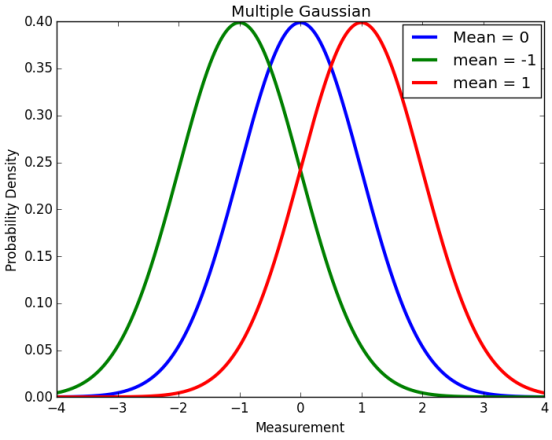

Figure 2. An example of three fitted normal distributions

The previously unseen particle is then attributed to the same group as the majority of its nearest neighbours. This method can be regarded as recognition rather than learning as it classifies a particle simply on how similar it is to particles that it has seen in the training data.

2.5 Support Vector Machines

A practical guide to Support Vector Classification is given in Hsu et al. (2003). The method transforms the data to a higher dimensional space and then splits the data using a linear decision function (Cortes and Vapnik, 1995). In 2 dimensions this would be a line, in three a plane etc. In 2-dimensions, points to one side of the line are classified as coming from one group; points on the other side of the line are classified as coming from the other group. Points to either side of the line correspond to positive and negative values of the decision function respectively. The line is selected on the basis of how well it splits the data without giving too much precedence to outliers.

In order to generalise this methods to multiple groups there are two methods: one-vs-rest and one-vs-one. One-vs-rest involves fitting a support vector machine for each of the groups against the rest of the groups and then attributing new particles to the group with the highest value of the decision function. One-vs-one fits a classifier to each pair of groups and then uses a voting scheme to attribute previously unseen particles to a group. LinearSVC uses the one-vs rest strategy whereas SVC uses the one-vs-one strategy.

How the data is transformed to a higher dimensional space is dependent on the kernel chosen. There are two implementations within scikit-learn (Pedregosa et al., 2011) that can be used for support vector machines: SVC (support vector classification) and LinearSVC (linear support vector classification). The former allows many different kernels, whereas the latter is a faster

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

version of the first but is limited to the linear kernel only. We test SVC using the RBF (radial basis function) kernel and use linearSVC for the linear kernel.

The SVC implementation has parameters γ and C. Since γ is a specific parameter for the RBF kernel, LinearSVC only requires the input of the value of C. Using a sample of 10% of the data, we test the values of C equal to $1, 10, 100, 1000$ and in the case of the SVC function we test all possible combinations of C with γ equal to $0, 1, 10, 100, 1000$. The values are selected to test a wide range of possible values of each of the parameters to allow for appropriate values to be selected. In future, it might be possible to get better performance by either conducting this initial parameter selection on a larger sample of the data, or to test more values, but within the scope of this manuscript we are intending to select parameters that perform fairly well which should give us an appropriate estimation on the effectiveness of the method. The values which perform best are selected to test over 100 samples which will form our final result.

3 Instrumentation

The MBS is a development of the Wideband Integrated bio-aerosol Spectrometer (WIBS) technology developed by the University of Hertfordshire (Kaye et al., 2005). Both instruments are designed to acquire data relating to the size, shape, and intrinsic fluorescence of individual airborne particles and use these data to detect and potentially classify those particles that are of biological origin. However, whereas the WIBS instrument records particle fluorescence over just two wavebands, approximately $310-400$nm and $420-650$nm (corresponding to the maximum emissions from tryptophan and NADH), the MBS records the fluorescence over eight equal wavelength bands from approximately 310 to 640nm. This is likely to provide better discrimination between biological particles and 'interferent' non-biological particles that may exhibit similar fluorescence properties. Similarly, while WIBS uses a simple 4-pixel detector to assess particle shape from the particle's spatial light scattering pattern (Kaye et al., 1996; Kaye, 1998), the MBS uses an arrangement of two 512-pixel CMOS detector arrays to record high-resolution details of the particle's spatial light scattering pattern, allowing both the macroscopic shape of the particle and potentially particle surface characteristics to be determined. Again, this can enhance the prospects of particle classification and reduces false-positive bio-particle detection. The key elements of the MBS are shown in Figure 3.

The MBS draws ambient aerosol through an inlet tube at a rate of approximately 1.5 litres/min. Part of this flow is filtered and used both as a 'bleed' flow (to maintain cleanliness of the inner optical chamber) and as a 'sheath' flow which surrounds and constrains the remaining 'sample' flow. Particles carried in the remaining 300 ml/min sample flow are forced to pass in single-file through the sensing volume defined by the intersection between the particle detection laser beam (see below) and the sample airflow column.

Each particle carried in the sample airflow is initially detected by a low-power laser beam (12mW at 635nm). The light scattered from the laser pulse is collected by the lens assembly shown at the upper-right of Fig. 3 and a small proportion of the light is directed by a pellicle beam-splitter to the photodiode trigger detector. The voltage output pulse of this detector is proportional to the intensity of light falling on it and is used to size the particle. The trigger signal also initiates the firing of a second, high-power, pulsed laser (250mW at 637nm) that irradiates the particle with sufficient intensity to allow elements of

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

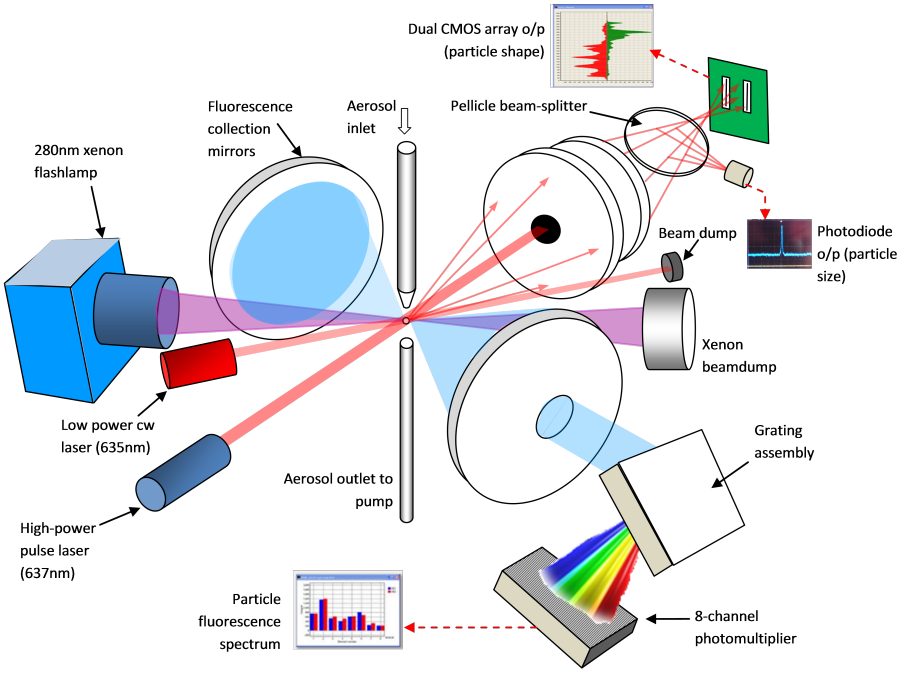

Figure 3. Schematic diagram of the Multiparameter bio-aerosol Spectrometer

the particle's spatial light scattering pattern, which relates to particle morphology and orientation (Kaye, 1998), to be captured by the arrangement of two CMOS linear detector arrays.

About 10μs after particle detection, the UV xenon source illuminates the particle for approximately 1µs with an intense UV pulse at 280nm wavelength. The resulting fluorescent light from the particle is collected by two spherical mirrors and directed through to the spectrometer optics. The fluorescence spectrum, covering 310-650 nm, is recorded by the 8-channel photomultiplier tube and the information is digitised and recorded by the electronics control unit. The particle then passes out of the chamber and the system is re-armed. The total measurement process takes 30μs. Despite the fact that the system is capable of counting particles at a rate greater than 1000 per second, the limiting factor is the xenon recharge time (approximately 5ms) which reduces the data acquisition rate to approximately 100 particles a second (this corresponds to measuring all particles to a concentration of 2×10^4 particles/litre.

Figure 4 below illustrates the type of data produced by the MBS for an individual airborne particle. The particle fluorescence spectrum, excited by the 280nm UV xenon flash, is denoted by the blue bars in the left-hand plot. The red bars represent the average fluorescence values for the previous 25 particles measured. The right-hand plot shows the relative intensity patterns of scattered light from the particle when illuminated by the high-power 637 nm laser pulse. The red and green plots, extending left and right from the centre, correspond to the relative intensities of light falling onto the two linear CMOS detector arrays shown

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

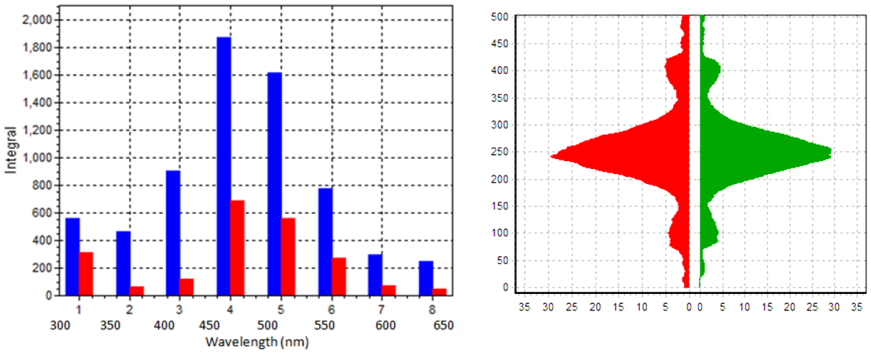

Figure 4. Typical fluorescence spectral data (left) and spatial light scattering data (right) recorded from a single aerosol particle by the MBS instrument

in Figure 3. The symmetry (or asymmetry), form and magnitude of these intensity distributions are related to particle shape and surface structure and are therefore characteristic of the morphology of the illuminated particle, thus offering additional parameters by which the particle may be classified.

4 Data

In order to evaluate the performance of the various different methods we use two different data sets. For each of the data sets we have included a parallel coordinate plot to allow the reader to see on average how each of the groups differ in their fluorescent intensity (see Figures 5 and 6).

4.1 Polystyrene Latex Spheres (PSLs)

From Figure 5 it should be clear that the PSLs should be highly separable by eye. This data set provides a benchmark of the simplest separation task. We would expect a good classification technique to perform well with this data set.

Six groups of spheres, five of which have been doped in colouring, of varying sizes are used. Details of the sizes and the doping of the different groups are given in Table 1. This data is very similar to that used in Crawford et al. (2015) where hierarchical agglomerative clustering was shown to effectively discriminate particles of this kind.

4.2 Laboratory Data

The Laboratory data is from aerosols more representative of the type of aerosol particles that could occur naturally in the environment, which a bio-aerosol sensor would need to be able to discriminate. These data contain examples of various different fungal spores, pollen, bacteria and non-fluorescent material that might be found within ambient data.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

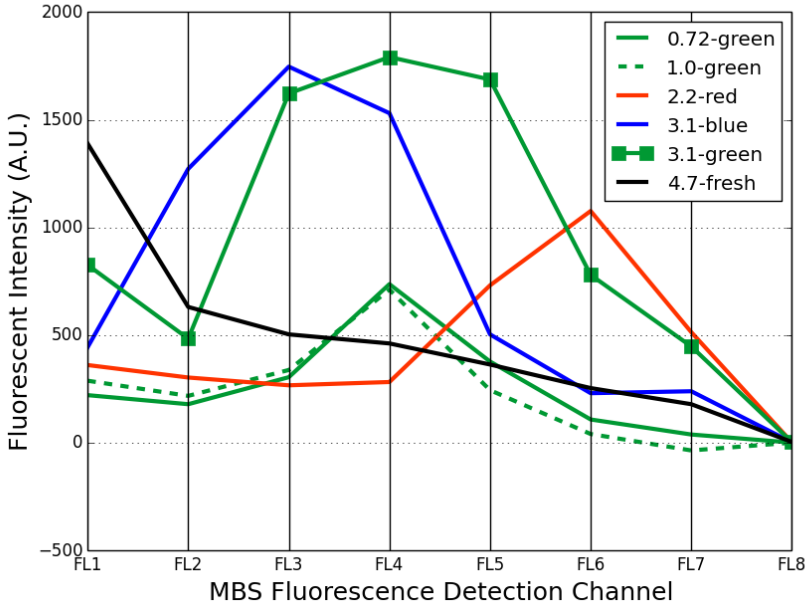

Figure 5. Average Fluorescent intensity given in arbitrary units (A.U) for the eight MBS channels for the PSL Data.

Table 1. Sample Sizes for PSLs

Size	Doping	Number of Particles
2.2 μm	Red	8715
3.1 μm	Green	9681
3.1 μm	Blue	10111
0.72 μm	Green	2917
1 μm	Green	7589
4.17 μm	None	1991

The materials listed in table 2 were aerosolised into a large, clean HEPA filtered containment chamber (incorporating a re-circulation fan), from which the aerosol inlet of the MBS sensor drew the measurement samples. Liquids and suspensions were nebulised using a medical mini-nebuliser (e.g. Hudson RCI Micro-Mist nebuliser), while the dry materials were aerosolised directly from small quantities of powder using a filtered compressed air jet.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

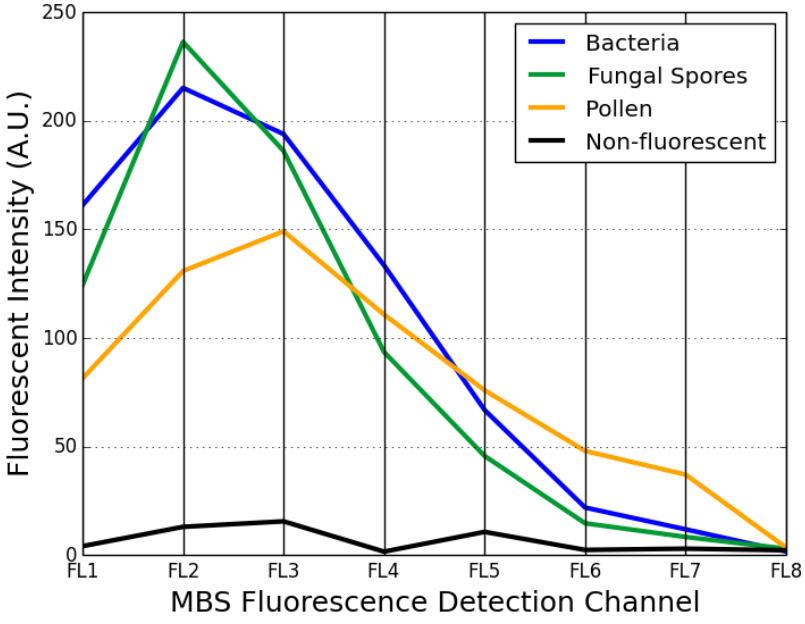

Figure 6. Average Fluorescent intensity given in arbitrary units (A.U) for the eight MBS channels for the laboratory Data.

Table 2. Classification, Generator Method and Sample Size of different samples from the laboratory data.

Material	Generator Method	Classification	# Particles
Bacillus Atrophaeus (BG) Spores (unwashed, in Lbroth)	Mini-nebuliser	Bacteria	1831
Phosphate-buffered saline	Nebuliser	Non-fluorescent	1388
Puff ball Spores	Dry	Fungal Spores	740
Poplar Pollen	Dry	Pollen	469
Paper Mulberry Pollen	Dry	Pollen	1041
Aspen Pollen	Dry	Pollen	466
Bacillus Atrophaeus (BG) Spores (washed, in distilled water)	Mini-nebuliser	Bacteria	1417
Escherichia coli MRE 162 (E.coli) cells (unwashed, in Lbroth)	Mini-nebuliser	Bacteria	1991
Sodium Chloride (Fluka #71376) 1% aqueous solution	Mini-nebuliser	Non-fluorescent	4502
Fuller's earth dust	Dry from mini-nebuliser	Non-fluorescent	3238

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

The BG and E.coli bacteria were generated from suspensions in L-broth growth media, so these aerosols also contain particles of L-broth. Some of the BG spores were also washed before use (by filtering the suspension and re-suspending the spores in distilled water) to obtain relatively clean aerosolised spores.

A second puff ball sample was taken 24 hours after the first. On average they indicated very little fluorescence meaning that the vast majority of the sample was indistinguishable from the non fluorescent material. For this reason we remove the sample prior to the analysis.

In addition, measurements of a rye grass pollen sample was taken but only consisted of approximately 50 particles, substantially less than the other samples, so was also removed.

The remaining particles were split into four broad groups: bacteria, fungal Spores, pollen and non-fluorescent material. Details of the sample sizes and group classifications are given in Table 2.

5 Results

After being split into training and testing data, as outlined in Sect. 2, the proportion of the testing data correctly classified for each of the supervised methods for each of the data sets is given in Table 3. In the case of the unsupervised method (HCA) it was not necessary to split the data into training and testing sets. Instead we applied the algorithm with all 7 available linkages to all the particles. The results for which, for ease of comparison, are also given in table 3.

When applying HCA we investigated whether removal of non-fluorescent particles and saturated particles gave better performance. For the PSL data set the best results were achieved by using all the data in the HCA analysis (including both saturated and non-fluorescent material), for the Laboratory data and the mixes however, it was beneficial to remove saturated and non-fluorescent material before conducting HCA analysis. Pre-filtering wasn't applied to the supervised methods as explained in the Sect. 2. Only the best results are listed in Table 3 i.e. for the PSLs the results listed are from when all the particles were included and for the Laboratory data the results obtained from the removal of non-fluorescent and fluorescent material are listed.

With the inclusion of the 1024 shape measurements we have a high dimensional data set, without we have a relatively low dimensional data set (9 dimensions). To give a good indication of each of the algorithms robustness to dimensionality as well as well as to ascertain whether the additional shape information yields any benefit, we provide results for both the full data set (1024 shape measurements, 8 fluorescent measurements and 1 size measurement) and the reduced data set (8 fluorescent measurements + 1 size measurement).

In Figure 5 we see that the dye-doped PSLs should be highly separable by eye whereas in Figure 6 it appears the laboratory data would present more of a challenge to the different algorithms. This is demonstrated also in our results where the percentage of data correctly classified for the laboratory data is much lower than that of the PSLs.

It is also clear from our results that supervised methods appear to be performing better than hierarchical agglomerative clustering.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

Table 3. Performance of the different methods for the different data sets. The best three methods for each of the columns are given in bold.

	Full Data		Reduced Data	
Method	**Lab Data**	**PSL**	**Lab Data**	**PSL**
Decision Trees	78.0	95.2	76.3	93.8
Adaboost	82.8	90.3	83.5	81.9
Random Forest	82.5	95.4	84.1	**95.0**
Gradient Boosting	**87.6**	**98.1**	**85.5**	**95.3**
SVC	**83.2**	**97.3**	**85.0**	**94.8**
Linear SVC	77.5	97.0	**84.9**	**94.8**
LDA	73.9	**98.5**	73.0	94.0
QDA	56.3	90.1	77.0	94.1
K-nearest neighbour	**84.4**	92.7	81.6	92.9
Gaussian Naïve Bayesian	44.7	74.3	70.2	91.4
HCA (Single)	72.7	24.7	65.2	24.7
HCA (Complete)	72.7	38.5	67.3	71.2
HCA (Average)	72.7	24.7	67.5	75.3
HCA (Weighted)	72.7	41.1	67.4	87.4
HCA (Ward)	75.3	48.4	67.5	91.1
HCA (Centroid)	72.7	24.7	67.5	90.3
HCA (Median)	72.7	24.7	67.2	46.2

HCA performs consistently across the different linkages, in general, with Ward linkage being the best performer by a few percent in all cases. This is consistent with the results obtained in Crawford et al. (2015). However, when applying the algorithms to data that more closely resembles data collected in an ambient situation such as the laboratory data, the unsupervised methods seems to struggle with discrimination.

It is important to note also that the reduction of the data set is clearly beneficial in the case of the PSLs, but the converse is true for the laboratory data. In the case of the PSLs, since all the particles are near spherical the shape information may be hindering results rather than improving them; whereas for the laboratory data where the morphology of the particles may be more complex the shape information may be of value for discrimination.

The problem of dimensionality is also clear with the Gaussian methods. Here, Quadratic Discriminant Analysis (QDA) performs much better with the reduced data set where a reduction of the dimensionality increases results substantially. This is likely due to the covariance matrix being numerically difficult to invert in the cases where the dimensions are close to the number of samples available. One would expect as larger samples are compiled, QDA would start to outperform LDA for the

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

full data as well as the reduced data. Gaussian Naive Bayesian on the other hand appears to only perform well for the reduced PSL data. It is likely that underlying assumptions of such a method will be adhered to in only rare circumstances yielding it inappropriate for the task at hand.

Decision trees and ensemble methods appear to be relatively robust to the introduction of the full data. This is to be expected since most of the methods undergo some kind of variable selection. Gradient boosting however does seem to offer improvements on the AdaBoost algorithm and random forests seem to improve on decision trees as is suggested in the literature.

Overall the best performing method was linear discriminant analysis (LDA) for the PSLs and for Lab Data the Gradient Boosting Algorithm performed better. Note however Gradient boosting only classified 0.4% less of the data correctly than LDA in the case of the PSLs data so overall our results indicate that Gradient Boosting is be the best performing algorithm.

6 Conclusions

UV-LIF is becoming a widely used and accepted method for collecting fluorescent signatures for bio-aerosols. However, the applicability of the method has yet to be demonstrated for routine real-time monitoring and reporting applications for airborne biological particles. In this manuscript we have combined the well developed and researched field of machine learning with the application of identifying atmospheric aerosol. We have demonstrated that previously used unsupervised methods may not be best at discriminating between aerosol using single particle broadband UV-LIF spectrometers and using the MBS we have identified the gradient boosting classifier as a possible supervised alternative.

Since these supervised learning algorithms have yet to be applied to the data produced using the WIBS it is not currently possible to draw any clear conclusions as to the performance of the MBS versus the WIBS. Instead the authors suggests that to provide direct comparison, further research needs to be undertaken whereby both instruments are used for identical samples.

In addition within this paper, we have not provided a thorough analysis of the time requirements of each of the methods. Hierarchical agglomerative clustering is known to be time consuming for a large data set. Even with the development of the Fastcluster package which offers substantial improvements on speed compared to other clustering packages, analysis of a data set with approximately 1 million particles will still take several hours on a modern computer. Hence further experiments need to be conducted to give a quantitative comparison of the newly introduced supervised methods in terms of their time.

Acknowledgements. Simon Ruske is funded by a NERC PhD project P118897 as part of the Manchester-Liverpool Doctoral Training Partnership. The MBS instrument was funded by the NERC research grant, NE/K006002/1 "Ice Nucleation Process Investigation and Quantification.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

References

Breiman, L.: Bagging predictors, Machine learning, 24, 123–140, 1996.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Cortes, C. and Vapnik, V.: Support-vector networks, Machine learning, 20, 273–297, 1995.

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., Robinson, N., Flynn, M., Foot, V., Gallagher, M., Huffman, J., Stanley, W., and Kaye, P. H.: Characterisation of bioaerosol emissions from a Colorado pine forest: results from the BEACHON-RoMBAS experiment, Atmospheric Chemistry and Physics, 14, 8559–8578, 2014.

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.

Freund, Y. and Schapire, R. E.: A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational learning theory, pp. 23–37, Springer, 1995.

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., Huffman, J., O'Connor, D., Pöhlker, C., Pöschl, U., and Sodeau, J.: Ambient measurements of biological aerosol particles near Killarney, Ireland: a comparison between real-time fluorescence and microscopy techniques, Atmospheric Chemistry and Physics, 14, 8055–8069, 2014.

Herich, H., Gianini, M., Piot, C., Močnik, G., Jaffrezo, J.-L., Besombes, J.-L., Prévôt, A., and Hueglin, C.: Overview of the impact of wood burning emissions on carbonaceous aerosols and PM in large parts of the Alpine region, Atmospheric Environment, 89, 64–75, 2014.

Hernandez, M., Perring, A., McCabe, K., Kok, G., Granger, G., and Baumgardner, D.: Composite Catalogues of Optical and Fluorescent Signatures Distinguish Bioaerosol Classes, Atmospheric Measurement Techniques Discussions, 2016, 1–17, http://www.atmos-meas-tech-discuss.net/amt-2015-372/, 2016.

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.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al.: A practical guide to support vector classification, 2003.

Huffman, J. A., Pöhlker, C., Prenni, A., DeMott, P., Mason, R., Robinson, N., Fröhlich-Nowoisky, J., Tobo, Y., Després, V., Garcia, E., et al.: High concentrations of biological aerosol particles and ice nuclei during and after rain, Atmospheric Chemistry and Physics Discussions, 13, 1767–1793, 2013.

Hummel, M., Hoose, C., Gallagher, M., Healy, D., Huffman, J., O'Connor, D., Pöschl, U., Pöhlker, C., Robinson, N., Schnaiter, M., et al.: Regional-scale simulations of fungal spore aerosols using an emission parameterization adapted to local measurements of fluorescent biological aerosol particles, Atmospheric Chemistry and Physics, 15, 6127–6146, 2015.

Jacobson, M. Z. and Streets, D. G.: Influence of future anthropogenic emissions on climate, natural emissions, and air quality, Journal of Geophysical Research: Atmospheres, 114, 2009.

Joly, M., Attard, E., Sancelme, M., Deguillaume, L., Guilbaud, C., Morris, C. E., Amato, P., and Delort, A.-M.: Ice nucleation activity of bacteria isolated from cloud water, Atmospheric environment, 70, 392–400, 2013.

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

Kaye, P. H.: Spatial light-scattering analysis as a means of characterizing and classifying non-spherical particles, Measurement Science and Technology, 9, 141, 1998.

Kaye, P. H., Alexander-Buckley, K., Hirst, E., Saunders, S., and Clark, J.: A real-time monitoring system for airborne particle shape and size analysis, Journal of Geophysical Research: Atmospheres, 101, 19 215–19 221, 1996.

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

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.

Morris, C. E., Conen, F., Alex Huffman, J., Phillips, V., Pöschl, U., and Sands, D. C.: Bioprecipitation: a feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere, Global change biology, 20, 341–351, 2014.

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.

Prenni, A., Tobo, Y., Garcia, E., DeMott, P., Huffman, J., McCluskey, C., Kreidenweis, S., Prenni, J., Pöhlker, C., and Pöschl, U.: The impact of rain on ice nuclei populations at a forested site in Colorado, Geophysical Research Letters, 40, 227–231, 2013.

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, 2013.

Schumacher, C., Pöhlker, C., Aalto, P., Hiltunen, V., Petäjä, T., Kulmala, M., Pöschl, U., and Huffman, J.: Seasonal cycles of fluorescent biological aerosol particles in boreal and semi-arid forests of Finland and Colorado, Atmospheric chemistry and physics, 13, 11 987–12 001, 2013.

Spracklen, D. and Heald, C. L.: The contribution of fungal spores and bacteria to regional and global aerosol number and ice nucleation immersion freezing rates, Atmospheric Chemistry and Physics, 14, 9051–9059, 2014.

Tobo, Y., Prenni, A. J., DeMott, P. J., Huffman, J. A., McCluskey, C. S., Tian, G., Pöhlker, C., Pöschl, U., and Kreidenweis, S. M.: Biological aerosol particles as a key determinant of ice nuclei populations in a forest ecosystem, Journal of Geophysical Research: Atmospheres, 118, 2013.

Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-214, 2016
Manuscript under review for journal Atmos. Meas. Tech.
Published: 13 July 2016

Atmospheric Measurement Techniques Discussions

Toprak, E. and Schnaiter, M.: Fluorescent biological aerosol particles measured with the Waveband Integrated Bioaerosol Sensor WIBS-4: laboratory tests combined with a one year field study, Atmos. Chem. Phys, 13, 225–243, 2013.