Response to reviewer comments on amt-2018-126

Simon Ruske

September 14, 2018

1 Introduction

The following document outlines the author's response to the two reviewer comments on the manuscript amt-2018-126. The following colours are used to differentiate between the reviewer comments, the response provided by the authors and the actions taken as a consequence of the comment.

Blue	Comments made by the reviewer
Black	Response from authors to address comment
Red	Action points (AP) that have been taken to address the comment

2 Response to Anonymous Referee

We thank the reviewer for taking the time provide such a thorough review which will aid substantially in developing the manuscript to the standard required for publication.

2.1 Main comments

2.1.1 Why have you clustered down to two or three clusters?

"1. Reasons for clustering in some cases to 2 or 3 clusters is not clear. Figure 6 with LAB 2008 which illustrates the worst rand index, has only two clusters. Why is a case of HAC with only two clusters shown here? There are quite a few papers in the literature applying HAC to atmospheric aerosol. I can't remember any which clustered down to two. There are 9 samples in the 2008 data set combined into four main categories (bacteria, fungal spores, pollen, and smoke). Wouldn't a reasonable number of clusters be expected to be 9 or somewhere between 4 and 9? I do not see how two clusters makes sense. Also, there are 10 samples in the 2008 data set in four main categories (bacteria, fungal spores, pollen, earth and two NaCl samples, with and without phosphate buffer). Wouldn't a reasonable number of clusters be expected to be 10 or close to 10? p.12, line 8-9 "In the worst case scenario two clusters are provided both primarily containing bacteria. In this case we can conclude the algorithm has failed to differentiate between any of the biological classes." I don't see how the failure is an intrinsic feature of HAC. The failure, at least in part seems attributable to the choice to use two clusters. Table 5 show the bacteria, spores, pollen and non-bio in each of the two clusters for the 2008 data. The discrimination of these clusters is remarkably poor. Why not first cluster to 9 and then show a table such as Table 5 but for the 9 particle types? The same applies to Fig. 4 why force these four sample types into three clusters? The results are confusing enough that I'd recommend showing dendrograms for the clusters of both 2008 and 2014 data sets, and discussing these dendrograms in relation to the data illustrated in Figs. 4 and 5."

We believe this confusion has arisen from the author's inadequate description of Figure 6, which we detail more thoroughly below. Since the response to this comment is substantial we have provided bullet points summarising the response, followed by a more detail explanation.

- A two cluster solution is presented since this is the solution for which the maximum of the CH index was attained.
- Clusterings containing between 1 and 10 clusters have been produced, but this was not made clear in our description of Figure 6.
- Figure 6 shows the maximum value of the adjusted rand score for clusterings containing between 1 and 10 clusters and the adjusted rand score for the clustering which produced the highest value of the Calinski-Harabasz index.
- A dendrogram alongside a heat map of the matching matrix for the clustering containing 10 clusters has been produced. From the plot we see material from two different broad categories e.g. fungal and pollen has been grouped together prior to the final stages of the hierarchy.

Some of the potential sources of error when using HAC are: errors due to the hierarchical agglomerative clustering routine, errors due to the clustering index used to determine the number of clusters and errors due to the data preparation used prior to the algorithm being applied. Figure 6 has been created in an attempt to differentiate between these errors.

Prior to HAC being applied, we label each particle depending on the broad category (1 for bacteria, 2 for fungal, 3 for pollen and 4 for non-biological). Hierarchical agglomerative clustering is then ran on every possible combination of the considerations provided in Table 3, producing a total of 96 hierarchies.

Subsequently, the clusterings containing between 1 and 10 clusters are extracted from each hierarchy and two statistics are calculated. First, the adjusted rand score is calculated as a measure of how similar each clustering is to the known labels. This measure is intended to provide an indication of performance and would be unavailable during an ambient campaign as cluster membership would not be known. We also calculate the Calinski-Harabasz index (CH index) for each of the clusterings containing between 1 and 10 clusters. This is a statistic usually calculated to determine the number of clusters for data collected in an ambient campaign. In Figure 6, we then present two values of the adjusted rand score. First, the maximum value of the adjusted rand score across the first 10 clusterings (presented in the dark bars). Second, the value of the adjusted rand score in the case of the clustering for which a maximum value of the CH index was obtained (in the light bars). This is intended to demonstrate errors that arise due to the CH index. For example, on the laboratory data collected in 2014 the CH index attains a maximum for the 3 cluster solution (shown in the light bar) and the maximum adjusted rand score was attained also for the 3 cluster solution (shown in the dark bar). This is an example where the CH index has worked at intended attaining a maximum for the clustering which is most similar to the known labels.

Scores are presented when using the data preparation strategy suggested in Crawford et al. (2015) modified to use a fluorescent threshold of either 3 (in blue) or 9 (in orange) standard deviations above the average forced trigger measurement as first suggested in Savage et al. (2017). The green bars show the best result across all 96 strategies tested.

In some of the cases presented in Figure 6, the adjusted rand score and calinski-harabasz attained a maximum for different clusterings. For example, in 2008 when using a fluorescent threshold of 3 standard deviations above the forced trigger measurement, we see that a maximum of the CH index was attained for the 4 cluster solution whereas the most similar clustering to the known labels was for the 5 cluster solution. In this case, the 4 cluster solution was nearly as similar to the known labels as the 5 cluster solution, so concluding that there are 4 clusters instead of 5 is reasonable since the 5 cluster solution was only marginally better than the 4 cluster solution.

However, in 2008 when using a fluorescent threshold of 9 standard deviations above the average forced trigger measurement, poor performance is observed where the CH index attains a maximum for the 2 cluster solution when the most similar clustering to the known labels was the 5 cluster solution. For this specific case, part of the poor performance is due to the CH index attaining a maximum at 2 clusters. But we also see that largest adjusted rand score for clusterings with between 1 and 10 clusters (presented in the dark bar) is still quite low. So the conclusion is that better performance could be attained if the index used concluded there were 5 clusters, but also better performance could be obtained should a different data preparation

strategy be used. The maximum of the adjusted rand score across all data preparation approaches tested (shown in the green bars) for 2008 however is still approximately 0.6 whereas for the same data Gradient Boosting attained an adjusted rand score of nearly 0.9.

To conclude, part of the poor performance is due to the index used to determine the number of clusters, part is in the choice of data preparation used, though we failed to provide a recommendation for a data preparation technique which performed consistently well across all the data sets tested, and part is due to selecting to use the HAC algorithm rather than one of the other the algorithms tested.

We have investigated the tendency of the CH index to conclude that there are 2 clusters in further detail and found that it is possibly due to a larger proportion of the data set consisting of bacterial samples. To demonstrate this point further we have simulated data from three normally distributed clusters centred around (0, 0), (5, 5) and (10, 10) and varied the proportion of the material placed in the cluster with the largest number of particles. As the proportion of the data which is sampled from the largest cluster increases, there is a point where the likelihood of the CH index to make the correct conclusion sharply drops. The proportion of data placed in the largest cluster before this sharp drop occurs is dependent upon the variability of the clusters. If we set the standard deviation of each of the clusters to $\sigma = 3$, it would only require approximately 70% of the data to be from one cluster, before this sharp drop in the accuracy of the CH index occurs. In an ambient environment where we would expect concentrations of bacteria to be an order of magnitude greater than the concentrations of fungal spores, this tendency of the CH index to conclude two clusters could be a significant disadvantage and investigating alternative indices for determining the number of clusters may be required in future studies.

We agree that dendrograms would aid in interpreting these results and the reviewers suggestion for providing Table 5 but for the 10 cluster solution has also been added to the same figure.

AP1 Produced dendrogram plots alongside a heat map of the matching matrix comparing the 10 cluster solutions to the known labels.

AP2 Rewritten the HAC results section to make clearer what analysis has been conducted and more adequately describe Figure 6.

AP3 Produced an additional section to indicate why the CH index has a tendency to conclude that there are 2 clusters.

AP4 Split the HAC section into subsections highlighting potential considerations for data preparation, the CH index, and potential issues in the hierarchy that could not be rectified by the selection of a different index to determine the number of clusters.

2.1.2 Why is it valid to cluster down to 2 or 3 clusters for some algorithms but not all?

"2. In comparing the value of classification/clustering approaches the justification for using different number of clusters for different methods is not clear. p.15, lines 7-10: "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)." In tables 8 and 9, four clusters (which are the minimum number that makes sense) were used in testing Gradient Boosting, while two or three clusters were used in testing HAC (1 or 2 less than the number of categories compared with) in Tables 4 and 5, and two or three clusters and an additional category for Unclassified were used in testing DBSCAN in Tables 6 and 7. (Table 6 has 2014 and Table 7 has 2008 data). Because of the use of smaller number of clusters than categories for HAC and DBSCAN, but the same number of clusters and categories for Gradient Boosting, I cannot see how these results say anything about the relative value of HAC, DBSCAN and Gradient Boosting. One cannot set the metric based on four categories, do HAC and DBSCAN down to two or three clusters, but generate four categories with Gradient Boosting, and then compare decide on the better algorithm based on the matched results."

We thank the reviewer for bringing the lack of clarity regarding how the number of clusters was set for each algorithm to our attention. Dependent on which algorithm is used the number of clusters is set in a different way.

In the case of HAC, the number of clusters has been determined by finding a maximum value of the CH index for the clusterings from 1-10. Since this is what would be used when analysing ambient data it is important to present the 2 and 3 cluster solutions since these solutions would be obtained if we analysed the laboratory data in exactly the same fashion as we would analyse ambient data.

In the case of DBSCAN, we have no direct control over the number of clusters as this is determined indirectly by the choice of epsilon and the number of points required to form a neighbourhood. We have performed a grid search for these parameters, testing a variety of combinations, and there are a number of selections for which a result with more than 3 clusters were produced. However, we attempted to select epsilon and the minimum number of points by inspection of Figure 7 on the basis of which parameters resulted in a consistent performance across the data sets tested and these selections resulted in 2-3 clusters.

Irregardless of the algorithm used, we have tested a fluorescent threshold of both 3 and 9 standard deviations above the average forced trigger measurement. The purpose of such threshold is to remove the instrument noise. Since the vast majority of the non-biological samples in 2014 will fail to exceed these thresholds and hence have been removed, we believe a 3 cluster solution would be reasonable for the 2014 data.

It is important to note in the case of HAC we have tested 96 combinations of data preparations, produced hierarchies for each, calculated the adjusted rand score for the clusterings between 1 and 10 clusters and at no point did we attain an adjusted rand score of greater than 0.75 for the laboratory generated aerosol collected in 2008 and 2014 (this can be seen in Figure 6). However, when using DBSCAN, with the exception of the 3 sigma threshold on the 2008 data, we were able to obtain an adjusted rand score of greater than 0.8 after removing between 25 and 35% of the data (as indicated in Figure 7). In the case of the Gradient Boosting algorithm we consistently attain an adjusted rand score of greater than 0.8. We believe that this should give a strong indication of the potential of considering alternatives to HAC for this particular application.

AP5 Produce a table that highlights the potential advantages and disadvantages highlighted within the current study and how the numbers of clusters are determined in each case.

2.1.3 Why combine the data into four categories?

"3. Why is there such confidence in the assumption that combining into categories is valid and appropriate for deciding between classification schemes? Why is there such a focus on combining all the bacteria into one category, pollens into one category, and fungal spores into one category? Why not differentiate into all the categories measured, test on that and then combine the results for each to obtain the results for all the pollens etc.? The two smut spore samples (2008) have similar features, but these are different from the puff ball spores (2014), and as far as I know, very different from the large majority of spore types. Maybe I'm misunderstanding what is done here. The two bacteria used here likely make sense to go into one category. Their FL look similar. I'm assuming the goal is to compare techniques for their capability to help understand atmospheric aerosol. Because of the way the conclusions are stated, this work implies that we can have some confidence that results made on clustering to a category "bacteria" makes sense. However, bacteria that survive in sunlight in the atmosphere tend to be more pigmented than E. coli. How about citing an article such as, Y. Tong and B. Lighthart, Solar Radiation is Shown to Select for Pigemented bacteria in the Ambient Outdoor Atmosphere, Photchem Photobiol 1997, pp 103-106, in at least acknowledging the two bacteria used here are not necessarily representative of bacteria in outdoor air. An explanation of the validity of the bacteria category, while taking into account bacterial pigments and fluorophores such as melanins and carotenoids could be helpful."

There are varying levels of complexity when attempting to discriminate between biological aerosol. In order of difficulty: 1) to be able to discriminate between biological and non-biological material 2) to be able to discriminate between the broad pollen, bacteria and fungal categories and 3) to be able to discriminate

between different species of pollen, bacteria and fungal spores. If an algorithm was capable of discriminating between the different samples, it seems logical that the algorithm would be capable of discriminating between the broad biological classes. Similarly if an algorithm could not discriminate between the broad biological classes we would believe that the algorithm would not be able to discriminate between the individual samples.

After producing the dendrograms suggested and matching matrices of the ten cluster solutions against the samples as suggested by the reviewer in Section 2.1.1, it has become more apparent that fungal material is being grouped with pollen material (2008) or with bacterial material (2014) prior to the final stages of the algorithm. So we do not believe HAC is not segregating by sample either. Instead, fungal material is being grouped with other classes prior to the final stages of HAC.

Given that there are still significant errors in classifying between fungal and pollen samples irregardless of algorithm used, we would suggest that attempting to discriminate between individual samples may be more successful using more recently developed instruments such as the MBS and the WIBS NEO or in future analyses where a wider variety of samples are collected.

We have read Tong and Lighthart (1997) which does indicate that pigmented bacteria is more prevalent in the presence of solar radiation, increasing to $50 - 60\%$ at noontime compared to approximately 33% at midnight. That being said, unless we are misunderstanding the article, these findings seem to indicate that non-pigmented bacteria do constitute between 40% and 67% of outdoor air dependent on time of the time of day. So, if pigmentation did significantly change the fluorescence response from the instrumentation it would seem that collection of both pigmented and non-pigmented bacteria would be required to characterise outdoor air.

A comment regarding the description of fluorophores within the current paper has also been made during the technical review stage, where it has been noted that citing Pöhlker et al. (2013) may be helpful. We also found an earlier article by the same author (Pöhlker et al., 2012), which we believe may also be of value in investigating the potential influence of pigments such as melanin and carotenoids.

In Table 1 in Pöhlker et al. (2012), a wide range of atmospherically relevant biological fluorophores are summarised, including a number of pigments. The excitation wavelengths reported for melanin and carotenoids, $469 - 471$nm and $400 - 500$nm respectively, are different to the excitation used by the WIBS which is 280nm and 370nm. Melanin is also noted to have relatively low fluorescence intensity and is estimated to have low relevance for fluorescent biological aerosol particle (FBAP) detection. Despite Pöhlker et al. (2012) suggesting that the role of carotenoids in FBAP detection is high, there are articles on the fluorescence of carotenoids that report very low fluorescence (Gillbro and Cogdell, 1989). It therefore does appear that carotenoids and melanins probably do not significantly influence the fluorescence response from the WIBS.

To further investigate pigmentation we also read an article specific to Bacillii (Khaneja et al., 2010). Colonies of *Bacillus atrophaeus* (one of the samples presented in the current study) were presented that appeared yellow-orange and others which appeared grey, although carotenoid production in the yellow-orange samples was low. *Bacillus subtilis*, which has been presented in Hernandez et al. (2016), is noted to carry a melanin-like compound to protect against solar radiation but shows very similar fluorescence response to the *E. coli* sample also presented in Hernandez et al. (2016). Given that a significant proportion of the bacterial content in the UK is believed to be Bacillii (Harrison et al., 2005), the inclusion of *Bacillus atrophaeus* as one of our samples seems sensible.

Whether additional bacteria with more varied pigmentation will be required to characterise the outdoor environment is an issue that could be investigated more thoroughly in further research through the collection of a wider range of bacteria and, if possible, by measuring the instrument response to pigments. Whilst the impact of pigmentation in bacteria cannot be fully addressed at this point in time, we recognise that our current description of fluorophores is lacking, and have attempted to make improvements in response to this comment.

AP6 Updated the description of the fluorophores in the introduction to include additional references.

2.1.4 Is size a useful measurement?

"4. Is size a useful measurement for classification of all these particle types? Why is size treated as a useful quantity in defining clusters when actual pollens of the species used here have sizes much larger than the sizes used in the study (as indicated in Tables 4 and 5)? It seems that the samples of pollens are of pollen fragments. Is there evidence that the size distributions of the pollens and fungal spores used in classification here are similar to those in atmospheric particles? The fungal spores also seem to be fragments. I'll assume the "size" is diameter or some effective diameter for non-spheres. Then puffball diameter (avg. approx. 2 um in Fig. 5), is less than half the value for puffball spores, as far as I know. I think smut spores are 6 to 9 um, much larger than the 4 um or smaller shown in Fig. 4. The hypothesis that these are fragments of spores seems more likely than the size calibration is incorrect? Some discussion of the relation to size and ambient sampling for pollen and fungal spores is needed, especially if fragments are the objective or part of the objective. Larger particles of one material should fluoresce more strongly than smaller particles, so I can see the usefulness of size or volume for normalising the FL. But if the algorithms used here benefit from clustering by size, some papers should be cited on the size distributions of pollen and fungal spore fragments measured in the atmosphere. In any case, the sizes in Fig. 4 and 5 need error bars."

We feel that size is definitely a useful measurement for all of these particle types. To better understand the relationship between size and fluorescence for each of the samples collected we have produced fluorescence versus size on scatter plots which we intend to include in the supplementary material. As we did in Section 2.1.1, we provide summary bullet points followed by a more thorough response.

- The WIBS measurement of size is different to other size measurements, but we feel that it is useful in discriminating between the particles collected.
- Whether the samples are fragments could be more thoroughly investigated in future by using microscopy in future experiments, but we agree with the hypothesis that a large quantity of the pollen samples collected are likely fragments.
- We have included additional references to compare the size ranges in the current study to other studies using the WIBS as well as other studies which use microscopy.
- As newer versions of the instrument are developed, measurements for larger particles, including more intact pollen will become possible.

The WIBS size measurement is an optical scatter calculation, calibrated with unit density polystyrene latex (PSL) spheres. If the particles are a different density (dry pollen) or refractive index (soots), or irregular in shape (dusts, clumps etc) then the resultant size measurement will likely be different from alternative measurements such as those from viewing the particles under a microscope. Whether or not the particles are fragments or not could be more thoroughly investigated through future research, i.e. by collecting filter samples of the particles after they are placed into the chamber.

We have now included several additional citations on the size ranges expected for the samples collected. Fortunately, samples of the same or similar species have previously been collected using the WIBS instrumentation across a number of studies (Healy et al., 2012; Hernandez et al., 2016; Savage et al., 2017). In Healy et al. (2012) a size range of $\sim 3 - 30\mu$m (low gain) is used when collecting pollen samples, whereas for the fungal samples they collected they used high-gain mode ($\sim 0.5 - 12\mu$m). The "low gain" mode, is available in the WIBS version 4 but the WIBS version 3 which is used to collect the data presented is limited to an approximate size range of between 0.5 and 12 microns according to Healy et al. (2012), although we did measure particles as large as 14 microns. In Hernandez et al. (2016), low-gain was used for the fungal and the pollen samples whereas the high-gain mode is used for the bacterial samples. In Savage et al. (2017), microscopy is used to support the hypothesis of a mixture of intact pollen and fragmented pollen being present in the samples collected. The size ranges for the pollen samples, with the exception of the mulberry sample, presented in the current study are similar to those presented in Hernandez et al. (2016) and would be consistent with the hypothesis that the sample is comprised mostly, if not entirely, of fragmented pollen. The mulberry sample has been also analysed in Healy et al. (2012) where an average size of $13.6 \pm 6.2\mu$m is presented, very similar to the value of 13.8μm which has been presented in other studies using microscopy

(Kang et al., 2007). The sizes of the paper mulberry samples presented here are 7.18 ± 4.74 and 3.40 ± 1.42 for two of the sample files from 2008 and then 11.27 ± 1.74 for the one sample file collected in 2014. For the first sample of Paper Mulberry collected in 2008, it would seem that we have a mixture of intact and fragmented pollen, whereas in the second sample could be entirely comprising of pollen fragments. In 2014, we may be viewing a sample consisting of primarily intact pollen, albeit only the smaller tail of the size distribution presented in Healy et al. (2012).

Inspecting the scatter plots of size versus fluorescence (which will be placed in the supplementary material for the re-submission), we can see in the case of one of the puffball samples there a number of particles just above the fluorescent threshold for a wide size range. However, there is a clear cluster of particles well above the fluorescent threshold for one particular file with a much narrower size range around 6 microns. A similar pattern is apparent in one of the Johnson-grass smut samples. The size range of the Bermuda grass sample is similar to the sizes presented in Healy et al. (2012) whereas the size range of the Johnson grass is substantially smaller. In both the current study and Healy et al. (2012) the size ranges are smaller than the dimensions quoted in a microscopy study on fungal smuts (Crotzer and Levetin, 1996). So it does seem possible that these samples will contain at least some fragments. The puff ball spores have been studied previously using a fluorescence particle counter where the size range was stated to be between 2-4μm for the particles that they believed to be puffball spores. Two of the puffball sample files produced in the current study had average particle size of 2.50 ± 0.85 and 2.45 ± 1.16, but only 35 and 16 measurements from these files exceeded a fluorescent threshold of three standard deviations above the average forced trigger measurement. Whereas the other puffball sample collected had a size range of $3.39 \pm 1.76\mu$m with 506 particles exceeding the 3σ threshold.

We recognise that sizes of the pollen collected may be different to those in atmospheric particles. However, the data collected should be representative, at least to some extent, of what would be collected using the instrument during an ambient campaign. Conclusions from this data should provide an enhancement of conclusions stated in Crawford et al. (2015) where PSLs alone have been used to inform our analysis approach.

Since there is a large amount of historic data, and measurements are still being collected using the same instrument, we believe the findings presented will be valuable at this point in time. Nonetheless, as newer instruments are developed for example the WIBS NEO from Droplet Measurement Technologies, particles will be collected over a larger size range which will likely include a larger quantity of intact pollen. We expect such measurements will be more representative of an atmospheric environment and the current study should also be somewhat helpful as starting point for future research using instrumentation that has been more recently developed.

AP7 Added a table in the appendix section comparing the average size of the particles presented with other studies.

AP8 Add error bars to size in Figures 4 and 5.

2.1.5 Add fluorescence tables for results

"5. Tables showing the same charts as in Figs. 4 and 5, but for the particles which were classified, should be shown for the cases on which the conclusions are based."

We thank the reviewer for this suggestion which will improve the paper. We have added tables of the average measurements of the particles classified to the appendix.

AP9 Added the tables suggested.

2.1.6 What happened to k-means?

"6. K-means is mentioned in the abstract, introduction, Section 2.4 and Fig. 1. But are any results shown? I'm not seeing any mention of k-means after section 2.4."

The results for K-means were generally very poor. We have added a sentence to the main text to indicate that this was the case and add further details to supplementary material.

AP10 Added sentence to describe the poor performance of k-means to the text and add details to the supplementary material.

2.2 Other Issues

2.2.1 Why is fraction of particles not used as a criteria for good performance

"1. There appear to be over 80, 000 lab-generated particles in the 2008 dataset and over 20,000 in the 2014 dataset. Why is the fraction-of-particles-classified not part of the criteria for best and worst cases? Is a capability to classify more particles a desired feature in studying atmospheric aerosol? It seems odd that 3/4 to 4/5 of lab-generated test particles are not matched."

There seems to be some level of instrument noise present in the samples that should be removed by the threshold imposed. Such measurements are definitely worth removing. The only algorithm which removes any additional particles not already removed by the fluorescent and size threshold is DBSCAN. But whether this is an advantage or disadvantage is rather subjective. None of the algorithms worked perfectly on the data tested. If it is the case that a particle cannot be correctly classified, whether it is better to classify the particle incorrectly or not classify it at all is debatable.

In any case, in similar studies, (e.g. Hernandez et al., 2016), a similarly large number of particles from a sample do not fluoresce in any of the channels and are removed. For example 19786 particles are collected for the *Bacillus subtilis* sample with only 100 particles being fluorescent in at least one channel.

2.2.2 Error bars or some indication of data variation are needed in Figs. 4 and 5

This issue has also been raised by Darrell Baumgardner and is addressed in our response to this review.

2.2.3 Why not combine the 2008 and 2014 datasets?

Why not combine the 2008 and 2014 datasets? Combining would help with the generality of the study and may help make it more realistic and applicable to ambient aerosol. The inorganic samples in 2008 are very different from those in 2014. And there are different pollens (except mulberry) in these two years. The WIBS instruments used here appear to have different sensitivities for the detectors, different filters (or something else?). But three sample (the two bacteria and mulberry pollen) are in both datasets, and so using the ratios of the measured fluorescences and assuming linearity it should be possible to find multiplication factors for the FL. If it is not possible to combine these datasets, an explanation of why it is not feasible should be presented.

We would agree that combining the data sets would perhaps be valuable. But on closer inspection of the data we see that the paper mulberry samples collected had different size ranges across the two years. It therefore would be quite difficult to combine the data sets as suggested. There is also the possibility that average forced trigger fluorescence measurements could be subtracted from each of the sample measurements in an attempt to combine the files. However, investigating whether data could be combined in such a fashion, would be more appropriate once further data is collected alongside measurements using other techniques such as microscopy, whereby we could be more certain of what particles are being measured by the instrument.

Furthermore, one of the findings from the study is that the conclusions that one makes as to how to prepare the data for HAC is dependent upon what data is used to make these conclusions. So keeping the data sets separate, may be beneficial in highlighting the importance of repeating experiments.

AP1 Update text to describe why the data sets have not been combined at this point

2.2.4 Define FL1, FL2 and FL3

"4. FL1, FL2 and FL3 are not defined, and yet they are shown in Figs. 4 and 5. They are important for understanding the data analysed here. These should be defined, for example in section 2.1 where the "four fluorescent measurements" are described."

The FL1, FL2 and FL3 notation has been used previously in other studies (e.g Healy et al., 2012), but may cause confusion between the current study and the notation used in Kaye et al. (2005) so has been replaced with $FL1\_280$, $FL2\_280$ and $FL2\_370$ respectively.

AP2 Updated text and figures to include alternative notation and define these in the data section.

2.2.5 Justify fully why you have omitted FL4.

"The justification for omitting FL4, i.e., that some particles saturate, is inadequate"

Consulting Kaye et al. (2005), the following description should be more appropriate.

"The particle is irradiated with UV light at 280nm and 370nm from the firing of two xenon sources. Fluorescence emission is collected via two collection channels in the ranges $310 - 400$nm and $420 - 600$nm. The 370nm xenon radiation lies within the first detection range and hence elastically scattered light from the particle, sufficient to saturate the detection amplifier, is received. This saturated signal is therefore discarded. "

AP3 Added this description to the text.

2.2.6 "Reference Particles"

"6. Abstract, line 14-16: "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%." This is unclear. Aren't all the particles studied here reference particles, e.g., mulberry pollen, E. coli. Even the smoke from the burning grass is a reference aerosol. I guess reference particle means PSL. How about "reference narrow-size distribution PSL particles" for clarity"

We thank the reviewer for this suggestion and agree that the wording proposed is clearer so have used this instead.

AP4 Add the suggestion

2.2.7 Describe Adjusted Rand Score

"p. 12 line 5: "The adjusted rand score is often quite difficult to interpret ..." That sounds correct. It is not defined in this paper. Even after looking it up, it is not clear what exactly is being done in this paper, especially when there are n categories and m clusters. A little more explanation is needed."

See response to 10.

2.2.8 Clarify drawbacks

"8. p. 16, line 10: "It is clear that Hierarchical Agglomerative Clustering certainly has it drawbacks." Almost everything has its drawbacks. But this paper does not demonstrate or clarify drawbacks for HAC, as far as I can understand."

We have added a table to the text to make this clearer to the reader, including potential advantages and disadvantages of the other algorithms which may not be apparent in the current submission.

2.2.9 Define the matching matrix

"9. How about defining the matching matrix as used here. What is the criterion of the match?"
See response to 10.

2.2.10 Add some references for ML

"10. The introduction cites general papers on aerosols and their importance, but the initial description of machine learning does not. How about a very few relevant citations in the initial ML descriptions."

We thank the reviewer for this suggestion. We agree that inclusion of these references alongside a description of the adjusted rand score and the matching matrices would be useful.

AP5 Add references suggested and additional information to the methods section.

2.2.11 What is fluorescent in I & J?

"What is fluorescing in the NaCl and NaCl+phosphate samples I and J in Fig. 5? Do pure samples of these fluoresce enough to give the values shown?"

The plot currently presented is of the fluorescence measurements after a threshold has been applied. In the case of the NaCl and NaCl+phosphate samples these measurements would be for 3 and 61 particles respectively. These particles are likely to be low level contamination. On reflection we realise that presenting this information is of little value to the readership so we have removed it. In addition we realise, especially in the case of the gradient boosting algorithm it would not be reasonable to train on this class and as such this has been removed from the analysis.

For the data collected in 2008, the diesel soot and grass smoke samples could be expected to fluoresce and it may be beneficial to be able to discriminate between the fluorescent particles within this sample and the remainder of the data so these particles remain in the analysis.

AP6 Remove salt samples from plot.

AP7 Rerun analysis using gradient boosting without the salt samples.

3 Response to Darrell Baumgardner

We would like to thank Darrell for taking the time to review our manuscript. Many of the comments provided highlighted some issues that we had not previously considered and will aid in improving the paper to publication standard. We would like to apologise for initially not including a citation for the Hernandez study in the submission. This choice was made in an attempt to keep the message of the paper succinct, but upon reflection we realise that its inclusion would aid improving the paper, not only in contextualising the findings but to compare and contrast results with previous studies.

3.1 Incorrect labelling of Tables

1) Table I and II are both labelled 2008

We thank you for bringing this incorrect labelling to our attention. This had been rectified during the technical review stage, but we believe you may be reading the previous version of the manuscript prior to the technical review. Nonetheless, it should be 2008 for Table I and 2014 for Table II and we will ensure that this labelling is correct in the revised submission.

3.2 Need to summarise variation in the data

"2) If Tables I and II are actually 2008 and 2014, then there needs to be a third table that summarizes the properties that are shown in Figs. 4&5, not only the averages but their variances as well. These need to be listed for both years for the same test particles because from an examination of the figures, it certainly appears that the properties are quite different for the same biotypes. If this is indeed the case, then it is no surprise that there are different results using the various different clustering methods for the two data sets."

The variation of the data does need to be explored in more detail. There are differences in the fluorescent properties that can be explained by differences in sample preparation. In particular, for the bacteria there are unwashed and washed samples, diluted and undiluted samples and some samples of vegetative cells. As a result we believe that Figures 4 & 5 do need re-plotting to segregate further by sample. We have also produced scatter plots of fluorescence against size in each of the three channels. We are considering providing some of these plots in the main text with a link to similar plots for the remaining samples. In addition, we have included the table you have suggested alongside ABC counts, similar to the table presented in the appendix of Hernandez et al. (2016), which should aid in comparing the studies.

You are right that a potential explanation of the algorithmic performance for the two data sets could be that there are differences in the fluorescence properties in each case. In addition, the different thresholds used would result in a difference proportion of the different samples being present in the data set tested which could also affect the performance. This consideration has been added to the main text.

However, it should be at least a slight concern that the performance of the unsupervised techniques seems to be dependent on what data they were tested on, as we would hope that HAC would be adaptable to a variety of different situations. Gradient boosting, the supervised technique tested, did provide a smaller classification error across all of the tests and did seem to perform well consistently across the variety of tests conducted, so long as a fluorescent threshold of either 3 or 9 standard deviations was applied.

AP1 Produce the table requested.

AP2 Update Figures 4 and 5.

3.3 Why not use the biomarkers suggested in Hernandez?

"3) Why are just FL1, 2 and 3 used. In the Hernandez study (not cited here, unfortunately), we found that FL1 & 2 and FL 1 & 2 & 3 are important markers. Leaving them out seems like a loss of useful information.

We are using the raw fluorescent measurements when conducting our analysis, so information would not be lost in this way. When a single decision tree is fit to the data all possible splits are considered, including the splits using the thresholds defined in the Hernandez study, and as such the performance of a single decision tree cannot be worse than the approach suggested in the Hernandez study. So when using decision and ensembles of decision trees information will not be lost as suggested.

In the case of the unsupervised algorithms it is indeed an interesting idea to see whether better performance could be attained by clustering the biomarkers indicated in the Hernandez study rather than the raw data. However at this point we would be informing our analysis using laboratory data, so arguably the analysis would cease to be unsupervised.

3.4 Why not use the variance?

"4) Bioaerosols are by their nature irregular in shape and in their fluorescing. Why isn't the variance also used as a parameter in the clustering"

We believe you are referring to either to the variance of each broad class or the variance of the samples. The variance could be used in the clustering of laboratory data, but during an ambient campaign we would not know the classification of each sample so the variance of each group or specific particle type could not be explicitly calculated.

3.5 More work?

"I think additional work remains to further separate by general categories within bacteria, fungi and pollen. In our analysis of the lab results we were able to quite clearly separate the bio types just by fluorescence and size without any sophisticated machine learning. I would assume that this can be improved upon using more sophisticated approaches like the current study."

To construct our response we have use two papers (Hernandez et al., 2016; Calvo et al., 2018). For the benefit of other readers we briefly describe these papers. First, a particle may be described as A, B or C if they exceed the fluorescent threshold in the first, second or third fluorescent channels respectively. These ABC labels can then be combined to make groups of A, B, AB, C, AB, AC and ABC. For example, "AB" would describe particles that exceeded the threshold in the first and second channel.

The grouping of the data in such a fashion, referred to as "ABC analysis",was first introduced in Hernandez et al. (2016) and has since been applied to ambient data in Calvo et al. (2018). In Calvo et al. (2018) more detail is provided on how these ABC counts could be combined with the equivalent optical diameter (EOD) to provide a classification of WIBS data e.g. "Type I: Having the characteristics of the library bacteria (category A or AB, EOD $< 1.5\mu$m)."

From Figure 3 in Hernandez et al. (2016) the average measurements of each of the samples can clearly be divided. However, such a plot does not include the variation of the data. If one further investigates Table A1 in Hernandez et al. (2016), there are some particles that are fungi for example that have fluorescence type ABC that may be incorrectly classified using the classification scheme suggested in Calvo et al. (2018), if they are larger than 2μm,. In addition, we believe if the size of the bacteria is log normally distributed with the mean and standard deviation presented in Hernandez et al. (2016) that some of the particles will exceed 1.5 microns which is the threshold set in Calvo et al. (2018) for the bacteria.

Without having access to the full data set from this study it is difficult to determine precisely what proportion of the data will be incorrectly classified using such an approach, to directly compare with the techniques tested in this study. However, we do see the value in the application of the approach suggested in Hernandez et al. (2016), and results using ABC analysis for the data presented has been conducted and added to the paper.

AP3 Ran ABC analysis on data and appended results to manuscript

References

Calvo, A., Baumgardner, D., Castro, A., Fernández-González, D., Vega-Maray, A., Valencia-Barrera, R., Oduber, F., Blanco-Alegre, C., and Fraile, R. (2018). Daily behavior of urban fluorescing aerosol particles in northwest spain. *Atmospheric Environment*, 184:262–277.

Crawford, I., Ruske, S., Topping, D., and Gallagher, M. (2015). Evaluation of hierarchical agglomerative cluster analysis methods for discrimination of primary biological aerosol. *Atmospheric Measurement Techniques*, 8(11):4979–4991.

Crotzer, V. and Levetin, E. (1996). The aerobiological significance of smut spores in tulsa, oklahoma. *Aerobiologia*, 12(1):177–184.

Gillbro, T. and Cogdell, R. J. (1989). Carotenoid fluorescence. *Chemical Physics Letters*, 158(3-4):312–316.

Harrison, R. M., Jones, A. M., Biggins, P. D., Pomeroy, N., Cox, C. S., Kidd, S. P., Hobman, J. L., Brown, N. L., and Beswick, A. (2005). Climate factors influencing bacterial count in background air samples. *International journal of biometeorology*, 49(3):167–178.

Healy, D. A., O'Connor, D. J., Burke, A. M., and Sodeau, J. R. (2012). 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.

Hernandez, M., Perring, A. E., McCabe, K., Kok, G., Granger, G., and Baumgardner, D. (2016). Chamber catalogues of optical and fluorescent signatures distinguish bioaerosol classes. *Atmospheric Measurement Techniques*, 9(7).

Kang, D.-Y., Son, M.-S., Eum, C.-H., Kim, W.-S., and Lee, S.-H. (2007). Size determination of pollens using gravitational and sedimentation field-flow fractionation. *Bulletin of the Korean Chemical Society*, 28(4):613–618.

Kaye, P. H., Stanley, W., Hirst, E., Foot, E., Baxter, K., and Barrington, S. (2005). Single particle multichannel bio-aerosol fluorescence sensor. *Optics express*, 13(10):3583–3593.

Khaneja, R., Perez-Fons, L., Fakhry, S., Baccigalupi, L., Steiger, S., To, E., Sandmann, G., Dong, T., Ricca, E., Fraser, P., et al. (2010). Carotenoids found in bacillus. *Journal of applied microbiology*, 108(6):1889–1902.

Pöhlker, C., Huffman, J., and Pöschl, U. (2012). Autofluorescence of atmospheric bioaerosols–fluorescent biomolecules and potential interferences. *Atmospheric Measurement Techniques*, 5(1):37–71.

Pöhlker, C., Huffman, J. A., Förster, J.-D., and Pöschl, U. (2013). Autofluorescence of atmospheric bioaerosols: spectral fingerprints and taxonomic trends of pollen. *Atmospheric Measurement Techniques*, 6(12):3369–3392.

Savage, N. J., Krentz, C. E., Könemann, T., Han, T. T., Mainelis, G., Pöhlker, C., and Huffman, J. A. (2017). Systematic characterization and fluorescence threshold strategies for the wideband integrated bioaerosol sensor (wibs) using size-resolved biological and interfering particles. *Atmospheric Measurement Techniques*, 10(11):4279–4302.

Tong, Y. and Lighthart, B. (1997). Solar radiation is shown to select for pigmented bacteria in the ambient outdoor atmosphere. *Photochemistry and photobiology*, 65(1):103–106.

amtd-2018-126 changes

simon.ruske

October 2018

Anon. Reviewer Significant Edits		
AP1	Produced dendrogram plots alongside a heat map of the matching matrix comparing the 10 cluster solutions to the known labels.	Presented in Figures 11 and 12 in the revised manuscript
AP2	Rewritten the HAC results section to make clear what analysis has been conducted and more adequately describe Figure 6.	Significant editing of Section 4.1
AP3	Produced an additional section to indicate why the CH index has a tendency to conclude that there are 2 clusters	See Section 4.1.2
AP4	Split the HAC section into subsection highlighting potential considerations for data preparation, the CH index, and potential issues in the hierarchy that could not be rectified by the selection of a different index to determine the number of clusters	Split into 4.1.1- 4.1.3
AP5	Produce a table that highlihgts the potential advantages and disadvantages highlighted with the current study and how the number of clusters are determined in each case	See Table 4
AP6	Updated the description of the fluorophores in the introduction to include additional references	Second paragraph of introduction
AP7	Added a table in the appendix section comparing the average size of the particles presented with other studies	Table A1
AP8	Add error bars to size in Figures 4 and 5	Figures 4 and 5 have been segregated by sample and now form Figures 5-7 with error bars included
AP9	Add the tables suggested	See Appendix C1-C4
AP10	Added sentence to describe poor performance of k-means to the text and add details to the supplementary material	The sentence has been added to Section 4.4 in the revised manuscript. Except we have elected to place the results for k-means with the rest of the additional material in the code repository rather than in the supplementary material.

Anon. Reviewer Other Issues		
AP1	Update the text to describe why the data sets have not been combined at this point	P7, L31-P8-L7
AP2	Updated text and figures to include alternative notation and define these in the data section	See Figures 5, 6, 7 and text P3, L56-62
AP3	Added this description of omitting FL4	P3, L51-55
AP4	Add this rewording of the PSL particles.	P1, L24
AP5	Add references suggested and additional information to methods section.	P2, 37-38 Includes reference to friedman et al., 2001 which is a good text on generic machine learning. Section 2.7 now includes a description of matching matrices and the adjusted rand score. Section 2.6 includes four additional references for Bagging, Random Forests, AdaBoost and Gradient Boosting which are mention in this section.

Darrell Baumgardner		
AP1	Produce the table requested	See Appendix C1-C4
AP2	Update Figure 4 and 5	Figures 4 and 5 have been segregated by sample and now form Figures 5-7 with error bars included
AP3	Ran ABC analysis on data and append results to manuscript	See Tables B1, 2 & 3.

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: 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~~ respond differently in the presence of ultraviolet light~~which potentially could be used to discriminate between~~ , potentially allowing for different types of biological aerosol to be discriminated. Development of ultraviolet light induced fluorescence (UV-LIF) instruments such as the Wideband Integrated Bioaerosol Sensor (WIBS) has ~~made is possible to collect~~ allowed for size, morphology and fluorescence measurements to be collected in real-time. However, it is unclear without studying ~~responses from the instrument~~ instrument responses in the laboratory, the extent to which ~~we can discriminate between~~ different types of particles can be discriminated. Collection of laboratory data is vital to validate any approach used to analyse ~~the data and to~~ data and 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~~ are tested on a range of ~~laboratory generated aerosols~~particles collected in the laboratory. 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~~ reference narrow-size distribution PSL 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 to attain consistent results across the different sets of laboratory generated aerosol tested.

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

~~Where~~ In the event that laboratory training data ~~is~~ in unavailable, DBSCAN was found to be ~~an~~ a 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 appeared to be heavily dependent on data preparation, requiring a different selection of parameters ~~dependent~~ depending on the preparation used. Further analysis will also be required to confirm our selection of the 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 to improve on current discrimination between pollen, bacteria and fungal spores and even between ~~their different types~~different species, 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. The~~ (Crawford et al., 2012; Cziczo et al., 2013; Gurian-Sherman and Lindow, 1993; Hader et al., 2014; Hoose and Möhler, 2012; Möhler . These 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~~

There are a wide range of biological molecules, commonly referred to as biological fluorophores, that are known to re-emit radiation upon excitation e.g. amino acids, coenzymes and pigments (Pöhlker et al., 2012, 2013). Ultraviolet-light induced fluorescence (UV-LIF) ~~spectrometer~~ spectrometers, such as the ~~Wideband Integrated Bioaerosol Spectrometer~~ 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~~ have received increased attention in recent years as a potential methodology for detecting biological aerosol (Kaye et al., 2005). The WIBS uses irradiation at 280nm and 370nm to target some of the most significantly fluorescent bioflorophores such as tryptophan (an amino acid) and NADH (a coenzyme). These measurements are combined with an optical measurement of size and shape ~~for each particle is taken~~ to further aid in discrimination.

~~These measurements~~ Measurements from the WIBS 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~~ there are a range of techniques that could be used to predict quantities of biological aerosol from these fluorescence, size and morphology measurements. Techniques that could be used to solve this classification problem, ~~and they can be divided broadly into two~~

groups: include field specific techniques such as ABC analysis (Hernandez et al., 2016) as well as supervised and unsupervised machine learning techniques that are broadly used (Friedman et al., 2001).

It is not clear whether the supervised or unsupervised approach is to be preferred as both approaches have their at this point what approach is preferred as all approaches have a range of advantages and disadvantages.

Supervised machine learning uses data , usually collected within laboratory settingscollected within the laboratory, where the correct classification is known. Subsequently, this data Data is split into training data and testing data . The where the training data is used to fit a model which is then validated using on the test set. Once the a model is fitted and validated it may then be applied to classify ambient data.

During unsupervised analysis, ambient data is classified without using training data from the laboratorylaboratory training data. Instead, an attempt is made to split the datainto groups using natural differences in the data. Ideallythe data would be naturally split naturally segregate the data. Ideally, we may expect data to naturally be segregated into broad biological classes , or into different groups of similar bacteria, fungal material and pollen, 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 might be collected during an ambient campaign. Particularly, in an urban environment, the instrument will collect may collect measurements for a large quantity of non-biological material that will still need to be either should be classified as such or removed from the analysis. 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 There are likely to be issues to be explored with either approach , and therefore it seems unlikely that we will be able to abandon either supervised or unsupervised techniques can justifiably be abandoned at this point in time . and it may well be the case that usage of a variety of techniques may be required to better understand the atmospheric environment. Nonetheless, it is still vital to investigate how these different techniques behave when analysing laboratory data to better understand how they can be most appropriately applied to ambient data.

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 relatively few studies have studied the usage of HAC on laboratory generated aerosoldata from the WIBS (Savage et al., 2017; Savage and Huffman, 2018) . Evaluating the effective-

ness 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~~ Such time requirements mean that not only is the method already quite slow, but will get increasingly slower as more data is collected~~. This~~ , which 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~~ Sci-kit learn 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~~ sci-kit learn implementation of DBSCAN by default is not memory efficient making it difficult to utilise for more than $30,000$ particles. ~~Sklearn~~ Sci-kit learn 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.

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

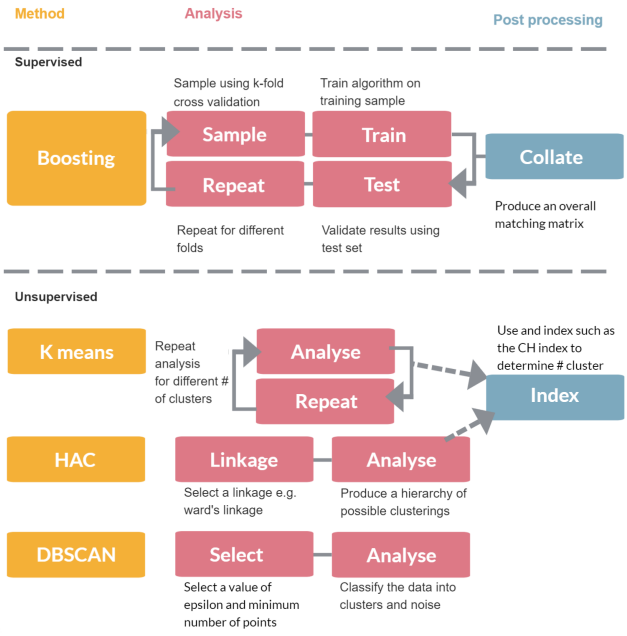

Figure 1. Overview of different analysis approaches

combined to produce a single asymmetry factor measurement. ~~Four fluorescencemeasurements are collected by firing a flash lamp~~ A more precise definition of asymmetry factor has been provided previously in the literature (Gabey et al., 2010).

To measure fluorescence, the particle is irradiated with UV light 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.~~ from the firing of two xenon sources. Fluorescence emission is collected via two collection channels in the ranges $310-400$nm and $420-600$nm. The 370nm xenon radiation lies within the first detection range and hence elastically scattered light from the particle, sufficient to saturate the detection amplifier, is received. This signal is therefore discarded.

After removal of this fluorescent measurement, there are three remaining fluorescence measurements~~which are~~. The notation FL1_280 is used to denote the measurement in the first detection channel when the particle is irradiated with ultraviolet light at 280nm and FL2_280 and FL2_370 are used to denote the measurements in the second detection channel when the particle is irradiated with ultraviolet light at 280 and 370nm respectively. These fluorescence measurements 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)~~ (e.g. Gabey et al., 2010; Healy et al., 2012a)

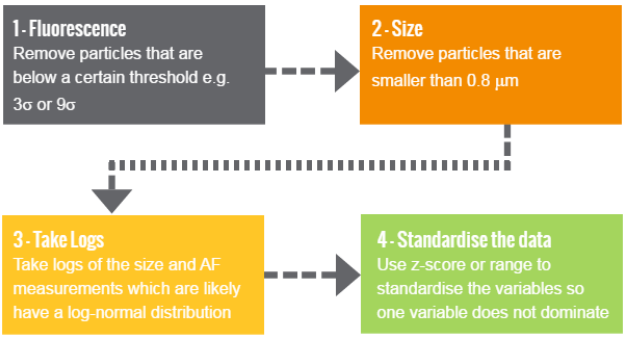

Figure 2. Overview of preprocessing steps for WIBS data

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.

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~~ as is the case in random forests and bagging classifiers (Breiman, 2001, 1996). In the case of random forests and bagging the data set is sampled with

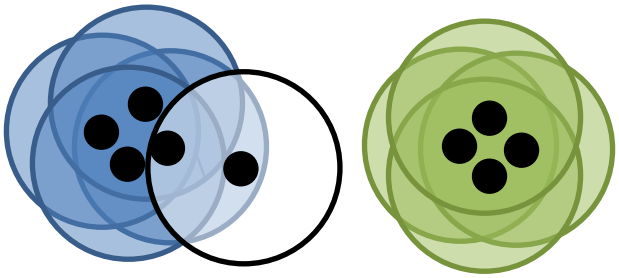

Figure 3. Visual representation of DBSCAN. Here each point is represented as a black dot and it's neighbourhood is represented by a circle. Here ϵ is the radius of the circle and the minimum number of points is 3. Four points have each been placed into the blue cluster and green cluster, all of which 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.

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, when using a random forest, 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 (Freund and Schapire, 1997). Gradient Boosting is an extension of AdaBoost to allow for other loss functions (Friedman, 2001) .

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)

2.7 Evaluation Criteria

To aid in evaluating how well methodologies performed we used two tools: the matching matrix (Ting, 2010) and the adjusted rand score (Hubert and Arabie, 1985).

3 ~~Data~~

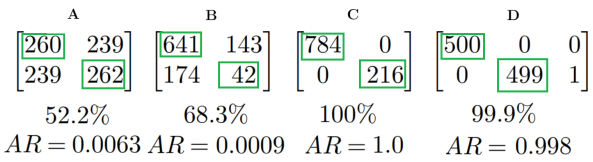

Figure 4. Four example matching matrices. Immediately below each matrix is the percentage of particles placed into the same cluster for both clustingers in each case. At the very bottom we have the adjusted rand score.

In Figure 4 we present four different matching matrices. To produce these matrices we compared: two random clusterings with approximately 50% of the data in each cluster (A); two random clusterings each with 80% and 20% of the data in each of the two clusters respectively (B); two identical clusterings (C); and two clusterings which were nearly identical except one data point had been placed into a third cluster for one of the clusterings.

2.0.1 Matching Matrix

The matching matrix, often referred to as a confusion matrix, can be used as an aid in comparing two clusterings.

In the case of the current manuscript, we use this to compare the output from an algorithm with labels assigned to each particle. We may assign labels to indicate what broad type the particle is (e.g. 1 if the particle is bacteria, 2 if the particle is fungal etc.) or we may assign labels to indicate what sample a particle is from (e.g. 1 if the particle is *Bacillus atrophaeus*, 2 if the particle is *E. coli* etc.)

Consider example C in Figure 4. This matching matrix compares two clusterings each containing two clusters. Each row corresponds to a cluster in the first clustering and each column corresponds to a cluster in the second clustering. The element in the first row and the first column (in this case 784) indicates the number of particles that were placed into the first cluster in the first clustering that were also placed into cluster 1 in the second clustering. Two identical clusterings will produce a matching matrix that has non-zero values only the diagonal.

A and B in Figure 4 are examples of poor performance and C and D are examples of very good performance.

2.0.2 Adjusted Rand Score

When evaluating a large number of clusterings, it may be useful to use a statistic to summarise the information in the matching matrix. In a previous study (Ruske et al., 2017), we used percentage of particles correctly classified as a statistic for indicating performance. This is an easy to interpret statistic, but can be misleading when used on imbalanced data. In both example A and B, we have two randomly generated clusterings. However in B we have 80% of the data points placed into the first cluster, whereas in A the data points are approximately equally distributed between the two clusters. The percentage of points which are placed into the same cluster for both clusterings are 52.2% and 68.3% for A and B respectively. We can see that the more imbalanced a data set is, the more likely data points are to be placed into the same clusters. It is for this reason we elect to use an alternative statistic: the adjusted rand score. This statistic attains a value of approximately zero for both A and B.

Comparing clusterings is a developing area of research and there are other alternative statistics such as the mutual information score (Vinh et al., 2010) that could be preferable to the adjusted rand score. However our initial tests (not presented), indicated that calculation of the mutual information often required an order of magnitude more time than the calculation of the adjusted rand. Therefore we elected to use the adjusted rand score for the current study.

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~~ data sets collected in 2008 and 2014 which are similar to data previously published using the Multiparameter Bioaerosol Spectrometer (Ruske et al., 2017). A subsection of the data collected 2014 has previously been analysed in the appendix of (Crawford 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). The bacterial samples were either washed or unwashed and diluted or undiluted.

We present a summary of the number of particles for each sample ~~in total as well as when using~~ after a fluorescent threshold of 3σ and 9σ ~~in Tables ?? and ??.~~ is applied in Tables 1, 2 and 3. In 2008 the thresholds are constructed using forced trigger data collected at the same time as the experiment, whereas in 2014 the thresholds are constructed using forced trigger data collected using the same instrument at an earlier date. Ideally, the threshold for the data collected in 2014 would be constructed using forced trigger data collected at the same time as the laboratory data, but we can see in Figure 8 that the threshold we have constructed is successful in removing the vast majority of NaCl samples collected.

Plots of the average fluorescent characteristics and size and shape for each sample are provided in Figures ~~?? and 7 .~~ 5, 6 and 7 after a fluorescent baseline of 3σ has been applied. Similar plots have been produced using a 9σ threshold and can be found in the repository released alongside the manuscript (see the code/data availability section for further details). Plots and tables for the polystyrene spheres previously published in Crawford et al. (2015) are omitted.

Table 1. ~~Counts~~ The number of ~~different aerosols collected in 2008 before and~~ particles remaining after a fluorescent threshold ~~is~~ of 3σ or 9σ was applied for each of the bacterial samples collected in 2008. Each sample was either washed or unwashed and diluted or undiluted. Washed samples are denoted by a check mark in the column "W" and diluted samples are mark in the column 'Dil.'.

ID	Sample	~~#~~ W	~~# (3σ)~~ Dil.	~~# (9σ)~~ $n > 3\sigma$	~~Classification~~ $n > 9\sigma$
A	*Bacillus atrophaeus* Spores	~~30946~~	~~12631~~	~~3239~~ 952	~~bacteria~~ 34
B	~~*E.coli*~~	~~15237~~	~~8332~~ ✓	~~3681~~ 52	~~bacteria~~ 4
C	~~Bermuda grass smut~~	~~5220~~ ✓	~~2681~~	~~423~~ 1171	~~fungal~~ 217
D	~~Johnson grass smut~~	~~7248~~ ✓	~~3882~~ ✓	~~637~~ 241	~~fungal~~ 38
E	~~Paper mulberry~~ —"— Vegetative Cells	~~1030~~	~~630~~	~~312~~ 4779	~~pollen~~ 1915
F	~~Ragweed pollen~~	~~569~~	~~332~~ ✓	~~151~~ 1488	~~pollen~~ 264
G		✓		1884	573
H		✓	✓	2064	194
I	*E coli.*			3684	1547
J			✓	1448	371
K		✓		2365	1461
L		✓	✓	835	302

Table 2. The number of particles remaining after a fluorescent threshold was applied for each of the non-bacterial samples collected in 2008.

ID	Sample	Category	$n > 3\sigma$	$n > 9\sigma$
M	Bermuda grass	Fungal	2681	423
N	Johnson grass I	Fungal	1209	259
O	Johnson grass II	Fungal	2673	378
P	Birch pollen	~~164~~ Pollen	111	56
Q	~~pollen~~ Paper mulberry I	Pollen	233	209
~~H~~ R	~~Grass smoke~~ Paper mulberry II	~~14457~~ Pollen	~~3357~~ 397	~~299~~ 103
S	~~interferent~~ Ragweed I	Pollen	123	34
T	Ragweed II	Pollen	209	117
~~I~~ U	Diesel smoke	~~7900~~ Interferent	11	5
V	~~interferent~~ Grass smoke I	Interferent	2542	231
W	Grass smoke II	Interferent	815	68

~~For most of the interferent particles~~To provide further clarity on the variation of the samples in terms of size and fluorescence we include scatter plots of each of the fluorescence channels against size for four of the samples in Figure 8. For the puffball and

Table 3. ~~Counts~~ The number of ~~different aerosols collected in 2014 before and~~ particles remaining after a fluorescent threshold ~~is~~ was applied to each of the samples collected in 2014. Whether a bacterial sample was washed (w) or unwashed (unw) is specified after the sample name.

ID	Sample	~~#~~ Category	~~#(3σ)~~ $n > 3\sigma$	~~#(9σ~~ $n > 9\sigma$
A	*Bacillus atrophaeus* (unw)	~~Classification~~ Bacteria	1728	684
~~A~~ B	*Bacillus atrophaeus* (w)	~~6217~~ Bacteria	~~3050~~ 1322	~~1292 bacteria~~ 608
~~B~~ C	*E. coli* (unw)	~~2534~~ Bacteria	1290	632 ~~bacteria~~
~~C~~ D	~~Puffballs~~ Puffball I	~~3919~~ Fungal	~~555~~ 504	~~252~~ 248
E	~~fungal~~ Puffball II	Fungal	35	3
F	Puffball III	Fungal	16	1
~~D~~ G	Aspen ~~pollen~~ Pollen	~~398~~ Pollen	74	31 ~~pollen~~
~~E~~ H	~~Poplar~~ Paper mulberry pollen	~~375~~ Pollen	~~104~~ 541	~~50 pollen~~ 537
~~F~~ I	~~Paper-Mulberry~~ Poplar Pollen	~~565~~ Pollen	~~541~~ 104	~~537 pollen~~ 50
~~G~~ J	Ryegrass pollen	~~47~~ Pollen	21	15 ~~pollen~~
~~H~~ K	Fullers ~~'~~ Earth	~~3226~~ Interferent	~~35~~ 61	~~3 interferent~~ 20
~~I~~ L	NaCl	~~2197~~ Interferent	3	0 ~~interferent~~
~~J~~ M	Phosphate Buffered Saline	~~3064~~ Interferent	~~61~~ 35	~~20 interferent~~ 3

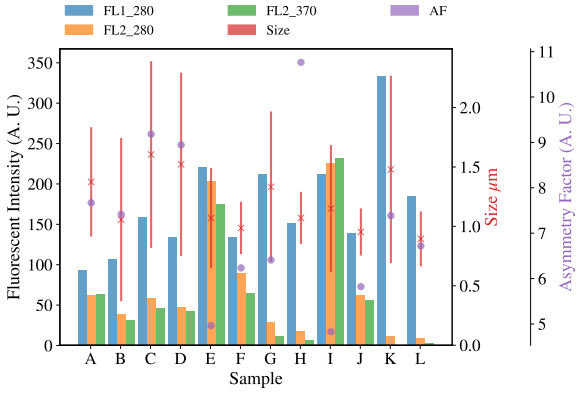

Figure 5. Average fluorescent characteristics for the ~~different aerosol~~ bacterial samples collected in ~~2008~~2008. The error bars in red indicate a range of $\pm 1\sigma$ for each sample.

rye grass samples, in particular we can see that we may be measuring both fragmented and intact particles. For the interferent samples we see that a ~~fluorescent threshold of either~~ threshold of 3σ ~~or 9σ will remove~~ removes 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 grass smoke~~

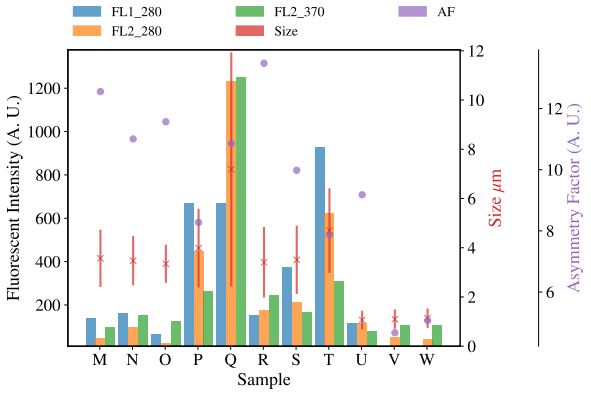

Figure 6. Average fluorescent characteristics for the ~~different aerosol~~ remaining samples collected in ~~2014~~2008. The error bars in red indicate a range of $\pm 1\sigma$ for each sample.

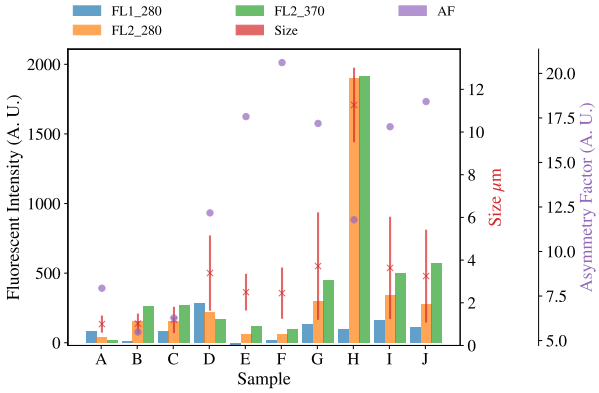

Figure 7. Average fluorescent characteristics for the different aerosol samples collected in 2014. The error bars in red indicate a range of $\pm 1\sigma$ for each sample.

~~samples even using a fluorescent thresholdof 9σ, providing an example of an interferent that does fluoresce in the instrument.~~ In fact the only interferent samples to measure a number of particles over a threshold of 3σ were the grass smoke samples.

The data collected is using a WIBS version 3 which is limited to a detection range of approximately 0.5μm-12μm, which limits the ability of the instrument to detect intact pollen grains. The vast majority of the equivalent optical diameters (EODs) for the pollen samples collected are much lower than the measurements for intact pollen grains and are therefore likely to be pollen fragments, as was the case in Hernandez et al. (2016). The exception is the paper mulberry samples where there are differences across each of the samples. In 2008, sample Q which shows a size range similar to the other pollen samples is most likely to consist entirely of pollen fragments, whereas sample R shows a much wider size range which is likely to comprise of both fragmented and intact pollen. The collection of both fragmented and intact pollen has previously been shown to occur in Savage et al. (2017). In 2014, for sample H, the size range is much larger, consistent with the hypothesis of measuring intact

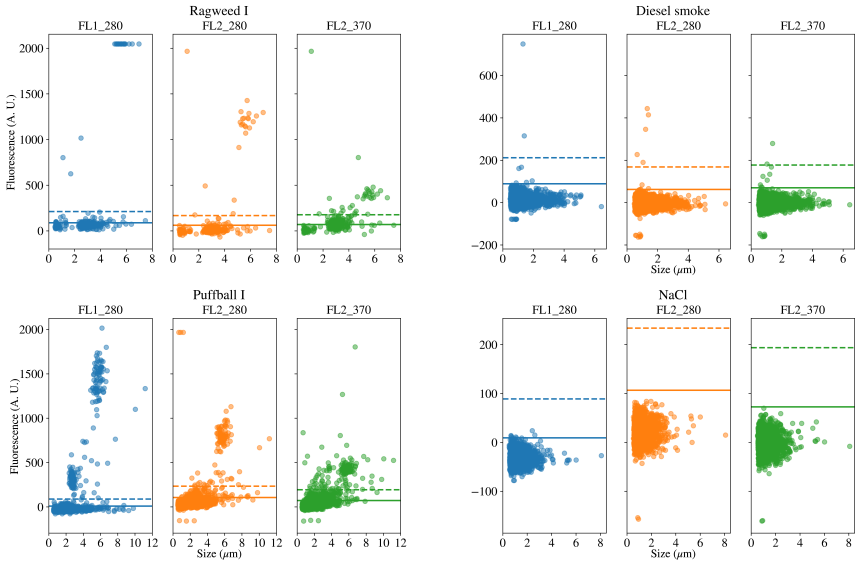

Figure 8. Scatter plots of fluorescence versus size for four of the samples. Two of the samples were collected in 2008 (top row) and two were collected in 2014 (bottom row); two are biological (left column) and two are non-biological (right column).

pollen. Paper mulberry has been previously been sampled in (Healy et al., 2012a), using a WIBS version 4 in a low-gain mode which allows for the collection of particles up to approximately 31μm. In this study, the size range of the paper mulberry was 13.6 ± 6.2, indicating that if sample H is intact pollen we may only be measuring part of the distribution.

It may have been possible to combine the data sets from 2008 and 2014. However, investigating if there are differences in conclusions when testing different methodologies using different laboratory samples could offer insight into the reproducibility of the research presented in the current study. We therefore elected to analyse the data sets separately and compare and contrast the findings when testing on the PSLs and each of the data sets collected in 2008 and 2014.

4 Results

In Sections 4.1, 4.2, 4.3 we present the results using HAC, DBSCAN and gradient boosting respectively. A summary of the findings for each method and an indication of how the number of clusters are determined are shown in Table 4.

4.1 Hierarchical Agglomerative Clustering

Prior to hierarchical agglomerative clustering (HAC) being applied, we labelled each particle from 1-4 to indicate whether the particle was bacteria, fungal, pollen or non-biological respectively. We then considered a variety of different approaches to prepare the data which are shown in Table 5. 96 possible combinations of these considerations were applied to the data and the hierarchical agglomerative clustering routine was used to cluster the resultant data in each case. For each of the ninety-six

Table 4. Summary of findings and considerations for method selection.

Method	Summary	# Clusters
HAC	– Does not rely on training data – The conclusion we make when using the CH index may be incorrect when a large proportion of the particles are from one broad class – How the data was prepared greatly impacted upon performance – Particles from different categories were sometimes clustered together e.g. pollen with fungal	Determined using the maximum value of the CH index produced for clusterings between 1 and 10 clusters.
DBSCAN	– Produced a clustering which contained three distinct clusters each containing primarily one broad class of bioaerosol in the case of one of the data sets – Data preparation greatly impacted upon performance – It is not clear at this point whether the values of epsilon and the minimum number of points would be applicable to ambient data	Naturally determined by setting epsilon and the minimum number of points required for a neighbourhood.
Gradient Boosting	– Performance was consistently good irregardless of data preparation provided that a threshold, either 3 or 9 standard deviations, was applied to the fluorescence – Relies on adequate training data being collected and it is not clear at this point whether the data collected will be sufficient.	Always the same as the number of groups in the training data

hierarchies produced, the clusterings containing between 1 and 10 clusters were extracted. Subsequently, a value of the adjusted rand score comparing each of these 10 clusterings to the known labels was calculated. These values of the adjusted rand score would be unavailable during an ambient campaign but are used here to measure the similarity of each clustering to the known labels in order to indicate overall performance and highlight which of the first 10 clusterings was most similar to the known

Table 5. 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

labels. Values of the Calinski-Harabasz index (CH index), an index which is usually used in an ambient campaign to determine the number of clusters, were also calculated. The number of clusters in the clustering for which the maximum value of the CH index was attained can then be compared to the clustering which is most similar to the known labels to determine if the CH index attains a maximum for the clustering which is most similar to the known labels.

~~Usually when using Hierarchical Agglomerative Clustering we use the following data preparation strategy:~~

4.1.1 Impact of data preparation

Figure 9 provides an overview of the results obtained using the 96 different strategies tested. The data preparation approach suggested in Crawford et al. (2015) (presented in blue) was to take logs of the size and ~~the~~ asymmetry factor, ~~apply~~ use a size threshold of ~~0.8 microns, apply~~ 0.8 microns, use a fluorescent threshold of ~~3 or more recently 9 standard deviations and~~ 3 standard deviations above the average forced-trigger measurement, 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 5.~~

~~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.~~

~~In Figure 9 we outline how well Hierarchical Agglomerative Clustering performed when using the standard strategy varying between 3σ and~~ and use Ward Linkage. It has also been suggested that a threshold of nine standard deviations may be more appropriate (Savage et al., 2017), so the approach suggested in Crawford et al. (2015) modified to use a threshold of 9σ ~~, and how well the algorithm worked with the best data preparation strategy across all 96 possible combinations of options for each data set.~~ is also presented (in orange)

~~First~~In the case of the PSL data set, we see that ~~the high performance achieved for the PSLs (AR =~~ HAC has produced a clustering with 5 clusters which is very similar to the known labels. The best performance occurred when using a fluorescent threshold of 9 standard deviations, albeit 3 standard deviations produced a similarly high value of the adjusted rand score (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~~.

The maximum adjusted rand scores attained for the the laboratory generate aerosol collected in 2008 and 2014 ~~respectively. This is~~ were 0.567 and 0.747. Lower scores are to be expected ~~as the~~ since we would anticipate laboratory generated aerosol ~~particles are much more complex , and therefore~~ to be more complex than polystyrene latex spheres and hence 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~~ discriminate. The adjusted rand score of the best data strategy of the 96 tested, as indicated by the height of the green bar, is larger than the corresponding adjusted rand score for the strategy suggested in Crawford et al. (2015), indicating that potentially a different strategy may yield better results. However, the best performing strategy was not consistent across both the 2008 and 2014 data.

In particular, the best strategy in 2008 was found to be: taking logs; using a size threshold of ~~0.8~~ 0.8 microns; using ~~3 standard deviations as a fluorescent threshold ;~~ 3 standard deviations and fluorescent threshold standardising using the range ~~;~~ and using Ward linkage. In 2014, the ~~best results were~~ highest value of the adjusted rand score was obtained by not taking logs, not applying a size threshold, using a fluorescent threshold of ~~9~~ 9 standard deviations and using the centroid linkage. Since our findings are inconsistent across the two laboratory generated aerosol data sets it becomes difficult to provide a better recommendation for data preparation other than the strategy suggested in Crawford et al. (2015).

In addition, there was a substantial difference between the quality of results attained ~~for~~ when using a fluorescent threshold of 3 ~~standard deviations vs.~~ or 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~~

It is possible that the difference in performance when using the different thresholds could be in part explained by the fluorescent threshold in 2014 being constructed using forced trigger data collected at a different time to the laboratory data, or by the fluorescence properties differing across the two data ~~preparation technique used, but the difference is inconsistent across different data~~ sets.

~~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~~But this differing behaviour when using different data preparation does need to be investigated further with additional laboratory data sets and in the context of ambient data. Nonetheless, the differing conclusions across the two data sets as to which data preparation is preferable does highlight the importance of repeating data collection and demonstrating conclusions are consistent across multiple experiments.

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 6 and 7. 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

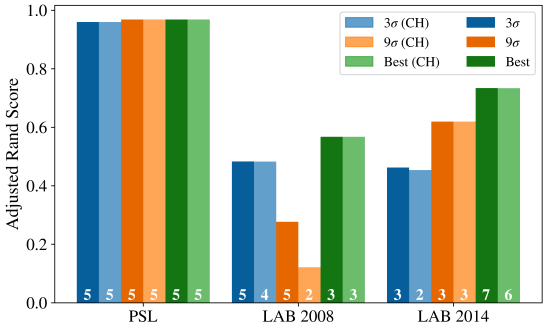

Figure 9. Performance of Hierarchical Agglomerative Clustering using the adjusted rand score for the data sets tested across different data preparation strategies. The number of clusters concluded in each case is indicated at the bottom of each bar.

conclude that algorithm has failed to differentiate between any of the biological classes, in part due to the CH index concluding there are 2 clusters.

~~From Figure 9~~

4.1.2 Impact of the Calinski-Harabasz Index

At the base of each bar in Figure 9 we provided the number of clusters in the clustering for which the adjusted rand score presented was obtained. For the darker bars, this number represents the number of clusters in the clustering for which the highest value of the adjusted rand score was obtained across the clusterings containing between 1 and 10 clusters. For the lighter bars, this number represents the number of the clusters in the clustering for which a maximum of the CH index was attained.

There are three different scenarios that occur. First, the Calinski-Harabasz index attains a maximum for the clustering which is most similar to the known labels e.g. for the PSL data using a fluorescent threshold of 3 standard deviations, the clustering which is most similar to the known labels (shown in the darker bar) contains 5 clusters which is the same as the clustering for which a maximum of the CH index is attained (shown in the lighter bar). Second, the Calinski-Harabasz index attains a maximum for a different clustering which is most similar to the known labels, but the conclusion does not have a large impact on performance. For example, in 2008 using a fluorescent threshold of 3 standard deviations, the clustering which is most similar to the known labels contains 5 clusters, whereas the clustering for which the CH index attains a maximum contains 4 clusters. However, the heights of bars are nearly the same. In this case, a very small cluster has been merged in the hierarchy from 5 to 4 clusters resulting in the 4 and 5 cluster clusterings being extremely similar and consequently the fact that the CH index has attained a maximum at 4 clusters instead of 5 is not concerning, since concluding 4 clusters instead of 5 has very little impact upon performance.

The final case is in 2008, using a fluorescent threshold of 9 standard deviations. Here the clustering which is most similar to the known labels is the clustering containing 5 clusters, whereas the CH index attains a maximum for the clustering containing only 2 clusters. The 2 cluster solution in this case is very dissimilar from the known labels.

In the cases where a maximum for the CH index was attained for a clustering containing 2 clusters i.e. in 2008 using 9σ and in 2014 using 3σ, ~~it is clear that data preparation strategy can have a substantial impact upon the quality of clustering results. From Tables 6 and 7 we demonstrate that for a particular data preparation approach the quality of~~ 78.6% and 76.5% of the particles were from a bacterial sample. Conversely in 2008 using 3σ and in 2014 using 9σ, 65.4% and 68.4% of the particles analysed were bacteria.

To investigate the possibility of a relationship between the proportion of the data which is contained in the category containing the largest number of particles and the tendency of the CH index to conclude that there are 2 clusters we produced data simulated from 3 normal distributions in 3 dimensions. Each of the clusters was centred around [0, 0, 0], [5, 5, 5], [10, 10, 10] and the co-variance matrix was set to σI_3, where I_3 is 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~~ 3 by 3 identity matrix. The value of σ was varied from 1-3 to produce a range of variation in the simulations. We elected to produce this simulated data from normal distributions rather than the laboratory data collected to remove any potential confounding issues such as the fluorescent threshold used. The proportion of the data that was contained in the dominant cluster was varied from 50% to 99%. Each simulation was repeated 100 times to provide an indication of the frequency the CH index attains a maximum for the 3 cluster solution.

In Figure 10 we see that there is a point where the frequency for which the CH index attains a maximum for the clustering contains 3 clusters starts to decrease. The proportion of data points that needs to be placed in the dominant cluster before this decrease in performance of the CH index is seen decreases as the variability in the data increases.

This incorrect conclusion when using the CH index when analysing data for which a large proportion of data is of one particular type is problematic when analysing biological aerosol, since we may expect the quantity of bacteria to be an order of magnitude greater than than the fungal spores, and for the quantity of fungal spores to be an order of magnitude greater than the pollen (Després et al., 2012; Gabey, 2011). In future studies it may therefore be necessary to explore the use of other indices for determining the number of clusters.

4.1.3 Breakdown of the hierarchies

To more clearly understand how data has been clustered using HAC we have presented dendrograms for the laboratory data collected in 2008 and ~~performance is repeatable.~~ 2014 in Figures 11 and 12 alongside heat maps of the matching matrices to indicate the cluster composition of the 10 cluster solution broken down by sample. The hierarchy produced using the strategy suggested in Crawford et al. (2015) is presented at the top of each plot whereas a modification of this strategy using a threshold of 9 standard deviations as suggested by Savage et al. (2017) is presented at the bottom.

Each row of the heat map corresponds to a particular cluster and each column corresponds to a particular sample. The intensity of each box corresponds to the quantity of particles placed into a particular cluster from a particular sample. Bacterial,

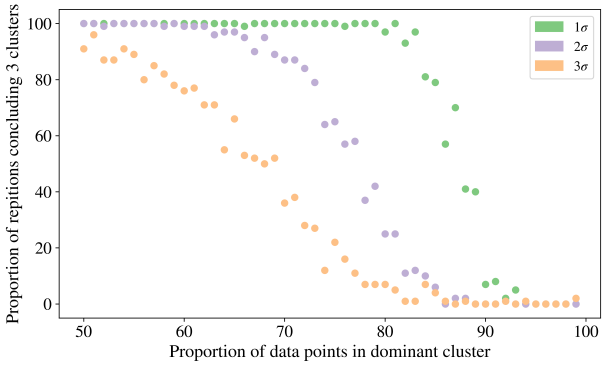

Figure 10. Percentage of simulations for which the CH index attained a maximum for the clustering containing 3 clusters against the proportion of the data which is placed into a dominant cluster.

fungal, pollen and non-biological samples are grouped together in blue, green, orange and black respectively. Different scales are used for the different groups to prevent the dominant class from obscuring information in the other classes.

In 2008, the majority of the Bacteria is placed into a single cluster for both 3σ and 9σ. The fungal and a number of pollen particles are placed into the same two clusters when using 3σ and into one cluster when using 9σ. The non-biological samples, consisting primarily of grass smoke, are clustered mostly with bacterial samples, possibly due to their similar size. In addition, there two clusters when using 3σ and three clusters when using 9σ containing primarily pollen.

In 2014, pollen has been placed primarily into 1 or 2 clusters. Some of the fungal samples have been placed into a singleton cluster. For both thresholds the bacteria is grouped with some of the fungal samples. The non-biological material has almost entirely been removed by the threshold and the remaining material has been divided among a number of the clusters.

In both 2008 and 2014, some of the material has been segregated into clusters containing primarily one broad class of biological aerosol. However, a number of fungal particles has been grouped with pollen samples in the case of 2008 and a number of the fungal samples have been grouped with bacterial particles in 2014. The more successful segregation of pollen in 2014 may be due to the much larger size range for the paper mulberry sample, whereas in 2008 the fungal and pollen material may be grouped due to presence of a larger number of pollen fragments. It is therefore important when interpreting results from an ambient campaign that it is possible that clusters may contain more than one broad biological class.

Also note that this potentially undesirable grouping of material from two different classes has occurred prior to the final stages of the algorithm and therefore will be apparent in the final solution regardless of the number of clusters concluded, and cannot be rectified by using a different validation index.

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 and 9σ using z-score standardisation, taking logs of the size and

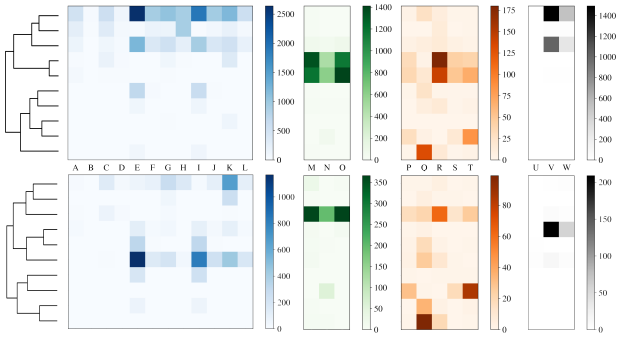

Figure 11. Dendrogram truncated at 10 clusters (left) for laboratory data collected in 2008 alongside heat map of matching matrix (right) indicating cluster composition by each sample segregated by bacteria, fungal, pollen and non-biological in blue, green, orange and black respectively. Separate scales are used for each broad class to prevent dominant class obscuring detail in the other classes. Hierarchies for 3σ (top) and 9σ (bottom) are presented.

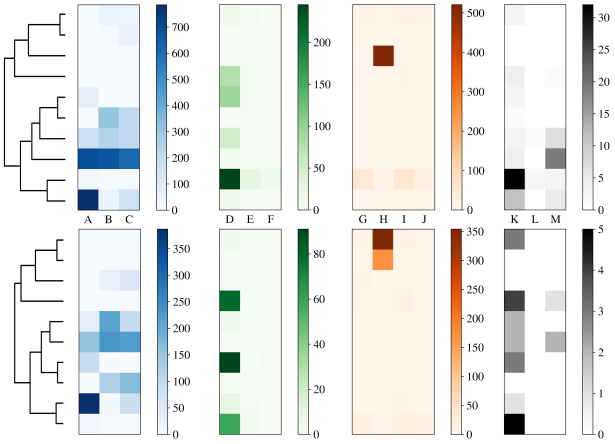

Figure 12. Dendrogram trunctated at 10 clusters (left) for laboratory data collected in 2014 alongside heat map of matching matrix (right) indicating cluster composition by each sample segregated by bacteria, fungal, pollen and non-biological in blue, green orange and black respectively. Separate scales are used for each broad class to prevent the dominant class obscuring detail in the other classes. Hierarchies for 3σ and 9σ (bottom) are presented.

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 $0.1\%, 0.2\%, \cdots, 1.0\%, 2.0\%, \cdots, 10.0\%$.

Table 6. 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 7. 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

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 13 we 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 14 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 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~~ The 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 8 and 9. 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.

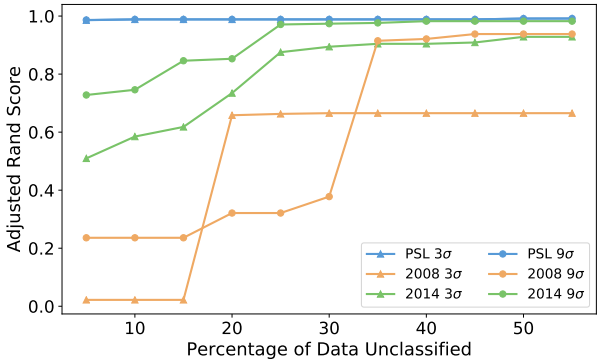

Figure 13. Adjusted rand score using different thresholds of percentage of points we allow to be left in the analysis for DBSCAN.

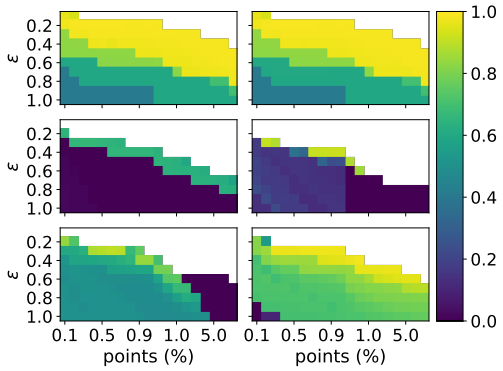

Figure 14. Adjusted rand score for DBSCAN, over a range different values of ϵ and minimum number of points required to form a neighborhood. The minimum number of points is expressed relative to the total number of points. The columns correspond to 3 and 9σ respectively. The rows correspond to the PSL, 2008 and 2014 data respectively.

In the worst case scenario i.e. using 3σ, on the 2008 data we fail to remove a ~~sizable~~ sizeable 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.

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

Table 8. 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 9. 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

a high value of the adjusted rand score was obtained regardless of whether we took logs, what ~~standardization~~ standardisation was used or the size threshold imposed.

Figure 15 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~~ gradient boosting using the current data preparation in Tables 10 and 11. 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.~~ 0.933).

In the worst case scenario a similar performance is achieved (AR = ~~0.877~~0.882). 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. The non-biological samples have been removed from this data set prior to gradient boosting being applied when using a fluorescent threshold of either 3σ or 9σ. We elect to remove these particles since too few of the non-biological samples that exceed either threshold to produce a viable training class.

4.4 K-means

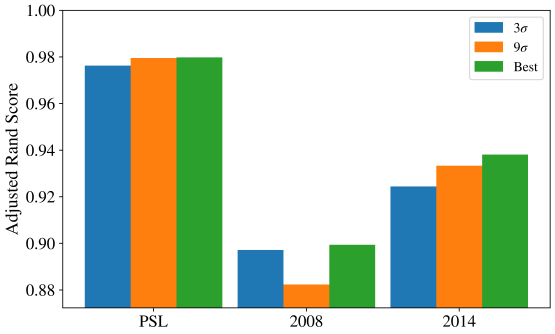

Figure 15. Performance of Gradient Boosting for the different data sets when using 3σ and 9σ.

Table 10. 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~~ 1911	8	~~18 10~~ 19
fungal ~~spores~~	~~6~~ 7	~~216~~ 219	~~31 7~~29
pollen	6	~~23~~ 25	~~584 5~~ 595

~~non-biological 4 5 0 1~~ height

Similar to the findings presented in Ruske et al. (2017), k-means performed poorly and hence the results are omitted from the main text. The results are available in the repository published alongside the manuscript (see the code/data availability section for further details).

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 the 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.

Should there not be sufficient training data available ~~we will have to use~~ an unsupervised approach may be required. In this case, a possible alternative to ~~Hierarchical Agglomerative Clustering is found~~HAC is provided. 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.

Table 11. 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~~ 85	~~79~~ 76	~~7~~ 8
fungal~~spores~~	~~51~~ 56	~~892~~ 898	~~148~~ 147	~~3~~ 2
pollen	~~9~~ 8	~~75~~ 72	~~288~~ 293	~~1~~ 0
non-biological	~~8~~ 4	~~4~~ 5	~~293~~ 3	294

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 of HAC was in general much lower than what ~~could be achieved for the PSLs~~ was achieved previously using the PSLs (Crawford et al., 2015). Performance was heavily dependent on the data preparation strategy used, and often results could vary substantially between different strategies and data sets~~.~~, potentially due to differences in the fluorescence measurements across the two data sets. A potential issue with the CH index is highlighted, whereby we see a failure of the index to determine the correct number of clusters as the size of the dominant class and variation in the data increases. Some of the pollen samples were clustered with the fungal samples when analysing the data from 2008. A number of the pollen particles may be fragmented which may explain why this grouping may occur. Similarly grass smoke was grouped with the bacterial samples, potentially due to their similar size. Caution will therefore be required when applying the HAC algorithm to ambient data, and it must be noted in particular that material from two different classes may be placed into the same cluster and that the CH index may indicate an incorrect number of clusters if the data collected contains a significant quantity of one particular type of particle.

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~~ spore samples analysed. It is therefore possible that higher spectral instruments such as the spectral intensity bioaerosol sensor (Nasir et al., 2018), will be required to provide a more accurate classification.

Code and data availability. Part of the code used produce the above manuscript is part of an ongoing development of a software suite for analysis of various UV-LIF instruments, are available 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 the current manuscript are available at https://github.com/simonruske/AMT-2018-126.

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

References

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

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

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

Carrera, M., Zandomeni, R., Fitzgibbon, J., and Sagripanti, J.-L.: Difference between the spore sizes of Bacillus anthracis and other Bacillus species, Journal of applied microbiology, 102, 303–312, 2007.

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.

Crawford, I., Gallagher, M. W., Bower, K. N., Choularton, T. W., Flynn, M. J., Ruske, S., Listowski, C., Brough, N., Lachlan-Cope, T., Fleming, Z. L., et al.: Real-time detection of airborne fluorescent bioparticles in Antarctica, Atmospheric Chemistry and Physics, 17, 14 291–14 307, 2017.

Crotzer, V. and Levetin, E.: The aerobiological significance of smut spores in Tulsa, Oklahoma, Aerobiologia, 12, 177–184, 1996.

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.

Freund, Y. and Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences, 55, 119–139, 1997.

Friedman, J., Hastie, T., and Tibshirani, R.: The elements of statistical learning, vol. 1, Springer series in statistics New York, NY, USA:, 2001.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Annals of statistics, pp. 1189–1232, 2001.

Fumanal, B., Chauvel, B., and Bretagnolle, F.: Estimation of pollen and seed production of common ragweed in France, Annals of Agricultural and Environmental Medicine, 14, 2007.

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.

Geiser, M., Leupin, N., Maye, I., Im Hof, V., and Gehr, P.: Interaction of fungal spores with the lungs: distribution and retention of inhaled puffball (Calvatia excipuliformis) spores, Journal of Allergy and Clinical Immunology, 106, 92–100, 2000.

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.

Hernandez, M., Perring, A. E., McCabe, K., Kok, G., Granger, G., and Baumgardner, D.: Chamber catalogues of optical and fluorescent signatures distinguish bioaerosol classes, Atmospheric Measurement Techniques, 9, 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.

Hubert, L. and Arabie, P.: Comparing partitions, Journal of classification, 2, 193–218, 1985.

Kang, D.-Y., Son, M.-S., Eum, C.-H., Kim, W.-S., and Lee, S.-H.: Size determination of pollens using gravitational and sedimentation field-flow fractionation, Bulletin of the Korean Chemical Society, 28, 613–618, 2007.

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.

Mäkelä, E. M.: Size distinctions between Betula pollen types—a review, Grana, 35, 248–256, 1996.

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.

Nasir, Z., Rolph, C., Collins, S., Stevenson, D., Gladding, T., Hayes, E., Williams, B., Khera, S., Jackson, S., Bennett, A., et al.: A Controlled Study on the Characterisation of Bioaerosols Emissions from Compost, Atmosphere, 9, 379, 2018.

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.

Pierucci, O.: Dimensions of Escherichia coli at various growth rates: model for envelope growth., Journal of bacteriology, 135, 559–574, 1978.

Pinnick, R. G., Hill, S. C., Nachman, P., Pendleton, J. D., Fernandez, G. L., Mayo, M. W., and Bruno, J. G.: Fluorescence particle counter for detecting airborne bacteria and other biological particles, Aerosol Science and Technology, 23, 653–664, 1995.

Pöhlker, C., Huffman, J., and Pöschl, U.: Autofluorescence of atmospheric bioaerosols–fluorescent biomolecules and potential interferences, Atmospheric Measurement Techniques, 5, 37–71, 2012.

Pöhlker, C., Huffman, J. A., Förster, J.-D., and Pöschl, U.: Autofluorescence of atmospheric bioaerosols: spectral fingerprints and taxonomic trends of pollen, Atmospheric Measurement Techniques, 6, 3369–3392, 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, 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. and Huffman, J. A.: Evaluation of a hierarchical agglomerative clustering method applied to WIBS laboratory data for improved discrimination of biological particles by comparing data preparation techniques, Atmospheric Measurement Techniques, 11, 4929–4942, 2018.

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.

Ting, K. M.: Confusion Matrix, pp. 209–209, Springer US, Boston, MA, https://doi.org/10.1007/978-0-387-30164-8_157, https://doi.org/10.1007/978-0-387-30164-8_157, 2010.

Vinh, N. X., Epps, J., and Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, Journal of Machine Learning Research, 11, 2837–2854, 2010.

Appendix A: Comparison of particle size with other studies

To contextualise the samples collected in the current study we examined the literature to find similar studies using the WIBS as well as other studies using microscopy. In the case of most of the samples we were able to find a paper on the same or similar species of particle which are presented in Table A1.

Appendix B: ABC counts / Average particle sizes

To aid in comparing the data presented with other studies, we have presented Tables B1 and B2 which are very similar to the table in the appendices of Hernandez et al. (2016). A, B and C are used to denote particles which exceed the fluorescent threshold in FL1_280, FL2_280, FL2_370 respectively. For example A is used to denote a particle that was only fluorescent in the FL1_280 channel only. Combinations such as AB, AC, BC and ABC are used to denote particles which exceed a

fluorescent threshold in more than one channel. For example, AB is used to denote a particle that exceeded the fluorescent in both the FL1_280 and FL2_280 channels.

The same information but using a 9σ threshold instead is presented in Table B3.

Appendix C: Summary of average properties of the different data sets

In the following section we summarise mean and standard deviations in each of the five measurements in each of the samples collected in 2008 and 2014. The properties presented in Tables C1, C2, C3 and C4 are after a size threshold of 0.8μm is imposed and a fluorescent threshold of either 3σ or 9σ has been applied. These summary statistics presented are prior to any log-transformations or data standardisation has been applied.

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.

Table A1. Average particle sizes for the current study compared with other studies. The sizes presented here are collated from the following studies [1] Healy et al. (2012a), [2] Savage et al. (2017), [3] Hernandez et al. (2016), [4] Pierucci (1978), [5] Carrera et al. (2007) , [6] Crotzer and Levetin (1996), [7] Geiser et al. (2000), [8] Pinnick et al. (1995), [9] Fumanal et al. (2007), [10] Mäkelä (1996), [11] Kang et al. (2007), [2008] & [2014] the current study.

Sample	Measurement type	Size (μm)	Reference
Paper mulberry	WIBS4 low-gain	13.6 ± 6.2	[1]
	WIBS3	7.18 ± 4.74	[2008]
	WIBS3	3.41 ± 1.43	[2008]
	WIBS3	11.27 ± 1.74	[2014]
	Miscroscopy	13.8	[11]
Ragweed pollen	WIBS4 low-gain	24.5 ± 7.6	[1]
	WIBS3	3.51 ± 1.38	[2008]
	WIBS3	4.70 ± 1.71	[2008]
	Microscopy	13.02 ± 0.12 - 14.86 ± 0.16	[9]
Birch pollen	WIBS4 low-gain	19.0 ± 9.2	[1]
Betula lenta, nigra & populifolia	WIBS4	2.5 ± 4.2	[3]
Birch pollen	WIBS3	3.98 ± 1.59	[2008]
Betula (various)	Microscopy	$17.31 \pm 0.08 - 24.36 \pm 1.59$	[10]
White poplar	WIBS4A	18.7 ± 1.9	[2]
White poplar fragments	WIBS4A	7.4 ± 4.0	[2]
Aspen pollen	WIBS3	3.72 ± 2.49	[2014]
Poplar pollen	WIBS3	3.63 ± 2.39	[2014]
Bermuda grass smut	WIBS4 high-gain	4.7 ± 2.2	[1]
	WIBS3	3.57 ± 1.16	[2008]
	Microscopy	6.7×6.5	[6]
Johnson grass smut	WIBS4 high-gain	8.9 ± 1.5	[1]
	WIBS3	3.47 ± 1.00	[2008]
	WIBS3	3.35 ± 0.78	[2008]
	Microscopy	13.9×12.6	[6]
Puffball spores	Microscopy	3.5 ± 0.24	[7]
	WIBS3	2.50 ± 0.85	[2008]
	WIBS3	2.45 ± 1.16	[2008]
	WIBS3	3.39 ± 1.76	[2008]
	Fluorescence particle counter	2-4	[8]
Bacillus atrophaeus spores	WIBS4A	2.2 ± 0.4	[2]
	WIBS3	1.00 ± 0.40 - 1.60 ± 0.78	[2008, 2014]
	Microscopy	1.22 ± 0.12 (length)	
		0.65 ± 0.05 (diameter)	[5]

Table B1. For the data collected in 2008, a summary of size and fluorescent measurements for each sample to include: the number of particles in the sample (total), average equivalent optical diameter (EOD), standard deviation of the size (σ), the number of points that exceeded a fluorescent threshold of 3 standard deviations above the average forced trigger measurement ($n > 3\sigma$), and ABC counts using a 3σ threshold.

	n	EOD	σ	$n > 3\sigma$	A	B	AB	C	AC	BC	ABC
Bacteria											
Bacillus atrophaeus (unwashed)	5778	1.4	0.5	1015	322	200	74	113	48	90	168
—"— (unwashed, diluted)	1525	1.1	0.7	82	65	6	3	4	0	3	1
—"— (washed)	4694	1.6	0.8	1246	728	107	191	18	29	5	168
—"— (washed, diluted)	1786	1.5	0.8	280	183	21	30	9	10	12	15
—"— vegetative cells (unwashed)	6142	1.1	0.4	5546	409	693	771	75	79	287	3232
—"— vegetative cells (unwashed, diluted)	2192	1	0.2	1739	484	279	326	30	26	67	527
—"— vegetative cells (washed)	6002	1.3	0.6	1961	1797	3	139	0	1	1	20
—"— vegetative cells (washed, diluted)	2827	1.1	0.2	2218	2178	3	36	0	0	1	0
E. coli (unwashed)	4956	1.2	0.5	4097	366	578	174	179	69	868	1863
—"— (unwashed, diluted)	2508	1	0.2	1778	751	309	82	99	27	263	247
—"— (washed)	5669	1.5	0.8	2627	2508	1	99	0	0	0	19
—"— (washed, diluted)	2104	0.9	0.2	1390	1383	0	5	0	0	2	0
Fungal											
Bermuda grass smut	5220	3.6	1.2	2681	1446	7	34	271	495	81	347
Johnson grass smut I	2157	3.5	1	1211	76	3	3	796	128	92	113
Johnson grass smut II	5091	3.3	0.8	2675	217	8	1	1939	270	132	108
Pollen											
Birch pollen	164	4	1.6	112	16	1	0	29	7	8	51
Paper mulberry pollen I	295	7.2	4.7	237	16	2	9	2	0	21	187
Paper mulberry pollen II	735	3.4	1.4	405	159	2	9	72	59	37	67
Ragweed pollen I	241	3.5	1.4	127	24	1	0	57	12	7	26
Ragweed pollen II	328	4.7	1.7	209	21	0	1	41	16	15	115
Non-biological											
Diesel smoke	7900	1.1	0.4	16	3	4	0	5	0	0	4
Grass smoke I	9212	1.1	0.4	2976	1	234	0	2004	0	737	0
Grass smoke II	5245	1.1	0.4	900	3	51	0	668	0	176	2

Table B2. For the data collected in 2014, a summary of size and fluorescent measurements for each sample to include: the number of particles in the sample (total), average equivalent optical diameter (EOD), standard deviation of the size (σ), the number of points that exceeded a fluorescent threshold of 3 standard deviations above the average forced trigger measurement ($n > 3\sigma$), and ABC counts using a 3σ threshold.

	n	EOD	σ	$n > 3\sigma$	A	B	AB	C	AC	BC	ABC
Bacteria											
Bacillus atrophaeus (washed)	3321	1	0.4	2685	2545	1	15	1	81	1	41
Bacillus atrophaeus (unwashed)	2896	1	0.5	2248	85	15	2	1166	88	350	542
E. coli (unwashed)	2534	1.2	0.6	1640	268	10	5	439	239	48	631
Fungal											
Puffball spores I	1739	2.5	0.8	35	3	1	0	27	1	3	0
Puffball spores II	553	2.5	1.2	16	2	0	0	12	1	0	1
Puffball spores III	1627	3.4	1.8	506	79	4	73	168	7	68	107
Pollen											
Aspen pollen	398	3.7	2.5	74	5	1	0	35	1	11	21
Poplar pollen	375	3.6	2.4	104	7	0	3	45	4	21	24
Paper mulberry pollen I	565	11.3	1.7	543	3	0	1	4	0	35	500
Ryegrass pollen	47	3.3	2.1	21	0	0	0	6	0	7	8
Non-biological											
Fullers' earth	3064	3.6	2.8	62	40	1	0	8	4	3	6
Phosphate buffered saline	3226	1.2	1.6	50	29	7	0	11	1	0	2
NaCl	2197	1.4	0.8	6	6	0	0	0	0	0	0

Table B3. Summary of properties for samples collected from 2008 and 2014 respectively using a fluorescent threshold of 9σ.

	EOD	σ	$n > 9\sigma$	A	B	AB	C	AC	BC	ABC
2008										
Bacteria										
Bacillus atrophaeus (unwashed)	2.2	0.6	34	9	2	3	13	1	3	3
—"— (unwashed, diluted)	2.7	1.8	4	1	0	0	0	0	2	1
—"— (washed)	2.6	1.1	217	182	0	9	0	5	0	21
—"—(washed, diluted)	2.6	1	38	28	0	1	1	0	6	2
—"— vegetative cells (unwashed)	1.3	0.6	2051	273	326	132	33	19	184	1084
—"— vegetative cells (unwashed, diluted)	1.2	0.3	278	121	72	24	2	2	14	43
—"— vegetative cells (washed)	1.7	0.9	581	567	0	13	0	0	0	1
—"— vegetative cells (washed, diluted)	1.3	0.3	196	196	0	0	0	0	0	0
E.coli (unwashed)	1.5	0.7	1676	343	97	23	92	30	334	757
—"— (unwashed, diluted)	1.1	0.2	413	333	23	4	12	4	17	20
—"— (washed)	1.7	0.9	1516	1506	2	7	0	0	0	1
—"— (washed, diluted)	1.1	0.3	349	348	0	0	0	0	1	0
Fungal										
Bermuda grass smut	4	1.5	423	118	10	14	133	19	37	92
Johnson grass smut	4.2	1.3	259	0	0	1	171	0	29	58
Johnson grass smut II	3.8	1	378	2	2	0	340	0	29	5
Pollen										
Birch pollen	4.5	1.8	57	7	0	0	9	0	2	39
Paper mulberry	7.8	4.6	212	22	0	7	3	0	30	150
Paper mulberry II	3.9	2.1	107	17	2	2	21	0	39	26
Ragweed pollen	4.7	1.4	34	0	0	0	11	0	2	21
Ragweed pollen II	5.5	1.5	117	2	0	0	9	0	10	96
Non-biological										
Diesel smoke	1.1	0.2	6	0	3	1	1	0	0	1
Grass smoke	2	0.5	236	0	1	0	218	0	17	0
Grass smoke	2	0.4	68	0	0	0	64	0	4	0
2014										
Bacteria										
Bacillus atrophaeus(washed)	1.3	0.5	735	721	0	0	0	12	1	1
Bacillus atrophaeus (unwashed)	1.5	0.5	679	2	0	0	262	7	264	144
E.coli(unwashed)	1.6	0.7	669	55	0	0	209	135	13	257
Fungal										

Table C1. Summary of particle measurements for the 2008 data set using a fluorescent threshold of 3σ.

Sample	$n > 3\sigma$		FL1_280	FL2_280	FL2_370	Size	AF
Bacillus atrophaeus (unw)	952	mean	94.6	63.9	65.4	1.4	7.7
		S. D.	47.3	36.1	44.6	0.4	3.8
—"— (unw, dil)	52	mean	110.3	51.8	43	1.3	8.1
		S. D.	84.8	79.7	76.5	0.8	6.1
—"— (w)	1171	mean	164.2	60.6	48.5	1.7	9.3
		S. D.	136.4	58.5	55.8	0.8	4.9
—"— (w, dil)	241	mean	140.7	50.3	46.1	1.7	9.4
		S. D.	97.9	57.1	58.4	0.8	5.9
—"— vegetative cells (unw)	4779	mean	239.3	221.2	192	1.1	4.7
		S. D.	287.5	293.3	284.9	0.4	2
—"— —"— (unw, dil)	1488	mean	140.5	94.5	68.8	1	6.1
		S. D.	71.6	70.1	60.4	0.2	3.8
—"— —"— (w)	1884	mean	214.5	29.8	12.4	1.4	6.4
		S. D.	156.7	44.9	33.6	0.6	3.4
—"— —"—(w, dil)	2064	mean	153.8	19	7.4	1.1	11
		S. D.	43.2	19.3	17.7	0.2	5.8
E coli. (unw)	3684	mean	222.5	240.4	247.7	1.2	4.7
		S. D.	301.6	351.1	375.5	0.5	2
—"— (unw, dil)	1448	mean	139.3	70	63.9	1	5.5
		S. D.	99.2	56.1	59.6	0.2	2.5
—"— (w)	2365	mean	351.4	12.5	0.8	1.6	7.5
		S. D.	317.8	30.7	22.5	0.8	4.7
—"— (w, dil)	835	mean	202.6	10.7	4.1	1	6.6
		S. D.	56.3	20.2	20.8	0.2	2.8
Bermuda grass	2681	mean	138.1	46.5	97.9	3.6	12.6
		S. D.	122.7	134.7	153.9	1.2	6.4
Johnson grass I	1209	mean	160.5	97.4	154.9	3.5	11
		S. D.	419.7	280.4	169.2	1	6
Johnson grass II	2673	mean	66.7	25.7	124.4	3.3	11.6
		S. D.	28.9	48.3	89.4	0.8	5.7
Birch pollen	111	mean	662	433.7	250.4	4	8.3
		S. D.	854.8	586.9	280	1.6	6.9
Paper mulberry I	233	mean	668.3	1228.4	1247.9	7.3	10.9
		S. D.	583	851.5	856.6	4.7	7.1
Paper mulberry II	397	mean	142.8	159.6	229.2	3.5	13.5
		S. D.	188.7	412.3	443.4	1.4	6.5
Ragweed I	123	mean	384.6	219.5	173.2	3.6	10
		S. D.	698.7	447.9	211.1	1.3	6.5

Table C2. Summary of particle measurements for the 2008 data set using a fluorescent threshold of 9σ.

Sample	$n > 9\sigma$	statistic	FL1_280	FL2_280	FL2_370	Size	AF
Bacillus atrophaeus(unw)	34	mean	214.2	142.5	163.3	2.2	10.2
		S. D.	79.3	60.2	72.9	0.6	5.6
—"— (unw, dil)	4	mean	230.8	259.2	242.5	2.7	14.4
		S. D.	243.8	162.1	142	1.8	11.3
—"— (w)	217	mean	358	121.6	110.1	2.6	12.4
		S. D.	218.1	105.4	95.5	1.1	6.1
—"— (w, dil)	38	mean	276	128	123.7	2.6	14.8
		S. D.	169.6	97.3	101.8	1	7.9
—"— vegetative cells (unw)	1915	mean	423.7	400.4	358	1.4	4.8
		S. D.	384	399.3	393.3	0.6	2.4
—"— —"— (unw, dil)	264	mean	244.9	166	123.1	1.2	7.5
		S. D.	94.5	117.3	103.9	0.3	4.6
—"— —"— (w)	573	mean	347.9	50.3	21.4	1.7	7.6
		S. D.	230.2	72	53.9	0.9	3.9
—"— —"—(w, dil)	194	mean	247.7	26.8	11.1	1.3	13.8
		S. D.	32.6	21.4	17	0.2	6.8
E coli. (unw)	1547	mean	413.7	447.1	470.3	1.5	4.8
		S. D.	388.8	467.3	498.4	0.7	2.4
—"— (unw, dil)	371	mean	254.1	75.4	68.8	1.1	6.6
		S. D.	111.1	86.8	92.2	0.2	3
—"— (w)	1461	mean	463.6	18.1	2.8	1.8	8.5
		S. D.	360.6	34.3	24	0.9	5.2
—"— (w, dil)	302	mean	260	12.8	6.4	1.1	7
		S. D.	47	22.5	24.7	0.2	3.2
Bermuda grass	423	mean	271.7	203.8	303.7	4	13.4
		S. D.	262.6	285.5	303.1	1.5	7.3
Johnson grass I	259	mean	510.2	380.2	344.8	4.2	9
		S. D.	814.8	513	289.3	1.3	6.1
Johnson grass II	378	mean	77.7	82.5	267.8	3.8	12.6
		S. D.	41.1	96.2	161	1	6
Birch pollen	56	mean	1229.8	828.8	406.1	4.6	5.4
		S. D.	891.9	605.8	324.5	1.7	5.8
Paper mulberry I	209	mean	730.8	1363.4	1384.5	7.9	11.2
		S. D.	583.7	794.4	798	4.5	7.1
Paper mulberry II	103	mean	258.1	556	690.7	4	13.4
		S. D.	340.3	663.8	682.1	2	7
Ragweed I	34	mean	1188	750.4	393.1	4.7	6.5
		S. D.	933.1	577.8	299.9	1.4	6.5

Table C3. Summary of particle measurements for the 2014 data set using a fluorescent threshold of 3σ.

Sample	$n > 3\sigma$	statistic	FL1_280	FL2_280	FL2_370	Size	AF
Bacillus atrophaeus (unwashed)	1728	mean	104.5	45.5	26.5	1.2	8.4
		S. D.	118	45.9	61.2	0.4	4.3
Bacillus atrophaeus (washed)	1322	mean	25.4	211.2	357	1.2	5
		S. D.	69.5	222.7	376.5	0.5	2.1
E. coli (unwashed)	1290	mean	104.3	174.9	317.4	1.3	6.1
		S. D.	187.3	207.1	395.6	0.6	2.8
Puffball I	504	mean	288.2	218.1	169.3	3.4	12.1
		S. D.	524.4	289	182	1.8	9.8
Puffball II	35	mean	-19.6	64.4	118.4	2.5	17.6
		S. D.	17.8	49.9	107.7	0.8	8.7
Puffball III	16	mean	19.4	64.2	100.2	2.5	20.6
		S. D.	165.4	68.4	60.3	1.2	12.3
Aspen pollen	74	mean	131.3	301	447.6	3.7	17.2
		S. D.	385.8	504.4	631.4	2.5	7.5
Paper mulberry pollen	541	mean	99.9	1907.9	1924.1	11.3	11.8
		S. D.	77.9	311.9	260.9	1.6	5.5
Poplar pollen	104	mean	163.2	338.2	496.2	3.6	17
		S. D.	488.6	525.4	643.3	2.4	9.1
Ryegrass pollen	21	mean	110.7	278.7	569.3	3.3	18.4
		S. D.	340	258.6	431	2.1	8.6
Fullers earth	61	mean	180.2	114.3	148.2	3.7	16
		S. D.	476.2	214.5	367.8	2.8	9.9
NaCl	3	mean	16.7	19.7	14.7	2	9.1
		S. D.	5.4	24.4	32.5	0.7	5.3
Phosphate buffered saline	35	mean	64.2	113.9	89.1	1.4	6.2
		S. D.	342.1	320	324.1	1.8	2.7

Table C4. Summary of particle measurements for the 2014 data set using a fluorescent threshold of 9σ.

Sample	$n > 9\sigma$	statistic	FL1_280	FL2_280	FL2_370	Size	AF
Bacillus atrophaeus (unwashed)	684	mean	195	60.5	46.2	1.4	9.8
		S. D.	144.6	65.3	90.3	0.5	4.7
Bacillus atrophaeus (washed)	608	mean	65.4	358.7	636.8	1.6	4.6
		S. D.	83.6	257.9	402.1	0.5	2
E. coli (unwashed)	632	mean	199.9	284.1	550	1.7	6.2
		S. D.	229.6	251.3	460	0.7	3.1
Puffball I	248	mean	599.7	380.6	252.3	4.3	8.7
		S. D.	606	341.3	226.1	1.8	8.2
Puffball II	3	mean	-20.7	176.7	417.3	2.4	19.7
		S. D.	17.4	76	146.6	0.7	9
Puffball III	1	mean	654	298	284	2	25.6
		S. D.	0	0	0	0	0
Aspen pollen	31	mean	338	643.5	952	5	18.7
		S. D.	529.9	635.9	716.1	3.2	8.6
Paper mulberry pollen	537	mean	101	1921.8	1937.7	11.4	11.8
		S. D.	77.2	268.5	209.1	1.4	5.5
Poplar pollen	50	mean	355.9	644.5	938.8	4.4	16.5
		S. D.	651.1	626.5	694.4	2.9	9.8
Ryegrass pollen	15	mean	168.1	361.5	753.3	3.6	17.8
		S. D.	387.7	263.5	375.9	2.4	9.5
Fullers earth	20	mean	521.1	274.6	411.7	4.5	15.7
		S. D.	719.7	317.2	552.1	3.2	11.3
Phosphate buffered saline	3	mean	748.7	725.7	711	4.6	4.7
		S. D.	918.1	878.5	888.2	5.1	1.6