Referee #1 – Darrell Baumgardner

For clarity the referee's comments are copied in black and our responses are given in blue.

Comment 1

Explain why one of the most common machine learning techniques was not tested, i.e. neural networks.

We thank the reviewer for his comment, from testing Neural Networks we found they perform very well and their inclusion has added further depth to the paper.

Originally neural networks were excluded from the study as they are not currently included in the Sci-kit learn package. We have nonetheless investigated an alternative package, pycaffe, and have tested neural networks with 1 and 2 hidden layers.

In Section 2, we have added a subsection (2.6) that details the method in similar detail to the other methods previously presented. Table 3 which has now been replaced with two figures (see comment 9 of referee #2) which also includes the results from our tests using the neural network method. Other minor changes have been made e.g. the method and conclusions regarding the method are now mentioned in the abstract. We have also added two paragraphs to the results section presented below:

> *"A possible alternative to Gradient Boosting is the 2-layer Neural Network that performs nearly as well as Gradient Boosting for the full data set. It may be possible to extend the number of layers in order to yield further improvements but we would suggest, due to the increased time requirements with additional layers, that this should be done using a Graphics Processing Unit (GPU) which could offer significant gains in the amount of time required to train the network. This is a benefit over the Gradient Boosting Algorithm, where it is far less clear how the algorithm might be parallelised.*
>
> *On the other hand, Neural Networks are very difficult for the user to tailor to achieve good performance. The results presented are achieved after a lot of experimentation, especially in terms of the learning rate. A learning rate that is too high often will overshoot a minimum for the loss function and a learning rate that is too low will fail to reach a minimum at all. Over-fitting, where the model fits very well for the training data but does not generalise well for the testing data, is also an issue. Over-fitting seems not to be a problem for the Gradient Boosting Algorithm, which also did not require any parameter selection by the user."*

Comment 2

Evaluate which of the particle metrics provides the most information on the different types of particles by each of the techniques, i.e. is it the size, a particular wavelength, a combination of

wavelengths? For example, in Hernandez et al., Fig. 3 shows that bacteria and fungi have similar fluorescence signatures but differ by size, whereas fungi and pollen have similar sizes but differ in fluorescence signatures. Given that the MBS has 8 fluorescence channels, and the purpose of more channels than the WIBS is that it provide more information, then this needs to be validated with these processing techniques

In order to include the additional information requested we have split the results section into general results and further analysis of the gradient boosting algorithm.

The difficulty we find with plots such as Fig. 3 from the Hernandez paper and Figures 5 and 6 of our own paper is that it is very difficult to demonstrate the variability in the data without making the graph unreadable. But without consideration of the variability it is difficult to come to accurate conclusions about differences between populations. We can see from our newly drawn Figure 5 and Figure 6 that on average there are differences between the different bio-groups, in size, fluorescent intensity and the channel in which the fluorescent signature peaks. But this doesn't necessarily translate to statistically significant differences between the populations when the variability of the data is considered. Nor do these differences in the data necessarily translate into any methods being able to distinguish between the groups with that variable alone.

We therefore propose to respond to this comment in conjunction with removing lesser important variables (see comment 5). Not all of the methods provide us with information on which variables are most important. But the Decision trees and ensemble methods do, and our best performing method was Gradient Boosting so we elect to use this method to produce a variable importance plot.

We have added several paragraphs to the new subsection 5.2.3 that outlines the importance of the variables and the implication of removing lesser important variables on the performance of the algorithm.

Comment 3

Which of the lab bio types was the most efficiently identified, bacteria, fungus or pollen?

We have provided a subsection (5.2.1), breaking down the results for the gradient boosting algorithm, which includes a table of the breakdown and accompanying text. From this breakdown it is clear that the largest cause of error is in the misclassification of fluorescent particles as non-fluorescent and that the error between the fluorescent classes was relatively small. We can also see from the breakdown for the individual bacteria or pollen species (see comment 4) that the biggest errors are from the fungal spore sample and the aspen and poplar pollen samples. Conversely the Paper mulberry sample was most efficiently identified (most likely to the particles being much larger and brighter than for the other sample), followed by the bacterial samples.

Comment 4

Could any of the techniques separate between the individual bacteria or pollen species?

We have repeated the analysis breaking down the bacteria into three samples (i) BG spores washed (b) BG spores unwashed (c) E-coli unwashed and the pollen samples into three samples (i) aspen pollen (ii) poplar pollen and (iii) paper mulberry. We can see that the washed sample could be in general distinguished from the unwashed samples, but the unwashed samples could only be distinguished from each other with limited success. For the pollen samples the aspen pollen was indistinguishable from the poplar pollen and the paper mulberry sample was distinguishable from all the other samples. This is likely to be because the paper mulberry particles are much larger and brighter than the other samples.

We have added a subsection 5.2.2 which details this information.

Comment 5

The real surprise is that removing the shape information not only didn't reduce the ability to separate bio types but in some of the techniques removing this improved their efficiency. So the question is what would happen if you reduced the number of wavelengths down to the three of the WIBS? Given the number of WIBS users, that would be of great interest. The Hernandez study suggests that maybe three wave lengths are sufficient to resolve the three general bio-types but probably not to differentiate within a species. As it is, the reader will conclude that the shape detector is not needed. Would that be a valid conclusion?

When we see a degradation of performance with the inclusion of the shape measurements, we believe this is a reflection of the method rather than of the data. Certain methods such as cluster analysis and quadratic discriminant analysis find it very difficult to cope with high dimensional data and hence yield poorer performance. We would conclude that the shape detector does provide valuable information just that not all of the information it provides necessarily improves performance by that much and not all of the methods are capable of coping with this extra information.

We must make clear that the MBS is a different instrument to the WIBS, in the sense it has only one excitation band, which means that we cannot reduce the variables as the referee suggests and it hence is not possible to reproduce the information that could be collected using a WIBS. Having said this we have considered the implication of removing variables in conjunction with our response to comment 2. Our results seem to demonstrate that a smaller shape detector may produce similar results to the full data set, but removal of the detector in its entirety may hinder the instruments' ability to distinguish between particles.

Comment 6

The paper concludes with a brief mention of processing time but I think that this is a critical topic that needs to be included since real-time identification of bio-types is an important application of this technology so there needs to be an evaluation of the detection efficiency versus processing time.

We have added in the time requirements for each of the methods as requested and the following text to the results section

"We have also provided a subset of our time results. In Figure 9 we have the training and testing times for the supervised methods and the full time taken for the HCA for the Full mixed PSL data. The timings for the reduced data set and for the laboratory generated aerosol are omitted as they show similar patterns. Note in particular that our training set is four times bigger than our testing set since we have used 5-fold cross validation (See Sect. 2)."

"Should these methods be applied to real-time applications, we would expect the testing data to contain much larger number of particles compared to the training data. For example we could collect between 10^4 *and* 10^5 *particles of laboratory data for training. But over the space of several months in an ambient contain we might collect* 10^6 *particles or more. It is for this reason we conclude that methods such as the ensemble methods and neural networks offer distinct advantages over HCA. For HCA the time requirements increase at a much faster rate than the number of particles (Mullner 2013). In other words if we double the amount of data, we will see more than a double in the amount of time required. A similar behaviour is true for the full support vector machines classifier.*

The behaviour of the neural networks and the Ensemble methods is much more desirable. While the training times are relatively large compared to other methods, once the model is fitted the testing time requirements are under a second for several thousand particles which is much faster than the flow rate of the instrument."

It is no longer necessary to include the information in the conclusion about the time requirements so we have removed this paragraph.

Additional Comments

Since the updates have provided a broader, more in depth analysis we felt it necessary to update the conclusion section with the following paragraphs:

"We have tested a variety of different methods that could be used for discriminating between different types of bio aerosol. Cluster analysis, while working well for the reduced data set for PSLs seems to struggle for the laboratory generated aerosol and when applied to the larger dimensional data set, so we suggest that more research for this method is required before it could be reasonably used on ambient data created using the MBS. For the Gaussian methods it seems that the methods work reasonably well for the PSLs, but we believe there are better alternatives when discriminating between atmospheric aerosol.

For the K-Nearest neighbours method we believe that a limiting factor is in the time it takes for the method to classify the testing data. Similarly, while the Full support vector machine performs very well, the time requirements would be inappropriate when larger samples are collected. Conversely while the linear version of the support vector machine performs much faster this is at the cost of performance so we suggest that Support Vector Machines not be used for this task.

Overall, the method we suggest for classification of atmospheric aerosol is the Gradient Boosting Algorithm which produces the best results, with limited user input but cannot easily be parallelised. Another possible alternative in the future, once more research is conduced, is the Neural Network which can be easily ported to a GPU for substantial speed up in training but requires a much larger input for the user, and produces slightly worse results compared to the Gradient Boosting Algorithm.

From our further analysis of the gradient boosting algorithm we also see that a disadvantage for the data we have collected is in the sample of fungal spores which is often misclassified as non-fluorescent as a good proportion of the particles are weakly fluorescent. We believe this issue can be circumvented with collection of a wider range of fungal spore samples in the future. Also we see that for the MBS we have reasonable success in discriminating between single bacterial samples. Finally we realise that performance can be maintained while removing a reasonable number of the lesser important variables, leading us to conclude that a smaller shape detector may be sufficient."