Public Comment- Simon Ruske (simon.ruske@student.manchester.ac.uk)

Received and published: 3 June 2018

Note regarding document formatting: black text shows original referee comment, blue text shows author response, and red text shows quoted manuscript text. Changes to manuscript text are shown as *italicized and underlined*. All line numbers refer to discussion/review manuscript.

[Public Comment] The study presented is an extremely well structured and written investigation into the use of Hierarchical Agglomerative Clustering for classification of biological aerosol using a UV-LIF sensor, and will make an excellent addition to the literature upon publication.

[Author Response] Simon, thanks for taking the time to read and comment on the manuscript. We appreciate the useful comments, which will help improve the quality of the manuscript. We respond to each comment in detail below.

However, the authors may have made a small error [L161-L162] where they state that the conclusions for Ruske et al. (2017) were for ambient data, whereas in the abstract they correctly state that the study was on standardised laboratory particles [L19-L20]. Please could you correct this prior to final publication.

I apologize for this mistake. I am not sure where this error came in our writing process, but I removed the incorrect statement, as requested: "Their conclusions, however, were based on ambient field data using unknown particle types and did not investigate laboratory generated particles of known origin."

In addition the authors may wish to consider the following comments prior to publication.
[L78-L79] Would it be possible to clarify the starting conditions for supervised learning you are referring to? Hyper-parameter selection is an extremely important consideration for neural networks, but other supervised techniques such as decision trees and ensemble methods do exist where low classification error can be attained without providing the algorithm with any initial conditions other than the training data.

This may have been a bit of a miscommunication. We do not deal with any supervised learning methods in this manuscript. We trust your team as the experts in this area. Nicole simply wanted to provide a few sentences of general contrast between supervised and unsupervised methods. That is also why we pointed to your 2017 paper in this section. We have also included citation of your manuscript currently being reviewed in AMT.

[L84-L85] Is it necessary to apply unsupervised techniques to assess the advantages of supervised methods? Do you mean that supervised techniques require laboratory data of known types to assess their advantages? A very important disadvantage of supervised techniques is that they rely on adequate training data, and it is not clear at this point how much training data will be required to adequately represent an ambient environment, which is the point I think you are alluding to here.

This is the way I understand some of the pros/cons of supervised and unsupervised. I agree that the community (probably you first) will continue to lean about how this all works together and how well lab-generated data can be useful to train supervised data algorithms. As you well know, the differences between nicely behaving lab particles and more complicated particles collected in the field confounds most areas of aerosol science to some degree. So these problems will not necessarily be trivial to solve, but I think collectively we are all learning little pieces that will help.

[L186 - 187] Does the z-score rely on the assumption of normality? The z-scores of a normal random variable will be normally distributed whereas the z-scores of a non-normal random variable will be non-normally distributed. **Applied to any data set, regardless of distribution, the resultant variables after z-scoring will have mean of 0 and standard deviation of 1.** Is the purpose of standardising the data to prevent one of the variables from dominating in the analysis or to produce normally distributed data?

Thanks to your prompting, we looked into these details and learned a bit more, which has been helpful to us. You are right that the way we characterized the z-scoring process was not correct. Talking back and forth with the university statistician, we now understand that values can indeed be input scaled to a normal distribution or not. We chose to standardize our variables to a mean of 0 and a variance of 1 so that the output variables would be on comparable scales, but this is also not the same as rigorously normalizing them in the rigorous sense. As a result, we have removed the statement you correctly indicated was inaccurate and updated the sentence as follows:

Original text: "Standardization using the z-score method compares results to a normal (Gaussian) population, ~~and therefore relies on the assumption that input data can be described by a normal distribution (Gordon, 2006).~~"

Updated text: "Standardization using the z-score method compares results to a normal (Gaussian) population, and we have chosen to standardize our variables to a mean of 0 and a variance of 1 so that the output variables would be on comparable scales."

[L203] It would be worth noting that in Crawford et al., 2015, there are particles for which negative measurement of fluorescence was recorded. The option of logtransformations may have been overlooked, as the logarithm is undefined for negative values. This was not intended to imply an assumption of normality, although this assumption has been stated explicitly in Robinson et al., 2013. In these cases would you recommend translating the fluorescence measurements to a range bounded below by 1, or alternatively would it be more appropriate to reject measurements for which the fluorescence produced was negative? It is also important to note that even if the data is log transformed, the data will still have a finite range due to the saturation point on the detector, and hence the data will have a truncated normal distribution rather than a normal distribution, and depending on how often saturation occurred there may still be a peak to the right hand side of the distribution. It is however, perfectly acceptable to apply HAC when the assumptions for best performance are not met as stated in Norusis, 2011.

My understanding is that negative fluorescence values can be observed after subtracting some threshold value from the fluorescence intensity data. Instead of subtracting the data and looking only at positive values, we did the same thing by filtering the data at several discreet thresholds. This gets around the problem of negative values. In any case, we looked at three thresholding scenarios (Table 3), i.e. no threshold, 3 sigma, and 9 sigma. The ultimate result is that we found the most consistently positive results to be as a result of 3 sigma filtering, but this could be different in other situations. You are correct about the fact that particles that exhibit saturation of the detector in any channel will truncate a normal distribution.

[L222] How often did the CH index conclude that there were 2 clusters? When the CH index concluded a number of clusters other than 2, how much of an impact did this have on the quality of the results? Were the two cluster solutions always the best solution?

We did not explore solutions that had more than 2 solutions, simply as a matter of limited time. There are certainly many scenarios in which individual bioparticle types (i.e. pollen, in many instances) can split into two reasonable clusters by themselves, and so independently allowing 3 or more cluster solutions could significantly improve results in many cases. We just didn't have the time to do this systematically, and so we chose to limit analysis to only 2 clusters in all cases. To help clarify this point, we added text at:

L227: "In order to reduce the length and complexity of *discussion, analysis of results in Sections 4.1-4.3 was limited to using cluster products only from the 2-cluster solution. In some cases a 3-cluster solution may have produced higher quality results, but these cases were not investigated.*"

[L267-270 & Figure 3] The HAC algorithm may not necessarily output clusters in the same order that they were inputted as demonstrated in Figure 5. In Figure 3 for preparation strategy A for bacteria and diesel for the 80:20 ratio, is it possible to attain 80% misclassification for a two cluster solution? Perhaps I have misunderstood, but would this not mean that there were more diesel particles in the bacterial cluster and more bacterial particles in the diesel cluster, and hence a better classification error could be attained simply by swapping the labels on the clusters?

You are correct that the order of cluster numbering is unrelated to the order of particles input and so the source of individual particles must be known already, but it is not possible to improve the results by swapping labels in the way you suggest. We independently tracked the source of each particle assigned to each cluster so we can rigorously calculate which particles were incorrectly assigned. The numbering of the clusters is arbitrary and the naming was assigned simply as a function of which particle was assigned in the largest concentration.

[Figure 3 & Table 2] Could you extend the results presented in Figure 3 to include at least one biological versus biological matchup? I notice when considering matching ups which contained only biological material the classification error is much higher. I believe that by not standardising the data this would cause the fluorescence to dominate more in the analysis. In the case of attempting to discriminate between fluorescent and non-fluorescent particles, this may be advantageous. However, in the case of attempting to discriminate between two different types of biological particle, it may be advantageous to give the size and shape measurements more weight, and hence it would be better in these cases to standardise the data. In addition other instruments such as the WIBS-NEO will have fluorescence measurements over a much larger range and fluorescent measurements are recorded often above 10000. What would the implication then be when not standardising the data in this case?

This is another interesting idea, but it was beyond the scope of what we were able to accomplish in the relatively short time we had available for this project. We chose to focus on the ability to separate bio from non-bio particles. While we didn't explore all Scenarios (e.g. A-F) for biological particles, we chose to look at bio-bio separations using Scenario B (i.e. Tables 2 and 3).