Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2018-109-SC1, 2018

Atmospheric Measurement Techniques Discussions

Interactive comment on "Evaluation of a Hierarchical Agglomerative Clustering Method Applied to WIBS Laboratory Data for Improved Discrimination of Biological Particles by Comparing Data Preparation Techniques" *by* Nicole Savage and J. Alex Huffman

S. Ruske

simon.ruske@student.manchester.ac.uk

Received and published: 30 June 2018

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.

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.

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.

[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.

[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?

[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 log-transformations 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.

[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?

[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?

[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?

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2018-109, 2018.