Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2018-109-RC2, 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

Anonymous Referee #1

Received and published: 1 June 2018

This paper builds on existing literature examining unsupervised learning techniques to improve the interpretation and classification of data obtained with WIBS UV-LIF spectrometers. As shown in previous publications, Hierarchical Agglomerative Clustering (HAC) can serve as a robust data analysis method for classification/interpretation of bioaerosol data but the accuracy of technique is highly sensitive to the choice of clustering linkage and data pre-treatment (e.g., Crawford et al., 2015); this is further explored in this paper which elucidates how data pre-treatment choices such as choice

of fluorescent threshold and log normalising data may influence clustering accuracy using laboratory samples of known particle types (Savage et al., 2017) in various synthetic mixtures, and thus the authors present tentative recommendations of data pre-treatment regimes depending on the analysis goals. Overall the paper is well written and the computational experiments well thought out. The findings here are useful and further validate the usefulness of Hierarchical Agglomerative Clustering for interpretation of WIBS data. The results also provide a useful framework for testing Hierarchical Agglomerative Clustering data pre-treatment regimes for other atmospheric science data problems and neatly demonstrate the potential pitfalls of not rigorously performing such tests. I recommend publication after the following comments have been addressed.

Specific comments

L73-77: The authors have conflated some of the terminology relating to unsupervised and supervised leaning methods. I'm uncomfortable with the use of the term clustering when discussing supervised methods as clustering specifically relates to cluster analysis. I suggest replacing "clustering techniques" with "classification algorithms" and "(trains) the clustering algorithm" with "(trains) the classification algorithm".

L120: Please state the bands and what they relate to.

L198: Can the authors please clarify why they have used log spaced bins. Do you mean that you have taken a log of the data and it is binned naturally by the discrete nature of the detector resolution (i.e., fine bins) or have you binned the data into specific (coarse) log bins? If it is the latter can you please state what the bins are and can you comment on how forcing the data to in bins may influence the clustering? My concern here is that too coarsely binning the data may create artificial hotspots due to reduced resolution and bias the clustering, reducing the capacity to differentiate between particles with similar properties. Can the authors comment on this and demonstrate the effect this may have by providing an example for comparison where the data is con-

verted to log space and not binned. I also wonder if the bins should be normalised by the bin width to further complicate matters.

L254: Can the authors comment on the environmental applicability of the chosen ratios. I would suspect that in an urban environment you may expect something closer to a ratio of 1:99 fungal to diesel particles with the converse being true in a forest environment. How does the clustering perform under such extreme mismatches?

L238: Would it be possible to show examples of the cluster centroids for a case where there is significant misclassification? This may illuminate why the algorithm is failing to correctly attribute particles. It may also be useful to examine the fluorescence/AF characteristics of each cluster as a function of size here. A 2D histogram or color density plot could show distinct hot spots that haven't been separated correctly and could provide a basis for manual separation based on sensible thresholds.

L312-315: Can you describe the method for producing the soot as they seem rather large as compared to that in the study of Toprak and Schnaiter (2013) which were also coincidently found to be weakly fluorescent in FL1. Perhaps the soot used in this study is larger and more fluorescent than we may expect of ambient/urban soot which may cause some of the difficulty in correctly attributing in in some cases?

L384: Would we expect to be able to differentiate between 2 different particles of the same type with such coarse spectral resolution?

L415: Again I wonder if the use of too coarsely separated bins may compromise the 9-sigma thresholding and cause misclassification?

L514: Can the authors comment on the applicability of their findings to new high resolution UV-LIF instruments that are beginning to become commercially available. Some of these new instruments have significantly more channels/greater fluorescent resolution than the WIBS.

Technical corrections

L63: instruments, not instrument.

L370: grains, not gains.

L112: Suggest "Experimental and Computational Methods"

L131: "each of the three"

L181: "was the best"

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