Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2018-126-RC2, 2018

Atmospheric Measurement Techniques Discussions

Interactive comment on "Machine learning for improved data analysis of biological aerosol using the WIBS" *by* Simon Ruske et al.

Anonymous Referee #2

Received and published: 12 July 2018

Evaluation of the usefulness of analysis techniques for WIBS data, the subject of this paper, is important. However, several problems make the paper unsuitable for publication without significant modifications and clarifications.

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, pollens, 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

Printer-friendly version

Discussion paper

main categories (bacteria, fungal spores, pollens, 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. Inthis case we can conclude that 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 shows 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 the 2008 and 2014 data sets, and discussing these dendrograms in relation to the data illustrated in Figs. 4 and 5.

2. In comparing the value of classification/clustering approaches the justification for using differentnumbers 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-casescenario and best case scenario when using Gradient Boosting using the current data preparation inTables 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 intesting 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 data and Table 7 has 2008 data). Because of the use of smaller numbers 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 match results.

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 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 Pigmented Bacteria in the Ambient Outdoor Atmosphere, Photochem Photobiol 1997, pp 103-106, in at least acknowledging that 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.

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 this 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 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-

Printer-friendly version

Discussion paper

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 that 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 normalizing 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 Figs. 4 and 5 need error bars.

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.

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.

Additional issues

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.

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

3. 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 samples (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.

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 analyzed here. These should be defined, for example in section 2.1 where the "four fluorescence measurements" are described.

5. The justification for omitting FL4, i.e., that some samples saturate, is inadequate.

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.

7. 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 should explanation is needed.

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.

9. How about defining the matching matrix as used here. What is the criterion of the match?

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.

11. 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?

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2018-126, 2018.

Printer-friendly version

Discussion paper