Atmos. Meas. Tech. Discuss., 5, C3378–C3383, 2012
www.atmos-meas-tech-discuss.net/5/C3378/2012/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Cluster analysis of WIBS single particle bioaerosol data" *by* N. H. Robinson et al.

N. H. Robinson et al.

niall.robinson@metoffice.gov.uk

Received and published: 21 December 2012

We thank the reviewer for their comments, which will undoubtedly make the paper clearer. We have addressed them in turn below.

6398 and Table 2. Where are these data from? Were these measured or made up? The text and table merely state here are the data input to the cluster analysis.

We hope this is clarified by 6398.12, which states "Five different PSL types were measured sequentially". We will change the caption of Table 2. to "Average modal centres of PSL measurements input to cluster analysis algorithm" for further clarity.

6399.1: Why THE 6 major clusters retained? This is a big reduction from 13 to 6

clusters in one sentence with little justification? Why were 13 clusters even chosen. Fig. 4 shows a significant drop in R2 and rise in RMS when 6 clusters are reached.

We direct the reviewer to the discussion of these issues between 6393.20 and 6394.11. We hope this addresses the issues surrounding which clusters to retain. As the discussion says, ultimately this is a decision to be made by the analyst which is informed by the presented statistics, and to a certain in extent their knowledge of PBAP types. As stated on 6399.1-4, the different types of PSLs are likely to be present in broadly similar concentrations (as they all have the same source i.e. they are nebulised in the laboratory). Incidentally, note that this is not the case in ambient datasets presented later in the manuscript. Due to this similarity of cluster size, the metric of "major" clusters detailed on p6393.21 can be used to define which clusters to retain. Only the minor clusters by this definition have been discarded, as stated in the caption of Table 3.

As for the choice of cluster solution, we direct the reviewer to the discussion highlighted above, specifically 6393.7. The point being that both the 13 and 6 (actually 7 seems to be more concomitant) cluster solutions are significant, by definition of the statistics. Kalkstein et al. and Cape et al. (referenced in the manuscript) encourage choosing the "first large step" in the statistics. It should be noted that either solution could justifiably be employed.

Table 3. Now, where do these values come from, measurements? If so why are they different than Table 2? The origin of cluster C at a size not in the original data is also not clear. There is something I do not understand which separates the origins of Tables 2 and 3.

Table 2 details the averages of separate sets of measurements of different PSLs. This entire dataset is then processed using the cluster analysis. Table 3 details the averages of the resultant clusters. Effectively Table 2 is the physical reality that we are trying to resolve the entire dataset into using the cluster analysis. An ideal cluster analysis

would reproduce Table 2 in Table 3.

To be clear, both tables are averages of PSL measurements performed in the laboratory. We appreciate that this is a subtle difference for someone unfamiliar with the work, but hope that the text between 6398.20 and 6399.1 explains this.

Table 3. Are all the PSL used spherical? Is so what is the origin of the large differences in asymmetry factor?

The PSLs are spherical. While there are certainly differences between AF, we would contest weather they are significant. It should be noted in any case that these differences are not to do with the cluster analysis technique, as the AF uncertainties are comparable between Table 2 (unprocessed by cluster analysis) and Table 3 (processed by cluster analysis). We would also highlight that the differences between different averages in these tables (a range of approximately 3-7) is low compared to the ambient datasets (a range of approximately 8-26) suggesting that the PSLs are more symmetric than ambient aerosol, as we might expect to be the case. Finally, the precision of AF measurement in the WIBS is relatively low, and some spread in values is to be expected. A broad point that our manuscript makes is that this lack of precision does not appear to be detrimental to the cluster analysis.

6399.11–18: It is hard to understand how two clusters (A and B) are defined as separated by 0.03 μm in diameter and both fluorescent. Is it really believed that the cluster analysis could make such a fine separation? In fact it can not as seen in Fig. 5. But Fig. 5 is a bit misleading. It only lists the diameter per cluster, when the real separation was probably based on asymmetry factor for this difference, but, again, why should these be different? See the question above about Table 3.

We would absolutely agree that clusters A and B are most likely separated by virtue of their different AFs (and make this point on p6399.14), and we hope that the manuscript did not imply that we believe such small differences in cluster size can be resolved (at least not with the present instrument precision). We have included the average size

on Figure 5 because these are most likely to be the definitive property of the data (as PSLs are defined by their size). As we try to say on p6399.14 without going in to too much detail, there are several possible explanations for this apparently significant difference in AF: either the PSLs physically have a slightly bimodal distribution; the WIBS AF response is artificially slightly bimodal; or its just a statistical anomaly. While this solution is not ideal it should be noted that it doesn't have any bearing on the interpretation of the data (as the clusters are qualitatively similar), and that 1um PSLs are close to the lower limit of the WIBS measurement range.

6399.19: Does not the "population normalised distance simple attribution" approach also have a problem with an inability to separate clusters C and D?

We would say that the "population normalised distance simple attribution" performs perfectly well in this respect, it is the cluster analysis that it is based on which is slightly erroneous. We accept that clusters C and D appear to represent the same PSL group in the clustering solution, that is they are split. However, given this, the attribution performs well.

6400.17: The justification for the 4 cluster solution is not obvious. The RMS does not significantly rise until cluster 3 is reached. R2 also drops more steeply then as well.

See the answer to the second comment: the literature suggests the best choice is at the first significant concomitant change, which we think is at the four cluster solution. Again, as stated on p6394.8-11, both solutions will be representative of the physical reality, but the less good choice may lead to the splitting/conflation of clusters. To a certain extent splitting (as would potentially in this case be the issue when choosing a four cluster solution to a three cluster reality) is also indicated by qualitatively similar clusters which have a correlating time series, which we do not observe.

6400.20-21: Here is a good explanation of how the 10 clusters were reduced to 6 based on sample size. Could this have been applied in going from 13 to 6 clusters in the example with PSL?

We will clarify the PSL case by stating: “The 13 cluster solution was chosen due to the observed decrease in R 2 and N, and the concomitant rise in RMS (Fig. 4). Of these 13 clusters, the six major clusters (as defined in Section 3) were retained for subsequent analysis (Table 3).” on p6399.1 to indicate that we are using the term “major” precisely to refer to the defined statistic.

6401.10: Somewhere about now the authors should reference Tables 4 and 5. In fact these tables are never called out in the text.

We thank the reviewer for bringing this to our attention and will add references to p6400.18 and p6400.22 respectively.

6401.15: I do not understand how two clusters in the 6 cluster solution would be agglomerated in a 9 cluster solution. It seems the resolution would only increase with cluster size. I also am not following how C4 and D4 are agglomerated in the 6 cluster solution when they are distinctly separated in Table 5 which I thought was the 6 cluster solution?

There is a subtle difference between the number of clusters of the solution, and the number of clusters retained from this solution for subsequent analysis. The “six cluster solution” that the reviewer refers to is in fact the ten cluster solution (p6400.19), of which the six largest clusters are retained. Therefore the nine cluster solution is one agglomeration further advanced than the chosen solution, and the six cluster solution is four agglomerations further advanced.

6402.10: Do you mean E4 instead of E3?

Yes. Thank you for bringing this to our attention. Along a similar vein, the column headings for Table 6 are incorrect. They should read: A3, A4+B4 | B3, C4+D4 | C3, E4 | D3, F4 (as in Figure 8), and will be changed in the final manuscript.

6402.12: What does the following phrase mean, “. . .if FL2_280 is typical of grass smut . . .”? Fl2_280 is a type of fluorescent measurement, which could be low or high,

depending on the particle

We will change the indicated sentence to read. “...high FL2_280 fluorescence levels are typical of grass smut....”

Fig. 8: The caption indicates that rainfall is displayed at the bottom of the figure, butthat is not the case. The bottom panel is fungal spores.

Thank you for noting this. We will remove this from the final manuscript.

Interactive comment on Atmos. Meas. Tech. Discuss., 5, 6387, 2012.