Anonymous Referee #1

For clarity the referees comments are copied in black and our responses are offset in blue.

This manuscript presents a new method for performing cluster analysis on data taken by the Wide band Integrated Bioaerosol Sensor (WIBS). Previous cluster analysis soft-ware for use with the WIBS has been severely limited in terms of the amount of data that could be processed (10ˆ4 particles). This typically requires heavy subsampling of the particles and brings into question the representativeness of the clusters. The authors test the performance of various clustering algorithms using laboratory calibration particles and then apply the best-performing algorithm to ambient data collected during the BEACHON-RoMBAS. They find that the bacterial concentrations are increased and the fungal concentrations decreased in the new clustering method relative to the old. In general this manuscript is well written and represents a nice advancement of the field. I do have some specific comments as outlined below

We thank the reviewer for their careful reading of the manuscript and the helpful comments and recommendations which we address below.

Specific comments:

1. Is this WIBS-4 from Droplet Measurements in Boulder? If so, that should probably be stated somewhere. If not, what does the "4" signify and how is this instrument different from other published WIBS measurements?

The WIBS used in this study was designed and built by the University of Hertfordshire. 4 denotes the revision number. The differences between the WIBS-3 and WIBS-4 are discussed in Crawford et al. (2014).

2. I would like to know a bit more about any calibrations that might have been done for the instrument used in BEACHON. First, was there any independent verification that the size derived from the lookup table was accurate? It is hard to see in the log-scale plots for the lab data how closely the WIBS-reported size matches with the PSL size for the different test particles. Also, I believe the users manual for the DMT WIBS states that only side-scatter is utilized for sizing rather than the forward to side ratio. Please check on this.

The sizing of the instrument used in BEACHON was verified using a series of standard and fluorescent PSLs as described in detail by Robinson et al. (2013) and which was cited. The fluorescent PSLs used in this study feature some additional complex scattering patterns (i.e. non-spherical/irregular) in some cases as measured with a Multiparameter Bioaerosol Spectrometer which may cause small differences and increased spread in optical sizing not detectable by standard optical particle counters or WIBS. This feature was not observed for non-fluorescent PSLs and we can speculate as to why this is. The WIBS instrument used here uses the sizing method described in Kaye et al., (2005) and Gabey et al. (2010) which were also cited.

3. Similarly, were there any calibrations for the asymmetry factor measurement? Has the WIBS's ability to report a reasonable AF been determined for small sizes? I thought it only really worked for some of the larger particles and even then I thought it was relatively unverified.

Yes it has, This has been answered in our response to referee #2 which we repeat here; Corn starch flour was used to represent irregular particles and ellipsoidal haematite particles were used as an analogue for rod-like bacterial particles as described in Kaye et al. (2007); Gabey et al. 2010 describes the effect of size on AF which is now briefly described. 1 µm and 3 µm polystyrene latex spheres were sampled with a WIBS-3 where they found the modal values of AF to be 2-3 units greater for the smaller particles. They suggested that the noise in the quadrant PMT causes smaller particles to register slightly greater AF, however the influence is small.

4. It seems like there is quite a lot of detail on the instrumental side and less on the statistics. This may also be driven by my expertise (which is not computation or statistics) but I would like to know a little bit more about what a "linkage" is, what the different linkages mean and what the different normalization strategies are. I realize that these definitions are likely in the literature and textbooks but, especially given that one of the major outcomes of the paper seems to be that the Ward linkage with either z-score or range normalization is the best performer, it would be appropriate to have a brief explanation for the layperson in the paper itself.

We thank the reviewer for their suggestion. We will include a brief description of how Hierarchical Agglomerative Cluster Analysis works, including a description of linkages and normalisations used in the revised manuscript. The description of HCA and the linkages will be inserted at the start of section 3 which is now given:

"Hierarchical Agglomerative Cluster Analysis (HCA) has been demonstrated to be a powerful tool to classify particles (Robinson et al., 2013; Crawford et al., 2014; Gabey et al., 2013), however, the available analysis toolkits are limited by heavy computational burdens making the analysis of large datasets problematic.

In Hierarchical Agglomerative Cluster Analysis (HCA) each data point is initially in its own single membered cluster. The clusters are sequentially combined into larger multi-membered clusters until all data points are in one large cluster at the end of the process. At each step through the process the two clusters which are separated by the shortest distance are combined where the inter-cluster distance is determined by the linkage algorithm. In this study we trialled several common linkages which are now described:

Single: The distance between two clusters is defined as the minimum distance between any single data point in the first cluster and any single point in the second cluster.

Complete: The distance between two clusters is defined as the maximum distance between any single data point in the first cluster and any single point in the second cluster.

Average (unweighted average distance): The distance between two clusters is defined as the average distance between all data points in the first cluster and all data points in the second cluster. The weight of each cluster is proportional to the cluster size.

Weighted (weighted average distance): Similar to average but the weight of each cluster is identical irrespective of size.

Ward: This linkage is a special case where the clusters to be merged is determined by finding the pair of clusters which yield the minimum increase in total within-cluster variance after merging, rather than by minimum distance between clusters.

Centroid: The distance between clusters is defined as the distance between the centres (mean vectors) of clusters.

Median: The distance between two clusters is iteratively defined as the distance between the cluster midpoints. Here the midpoint is defined as the point itself in a singleton cluster or the average of the midpoints of the clusters to be merged.

A full mathematical description of these linkages is provided in Müllner (2013)."

The descriptions of the normalisations in section 3.4.2 will be revised to the following:

"In the Robinson et al. (2013) study the prepared data was z-score normalised prior to analysis. This was performed to minimise the effect of the different ranges of scale of each parameter biasing the clustering, i.e. the fluorescent intensities are of the scale 0-2092, Size 0.8-20 and AF 0-100. We investigate the effect of normalisation on clustering performance using the following standard procedures:

1. No normalisation.

2. Subtract mean, divide by standard deviation (z-score). The mean value of the normalised distribution is 0, where the z-score value of a data point is the number of standard deviations from the mean and this can be positive or negative.

3. Standardise by range. Subtract minimum value, divide by the maximum value. Normalises data to new range of 0-1.

4. Divide by sum. Divide each of the variables by its sum. The sum of the normalised distribution is 1. Since our original data is positive the normalised values will also be positive

5. Rank. Replace each data point by its rank. The data under this normalisation will be integers from 1 to N where N is the number of data points.

These are the procedures considered in Milligan and Cooper (1988) excluding procedures which produce identical results for the Euclidean metric. They concluded that the range normalisation to be the best performing. We considered procedures proposed by Gnanadesikan et al. (2007) which considered better performing alternatives to the above procedures. However it seems unlikely that the procedures will scale in terms of performance for large data."

5. It also seems really odd to me to only test the clustering using PSL spheres which are so obviously and easily differentiated by eye without any fancy analyses. Couldn't you test the clustering

performance with at least lab-generated bacteria and fungal populations? That would improve confidence in the ambient clusters greatly.

> The fluorescent doped PSL spheres were chosen to add additional diagnostics to assess the capacity of the tested methods to discriminate between the different samples as they can be readily nebulized with little background contamination and have unique fluorescent signatures, making them ideal for this purpose. Conversely the datasets of laboratory generated bacteria and fungal spores available to us have shown a strong spectral dependence on the growth medium used during the culturing process and the nebulised bioaerosol has contained other contaminants produced during nebulisation (i.e. contains nebulised growth medium). Further bioaerosol characterisation experiments with improved sampling methodologies are planned to overcome these sampling errors and this data will be used to investigate machine learning techniques, including the methods described in this study.

6. I would love to see some size distributions for the clusters from the BEACHON data. Does the "bacterial" cluster also actually look like bacteria in addition to "behaving" like bacteria? Similarly for the fungal clusters which I would expect to also have well-behaved size distributions at larger sizes than you see for the bacterial populations.

> We thank the referee for their suggestion and we will include the following figure in the revised manuscript to highlight the expected difference in size of the BEACHON clusters. It can be seen that the clusters attributed to bacteria are generally smaller than the fungal cluster as would be expected.

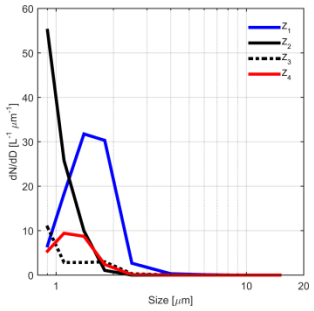

Figure 1. Size distribution of BEACHON Z-score normalised clusters produced using the Ward linkage for the period 00:00 to 06:00 27 July 2011.

7. I think you should be careful not to present the increase in bacterial concentrations and decrease in fungal concentrations with the new clustering methods as closer to "true" than the WASP parameterization. Right now these clustering methods are different statistical treatments with little "ground-truth" for either although this paper will likely convince the reader that the new methods are better. It would be best to simply describe how they differ and why you think that might be. Also, I believe the explanation for how the new clustering generates more bacteria-attributed particles is that WASP is miscategorizing some bacteria as fungi? But the fungal concentrations dropped by 10 /L while the bacterial concentrations increased by 80 or 90 per liter so reclassification of WASP fungal signals can only explain a minor fraction of the bacterial increase in the new algorithm. Are there also many many more unattributed particles in WASP?

We feel that we have been careful in our explanation of the differences between the two methodologies to not present the new method as "true". We have described how and why we think they differ as suggested. In the analysis we have presented we have covered the approach used in Robinson et al. (2013) which we have extended to provide a more generic sensitivity analysis of HCA linkages and normalisation methods. We agree to revise this section to make sure that this is clear to the reader as advised by the referee.

The new method generates more bacteria attributed particles by the inclusion of two more bacterially consistent clusters to assign particles to which are not present in WASP, rather than some bacteria being erroneously classified as fungi. This would mean that particles consistent with these missing clusters would be left unclassified by WASP leading to reduced bacterial cluster concentrations. Unfortunately WASP does not return diagnostic information about the cluster attribution, however, the sum of the concentration of WASP clusters B_3, C_3 and D_3 only accounts for approximately 24% of the fluorescent aerosol concentration suggesting that many particles are unattributed in WASP. We will clarify this in the revised manuscript in section 5.2.

Smaller technical comments:

1. You could use some references for the offline techniques in your "Detection methods" section of the introduction.

We will include references as requested, however this is not the focus of this study.

2. In figures 3 and 4 it would be nice if the colors were consistent for a given calibration particle.

We will standardise the colours across both figures in the revised manuscript.

3. Tables 2 and 3 seem not quite harmonized. The point in table 2 was that for large data sets the z-score normalization slightly outperforms the range but then in table 3 the range normalization looks better for all sample sizes tested.

The point in table 2 is that when using the full dataset, without any sampling, the z-score normalisation performed slightly better. However, the range normalisation seems to be much more robust to sampling of the data. We extended our original tests for table 3 to include samples of up to 90% of the data where the range normalisation outperforms the z-score normalisation on average as shown in table 3. This is only of concern if it is necessary to perform analysis on a sample of a very large dataset.

4. P 7316, line 19, I believe you mean that the range-normalized result has 4 clusters not 5.

We thank the referee for bringing this typographical error to our attention and we will correct this in the revised manuscript.s such that the range normalised results are described as having 5 clusters and the z-score normalised results as having 4.

5. I believe that on the right side of figure 7 the blue points represent Z1 vs B3? Also these labels are a little unfriendly. Perhaps also include in parenthesis the identities (bacteria and fungi) that you attribute to the sum of the clusters.

We thank the referee for bringing this typographical error to our attention and we will correct this in the revised manuscript. We will also include the identities in the label as suggested.

References:

Crawford, I., Robinson, N. H., Flynn, M. J., Foot, V. E., Gallagher, M. W., Huffman, J. A., Stanley, W. R., and Kaye, P. H.: Characterisation of bioaerosol emissions from a Colorado pine forest: results from the BEACHON-RoMBAS experiment, Atmos. Chem. Phys., 14, 8559-8578, doi:10.5194/acp-14-8559-2014, 2014. 7307, 7309, 7313, 7314, 7317, 7318

Gabey, A. M., Gallagher, M. W., Whitehead, J., Dorsey, J. R., Kaye, P. H., and Stanley, W. R.: Measurements and comparison of primary biological aerosol above and below a tropical forest canopy using a dual channel fluorescence spectrometer, Atmos. Chem. Phys., 10, 4453-4466, doi:10.5194/acp-10-4453-2010, 2010.

Kaye, P. H., Stanley, W. R., Hirst, E., Foot, E. V., Baxter, K. L., and Barrington, S. J.: Single particle multichannel bio-aerosol fluorescence sensor, Opt. Express, 13, 3583, doi:10.1364/OPEX.13.003583, 2005.

Robinson, N. H., Allan, J. D., Huffman, J. A., Kaye, P. H., Foot, V. E., and Gallagher, M.: Cluster analysis of WIBS single-particle bioaerosol data, Atmos. Meas. Tech., 6, 337-347, doi:10.5194/amt-6-337-2013, 2013.