Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-733-RC2, 2019

Atmospheric Chemistry and Physics
Discussions

Interactive comment on "A robust clustering algorithm for analysis of composition-dependent organic aerosol thermal desorption measurements" *by* Ziyue Li et al.

Anonymous Referee #2

Received and published: 7 October 2019

Note: This review is focused on the clustering aspects of the paper only. I make no comment on whether the generated clusters constitute a useful and valid contribution to the analysis of compositionâĂŘdependent 2 organic aerosol thermal desorption measurements as this is not in my area of expertise. I leave that for others to comment on.

From the title and abstract I was expecting to read about a new clustering algorithm. However, this paper presents a data pre-processing step before using DBScan, followed by an application-specific post-clustering process.

Summary The section on the clustering process should be re-written. This is not a new clustering algorithm, but a method of pre- and post-processing data to provide useful analysis. I do not deny the usefulness of the analysis, its methodology or results. There are also points below that should be clarified.

1. DBScan is compared to other clustering algorithms, however these alternatives are significantly different in their intended use, i.e. they generate hype-ellipsoidal clusters, whereas DBScan is specifically targeted at arbitrarily shaped clusters. It would be better to compare DBScan with similar algorithms. For the purposes of this paper, I am not sure that comparison to other algorithms is required.
2. Line 111 states 'a novel variant of DBScan'. The clustering algorithm used appears unchanged from DBScan, but rather a pre-processing technique of ordering the data is utilised before clustering.
3. Line 132 'absolute magnitude' is moot, magnitude has no direction.
4. The paragraph from lines 134-144 describes a process based on a number of factors which are not justified. This leaves the questions: why 100 points; why +/- 50 points; and why is an anomaly only outside of -3σ and not outside of $+3\sigma$ also?
5. Similarly, lines 171-182 use a smoothing over 35 points. It is not clear how this smoothing is carried out, e.g. a mean of 35 points? Is it +17/-18, or -35, or +35 points that are smoothed? If the thermograms have peaks around 40 points wide are we seeing the mean of the value in the peak? If so how does this correlate with 'retaining the peak shape'? I think this should be clarified.
6. Similarly, line 190 recommends a weighting of 4:1. How has this value been arrived at?

7. Lines 216-218 discuss the removal of noise. DBScan is considered deterministic for core points and noise. Noisy data would normally be identified and be members of clusters $< minpts$. Is there a danger that data identified with 'high levels of noise' is excluded when, despite the noise, it is similar enough to be included in a cluster?

8. Lines 226-235 appear to form the work being considered as 'a novel variant of DBScan'. This describes DBScan with no alterations, except to force the order of data to consider 'seed thermograms' first. This is a pre-processing stage of the data, rather than a novel clustering algorithm. DBScan is deterministic if data order is preserved. If data order is not preserved then it is deterministic for core and noise, with only border points varying in some cases. I'd be interested to see how the border points vary to justify forcing the data order. Are the results generated by forcing the data order better simply because they consistent with each run, whereas random initialisation is not? If so, is it possible to identify which thermograms change cluster and consider why?

9. Lines 240-255 describe the DBScan algorithm. I am unclear how this varies from standard DBScan. Cycling through those thermograms identified as 'seeds' could equally be done by ordering the data by order of noise, then using the data, in order, to run standard DBScan.

10. Line 240 refers to Figure 2. This is not a suitable method for presenting an algorithm and a formal pseudo-code should be used. This may help clear up any confusion over the similarities or differences between DBScan and the method proposed.

11. Lines 256 – 269 describe a 'second round' of clustering. This generates new data for each cluster in the form of a 'signal weighted average', presumably of the cluster members. A thermogram that is within ϵ of the average, but not already clustered suggests that it is a border point, but below the $minpts$ threshold for

inclusion? (I am unclear how a thermogram can be within ϵ of the average, but not within ϵ of $minpts$ of other cluster members?) This part appears to be a novel 'second stage', however I would not consider this to be a clustering algorithm in itself, but rather a post-processing step to tidy up 'stragglers', which is application specific.

12. Section 2.2.3 Describes a process for selecting an optimal ϵ value. The selection of ϵ is based on fuzzy terms such as 'small' and 'near the maximum'. Figure 2 shows a clear value of ϵ in this case, is this the same for all analyses?

13. I am also unclear from section 2.2.3 whether this selection of optimal ϵ is generic to all future datasets of this type, or whether this optimal selection process is required for each new set of data?