Atmos. Chem. Phys. Discuss.,
doi:10.5194/acp-2016-632-RC1, 2016

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Chemometric analysis of aerosol mass spectra: exploratory methods to extract and classify anthropogenic aerosol chemotypes" *by* Mikko Äijälä et al.

Anonymous Referee #1

Received and published: 21 September 2016

This paper presents the application of k-means clustering to AMS data recorded at Hyytialla. While techniques like PMF are more commonly used as a data reduction tool in the AMS community, clustering presents some interesting possibilities, specifically for the purposes of plume classification. The paper uses an interesting technique whereby PMF is used to screen out discrete plumes first and then clustering applied to the outputs, so that plumes can be grouped and examined free of the influence of background aerosols.

This work is relevant to ACP and there are a lot of good features in this manuscript, such as a thorough evaluation of different distance metrics and determination of the correct number of clusters, two things that are absolutely crucial and yet frequently

missed from some of the more naïve applications of clustering within atmospheric science. However, the paper is not without its weaknesses; it is written in a very rambling, overly conversational and at times woolly tone, which made it very difficult for me (as a reviewer) to get at the hard-and-fast science. Certain key details regarding the methodology are also not covered in sufficient detail. In short, I would say there is good science in here, but it does need more work to turn it into a good paper.

General:

This paper is very technical in nature and risks being outside of the scope of ACP and more suited to something like AMT (especially seeing as the authors seem to imply in a number of places that the technique needs more work). In order to remain in scope, I recommend that the abstract and conclusions contain more of an emphasis on the new insights to atmospheric science that this work has offered.

The language used in this paper is very conversational and more in the style of a magazine article or opinion piece, with the insertion of many words that serve no tangible purpose to the paper (a few are picked out in the technical comments). While this would mainly be considered cosmetic and probably not worth making too big a fuss over, the authors at times risk crossing the line to using 'weasel words', i.e. the insertion of adjectives that convey an opinion-based or otherwise unsubstantiated point to the reader. Examples of this would include describing the tools as "somewhat underused" on page 2 or the use of the CTOF as "advantageous" on page 4. This practice is to be discouraged in scientific literature, so I would suggest the authors generally revise the text to a more formal style, sticking to the facts as much as possible.

Also regarding presentation style, there is a tendency to start sections with a loquacious preamble describing previous work or setting out the need for a particular technique to be applied, however in a number of cases (see specific comments) this level of detail is completely unnecessary because many of these motivations are so well established in the literature they would be considered common knowledge in the com-

munity. While this too could be considered cosmetic, in places it seems that this writing is done at the expense of necessary technical detail. An example (described below) would be the extensive text devoted to comparing the CTOF with HRTOF (which could be handled by a decade-old citation) but insufficient detail on the unique features of the specific instrument used here.

Finally regarding writing style, the supplement is very rambling in its opening sections. While collating quotes from old textbooks would be a good primer for a non-specialist, I would tighten the text up a bit and focus more on what is specifically important for this work. I would also try to avoid repeating material that is already covered in the main article.

Contrary to what is frequently implied, this paper does not represent the first use of clustering applied to AMS data; people were trying it long before PMF was used, an example of which is the Marcolli et al. (2006) paper cited in both the manuscript and supplement (although it is incorrectly presented as an example of factorisation in the introduction). As a technique for analysing ambient data, clustering failed to gain traction within the AMS community (reflected in the low number of publications) because unlike SPMS, AMS mass spectra do not (generally) represent discrete events so therefore interpreting clustering outputs carries with it many inherent problems. While this paper addresses many of these limitations, the authors would do well to tone down much of the text (in particular in the introduction) that seems to work off the principle that the application of clustering to AMS data is completely new. The real novelty of this work is the combination of clustering with discrete PMF analysis to get at data from specific plumes, which should be better suited to clustering than the blanket application to all recorded spectra, so I would spend more time focusing on this aspect of the work when demonstrating novelty.

A general fundamental weakness with clustering as applied to AMS data, even as applied here, is that it is not capable of identifying individual components when a measured mass spectrum is composed of an indeterminate combination of different com-

ponents, rather than a single type. While this would not be the case here if individual plumes could be attributable to single point sources, it would not be true of plumes from mixed sources, which may occur with urban plumes consisting of a mixture of traffic, cooking, etc. However, this very fundamental limitation is not really discussed, in particular in section 3.4.2, where the analysis appears to have been approached from the perspective that discrete clusters representing these types should be expected. I would argue that good clustering for these potentially overlapping sources should not be expected and the fact that these are represented by 'weak' clusters should come as no surprise. It is also completely overlooked when suggesting that the variability within clusters could inform the a-values used in ME-2. The text should really take this whole issue into account better. I would note that the use of a fuzzy clustering algorithm (e.g. c-means) may at least partially overcome this issue, but this presents an entirely new avenue of work outside of the scope of this paper.

Specific comments:

Title: The title of the paper is very obtuse and says very little about the actual content. Recommend rewording as something that includes the words 'clustering' and 'plumes'.

The first two paragraphs of the introduction are a little superfluous considering how well established mass spectrometry and the need for associated data mining and reduction is within atmospheric science. Given that there is a wealth of articles already published in ACP on mass spectrometric data reduction, I would remove this text.

Pages 3-4: It is difficult to sell the CTOF as advantageous given that the more diverse variable set provided by the HRTOF would almost certainly lead to a better statistical treatment (this is pretty much said later in the manuscript). But then the authors shouldn't have to justify using a CTOF over a HRTOF because the choice of the specific pre-existing dataset is justified later in the paper and the focus of the paper is on the analysis technique anyway. More generally, given how well established both instruments are, it is really not necessary to describe the mass spectrometry technology in

this much detail; a simple citation of the literature would suffice.

Page 4: Regarding the differences between this instrument and a standard CTOF, is this described elsewhere in the literature? If so, these should be cited. If not, much more detail should be given here, particular as regards the helium bleed system, ideally with a technical schematic.

Page 5: A description of the diagnostic that lead the authors to be concerned of the airbeam linearity would be appropriate. Was the airbeam affected by the helium feature of the instrument?

Page 8: The criteria given for plume identification are very qualitative and therefore subjective. Can some quantities be assigned to any of the criteria, such as rise rate or duration? These would contribute to the general goal of a truly objective system of data reduction, even if it is not achieved here.

Page 8: The justification for using k-means seems a little overwrought. To be clear, k-means is not the most simple algorithm in existence (hierarchical agglomerative clustering can probably claim that), but it is nevertheless generally treated as the 'default go-to' algorithm by most people in absence of a reason to use anything else because of its simplicity of operation and low computational cost. It's difficult to see this being any different in this case, so it would be better to simply state that you chose k-means for this reason and that a comparison with other algorithms could be done as future work.

Page 10, line 11: Saying that rotational ambiguity is 'mostly avoided' is a strong statement. What evidence can the authors present to back this up?

Page 12: Why not use the same weighting function as the error model used to weight the PMF residuals?

Page 14, line 21: "We hope" is a very odd thing to say. Can the robustness of the method not be tested somehow?

Page 20: After all the discussion regarding the selection of the correct distance metrics for mass spectra during clustering (particularly in the supplement), why use Pearson's R here?

Page 24: Referring to a fundamental limitation of clustering as a technique (see general comments), the authors should take account of the fact that some plumes may hypothetically consist of a mixture of individual sources.

Section 3.4.3 seems rather long and tangential considering that it fails to reach a definitive conclusion. Given that this is by no means the first time amines have been reported at Hyytialla, I would shorten this section for the sake of brevity.

Page 29: Again, the authors fail to acknowledge that the within-cluster variability can be caused by the varying influences of different sources within mixed plumes. Following from this, the later statements that "...the variabilities implied by this study can be used as an indicator of what the likely magnitude of the underlying natural variability within the observed classes of aerosols..." and "...the natural variability within an aerosol type may be significantly larger than what is often allowed in conjunction with the constrained PMF/ME-2." should have the caveats added that this will only work if the plumes can be absolutely verified as being of a single source.

Page 29: A frequently-used tool for quantifying rotational ambiguity is the PFEAK parameter in PMF, yet this is not even mentioned. Why was this not used? This would seem particularly appropriate here because when looking at 2-factor solutions, the limitations of applying a global parameter to explore the solution space are significantly reduced.

Page 31: The comparison with PMF in the conclusions is extremely disingenuous because the authors fail to distinguish between the two very different data models employed by the different algorithms and the very different way in which they can be used. It also seems strange to compare these like this because the clustering technique used here relies on PMF to extract the plumes in the first instance. While clustering is good

at analysing discrete plumes, its data model cannot handle arbitrary mixtures, which can make up the majority of AMS data in many cases. It is wrong to say that clustering eliminates the problem of rotation (there may still be some rotation in the plume extraction part – the authors have not discounted this) or that the cluster variabilities can be used to estimate source variability (it will only work for pure-component plumes). I recommend reworking this section to focus on what scientific insights this offers in addition to PMF, rather than pitching the two against each other.

Technical corrections:

Page 3, line 21: What is so "regrettable" about a full review of the statistical techniques being out of scope? As a reviewer of an atmospheric science paper, I confess I was actually quite relieved it wasn't in there.

Page 4, line 9: The word "specimen" is a very peculiar choice and not necessary. Please remove.

Page 4: The need to process AMS data correctly is very well established in the literature, so the opening text of section 2.1.2 is unnecessary.

Page 8, line 22: "...to thoroughly evaluate events' satisfaction our selection criteria" makes no sense. Please revise.

Page 9, line 8: "Achilles heel" is inappropriate language, seeing as it is an inherent feature of the data model applied and associated constraints, not PMF specifically. Please remove.

Page 12, line 1: Remove "unfortunately"

Page 13, line 28: Remove "with its own unavoidable weaknesses". It's not possible to make this statement without an algorithm in mind.

Page 14, line 4: Remove the word "Obviously". It would not be obvious to a reader with no experience of this.

Page 20, line 6: Remove "on the other hand" but also generally check the wording of the sentence; I'm not 100% sure what it is that is being said.

Page 21, line 20: The use of the future tense in "We will call..." is again overly conversational in tone.

Page 22, line 19: Please be specific when referring to "the last hypothesis". I had to read this several times before I thought I understood it.

Figure 1: This needs extensive tidying up, specifically to avoid lines overlaying the axis labels of inset graphs. Also, I would use legends rather than referring to colours in the text.

Supplement: This would be much easier to follow if the figures and tables were presented alongside the associated text.

Page S4, line 15: Should be 'in practice' ('practise' is the verb form in UK English)

Interactive comment on Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2016-632, 2016.