Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., author comment AC1
https://doi.org/10.5194/amt-2022-16-AC1, 2022

Reply on RC1

Sungwoo Kim et al.

Author comment on "Comprehensive detection of analytes in large chromatographic datasets by coupling factor analysis with a decision tree" by Sungwoo Kim et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2022-16-AC1, 2022

We thank the reviewers for their careful consideration and recognition of the value of this work. Please find below our responses to all comments and associated changes to the manuscript. Original comments are included in **bold**, and changes to the manuscript are excerpted in *italics*.

Reviewer 1

1. Although the method developed by the authors can identify more than 10 times of analytes than the conventional one (manual inspection), what scientific problems can this method help us to solve? It seems that we still do not know most of the identified analytes. I think the limit of identified analytes in previous studies is primarily due to the lack of authentic standards. The other big problem in investigating organic aerosols is the lack of organic tracers specifically related to their sources and/or transformation. If the total ion chromatographs of environmental data were inspected manually based on individual m/z ratios, at least several hundred of analytes will be identified, but mostly unknown. Here, I have to admit that the method developed by the authors really simplified the time-consuming work for chromatograph inspection. But scientific problems should be solved to enlighten the significance of this work.

We agree that a lot of work remains to be done to identify the unknown compounds in the atmosphere, and indeed agree that the limit is in many cases the availability of authentic standards. We note that many tracers now commonly used by the community started out as components with unknown structure or origin (for example, C5 alkene triols that are commonly measured as isoprene oxidation tracers required significant dedicated effort to identify (Wang et al. 2005)). Recently, a focus of the atmospheric community of GC/MS users has been to build up libraries of these types of compounds and works show that unknowns can be useful tracers (e.g., multiple libraries available through Dr. Allen Goldstein at UC Berkeley: https://nature.berkeley.edu/ahg/resources/). We also highlight previous work that used correlation to known tracers to identify the likely sources of unknown compounds (e.g., Isaacman-VanWertz, G. et al. 2016), and in some cases used this information to quantitatively attribute sources of aerosol (Zhang, H. et al. 2018). Even in the lack of current identification, we therefore believe it is useful to integrate and investigate all analytes, even those that are unknown at the moment, and examine the data as a whole to determine if any minor components may serve as useful tracers or provide unique information.

Further, we agree that manual inspection can capture many of the analytes in a dataset as proposed, and a central goal of the current work is to increase the efficiency of this process. However, there are also advantages to the proposed process relative to manual inspection other than simple decreases in processing time (which is nevertheless a major advance of this work). Perhaps most importantly, the current work outputs not only the location of an analyte, but also a "clean" mass spectrum that can be used comparison to known libraries (and identification, though that is a rare case as discussed). By using multiple chromatograms to extract this spectral information, this output is more robust that could be achieved through manual inspection. Furthermore, by including multiple chromatograms in the analysis, components that are present in only one or one type of chromatogram are also cataloged, which might be overlooked in manual inspection (which is not practical to do thoroughly for many chromatograms). Lastly, as illustrated in Fig. 4, this method is able to resolve peaks that would otherwise be nearly impossible under manual inspection. This is useful in identifying low signal analytes which can yield new insight into sources and transformations of organic compounds in the samples.
For the reasons described here, we believe the method has significant value to address research questions, ranging of quantifying aerosol sources to a more complete understanding of the impacts of atmospheric processes on atmospheric composition. However, given the complexity of the analytical approach, we believe it is best to focus this manuscript on the approach itself and allow future work to focus on the scientific advances enabled.
The following sentences have been added to clarify the issue.
"Significant work remains to be done to identify the unknown compounds in the atmosphere, however many tracers commonly used by the community started out as components with unknown structure or origin. For example, C5 alkene triols that are commonly measured as isoprene oxidation tracers required significant dedicated effort to identify (Wang et al. 2005). Previous work has also been done wherein correlation to known tracers was used to identify the likely sources of unknown compounds (Isaacman-VanWertz, G. et al. 2016), and in some cases, this information was used to quantitatively attribute sources of aerosol (Zhang, H. et al. 2018). Therefore, despite the lack of current identification, we believe it is useful to integrate and investigate all analyte, and examine the data as a whole."

Wang, W. et al. Characterization of oxygenated derivatives of isoprene related to 2-methyltetrols in Amazonian aerosols using trimethylsilylation and gas chromatography/ion trap mass spectrometry. Rapid Commun. Mass Spectrom. 19, 1343–1351 (2005).
Isaacman-VanWertz, G. et al. Ambient Gas-Particle Partitioning of Tracers for Biogenic Oxidation. Environ. Sci. Technol. 50, 9952–9962 (2016).
Zhang, H. et al. Monoterpenes are the largest source of summertime organic aerosol in the southeastern United States. Proc. Natl. Acad. Sci. U. S. A. 115, 2038–2043 (2018).

2. To keep column conditions, the GC column close to the inlet might need to be cut for quite a few centimeters after a batch of sample analysis, and retention times of analytes will vary differently. In this case, will the method be able to match the same analytes from different chromatograms in different batches?

As the reviewer notes, aligning chromatographic signals is necessary before applying statistical data reduction methods such as principal component analysis (PCA) and positive matrix factorization (PMF) (Eilers 2004). In the present work, a coarse retention time adjustment was applied manually, with a subsequent fine-scale retention time adjustment that was conducted as part of the automated cataloging approach. The current implementation of this process does not include the initial global adjustment, which we agree would be necessary for application to chromatograms from different batches. Others have described relatively robust global retention time alignment methods that could be automated (e.g., parametric time warping, Eilers 2004), and we are actively

working to integrate such an approach into the peak cataloging software as a pre-processing step to make the entire process more automated. However, doing so would be primarily an implementation of previously demonstrated methods, so a detailed discussion is not included in this manuscript. Instead, we have revised the discussion on line 131 to make this issue clear.
"Each chromatogram is first aligned to the same retention time basis by using a small number of known compounds or introduced standards in each sample to define known retention times. *Strictly speaking, this preprocessing is not necessary for factor analysis. However, interpretation of the outcome of data reduction techniques such as PARAFAC(2) and PMF can be unreliable when chromatograms are used directly as input (Eilers, 2004; Van Nederkassel et al., 2006), as it may be difficult or impossible to determine if unaligned peaks in each chromatogram represent the same analyte. Chromatogram alignment may occur through manual adjustment by users or may be automated using any of multiple solutions (Eilers, 2004; Kassidas et al., 1998; Nielsen et al., 1998), but the cataloging approach described here is independent of the details of any such approach (a manual approach is used in this work), so details are not included."*

3. How does the developed method deal with background peaks like pollution, column blood, etc? How does the method perform field blank correction?

Multiple samples are analyzed simultaneously using this method and positive matrix factorization (PMF) extracts a commonly observed pattern from all the samples. With the assumption that a high enough number of factors were used for the analysis, any existing analytes with a significant level of signal should be identified as separate analytes by this method, regardless of whether such an analyte is a compound present in the sampled air or is a contaminant. It is up to the user on how to utilize the outcome analyte information. In some analyses our approach has been to include a chromatogram from a background or blank to identify components that are present in these background samples as well. The manuscript has been revised to clarify the issue.
"In contrast to other PMF applications, the primary goal in this work is not to optimally describe the complete data set, but rather to increase the number of factors to a point where even minor components are extracted as separate factors, even at the risk of over-fitting the data (which will be rectified by a subsequent decision tree). *With this approach, any existing analytes with a significant level of signal should be identified as separate analytes, regardless of whether such an analyte is a compound present in the sample or is a contaminant."*