Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2020-79-RC1, 2020

Atmospheric Measurement Techniques Discussions
Open Access

Interactive comment on "Estimating mean molecular weight, carbon number, and OM/OC with mid-infrared spectroscopy in organic particulate matter samples from a monitoring network" *by* Amir Yazdani et al.

Anonymous Referee #1

Received and published: 8 June 2020

The paper describes a new method for obtaining important characteristics of Organic Aerosol (OA), such as mean carbon number, molecular weight and organic-mass-to-organic-carbon (OM/OC) ratios, using mid-infrared spectroscopy (also referred to as Fourier transform infrared spectroscopy FTIR). The technique is applicable to spectra acquired non-destructively from Teflon filters used for particulate matter sampling and it is tested on a relevant set of samples (more than 800) coming from the Interagency Monitoring of PROtected Visual Environments (IMPROVE) network in US. The approach involves multivariate statistical analyses (namely Partial Least Squares Re-

gression – PLS) and classification by CART (classification and regression trees) applied on the absorbance profiles and linking them to molecular structures in OM. The multivariate statistical models are trained on calibration spectra prepared from laboratory standards and are then applied to the ambient samples. The results of the models are consistent with previous OM/OC values estimated using different approaches and with temporal and spatial variations in these quantities associated with aging processes, and different source classes (anthropogenic, biogenic, and burning sources).

This is an overall well-written paper even if in some parts (the description of statistical methods for instance) is quite hard to digest and follow and could be improved (look my suggestions below). The method is anyway innovative and informative and so the manuscript is in my opinion worth of publication on AMT after minor changes which are detailed below.

General Comments:

Section 2.4, P9: this section is quite difficult to follow: the steps of the analysis are not clear enough and there are for example some abbreviations not explained (i.e., what does "RMSE" means?) or misleading (i.e., PLS or PLSR?) and some definitions poorly explained. All this makes difficult to follow the statistical methodology, its steps and their meaningfulness. My suggestion is to rephrase the Section, spending time in clarifying the methodology and its steps to make sure the readers can follow your process.

An (incomplete) list of the misleading elements in the section is reported here:

-you define the partial Least Squares Regression as "PLSR", but then you use always "PLS" as abbreviation in the subsequent text. Please decide your favorite abbreviation and check for consistency;

- "RMSE" is not defined;

-readers not familiar with multivariate statistical analysis can find difficult to understand the concept of “different number of latent variable (LVs)”: please explain better what is a latent variable in the context of this analysis and/or the motivation for repeating the analysis with a different number of LVs;

- the unbalanced use of the word “model” (in this section but in general in all the text) makes sometimes difficult to follow the discussion and to understand the different steps of the methodology: the “model” is both the statistical analysis and its results, the calibration process as well as the complete process to determine molecular weight and number of carbon-atoms. The word “model” in the abstract and in the Introduction refers also to thermodynamics and chemical numerical models, making even more confusing the discussion. I suggest to use more carefully the word “model” distinguish between the different types of “models” considered. Other words like “regression” when you are talking of PLS or sometimes simply “analysis” can be used to clarify the steps.

Detailed Comments:

P3, L10: consider to add the article “the” before “spectrum”.

P4, L11 & L15: there are question marks inside the brackets: add reference or remove the symbols.

P4, L12: equation (1): consider to move the definition of ïĄ■ in a different row.

P6, L5-7: how long is the sampling time? Not clear, even if important to understand possible advantages/disadvantages of the technique. This is especially true because some sentences later (L13 as also shown in Figure 4) it is stated that rural samples have very low recovery. Is this problem possibly fixed by longer sampling time in rural/remote sites? Consider to add 1-2 sentences discussing this here or in the conclusion as a suggestion to make the methodology more robust also in non-urban sites.

Figure 4: what is the number inside the histogram’s bars representative for? I suppose it is the number of samples of each category, but this should be described explicitly in the caption.

Section 2.2, P6, l8-: is the choice of the laboratory standards linked to natural abundance of species and/or functional groups? Or what is the rationale in the choice of the laboratory standards? Looking at table 1, why for example only one species of dicarboxylic acid has been tested? Or why only Fructose and not Glucose or Galactose? Or other Sugars with different numbers of C-atoms/molecular weight? I can understand that the choice is made also based on availability of standards and of already existing spectroscopic data, but this should be acknowledged better in the text in my opinion.

P9, L28: “to the classify . . .”, please remove “the”;

P15, L12: “The is not a concern. . .”, not meaningful sentence, probably misspelled;

P16, L2-3: “we used all laboratory standards to produce PLS models to applying to ambient samples”, here maybe a passive form is needed. Please, replace with “we used all laboratory standards to produce PLS model to be applied to ambient samples”.

P22, L7: other inconclusive question marks. Please replace with the number of figure or explain.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2020-79, 2020.