Atmos. Meas. Tech. Discuss., 7, C4548–C4556, 2015
www.atmos-meas-tech-discuss.net/7/C4548/2015/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Predicting ambient aerosol Thermal Optical Reflectance (TOR) measurements from infrared spectra: organic carbon" *by* A. M. Dillner and S. Takahama

A. M. Dillner and S. Takahama

amdillner@ucdavis.edu

Received and published: 13 January 2015

Interactive comment on "Predicting ambient aerosol Thermal Optical Reflectance (TOR) measurements from infrared spectra: organic carbon" by A. M. Dillner and S. Takahama R. Subramanian (Referee) randomsubu@gmail.com

Response to the reviewer's comments and added text are shown immediately following each reviewer comment. The revised supplemental material is uploaded. No changes were made to the figures so they are not included here.

1. This is an interesting submission that aims to complete the chemical composition

possible from a single teflon filter, which is useful to air quality monitoring networks. The idea is that a single filter can be used to provide PM mass, metals (with XRF), BC (optical methods have been developed, but are not widely used), inorganic composition (destructive methods), and now with this manuscript, OC using non-destructive FT-IR. So it is very relevant to AMT.

Response: Thank you for this nice framing of our work.

2. However, I have one major concern. The authors have ïňĄxed the calibration data set at 2/3 of the field data. This is not the parameter I would have chosen to keep constant; it is also not how the real world application would work. That is because if to get the OC data from one hundred samples, I have to calibrate the technique with two hundred samples, then what is the cost advantage of using FTIR? If I understand correctly, the calibration set is 2/3 of the ïňĄeld data set. That seems excessive, and signiïňĄcantly advantages the model. The authors find that when aerosol composition is different - higher OM/OC for test set than for calibration set, or apparently at Phoenix, where OM is less oxidized - the errors are higher. Figure 8 shows that the normalized error can be 20-25% at a single site (Olympic), even when data from all other sites is available to train the method, and the OC distributions are not signiïňĄcantly different between the calibration and test sets. This is even when the authors have two-thirds of the data set to train their model! What happens if samples are ordered at random, and the ïňĄrst one-third or even ïňĄrst one-tenth of the samples are used for calibrating the method? What would be the errors and bias in the remaining test samples then? This is how I would expect the real world application to play out. So in essence, the authors need to answer the question: Is an empirical calibration is required for each new data set, whether from a new site or a different year?

Response: We neglected to discuss how this work would be applied in the real world which led to your concerns. We will discuss that here and have added text to address this. This paper is intended to show feasibility of the method and is a first step towards proposing a method for a monitoring network. It is not intended to be the final model

for the network nor do we expect to use 2/3 of every sampling site for every year to develop a model for the other third. The method and proportion of sample selection in this work follows a statistical convention for evaluation of a proposed model. With this approach, we demonstrate the possibility to reproduce an operationally defined "species" using FTIR and for understanding what is important in model development in terms of sample composition for the calibration set. We expect that once a model is developed, it will be applied to all samples in the future (if a monitoring network choses to implement the method) so that the incremental cost of analysis will be small. The following text has been added to more clearly describe the purpose of our work.

Abstract: This work marks an initial step in proposing a method that can reduce the operating costs of large air quality monitoring networks with an inexpensive, non-destructive analysis technique using routinely collected PTFE filter samples which, in addition to OC concentrations, can concurrently provide information regarding the composition of organic aerosol. This feasibility study suggests that the minimum detection limit and errors (or uncertainty) of FT-IR predictions of TOR OC are on par with TOR such that evaluation of long-term trends and epidemiological studies would not be significantly impacted. Introduction: This work is the first step in proposing a non-destructive method for reducing sampling and analysis costs for large particulate speciation monitoring networks. The method also provides a means of obtaining information about the carbonaceous aerosol at sampling sites that have only Teflon filter samples provided that new samples have similar aerosol composition to the samples in the calibration set.

Conclusions:

Future work includes establishing that the calibration developed using samples from one year can be used to predict TOR OC during other years and developing a calibration that includes samples with a broader range of aerosol composition.

In establishing the feasibility of this method, we followed the common practice of using

2/3 of the samples for calibration (Arlot and Celisse, 2010; Hastie et al., 2009). Section 2.3 has been revised to state this more clearly. As you mention, this does advantage the model and this is done intentionally to obtain the best model for these samples and future samples. However, to evaluate how many samples are needed to represent the variability present in this data, we developed several calibrations with fewer samples in the calibrations. We determined that including 250 filters in the calibration set gives results similar in quality as using over 500 (2/3 of the total number of filters). This number is not an absolute number but only pertains to this variability of samples used in the calibration and test sets. Text is added to section 3.1 in the paper and text and a figure are added to section S5 of the supplemental material. We will explore this further when we have a larger data set. Rather than selecting samples for the calibration set at random, we have used the information we have from the IMPROVE network and other sources to test extreme cases (using low OC samples to predict high OC samples) to determine what causes error and bias in the calibration. This also provides a basis for understanding how well samples from new sites and a different year will be predicted (i.e., if a new site or different year has similar distribution of OC, OM/OC and ammonium/OC, it will be well predicted).

Section 2.3

We follow the common approach of using 2/3 of the total filters in the calibration set (Arlot and Celisse, 2010; Hastie et al., 2009) for the "base case" (described in the following paragraph) and other cases used to evaluate which parameters impact prediction quality. Section 3.1

Additional calibrations are created using fewer samples in the calibration set and the error in the test set is independent of the number of samples in the calibration set as long as there are at least 1/3 of the total samples ($\sim$250 samples) in the calibration set (see section S5 in supplemental material) indicating that the calibration is robust with respect to the number of samples used to calibrate between 1/3 and 2/3 of the sample set. The number of samples is not, however, an absolute number but is dependent on

the specific set of samples in the calibration and test sets. Supplemental section S5 added text. For added figure, see supplemental material.

The number of samples in the calibration was varied from 507 samples (2/3 of the total samples) to less than 100 to evaluate the quality of calibration with decreasing calibration samples. Figure S5b shows that the mean error for the truncated and baseline corrected cases increase significantly when the number of samples falls below 200 and 100 respectively. For the raw case, the mean error is less sensitive to the number of samples although the error is highest when the number of samples is lowest. The MDL is independent of the number of samples in the calibration set.

The question posed: "Is an empirical calibration required for each new data set whether from a new site or a different year?" is the next step in developing the calibration model. We have begun to address this question in this paper by analyzing individual sites without including samples from the site in calibration (section 3.5). This analysis shows good results when the aerosol composition is similar for the new site (the site not included in the calibration) but has additional errors when they are different, and these interpretations are supported by the preceding section of the manuscript ascribing prediction errors to particular differences in aerosol loading and composition. Errors at sites with low EC (for example Olympic, figure 8) are on the same order as TOR collocated error for the same masses so this is not a limitation of the FTIR method but of the TOR method. We will further address this question in a separate paper by applying the calibration we developed in this paper to hundreds more samples collected two years after the samples used in this paper, some from the same sites as this paper, some from others. We will evaluate how well the model works without using any of the new samples to address the question of samples collected "in the future" and improve upon it if needed and possible. Finally, we will be analyzing one year of samples from the entire IMPROVE network for developing a calibration for the network. This will provide more insight into how many samples are needed to capture the variability in OC and the associated aerosol composition for 170 sites.

2. Minor subquestion: which type of spectra do the authors recommend, since they have evaluated three types?

Response: All spectral types work equally well. Since using full spectra is easiest, that would be the simplest to use.

Other speciïňĄc comments:

3. Introduction - can skip some of the details of the two other details, because it is not clear those are relevant to the study at hand beyond "similar work has been done in other ïňĄelds."

Response: This is the first time in the aerosol community that FTIR and PLSR have been used to predict data obtained from costly, destructive methods for field samples although it is done somewhat routinely in other fields. We feel it is important to reference and describe the works in related fields to establish the credibility of this method.
4. Ruthenberg et al. (2014) seems more relevant. But if R2014 already determined OM and OM/OC ratios with FTIR, doesn't that lead to OC? So what's new in this paper? The authors could help the reader understand the difference between R2014 and this submission, instead of discussing the two works from unrelated ïňĄeld "inexpensive" - compared to TOR OC?

Response: Ruthernberg 2014 and others FTIR methods can calculate OC from functional group measurements based on laboratory standards. Text has been added to clarify this in the introduction.

These studies use laboratory-generated standards as reference material to develop calibration models for quantifying functional group abundance which can be used to calculate OC and OM.

This method calibrates to TOR OC to predict TOR OC data. The value of predicting TOR OC is that there is an historical record of that data that is used for trends analysis and epi studies. Discontinuing that measurement all together would have large conse-

quences on these types of studies. We have added text in the concluding paragraph of the introduction (shown above) to state this.

5. Can the authors provide a relative estimate? TOR OC is direct, and only requires someone who can take a punch from a ïňĄlter and stick it in a TOR oven - not very specialized tasks. FT-IR analysis appears much more specialized.

Response: FT-IR is similarly simple. The whole Teflon filter is placed into the instrument and then a scan is collected. With our current setup, it takes about 5 minutes to collect one spectrum. Text has been added to section 2.2.1 stating the time required for analysis and that no sample pre-treatment is done. Another cost savings is the elimination of quartz filters, sampler and sample handling of the quartz filter.

Each sample or reference spectrum takes about 1 minute to collect so that the total analytical time per filter is about 5 minutes. No sample pre-treatment is performed.

6. Why is OC mass for the blank ïňĄlters assumed to be zero? The values are available from the monthly mean blanks.

Response: The TOR OC values for the blank Teflon filters are assumed to be zero because these are laboratory blanks with no corresponding quartz blanks for obtaining TOR OC. The text has been changed to state that these are laboratory blanks in section 2.1 and in section 2.3 where the statement is made that the OC is assumed to be zero.

7. Sec 2.4: “Evaluation of the quality of calibration” - I expected to actually see an evaluation of the calibration here, not just how the calibration is evaluated. The title should be revised to “methodology for evaluating the quality of calibration.”

Response: We changed the title to Methods for evaluating the quality of the calibration.

8. Sec 3.3: “ammonium is often correlated with OC” - while this may be true for fairly regional/aged pollution, I would not expect it to be the case for urban environments. Second, the authors should provide references for this statement (which could also prove me wrong.)

Response: We have removed that statement.

9. If I understand correctly, the authors do not measure ammonium, but assume nitrate and sulfate exist as NH4NO3 and (NH4)2SO4. How good is this assumption across all sites?

Response: IMPROVE does not measure ammonium so it is difficult to evaluate the validity of this assumption. This assumption is used routinely in the IMPROVE network for mass closure across all sites and the sufficient correlation between gravimetric and the sum of the species suggests that, on average, this is a reasonable one. It is an underestimate when the aerosol is not fully neutralized and an over estimate when there are significant quantities of other ammoniated species present. However, we expect that for the purpose of our study, the errors in this assumption will not significantly alter our evaluation.

10. Could the authors use TOR EC or an optical measure of BC from TeïňĆon ïňĄlters to perform a similar sensitivity analysis? The EC would be an indicator of the primary vs secondary nature of the sample aerosol.

Response: We used OC/EC to do a similar sensitivity analysis which is in the supplemental material. It shows that OC/EC sensitivity is similar to OM/OC.

11. Please clarify non-uniform cases - for example, the lowest 2/3 of what parameter predict the highest 1/3? Similarly for the other non-uniform cases. Response: This has been clarified in the following text in section 3.3.

Three non-uniform cases are also considered: samples with TOR OC in the lowest 2/3 of the TOR OC range are used to predict samples with TOR OC in the highest 1/3 of the TOR OC range (Non-uniform A), samples with the highest and lowest 1/3 TOR OC are used to predict samples in the middle 1/3 TOR OC (Non-uniform B), and samples with the highest 2/3 TOR OC are used to predict samples with the lowest 1/3 TOR OC mass (Non-uniform C). Similarly, three non-uniform cases are modeled for OM/OC and

Ammonium/OC.

12. The ïňĄgures and analysis/write-up could be made simpler and direct by showing just one recommended spectra type (raw, baseline-corrected, or truncated), as the results do not signiïňĄcantly change based on spectra type. The discussion of the remaining spectra types can be relegated to the supplemental information for the very interested reader. What spectra type do the authors recommend?

Response: Including all three spectra in the results highlights the robustness of the method and provides the foundation for future work in which one or the other spectra may be used by IMPROVE or other networks due to operational or other considerations.

References

Arlot, S., and Celisse, A.: A survey of cross-validation procedures for model selection, Statistics Surveys, 4, 40-79, 2010.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition, 2nd ed., Springer Series in Statistics, Springer, 745 pp., 2009.

Please also note the supplement to this comment:
http://www.atmos-meas-tech-discuss.net/7/C4548/2015/amtd-7-C4548-2015-supplement.pdf

---

Interactive comment on Atmos. Meas. Tech. Discuss., 7, 10931, 2014.