Supplement of Atmos. Meas. Tech., 12, 3885–3906, 2019
https://doi.org/10.5194/amt-12-3885-2019-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

Classification of iron oxide aerosols by a single particle soot photometer using supervised machine learning

Kara D. Lamb

Correspondence to: Kara D. Lamb (kara.lamb@noaa.gov)

Supplementary materials

This Supplementary contains additional information and figures to support the conclusions and discussion in the main manuscript.

In addition to the random forest approach discussed in the main manuscript, we initially considered other machine learning algorithms, including linear discriminant analysis and the multi-layer perceptron implementation in the sci-kit learn Python package (Pedregosa et al., 2011). While the neural network approach was promising (and generally demonstrated only slightly worse performance to the random forest we have chosen to focus on here), we had less success in applying linear discriminant analysis, likely because the large variability within classes for some of the features.

Figure S1 shows the normalized histograms of the color ratios for the laboratory samples used as the training/testing data set. These histograms indicate that ambient aerosols (mainly composed of rBC) have a greater prevalence of particles with a slightly lower color ratio than the fullerene soot samples. These differences in distribution of features in the laboratory samples used to train the algorithm can impact the application of the trained algorithm to atmospheric observations of rBC, as discussed in greater detail in Section 4.4.

Figure S2 shows the size distributions of the laboratory samples of Fe_2O_3 and of Fe_3O_4 sampled by the SP2. For comparison, fits to the size distributions of FeO_x previously sampled at a ground-based site in East Asia is shown (Yoshida et al., 2018). These distributions indicate that the nebulized iron oxide samples were predominately greater than 300 nm, even though ambient observations of FeO_x demonstrated a number distribution peaked below 200 nm volume equivalent diameter. We have assumed 5.17 g/cc for the void free density of the iron oxide samples.

Figure S3 demonstrates the mass to color temperature ratio relationship for fullerene soot measured on three separate occasions with the NOAA SP2. In each of these three cases, between 6,000 to 13,000 particles were sampled, and the SP2 optical head had been independently aligned. A greater variability in color temperature ratio as a function of incandescent peak height was observed when the alignment of the PMTs was not ideal. This misalignment was not evident from the single particle signals, but could be observed in the relative distribution of color ratios as a function of incandescent blue amplitude (mass) for fullerene soot (See Figure S3a). This dependence suggests that color temperature ratio variability for fullerene soot samples can be used to optimize alignment of the red and blue detectors in preparing the modified SP2 to detect different types of incandescent aerosols.

Table S1 summarizes the pre-processing of the features. We have chosen to take the natural logarithm of features that are log-normally distributed, as machine learning algorithms typically perform better with normally distributed values for the features. This table also shows the imputation of values in cases where the features derived from the single particle SP2 signals do not have valid values, as the machine learning algorithm requires values for every element in the feature vector. For features that do not have valid values because the detector is saturated, we chose a value higher than the maximum possible value for that

feature. For the other features, we have chosen to impute a dummy value outside the typical range of values for that feature. These choices are somewhat arbitrary, so as discussed in Section 3.2, we also explored using a mean value for the features instead, and saw little difference in classification performance.

Table S2 and Table S3 provide the classification accuracy for each of the 3 class and 6 class cases, respectively, after optimizing the algorithm using only the n most important features. Precision and recall are defined in Section 4.3 in the main manuscript. We tested n =17, 9, and 5 features for the 3 class case, and n =17 and 11 features for the 6 class case. In each case, the selection of features was based on the ranking of the relative importance of features, as given in Table 3. The effective fraction of the samples for which a feature contributes to the decision pathway can be used as an estimate of the relative importance of that feature; averaging over multiple randomized trees (as in a random forest) improves the accuracy of this estimate (Pedregosa et al., 2011).

References

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, Journal of machine learning research, 12, 2825–2830, 2011.

Yoshida, A., Ohata, S., Moteki, N., Adachi, K., Mori, T., Koike, M., and Takami, A.: Abundance and emission flux of the anthropogenic iron oxide aerosols from the East Asian continental outflow, Journal of Geophysical Research: Atmospheres, 2018.

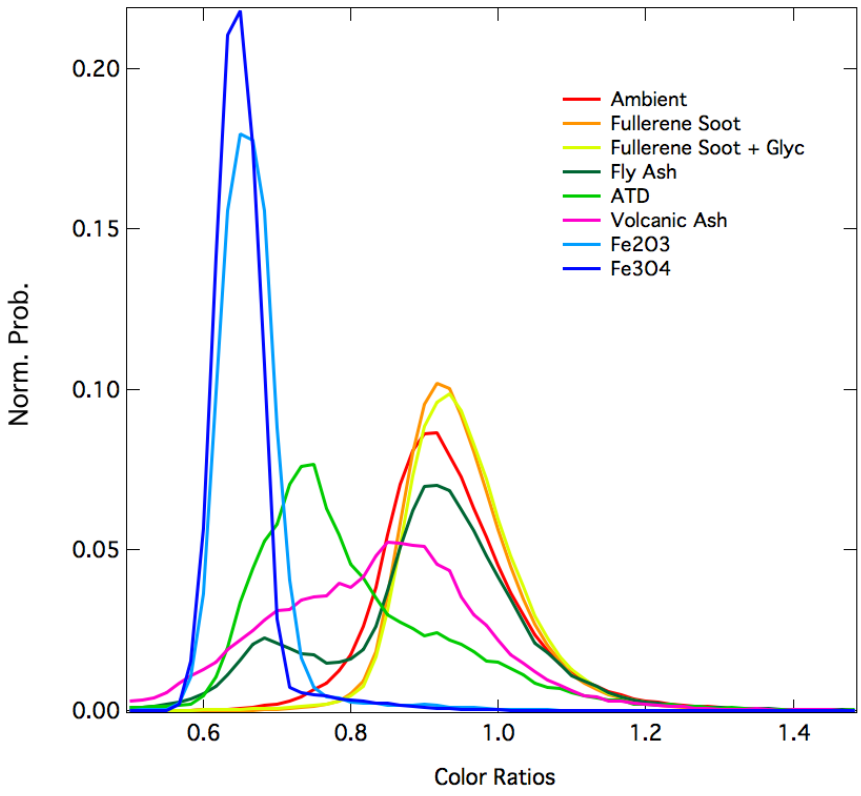

Figure S1. Normalized histograms for the color ratios of laboratory samples and ambient aerosols shown in Figure 2 in the manuscript.

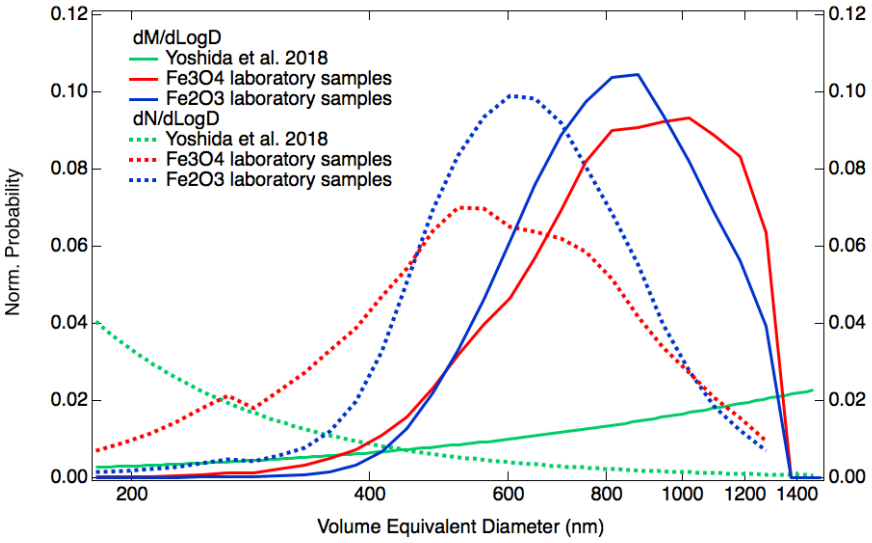

Figure S2. Normalized dM/dLogD and dN/dLogD size distributions for the iron oxide samples observed by the SP2, compared with ambient observations of FeO_x in East Asia

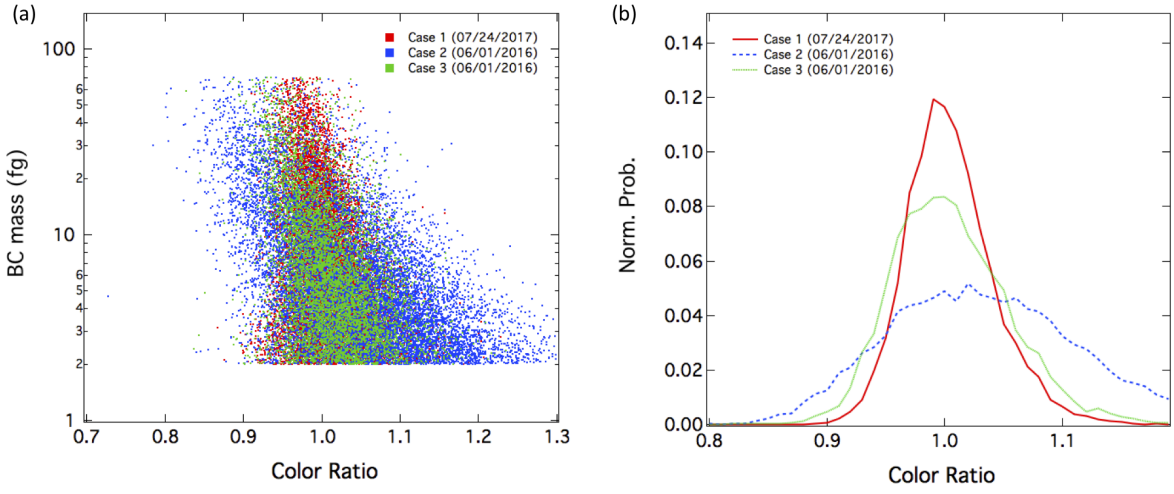

Figure S3. Influence of detector alignment on the incandescent blue amplitude (mass) to color temperature ratio relationship for fullerene soot samples (a) Mass vs. color ratio for fullerene soot sampled by the NOAA SP2 on three different occasions, with three independent alignments for the blue and red PMT's. The color ratio in each case was normalized to 1.0 for fullerene soot with a mass of 10 fg. Greater variability in color ratio was observed when the detectors are not well-aligned (as in case 2) (b) Normalized histograms of the color ratios for fullerene soot for particles with masses between 2 and 70 fg for the three different optical alignments demonstrate differences in the width of the distributions.

Table S1. Summary of preprocessing steps for each of the features considered in this study

Symbol	Feature	Pre-processing	Imputation for particles without valid values
x_0	Blue peak amplitude	Natural Log.	Particles without valid values were rejected
x_1	Color ratio		Particles without valid values were rejected
x_2	Core scattering	Natural Log.	110% of maximum possible value
x_3	Total scattering max.	Natural Log.	110% of maximum possible value
x_4	Post incandescent scattering		Dummy value - low negative value
x_5	Evaporation scattering size	Natural Log.	110% of maximum possible value
x_6	Position sensitive wideness		Dummy value - low negative value
x_7	Min. scattering before incandescence		Dummy value - low negative value
x_8	Position sensitive trigger position		Dummy value - low negative value
x_9	Scatter peak location		Dummy value - low negative value
x_{10}	Saturation width		Dummy value - zero
x_{11}	Incandescent start position		Dummy value - low negative value
x_{12}	Evaporation point		Dummy value - low negative value
x_{13}	Incandescent total length		Dummy value - zero
x_{14}	Incandescent used length		Dummy value - low negative value
x_{15}	Light on laser intensity	Natural Log.	110% of maximum possible value
x_{16}	Width fraction from center		Dummy value - low negative value

Table S2. Summary of classification accuracy for the 3-class case, using different subsets of feature space. We provide classification accuracy for the optimized algorithm when using only the n most important features for the 3 class case.

Class	17 features		9 features		5 features	
	Precision	Recall	Precision	Recall	Precision	Recall
rBC	0.94	1.00	0.93	0.99	0.84	0.99
dust-like	0.97	0.76	0.96	0.73	0.91	0.36
FeO_x	0.99	0.99	0.99	0.99	0.98	0.99

Table S3. Summary of classification accuracy for the 6-class case, using different subsets of feature space. We provide classification accuracy for the optimized algorithm when using only the n most important features for the 6 class case.

Class	**17 features**		**11 features**	
	Precision	**Recall**	**Precision**	**Recall**
rBC	0.94	1.00	0.94	1.00
ATD	0.80	0.64	0.79	0.61
VA	0.81	0.66	0.81	0.66
FA	0.75	0.58	0.73	0.56
Fe_3O_4	0.92	0.99	0.91	0.99
Fe_2O_3	0.87	0.50	0.87	0.50