Supplementary material for:

Chemometric analysis of aerosol mass spectra: exploratory methods to extract and classify anthropogenic aerosol chemotypes

Mikko Äijälä[1], Liine Heikkinen[1], Roman Fröhlich[2], Francesco Canonaco[2], André S.H. Prévôt[2], Heikki Junninen[1], Tuukka Petäjä[1], Markku Kulmala[1], Douglas Worsnop[1,3] and Mikael Ehn[1]

1 Department of Physics, University of Helsinki, Helsinki, Finland
2 Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, Villigen, Switzerland
3 Aerodyne Research Inc., Billerica, MA, USA

correspondence to: Mikael Ehn (mikael.ehn@helsinki.fi)

S.1 PMF – on robust mode and rotational ambiguity

Since the amount of weight given in PMF to an observation (here: the value of a variable *i* at a certain time *j*) by the iterative process is proportional to the square of $E_{i,j} / \sigma_{i,j}$, outliers with abnormally high squared signal or low variance may end up dominating the model solution. This phenomenon is especially relevant in environmental observations, as there are several types of outliers that would conceivably cause this behaviour, such as errors in the functioning of the measurement instrument or extreme, rare events that are considered contamination from the point of view of the analysis.

Therefore a "robust mode" for PMF was introduced (Paatero, 1997). The approach in short is to introduce a limit α, for the weight given to a point ($E_{i,j} / \sigma_{i,j}$) beyond which the point is considered an outlier, and dynamically down-weighted to negate its disproportional effect on the objective function *Q*. For a complete explanation on outliers and the robust mode, we refer the reader to the original work (Paatero, 1997).

The main weaknesses of PMF and indeed most factor analytic or linear algebraic methods are: 1) The "rotational ambiguity" of the solutions, i.e. the existence of multiple, sometimes very different, mathematical solutions with equally high rate of explanation of the observed (weighted) variance (Paatero, 1997; Paatero et al., 2002). Exploring the rotations and selecting the best solution from the "solution space" needs to be done by the analyst, often based mainly on interpretability of the results in the context of the particular research topic at hand. 2) The selection of number of factors, *f*. While exploring the rate of decrease of *Q* when increasing *f* can be considered an indicator of the amount of factors present in the set of data (Paatero and Tapper, 1993; Ulbrich et al., 2009; Reff et al., 2007), it rarely gives unambiguous answers. In the end it is up to the analyst to decide *f* based on both the diagnostics offered by *Q* and the interpretability of the result. These two subjective selections are often considered the most debatable part of a factor analysis (Ulbrich et al., 2009; Reff et al., 2007; Kim and Mueller, 1978). An additional constraint of the method is that is regards the chemical composition of a factor invariable, and

as such is less than ideal for atmospheric conditions where physicochemical processes constantly alter the aerosol composition (Canonaco et al., 2015).

In this work we utilise PMF in a non-standard way, to resolve the time series and mass spectral profiles explaining "anomalous" observations often discarded from a PMF analysis: the periods with air pollution spikes and plumes. PMF analysis is done for each air pollution event individually, altering the time window of the analysis around the event to include both the pollution episode and some background before and after the event. The advantage of studying this type of short relatively short term phenomena is, that we can easily evaluate fulfilment of the criteria outlined in Sect. 2.2.2, and we can additionally discriminate between mathematically equal solutions, mostly evading the issue of rotational ambiguity. Essentially knowing beforehand what the (qualitative) temporal behaviour of a pollution and background factors should be like, (*i.e.* the time series of the factors should be uncorrelated), we explore the number of factors and the solution space to select the solution best fulfilling our criteria for a physically correct solution. Adhering to these criteria, we strive to minimise the ambiguity related to our selection of solutions, as well as considerably reduce the effect of subjectivity with regard to selection of solutions.

We note the inclusion of the robust mode, hard coded in our user interface of choice (SoFi 4.8; Canonaco et al., 2013) is a potential issue for events with temporally very short plumes of only a few time points, but our testing confirmed it did not noticeably hinder the algorithm finding the expected, physically realistic solutions. It was noted if the solution returned by the algorithm was not driven by the pollution plume-like temporal behaviour, the time window of the analysis was often too wide, and applying a narrower window reproduced the plume factor in most such cases. In the few cases, when despite using a narrow window an acceptable solution was not reached, the event was discarded from further analysis to avoid any errors in extraction of pollution features.

S.2 k-means clustering, parameter selection

The iterative method to achieve this operates as follows:

1. (Initialisation) In the very first step, called initialisation, a pre-set amount of k starting cluster centres are defined: this can be done in a number of ways, but often involves random selection. Distances between each object x and every cluster centre c are calculated
2. (Step 1: Assignation) All sample objects are assigned to their nearest cluster centre, based on the selected metric of distance (dissimilarity). Every object now belongs to exactly one cluster.
3. (Step 2: Updating) Cluster centres c are re-calculated as a mean μ of the now updated cluster members x_i of each cluster. After updating the cluster centres the assignation is performed again (move to step 1).
4. (Convergence) This cycle is repeated until no more changes are made in the assignments in consecutive iterations. The algorithm has reached a convergence.

The input parameters required for the operation of k-means include:

1. Number of clusters, *k*. This value must be selected by the user. It can be based on *a priori*, external information revealing the number of clusters to be expected present, or in an exploratory analysis it can be set *a posteriori*, based on diagnostics values indicating the quality of solution for various *k* values and/or based on analyst expert opinion (reasonability and interpretability of the result). This requires calculating clustering solutions for a range of *k* values.

2. A metric for distance used in calculation of $J(C_n)$. Typically Euclidean squared "distance" $\|x_i - \mu_n\|^2$ is used, (as in Eq. (5), main text), but other option also exist. In lieu of an actual distance between two objects, a metric describing the similarity or (conversely dissimilarity) of the said objects may be used instead (Anderberg, 1973).

3. Initialisation cluster centres. They can either be selected by the user or randomly chosen. Random selection can be obtained sampling from among all the objects at random (or uniformly at random), or by performing pre-clustering with a subset of data (*e.g.* 10% of the objects selected randomly) and using these as initialisation for the final clustering. The selection of initialisation, (along with the number of random repetitions of the clustering) may influence the likelihood of finding the global minimum of *J(C)* instead of a local one.

Regarding initialisation, it has been shown by Arthur and Vassilvitskii (2007) that selecting the initialisation points not uniformly at random but spreading them out via stepwise selection from a weighted distribution improves the performance of the k-means algorithm. We also adopt this k-means initialisation method, resulting in an algorithm commonly known as "k-means++".

As there are no general rules for the selection of metric or number of clusters, but they rather depend on the type of data and application at hand, we will experimentally study their effects and try to select the settings fitting the classification of AMS mass spectra.

S.3 On spectral (dis)similarity

When applying clustering methodologies, the analyst's choice of an appropriate distance or dissimilarity metric, constructed from the variables studied in the analysis, is inherently dependent on the class of data variables (i.e. nominal, ordinal, interval, ratio scale) on hand. When dealing with aerosol mass spectrometric results, the data consists of a set of "ratio scale" variables, i.e. fractions of signals at various *m/z* ratios that together form a mass spectra (an "observation" or "object"). Ratio scale measurement fulfil the strictest requirements of quantitative measurements (i.e. the unit difference between two values is meaningful as is the zero point of the scale), and contain the most information of these metric categories (Kaufman and Rousseeuw, 2009). They also impose the least restrictions on the choice of a distance metric.

A significant benefit of the AMS as an instrument is, that the mass spectrometric signal is well quantified and linear. All the variables in our data adhere the same units and scale, so they can be considered homogenous and harmonised relative to each other. While the variables are not completely independent, when dealing with fractions of total signal, the dependence is generally not dominant over the actual variations in the signal. Typically for the AMS the dynamic range of individual normalised signals is quite moderate compared to many other mass spectrometers, which partly avoids the problem of large signals completely dominating the clustering outcome – and also with the aforementioned issue with the signal fractions being co-dependent. Still, we are far from immune to these effects, and will discuss and test the effect signal intensity and possible scaling options.

Authors of classic textbooks such as Anderberg (1973), Kaufman and Rousseeuw (2009) and Spath (1980) also discuss the dissimilarity metrics' theoretical background. It is noted *e.g.* by Anderberg (1973) that the correlation and cosine methods are very closely related, and their advantage is they are invariant to uniform multiplicative scaling, unlike the Euclidean and cityblock distances. For mass spectra this means if a spectra B is a scalar multiplicative of spectra A (as a thought experiment we disregard breaking of normalisation here), they are rightly considered identical by the cosine and correlation methods while the cityblock and Euclidean distances would find them different. With proper normalisation, this is less of a problem in practise, although, there are cases when this could conceivably cause issues. As an example, consider a case when we have two otherwise identical spectra, but in the other we let us say double *m/z* 44 Th, diminishing the other signals respectively via the normalisation. The single difference in aerosol chemistry would now be considered the change of oxidation level, but the Euclidean distance method would now find all the *m/z* signals to be dissimilar, the correlation and cosine methods would (perhaps more correctly) consider the dissimilarity increased only with regard *m/z* 44 Th. The feature is further exacerbated by Euclidean distances' proneness to giving excessive weight to outlying values of single variable (Cormack, 1971). We therefore hypothesise the Euclidean distance will discriminate the AMS spectra more with regard to their highest signals, such as 44 and 43 Th ions often linked to aerosol oxidation level (Aiken et al., 2007; Aiken et al., 2008), and possible outliers. Conversely, cosine and correlation would rather focus on the (dis)similarity of the higher end of the spectrum. This hypothesis is supported in light of the results in Sect. (3.3), where it is shown the squared Euclidean algorithm does best in sorting aerosols types by their oxidation state and finds outlier groups, but struggles to separate classes with minor variation in higher up spectral structures.

On the more subtle differences between the closely related methods of cosine and correlation, Anderberg (1973) states:

> *"[…]the distinction [between cosine and correlation] is precisely the difference between ratio and interval scale variables, respectively. Thus, the cosine makes use of ratio scale information, while the correlation coefficient only uses interval scale information.",*

and recommends cosine to be used when the origin is meaningful and well established, which is the case with our AMS mass spectra (and derivatively the PMF results). This can be considered as a good argument for selecting the cosine metric. Anderberg (1973) additionally notes the correlation metric is invariant to any linear transformations, such as uniformly adding a constant to all the elements (here: spectrum *m/z*'s) and therefore less discriminating, which in the case of clustering

isn't a favourable quality, than the cosine metric. Previous AMS clustering studies (e.g. Marcolli et al., 2006) have also utilised a "dot-product" similarity metric

$$d(u, v) = 1 - u \cdot v, \quad \text{(S.1)}$$

which when normalized by the vector lengths, as is done for our mass spectra from PMF results, becomes exactly equal to the cosine metric (Eq. 7).

There also exist experimental evaluations for the performance of different metrics, the most comprehensive known to us being the aforementioned work of Stein and Scott (1994), focusing on evaluating the (dis)similarity metrics used to automatically identify mass spectra. The compared metrics include the dot product (cosine) metric and the Euclidean distance, and finds the dot product metric to perform highest in matching the primary NIST library spectrum and an alternative spectrum from the same compound, with 75% identification accuracy. Euclidean distance is the runner-up among the five metrics included, with 72% accuracy. (The three others being absolute value distance (68%), probability based matching (68%) and the Hertz et al. method (64%); Stein and Scott, 1994). Stein and Scott therefore conclude dot product (cosine) to be the best algorithm for mass spectra matching.

Additionally, in connection to the hypothesis about oxidation level indicator signals presented above, and in the spirit of Stein & Scott's mass weighting rationale *("[mass weighting] deemphasizes the more variable and less characteristic lower mass range in a spectrum and emphasizes the more informative higher mass ions near the molecular ion*") we also tested completely omitting the oxidation-sensitive signal range, below 45 Th, with proper re-normalisation, of course, to see if we can alter the basis of discrimination in the clustering. The results proved intriguing, as can be found in Figure S.1, but were considered similar to the ones derived with the more elegant mass weighting processing, which we consider preferential over the omission method. The silhouette maximum seen at k = 4 (Figure S.1, left panel) for *m/z* < 45 Th omission derives from discrimination against outlier groups, discussed in main text (Sect. 3.4.3) and was therefore seen as not incorporating new information over the selected method of mass scaling.

S.4 Additional methods for evaluating clustering solutions

The alternative methods of solution quality evaluation were briefly examined. Evaluation results and references are presented in Figures S.2 to S.4.

S.5 Notes on silhouette value

From the main defining Eq. (11) it follows that

$$-1 \leq s(i) \leq 1. \quad \text{(S.2)}$$

The definition of *s(i)* in this fashion is invariant to multiplicative matrix operations, *e.g.* multiplying the distances by a positive number, but not additive operations such as adding a positive constant to the distances. The silhouette method not

only provides a metric for the robustness of a clustering solution but also introduces a graphical display for the description of n-dimensional results, an immensely helpful feature for an analyst when having to judge the "goodness" of a particular solution offered up by the clustering algorithm. A typical way to display silhouette values is a horizontal bar plot, such as the graph presented for the solution "corr" $k = 8$ of this study in Figure S.5.

S.6 Posteriori weighting by silhouette value, effect on mass spectra

Posteriori weighting by silhouette values had minimal effect on cluster centroids, but did affect calculated intra-cluster variability values somewhat. The difference between unweighted and weighted cases is shown in Figure S.6.

S.7 Mass and intensity weighting, effect on solution quality

Scaling effect on mass spectra is further illustrated in Figures S.7 to S.9 below.

S.8.Clusters' cross correlations and diagnostics values

Some of the main diagnostics values and correlations for the clustering result are presented in Tables S.1 to S.3.

S.9 Wind direction dependences and diurnality of pollution events:

Average wind directions at 0-64 m height and local time at peak concentration was recorded for all events, and is shown in Figure S.11. Sawmill-SOA seems connected to directions with lumber mills (see Liao et al., 2011) and the A-SV-OOA similarly to the Juupajoki/Korkeakoski direction. HOA seems to originate from the direction with the nearest road (western sector), and the A-LV-OOA that we find connected to wood burning, is mostly seen with incoming air passing over the Hyytiälä forestry station buildings, saunas and the nearby cottages and houses. COA directionality is already more evenly distributed, but includes the forestry station as one major direction of origin. For the outlier spectra there are too few data points available, but we note there is a cattle farm to the east direction, and agricultural fields to the south, which would be possible sources of the amine compounds. Also the forest clear-cut area is to the south.

As for the diurnality analysis presented in a histogram in Figure S.12., the results are inconclusive at best. For sawmill-SOA the most peaks arrive after midnight, but include observations at all times of a day. For A-LV-OOA and A-SV-OOA the observations are more frequent during night time. In the case of A-LV-OOA domestic heating and cooking in the evening may play a part. HOA plumes occur mostly daytime, while the COA (potentially mixed with fresh BBOA) is evenly distributed. The few amine plumes occur night-time or in the early morning, potentially suggesting the compounds could be semi-volatile and therefore sensitive to diurnal temperature changes. Again, all of the aboveis speculative due to statistically

too low sample sizes. Also it should be noted these are times when the plume is observed at the receptor location (SMEAR II), and therefore delayed by an unknown amount of time from the actual time of the emission.

S.10 Local and regional sources identification

To support source identification, wind direction analysis (Figure S.11) was combined with geographical information (Figures S.13 to S.16) of the area nearby SMEAR II (Hyytiälä) measurement station.

S.11 On additional important dimensions driving the clustering results

The source-indicative *f55:f57* dimension is what we interpret separates the sawmill aerosol type from the other semi-oxidised aerosol types with similar O:C ratios and *f44* contributions (WI-III, A-SV-OOA types). While we should not draw stretched conclusions based on the separation of this one instance of a biogenic source, we find the sawmill aerosol type is additionally characterised by the high 53 to 57 Th ratio (or "*f53:f57*"); the biogenic sawmill aerosol has very low *m/z* 57 Th contribution, while the relatively high 53 Th signal clearly sets it apart from e.g. COA with sometimes similarly low *f57* but high *f55*. We find a ratio of *f53*:*f57* > 2 seems indicative of aerosol originating from the SOA conversion of sawmill monoterpene emissions (Figure S.17).

As for the HOA vs COA separation, we feel equally conclusive indicators as *f55:f57* of the same division are found higher up in the mass spectrum *m/z* scale; the odd *m/z* value structures from 65 to 71 Th and 77 to 85 Th seem to offer equally good markers. Specifically we suggest a low ratio of *f65* to *f71* and *f85* to *f77* can be used as an alternative a marker for HOA as opposed to COA (possibly mixed with fresh BBOA) (Figure S.18).

References

Aiken, A. C., DeCarlo, P. F., and Jimenez, J. L.: Elemental analysis of organic species with electron ionization high-resolution mass spectrometry, Analytical Chemistry, 79, 8350-8358, 2007.

Aiken, A. C., Decarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., and Sueper, D.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, Environmental Science & Technology, 42, 4478-4485, 2008.

Anderberg, M. R.: Cluster analysis for applications. Monographs and textbooks on probability and mathematical statistics, in, Academic Press, Inc., New York, 1973.

Arthur, D., and Vassilvitskii, S.: k-means++: The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, 1027-1035.

Caliński, T., and Harabasz, J.: A dendrite method for cluster analysis, Communications in Statistics-theory and Methods, 3, 1-27, 1974.

Canonaco, F., Crippa, M., Slowik, J., Baltensperger, U., and Prévôt, A.: SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data, Atmospheric Measurement Techniques, 6, 3649-3661, 2013.

Canonaco, F., Slowik, J., Baltensperger, U., and Prévôt, A.: Seasonal differences in oxygenated organic aerosol composition: implications for emissions sources and factor analysis, Atmospheric Chemistry and Physics, 15, 6993-7002, 2015.

Cormack, R. M.: A review of classification, Journal of the Royal Statistical Society. Series A (General), 321-367, 1971.

Davies, D. L., and Bouldin, D. W.: A cluster separation measure, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 224-227, 1979.

Kaufman, L., and Rousseeuw, P. J.: Finding groups in data: an introduction to cluster analysis, John Wiley & Sons, 2009.

Kim, J.-O., and Mueller, C. W.: Introduction to factor analysis: What it is and how to do it, 13, Sage, 1978.

Liao, L., Dal Maso, M., Taipale, R., Rinne, J., Ehn, M., Junninen, H., Äijälä, M., Nieminen, T., Alekseychik, P., and Hulkkonen, M.: Monoterpene pollution episodes in a forest environment: indication of anthropogenic origin and association with aerosol particles, Boreal environment research, 16, 288-303, 2011.

Marcolli, C., Canagaratna, M., Worsnop, D., Bahreini, R., De Gouw, J., Warneke, C., Goldan, P., Kuster, W., Williams, E., and Lerner, B.: Cluster analysis of the organic peaks in bulk mass spectra obtained during the 2002 New England Air Quality Study with an Aerodyne aerosol mass spectrometer, Atmospheric Chemistry and Physics, 6, 5649-5666, 2006.

Paatero, P., and Tapper, U.: Analysis of different modes of factor analysis as least squares fit problems, Chemometrics and Intelligent Laboratory Systems, 18, 183-194, 1993.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, Chemometrics and intelligent laboratory systems, 37, 23-35, 1997.

Paatero, P., Hopke, P. K., Song, X.-H., and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, Chemometrics and intelligent laboratory systems, 60, 253-264, 2002.

Reff, A., Eberly, S. I., and Bhave, P. V.: Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods, Journal of the Air & Waste Management Association, 57, 146-154, 2007.

Spath, H.: Cluster analysis algorithms: for data reduction and clasification of objects, John Wiley & Sons, 1980.

Stein, S. E., and Scott, D. R.: Optimization and testing of mass spectral library search algorithms for compound identification, Journal of the American Society for Mass Spectrometry, 5, 859-866, 1994.

Tibshirani, R., Walther, G., and Hastie, T.: Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 411-423, 2001.

Ulbrich, I., Canagaratna, M., Zhang, Q., Worsnop, D., and Jimenez, J.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmospheric Chemistry and Physics, 9, 2891-2918, 2009.

Figures and Tables

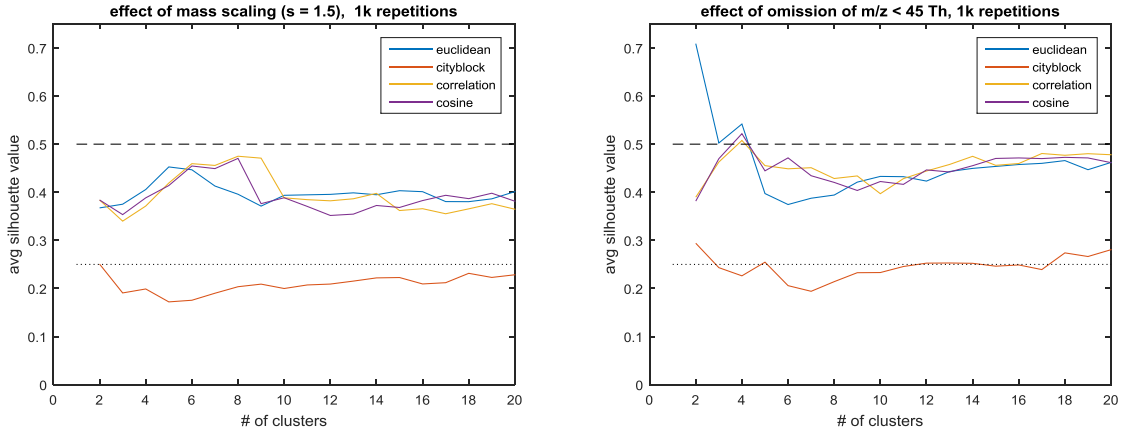


Figure S.1. Silhouettes for clustering solutions over *k* values from 2 to 20. Left: dataset processed with mass scaling. Right: *m/z* < 45 Th omitted. These curves can be further compared to the case of the original set of data (Figure 2), which can be considered a baseline for any pre-processing tests.

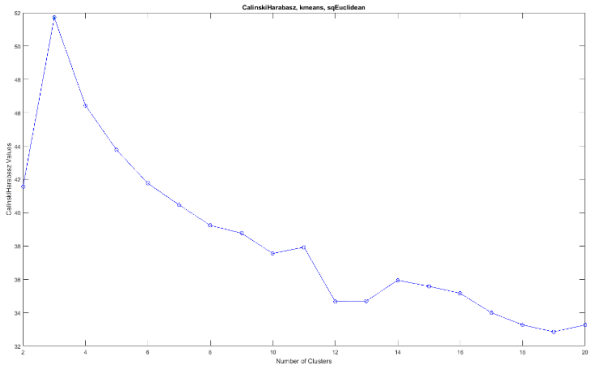


Figure S.2. Calinski-Harabasz criterion; (suggests max value *k* = 3; Caliński and Harabasz, 1974).

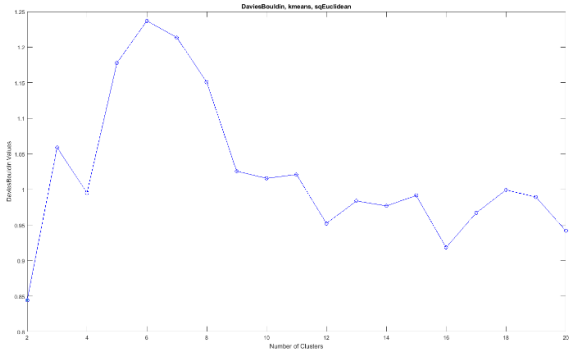


Figure S.3. Davies-Bouldin criterion; (suggests min value *k* = 2; Davies and Bouldin, 1979).

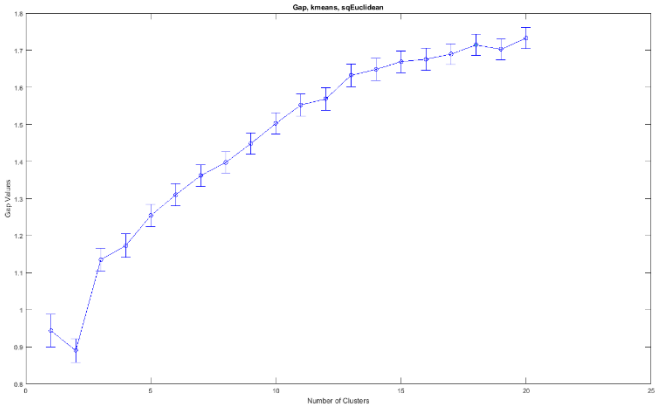


Figure S.4. Gap criterion; (suggests elbow value *k* = 18; Tibshirani et al., 2001).

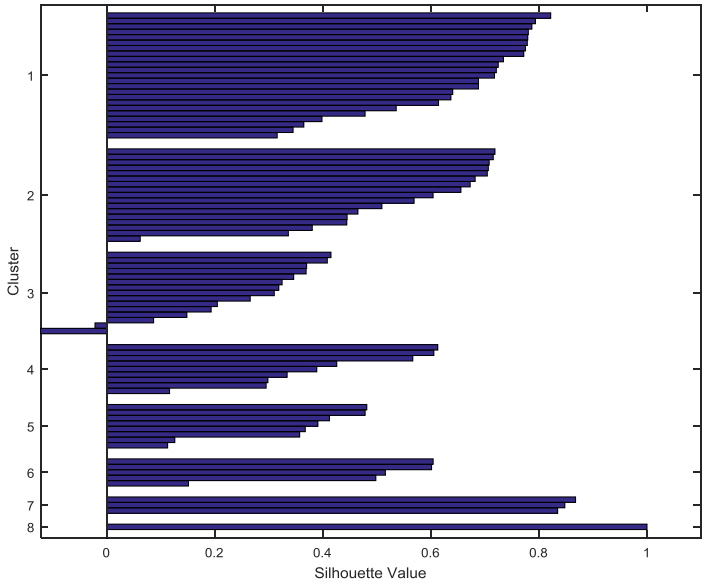
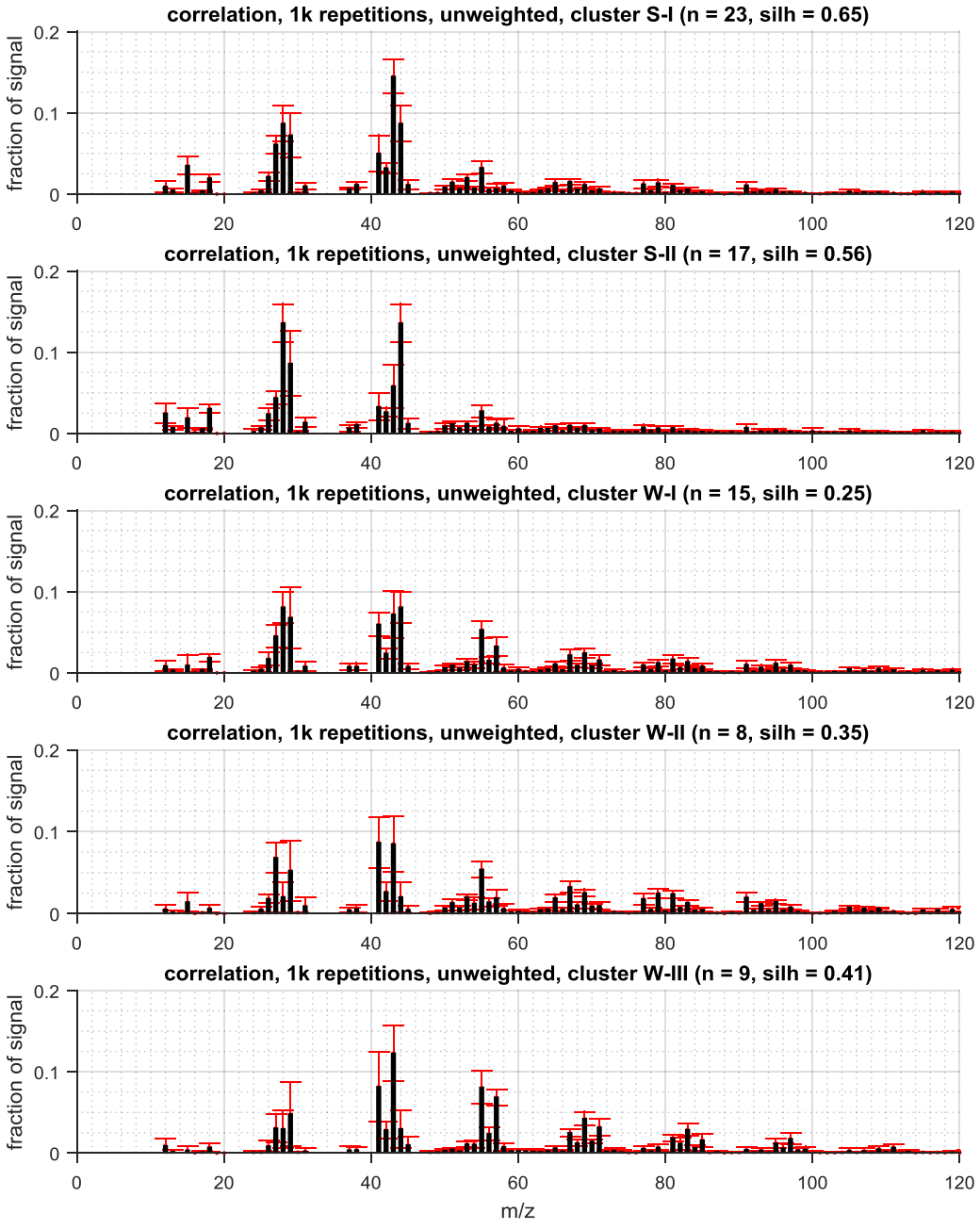


Figure S.5. Silhouette plot for "corr" *k* = 8 (s_m=1.21) solution. Numeric clusters 1 – 8 correspond (from top down) to clusters S-I, S-II, W-I, W-II, W-III, O-I, O-II and O-III, interpreted further in Sect. 3.4.


correlation, 1k repetitions, unweighted, cluster S-I (n = 23, silh = 0.65)
correlation, 1k repetitions, unweighted, cluster S-II (n = 17, silh = 0.56)
correlation, 1k repetitions, unweighted, cluster W-I (n = 15, silh = 0.25)
correlation, 1k repetitions, unweighted, cluster W-II (n = 8, silh = 0.35)
correlation, 1k repetitions, unweighted, cluster W-III (n = 9, silh = 0.41)
fraction of signal
0
0.1
0.2
0
20
40
60
80
100
120
m/z

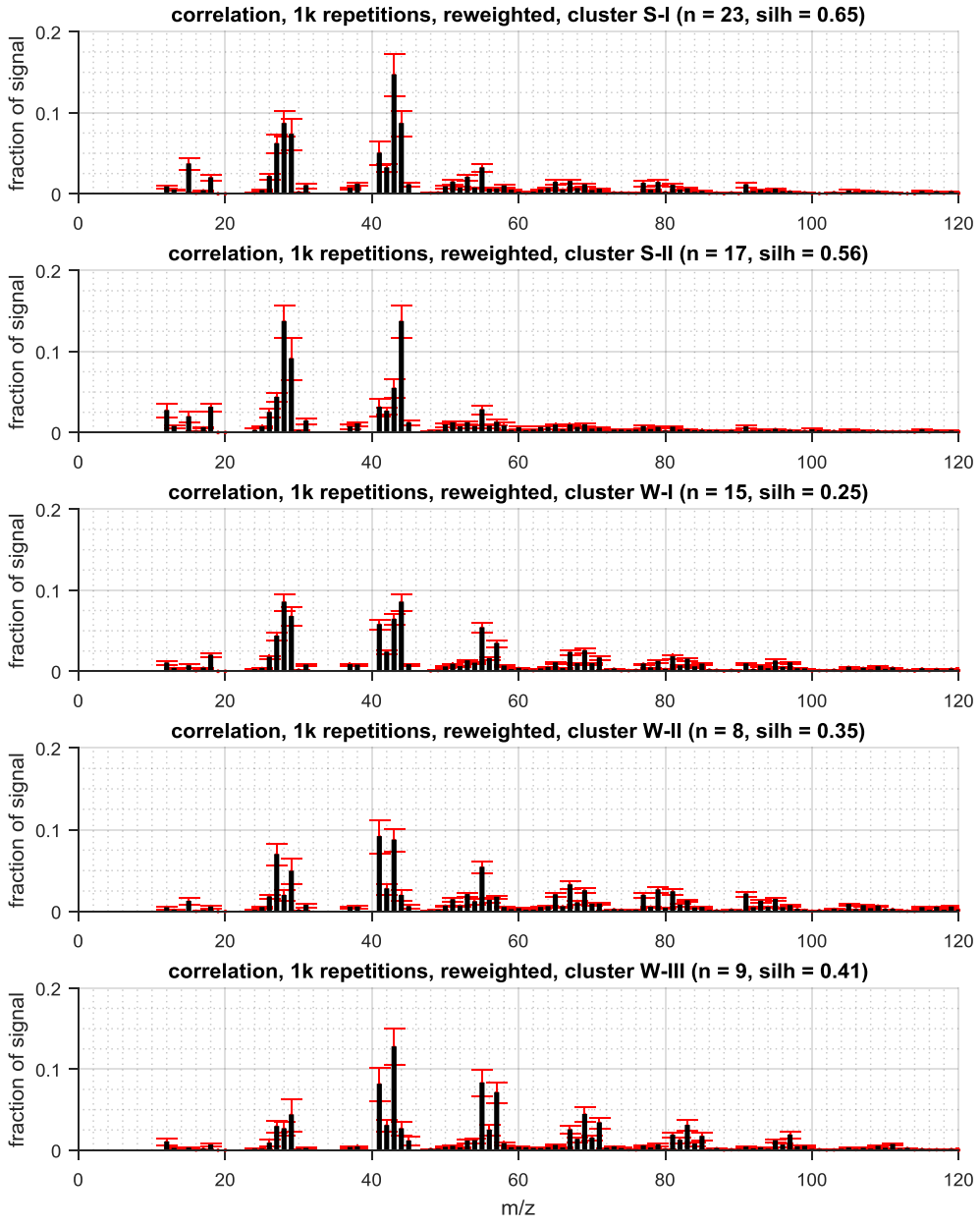


Figure S.6. Effect of cluster centroid weighting by silhouette for the S-X and W-X clusters' mass spectra. Above (5 uppermost spectra): unweighted cluster centroid spectra (mean of objects in cluster). Below (lowest 5 spectra), similar spectra weighted by the silhouettes of the objects. Correlation between the respective clusters is effectively unity ($r_s^2 > 0.994$), so the main difference is seen for the intra-cluster variabilities, depicted by the error bars.

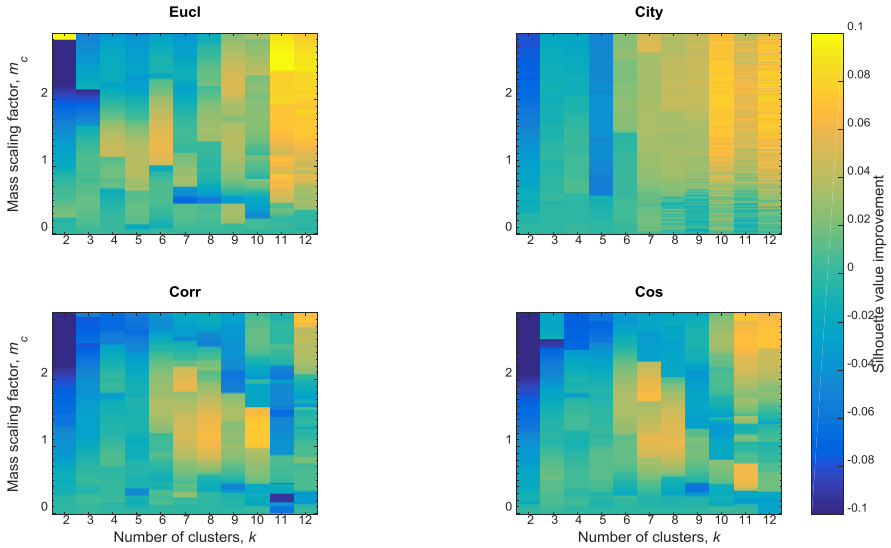


Figure S.7. Improvement in (solution) absolute silhouette values for mass scaling.

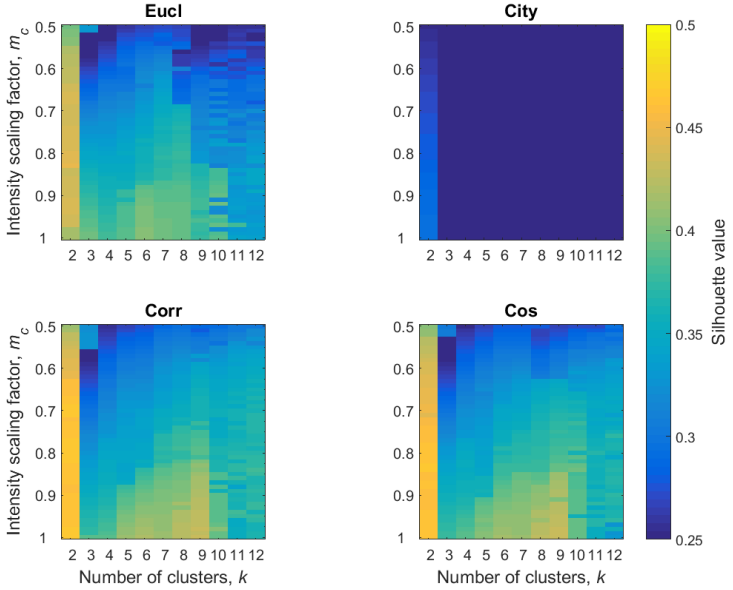


Figure S.8. Solution absolute silhouette values with intensity scaling.

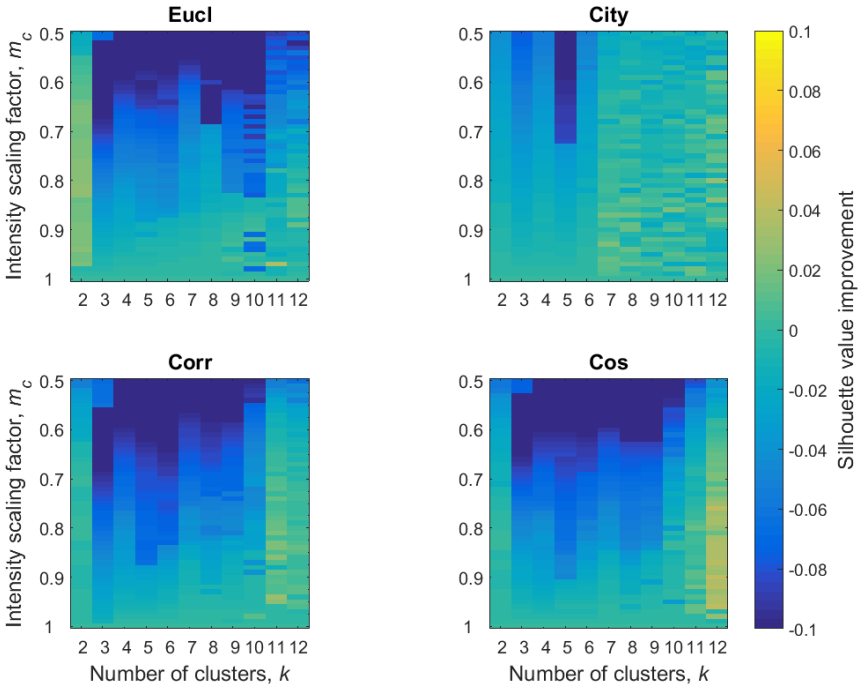


Figure S.9 Silhouette improvement / degradation for intensity scaling.

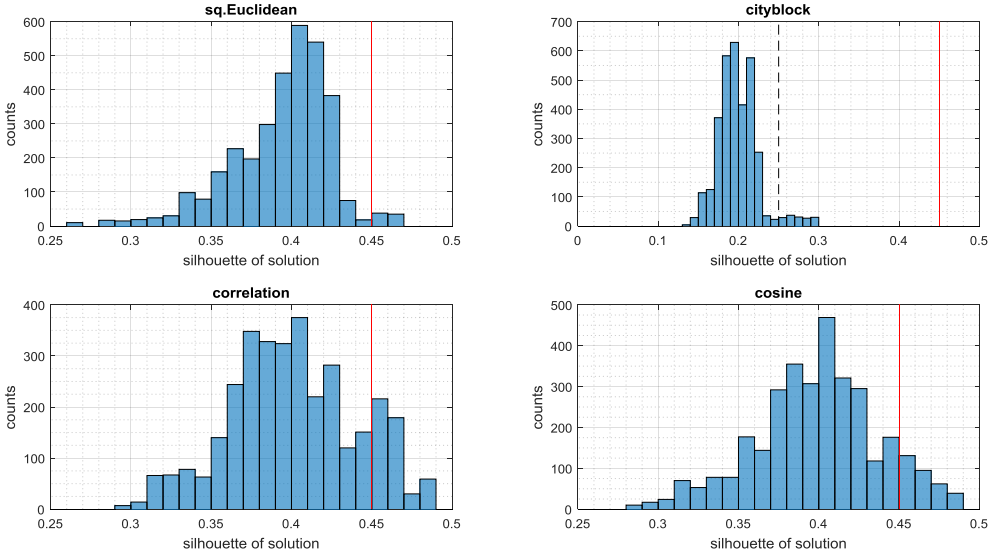


Figure S.10. Histograms of total solution silhouette value distributions for mass scaled data, corresponding to results presented in Figure 3. Value of 0.45 (red line) was chosen as a lower limit for solution to be included in a more detailed examination (Sect. 3.3; Table 2). Note: for "cityblock" metric the x-axis scaling differs – the minimum x-axis value of other panels, 0.25, is marked with a dashed line.

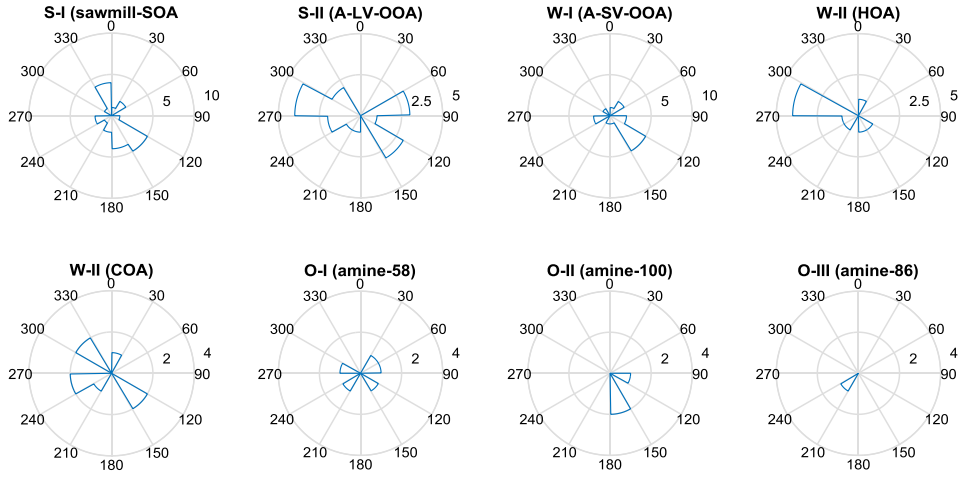


Figure S.11. Wind rose "histogram" plot for ("corr" *k*= 8) clusters: sawmill-SOA, A-LV-OOA, A-SV-OOA, HOA, COA, "amine-58", "amine-100" and "amine-86". Bar length depicts number of pollution cases with wind direction from the sector in question.

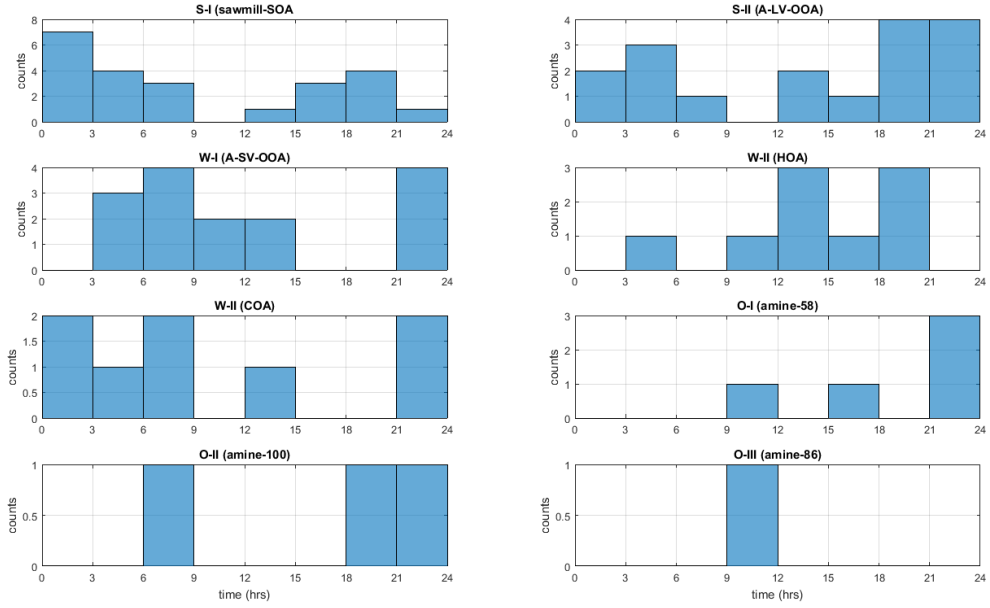


Figure S.12. Diurnal plots for the "corr" *k*=8 clusters. x-axis: local time when the plume is observed at SMEAR II, y-axis: plume count. Due to very low sample sizes not many conclusions can be drawn.

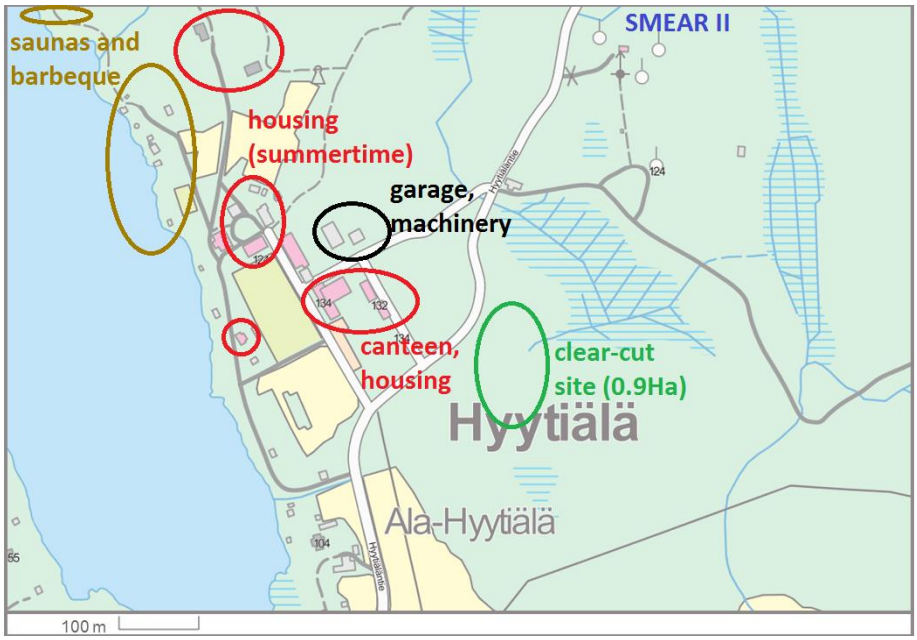


Figure S.13. Potential close-proximity aerosol sources – the main close range sources are related to the activities at the Hyytiälä forestry station, and the local road "Hyytiäläntie".

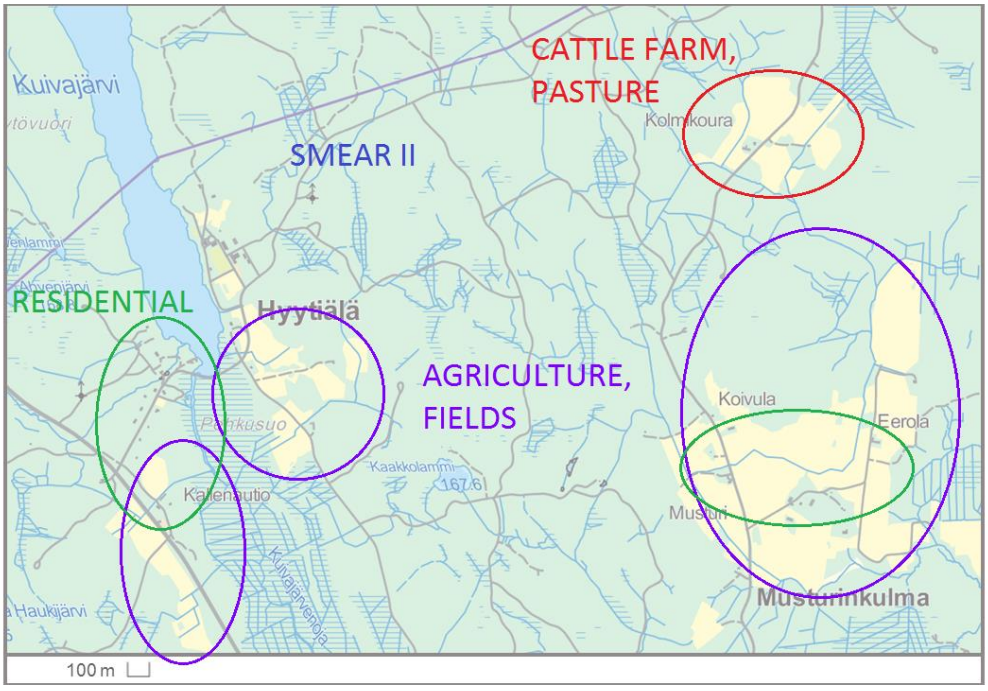


Figure S.14. Potential local aerosol sources – despite the remote location of SMEAR II there is scattered housing and small scale agriculture nearby.

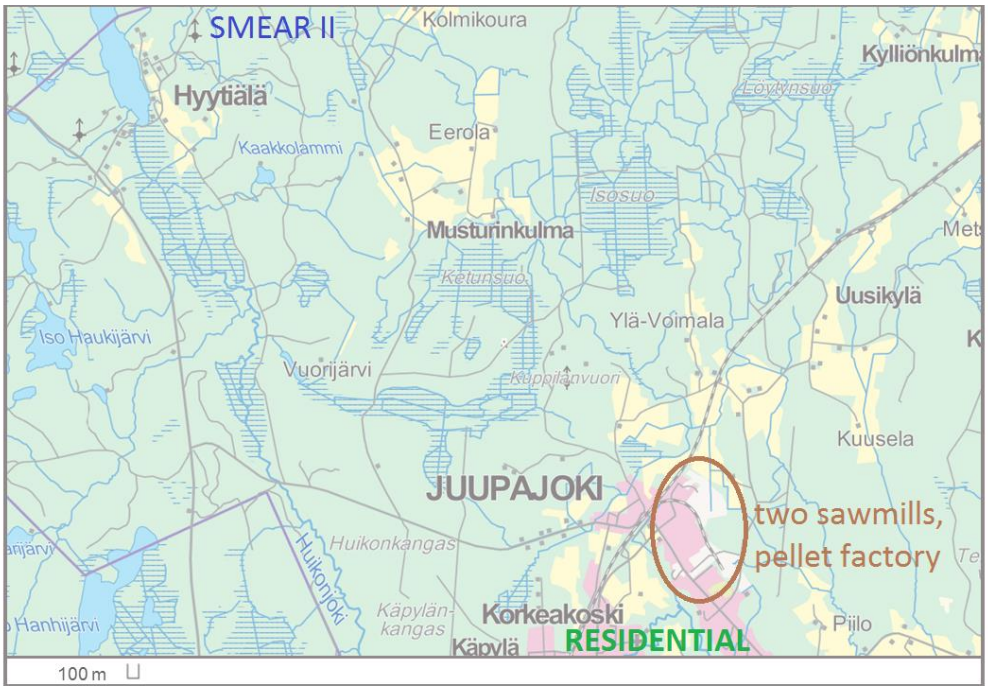


Figure S.15. The two sawmills and adjoined pellet factory are situated some seven kilometres south-east of SMEAR II, at the village of Korkeakoski.

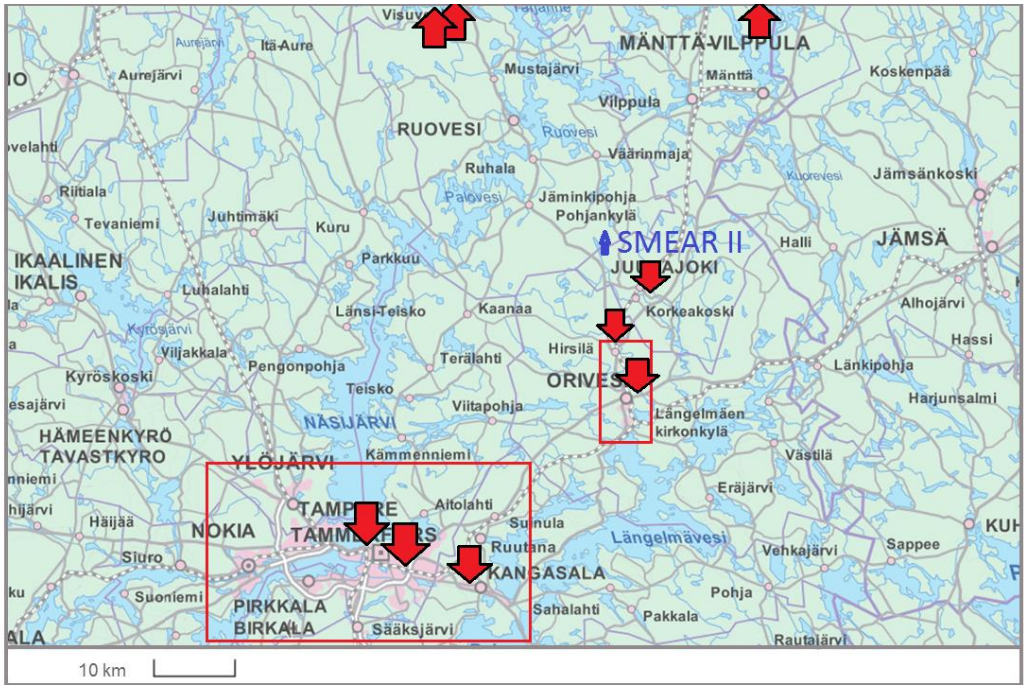


Figure S.16. Likely regional sources of anthropogenic aerosols. The town of Orivesi and the city of Tampere are highlighted. Sawmill locations taken from Liao et al. (2011) are marked by the arrows.

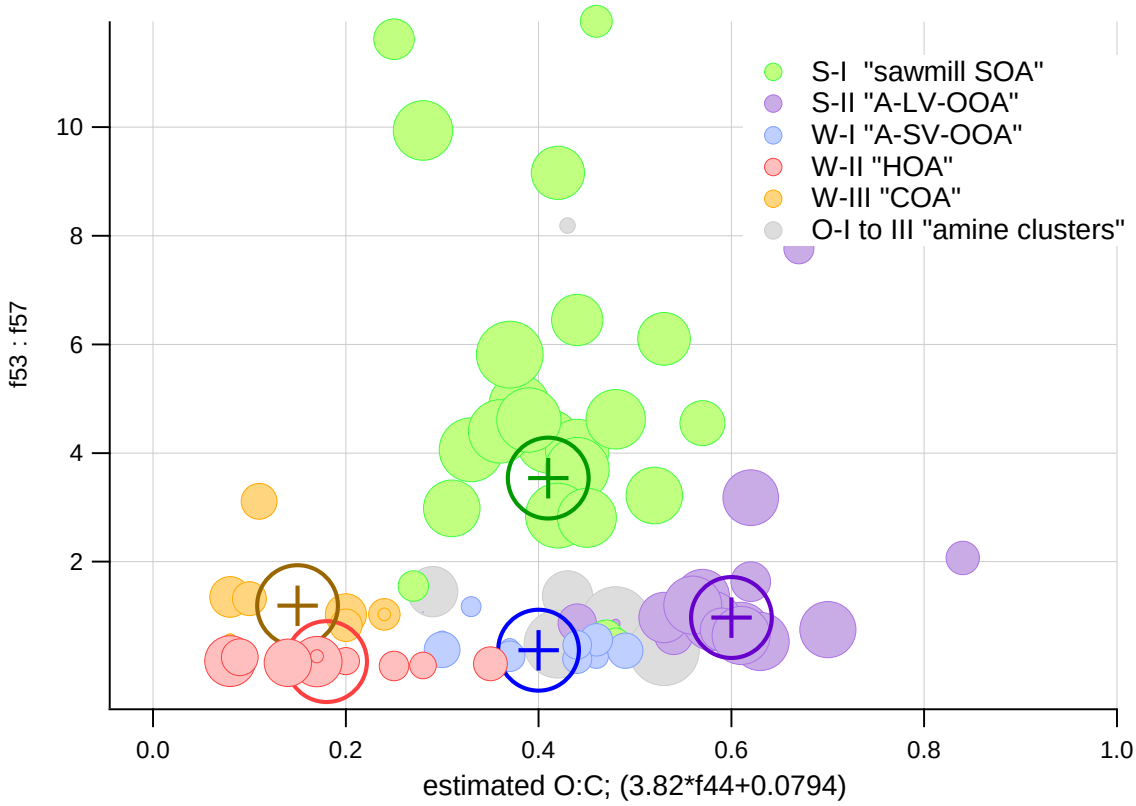


Figure S.17. "corr" *k* = 8 clustering solution projected onto 2-d axes corresponding to *f44* derived oxidation level (estimated O:C; Aiken et al., 2008) and *f53:f57*. Marker size corresponds to silhouette value of the point, ranging from zero to one. Cluster centroid locations are marked separately with darker colours. Outlier clusters are shown in grey, without centroids.

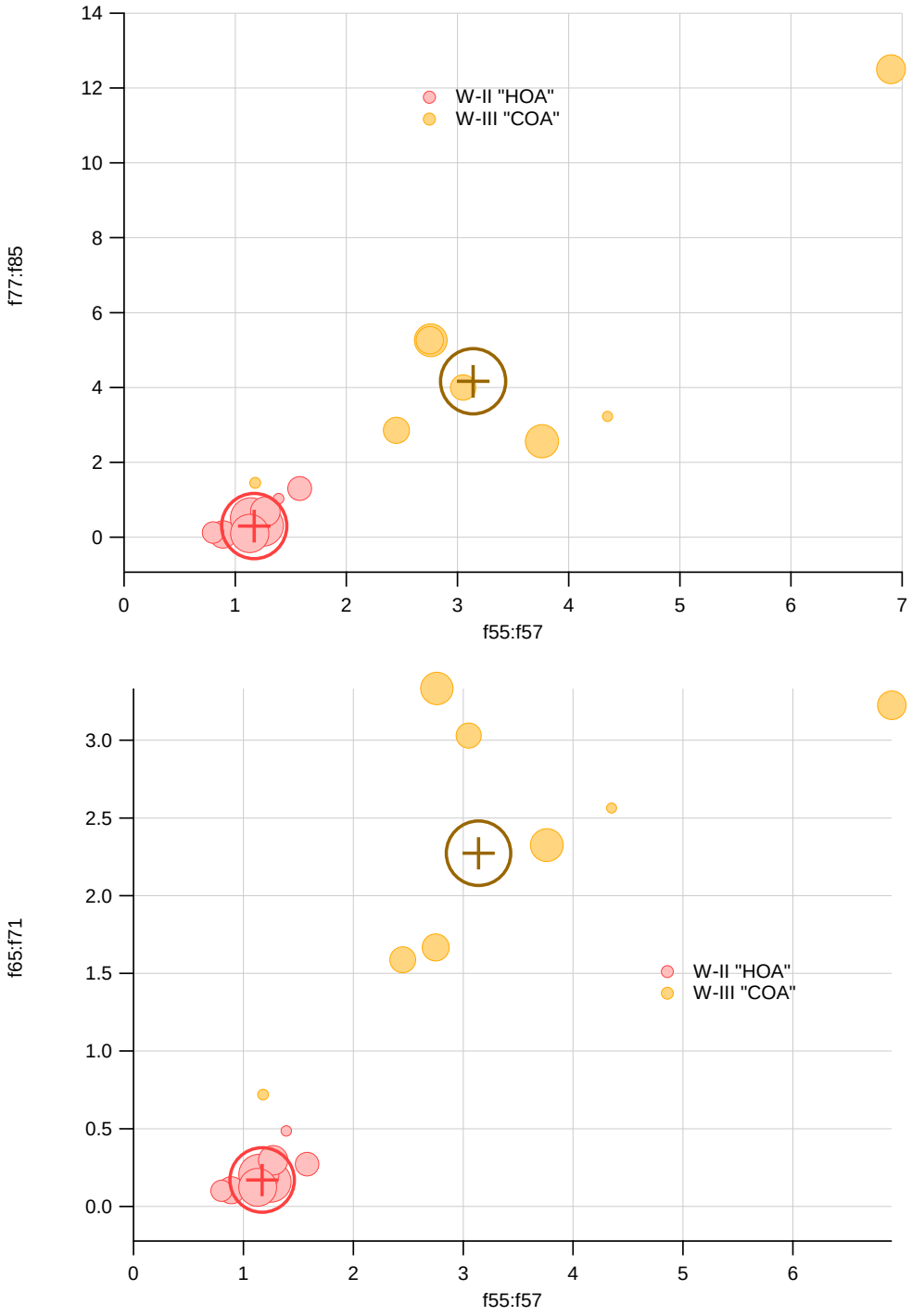


Figure S.18. The ratios *f77:f85* (upper panel) and *f65:f71* (lower panel) seem to offer good additional indicators for separating the COA (W-III) and HOA (W-II) clusters. Marker size corresponds to silhouette value of the point, ranging from zero to one. Cluster centroid locations are marked separately with darker colours.

Table S.1. Similarity matrix for clustering results, "corr" k = 8. Scaled similarity values (r_s^2 ; s_m=1.36; p < 0.05) are used.

	sawmill-SOA	A-LV-OOA	A-SV-OOA	HOA	COA	"amine-58"	"amine-100"	"amine-86"
sawmill-SOA	-	0.63	0.63	0.39	0.53	0.45	0.33	0.07
A-LV-OOA	0.63	-	0.62	0.16	0.18	0.48	0.47	-
A-SV-OOA	0.63	0.62	-	0.69	0.63	0.33	0.35	0.12
HOA	0.39	0.16	0.69	-	0.54	0.13	0.16	0.15
COA	0.53	0.18	0.63	0.54	-	0.14	0.13	0.13
"amine-58"	0.45	0.48	0.33	0.13	0.14	-	0.27	0.17
"amine-100"	0.33	0.47	0.35	0.16	0.13	0.27	-	0.08
"amine-86"	0.07	-	0.12	0.15	0.13	0.17	0.08	-

Table S.2. Cluster diagnostics values: population, silhouette, *f*-values, estimated O:C ratio

cluster	n	silhouette	*f43*	*f44*	*f55*	*f57*	*f60*	*f58*	*f72*	*f86*	*f100*	O:C(est)	*f55/f57*
sawmill-SOA	27	0.60	0.15	0.08	0.03	0.01	0.002	0.01	0.00	0.00	0.00	0.41	5.59
A-LV-OOA	21	0.62	0.06	0.15	0.03	0.01	0.005	0.01	0.00	0.00	0.00	0.60	2.38
A-SV-OOA	19	0.27	0.06	0.09	0.05	0.03	0.003	0.01	0.00	0.00	0.00	0.40	1.68
HOA	11	0.36	0.13	0.03	0.08	0.07	0.003	0.01	0.00	0.00	0.00	0.18	1.18
COA	9	0.37	0.09	0.02	0.05	0.02	0.004	0.00	0.00	0.00	0.00	0.15	3.48
"amine-100"	4	0.81	0.06	0.09	0.03	0.02	0.003	0.02	0.00	0.02	0.05	0.45	1.43
"amine-58"	7	0.46	0.07	0.10	0.02	0.01	0.003	0.09	0.00	0.00	0.00	0.48	4.22
"amine-86"	1	NaN	0.04	0.00	0.04	0.04	0.004	0.08	0.02	0.08	0.00	0.08	1.15

Table S.3. Library spectra similarities related to SV-OOA sub-species differentiation. Mass scaled (s_m=1.36) squared Pearson correlation coefficients against reference spectra from AMS spectral database (Ulbrich et al., 2009).

spectrum name	reference	r_s^2 vs	SV(HOA)	SV(COA)	SV(BBOA)
HOA cluster (W-IV)	(this study)		0.70	0.51	0.35
COA cluster (W-V)	(this study)		0.79	0.58	0.68
A_DEC_Q_010_HOA	Lanz et al., 2007		0.81	0.53	0.39
A_DEC_Q_012_PittsHOA	Ulbricht et al., 2009		0.76	0.67	0.51
A_DEC_Q_015_HOA_avg	Ng et al., 2011		0.67	0.57	0.40
A_DEC_C_032_HOA	Hersey et al., 2011		0.66	0.65	0.44
A_DEC_W_037_HOA	Crippa et al., 2013		0.63	0.60	0.41
A_DEC_Q_001_HOA_Pittsburgh	Zhang et al., 2005		0.63	0.60	0.41
A_HR_015_HOA_HOA	Mohr et al., 2013		0.53	0.50	0.32
A_DEC_Q_005_HOA'	Lanz et al., 2008		0.50	0.27	0.18
A_DEC_W_036_COA	Crippa et al., 2013		0.65	0.77	0.65
A_HR_014_COA_COA	Mohr et al., 2013		0.71	0.86	0.76
A_DEC_Q_011_Wood_burning	Lanz et al., 2008		0.69	0.50	0.74
A_DEC_W_035_BBOA	Crippa et al., 2013		0.56	0.29	0.72
A_DEC_Q_019_BBOA_avg	Ng et al., 2011		0.71	0.62	0.70
A_HR_013_BBOA_BBOA	Mohr et al., 2013		0.69	0.72	0.58