Analysis of Ambient Particle Size Distributions Using Unmix and Positive Matrix Factorization

EUGENE KIM
Department of Civil and Environmental Engineering, Clarkson University, Box 5708, Potsdam, New York 13699

PHILIP K. HOPKE
Department of Chemical Engineering, Clarkson University, Box 5708, Potsdam, New York 13699

TIMOTHY V. LARSON
Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington 98195

DAVID S. COVERT
Department of Atmospheric Science, University of Washington, Seattle, Washington 98195

Hourly averaged particle size distributions measured at a centrally located urban site in Seattle were analyzed through the application of bilinear positive matrix factorization (PMF) and Unmix to study underlying size distributions and their daily patterns. A total of 1051 samples each with 16 size intervals from 20 to 400 nm were obtained from a differential mobility particle size operating between December 2000 and February 2001. Both PMF and Unmix identify four similar underlying factors in the size distributions. Factor 1 is an accumulation mode particle size spectrum that shows a regular nocturnal pattern, and factor 2 is a larger particle distribution. Factor 3 is assigned as a traffic-related particle distribution, based on its correlations with accompanying gas-phase measurements, and has a regular weekday-high rush-hour pattern. Factor 4 is a traffic-related particle size distribution that has a regular rush-hour pattern on weekdays as well as weekends. Conditional probability functions (CPF) were computed using wind profiles and factor contributions. The results of CPF analysis suggest that these factors are correlated with surrounding particle sources of wood burning, secondary aerosol, diesel emissions, and motor vehicle emissions.

Introduction

Many airborne particle measurements and epidemiology studies have been undertaken since the association between ambient particle concentrations and adverse health effects has been shown in many studies (1–3). Recent epidemiological studies suggested that fine particles (particle diameter < 2.5 μm), especially ultrafine particles (~100 nm), may cause negative health effects because of their high number concentrations in the ambient, high penetration, and deposition efficiencies in the lungs and their acidity or catalytically active chemical properties (4–8). Many studies are monitor-

Positive matrix factorization (PMF) (14) and Unmix (15) have been shown to be powerful alternatives to traditional receptor modeling methods for airborne particulate matter source identification (16, 17). PMF has been used to assess particle source contributions in the Arctic (18), Hong Kong (19), Phoenix (20, 21), Thailand (22), Vermont (23), Atlanta (24), and three northeastern U.S. locations (25). Unmix has been applied to several aerosol data sets from Los Angeles (26) and Phoenix (27). Also, PMF and Unmix were compared in the northern Vermont aerosol study (28). All of these applications have been to particle composition data. Particle size distribution data measured in European cities (29, 30) have been studied using principal component analysis (PCA) as well as a constrained linear receptor model that has some similarities to PMF. The results of these studies suggest that particle sources produce characteristic size distributions that are sufficiently stable in the atmosphere when measured at a particular site to provide a problem suitable for factor analysis.

Particle size distributions in the atmosphere are dynamic with the potential for coagulation and size-dependent deposition. Zhu et al. (31, 32) performed studies near major highway sources. These results suggest that within a few hundred meters of the roadway there are significant changes in the size distributions as the smallest sizes coagulate and dry deposit. However, at a reasonable distance from the road, these processes slow significantly as the particle numbers decrease, and thus, a quasistationary profile can be anticipated if the sampling location is not too near active sources of ultrafine particles. Similar results have been observed in studies of the size distributions in a dilution sampler (33); examining the emissions of a boiler burning coal, oil, or natural gas shows that at a sufficient residence time, a stable size distribution appears to have developed. Previous laboratory studies of a variety of combustion sources showed that they emitted monomodal, log-normal distribution (34). Thus, although the potential for additional change in the size distributions while the aerosol is in transit from the source to the receptor site, it appears that there is sufficient stability in the size distributions from the various sources that the ambient aerosol size distribution can be considered to be a superposition of the distributions arising from multiple sources.

The objectives of this study are to apply these new multivariate receptor modeling methods to particle size distribution data and to estimate possible sources from model-identified particle size distributions. In the present paper, PMF and Unmix were applied to a particle size distribution data set from 20 to 400 nm collected during a ca. 2-month period at a monitoring site in Seattle, WA. The same number of factors were extracted from both PMF and Unmix to ensure that these factors were stable and strong underlying features. The resolved particle size distributions and their hourly patterns as well as their possible sources are discussed. Conditional probability functions were calculated using local wind data to help identify the possible sources of particles. The results could be used for the correlation study of local fine particle sources and their adverse health effects.

Experimental Methods

Sample Collection. The particle size distribution data used in this study consisted of hourly measurements collected from December 2000 to February 2001 at an urban monitoring site (Beacon Hill). As shown in Figure 1, the Beacon Hill site
is centrally located within the Seattle urban area on a hilltop, 99 m above sea level. The monitoring site is located inside a water reservoir impoundment, located 5 km southeast of the downtown business district. The sampling height of particle size and gaseous measurements was 4 m above ground. All instruments were closely located within a 2 m radius. Wind data were measured at a 10 m tower located in the site. The area to the immediate north and east of the hill is part of a larger ridge defining the eastern edge of an industrialized valley. The Port of Seattle, reservoir is residential. The hill is centrally located within the Seattle urban area on a hilltop, WA.

FIGURE 1. Location of the Beacon Hill monitoring site in Seattle, WA.

A total of 1051 particle size distributions, each with 17 size intervals from 20 to 600 nm, were measured using a differential mobility particle sizer (DMPS). The last size interval (Dp, 0.571 nm) was not included in the analysis due to the collection efficiency drop. The DMPS consists of a TSI model 3080 Electrostatic Classifier with model 3081 long column differential mobility analyzer (DMA) and a TSI model 3010 condensation nuclei counter (CNC). The DMPS steps through the mobility size range by changing the voltage on the DMA center rod from 40 to 10 000 over a period of 10 min. At the end of stepped scan sequence, the mobility data are inverted with the charge probability matrix to get the Stokes diameter (Dp) size distribution. The sample and sheath flows, temperature, pressures, relative humidity in the sensing volume, and scan voltages were checked twice weekly. It is possible that significant changes in the ambient concentration at any size during a scan lead to distorted measurements (36). In this study, the inversion scheme checks the number concentration from one size interval to the next size interval for continuity. The scan is discarded if the change exceeds a given threshold. The inversion will not converge if the measured size distribution is unrealistic. Most of the missing values in the size distributions occurred when the inversion failed or when the sampling pumps had problems.

The particle volume size distributions were used in this study for the direct comparisons with measured source size distributions. It would be possible to use either number or volume distributions since one can be derived from the other. Since current particulate matter regulations are mass based, it is more useful to examine volume distributions. Every 10-min measured particle number concentration (dN/dlogDp) for each size interval i was integrated to estimate hourly particle number concentration, and then the particle volume concentration (dV/dlogDp) was calculated for the analysis. In addition, the local air pollution control agency (Puget Sound Air Pollution Control Agency) provided hourly measurements of oxides of nitrogen (NOx, chemiluminescence method) and carbon monoxide (CO, nondispersive infrared method), light scattering coefficient (bap, light absorption photometer), and light absorption coefficient (ba, light absorption photometer). Summaries of particle volume concentrations and other measurements used in this study are presented in Tables 1 and 2, respectively.

Multivariate Receptor Modeling. The general receptor modeling problem can be stated in terms of the contribution from p independent sources to all size intervals in a given sample (37, 38) as follows

\[ x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} \] (1)

where \( x_{ij} \) is the concentration in the jth size interval measured in the ith sample, \( g_{ik} \) is the particle concentration from the kth source contributing to the ith sample, and \( f_{kj} \) is the particle volume fraction in the jth size interval from the kth source. This equation then assumes that there are stable size
distributions that are emitted from the various source types that contribute to the observed ambient aerosol size distribution.

As pointed out by Henry (39), there are infinite numbers of possible solutions to the factor analysis problem outlined in eq 1 (rotation of matrices). However, this problem is the same as what exists in other areas of science. Particle size distributions represent data that are similar to the problem of spectrochemical separations. The absorption spectra for individual components are sums of 1 or more Gaussian-like peaks. These peaks can also be resolved by regression of the spectral profiles if they are known and not changed by the presence of the other absorbing species. However, for many years back to the early work of Lawton and Sylvester (40), factor analysis has been applied to these data with a high level of success. Given the prior results of the analysis of particle size data (29, 30), there do appear to be reasonably stable source size profiles. There can be expected to be substantial differences in the emission rates of various sources at different times that will lead to the different sources contributing quite different numbers of differently sized particles. Because we have quite high time resolution, variations in emissions such as morning and evening rush hours, evening and weekend wood fires, and other time-varying activities can be expected to provide inputs of particles at differing times. It was previously demonstrated that higher time resolution sampling provides the ability to identify increased number of sources (41). Thus, a series of measurements that represent the superposition of the particle size profiles with differing intensities as functions of time will always permit a factor analysis to separate them.

To decrease rotational freedom, PMF and Unmix (version 2.3) use nonnegativity constraints on the factors. In addition, parameter FPEAK is used to control the rotations in PMF (42). By changing FPEAK values as one of the model input parameters, rotations of matrices are visualized and then best solutions can be found. Unmix uses the edges for which parameters, rotations of matrices are visualized and then best solutions can be found. Unmix identifies four profiles from the given data set. In this study, FPEAK was set to be 11.3°. Calm winds (<1 m/s) were excluded from this analysis due to the isotropic behavior of wind vane under calm winds. From the trials of several different percentile of the fractional contribution from each source, the threshold criterion of upper 25% was decided to show clear directionality.

Results and Discussion

Unmix identified four profiles from the given data set. In Unmix, the tracer variable was not chosen for this study. In the PMF, different numbers of factors need to be tested and the optimal number is the one that adequately fits the data with the most physically meaningful results. In this study, Δφ was set to be 11.3°. Calm winds (<1 m/s) were excluded from this analysis due to the isotropic behavior of wind vane under calm winds. From the trials of several different percentile of the fractional contribution from each source, the threshold criterion of upper 25% was decided to show clear directionality.

Unmix provided a solution that minimizes an objective function, Q(E), based upon uncertainties for each observation (14, 44). This function is defined as

\[
Q(E) = \sum_{i} \sum_{j=1}^{n} \frac{u_{ij}^2}{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}
\]

where \( u_{ij} \) is an uncertainty estimate in the \( j \)th size interval measured in the \( i \)th sample. The application of PMF depends on the estimated uncertainties for each of the data values. The uncertainty estimation provides a useful tool to decrease the weight of missing and below detection limit data in the solution as well as account for the variability in the source profiles.

To use same input data for both PMF and Unmix, the hourly particle size distributions in which concentrations are missing are excluded from this analysis. Also, the treatment of below detection limit values (45) for PMF and Unmix were not used for the same reason. For the input data for the PMF, the associated uncertainty with measured value for each size interval is estimated by taking into account the temporal variability of both measured particle number and sampled air flow rates.

The results of PMF were then normalized by a scaling constant, \( s_i \), so that the quantitative source contributions as well as profiles for each source were obtained. Specifically

\[
x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} s_i
\]

where \( s_i \) is determined by regressing sum of volume concentrations in the \( i \)th sample, \( v_i \), against estimated source contribution values.

In the Unmix, sum of volume concentrations shown in eq 4 were included as an input variable so that the model results were normalized by Unmix directly.

Conditional Probability Function. The conditional probability function (CPF) (46) was calculated using source contribution estimates from PMF and Unmix coupled with wind direction values measured at the site. The source impacts from various wind directions can be analyzed by CPF. The sources are likely to be located to the direction that have high conditional probability values. To minimize the effect of atmospheric dilution, the hourly fractional particle volume contribution from each source relative to the total of all sources was used rather than using the absolute source contributions. Specifically, the CPF is defined as

\[
CPF = \frac{m_{\Delta\phi}}{n_{\Delta\phi}}
\]

where \( m_{\Delta\phi} \) is the number of occurrences from wind sector \( \Delta\phi \) that exceeded the threshold criterion and \( n_{\Delta\phi} \) is the total number of data from the same wind sector. In this study, \( \Delta\phi \) was set to be 11.3°. Calm winds (<1 m/s) were excluded from this analysis due to the isotropic behavior of wind vane under calm winds. From the trials of several different percentile of the fractional contribution from each source, the threshold criterion of upper 25% was decided to show clear directionality.

Results and Discussion

Unmix identified four profiles from the given data set. In Unmix, the tracer variable was not chosen for this study. In the PMF, different numbers of factors need to be tested and the optimal number is the one that adequately fits the data with the most physically meaningful results. In this study, however, only four factor solutions were used in order to extract stable and common factors to compare the results between PMF and Unmix. Also, since rotational ambiguity exists in the PMF solutions, PMF was run several times with different FPEAK values to determine the range within which the objective function \( Q(E) \) value in eq 2 remains essentially constant (38). The optimal solution should lie in this FPEAK range. As shown in Figure 2, FPEAK values between –0.3 and

FIGURE 2. FPEAK versus Q value for PMF.
0.3 were likely to provide relatively constant $Q(E)$ values, so that the PMF solution from default rotation ($FPEAK = 0$) as well as from six $FPEAK$ values between $-0.3$ and $0.3$ were considered in this study. For the PMF, the robust mode was used to reduce the influence of extreme values on the solution. Table 3 shows the Pearson correlations of the contributions from PMF and Unmix against on-site measured variables. The comparisons of the hourly volume contributions from all factors for 16 size intervals reconstructed by PMF and Unmix with measured values show that the resolved factors by both methods effectively reproduce the measured values and account for most of the variation in the particle concentrations (slope $= 0.988$ and $r^2 = 0.99$ for PMF; slope $= 0.992$ and $r^2 = 0.99$ for Unmix). Figure 3 presents the PMF identified feature profiles. Mean and standard deviation values are estimated based on the variability in seven $FPEAK$ solutions. Figure 4 presents the Unmix identified profiles (value ± standard deviation resolved by Unmix). In Unmix, “edges” are used to fix the profiles. Edges are defined by points of nearly pure contributions of only a single source. The idea is discussed by Henry and co-workers (15, 25, 26). Thus, Unmix has reduced the rotational ambiguity in the solution by assuming the edges serve as constraints on several factors. In the PMF analysis, the range of rotational ambiguity (i.e., the region over which $FPEAK$ can be varied with essentially no change in the value of $Q$ in eq 2) has been explored. Thus, the PMF error bars represent the range of uncertainty including measurement error and rotational ambiguity while the Unmix solution only reflects the measurement errors, which are much smaller. When the errors associated with profiles are considered, Figures 3 and 4 indicate that both methods resolved very similar profiles. Since the conditional probabilities for the possible source locations between PMF and Unmix are the same, only PMF resolved conditional probabilities are presented in Figure 5. In Figure 6, the hourly contribution patterns estimated for both weekdays and weekends by PMF are shown for the same reason.

**TABLE 3. Pearson Correlation Coefficients between Factor Contributions and Independently Measured Variables**

<table>
<thead>
<tr>
<th>variable</th>
<th>factor 1</th>
<th>factor 2</th>
<th>factor 3</th>
<th>factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PMF</td>
<td>Unmix</td>
<td>PMF</td>
<td>Unmix</td>
</tr>
<tr>
<td>$d_{ap}$</td>
<td>0.74</td>
<td>0.60</td>
<td>0.61</td>
<td>0.45</td>
</tr>
<tr>
<td>$d_{ap}$</td>
<td>0.69</td>
<td>0.62</td>
<td>0.95</td>
<td>0.86</td>
</tr>
<tr>
<td>NO$_x$</td>
<td>0.69</td>
<td>0.48</td>
<td>0.56</td>
<td>0.42</td>
</tr>
<tr>
<td>CO</td>
<td>0.60</td>
<td>0.42</td>
<td>0.64</td>
<td>0.52</td>
</tr>
</tbody>
</table>

* $FPEAK$ values $= 0$. 

**FIGURE 3.** Normalized factor profiles (mean ± standard deviation) resolved from particle size distributions using PMF. $FPEAK$ values from $-0.3$ to $0.3$ were used.

**FIGURE 4.** Normalized factor profiles and standard deviation resolved from particle size distributions using Unmix.

**FIGURE 5.** CPF plots for the highest 25% of the volume contribution from model resolved underlying factors.

**FIGURE 6.** Hourly contribution patterns estimated for both weekdays and weekends by PMF.
nocturnal pattern is distinct during weekends. The origin of this factor is likely to be residential wood burning. The CPF plots in Figure 5 indicate that the source of this factor is located in a residential area that is immediately north and east of the site. Wood burning is prevalent in the northwestern United States (east of the site). Wood burning is prevalent in the northwestern United States (47). In this study, wood burning in winter is consistent with emissions from fireplace and woodstoves, which are ubiquitous ground-level sources whose emissions are frequently trapped by nighttime temperature inversion. The daytime contributions from this factor during the weekends are lower than those of weekdays. Both PMF- and Unmix-estimated contributions from this factor have association with NOx and SO2, which are ubiquitous ground-level sources whose emissions are frequently trapped by nighttime temperature inversion.

The origin of this factor can be diesel emissions. The CPF plots show that the source of this factor is located north of the site. Although sea salt and traffic produced airborne soil can be possible sources of these larger particles (45), the source associated with this factor is likely to be secondary aerosol. This assignment is made because the extracted size distributions compare well with those from prior measurements (49). Sulfate, nitrate, and ammonium measured in Claremont, CA, have the highest concentrations between 0.6 and 0.7 μm (47). Nitrate-rich secondary aerosol, in which most of nitrates originate from combustion emissions, has seasonal variations with its maximum in the winter when these samples were obtained. Low temperature and high relative humidity in winter help the formation of nitrate aerosols in urban areas (24). In addition, the direction for the ocean (west of the site) and nearby highways are not consistent with the CPF plot.

Factor 3 has the highest concentration at approximately 0.1 μm. The association with NOx, CO, and bp indicate that this is a traffic-related particle feature. The hourly pattern in Figure 6 shows that this factor has a regular weekday rush-hour pattern. The source of this factor can be diesel emissions. In Figure 5, the high values of conditional probability point northwest of the site, indicating the possibility of contributions from heavy-duty diesel emissions from the Port of Seattle which is located about 5 km northwest of the site. In addition, the roads and freeways close to the Beacon Hill site have high traffic of commuter buses and delivery trucks that contribute diesel emissions to the site. Factor 3 shows similar size distribution to measured mass size distribution of heavy-duty diesel (50) and diesel passenger car exhaust particles (51) that has the highest concentration at approximately 0.1 μm. Park et al. (52) found that diesel engine exhaust is monomodal with a mass mode mean around 100–200 nm. Also, particle volume distribution of 1995 diesel engine exhaust shows the geometric volume mean diameter of 0.125 μm (53).

Factor 4 has high concentrations in nucleation (Dp < 0.05 μm) and accumulation mode. As shown in Table 3, the association with NOx indicates that this is also a traffic-related particle feature. Figure 6 shows that this factor has a regular rush-hour pattern during weekdays as well as on weekends. In the CPF plots, three high peaks of conditional probability point to the nearby freeways and junctions. For the possible source of this factor, motor vehicle emissions are suggested. The volume size distribution of this factor shows bimodality that has the highest concentration at approximately 0.04 and 0.3 μm. The peak in the accumulation mode is very similar to those in a previous study (44). From the measurements near the freeway during peak traffic condition, Morawska et al. (44) showed two peaks at 0.3 and 3 μm in volume size distribution of the traffic-influenced aerosol. As they expected to see near a major motor vehicle traffic concentration, a nucleation mode peak is shown in the distribution of factor 4. As discussed with respect to factor 3, diesel engines typically emit particles in the accumulation mode range. Maricq and co-workers (54) found on-road vehicles emitted particles with number distributions that would correspond well to the volume size found in this factor. Thus, a mixture of vehicles...
on local highways appears to be a good assignment for this factor.

The PMF and Unmix estimated particle volume contributions are compared in Figure 7, indicating high correlations ($r^2 = 0.89-0.98$) and similar contributions (slope = 0.8-1.1), except for factor 3 (slope = 2.1). From the principal component analysis of particle number size distributions from 10 to 700 nm measured at the urban street canyon, Wahlin et al. (29) identified three sources: traffic, diesel, and secondary particles. The traffic source contributed to the total number of particles the most, and the secondary particles contributed to the total number of particles the least. For this study, the average source contributions of each factor to the particle volume concentration for 2-month period are summarized in Figure 8. Factor 1 (residential wood burning) has the highest contribution to the particle volume concentrations (PMF 48%; Unmix 44%). For the PMF, the second contributor is factor 2 (secondary aerosol), accounting for 21% of the particle volume concentration, and the third contributor is factor 3 (diesel emissions), accounting for 20%. For the Unmix, the second contributor is factor 3 (diesel emissions), accounting for 30% of the particle volume concentration, and the third contributor is factor 2 (secondary aerosol), accounting for 16%. Factor 4 (motor vehicle) is the fourth contributor (PMF 11%; Unmix 10%). When the errors associated with average contributions are considered, significant differences only occur for factor 3, in which Unmix estimates higher contribution than PMF does.

In the analysis of particle size distribution, a factor can be a combined size distribution of surrounding sources if those are co-located and the temporal emission patterns are similar. Unmix did not find a five-factor solution, but it is possible to extract an additional factor by PMF. The factor 5 extracted from the five-factor PMF model is shown in Figure 9. There is a peak in the volume distribution at 20–30 nm. Thus, there is a very high number count in sizes near 10 nm.

It appears that this factor is another representation of freshly nucleated particles. Examination of the distributions has shown that there are no nucleation and growth events, as seen at other locations (55). Motor vehicle emissions do produce different sized particles including nucleation mode particles. The CPF plots for factors 4 and 5 (Figure 10) show similar directional patterns. The contributions of factors 4 and 5 are plotted against the wind speed in Figure 11. It can be seen that factor 5 contributed less than factor 4 at low wind speeds, suggesting that factor 5 is fresher motor vehicle emissions from nearby freeways. Although this factor contributes very little to the total volume of the ambient particulate matter, it is the largest factor contributing to the particle number concentrations.
FIGURE 9. Normalized factor profiles (mean ± standard deviation) resolved from particle size distributions using PMF 5-factor model. FPAKE values from −0.3 to 0.3 were used.

FIGURE 10. CPF plots for the highest 25% of the volume concentrations deduced by PMF 5-factor model.

FIGURE 11. Wind speed versus particle volume concentrations deduced by PMF 5-factor model.

The sources and their profiles are not clearly identified in detail without costly chemical speciation data, and therefore, the source identifications of the factors shown in this study are tentative. However, the comparisons between PMF- and Unmix-predicted and measured volume concentration over all samples show that the extracted factors from both methods effectively reproduce the original size distributions. This study suggested that PMF and Unmix are useful methods to extract underlying features from ambient particle size distribution data and possibly cost-effective tools to identify their possible sources.

Acknowledgments
This study was partly funded by the University of Washington/ EPA Northwest Research Center for Particulate Air Pollution and Health (R827355) and by the University of Rochester/ EPA Participulate Matter and Health Center (R827354). This support does not constitute an endorsement by the U.S. EPA of the views expressed. We also thank Dr. Pentti Paatero for his helpful discussions of the work presented in this paper. We would also like to thank the anonymous reviewer A for the useful comments on this manuscript.

Literature Cited