Aerosol Science and Technology, 38(S1):118–132, 2004

ISSN: 0278-6826 print / 1521-7388 online
DOI: 10.1080/02786820390229589

Source Apportionment

Advanced Factor Analysis on Pittsburgh Particle Size-Distribution Data

Liming Zhou,[1] Eugene Kim,[1] Philip K. Hopke,[1] Charles D. Stanier,[2] and Spyros Pandis[2]

[1]*Department of Chemical Engineering, Clarkson University, Potsdam, New York*
[2]*Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania*

Positive matrix factorization (PMF) method was applied to particle size-distribution data acquired during the Pittsburgh Air Quality Study (PAQS) from July 2001 to August 2001. After removing those days with nucleation events, a total of 1632 samples, each with 165 evenly-sized intervals from 0.003 to 2.5 μm, were obtained from scanning mobility particle spectrometer (SMPS) and aerodynamic particle sampler (APS). The temporal resolution was 15 min. The values for each set of five consecutive-size bins were averaged to produce 33 new size channels. The size distributions of particle number as well as volume were analyzed with a bilinear model. Three kinds of information were used to identify the sources: the number and volume size distributions associated with the factors, the time frequency properties of the contribution of each source (Fourier analysis of source contribution values) and the correlations of the contribution values with the gas-phase data and some composition data. Through these analyses, the sources were assigned as sparse nucleation, local traffic, stationary combustion, grown particles and remote traffic, and secondary aerosol in a sequence of decreasing number concentration contributions. Conditional probability function (CPF) analysis was performed for each source so as to ascertain the likely directions in which the sources were located.

Received 23 October 2002; accepted 31 March 2003.

This research was conducted as part of the Pittsburgh Air Quality Study which was supported by US Environmental Protection Agency under contract R82806101 and the US Department of Energy National Energy Technology Laboratory under contract DE-FC26-01NT41017. This paper has not been subject to EPA's required peer and policy review, and therefore does not necessarily reflect the views of the Agency. No official endorsement should be inferred.

Address correspondence to Philip K. Hopke, Department of Chemical Engineering, Clarkson University, P. O. Box 5708, Potsdam, NY 13699-5708. E-mail: hopkepk@clarkson.edu

INTRODUCTION

Recently, substantial attention has been paid to the airborne particulate matter (PM), which is believed to be associated with increased morbidity and mortality, especially to high-risk groups (van Bree and Cassee 2000). Particles of different size have different deposition pattern in the airways. Ultrafine particles ($<$0.1 μm) have a higher deposition fraction than fine particles (0.1–2.5 μm). In general, pulmonary deposition increases with decreasing particle size (Venkataraman and Kao 1999) and number concentrations of ultrafine particles were also shown to be more closely associated with variations in lung function (Peters et al. 1997).

To implement effective strategies to control the emission of PM, a comprehensive data set is needed. The Pittsburgh Air Quality Study (PAQS) is a multidisciplinary set of projects in the Pittsburgh region that addresses issues including understanding of the PM health effects, establishing the PM source–receptor relationships, and finding the limitations of existing instrumentation for PM measurements (Wittig et al. 2004).

Previously, principal component analysis (PCA) has been widely used for the source apportionment and recently has been applied to size-distribution data (Ruuskanen et al. 2001; Wahlin et al. 2001; Kim et al. 2004). An alternative to PCA, positive matrix factorization (PMF) is a powerful factor analysis method and has been successfully used to solve the receptor model for the source apportionment of the aerosol particles (Xie et al. 1999; Lee et al. 1999; Ramadan et al. 2000; Chueinta et al. 2000; Polissar et al. 2001; Song et al. 2001). In this study, particle size-distribution data from PAQS will be analyzed by PMF method in terms of both the number and volume distributions.

The continuous measurement of aerosol size distributions has low labor costs and can provide very large data sets with a time resolution of minutes for a moderate investment. The number of sites around the world measuring aerosol size distributions

has been increasing. The present article explores the possibility of using this information to gain insights about the sources contributing to the ambient PM levels. While it is possible to increase the power of the proposed method by combining the aerosol size distributions with additional measurements of the aerosol composition (metals, speciated organics, single-particle composition), here we focus on the use of the size distributions only. In future work the results of the present study will be compared with the more comprehensive analysis for a stricter test of the source apportionment power of the proposed method.

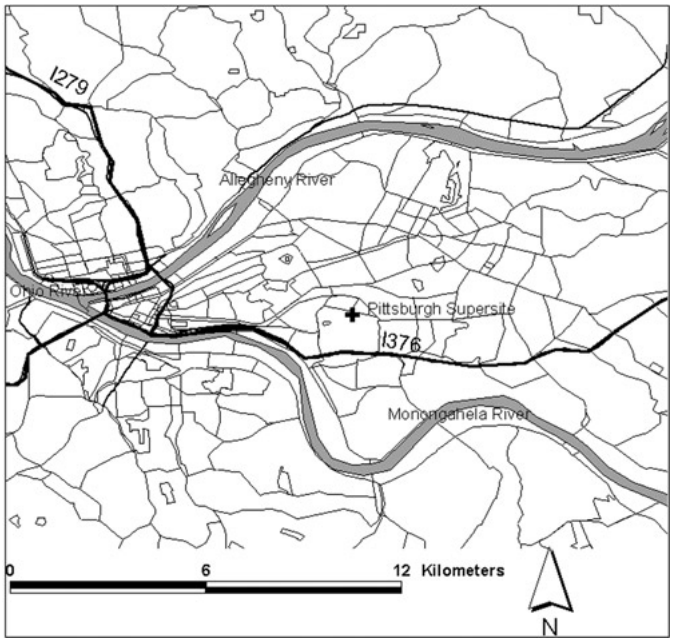
The aerosol may experience some changes in the size distribution during transport from the sources to the receptor site. The size distribution will change as particles coagulate and dry deposit. However, after some initial changes in the vicinity of the sources the most mobile particle sizes will be depleted, and the particle number concentrations will be sufficiently small so that further deposition and coagulation will be slow processes. Thus, it is reasonable to expect that the particle size distributions will become relatively stable at some appropriate distance from the emission sources. In this study, the events in which there is active growth of the particle size distribution have been explicitly eliminated from the analysis. Gas-phase data and some particle-mass data, such as $PM_{2.5}$, sulfate, and Organic carbon/elemented carbon (OC/EC), were also used to assist the identification of the potential sources.

DESCRIPTION OF THE DATA SET

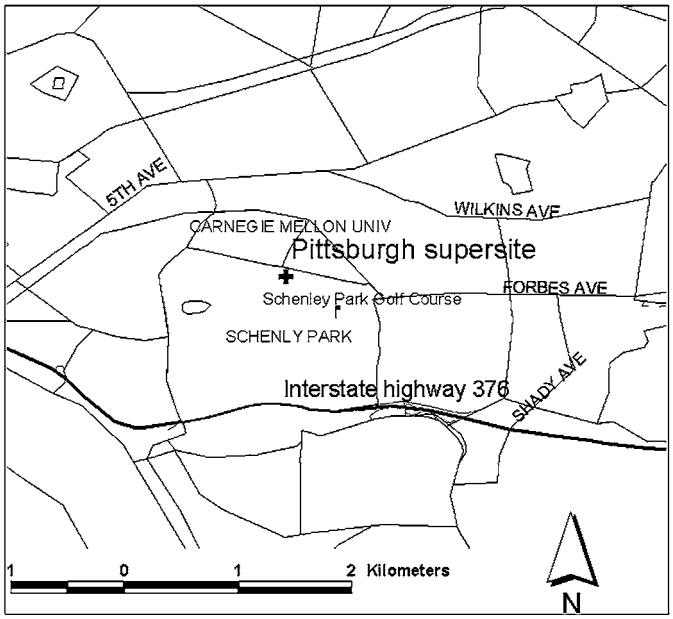
The size-distribution measurements were conducted from 30 June to 4 August 2001 at the Pittsburgh Supersite (Latitude 40.4395 , Longitude −79.9405). The location of the receptor site is shown in Figure 1 and is far from any major sources. Ambient air passed through Nafion dryers where the relative humidity (RH) was controlled to around 15%, and thus the data obtained through these measurements were "dry" data in comparison to the data measured at the ambient RH. The frequency was four 15 min samples during each hour. The results obtained from three instruments were used to produce a single size distribution.

Measurements were made with a scanning mobility particle spectrometer (nano-SMPS; TSI 3083 DMA and TSI 3025 CPC) sampling from 3–83 nm, an SMPS (TSI 3081 DMA and TSI 3010 CPC) sampling from 13–583 nm, and an aerodynamic particle size spectrometer (APS, TSI 3320) sampling from 0.56–2.5 microns. The SMPS instruments measured electrical mobility diameters. The APS reported aerodynamic diameters with the time of flight calibrated to a density of 1.0. The APS and long differential mobility analyzer (LDMA) data were merged by solving for an effective density for the APS particles, which gave the smallest error (in a least squares sense) in the APS–LDMA overlapping region (560–583 nm). Thus, above 583 nm the data represent electrical mobility diameter inferred from aerodynamic mobility and estimated density (Khlystov et al. 2004).

The meteorological (wind direction and wind speed), gas phase, and particle mass (including $PM_{2.5}$, sulfate, and OC/EC) data were measured at the same time and location as the particle



(a)



(b)

Figure 1. The map of the Pittsburgh region around the supersite.

number concentration measurements. The temporal resolution for meteorological data was 15 min, and for gas phase and particle mass data was 10 min. The sampling period for OC/EC was 4–5 h.

It appears that most of the sampling days have some degree of homogeneous nucleation occurring. In a typical nucleation event, the particle size initially grows rapidly (4–5 nm/h) and then the growth slows as it reaches a size of 30–120 nm. The

Table 1
Sampling periods without nucleation events

No.	Date	Day
1	Jul. 3	Tue.
2	Jul. 4	Wed.
3	Jul. 7	Sat.
4	Jul. 8	Sun.
5	Jul. 10	Tue.
6	Jul. 16	Mon.
7	Jul. 18	Wed.
8	Jul. 19	Thu.
9	Jul. 20	Fri.
10	Jul. 21	Sat.
11	Jul. 23	Mon.
12	Jul. 25	Wed.
13	Jul. 28	Sat.
14	Jul. 29	Sun.
15	Jul. 31	Tue.
16	Aug. 1	Wed.
17	Aug. 2	Thu.

basic assumption of the receptor model is that the ambient data is the sum of constant profiles (size distributions from the contributing sources). Thus, the days with intense nucleation events (usually having particle growth) were excluded in this study. The definition of an intense nucleation is by investigating the temporal variation rate of the total particle number concentration between 3 nm and 10 nm, denoted as dN_{10}/dt. If one day has a dN_{10}/dt value over 4,000/(cm^3 h), then it is thought that an intense nucleation events happens (Stanier et al. 2004) and that day is excluded from this study. Most of these days excluded have particle growth after the new particle formation. The remaining 17 days are indicated in Table 1, and they may still contain some weak and shortlived nucleation without observable particle growth.

The wind was generally from two directions during July, the northwest and southeast (0° or 360° for north, 90° for east, etc.). After those days with nucleation events were removed the wind from the northwest also disappeared, as shown by the wind profile subfigure in Figure 8. The reason may be that the transport from the Great Lakes area (northwest direction) brings clean and cooler air masses that provide good conditions for nucleation.

The missing values in the 165 size bins are replaced by the mean value of the samples in the same size interval. Since the maximum number of missing values for each size bin is 117, about 7% of the total sample number, there is no necessity to omit any size bins. The volume concentration is then calculated by the number concentration.

ESTIMATION OF MEASUREMENT ERRORS

Since no measurement errors were provided with the original experimental data, the following method is used to estimate the measurement errors to provide the needed inputs to the PMF program.

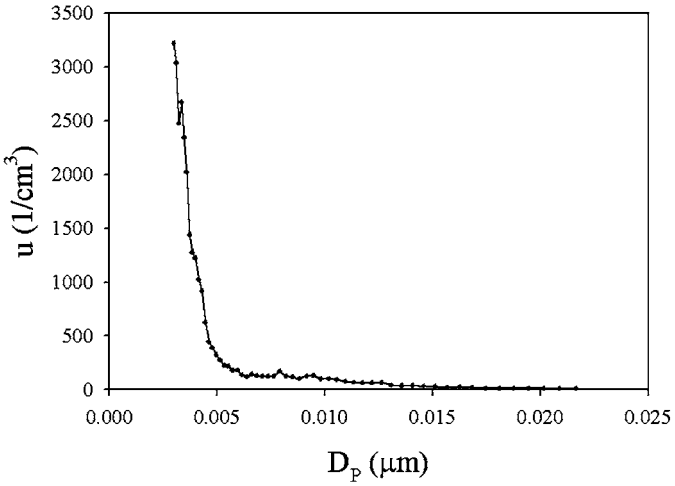


Figure 2. Estimated instrumental errors of small particles.

For the size range of 3–22 nm measured by the nano-SMPS, the smaller of the minimum nonzero value and the minimum difference of the concentration values within each size bin, u_h, was taken as the instrumental error for the hth size interval. For larger particles, the instrumental error cannot be estimated through this method. Figure 2 illustrates these values, and it can be seen that the numbers of particles smaller than 5 nm have large u values while they also have large particle numbers. Reischl (1991) discussed the instrumental error of the differential mobility analyzer (DMA) method and reported mean concentration determination errors for his DMA–Faraday cup electrometer (FCE) system. The condensation particle counter (CPC) has a counting efficiency that drops quickly toward zero for smaller particles and consequently has a higher instrumental error than the FCE, but the performance of CPC is much better for large particles, especially around 100 nm.

Comparing the estimated instrumental error in this study and the instrumental errors reported by Reischl (1991), it can be found that in this study the errors for 3 nm particles are higher, but they are lower for the larger particle sizes.

Besides the instrumental error, there are other sources of measurement errors. A discussion about the measurement errors can be found in Wong (1997). Based on statistics, the true value of a quantity is given by the average of a large number of measurements. "If the uncertainties are associated with the measuring process are random, the values obtained will most likely be scattered around the true value with some definite distribution" (Wong 1997).

For the measurement of the ith sample at the jth size bin, the concentration increased from 0 with a step length u_j, and each increase is independent. Let $x_{ih} = Nu_h$. Thus the probability distribution follows Poisson distribution with a mean of N and a variance N. The best estimation of the measurement error is $\sqrt{N}u_h$ as indicated in Equation (1). To add 1 in the parentheses

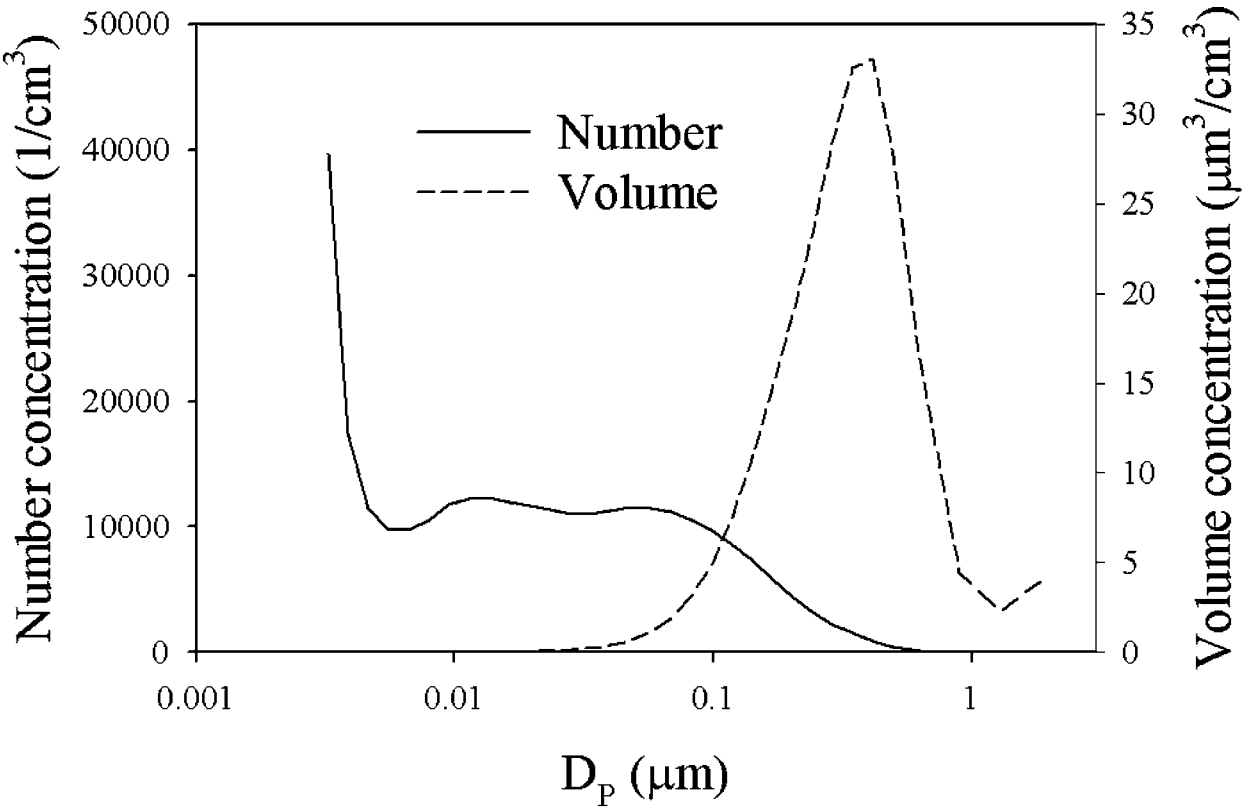


Figure 3. Average number and volume size distribution of all the samples.

is to avoid a zero uncertainty for a zero concentration value.

$$\sigma'_{ih} = \left(\sqrt{\frac{x_{ih}}{u_h}} + 1\right) u_h, \quad h = 1, 2, \ldots, 55. \qquad [1]$$

Every consecutive five size intervals were combined into a single new interval to make the size distribution smoother and minimize the error caused by the discontinuity between instruments, and thus 33 new size channels are created. Since in the original dataset the size bins were evenly spaced (in the sense of logarithm), the obtained 33 size channels are also evenly spaced. The diameter of each new size channel is the diameter from the middle one of the original five size bins. The measurement error for x_{ij} after the combination is determined by the following equations:

$$D_P \leq 0.022\,\mu\text{m}, \ \sigma_{ij} = \sum_{h=5j-4}^{5j} \sigma'_{ih} \cdot \frac{1}{5}, \ j = 1, 2, \ldots, 11, \qquad [2]$$

$$D_P > 0.022\,\mu\text{m}, \ \sigma_{ij} = [\max(x_{i,5j-4}, \ldots, x_{i,5j}) - \min(x_{i,5j-4}, \ldots, x_{i,5j})] \times 0.5, \quad j = 12, 13, \ldots, 33. \qquad [3]$$

For the 12 to 33 new size channels, the variation of the concentration is smooth. Since the variations between neighboring size bins are small, the concentrations of the five consecutive bins can be regarded as five measurements of the same true value. According to the central limit theorem, a large number of measurements conform a normal distribution whose mean is the true concentration value no matter what distribution each measurement follows. The five measurement are five samples of this normal distribution. The average of the five concentrations is then the best estimation of the mean. In Equation (3), the width of this distribution is approximated by the difference of the maximum and minimum concentration values, and the half width is referred to as the standard deviation of the normal distribution, which is taken as the measurement error. If any of the five original concentration values (165 size bins) is a missing value, the corresponding measurement error is assigned as 4 times the new-formed concentration value (33 size intervals).

Figure 3 gives the mean number and volume size distributions averaged over all of the samples used in these analyses. Table 2 contains detailed information about the maximum, minimum, and mean values for each combined size channel.

POSITIVE MATRIX FACTORIZATION (PMF)

The basic receptor model is expressed as

$$X = GF + E, \qquad [4]$$

Table 2
Minimum, maximum, and mean value of the 33 size channels

	Volume concentration $dN_V/d\log D_P$ ($\mu m^3/cm^3$)			Number concentration $dN/d\log D_P$ (cm^{-3})		
$D_P(\mu m)$	Minimum	Maximum	Mean	Minimum	Maximum	Mean
0.003	0	0.0144	0.0006	0	8.21e + 05	3.97e + 04
0.004	0	0.0122	0.0006	0	4.22e + 05	1.75e + 04
0.005	0	0.0066	0.0006	0	1.34e + 05	1.15e + 04
0.006	0	0.0096	0.0008	0	1.06e + 05	9.74e + 03
0.007	0	0.0216	0.0014	0	1.39e + 05	9.72e + 03
0.008	0	0.031	0.0028	0	1.19e + 05	1.06e + 04
0.009	0.0002	0.0638	0.0052	214	1.37e + 05	1.17e + 04
0.011	0.0006	0.07	0.0094	776	9.43e + 04	1.22e + 04
0.014	0.0014	0.101	0.0162	998	7.72e + 04	1.23e + 04
0.016	0.0024	0.182	0.027	1043	8.02e + 04	1.19e + 04
0.019	0.0022	0.327	0.0452	563	8.53e + 04	1.16e + 04
0.023	0.006	1.28	0.0758	880	1.87e + 05	1.13e + 04
0.028	0.011	1.04	0.126	979	9.01e + 04	1.10e + 04
0.033	0.0168	1.88	0.217	857	9.50e + 04	1.10e + 04
0.040	0.0316	3.61	0.382	930	1.07e + 05	1.13e + 04
0.048	0.0724	5.75	0.670	1242	1.01e + 05	1.16e + 04
0.057	0.179	5.93	1.14	1786	6.13e + 04	1.15e + 04
0.069	0.365	9.10	1.90	2120	5.34e + 04	1.12e + 04
0.082	0.457	16.1	3.07	1600	5.50e + 04	1.05e + 04
0.098	0.530	27.0	4.86	1070	5.41e + 04	9.73e + 03
0.118	0.713	37.4	7.39	830	4.40e + 04	8.63e + 03
0.141	1.33	42.23	10.8	893	2.93e + 04	7.34e + 03
0.169	2.42	51.2	14.6	958	2.05e + 04	5.84e + 03
0.202	2.37	50.8	18.7	564	1.21e + 04	4.36e + 03
0.241	1.79	70.2	23.3	252	9.51e + 03	3.17e + 03
0.289	1.13	87.3	28.4	91.7	6.92e + 03	2.25e + 03
0.346	0.695	111.7	32.6	33.1	5.14e + 03	1.51e + 03
0.414	0.450	134	33.1	12.3	3.62e + 03	9.00e + 02
0.496	0.208	164	27.7	3.27	2.64e + 03	4.45e + 02
0.626	0.222	84.7	16.4	2.01	8.04e + 02	1.46e + 02
0.898	0.116	28.0	4.42	0.328	8.80e + 01	1.35e + 01
1.286	0.137	11.98	2.28	0.122	1.10e + 01	2.15e + 00
1.843	0.280	19.0	3.98	0.082	5.87e + 00	1.18e + 00

or in the element form

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}, \qquad [5]$$

where X is the matrix of observed data and the element x_{ij} is the concentration value of the ith sample at the jth size bin. G and F are, respectively, the source contributions and size distribution profiles of the sources that are unknown and to be estimated from the analysis. To be specific, g_{ik} is the concentration of particles from the kth source associated with the ith sample and f_{kj} is the size distribution associated with kth source. E is a matrix of residuals. The model is solved by a least-square method using the program PMF2 (two-way PMF) (Paatero 1997). The mathematical expressions are

$$\min_{G,F} Q \qquad [6]$$

and

$$Q = \left\| \frac{(X - GF)}{s} \right\|_{F,G}^{2} = \sum_i \sum_j \left(\frac{e_{ij}}{s_{ij}} \right)^2, \qquad [7]$$

where s_{ij} is the uncertainty of each x_{ij} value and the reciprocal of s_{ij} serves as the weight.

The details of the algorithm of PMF2 can be found elsewhere (Paatero 1997). The uncertainties inputted into PMF2 were computed based on the measurement errors with the expression

$$s_{ij} = \sigma_{ij} + C_3 \max(|x_{ij}|, |y_{ij}|), \quad [8]$$

where y_{ij} is the calculated value for x_{ij}, σ_{ij} is the measurement error estimated in the previous section, and C_3 is a dimensionless constant value between 0.1–0.2. The additional uncertainty beyond the measurement errors estimated above is included to take into account in part the variability of the source profiles. The size distribution of particles emitted from a source cannot be expected to be perfectly constant, and some additional variation in the fit needs to be allowed to accommodate this variability. PMF2 was run separately for number and volume size distributions. The results of PMF2 analysis should be rescaled to satisfy the mass apportionment conditions (Hopke et al. 1980). For the sake of completeness, a brief introduction is given, as shown in the following equation:

$$x_{ij} = \sum_{k=1}^{p} f_{ik} \cdot \frac{w_k}{w_k} \cdot g_{kj}. \quad [9]$$

The scaling constant in the above equation, w_k, is determined by regressing the total number or volume contribution against the estimated source contributions:

$$v_j = \sum_{k=1}^{p} w_k \cdot g_{kj}. \quad [10]$$

The G values discussed in this article are all rescaled for a number apportionment by this method.

CONDITIONAL PROBABILITY FUNCTION (CPF)

A conditional probability function (Ashbaugh et al. 1985; Kim et al. 2004) was calculated with the source contributions obtained by PMF2 and wind direction values as the following equation:

$$\mathrm{CPF} = \frac{m_{\Delta\theta}}{n_{\Delta\theta}}, \quad [11]$$

where $m_{\Delta\theta}$ is the number of occurrences in the direction sector that exceed the threshold, upper 25% of the fractional contribution from each source; and $n_{\Delta\theta}$ is the total number of wind occurrences in the same direction sector. The fractional contribution is used instead of the volume or number contribution to avoid the influence of atmospheric dilution. Each direction sector is set 10 degrees, and thus there are 36 direction sectors. Those winds below 1.0 m/s are excluded from this study. The sources are thought to be located in the direction sectors with high CPF values. It has to be pointed out that the CPF is not so dependable in finding the directions of far sources since the air mass may travel through a circuitous pathway.

The wind directions are concentrated within a few adjacent sectors; other direction sectors have a very low frequency of occurrence, and some have zero occurance. Those high CPF values with low values of $n_{\Delta\theta}$ are not reliable because of the uncertainties between the calculated contributions and the real contribution of each source. A threshold criteria, n_c, is needed. When $n_c > n_{\Delta\theta}$, the directional sector is neglected and CPF value is set to zero. In this study, n_c is arbitrarily chosen to be 10. The proper choice of the n_c value remains an open question.

RESULTS AND DISCUSSION

Different numbers of factors and *Fpeak* values have been explored to obtain the most meaningful results. *Fpeak* is a parameter used in PMF2 to control the rotation. When the *Fpeak* is set to a positive value, the program adds one G vector to another and subtracts the corresponding F vectors from each other to obtain a more physically realistic solution. The details of the mathematical process can be found in Paatero et al. (2002). Five factors were selected, and the *Fpeak* value was set to 1.2 for number size-distribution and 0.2 for volume size-distribution analyses.

If the uncertainties are well estimated, the expected Q value is approximately equal to the element number of X matrix, which is 50,000 in this study. Since larger uncertainties were constructed (C3 is chosen as 0.2), the final Q value is 25,013, much smaller than the expected Q value. The Q value decreases when the factor number increases. When the factor number increases from 5 to 6, the Q value is not improved as much as it is when the factor number increases from 4 to 5. Therefore, the final factor number was chosen as 5.

Only the variation and time series analysis of the contribution values for the number concentration are presented here since the instruments directly measured the number concentration. The number and volume contributions of each factor were found to have quite similar temporal variations. For a given source, the size distribution profile for number concentration can be converted to obtain the profile for volume concentration, as indicated in Equation (12). In this equation, N stands for number, V stands for volume, d_j is the diameter for the jth size channel and f_j is the fraction for the jth channel. Multiplying the number contribution by the constant C provides the corresponding volume contribution and vice versa. Since the source profiles and contributions for the number and volume distributions of each source are obtained through PMF with different rotations, the conversion is an approximate one. The match of the G values by number and volume contribution is also based on the correlation coefficient of the G values with gas-phase data and frequency properties:

$$f_j^V = \frac{f_j^N d_j^3 \cdot \pi/6}{\sum_{j=1}^{n} f_j^N d_j^3 \cdot \pi/6} = \frac{f_j^N d_j^3}{\sum_{j=1}^{n} f_j^N d_j^3} = \frac{f_j^N d_j^3}{C}. \quad [12]$$

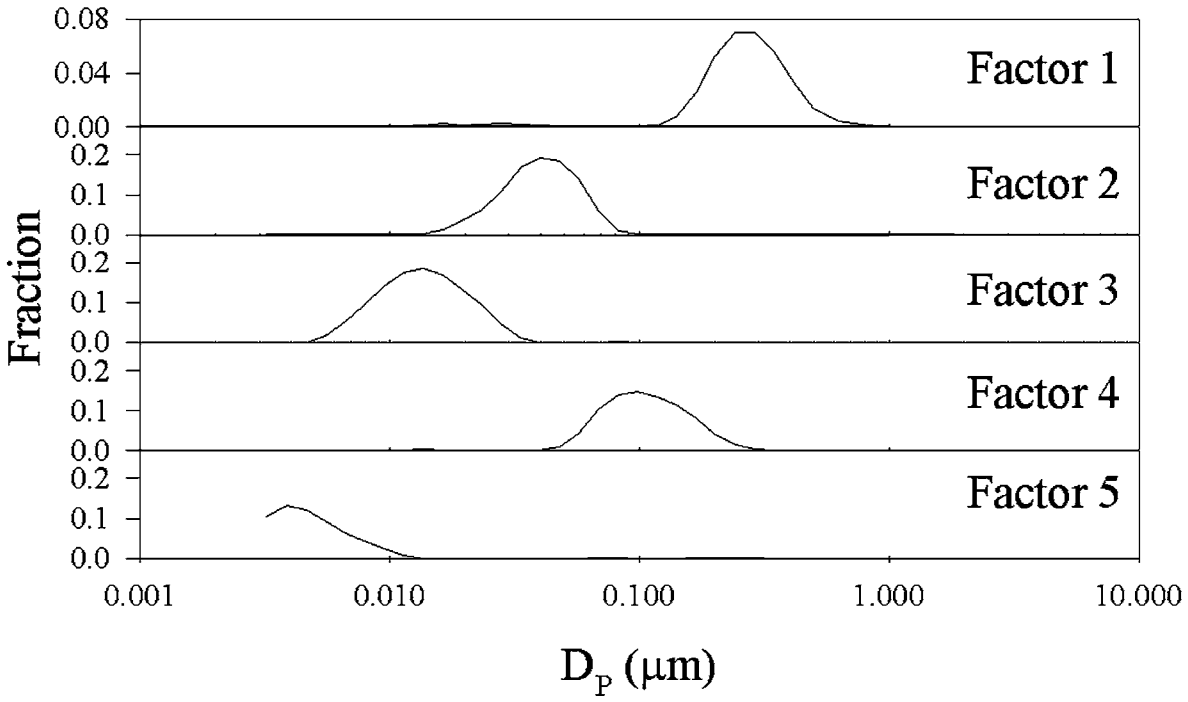


Figure 4. Normalized feature profile for number size distribution.

The sources were identified based on the profiles for both the number and volume size distributions of the factors, the time frequency properties of the contribution of each factor (Fourier analysis of *G* values), and the correlations of the *G* values with the gas phase and some composition data. The Fourier analysis identifies the strong frequencies in the data associated with periodic behavior, such as reoccurrences of rush hour traffic each day.

The number and volume size-distribution profiles of the five factors are shown in Figures 4 and 5, respectively. However,

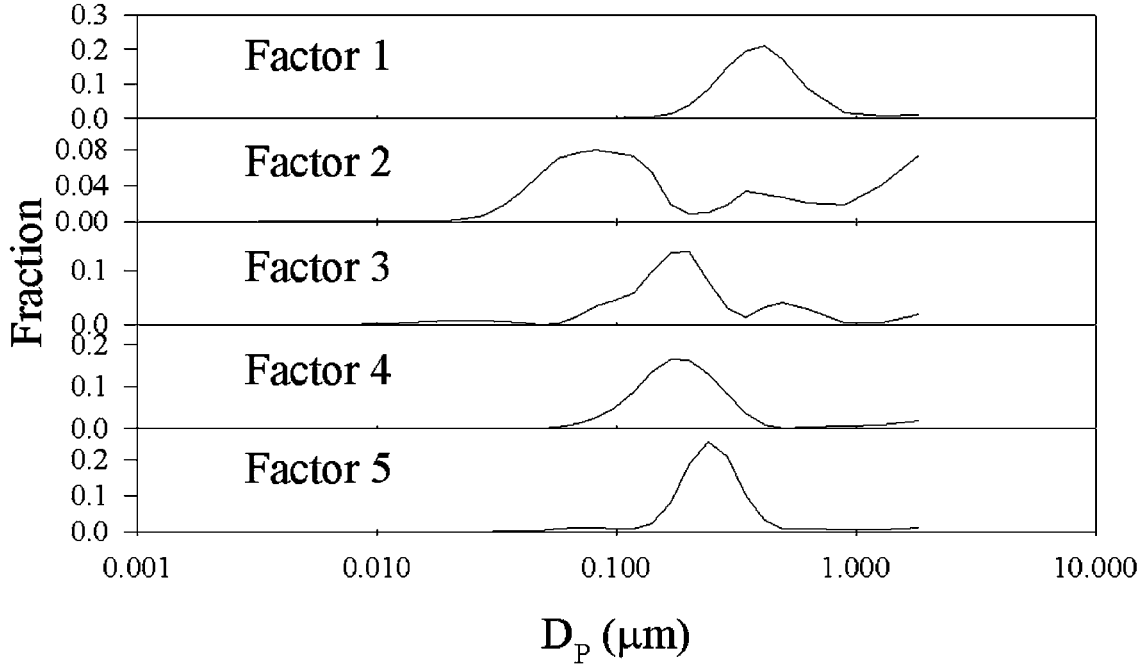


Figure 5. Normalized feature profile for volume size distribution.

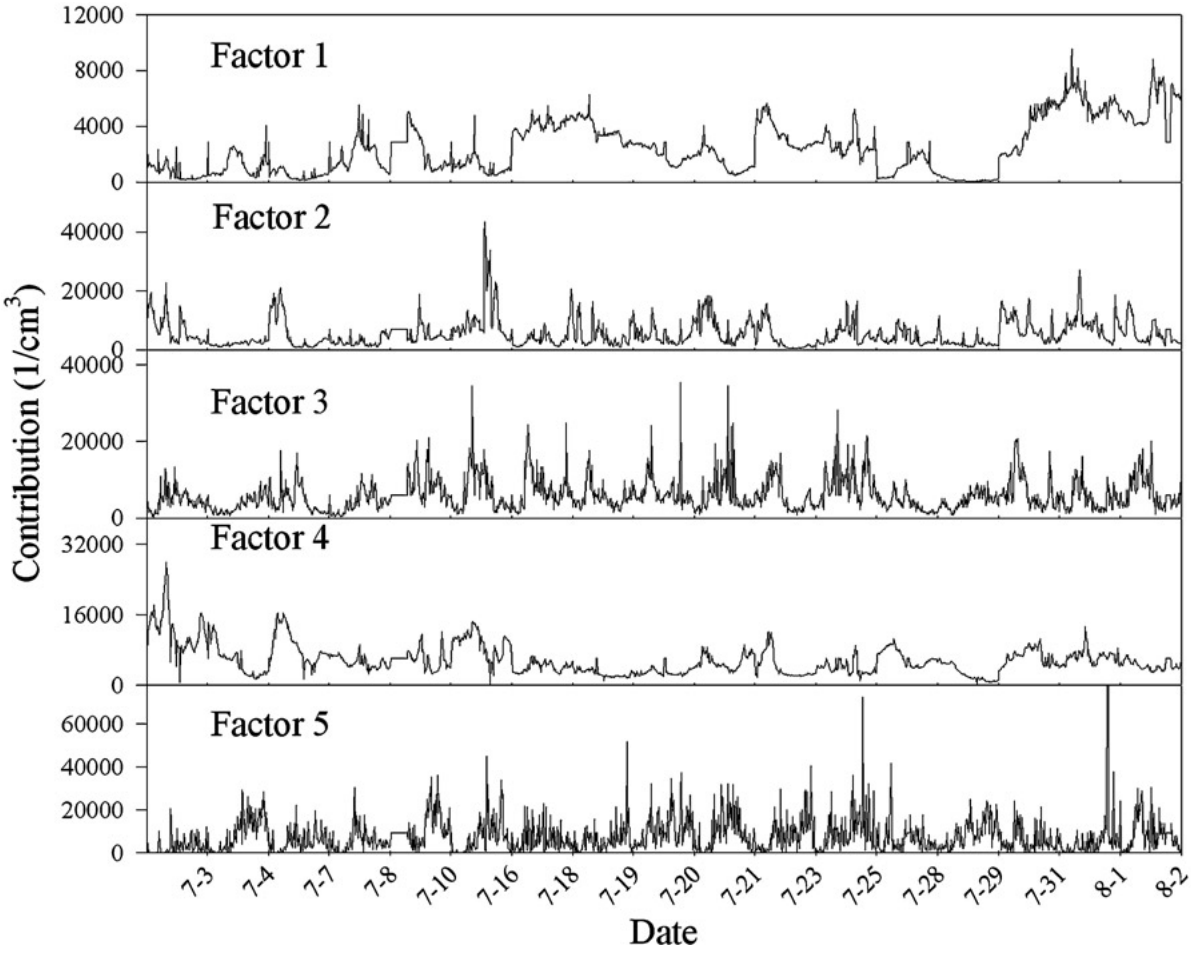


Figure 6. Number contribution for each factor.

size-distribution profiles alone are not sufficient since more than one source may have similar size-distribution profiles. The number of contributions for each factor are presented in Figure 6. The high temporal resolution of the experimental data make it possible to thoroughly investigate the frequency behavior of the contributions. It should be noted that those frequency peaks obtained by Fourier analysis do not absolutely reflect the actual temporal variations except those which are integral multiple of $1/24\ h^{-1}$ because of the discontinuity of the dates resulting from the elimination of the nucleation events. However, the results of the Fourier transformation can still be used to identify daily patterns in the factors and to investigate the frequency properties qualitatively. The results of the Fourier transformation of the G values (number) are shown in Figure 7. The results of the correlations of the G values (number) with the gas-phase data and CPF function are presented in Table 3 and Figure 8, respectively. In Figure 8, the first subfigure indicates the number of occurrences in each wind direction sector is presented. Figure 9 shows about the CPF of gases for the same days as the particle. Because of their different temporal resolutions, the

Table 3
Correlations of G (number) factors with gas-phase data and particle-mass data

	O_3	NO	NO_x	SO_2	CO	$PM_{2.5}$	Sulfate	OC	EC
Factor 1	0.07	0	0.146	0.332	0.148	0.907	0.858	0.527	0.463
Factor 2	−0.24	0.394	0.52	0.21	0.495	0.17	0.06	0.373	0.643
Factor 3	−0.19	0.198	0.31	0.01	0.145	0.113	0.04	0.09	0.394
Factor 4	−0.31	0.658	0.666	0.367	0.474	0.09	−0.1	0.289	0.492
Factor 5	0.221	−0.19	−0.22	−0.12	−0.2	−0.16	−0.14	−0.43	−0.21

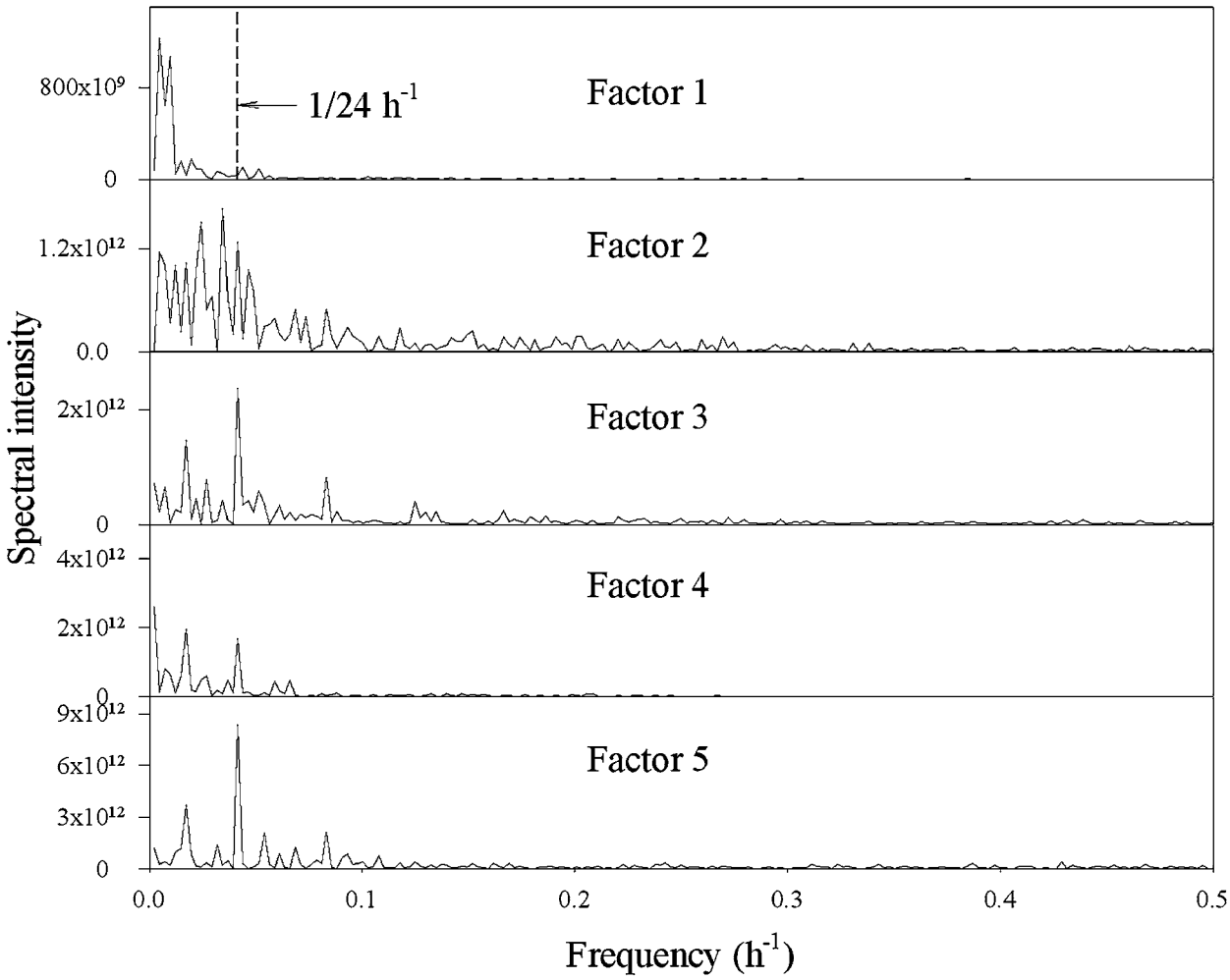


Figure 7. Fourier transformation of number contributions.

gas-phase data set and wind data were averaged to 30 min, which caused some minor differences in the wind profiles of the two figures.

Factor 1 has a peak around 0.3 μm and high correlations with $PM_{2.5}$ and sulfate. The Fourier transformation shows that this factor has no obvious frequency peak. The total contribution to the receptor site does not change periodically with time. This information indicates a secondary aerosol. This factor is from distant sources; the particles were produced far from the receptor site and have accumulated secondary aerosol components (sulfate and organics) growing from their original size. The CPF function shows that these particles are from south of the site. Comparisons of the $PM_{2.5}$ concentrations measured in Pittsburgh with the measurements in satellite sites around the city suggest that during July 2001 more than 80% of the Pittsburgh $PM_{2.5}$ was transported into the city (Wittig et al. 2004). The present analysis also indicates that this factor is responsible for most of the aerosol volume as indicated in Table 4.

Factor 2 has a number mode at 0.04 μm, a volume mode around 0.09 μm, and is correlated with NO, NO_x, and CO. The correlation with EC is also strong, suggesting emissions from diesel trucks on the highway. The fact that the heavy-duty diesels also emit particles having this size range (Shi 2000) is consistent with the correlation with EC. There is a peak at 1/24 h^{-1}, indicating a daily pattern. The CPF of the NO, NO_x, and CO indicate that they are from the south, as shown in Figure 8, probably

Table 4
Average volume and number contribution of the sources

Factor	Volume ($\mu m^3/cm^3$)	Number ($\#/cm^3$)
1	14.09	2445
2	0.93	5411
3	0.36	5882
4	3.26	5741
5	0.24	7659

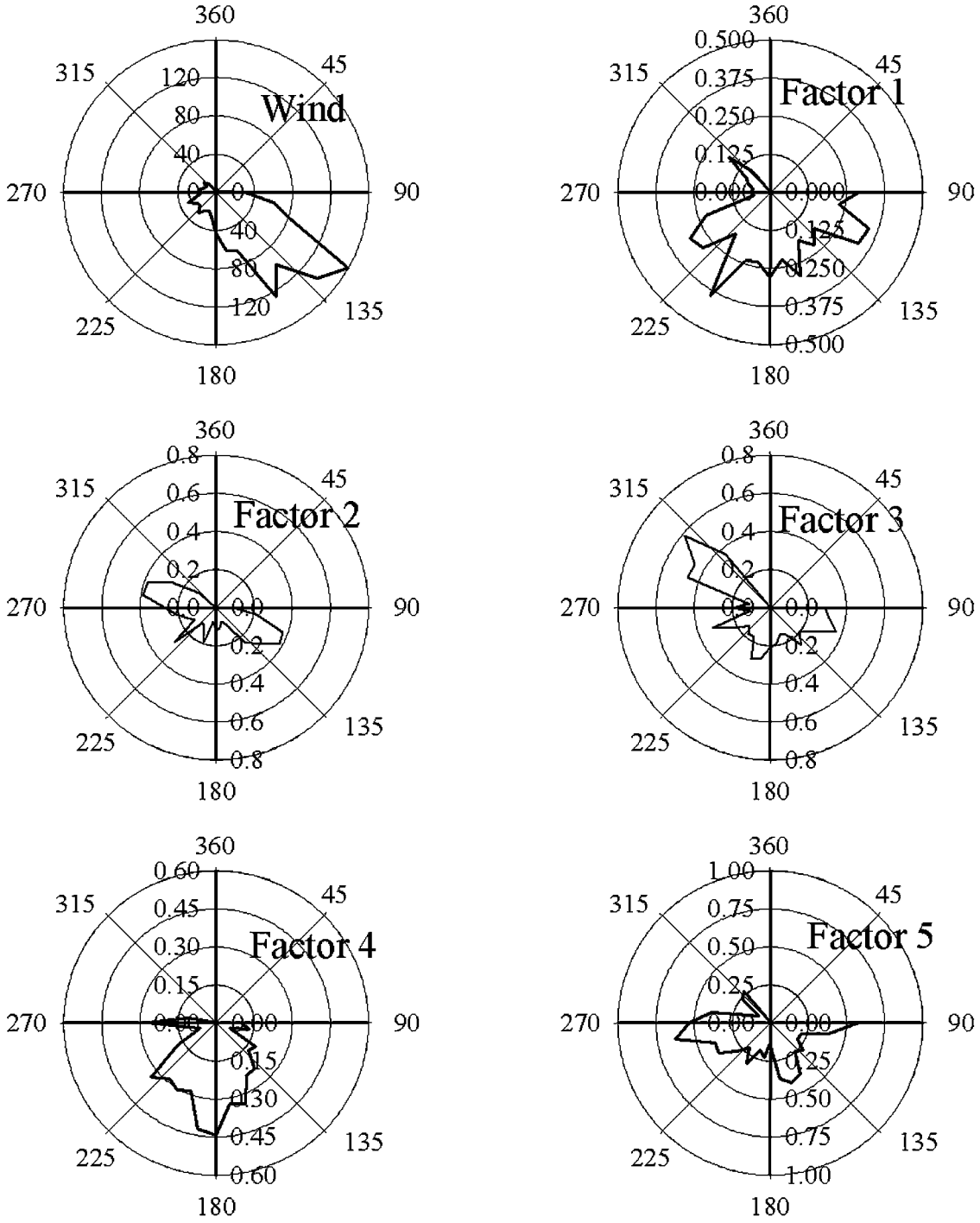


Figure 8. The conditional probability function for the number contributions.

from highway 376, 1 mile south to the site, extending from west to east. This factor may contain particles from highway 376. The distance from the source may explain the lack of a clear "transportation" diurnal profile. Meteorological variables such as mixing height and wind speed influence the profile more than the actual traffic pattern.

The number and volume modes are 0.015 and 0.2 μm respectively, for Factor 3. Factor 3 is only weakly correlated with NO and NO_x, but it has an obvious frequency peak at 1/24 h^{-1}. The two rush-hour peaks can be identified in Figure 10. Therefore, Factor 3 is assigned as a traffic aerosol. Because the particles are small (15 nm number size mode), this factor may be from the roads in the immediate vicinity of the site (Forbes Avenue, Schenley Drive, etc.), as shown in Figure 1. There is almost no diesel traffic in these places, which could explain the lack of correlation with EC and the weak correlation with NO_x. The

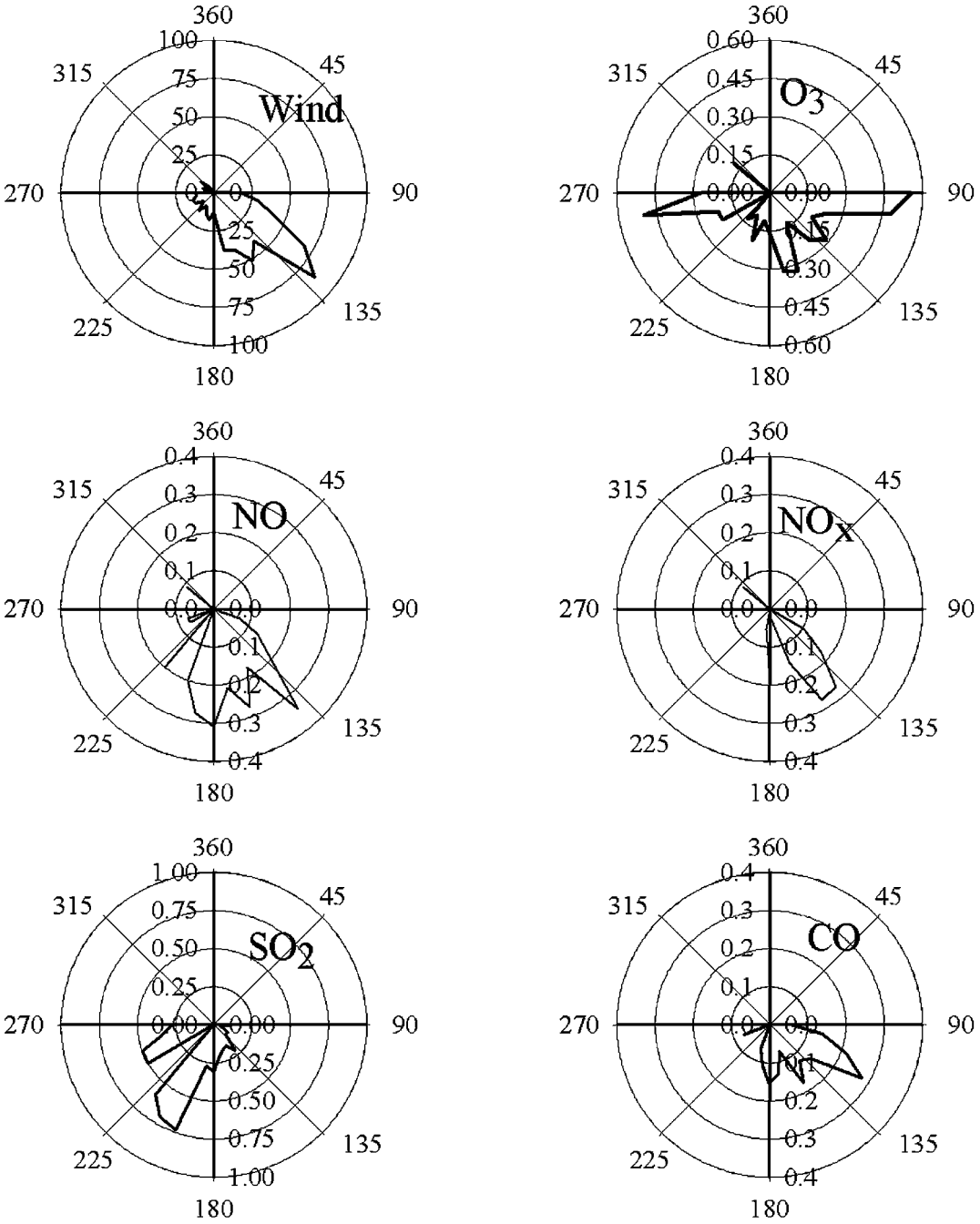


Figure 9. The conditional probability function for gases.

CPF has peaks at east and northwest, the directions of Forbes Avenue and other roads in the vicinity. The good diurnal traffic profile, as indicated in Figure 10, also suggests good proximity to the source. Figure 11 indicates the temporal variations of the gases (NO, NO_x, and CO). The gases sometimes have high values at night, which could be caused by the inversion layer. These phenomena also occurred for factor 2. The particles captured by the inversion layer were probably particles from the local traffic, which also weakens the "transportation" profile of factor 2. The traffic volume is lower on these smaller roads than on the highway, leading to a small contribution to the primary pollutants, which is also consistent with the small volume contribution of Factor 3. The low overall concentration contribution may explain the weak correlation of factor 3 with NO, NO_x, CO, and EC in another aspect. Conversely, factor 2 is named *grown particles* and *remote traffic* and factor 3 is titled *local traffic*.

Factor 4 has high correlations with NO, NO_x, and CO but has no peak at 1/24 h^{-1}. This factor may be related to combustion

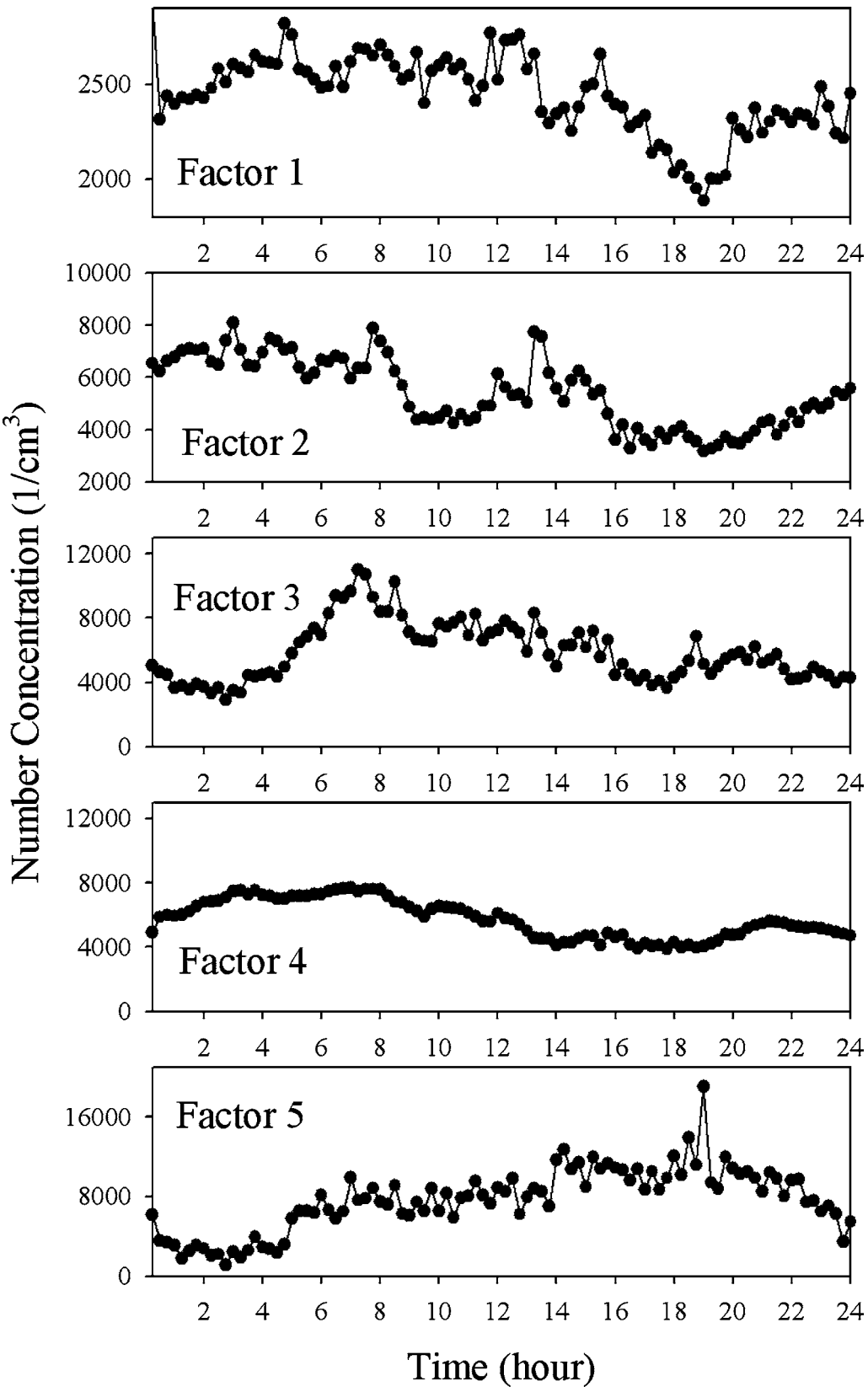


Figure 10. The average diurnal variation of the factors.

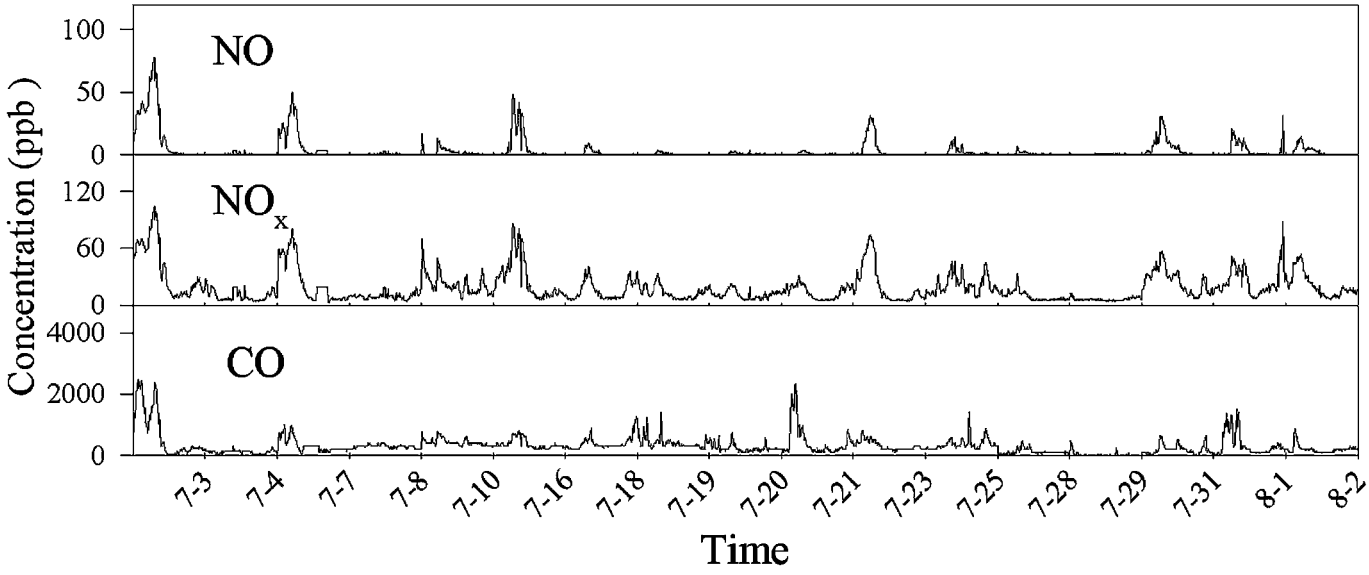


Figure 11. Temporal variations of NO, NO_x, and CO.

sources such as a power station or biomass fires (Morawska et al. 1998), but it is unlikely to be motor vehicle emissions. The mode around 0.1 μm for the number distribution was suggested as representative of vegetation-burning–influenced aerosols by Morawska et al. (1999). The location of the volume mode is also in accordance with the reported measurement of wood burning (Morawska et al. 1999; Le Canut et al. 1996).

Wehner et al. (1999) reported that the number size distribution of a coal-fired heating power plant has a mode between 45 and 100 nm. The mass mode of coal combustion is at 10 μm (Lind et al. 1994), which is beyond the size range of this study. Factor 4 may be from local combustion sources and possibly vegetation burning as well. The CPF shows the sources of factor 4 are located to the south. The correlation of factor 4 with SO_2 indicates coal combustion sources. Factor 4 has no correlation with sulfate, which means SO_2 has not had enough time to oxidize to sulfate. Figure 12 is a map of the major emission sources in the eastern US. There are two major power plants to the south of Pittsburgh, close to the border with West Virginia, at a distance of less than 100 km. A calculation is provided in the appendix to estimate the conversion rate of SO_2. The result shows that within this range most SO_2 is not reacted.

Factor 5 has a number mode at 3 nm and a volume mode at 0.25 μm. Factor 5 has a peak at 1/24 h^{-1} but no correlation with NO, NO_x, or CO. The contribution reaches a peak at mid- to late afternoon. The small particles, several nanometers large, with high number concentration, might be caused by nucleation events that were often observed to have peak activities in the afternoon. Photochemical reaction intensity increases in the afternoon and oxidants are produced, which can explain the positive correlation of factor 5 with ozone. These oxidants oxidize SO_2 to form sulfuric acid which then nucleates, probably with water and possibly with ammonia. The volume mode at 0.25 μm, corresponding to the small "bumps" in the number distribution at the same size range, cannot be from nucleation. This volume mode can only be created locally together with the new particles by condensation of sulfuric acid and organic compounds. The anticorrelation of factor 5 with OC could be the result of the fact that most OC is primary. Factor 5 is called "sparse nucleation" since the nucleation occurs over limited time intervals with insufficient material present to permit growth in size up to the accumulation mode range (the excluded nucleation events). Factor 5 may have a composition similar to factor 1, and they are both secondary aerosol. The difference is that factor 5 is new-formed particles that have a short life time, while factor 1 are aged particles and more stable in the air.

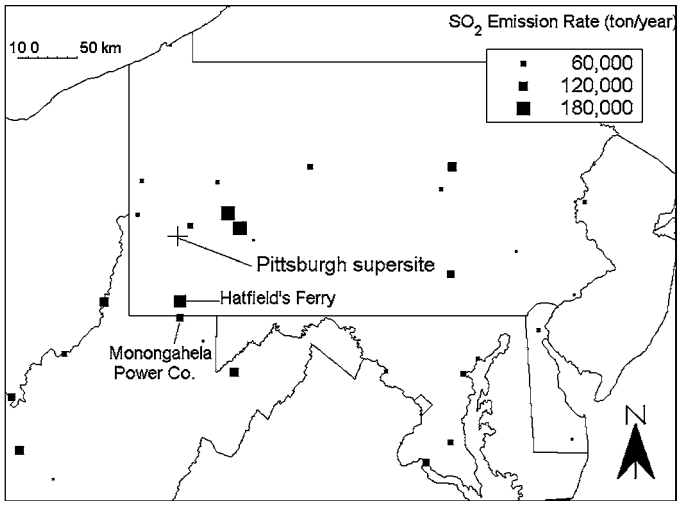


Figure 12. Map of region around the Pittsburgh area showing locations of power plants.

Figure 5 shows that factors 1 and 4 have small peaks at high frequencies (over 0.1 h^{-1}). This result may suggest that factors 1 and 4 form a kind of background that is stable and has no regular periodicity.

Table 4 summarizes the average contribution of those factors to the volume concentration and number concentration. The number concentration contributions are at the same order of magnitude, but the volume contributions are quite different. The volume contribution is mainly composed of secondary aerosol and stationary combustion.

CONCLUSIONS

Positive matrix factorization was used to explore size-distribution data from Pittsburgh for source apportionment. From the investigation of the number and volume modes of the size distributions of the factors, the frequency properties, the correlations of the G factors with gas phase data, and the CPF functions, five sources are identified: sparse nucleation, local traffic, stationary combustion, grown particles and remote traffic, and secondary aerosol. Although these factors are thus titled, they may still be contaminated by other unknown sources, especially the two traffic factors. Future analysis of the composition size data, such as micro-orifice uniform deposit impactor (MOUDI) and single-particle spectrum data, might provide additional, more specific source information.

NOTATION

C_3	Coefficient used for a heuristic uncertainty in PMF2.
D	Particle diameter, μm
E, e	Regression residual
F, f	Size distribution profile associated with each source
G, g	Source contribution, 1/cm^3 or μm^3/cm^3
m	The number of occurrences in the direction sector that exceed the threshold
n	The total number of wind occurrences
N	Number concentration, 1/cm^3
Q	Sum of the residual squares
s	Uncertainty for each datum point used in PMF2
t	Time
u	Instrumental error for small particles, 1/cm^3
V	Volume concentration, μm^3/cm^3
v	Total mass, number or volume concentration for each sample
w	A constant for mass, number, or volume apportionment
X, x	Data to be regressed
y	Regressed concentration value used in PMF2

Greek Letters

θ	Wind direction
σ	Measurement error for the combined size bins, 1/cm^3
σ'	Measurement error for the original-size bins, 1/cm^3

REFERENCES

Ashbaugh, L. L., Malm, W. C., and Sadeh, W. Z. (1985). A Residence Time Probability Analysis of Sulfur Concentrations at Ground Canyon National Park, *Atmos. Environ.* 19(8):1263–1270.

Chueinta, W., Hopke, P. K., and Paatero, P. (2000). Investigation of Sources of Atmospheric Aerosol at Urban and Suburban Residential Area in Thailand by Positive Matrix Factorization, *Atmos. Environ.* 34:3319–3329.

Hopke, P. K., Lamb, R. E., and Natusch, D. F. C. (1980). Multielemental Characterization of Urban Roadway Dust. *Environ. Sci. Technol.* 14:164–172.

Khlystov, A., Stanier, C., and Pandis, S. N. (2004). An Algorithm for Combining Electrical Mobility and Aerodynamic Size Distributions When Measuring Ambient Aerosol, *Aerosol Sci. Technol.* 38:229–238.

Kim, E., Hopke, P. K., and Larson, T. V. (2004). Analysis of Ambient Particle Size Distributions Using Unmix and Positive Matrix Factorization, *Environ. Sci. Technol.* 38:202–209.

Le Canut, P., Andreae, M. O., Harris, G. W., Wienhold, F. G., and Zenker, T. (1996). Airborne Studies of Emissions From Savanna Fires in Southern Africa: Aerosol Emissions Measured with a Laser Optical Particle Counter, *J. Geophys. Res.* 101(D19):23615–23630.

Lee, E., Chun, C. K., and Paatero, P. (1999). Application of Positive Matrix Factorization in Source Apportionment of Particulate Pollutants, *Atmos. Environ.* 33:3201–3212.

Lind, T. M., Kauppinen, E. I., Jokiniemi, J. K., Lillieblad, L., and Klippel, N. (1994). Coal Combustion Aerosol Particle Size Distribution Determination Using Low-Pressure Impactor and CCSEM Methods, *J. Aerosol Sci.* 25(Suppl. 1):S327–S328.

Morawska, L., Thomas, S., Jamriska, M., and Johnson, G. (1999). The Modality of Particle Size Distributions of Environmental Aerosols, *Atmos. Environ.* 33:4401–4411.

Morawaska, L., Thomas, S., Bofinger, N., Wainwright, D., and Neale, D. (1998). Comprehensive Characterization of Aerosols in a Subtropical Urban Atmosphere: Particle Size Distribution and Correlation with Gaseous Pollutants, *Atmos. Environ.* 32:2467–2478.

Paatero, P. (1997). Least Squares Formulation of Robust, Non-Negative Factor Analysis, *Chemometrics Intelligent Lab. Syst.* 37:23–55.

Paatero, P., Hopke, P. K., Song, X. H., and Ramadan, Z. (2002) Understanding and Controlling Rotations in Factor Analytic Models, *Chemometrics Intelligent Lab. Syst.* 60:253–264.

Peters, A., Wichmann, H. E., Tuch, T., Heinrich, J., and Heyder, J. (1997). Respiratory Effects Are Associated with the Number of Ultrafine Particles, *Am. J. Resp. Crit. Care Med.* 155:1376–1383.

Polissar, A. V., Hopke, P. K., and Poirot, R. L. (2001). Atmospheric Aerosol over Vermont: Chemical Composition and Sources, *Environ. Sci. Technol.* 35:4604–4621.

Ramadan, Z., Song, X. H., and Hopke, P. K. (2000). Identification of Sources of Phoenix Aerosol by Positive Matrix Factorization, *J. Air Waste Manag. Assoc.* 50:1308–1320.

Reischl, G. P. (1991). Measurement of Ambient Aerosols by the Differential Mobility Analyzer Method: Concepts and Realization Criteria for the Size Range Between 2 and 500 nm, *Aerosol Sci. Technol.* 14:5–24.

Ruuskanen, J., Tuch, T., Brink, H. T., Peters, A., Khlystov, A., Mirme, A., Kos, G. P. A., Brunekreef, B., Wichmann, H. E., Buzorius, G., Vallius, M., Kreyling, W. G., and Pekkanen. J. (2001). Concentrations of Ultrafine, Fine and PM2.5 Particles in Three European Cities, *Atmos. Environ.* 35:3729–3738.

Seinfeld, J. H., and Pandis, S. N. (1998). *Atmospheric Chemistry and Physics*. John Wiley & Sons, New York, p. 365.

Shi, J. P., David, M., and Harrison, R. M. (2000). Characterization of Particles from a Current Technology Heavy-Duty Diesel Engine, *Environ. Sci. Technol.* 34:748–755.

Song, X. H., Polissar, A. V., and Hopke, P. K. (2001). Source of Fine Particle Composition in the Northeastern U.S., *Atmos. Environ.* 35:5277–5286.

Stanier, C., Khlystov, A., and Pandis, S. N. (2004). Nucleation Events During the Pittsburgh Air Quality Study: Description and Relation to Key Meteorological, Gas Phase, and Aerosol Parameters, *Aerosol Sci. Technol*. 38:253–264.

van Bree, L., and Cassee, F. R. (2000). *A Critical Review of Potentially Causative PM Properties and Mechanisms Associated with Health Effects, National Institute of Public Health and the Environment (RIVM) Research Report*, report no. 650010015, Bilthoven, the Netherlands.

Venkataraman, C., and Kao, A. S. (1999). Comparison of Particles Lung Doses from the Fine and Coarse Fraction of Urban PM-10 Aerosols, *Inhalation Toxic*. 11:151–169.

Wahlin, P., Palmgren, F., Dingenen, R. V., and Raes, F. (2001). Experimental Studies of Ultrafine Particles in Streets and the Relationship to Traffic, *Atmos. Environ*. 35(S1):S63–S69.

Wehner, B., Bond, T. C., Birmili, W., Heintzenberg, J., Wiedensohler, A., and Charlson, R. J. (1999). Climate-Relevant Particulate Emission Characteristics of a Coal Fired Heating Plant, *Environ. Sci. Technol*. 33:3881–3886.

Wittig, B., Anderson, N., Khlystov, A. Y., Davidson, C., Robinson, A., and Pandis, S. N. (2004). The Pittsburgh Air Quality Study (PAQS), *Atmos. Environ*. in press.

Wong, S. S. M. (1997). *Computational Methods in Physics and Engineering*, 2nd ed. World Scientific Publishing Co. Pte. Ltd., Singapore, p. 237.

Xie, Y. L., Hopke, P. K., Paatero, P., Barrie, L. A., and Li, S. M. (1999). Identification of Source Nature and Seasonal Variations of Arctic Aerosol by Positive Matrix Factorization, *J. Atmos. Sci*. 56:249–260.

APPENDIX: AN ESTIMATION OF SO_2 CONVERSION RATE

The distance of the coal power plant and other combustion sources emitting SO_2 is 10–100 km. Suppose the wind speed is 10 km/h. The traveling time of the pollutant is 1–10 h.

The following equation is used to estimate the conversion rate:

$$\frac{dC}{C} = -Rdt, \qquad \text{[A1]}$$

where C is the concentration of SO_2 and R is the reaction rate with a unit of $\%h^{-1}$. The conversion rate of SO_2 X is

$$X = 1 - \frac{C}{C_0} = 1 - \exp(-Rt). \qquad \text{[A2]}$$

A typical gas-phase SO_2 oxidation rate by OH is on the order of $1\%h^{-1}$ (Seinfeld and Pandis 1998). By substituting $R = 1$ into Equation (A2), the conversion rate is $0.006 \sim 1\%$.

The oxidation of SO_2 is dominated by dissolved hydrogen peroxide for $pH < 5$. Assume that the frequency of cloud occurrence is 0.4, the areal coverage by cloud is 0.2, and the average liquid water content of cloud is 0.2 g/m^3. In Seinfeld and Pandis (1998) the oxidation rate is estimated at around 500%/h when the liquid water content is 1 g/m^3 and hydrogen peroxide has a concentration of 1 ppb. Thus the reaction rate of SO_2 under the former assumptions should be $0.4 \times 0.2 \times 0.2 \times 500\%/h = 8\%/h$.

In this situation, the conversion rate of SO_2 is 7.7–55%. In fact, the production of sulfuric acid decreases the pH value, which causes less dissolved SO_2, so the conversion rate should be lower.