Final Report
ARB Contract 01-348

Initial Exploration of Advanced Data Analysis Methods to Assist Air Quality Management

Prepared by
Philip K. Hopke

The Bayard D. Clarkson Distinguished Professor
Center for Air Resources Engineering and Science
and
Department of Chemical Engineering
Clarkson University
Box 5708
Potsdam, NY 13699-5708

Legal Notice

This report was prepared as a result of work sponsored by the California Air Resources Board. It does not necessarily represent the views of the Commission, its employees, or the State of California. The Commission, the State of California, its employees, contractors, and subcontractors make no warranty, express or implied, and assume no legal liability for the information in this report; nor does any party represent that the use of this information will not infringe upon privately owned rights. This report has not been approved or disapproved by the Board nor has the Board passed upon the accuracy or adequacy of this information in this report.

Abstract

This project had two major data analysis tasks. The first task was to perform advanced factor analysis using Positive Matrix Factorization (PMF) on three sets of IMPROVE data, Crater Lake National Park (CRLA), Lassen Volcano National Park (LAVO), and San Gorgonio National Wilderness (SAGO). Two of these IMPROVE sites, CRLA and LAVO, are at relatively high altitude and the objective is to separate and quantify the influence of Asian dust on the observed mass concentrations. Among the sources resolved at the two sites, six are common. These six sources exhibit not only similar chemical compositions, but also similar seasonal variations at both sites. The Asian dust represented by Al, Ca, Fe, NO_3, S, K, and Ti. with strong seasonal variation; secondary sulfate with a high concentration of S and strong seasonal variation correlated with the Asian dust; wood smoke represented by organic carbon (OC), elemental carbon (EC) and K; sea salt with the high concentrations of Na, S and $NO_{3;}$ nitrate dominated by NO_3 and motor vehicle with high concentrations of OC, EC and dust elements. A incinerator source with the presence of Cu and Zn also was resolved from Crater Lake site. Generally, most of the sources at these two sites showed similar chemical composition profiles and seasonal variation patterns. The source profile of Asian Dust resolved from this study agreed reasonably well with the source characteristics found in other Asian Dust studies.

The third site is downwind of Los Angeles (SAGO) and the primary objective is to determine if gasoline and diesel emissions can be separately identified and quantified using these chemically speciated data. The results demonstrate the feasibility of separating diesel/gasoline emission profiles based on concentration data including OC/EC fractions. Also in the analysis of these data, two crustal factors were identified with one being associated with local suspended soil and the other being associated with transported Asian desert dust.

The second task is to expand the existing capabilities of Aerosol Time-of-Flight Mass Spectrometry by ascertaining our ability to apply factor analysis to separate diesel from gasoline motor vehicle emissions and to develop and test calibration models that permits the estimation of the composition of the bulk ambient aerosol composition from single particle data. In the study of data from Fresno, 52 samples were created to build a calibration model. Compared with an earlier study (Fergenson, et al., 2001), significant improvements were obtained in this work, which fully demonstrated the ability of the calibration model based on ART-2a and PLS to estimate the chemical composition from ATOFMS data and also provided a good base to testing the transferability of calibration models of neighbor sites. In addition, some important steps to building a successful calibration mode, like how to determine the PLS components number, are presented in detail, and the corresponding guidance is provided.

In order to use single particle data obtained from ATOFMS measurements in PMF models, it is essential to have effective uncertainty estimates for the numbers of particles in each identified particle class in each time interval sample. An approach to developing these uncertainties is presented. Data from Fresno has been analyzed by PMF and results are presented. However, the interpretation of these results is currently incomplete and will require collaboration with Prof. Prather of the University of California, San Diego over the next several months to produce final, interpreted results.

Table of Contents

Abstract iii

INTRODUCTION 1

OBJECTIVES 2

ORIGINS OF FINE AEROSOL MASS USING PMF 3
Sampling and Analyses 3
Data Analyses 4
Results and Discussion 8

PREDICTING BULK AMBIENT AEROSOL COMPOSITIONS FROM ATOFMS DATA . . 27
Introduction 27
Method Description 28
Description of ART-2a 28
Description of PLSR 29
Data Treatment and Analysis 30

FUTURE STUDIES 40
Bootstrap Uncertainty Estimation 40
PMF Studies 41

REFERENCES 41

INTRODUCTION

Both Federal and State law require California to control concentrations of airborne particulate matter (aerosols) to protect public health and prevent visibility impairment. In order to accomplish this, ARB must identify the sources of ambient aerosols so that appropriate control programs can be devised. More specifically with the new $PM_{2.5}$ standards, long range transport of particles into California represents a significant contribution to the background particle concentrations. These particles can include materials from both natural and anthropogenic sources including transport across the Pacific Ocean. These source must be identified so that they can be properly accounted in setting air quality goals and developing control programs.

Ambient aerosols are complex mixtures of material derived from multiple natural and anthropogenic sources both near and distant from sampling sites. Daily and seasonal variations in source strength and meteorological conditions cause the composition and concentration of aerosols at any one sampling site to vary over time. This project proposes to use a newly developed statistical technique, Positive Matrix Factorization (PMF) to exploit the variation in long time series of aerosol data to resolve ambient mixtures into their component sources. As part of this project, this technology will be transferred to the ARB staff

Finally, ARB along with other agencies have supported the development of important new monitoring tools such as the Aerosol Time-of-Flight Mass Spectrometer. This instrument developed originally at the University of California at Riverside by Prof. Kimberly Prather provides qualitative characterization of individual particles in real-time. These very large data bases require sophisticated data analysis tools to fully utilize information content of this rich data base. In a prior ARB-supported project, we showed that these particles could be sorted with a dynamic classification tool, the Adaptive Resonance Theory Neural Network. The mass fractions in the identified classes can be used as quantitative variables. Thus, there is the potential for the instrument to provide real-time information on the ambient aerosol composition as well as providing quantitative information on the relative contributions of sources to the ambient aerosol mass. The work under this proposal will continue the development and testing of these data analysis tools that should enhance the substantial investment that has already made in the development of this important aerosol monitoring instrument.

OBJECTIVES

This project will demonstrate the utility of Positive Matrix Factorization (PMF) analyses to resolve the sources of ambient aerosols. Two priority areas have been selected for study: the contribution of Asian sources to “background” aerosols in northern California, and discriminating between the contributions of gasoline and diesel engines to ambient aerosols sampled in California’s Central Valley. In addition to distinct research questions, these experiments will utilize different aerosol data sources as well.

Routine filter-based data analysis will employ data from the National Park Service / USEPA IMPROVE aerosol network. Two experiments are planned. First, PMF will be applied to resolve the sources of combustion material accompanying Asian dust arriving at the Crater Lake National Park (CRLA) and Lassen Volcanic National Park (LAVO) monitoring sites. Second, PMF will be applied to IMPROVE data from the San Gorgonio National Wilderness (SAGO) site in southern California to explore the complex sources influencing that site, with a special focus on attempting to develop source profile(s) for gasoline-engine particles that are distinct from profile(s) for diesel engines.

Single-particle data analysis will use data from Aerosol Time of Flight Mass Spectrometry (ATOFMS) sampling conducted by Prof. Kimberly Prather at Angiola and Fresno to attempt to distinguish ambient air particles produced by gasoline-powered vehicles from those produced by diesel vehicles.

The objective of this project is to demonstrate the usefulness of the data analysis tool - positive matrix factorization (PMF) - to resolve the sources of ambient aerosols. This objective will be realized through two major data analysis tasks. The first task is to perform PMF on three sets of Interagency Monitoring of Protected Visual Environments (IMPROVE) data. Two IMPROVE sites are at high altitude and the objective is to separate and quantify the influence of Asian dust on observed mass concentrations. The third site is downwind of Los Angeles and the primary objective is to determine if gasoline and diesel emissions can be separately identified and quantified using the chemically speciated data.

The second task is to apply factor analysis to Aerosol Time-of-Flight Mass Spectrometry data and ascertain the ability of this method to separate diesel from gasoline motor vehicle emissions. As part of this second task, test calibration models will be developed that permit estimation of the composition of the bulk ambient aerosol composition from single particle ATOFMS data.

ORIGINS OF FINE AEROSOL MASS USING PMF

Sampling and Analyses

The Interagency Monitoring of Protected Visual Environments (IMPROVE) program (Malm et al., 1994) is a cooperative measurement effort governed by a steering committee composed of representatives from Federal and regional-state organizations. The IMPROVE monitoring program was established in 1985 to aid the creation of Federal and State implementation plans for the protection of visibility in Class I areas as stipulated in the 1977 amendments to the Clean Air Act.

The Crater Lake (CRLA)(42.89^0 N,-122.14^{0}W) field site is located in southwestern Oregon with elevation of 1981m. The Lassen Volcanic (LAVO)(40.54^0 N,-121.58^{0}W) field site is located in northern California with elevation of 1798m. The main advantages of these sites for atmospheric sampling are that it is far removed from any major local air pollution sources because of their high elevation. They are suitable for evaluating the impacts of Asia dust to the air quality of North America.

Samples were also collected at the San Gorgonio Wilderness IMPROVE site (Latitude: 34.1924N, Longitude: 116.9013W, Altitude 1705 m), which is downwind of the Los Angeles area, an area dominated by mobile source emissions beginning in March 1988. It is north of the city of San Bernadino.

The IMPROVE sampler include: Module A: $PM_{2.5}$ particles collected on Teflon. These are analyzed by five methods at for gravimetric mass for $PM_{2.5}$, hydrogen by particle elastic scattering (PESA), elements from Na to Mn by particle induced X-ray emission (PIXE), Elements from Fe to PB by photon-induced x-ray fluorescence (XRF) [Cohen 1999]; Module B: $PM_{2.5}$ particles collected on nylon. A denuder before the nylon filter removes nitric acid vapors. These are analyzed by ion chromatography for nitrate, chloride, sulfate and nitrite; Module C: $PM_{2.5}$ particles collected on quartz. These are analyzed for carbon using the Thermal Optical

Reflectance (TOR) [Chow *et al.*, 1993]. The duration of aerosol sampling was 24 h, and the samples were collected on Wednesdays and Saturdays. After 2000, the IMPROVE program changed the sampling schedule from two 24-hour samples per week (on Wednesday and Saturday) to one 24-hour sample every three days.

Data Analyses

Mass Balance Model

The fundamental principle of source/receptor relationships is that mass conservation can be assumed and a mass balance analysis can be used to identify and apportion sources of airborne particulate matter in the atmosphere. This methodology has generally been referred to within the air pollution research community as *receptor modeling* [Hopke, 1985; 1991]. The approach to obtaining a data set for receptor modeling is to determine a large number of chemical constituents such as elemental concentrations in a number of samples. Alternatively, automated electron microscopy can be used to characterize the composition and shape of particles in a series of particle samples. In either case, a mass balance equation can be written to account for all m chemical species in the n samples as contributions from p independent sources

$$x_{ij} = \sum_{k=1}^{p} f_{ik} \cdot g_{kj} \qquad \textbf{(1)}$$

where x_{ij} is the ith elemental concentration measured in the jth sample, f_{ik} is the gravimetric concentration (ng mg^{-1}) of the ith element in material from the kth source, and g_{kj} is the airborne mass concentration (mg m^{-3}) of material from the kth source contributing to the jth sample.

There exist a set of natural physical constraints on the system that must be considered in developing any model for identifying and apportioning the sources of airborne particle mass [Henry, 1991]. The fundamental, natural physical constraints that must be obeyed are:

1) The original data must be reproduced by the model; the model must explain the observations.

2) The predicted source compositions must be non-negative; a source cannot have a negative percentage of an element.

3) The predicted source contributions to the aerosol must all be non-negative; a source cannot emit negative mass.

4) The sum of the predicted elemental mass contributions for each source must be less than or equal to total measured mass for each element; the whole is greater than or equal to the sum of its parts.

When developing and applying these models, it is necessary to keep these constraints in mind in order to be certain of obtaining physically realistic solutions.

The critical question is then what information is available to solve equation (1). It is assumed that the ambient concentrations of a series of chemical species have been measured for a set of particulate matter samples so that the x_{ij} values are always known. If the sources that contribute to those samples can be identified and their compositional patterns measured, then only the contributions of the sources to each sample need to be determined (e.g. Kowalczyk et al., 1982; Chow et al., 1992). These calculations are generally made using the effective variance least squares approach incorporated into the EPA's CMB model. However, for many locations, the sources are either unknown or the compositions of the local particulate emissions have not been measured. Thus, it is desirable to estimate the number and compositions of the sources as well as their contributions to the measured PM. The multivariate data analysis methods that are used to solve this problem are generally referred to as ***factor analysis***. Factor analysis methods (e.g. Koutrakis and Spengler, 1987; Chueinta et al., 2000; Song et al., 2001a) do not utilize knowledge of sources but rely on a large number of measurements to provide a data set from which the source information can be derived.

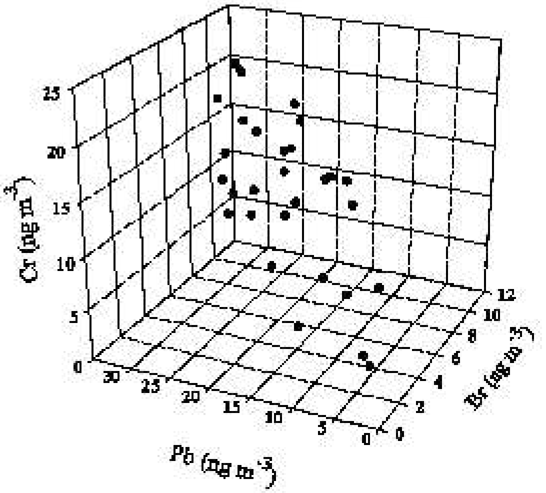


Figure 1. Three dimensional plot of simulated data.

The factor analysis problem can be visualized with the following example. Suppose a series of samples are taken in the vicinity of a highway where motor vehicles are using leaded gasoline and a steel mill making specialty steels.

For these samples, measurements of Pb, Br, and Cr are made. This set of data can then be plotted in a three dimensional space as in Figure 1. A cloud of points can be observed.

However, it is known that there are only two particle sources. The problem is then to determine the true dimensionality of the data and the relationships among the measured variables. That is goal of a factor analysis. In the case of this example, the relationships can be observed with a simple rotation of the axes so that we look down onto the figure so that the Cr axis sticks out of the page. This view is seen in Figure 2. Now it can be seen that the data really cluster around a line that represents the Pb-Br relationship in the particles emitted by the motor vehicles. The Cr values are distributed vertically and are independent of the other two elements. Factor analysis of this problem would find two sources and provide the relationship between the lead and bromine.

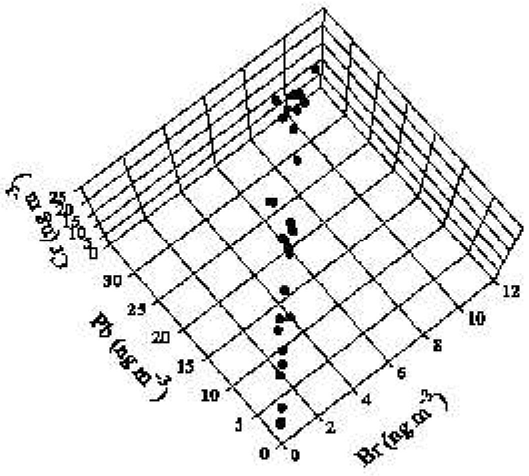


Figure 2. Plot of the simulated data as viewed from above relative to the view in Figure 1.

Positive Matrix Factorization

Multivariate approaches are based on the idea that the time dependence of a chemical species at the receptor site will be the same for species from the same source. Chemical species are measured in a large number of samples gathered at a single receptor site over time. Species of similar variability are grouped together in a minimum number of factors that explain the variability of the data set. It is assumed that each factor is associated with a source or source type. Among the multivariate receptor modeling used for aerosol source identification, positive matrix factorization (PMF) developed by Paatero and Tapper (1993, 1994) and Paatero (1997) is a relatively new technique. PMF has special features of the use of realistic error estimates to weight the data values and the imposition of non-negativity constraints in the factor computational process. PMF have been successfully applied in many atmospheric studies [Juntto and Paatero, 1994; Anttila et al., 1995; Polissar et al.,1996, 1998, 2001; Xie et al., 1999; Paterson et al., 1999;

Chueinta *et al.*, 2000; Song *et al.*, 2001a; Polissar *et al.*, 2001; Lee *et al.*, 2002; Kim *et al.*, 2003a, b, 2004; Kim and Hopke, 2004].

In this study, PMF was applied to the various IMPROVE data sets. PMF is a described in detail by Paatero [1997]. Only a brief description of the technique is given here. PMF uses a weighted least-squares fit with the known error estimates of the elements of the data matrix used to derive the weights. The factor model (PMF2) can be written as

$$X = GF + E \tag{1}$$

where *X* is the known *n* x *m* matrix of the *m* measured chemical species in *n* samples. *G* is an *n* x *p* matrix of source contributions to the samples (time variations). *F* is a *p* x *m* matrix of source compositions (source profiles). Both *G* and *F* are factor matrices to be determined. *E* is the residuals matrix, i.e., the difference between the measurement *X* and the model Y as a function of factors G and F.

$$E = X - Y = X - G \times F \tag{2}$$

The "object function," Q, that is to be minimized as a function of G and F is given by

$$\mathrm{Q(E)} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{e_{ij}}{s_{ij}}\right]^2 \tag{3}$$

where s_{ij} is an estimate of the “uncertainty” in the ith variable measured in the jth sample. The factor analysis problem is then to minimize Q(E) with respect to G and F with the constraint that each of the elements of G and F is to be non-negative.

The solution to the PMF problem depends on estimating uncertainties for each of the data values used in the PMF analysis. There are three types of values that are typically available. Most of the data points have values that have been determined, x_{ij}, and their associated uncertainties, s_{ij}. There are samples in which the particular species cannot be observed because the concentration is below the method detection limit. Finally, there are samples for which the values were not determined. These latter two types of data are often termed “missing” data. However, there are qualitative differences between them. In the below detection limit samples, the value is known to be small, but the exact value is not known. In the case where values could not be determined, the value is totally unknown. Polissar *et al.* (1998) has suggested an approach for estimating the concentration values and their associated error estimates including values below detection limits or missing for IMPROVE data from Alaska and we have use this approach in this study.

Another important aspect of weighting of data points is the handling of extreme values. Environmental data typically shows a positively skewed distribution and often with a heavy tail. Thus, there can be extreme values in the distribution as well as true "outliers." In either case, such high values would have significant influence on the solution (commonly referred to as leverage). This influence will generally distort the solution and thus, an approach to reduce their influence can be a useful tool. Thus, PMF offers a "robust" mode. The robust factorization based on the Huber influence function [Huber, 1981] is a technique of iterative reweighing of the individual data values. The least squares formulation, thus, becomes

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}(e_{ij}/h_{ij}s_{ij})^2 \qquad (4)$$

where

$$h_{ij}^2 = \begin{cases} 1 & \text{if } \left|e_{ij}/s_{ij}\right| \le \alpha, \text{ and} \\ \left|e_{ij}/s_{ij}\right|/\alpha & \text{otherwise} \end{cases} \qquad (5)$$

where α = the outlier distance and the value of α = 4.0 was chosen. It is generally advisable to use the robust mode when analyzing environmental data. Our experience has generally found that the robust mode provides the best results for typical particulate composition data.

With the total $PM_{2.5}$ mass concentration measured for each sample, multiple linear regression (MLR) were used to regress the mass concentration against the factor contributions. The regression coefficients were then used to scale the values into physically meaningful units.

Results and Discussion

The discussion of the results will be focused on the two different types of site locations. Crater Lake and Lassen are relatively remote from populated areas and the primary purpose for examining these sites was to ascertain the effect of Asian dust events at times other than the March to May period that is normally considered as the interval during which such events occur. Prior work [Vancuren and Cahill, 2002] had suggested that there were more frequent impacts of Asian dust at elevated sites in the western US.

The analysis of the other site, San Gorgonio, was focused on the separation of gasoline from diesel exhaust emissions. Prior work in Seattle [Maykut *et al.*, 2003], Atlanta [Kim *et al.*, 2004a], Brigantine NJ [Kim and Hopke, 2004a] and Washignton, DC [Kim and Hopke, 2004b] found that the analysis of IMPROVE data using the individual organic and elemental carbon thermal fractions separated factors that were assigned to diesel and spark-ignition vehicle emissions. San Gorgonio is near San Bernardino and could be expected to be influenced by motor vehicles to a much greater extent than the other two sites.

Crater Lake and Lassen Volcano

The results of this work have been published [Liu *et al.*, 2003] and will be summarized in this report. The samples were characterized by the concentrations of 33 chemical species as shown in Table 1. Table 1 shows geometric means along with geometric standard deviations for the aerosol species measured at both sites. The percentage of observations with reported concentrations below the detection limit is also shown in Table 1.

In summer 1992, analysis of IMPROVE filters for elements with atomic weights from Fe to Pb was changed from PIXE to XRF to decrease their minimum detection limits (MDL). The cyclotron time for the PIXE analysis was reduced and MDLs for elements below Fe. As mentioned in section 3.2, the detection limit was a very important parameter in application of PMF. In order to get consistent result, only the data from July 1st, 1992 to Feb, 2000 was employed for the data analysis in this study. A total of 731 and 740 samples were obtained and analyzed for the CRLA data and for the LAVO data, respectively.

A critical step in PMF analysis is the determination of the number of factors. The results of a PMF analysis are not hierarchical, i.e. a higher dimension solution does not necessarily contain all the factors of the lower dimensions, because orthogonality is not required. Thus, is normal practice to experiment with different numbers of factors and find the optimal one with the most physically meaningful results. Analysis of the goodness of model fit, Q, as defined in Equation (3), can help determine the optimal number of factors. Assuming that reasonable error estimates of individual data points are available, then fitting each value should add one to the sum and the theoretical value of Q should be approximately equal to the number of data points in the data set. However, the resulting solution also has to make physical sense within the system being studied.

The factor numbers from 5 to 9 were tested for both the CRLA and the LAVO data. The results from each trial run were examined, e.g., Q values and presented source profiles. For the CRLA data, seven factors were resolved with Q value equaling to 21454 which is close to the data dimension of 24123. For the LAVO data, six factors were resolved with Q value equaling to 22877 which is also close to the the data dimension of 23680. The resolved factors are shown in Figures 3 and 4 for the CRLA and LAVO sites, respectively. Accompanying the factors, individual error estimates were also computed for all of the factor elements. The time series plots for the source contributions are shown in Figures 5 and 6 for the data sets in the same order. The average mass contributions of each factor to the measured total mass are shown in Figures 7 and 8 for the CRLA and LAVO data, respectively.

Factor F1, on average, contributes 16% and 11% of the $PM_{2.5}$ mass at CRLA and LAVO sites, respectively. This factor represented soil factor with high concentration of Al, Si, Ca, Fe, NO_3, S, K, and Ti. It represents wind-blown Asian Dust since it occurs most frequently in spring, between March and May and showing a clearly seasonal cycle. The dust storms occur in East Asia, mainly from the Taklamakan, Gobi, and Ordos deserts and the Loess plateau. Since the phenomenon of Asian Dust usually occurs with the migratory cyclone, the frequency of the occurrence of yellow sand largely depends on the number of weather systems passing through the sources regions from the end of March to the early part of May. The Asian Dust factor contributions for these two sites is show with an expanded scale in Figure 9. It can be seen that Asian dust is less frequently observed from June to February. One of the reasons for this appears to be that cyclones traveling along the polar front contain relatively more abundant moisture than the ones that form in spring. For the high mass contribution peak in 1998 (as shown in Figure 9), in particular, Asian dust was observed all over Korea including Seoul and Anmyon Island from 14 to 22 April. Husar et al. [2001] reported that in April 1998, several usually intense dust storms occurred over the Gobi Desert in Western China and Mongolia. The storm on April 19, 1998 produced a dust cloud that crossed the Pacific and caused aerosol concentration near the health standard over much of the west coast of North America. The high mass contribution peak resolved from CRLA and LAVO on April 29, 1998 corresponded to the Asian Dust occurrence in China on April 19.

Table 1. Geometric Means (GM), Geometric Standard Deviations (GSTD), Percentage of Data below Detection Limits(BDL) for both sites

		CRLA			LAVO	
Species	BDL(%)	GM(ngm^{-3})	GSTD	BDL(%)	GM(ngm^{-3})	GSTD
AL	33.7	10.64	5.54	29.7	15.02	5.87
AS	65.8	0.05	1.85	61.2	0.05	1.95
BR	1.9	0.60	2.29	0.9	0.79	2.21
CA	5.3	13.15	3.17	2.7	14.16	2.74
CL	75	1.25	4.13	84.1	0.85	3.05
CR	66.8	0.34	2.18	71.8	0.31	2.00
CU	26.1	0.23	3.10	26.6	0.39	2.49
FE	0.8	13.79	4.08	0.4	16.07	3.44
H	1.1	85.2	2.27	0.3	103.23	2.16
K	4.5	14.73	3.01	2.3	19.67	2.71
PM$_{2.5}$ Mass	1.6	2042.3	2.23	0.7	2451.8	2.29
MG	80.5	1.98	2.54	78	1.93	2.74
MN	57.5	0.53	3.00	58.4	0.45	2.47
NA	42.3	14.43	3.52	35.8	17.34	3.72
NI	72.7	0.05	1.79	71.1	0.05	1.78
NO3_	2.9	49.38	2.83	0.3	66.22	2.99
P	92.8	0.65	1.48	95.0	0.62	1.44
PB	14.6	0.43	2.20	12.6	0.48	2.18
RB	51	0.08	1.52	45.9	0.09	1.58
S	0.7	82.09	2.49	0.1	98.60	2.42
SE	63.9	0.04	1.56	49.7	0.05	1.87
SI	3.7	43.78	3.50	1.9	55.71	3.05
SR	30.9	0.14	1.88	25.3	0.15	1.84
TI	27.6	1.92	3.10	22.8	2.21	3.02
V	63.3	0.41	2.14	59.7	0.40	2.25
ZN	3	1.12	2.92	2.0	0.77	2.38
ZR	81	0.08	1.22	82.3	0.09	1.14
OC1	17.9	48.98	1.76	2.0	53.85	2.08
OC2	2.6	104.44	2.30	0.1	123.02	2.40
OC3	3.3	162.93	2.79	0.1	215.66	2.57
OC4	2.6	94.53	2.77	0.1	121.82	2.54
EC1	2.7	144.30	2.94	0.1	172.89	2.83
EC2	3.1	63.60	2.21	0.8	46.66	1.95
EC3	34	9.84	1.54	14.5	8.84	1.62

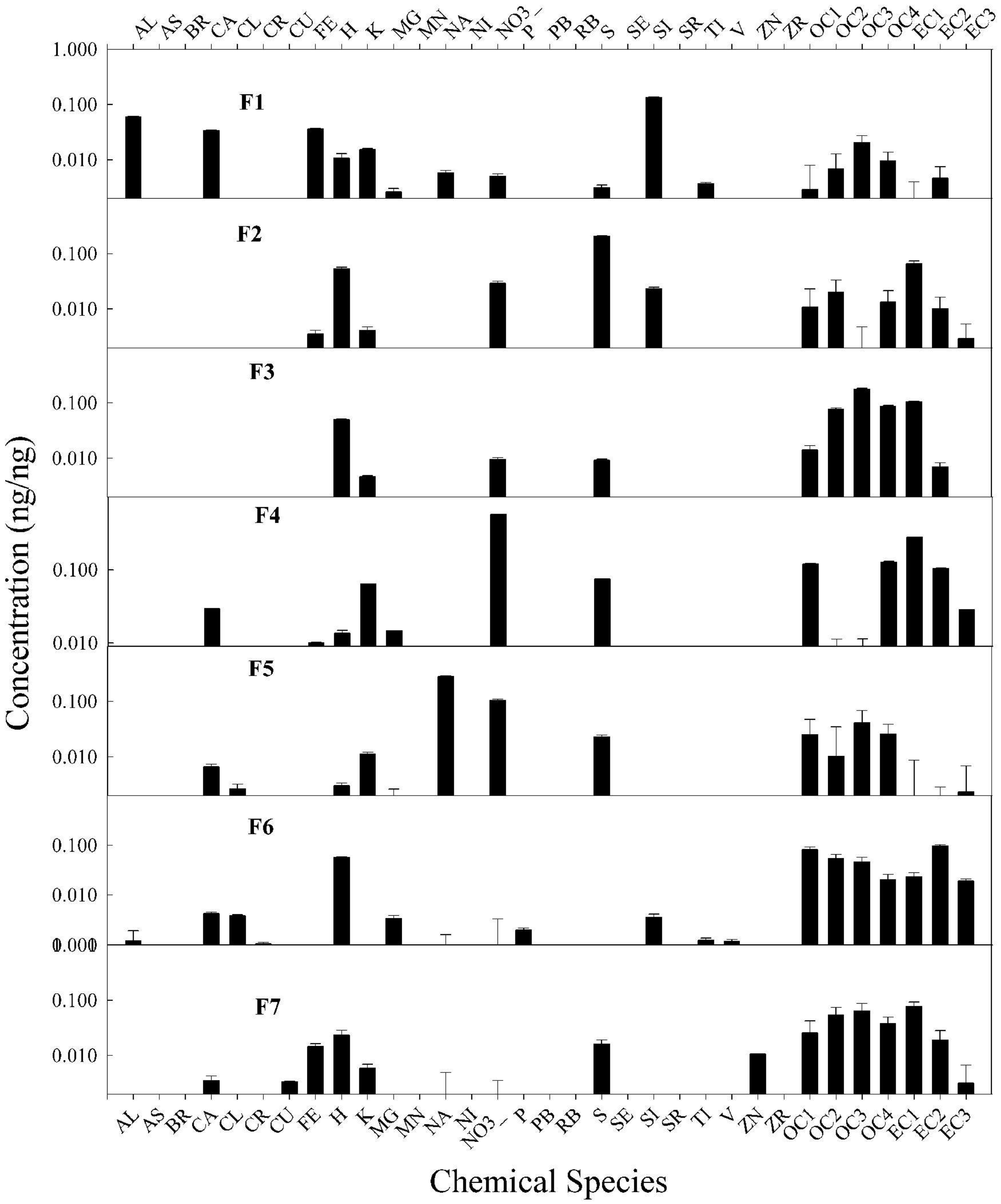


Figure 3. Source profile resolved from CRLA, OR.

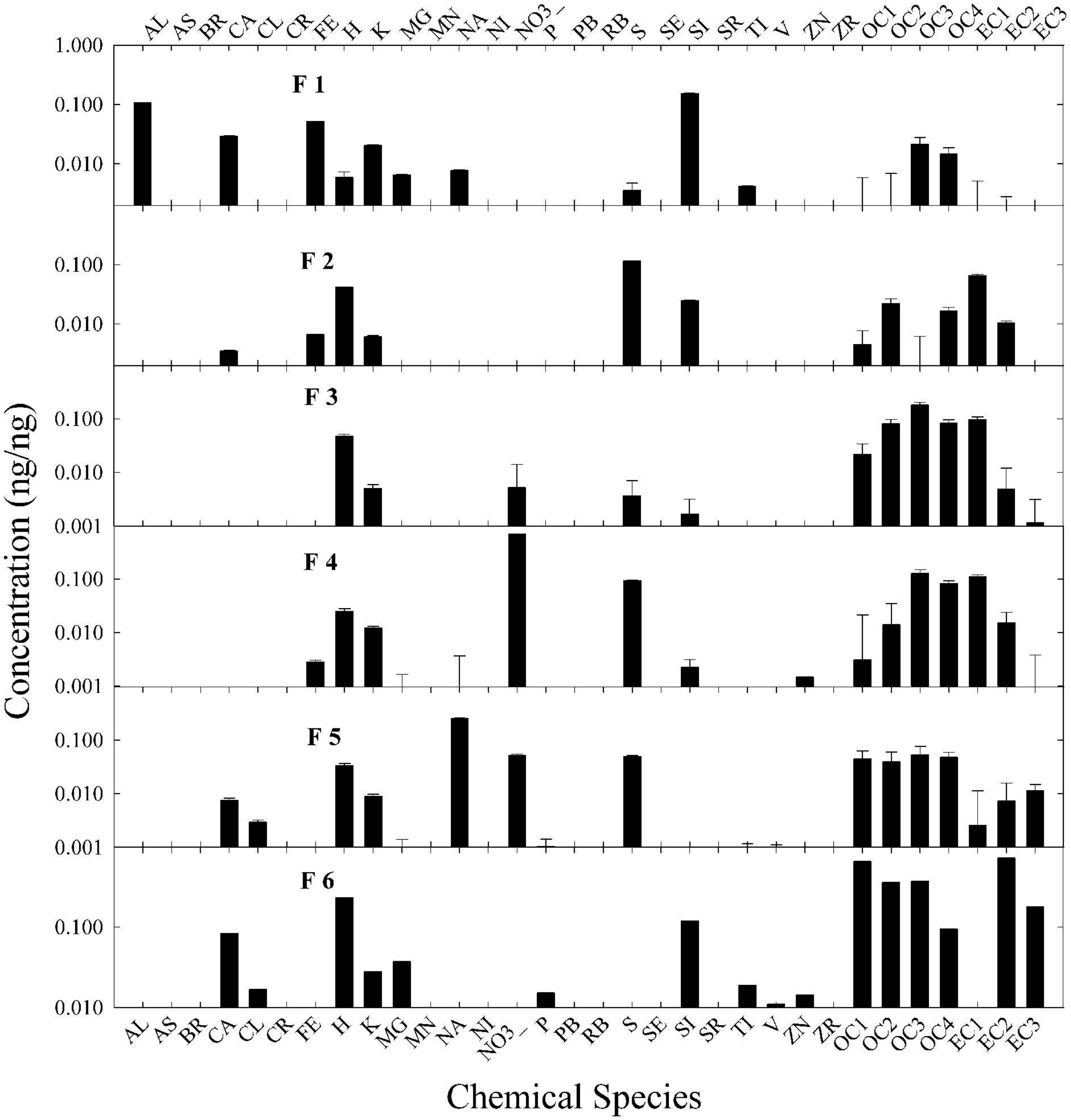


Figure 4. Source profile resolved from LAVO, CA.

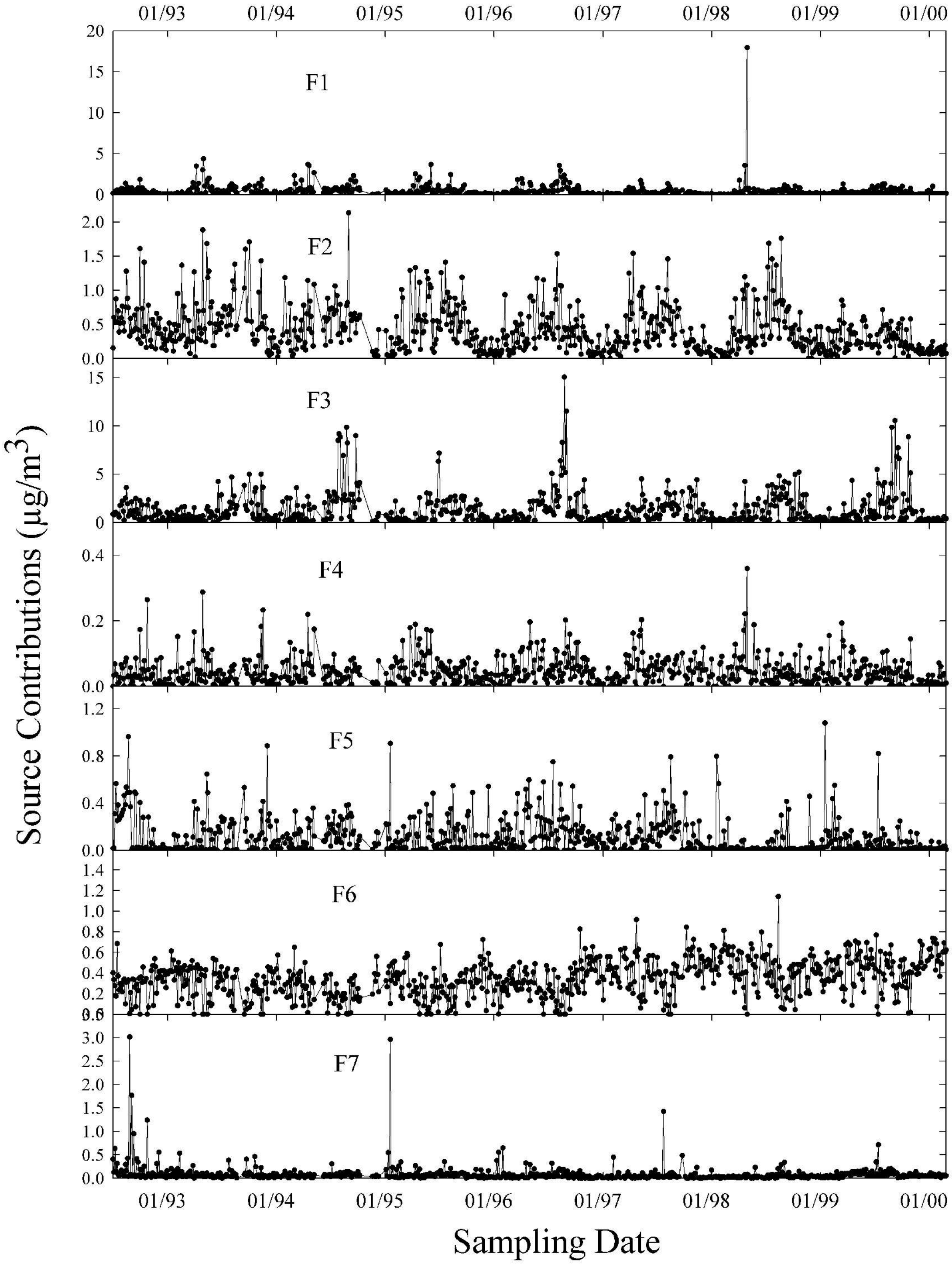


Figure 5. Time series of the source contributions at CRLA, OR.

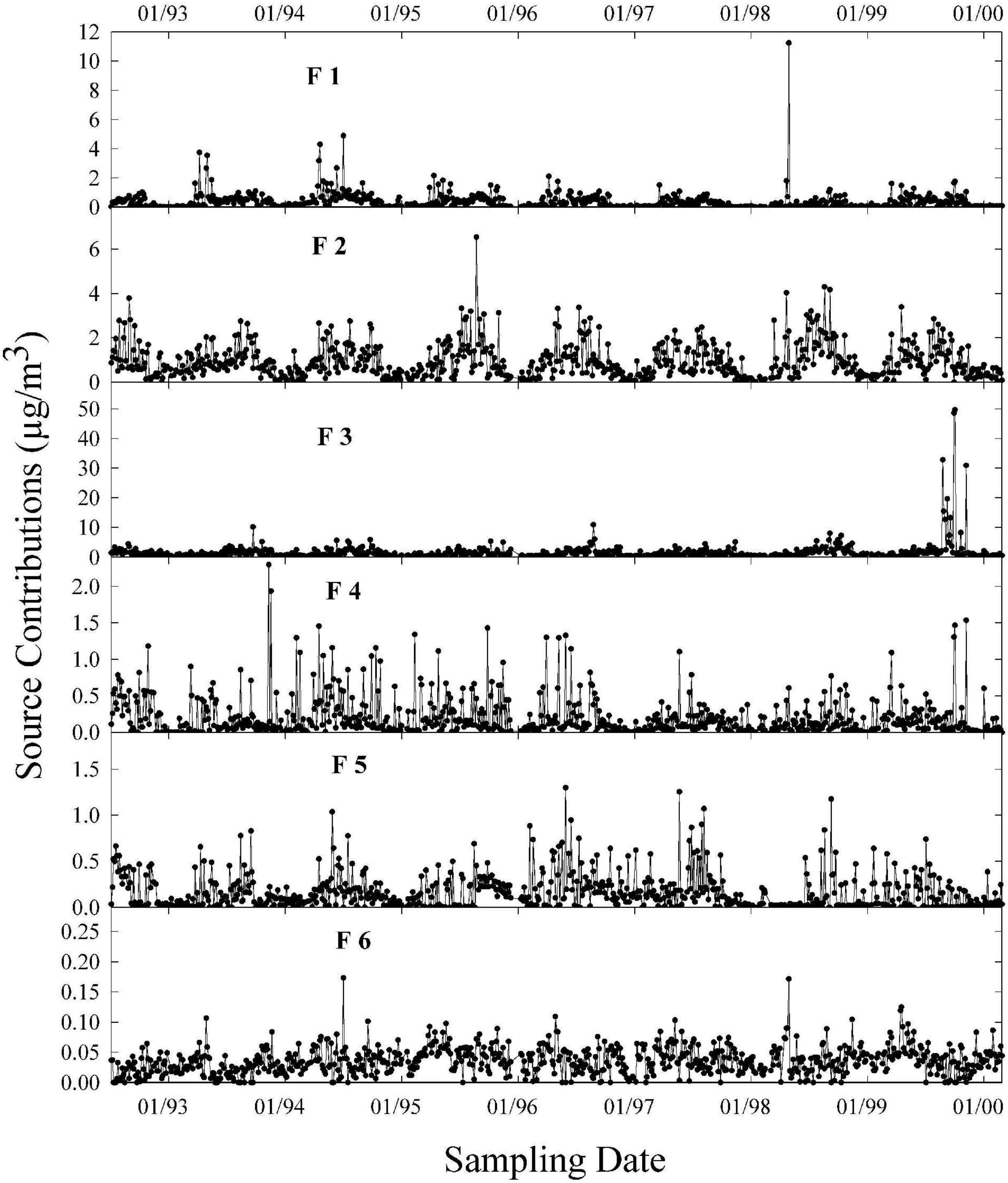


Figure 6. Time series of the source contributions at LAVO, CA.

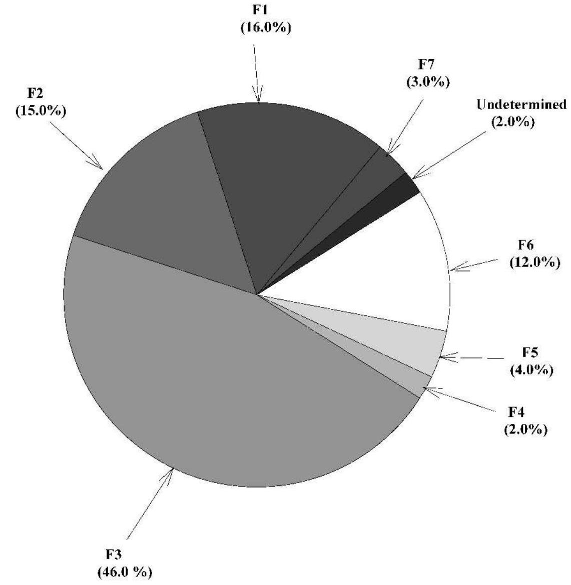


Figure 7. Average source contributions for each factor at CRLA, OR.

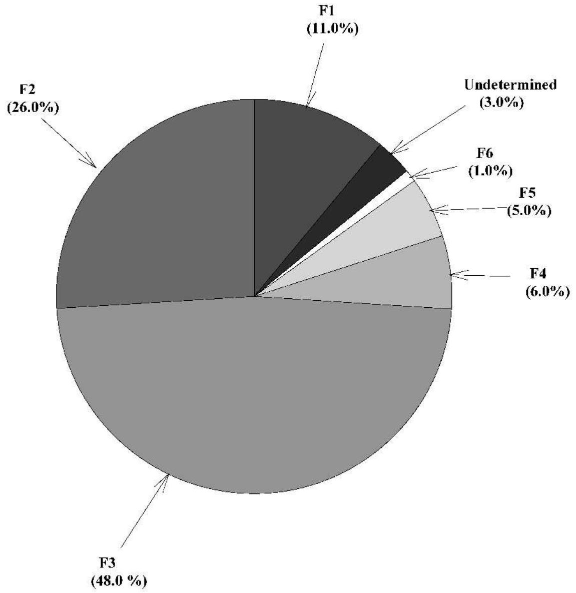


Figure 8. Average source contributions for each factor at LAVO, CA.

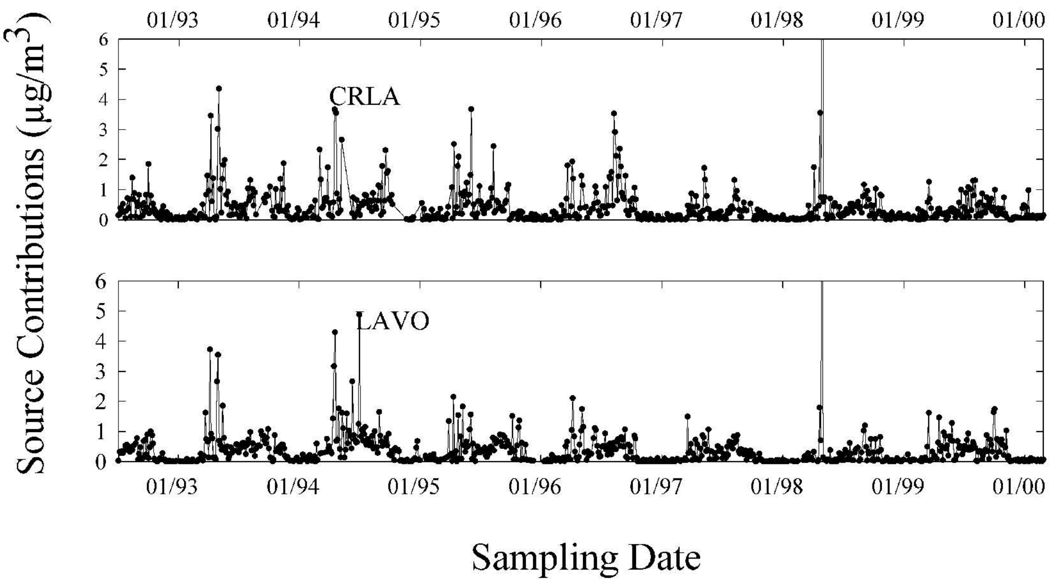


Figure 9. Comparison of the time series of Asian Dust observations at Crater Lake (top) and Lassen Volcano (bottom).

The chemical composition of F1 compared fairly well with Asian Dust Material (ADM) [Nishikawa et al., 2000] except for S, Ca and Si at both sites. The comparison results are shown in Figure 10 and Table 2 for both sites. Only 20 elements were compared between F1 and ADM because only these elements were determined in the Asian Dust Material. The biggest difference between the composition of factor F1 and the reference material is the sulfur concentration. The sulfur concentration in factor F1 is more than six to seven times of that in the reference materials at CRLA and LAVO, respectively. This high value can be explained by accumulation of sulfur during the transportation of Asian Dust through industrial areas such as the eastern part of China. The Ca and Si concentration in factor F1 is much less than that of in the reference material. The reduction in Ca and Si in the downwind samples is consistent with the modeling and aerosol analysis reported by Carmichael and coworkers for the PEM-West Asian plume field program [Chen et al, 1997, Xiao et al., 1997]. Additionally, the particle size distribution of the ADM includes particles up to 40 µm while the sampling process collects particles under 2.5 µm. There could be significant changes in composition for the finer sized particles.

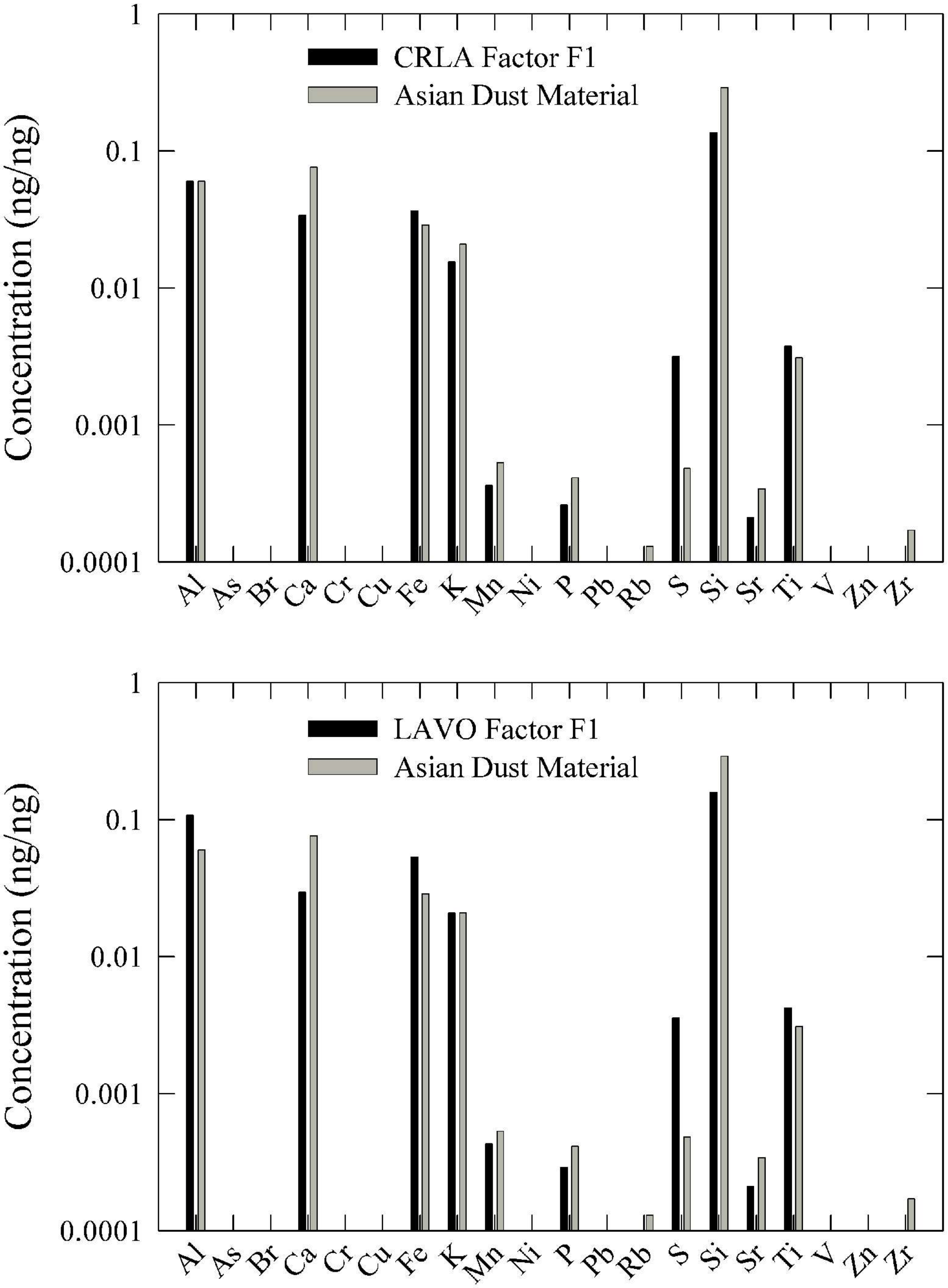


Figure 10. Factor F1 species concentrations resolved in this work compared with Asian Dust Material (ADM) [Masataka et al., 2000]).

Table 2. Factor F1composition concentration resolved from this work compared with the Asian Dust Material (ADM) [Nishikawa et al., 2000]

Element	CRLA F1 (ng/ng)	LAVOF1 (ng/ng)	ADM (ng/ng)	CRLA F1 /ADM	LAVO F1 /ADM
Al	0.060 ± 0.00150	0.107 ± 0.00100	0.060	1.00	1.78
As	0.00001 ± 0.00000	0.00001 ± 0.00000	0.00002	0.72	0.68
Br	0.00001 ± 0.00000	0.00001 ± 0.00000	0.00001	0.83	1.35
Ca	0.034 ± 0.00060	0.029 ± 0.00060	0.076	0.45	0.39
Cr	0.00003 ± 0.00000	0.00004 ± 0.00000	0.00005	0.65	0.78
Cu	0.00002 ± 0.00000	N/A ±	0.00003	0.67	N/A
Fe	0.036 ± 0.00090	0.053 ± 0.00050	0.029	1.27	1.80
K	0.015 ± 0.00050	0.021 ± 0.00060	0.021	0.74	0.98
Mn	0.00036 ± 0.00003	0.00043 ± 0.00003	0.00053	0.68	0.81
Ni	0.00002 ± 0.00000	0.00003 ± 0.00000	0.00004	0.55	0.66
P	0.00026 ± 0.00003	0.00029 ± 0.00002	0.00041	0.64	0.71
Pb	0.00001 ± 0.00000	0.00002 ± 0.00000	0.00002	0.67	0.83
Rb	0.00006 ± 0.00001	0.00007 ± 0.00001	0.00013	0.50	0.55
S	0.0032 ± 0.00030	0.0036 ± 0.00030	0.0005	6.59	7.41
Si	0.136 ± 0.00200	0.157 ± 0.00210	0.290	0.47	0.53
Sr	0.00021 ± 0.00001	0.00021 ± 0.00001	0.00034	0.62	0.60
Ti	0.0037 ± 0.00020	0.0042 ± 0.00020	0.0031	1.20	1.35
V	0.00004 ± 0.00000	0.00004 ± 0.00000	0.00006	0.66	0.72
Zn	0.00004 ± 0.00000	0.00006 ± 0.00000	0.00007	0.64	0.79
Zr	0.00006 ± 0.00001	0.00006 ± 0.00001	0.00017	0.35	0.35
Average				1.00	1.10

Factor F2 represents the secondary sulfate profile with a high concentration of sulfur. This factor, on average, contributes 15% and 26% of the $PM_{2.5}$ mass in CRLA and LAVO, respectively. This factor shows a strong seasonal variation trend corresponding to the Asian Dust with high concentration in summer time. This factor may represent the sulfur emission from the industrial area as well as marine sources. Dimethylsulphide (DMS, CH_3SCH_3) emitted from marine phytoplankton, globally contributes to 30% of the total emission of sulphur to the atmosphere [Andreae, 1990; Spiro et al., 1992; Bates et al., 1992; Berresheim et al.,1995]. In remote oceanic areas near the CRLA and LAVO, DMS is considered to be the main source of climatically active sulphate aerosols [Charlson et al., 1987]. Sulfate can also be formed from SO_2 emitted from the industrial areas across eastern Asia. The big difference of mass contribution between these two sites can be explained by the location of these two sites. The LAVO is located in northern California while the CRLA is located in southwestern Oregon. The industrial areas surrounding San Francisco and Sacramento could contribute the additional sulfur to LAVO when compared to CRLA.

Factor F3 is the highest contributor of the $PM_{2.5}$ mass, on average, contributes about 46% and 48% to the $PM_{2.5}$ mass at CRLA and LAVO, respectively. It is characterized by OC, EC and K. It is connected with local residential wood burning and occasional forest fire impacts. Figure 5 shows that there were significant fires in the summers of 1994, 1996, and 1999 affecting CRLA. At LAVO, there was a major event that appears in 1999 (Figure 6). A limited number of very high values raised the average contributions to the wood combustion factor.

Factor F4 is relative minor, on average, contributes 2% and 6% to the $PM_{2.5}$ mass of CRLA and LAVO, respectively. It is the dominant source of secondary NO_3. Nitrate is formed in the atmosphere predominantly through oxidation of NOx. The suspected origin of this source is the mobile emissions. The particle partitioning of total nitrate ($HNO_{3(g)}$+NO_3^-) depends on ambient temperature and relative humidity [Seinfeld and Pandis, 1998].

Factor F5, on average, contributes 4% and 5% to the total $PM_{2.5}$ mass for CRLA and LAVO, respectively. It contains high concentrations of Na, S and NO_3. The Cl concentration is very low which may be due to the conversion from NaCl to $NaNO_3$ and Na_2SO_4. The temporal variation does not exhibit strong seasonal pattern.

Factor F6, on average, contributes 12% and 1% to the total PM2.5 mass for CRLA and LAVO, respectively. It was characterized by high OC and EC concentrations, accompanied by some soil components (Al, Fe, Si, Ca, Mg) entrained by passing traffic. The temporal variations shown seasonable variations with high impacts during summer time when the tourist most likely visited these two sites.

Factor F7 resolved from CRLA is relatively minor contributor of the $PM_{2.5}$ mass which on average, contributes 3% of the $PM_{2.5}$ mass. This profile suggests a source rich in Zn, Cu, Fe with high concentration of EC and OC. This factor could be caused by industrial emissions, such as iron/copper smelters. Only sporadic high mass contribution events occurred from this source during these years.

Through the regression of the measured total $PM_{2.5}$ mass concentration against the factor scores, on average, 98% of the $PM_{2.5}$ mass can be explained by the factors in both sites. The results also showed that PMF was a powerful and useful factor analysis method to extract emission sources out of ambient concentration data.

San Gorgonio

The results of the analysis of the San Gorgonio National Wilderness have been presented in a manuscript that is currently under review at *Atmospheric Environment.* We are currently waiting to receive the reviews. Seven sources were resolved for the ambient aerosols at the San Gorgonio IMPROVE site. They were 1: soil, 2: sea salt, 3: diesel emission, 4: gasoline emission, 5: secondary sulfate with secondary organic, 6: secondary nitrate, and 7: Asian dust. Multiple linear regression (MLR) was applied to regress the total $PM_{2.5}$ mass against the estimated source contributions [Hopke et al., 1980]. The regression coefficients were used to scale the source profiles and contributions to make them more physically meaningful. The corresponding source profiles and contributions are shown in Figures 11 and 12, respectively. An interesting feature of this solution is the existence of two sources with different OC/EC fractions that have been tentatively identified as diesel and spark ignition vehicle. The separation of diesel and gasoline emission will be discussed first and then the description of other sources will be given.

Diesel oil and gasoline are composed of many kinds of compounds, and their emissions are also a complex mixture of compounds. To provide a characterization of gasoline and diesel vehicle emissions, Chow et al. [1993] and Watson et al. [1994] applied thermal/optical reflectance (TOR) method to the carbon analysis to the diesel/gasoline emission obtained in dynamometer tests. The same seven carbon fractions (four OCs and three ECs) were detected as have been measured for the quartz filter in module of the IMPROVE sampling system. Through the results of their experiments [Chow et al. 1993 and Watson et al. 1994], the similarities and differences in the profiles of diesel and gasoline emissions with respect to the organic and elemental carbon fractions are presented below.

The diesel fueled OC1 abundance is significantly higher than the gasoline fueled OC1 abundance. The EC1 fraction in gasoline emission is generally more abundant than that in diesel emission. The EC2 fraction in the diesel emission is significantly higher than that in the gasoline emission. There is very little EC3 in either diesel emission or gasoline emission.

It can be seen from Figure 11 that a source attributed to diesel emissions was separated from the gasoline emission source in this study. Source 3 with high concentrations of OC1 and EC2 was assigned to be diesel emission while source 4 with high OC3 and EC1 represents gasoline emission. Some Fe, Si, and other elements were also included in the source profiles possibly because gasoline/diesel emission might be mixed with soil dust constituents during transportation or these may be elements in fuel additives [Schauer et al., 2002]. To highlight these source profiles in comparison with direct measurements, the profiles of diesel and gasoline with relative concentrations of OCs and ECs are shown in Figure 13 along with the respective profiles measured in dynamometer studies by Watson et al. [1994]. The results from San Gorgoio are in very good agreement with those of Watson et al. [1994], especially with respect to the key fractions (i.e., OC1, EC2 for diesel, and EC1 for gasoline).

In addition to diesel and gasoline emissions, other five sources were identified in this study. They are soil, aged sea salt, secondary sulfate with secondary organics, secondary nitrate, and desert dust, respectively. Source 1 represents soil with high concentrations of of Al, Si, Ca, and Fe, and contributes 8.2% to the total PM 2.5 mass in this site. The ratio of Al to Si in this source is 0.89 much higher than the typical ratio observed in soils, 0.293 [Mason, 1966]. This source appears to include some Al that should have been assigned to the other sources.

Source 2 appears to be aged sea salt with high concentrations of Na, NO_3, and S. This source contributes 7.3% to the total PM 2.5 mass. The low concentration of chloride in this source may be due to the conversion from NaCl to Na_2SO_4 or $NaNO_3$ during transportation [Liu et. al., 2003].

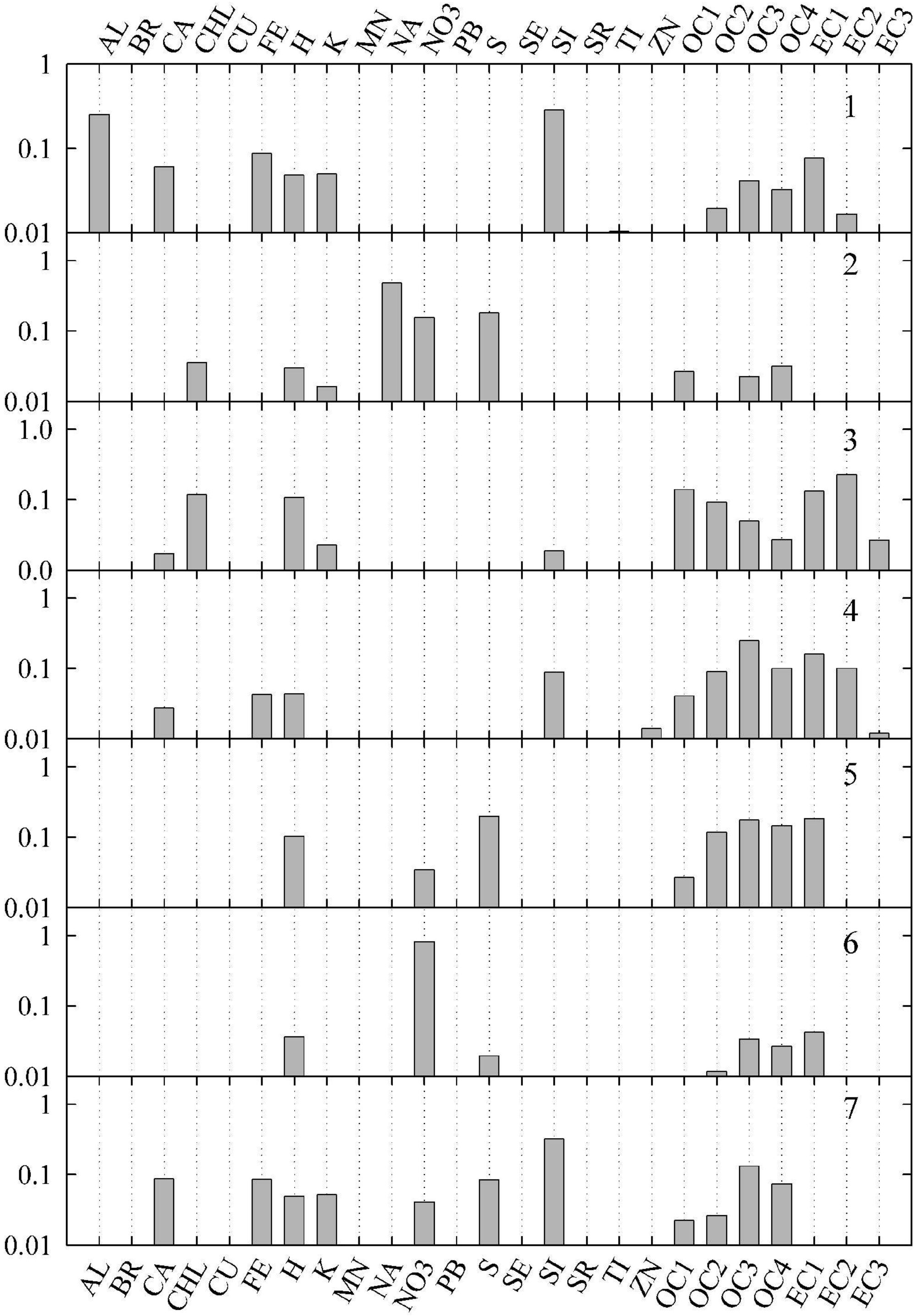


Figure 11. Resolved source profiles based on the data from San Gorgonio, CA

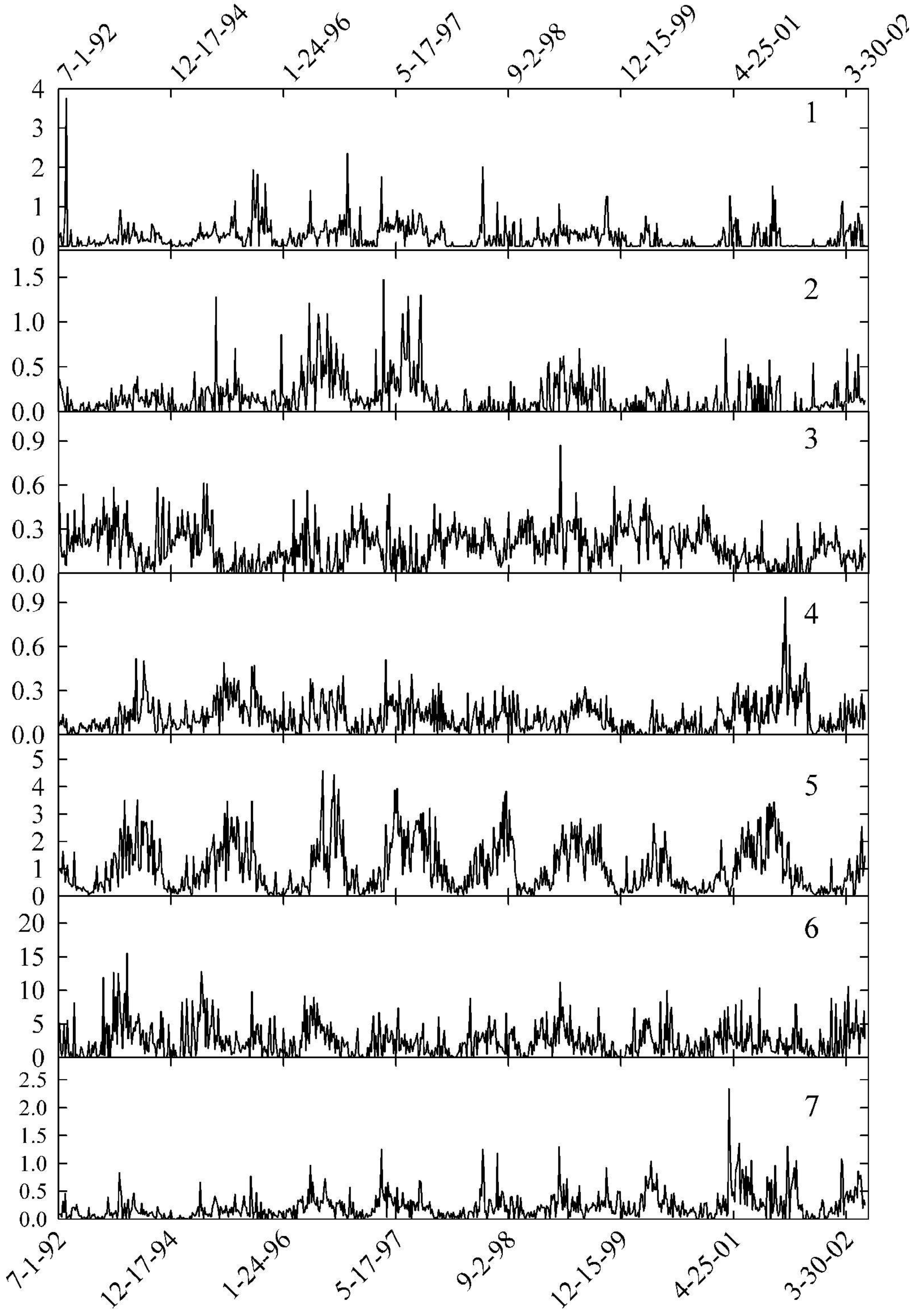


Figure 12. Time series of the source contributions of the resolved factors for San Gorgonio, CA.

Source 6 shows secondary nitrate with high concentration of nitrate and contributes 39.3% to the total PM 2.5 mass. Nitrate is formed in the atmosphere predominantly through oxidation of NOx and the suspected origin of this source is the mobile emissions. This source shows a seasonal trend with high contributions in winter because low temperature shifts the equilibrium system of $NO^{-}{}_{3}$ and HNO_3 toward the particle phase, increasing the mass of NH_4NO_3 [Seinfeld and Pandis, 1998]. However, the temporal contribution plot for this source also shows some peaks in the summer. This result reflects the temporal variation of measured NO_3 concentration. Actually, it has been observed that some high peaks of nitrate concentration occurred in summer time [Liu, et al., 2000].

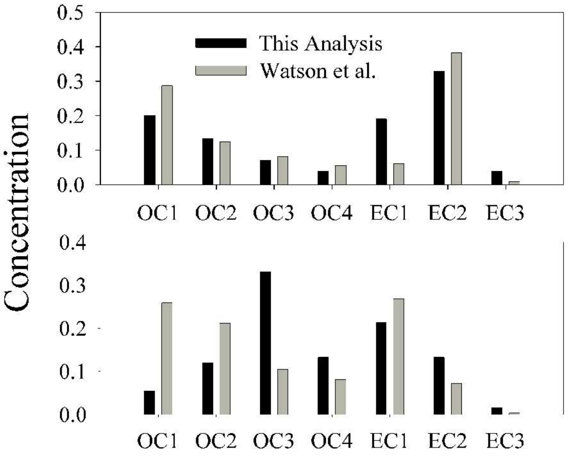
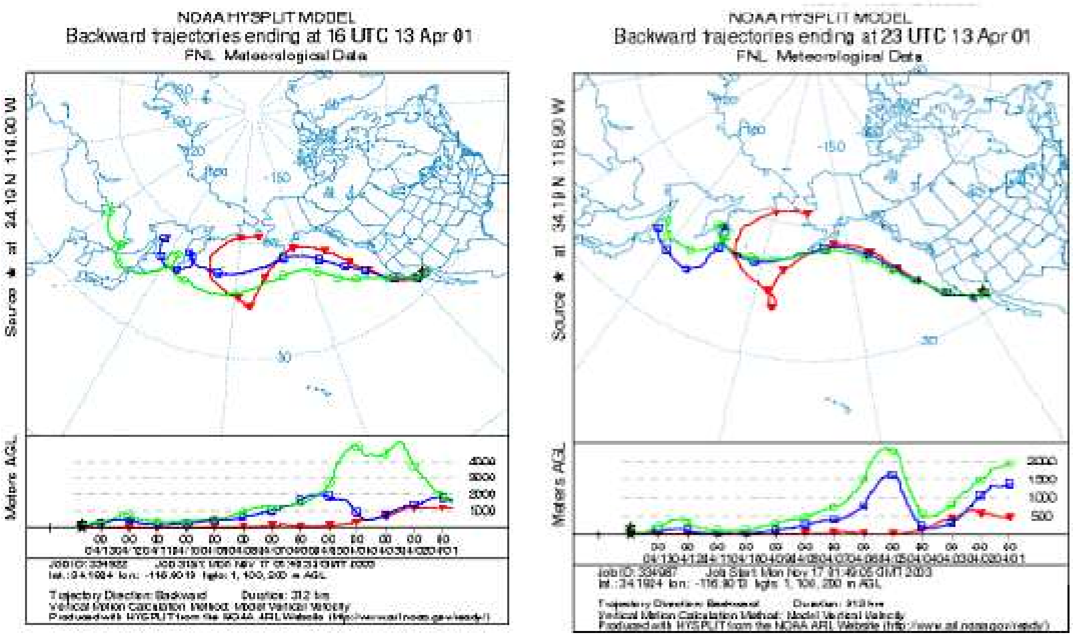


Figure 13. Comparison of diesel (top) and gasoline (bottom) carbon thermal fractions extracted from the San Gorgonio site data and those reported by Watson et al. (1994).

Source 7 shows high concentration of Si and contributes 5.9% to the total $PM_{2.5}$ mass. The Al concentration is very low in this source and represents the rotational problem noted for source 1. This source is assigned as desert dust. It can be seen from the temporal contribution plot of this source that almost every spring (from March to May) shows a high peak, especially in the spring of 2001. Dust storms occur almost every spring in the deserts of Western China, such as Taklamakan, Gobi and Ordos Deserts. Husar et al. [2001] reported that in April 1998, several intense dust storms occurred over the Gobi Desert in Western China and Mongolia. In particular, the storm on April 19, 1998 produced a dust cloud that crossed the Pacific and reached much of the west coast of North America. Thus, the peak of April 29, 1998 in the contribution plot of this source may correspond to this sand storm. Recently, the studies of the aerosol of the Crater Lake and Lassen Volcanic National Parks have reported the similar result and the corresponding date was also April 29, 1998 [Liu et al., 2003]. They suggested that such events may be observed more frequently at high altitude sites such as San Gorgonio.



NOAA HYSPLIT MODEL
Backward trajectories ending at 16 UTC 13 Apr 01
FNL Meteorological Data
Source ★ at 34.10 N 116.90 W
-150
30
Meters AGL
4000
3000
2000
1000
Trajectory Direction: Backward
Vertical Motion Calculation Method: Model Vertical Velocity
Produced with HYSPLIT from the NOAA ARL Website
NOAA HYSPLIT MODEL
Backward trajectories ending at 23 UTC 13 Apr 01
FNL Meteorological Data
Source ★ at 34.10 N 116.90 W
-150
30
Meters AGL
2000
1500
1000
500
Job ID: 334987 Job Start: Mon Nov 17 21:49:05 GMT 2003
lat.: 34.1924 lon.: -116.9013 hgts: 1, 100, 200 m AGL
Trajectory Direction: Backward
Vertical Motion Calculation Method: Model Vertical Velocity
Produced with HYSPLIT from the NOAA ARL Website (http://www.arl.noaa.gov/ready/)

Figure 14. Calculated backward air parcel trajectory plots for three different heights and two different times on April 13, 2001.

In addition, the peak concentration on April 13, 2001 in the contribution plot may also correspond to a major sand storm occurred in China. In order to explore this hypothesis, air parcel back trajectories with three different heights (1, 100, and 200 meters above model ground level) and two different end times (16:00 pm and 23:00 pm UTC) on April 13, 2001 were calculated using the NOAA HYSPLIT model [http://www.arl.noaa.gov/ready/hysplit4.html, Draxler and Hess, 1998]. All the trajectories were across the northern Pacific Ocean, which suggests eastern Asia as the source area (Figure 14). This source profile also shows some S and OC probably because the Asian dust was mixed with anthropogenic air pollutants during the transport across China. Sources 1 and 7 both show high concentration of Si, but it can be seen from Figure 12 that their temporal contribution plots are obviously different, suggesting the concept of two different sources called "local soil" and "Asian dust". As for the low concentration of Al in source 7, the possible reasons are 1) Al might be incorporated into other sources that represent stronger sources of Al, and 2) more than 1/3 of the Al concentration measurements were below the detection limit, which is adverse to the analysis.

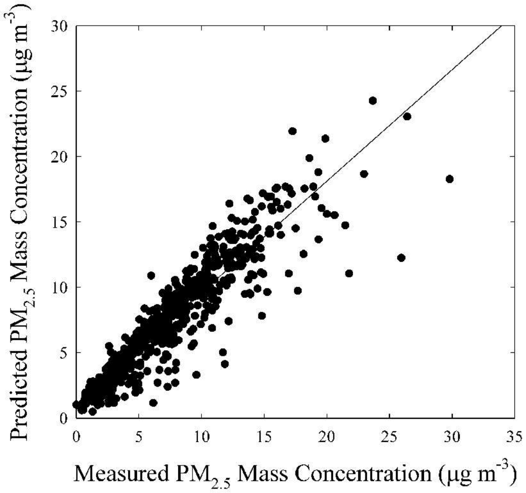


Figure 15. Comparison of the predicted total $PM_{2.5}$ mass from the PMF analysis with the measured $PM_{2.5}$ mass concentrations

Another test of the effectiveness of the PMF analysis is the comparison of the predicted PM mass vs. the measured PM mass. The predicted PM mass of each sample was obtained from the sum of scaled source contribution values. It can be seen from Figure 15 that there is a high correlation between the predicted PM mass and the observed values with a squared correlation coefficient R^2 of 0.86.

Figure 16 shows that the average contributions of the sources to total $PM_{2.5}$ mass. The largest sources are secondary sulfate and nitrate. These secondary factors also include carbonaceous species so that the mass contributions are greater than just the mass of ammonium sulfate or ammonium nitrate. The contributions of diesel and gasoline emissions are 8.1% and 2.2% respectively, so they are not large contributors of the aerosol particles in the San Gorgonio site. However, the proportions are similar to those observed in earlier studies of the Los Angeles basin [Schauer *et al.*, 1996]. Again at this elevated site, there is an apparent impact of Asian dust events as well as aerosolization of local soils.

A test of the effectiveness of the PMF analysis is the comparison of the predicted PM mass vs. the measured PM mass. The predicted PM mass of each sample was obtained from the sum of scaled source contribution values. It can be seen from Figure 14 that there is a high correlation between the predicted PM mass and the observed

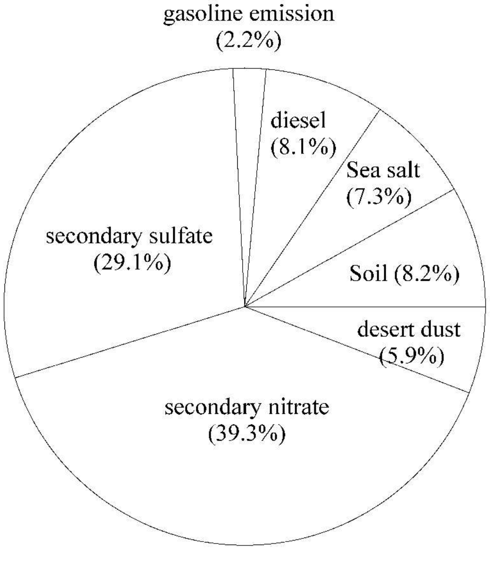


Figure 16. Relative contributions of the identified sources to the $PM_{2.5}$ mass

values with a squared correlation coefficient R^2 of 0.86. The slope is 1.07 ± 0.11 and an intercept of 0.85 ± 0.01 μg m^{-3}.

PREDICTING BULK AMBIENT AEROSOL COMPOSITIONS FROM ATOFMS DATA

Introduction

Motor vehicle exhaust, road dust, industrial emissions, biogenic emissions and other pollution sources make the exposure to ambient aerosols unavoidable, so more and more, new measurement techniques and data analysis tools have been applied into the studies on ambient particles (Hughes *et al.*, 1999; Song *et al.*, 2001b). First developed in 1994, aerosol time of flight mass spectrometry (ATOFMS) represents a particle analysis technique with the ability to provide the size and composition of individual aerosol particles in real time (Prather, et al., 1994). This technique provides information to understand the size and composition distribution of particles (Hughes, et al., 1999). However, one of the criticisms of ATOFMS is that it cannot provide quantitative estimation of the bulk chemical composition of particles. Actually bulk chemical compositions are very helpful for studying the relationship between ambient aerosols and human diseases and designing pollution control strategies, so solving this estimation problem will make more uses of the advantages of ATOFMS and extend the application fields of ATOFMS. Fergenson et al. (2001) performed the initial study on calibration models to estimate the ambient aerosol chemical composition from single particle data. However, because of the limited samples for that research (only 12 samples), the ability of the multivariate calibration model to predict bulk chemical compositions was not demonstrated sufficiently and there were salient errors between the predicted and measured concentrations, so the goals of this work are: 1) fully prove the feasibility and predicting effect of the multivariate calibration model based on adaptive resonance theory (ART) neural networks and partial least square regression (PLSR) on estimating bulk aerosol chemical composition from ATOFMS data, 2) make detailed study on some key steps to building a successful calibration model, and 3) make preparation for testing the transferability of calibration models and thereby determine if the calibration models of one location can be applied to data from another location.

Method Description

The data analysis consists of two major parts. First the individual particles need to be classified based on their individual mass spectra. The mass of particles in each identified class then become part of the input into the multivariate calibration model to permit the prediction of bulk aerosol compositions from new single particle data. These methods have been described in detail elsewhere (Song *et al.*, 1999; Fergenson *et al.*, 2001) and only a brief introduction to these methods are presented in the subsequent sections.

Description of ART-2a

With the developments of single particle mass spectrometer, various cluster analysis methods, typically ART-2a (a kind of ART networks), have been applied into the on-line particle composition analysis. There are a number of reports about the application of ART-2a into the classification of single particle mass spectrometry data (Song et al., 1999; Phares et al., 2001). In the classification of ATOFMS data, the input to the ART-2a algorithm are the vectors of the areas for each mass-to-charge ratios of all of the peaks in each particle and the outputs are the index of the class each particle belongs to. Compared with most clustering methods, the significant advantage of ART is the ability to add a new cluster without disturbing any existing clusters, so it is very useful for the online particle analysis.

Training for the ART-2a algorithm is briefly described as below. The details can be found in the literature (Carpenter et al., 1991; Xie, et al., 1994).

1. Randomly select an input vector and scale it into unit length.

$$\mathbf{p}_i = \mathbf{x}_i / \|\mathbf{x}_i\|$$

2. Contrast enhancement: transfer all elements of p_i through a nonlinear transfer function.

$$q_{ij} = \begin{cases} p_{ij}, & \text{if } p_{ij} > \theta \\ 0, & \text{otherwise} \end{cases}$$

where θ is a threshold value. In this study, θ is 0.005.

3. Rescale $\mathbf{q}_i$ to unit vector $\mathbf{r}_i$.

4. Evaluate the competitions among all existing *l* output neurons, and select the winner neuron. The competition denotes the resonance between the input vector and the existing clusters and is measured by their dot product.

$$\rho_k = \mathbf{r}_i \mathbf{w}_k \quad (k = 1 \ldots l) \quad \text{and} \quad \rho_{win} = \max(\rho_k)$$

5. If the resonance of the winner neuron is larger than the vigilance limit, ρ_{vig}, that is determined before training (in this study, ρ_{vig} is set as 0.6), modify the cluster vector of the winner neuron toward the input vector according to the following procedure;

$$v_{ij} = \begin{cases} r_{ij}, & \text{if } w^{old}_{(win)\,ij} > \theta \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{u}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$$

$$\mathbf{t}_i = \mathbf{w}^{old}_{win} + \eta(\mathbf{u}_i - \mathbf{w}^{old}_{win})$$

$$\mathbf{s}_i = \mathbf{t}_i / \|\mathbf{t}_i\|$$

$$\mathbf{w}^{new}_{win} = \mathbf{s}_i$$

where η is learning rate. In this study, η is 0.1. Otherwise, create a new cluster as below.

$$\mathbf{w}_{new} = \mathbf{r}_i$$

Repeat the above steps until the change between the cluster vectors of two consecutive cycles is sufficiently small or the number of repetitions reaches the pre-defined number. Finally, the assignments to clusters obtained from the ART-2a analysis are then used to estimate the mass fractions of all of the classes during each sampling period. The detailed procedure will be shown in next section.

Description of PLSR

PLSR is a recently developed generalization of multiple linear regression (MLR). The significant advantage of PLSR over traditional MLR is that PLSR can analyze strongly collinear and noisy data, and also simultaneously model a number of response/dependent variables (Wold, et al., 2001). A linear regression model can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} \tag{6}$$

where **B** are the regression coefficients. In PLSR, **X** can be transformed into

$$\mathbf{X} = \mathbf{T}\,\mathbf{C}^{T} \tag{7}$$

where **T** is the matrix of PLS components and **C** is the loading matrix. Let $\mathbf{W} = (\mathbf{C}^{T})^{-1}$. Then the PLS transformation is given by

$$\mathbf{T} = \mathbf{X}\mathbf{W} \tag{8}$$

Thus, equation (6) can be re-written as $\mathbf{Y} = \mathbf{T}\,\mathbf{C}^{T}\mathbf{B} + \mathbf{E} = \mathbf{X}\mathbf{W}\,\mathbf{C}^{T}\mathbf{B} + \mathbf{E}$. Let $\Lambda = \mathbf{C}^{T}\mathbf{B}$. Then equation (6) can be expressed as

$$\mathbf{Y} = \mathbf{T}\Lambda + \mathbf{E} = \mathbf{X}\mathbf{W}\Lambda + \mathbf{E} \tag{9}$$

Let $\hat{\Lambda}$ be the estimation of Λ., Thus, the predictor for new input, $\mathbf{X}_{\mathbf{new}}$, can be represented by $\mathbf{Y}_{pred} = \mathbf{X}_{new}\mathbf{W}\hat{\Lambda}$. The detailed procedure can be found in Hoskuldsson (1988). In this study, **X** and **Y** denote mass fraction of each class and species concentrations respectively.

Data Treatment and Analysis

In this study, the ATOFMS data were collected in Fresno, CA. The sampling period was from Dec. 2000 to Feb. 3, 2001. Both positive and negative ions were measured in the ATOFMS, so the range of mass-to-charge (m/z) for this study was set to [-350, +350]. This means the input of ART-2a is a 700 dimensional vector. The bulk aerosol species concentrations were measured as part of the California Regional Particulate Air Quality Study (CRPAQS). The California Regional PM10/PM2.5 Air Quality Study is a comprehensive public/private sector collaborative program with two main goals: 1) to provide an improved understanding of particulate matter and visibility in central California, and 2) to provide decision-makers with the tools needed to identify equitable and efficient control methods (CRPAQS, 2003). The species concentration data of PM2.5 for the Fresno site were collected from Dec. 15, 200 to Feb. 3,

2001. Only 11 days (Dec. 15, 16, 17, 18, 26, 27 and 28, Jan. 31, and Feb. 1, 2, and 3) contained both ATOFMS data and species concentration data, and the concentration data of each day were collected in 5 time intervals 0:00 to 5:00, 5:00 to 10:00, 10:00 to 13:00, 13:00 to 16:00 and 16:00 to 24:00. So the total number of the time periods for the calibration model was 11*5=55. The total number of measured particles in these 55 time periods was 249, 674.

The problem is how to classify such a large number of particles with ART-2a networks. In Fergenson *et al.* [2001], a total of 12, 479 particles were grouped into 12 samples. Each sample was classified individually by ART-2a, but the weight matrix (cluster vectors) was preserved from one analysis to the next. Any particles that did not fit into the existing classes nucleated their own new classes. The feature of ART-2a network that ART networks can create new classes without disturbing the existing classes permits this procedure to function, but in term of system completeness, it would be better to classify the particles in one analysis. Thus, in this study, the 249, 674 particles were classified at the same time. A total of 1325 classes were created. However, most of 1325 classes only contained a very few particles. Therefore, only 117 classes were retained for further analysis after the following screening. The particles of the selected 117 classes accounted for 95% of total particle mass. Each of the remaining classes accounted for less than 1% of total particle mass.

The classification results were used to estimate the mass fraction of each class in each time period. There are two corrections that need to be made during the mass fraction calculation. The first is time scaling. Because the length of the sampling intervals varied, the sampling periods were scaled to give an estimate of the actual number of particles that would have been acquired in the full time period. The other is the inlet transmission efficiency. The ATOFMS instruments do not detect particles of all aerodynamic diameters equally. Larger particles are detected with a higher efficiency than smaller particles, so a scaling equation was applied to relate the particle detection efficiency to the aerodynamic diameter of a particle (Fergenson et al., 2001). The detection efficiency as a function of particle size can be expressed as

$$N = \alpha D_a^{\beta} \qquad (10)$$

where N is the number of particles in a given volume of air per particle observed by ATOFMS in that volume, D_a is the aerodynamic diameter in micrometer of the particle, and α and β are parameters that are determined through calibration measurements. They were set as 4999 and –3.236, respectively [Allen, et al, 2000]. The physical diameter for calculating the particle mass was estimated from the aerodynamic diameter by assuming the spherical particle with a density of 1.3 g.cm^{-3} [Allen, et al, 2000]. In addition, in order to ensure the statistical reliability of the clustering results among 55 time periods, those that contained less than 1000 particles were excluded from analysis. Thus, 52 time periods were retained for analysis, i.e., 52 mass fraction vectors each of which was of dimension 117, were created to build the calibration model.

In the species concentration data, the species whose missing or below detection limit measurements were more than one third of the total measurements were excluded from the analysis. Ultimately, 35 species (NH_3, Cl^-, NO_3, SO_4, NH4, Na(a) (a: soluble), K(a) (a: soluble), OC1 (OC: organic carbon), OC2, OC3, OC4, OP (organic pyrolized carbon), OC (total organic carbon), EC1 (EC: element carbon), EC2, EC, TC (total carbon), Na, Mg, Al, Si, S, Cl, K, Ca, Mn, Fe, Ni, Cu, Zn, As, Se, Br, Rb, and Pb) each of which had 52 measurements were retained to build the calibration model. Thus, both independent and dependent variables for PLS calibration model were available.

Initial Calibration Model

In PLS calibration model, how to determine the number of PLS components is one of the keys to the success. In order to avoid the over-fitting, cross-validation was applied in this study to test the predictive significance of each PLS component. "Leave one out" strategy was used for cross-validation, so 52 runs were executed for each candidate number of PLS components. The predictive error of sum of squares (PRESS) was used as the criterion of PLS model in this study, and could be expressed as

$$PRESS = \sum_{j=1}^{35} \sum_{i=1}^{52} (y_{ij} - \hat{y}_{ij})^2 \qquad (11)$$

where y_{ij} is the concentration of species *j* in sample *i* (i.e., the sample that is left out for testing in the *i*th run), and $\hat{y}_{ij}$ is the prediction of y_{ij}. In this study, the independent and dependent variables were scaled and centered to make their distribution more symmetrical. Such scaling helps to ensure that one variable does not overwhelm other variables.

Different numbers of PLS components produce different PRESS values as shown in Figure 17. First, PRESS decreased and then stayed relatively flat, but then increased quickly after the number of PLS components reached 15. This rise is because noise began to be included into the model when extra PLS components were added. This result is called "over-fitting", which could create a well-fitted model with little or no predictive ability. Thus, the proper PLS components number should lie in the relatively flat range. In this case, the PLS number was selected as 6, since it produced the smallest PRESS.

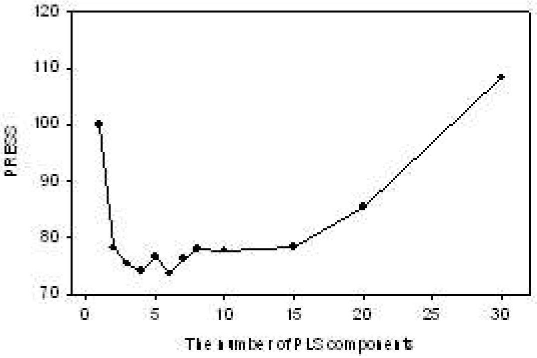


Figure 17. The PRESS value plotted as a function of the number of PLS components.

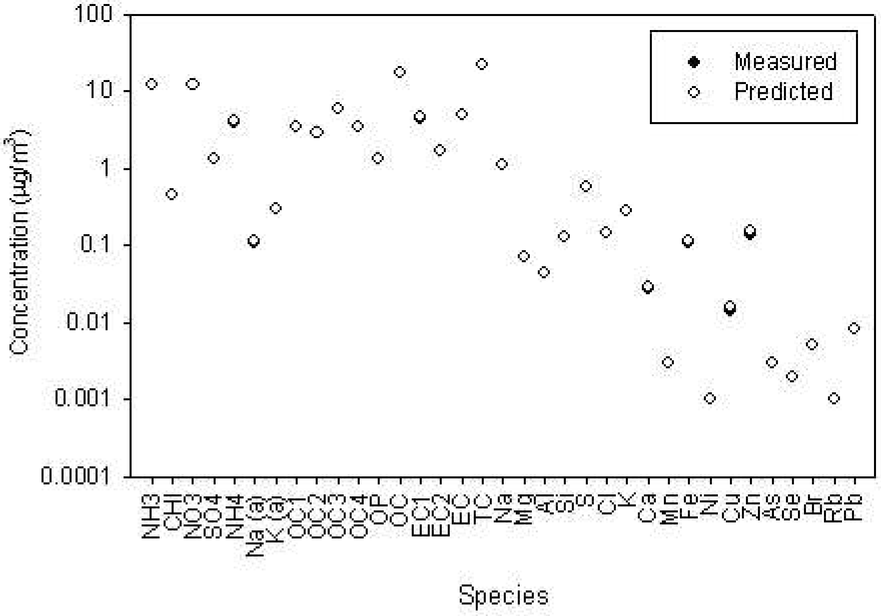


Figure 18. Average predicted concentration vs. average measured concentration for each species.

The main results of 6 PLS components are shown in Figures 18, 19, 20, and 21, and Tables 3 and 4. Figure 18 shows that the average of the predicted concentrations for each species, i.e., $\hat{A}E_j = \underset{i}{mean}(\hat{y}_{ij})$, almost exactly equaled to the average of the measured concentrations, i.e., $AE_j = \underset{i}{mean}(y_{ij})$. Let $REA = \underset{j}{mean}\left(\left|\hat{A}E_j - AE_j\right| / AE_j\right)$, then *REA* was only 3.4%. Here, "$\underset{i}{mean}$ " and "$\underset{j}{mean}$ "denote the average over "*i*: sample" and "*j*: species". In fact, $\underset{i}{mean}(y_{ij})$ and $\underset{i}{mean}(\hat{y}_{ij})$ provide representations of the distributions of the measurements and predictions, respectively. Thus, Figure 18 shows the predictive power of the PLS calibration model from this aspect.

In this initial study, there were 52 samples, but there were some missing and below detection limit values in the species concentration data, so 11 samples (none of which contained missing or below detection limit values) were used as typical examples to test the prediction effect of PLS calibration model on individual sample. The corresponding results were shown in Figures 19 to 21. It can be seen that almost every black circle (denoting measurement) was covered by or overlapped a white circle (denoting predicted value). Compared with Figures 2 to 5 of the previous study (Fergenson et al., 2001), the prediction effect on individual sample were substantially improved in this study. There is one extremely large error between the predicted and measured concentration of OP in sample 32. A reason for this large discrepancy may be the very low measured concentration for OP in this sample. The average concentration of OP was 1.349 μg/m^3 while the concentration of sample 32 was only 0.014 μg/m^3. Thus, this measurement could be considered as an outlier.

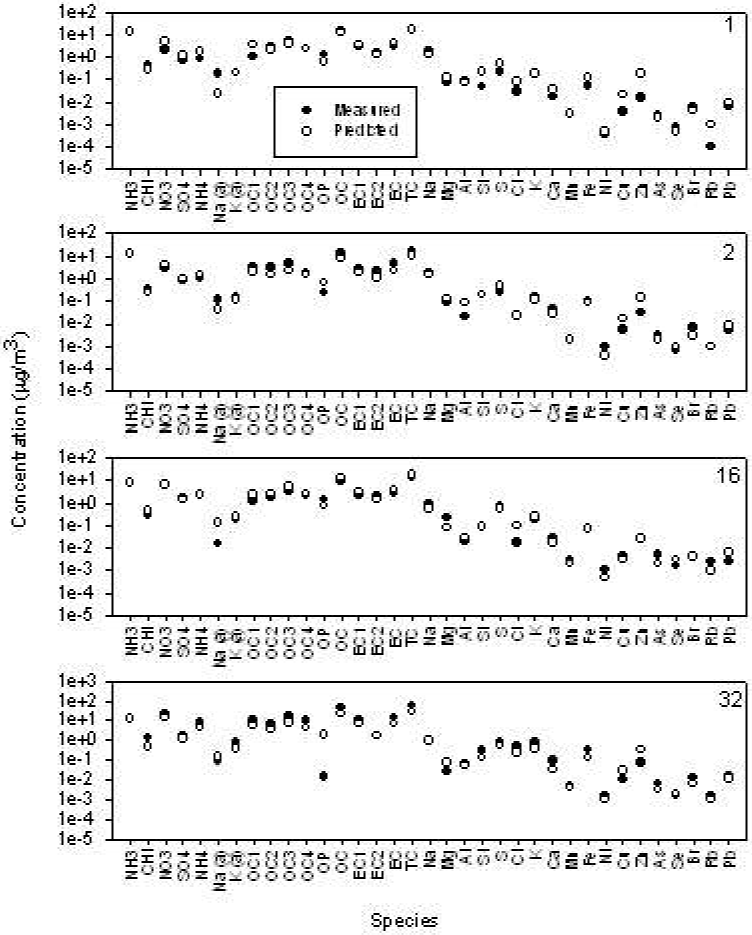


Figure 19. Predicted concentration vs. measured concentration for each species of individual sample (a). The number at right upper corner denotes the sample index.

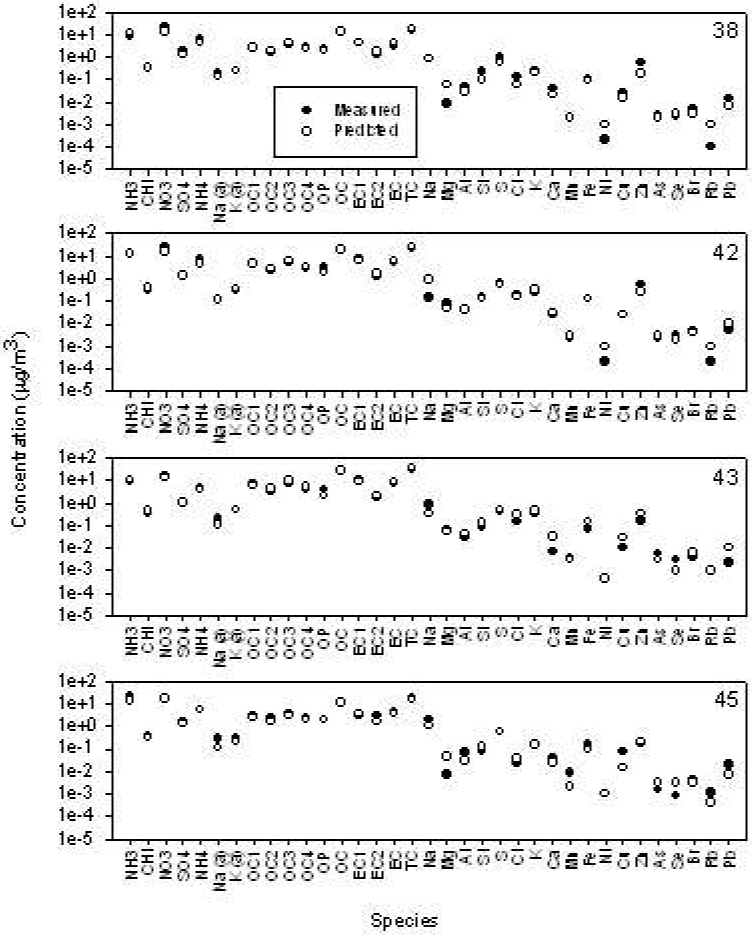


Figure 20. Predicted concentration vs. measured concentration for each species of individual sample (b). The number at right upper corner denotes the sample index.

Table 3. The average of the relative errors between the predicted and measured concentrations of 11 samples.

Species	TC	OC	NH_3	Rb	Ni	Se
REj	19.6%	18.4%	13.7%	230.5%	124.5%	79.5%

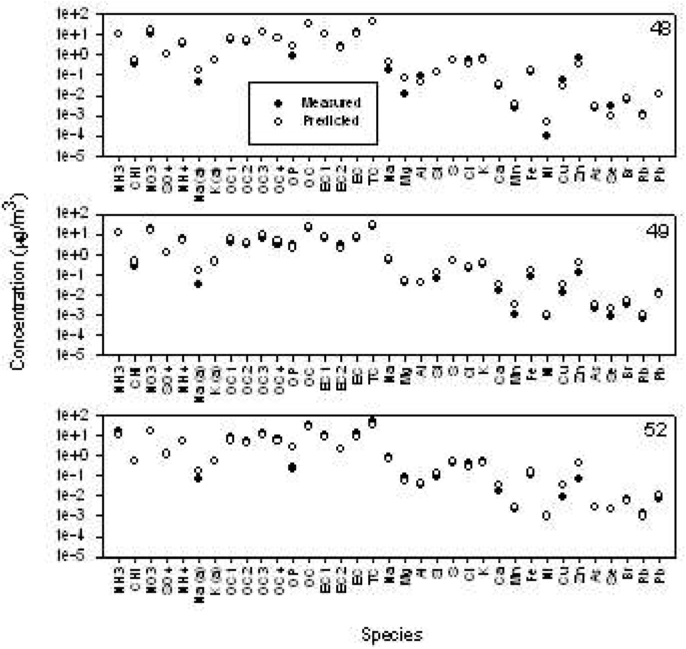


Figure 21. Predicted concentration vs. measured concentration for each species of individual sample (c). The number at right upper corner denotes the sample index

Besides checking the prediction effect on the average concentration, **r**elative **e**rror (RE) of each predicted concentration was also taken into account. The corresponding criterion was defined as $RE_j = \underset{i}{mean}\left(\left|(\hat{y}_{ij} - y_{ij})\right| / y_{ij}\right)$, i.e. the average value of the relative errors of the predicted concentrations for species *j*. The variables for PLS model were scaled and centered in this study [Wold et al., 2001], but a problem similar to that seen in Fergenson et al. [2001] occurred. The problem is the species with larger quantities were fit much better than those with smaller values. Table 4 showed the RE_j values of six species. TC (total carbon), OC (organic carbon), and NH_3 were the 3 species with the highest average measured concentrations, while Rb, Ni, and Se were the 3 species with the lowest values. Eleven samples for the Figures 19-21 were again used in this comparison. It could be seen from Table 4 that the RE_j values of the smallest 3 species were much larger than those of the largest 3 species.

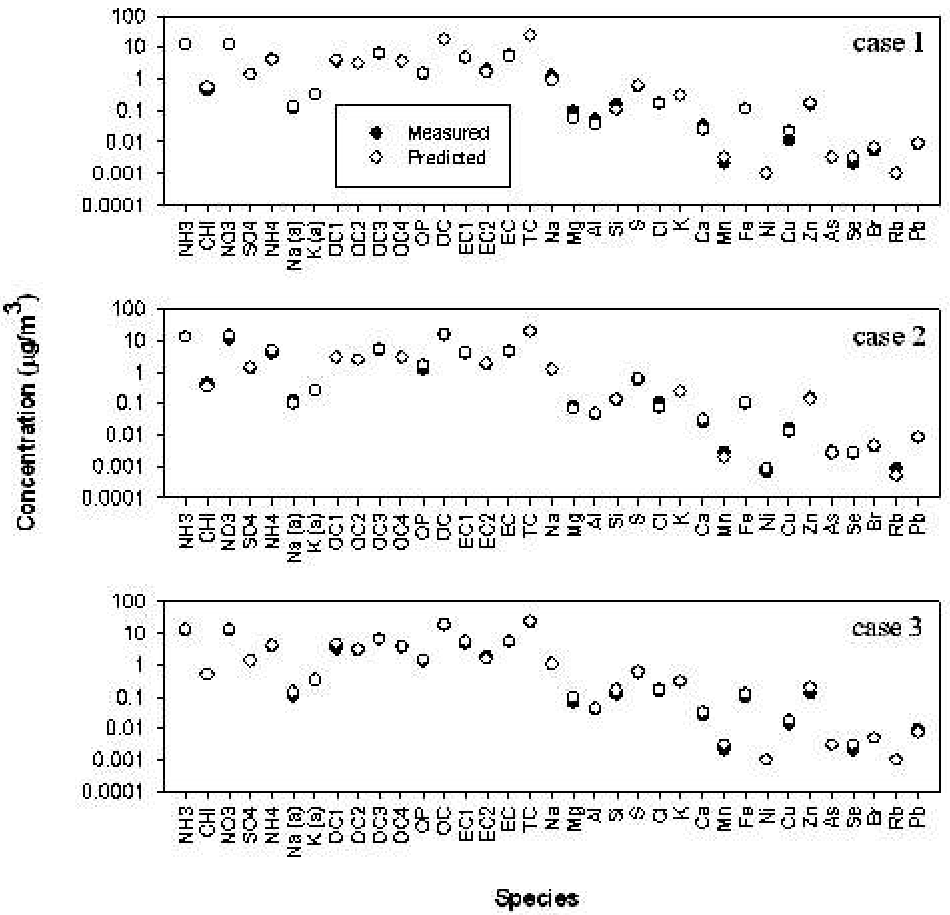


Figure 22. Average predicted concentration vs. average measured value for each species.

One possible reason was that those species in the larger quantities exerted the greater influence on the ART-2a analysis, ensuring that the identified classes would have a greater dependence on their concentrations [Fergenson, et al., 2001]. The parameters for calculating the detection efficiency and the estimated physical density of particle may also be reasons for the prediction errors of the whole system.

Another criterion for testing prediction ability is the correlation coefficient R^2 between the predicted and measured values. Table 4 presents the values for this study and the previous study [Fergenson et al., 2001]. R^2 (F) was the result of Fergenson et al. [2001], and R^2 (1820) and R^2 (385) corresponded to all the measurements (52*35=1820) and the measurements of 11 normal samples (11*35=385), respectively. Both R^2 (1820) and R^2 (385) were better than that of the previous study, in agreement with the better prediction results of the calibration model in this study. However, R^2 (385) was better than R^2 (1820). This proved the noise resisting ability of PLS model, because the prediction effect on the normal samples was not disturbed by the samples that contained missing and below detection limit values.

Table 4. The correlation coefficient of the predicted and measured concentrations.

R^2 (F)	R^2 (1820)	R^2 (385)
0.83	0.87	0.9

These results suggest that the single particle data can provide good estimates of bulk aerosol composition. There are additional studies that need to be made. In the current study, all but one of the samples are used to develop a calibration model. We will explore the size of the calibration data set necessary to produce adequate calibration models.

Realistic Calibration Model

The model presented above worked well, but it does not make sense to use all but one sample to predict the remaining sample. To be an effective method, it has to be possible to develop an effective calibration model from an appropriate subset of the available data. Thus, a second study has been made. In order to test the prediction ability of PLS calibration model, 20 samples were randomly selected from the 52 samples to build the model and the remaining 32 samples were for testing.

Different numbers of PLS components produced different PRESS values as shown in Figure 22. First, PRESS decreased and then stayed relatively flat, but finally increased quickly with the increasing on PLS components number. This rise is because noise began to be included into the model when extra PLS components were added into model. This is called “over-fitting”, which could create a well-fitting model with little or no predictive ability. Thus, the proper PLS components number should lie in the relatively flat range. In this study, the PLS components numbers for 3 cases were set as 4, 3 and 5 respectively, since they produced the smallest PRESS.

The main results of the three cases with the proper PLS components numbers are shown in Figure 23 and Tables 5 and 6. Figure 23 showed that the average of the predicted concentrations for each species, i.e., $\hat{AE}_j = \underset{i}{mean}(\hat{y}_{ij})$, almost exactly equaled to the average of the measured concentrations, i.e., $AE_j = \underset{i}{mean}(y_{ij})$. Here, "$\underset{i}{mean}$ " denotes the average over "i: sample". In fact, $\underset{i}{mean}(\hat{y}_{ij})$ and $\underset{i}{mean}(y_{ij})$ provide a representation of the distributions of the measurements and predictions, respectively, so Figure 23 showed the predictive power of the PLS calibration model from this aspect.

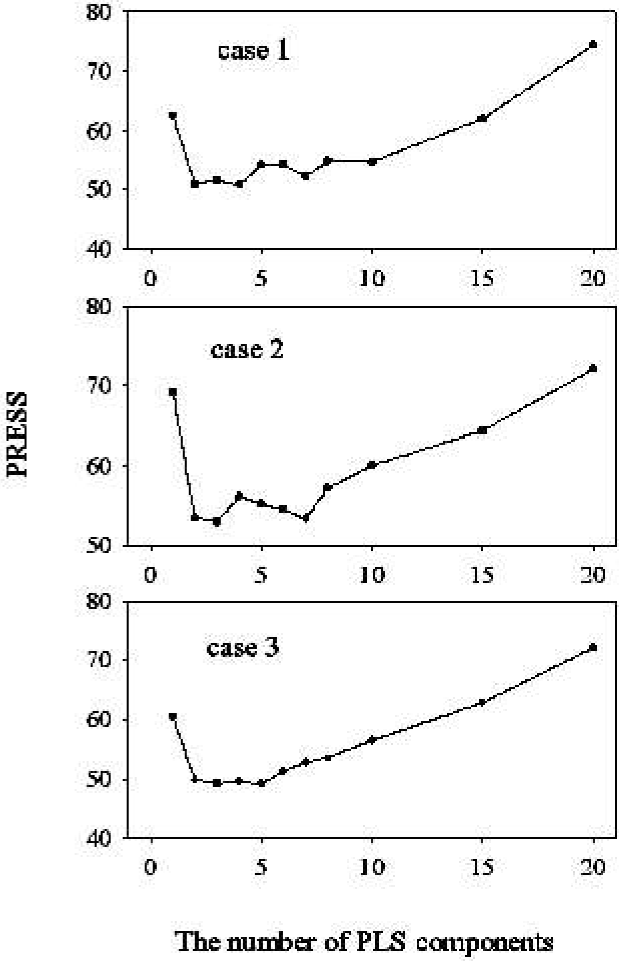


Figure 23. PRESS vs. PLS components number

Among the 52 samples, 11 samples did not contain any missing and below detection limit values in the species concentration data (they were called "normal samples set"), so in the testing samples of each of 3 cases, those belonging to the normal samples set were selected as typical examples to discuss the prediction effect of PLS calibration model on individual sample. In each case here, there is very good agreement between the predicted and measured values. Compared with Figures 2 to 5 of the previous study [Fergenson, 2001], the prediction effect on individual sample was substantially improved in this study. There are extremely large errors between the predicted and measured concentration for OP in one particular sample. This sample has a measured OP concentration of only 0.0143 µg/m^3. The average OP concentration of the data set was 1.349 µg/m^3, and this measurement can be considered as an outlier.

Table 5. The average of the relative errors between the predicted and measured concentrations of testing samples.

Species	TC	OC	NH_3	Rb	Ni	Se
Case 1	23.0%	22.0%	22.3%	161.5%	63.6%	98.1%
Case 2	15.7%	16.0%	20.1%	101.9%	128.8%	107.3%
Case 3	24.2%	23.7%	29..5%	316.9%	124.1%	107.2%

Table 6. The correlation coefficient of the predicted and measured concentrations.

Items	R^2(F)	R^2(TA)	R^2(TN)
Case 1	0.83	0.838	0.874
Case 2		0.807	0.842
Case 3		0.8610	0.9437

Besides checking the prediction effect on the average concentration, **r**elative **e**rror (RE) of each predicted concentration was also taken into account. The corresponding criterion was defined as $RE_j = \underset{i}{mean}\left(\left|\hat{y}_{ij} - y_{ij}\right| / y_{ij}\right)$, i.e. the average value of the relative errors of the predicted concentrations for species *j*. The variables for PLS model were scaled and centered in this study, which could ensure the PLS model effect in some extent [Wold et al., 2001], but a problem similar to that seen in Fergenson et al. [2001] occurred. The problem is the species with larger quantities were fit much better than those with smaller quantities. Table 5 showed the RE_j values of six species of 3 cases. TC (total carbon), OC (total organic carbon), and NH_3 were the 3 species with highest average measured concentrations, while Rb, Ni, and Se were the 3 species with lowest values. It can be seen in Table 5 that the RE_j values of the smallest 3 species were much larger than those of the largest 3 species.

One possible reason was that those species in the larger quantities exerted the greater influence on the ART-2a analysis, ensuring that the identified classes would depend on their concentrations [Fergenson, et al., 2001]. The parameters for calculating the detection efficiency and the estimated physical density of particle were also the reasons for the prediction errors of the whole system.

Another criterion for testing prediction ability is the correlation coefficient R^2 between the predicted and measured values. Table 6 showed the values of this study and the previous study [Fergenson, 2001]. R^2(F) was the result of Fergenson et al. [2001], and R^2(TA) and R^2(TN) corresponded to all the testing samples in each case and those belonging to the normal samples set, respectively. In the three cases, the numbers of the measurements for R^2(TA) were all 32*35=1120, and those for R^2(TN) were 8*35=280, 8*35=280, and 7*35=245, respectively. All the R^2 values but the R^2(TA) for case 2 are better than that of the previous study, in agreement with the better prediction results of the calibration model in this study. Moreover, all the R^2(TN) values are better than R^2(TA) values. This proved the noise resistance of the PLS model because the prediction effect on the normal samples was not disturbed by the samples that contained missing and below detection limit values.

PMF ANALYSIS OF ATOFMS DATA

In addition to completing the calibration modeling studies, we have applied PMF to the ATOFMS data from Fresno, CA to ascertain the ability to obtain source resolutions. The approach that was taken is to use the number of particles in each particle class during a given sampling interval (sample) as the input to the PMF analysis. As we began the study, we recognized the need for an improved estimation of the uncertainties of the numbers of particles in each particle class since the performance of PMF is critically dependent on those uncertainties estimates. To solve this problem we are developing a new approach using a bootstrapping method [Efron, 1982].

Bootstrap Uncertainty Estimation

The bootstrap is a resampling method in which replicate data sets are produced by sampling of the original data with replacement. In this case, approximately 250,000 particles were measured during the sampling program in Fresno. Thus, a new sample of 2500,000 particles is selected from the original data set such that some particles will be present multiple times and some particles will be absent. The particles are then parsed into their classes and into the time sampling intervals so that a matrix of number of particles in each of 117 classes in each of the sampling intervals are available. Then the process is repeated to produce a second data set. This is repeated a large number of times (e.g. 100 replicate data sets). Then the data set for the PMF analysis would be the average number of particles in each class obtained by averaging the class membership numbers for the 100 replicate samples. The uncertainty is the standard deviation associated with that average value. From this process, we then have the values and uncertainties needed as the input values to a PMF analysis.

PMF Studies

In this study, the ATOFMS data were collected from Jan. 9, 2001 to Feb. 2, 2001 in Fresno, CA. Not only the positive ions but also the negative ions were provided in the ATOFMS data, so the range of mass-to-charge (m/z) for this study was set as [-350, +350]. This means the input of ART-2a is a 700 dimensional vector. The total number of particles in this period is 509, 600.

Now the problem is how to classify such a great number of particles with ART-2a networks. In Fergenson et al. (2001), a total of 12, 479 particle samples were grouped into 12 cohorts which finally generated 12 samples for the calibration model. Each cohort was classified individually by ART-2a, but the weight matrix (cluster vectors) was preserved from one analysis to the next. Any particles that did not fit into the existing classes nucleated their own classes. The feature of ART-2a network that ART networks can create new class without disturbing the existing classes permit this procedure to function, but in term of system completeness, it would be better to classify all of the particles at the same time. Thus, in this study, 509, 600 particles were classified at the same time. The vigilance of ART-2a is the key parameter to control the classification results. The bigger it is, the more the classes are generated. In the initial experiments, some of the clustered centers of vigilance=0.6 were very similar, which means 0.6 may be too large. The vigilance was then set to 0.4 and a total of 559 classes were created. However, most of 559 classes just contained a very few particles, so only 37 classes were retained for further analysis based on three rules: 1) the selected 37 classes were the highest with respect to the particle masses of all classes, 2) the selected 37 accounted for over 90% of the total particle mass, and 3) each of the rest 522 classes accounted for less than 1% of the total particle mass.

The classification results were used to estimate the particle number concentration or mass concentration of each class in each time period. The time interval for each PMF sample was set as one hour, thus there were 600 (24 hours/day * 25 days) time periods. In this study, the mass concentrations generated more reliable results of the PMF model than the particle number concentrations. Detection efficiency is the key to the accurate estimation. The ATOFMS instruments do not detect particles of all aerodynamic diameters equally. Larger particles are detected with a higher efficiency than smaller particles, so a scaling equation was applied to relate the particle detection efficiency to the aerodynamic diameter of a particle [Fergenson et al., 2001]. The detection efficiency as a function of particle size can be expressed as

$$N = \alpha D_a^{\beta} \tag{13}$$

where N is the number of particles in a given volume of air per particle observed by the ATOFMS in that volume, D_a is the aerodynamic diameter in micrometer of the particle, and α and β are the parameters. In this case, the initial values of α and β were taken to be 4999 and –3.236, respectively [Allen, et al, 2000]. The physical diameter for calculating the particle mass was estimated from the aerodynamic diameter by assuming the spherical particle with a density of 1.3 g.cm^{-3} [Allen, et al, 2000].

The samples (inputs) for the PMF model were generated as below. In order to ensure the statistical reliability of the clustering results, among 600 time periods, those that contained less than 400 particles were excluded from analysis. Thus, 462 time periods were retained for analysis and the input of the PMF model was a 462 (samples) * 37 (classes) mass concentration matrix.

Now the problem is how to estimate the uncertainties of these calculated mass concentrations. In this study, bootstrapping was applied to estimate the uncertainties. However, the particle numbers in some periods for some classes are less than 10, so it is almost impossible to pick up the particles of these periods from the total 509, 600 particles by bootstrapping. Therefore, the values for these periods in the bootstrapped mass concentration matrix were zero or very close to zero, which influenced the PMF results considerably. For the convenience in discussion, this kind of period is called "abnormal" period. Finally, the estimated mass concentrations and uncertainties using bootstrapping were retained for the "normal" periods while those for the "abnormal" periods were replaced with the directly calculated mass concentrations. The uncertainties for the "abnormal" periods were 15% of the corresponding mass concentrations.

The positive ion spectra of centroid of the 37 particle classes are shown in Figures 24 -27 and the negative ion spectra are given in Figures 28 - 31. These cluster centroid spectra denote different types of particles emitted by the various sources. It can be seen that there may be no observable positive ions for several of the particle types. We need to consult with Professor Prather with respect to the interpretation of these spectra and as of this time, we have not had time to do so.

As for the PMF model, one of the keys to ensuring the model effect is to determine the number of factors. In this study, different numbers of factors varying from 6 to 21 were tried. The Q value showed a change in slope at 18 factors. In addition, the 18 factor Q is close to the theoretical value (462*37). Figure 32 shows the distribution of the scaled residuals for the identified classes. For almost all of the classes, the distributions are symmetric with acceptably small values. Therefore, it appears that 18 factors provides a reasonable solution. The corresponding profiles are shown in Figures 33-34 and the contributions in Figures 35-36.

In the profile plots, most of 18 factors are represented by one dominant class. For example, factor 2 is represented by class 10. In general speaking, each class obtained by ART-2a corresponds to one source, for example, sea salt, gasoline emission and so on. Therefore the factor which has more than 1 dominant classes may mean that the dominant classes are very similar or co-existing. For example, the positive ion spectra of classes 7 and 15 in factor 1 are very similar, so are classes 1 and 3 in factor 3 and classes 6 and 7 in factor 12.

In the contribution plots, most factors show distinct features, but the contributions of factors 4 and 5, 11 and 12, and 17 and 18 are very similar. The possible reasons are the factors represented by these classes can be integrated or may have the same wind direction effect.

We are working on interpreting the obtained classes to make sure all the classes have reasonable explanation. After that, it can be easier to determine the proper number of factors and more advance techniques of PMF model like rotation by "fpeak" and pulling down by "fkey" can be applied. Then more reasonable PMF results will be presented.

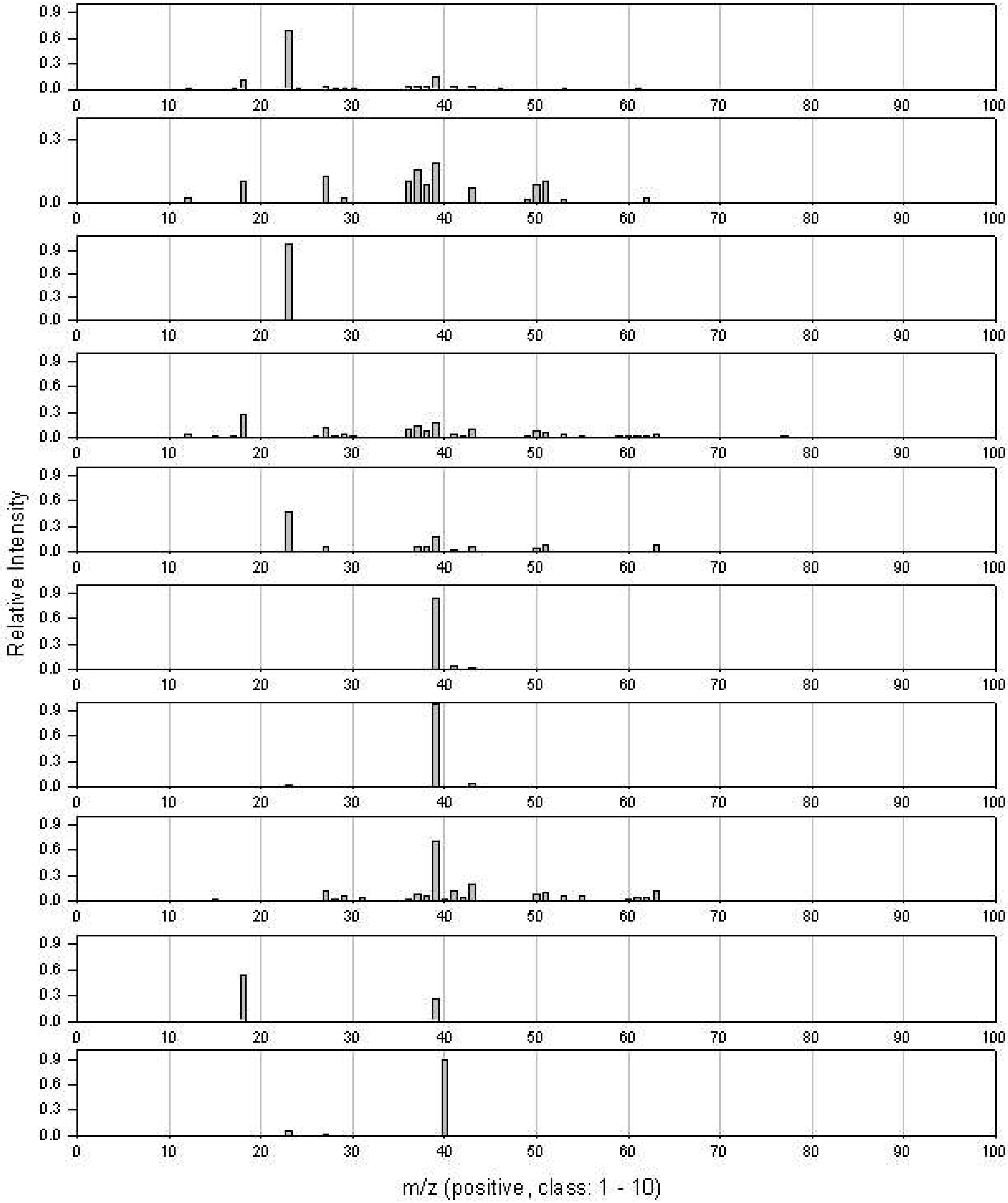


Figure 24. The positive ion spectra for classes 1 – 10.

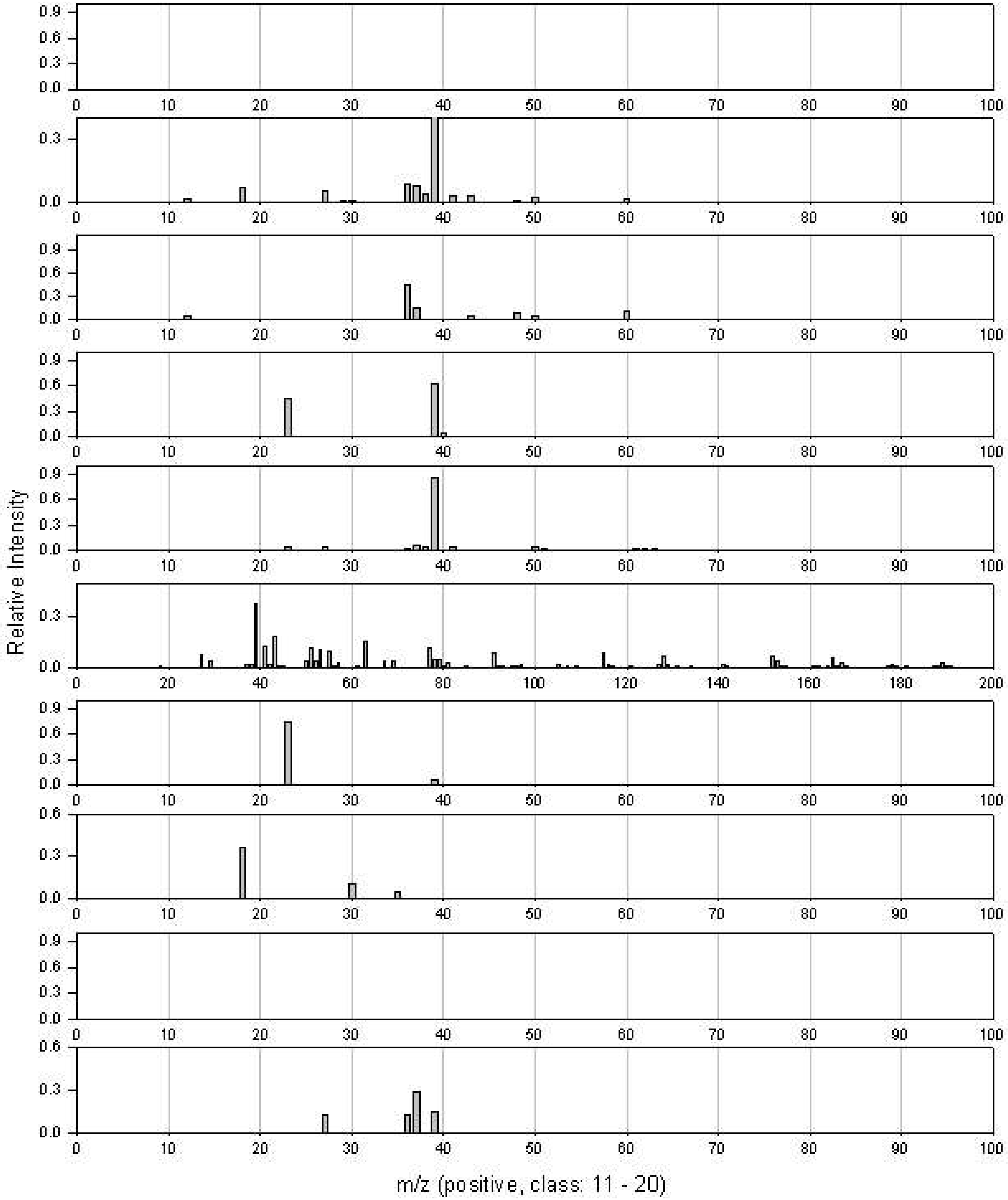


Figure 25. The positive ion spectra for classes 11 – 20

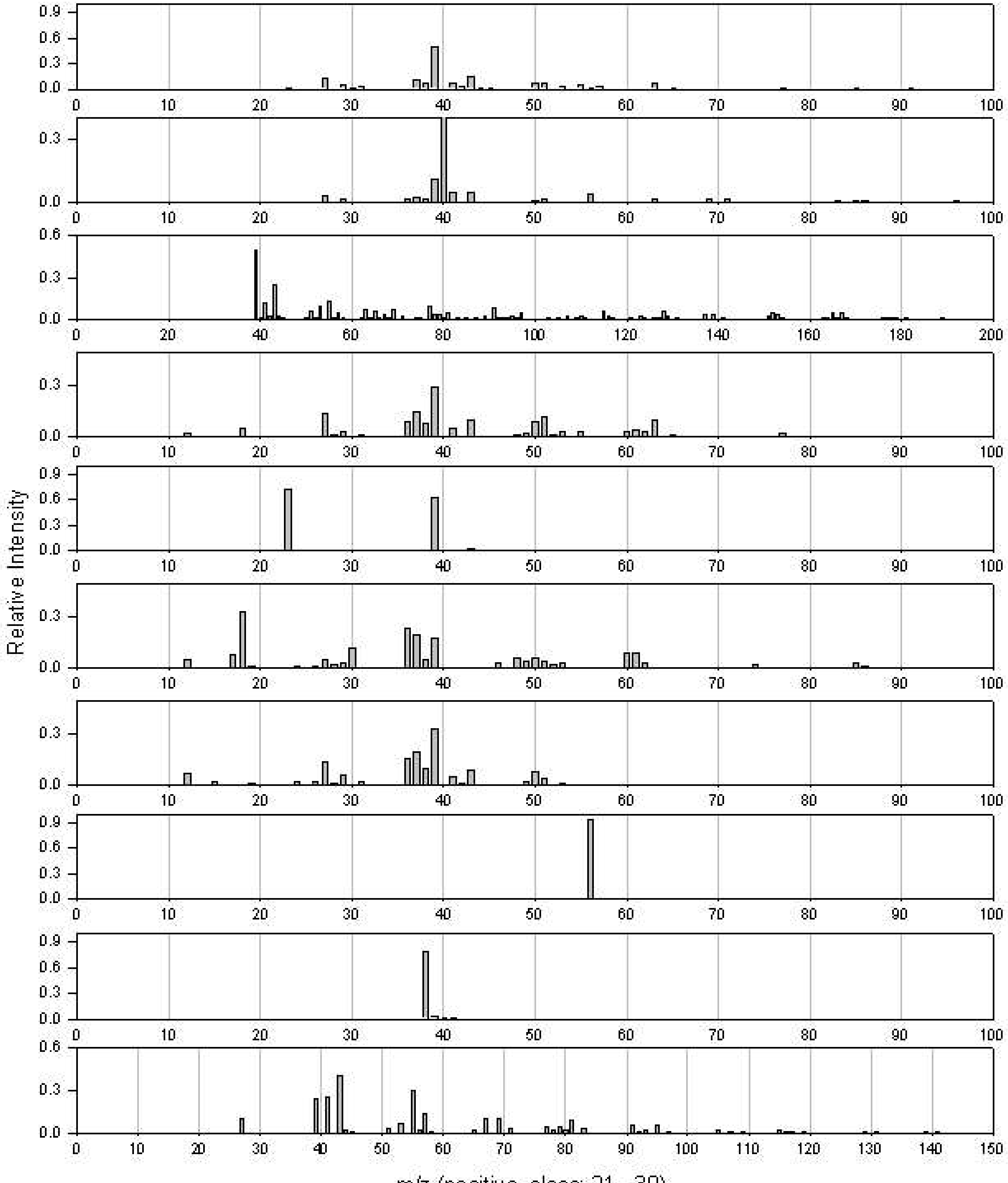


Figure 26. The positive ion spectra for classes 21 – 30

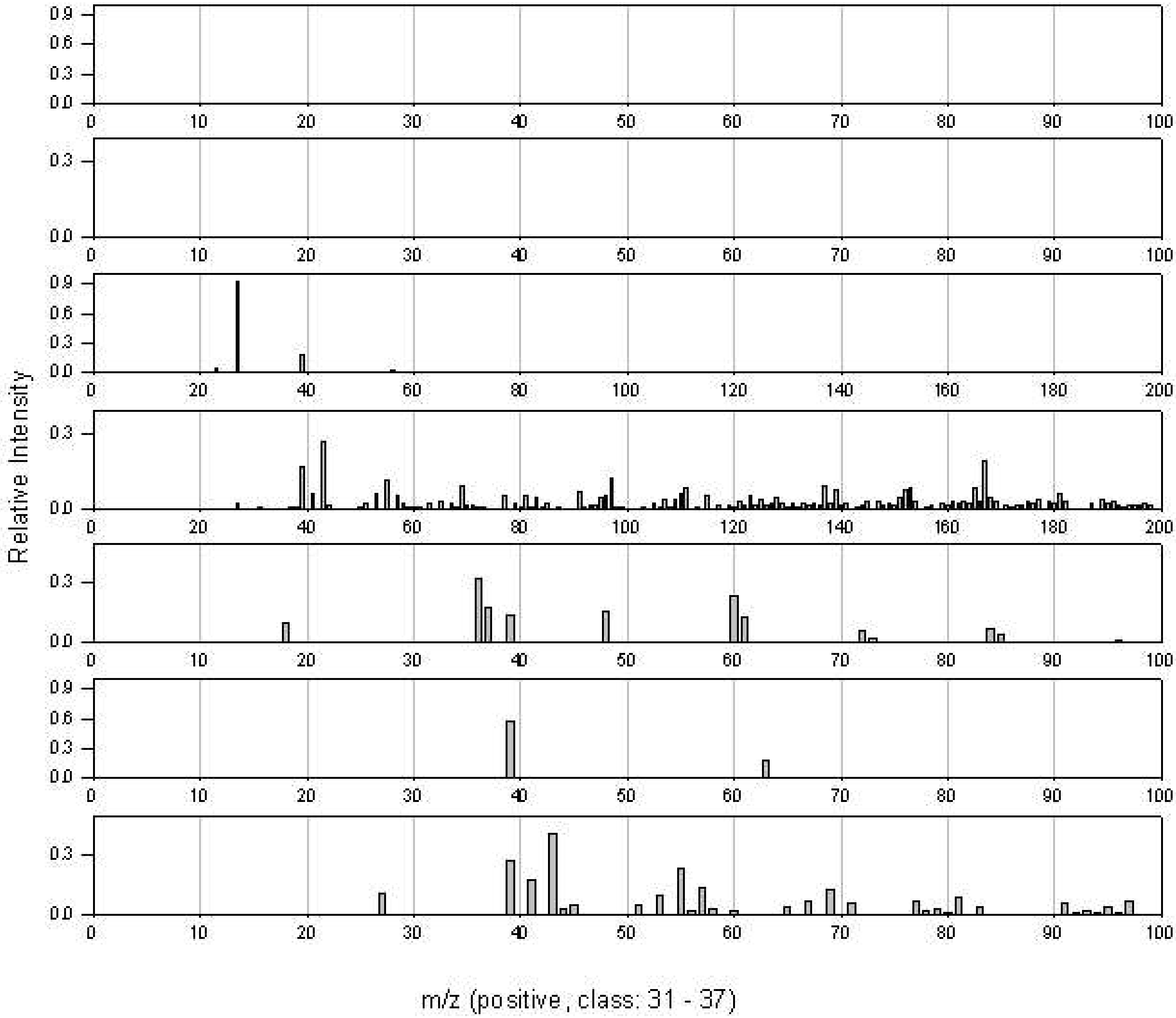


Figure 27. The positive ion spectra for classes 31 – 37

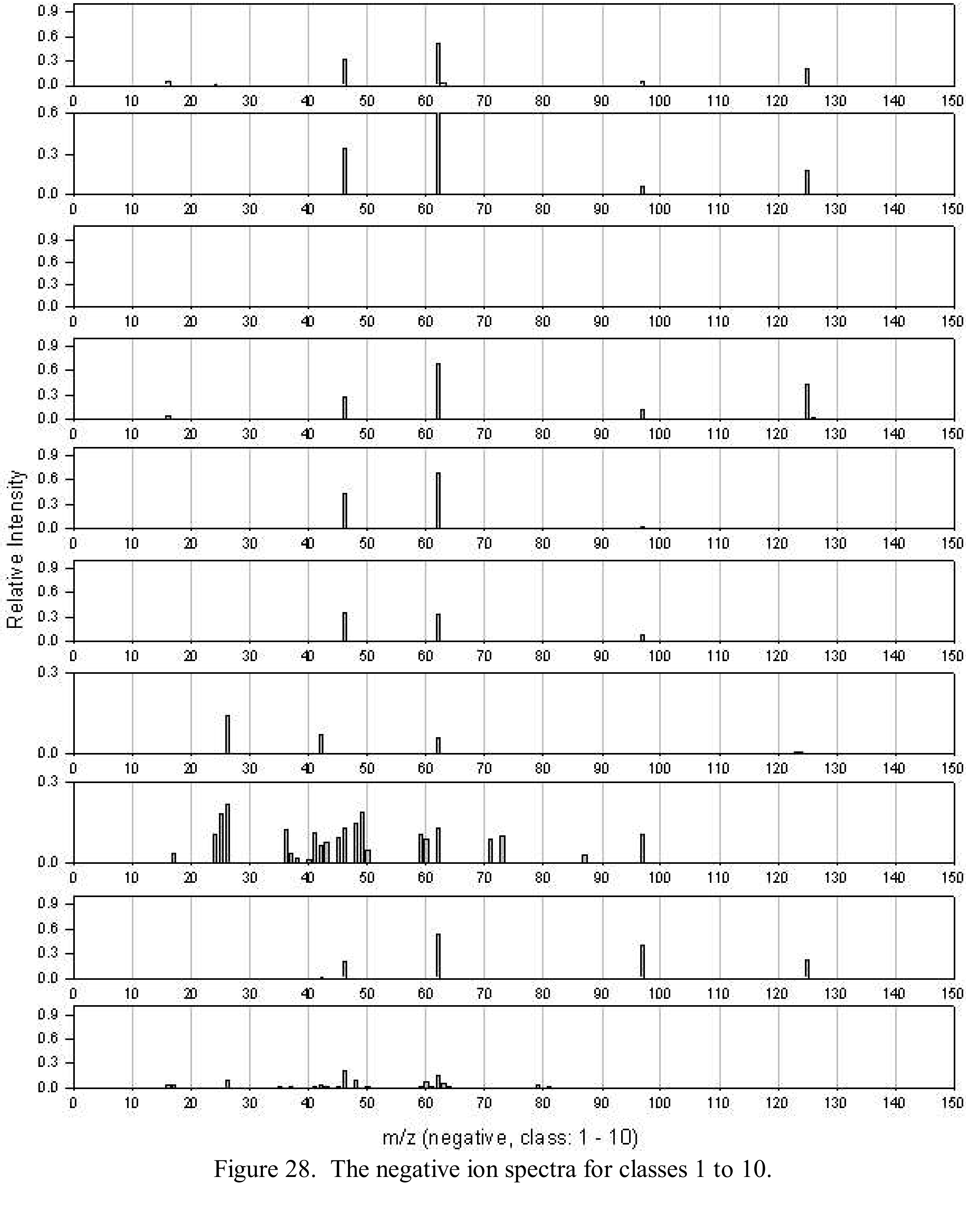


Figure 28. The negative ion spectra for classes 1 to 10.

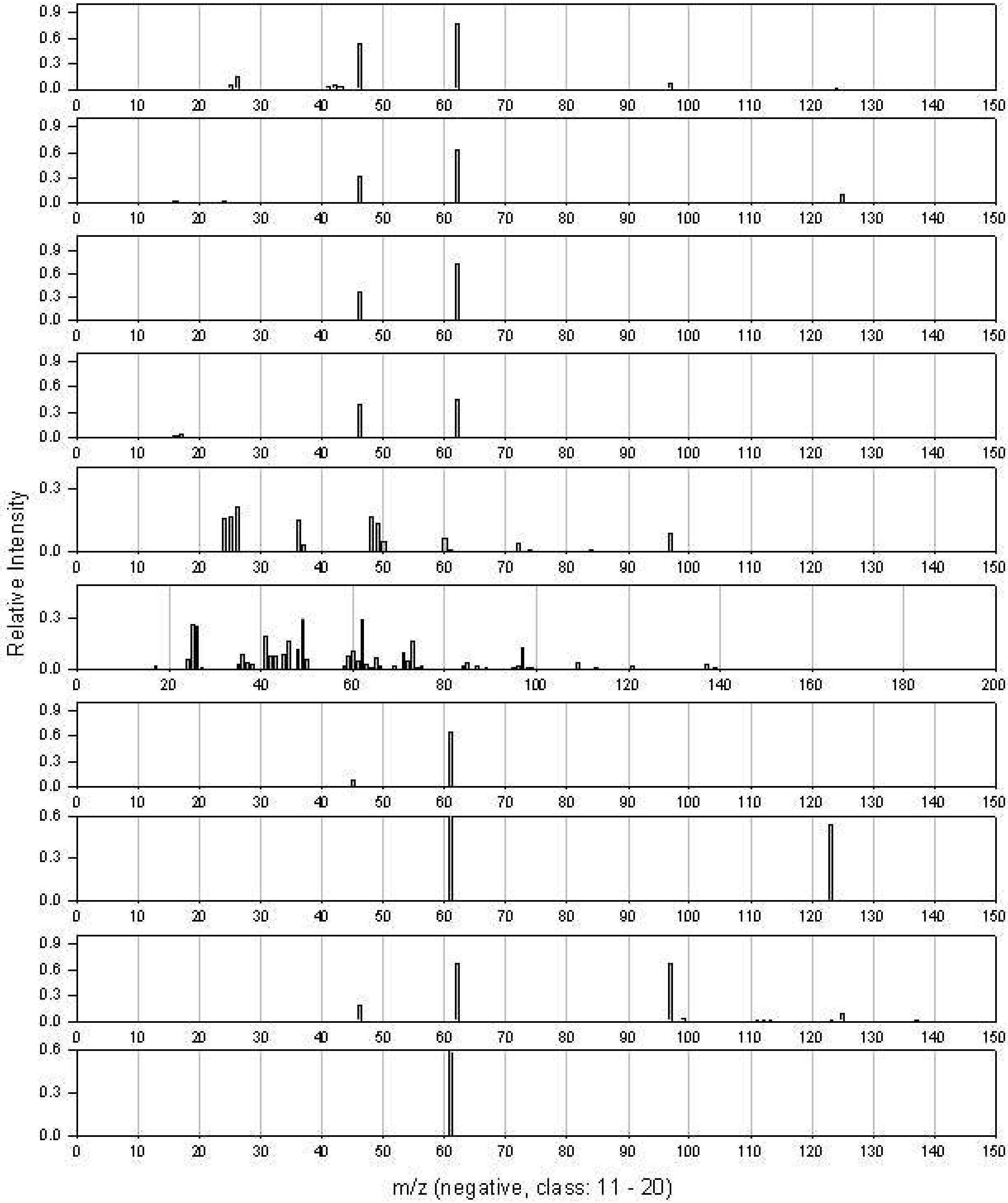


Figure 29. The negative ion spectra for classes 11 - 20.

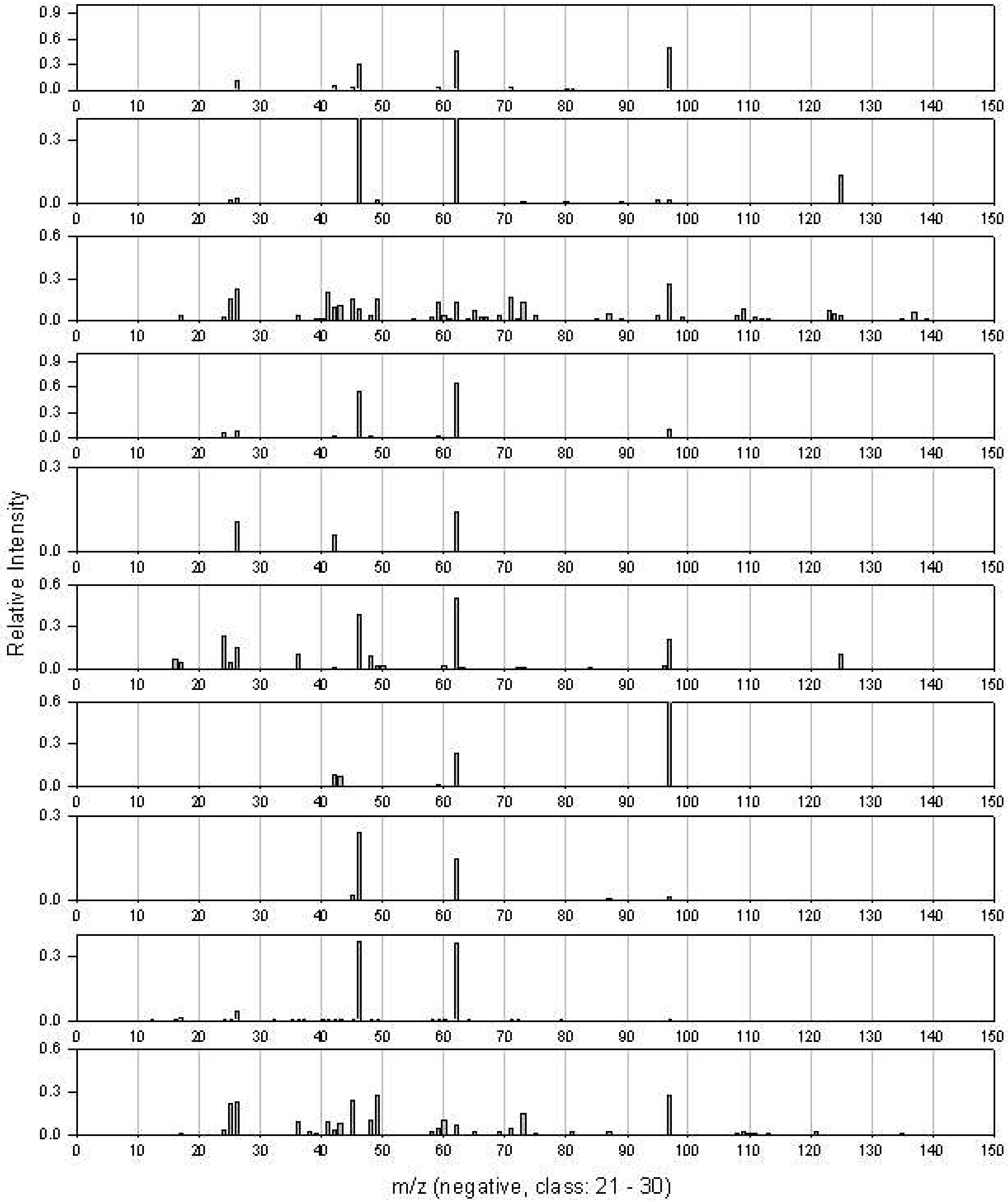


Figure 30. The negative ion spectra for classes 21 - 30.

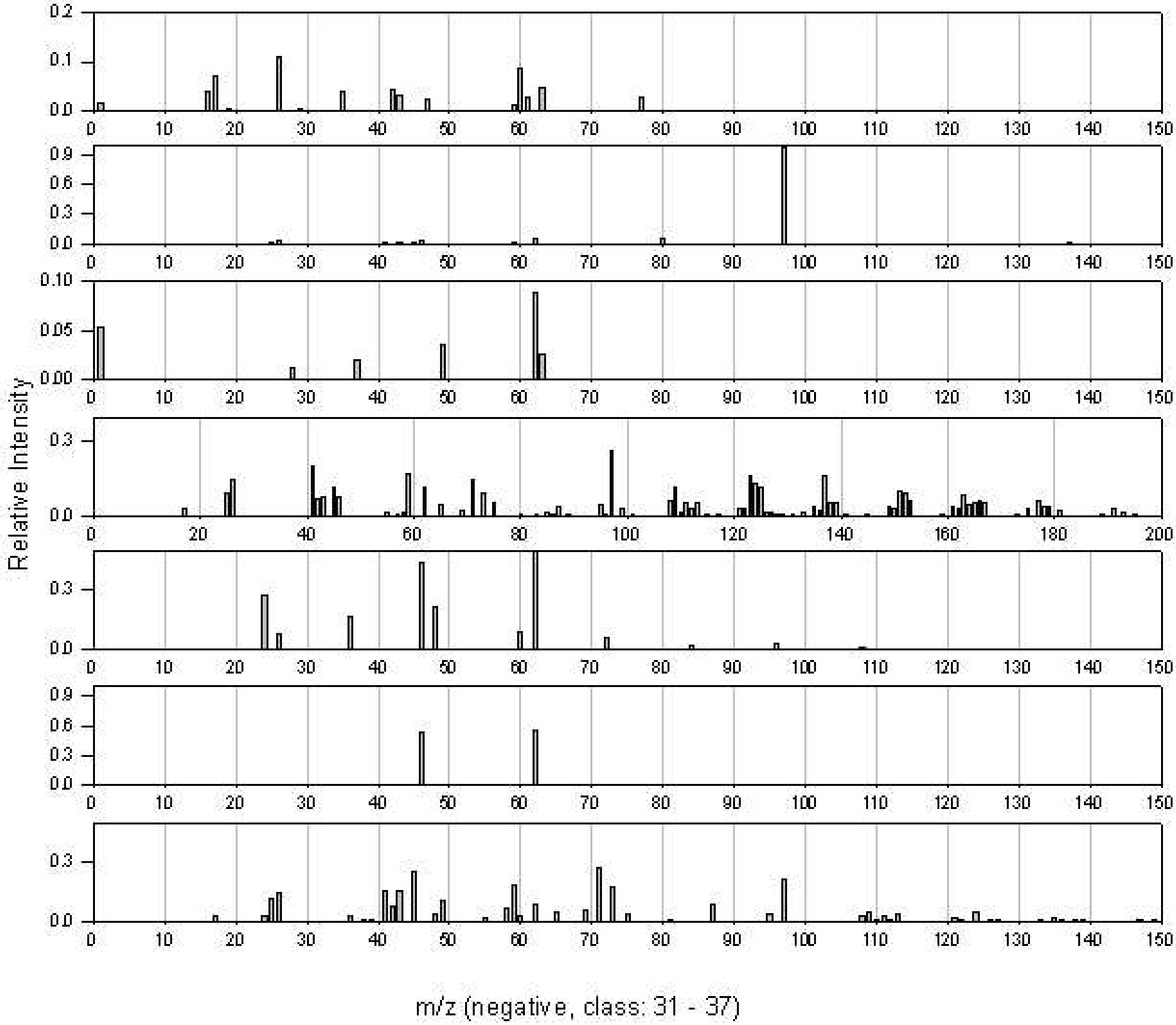


Figure 31. The negative ion spectra for classes 31 - 37.

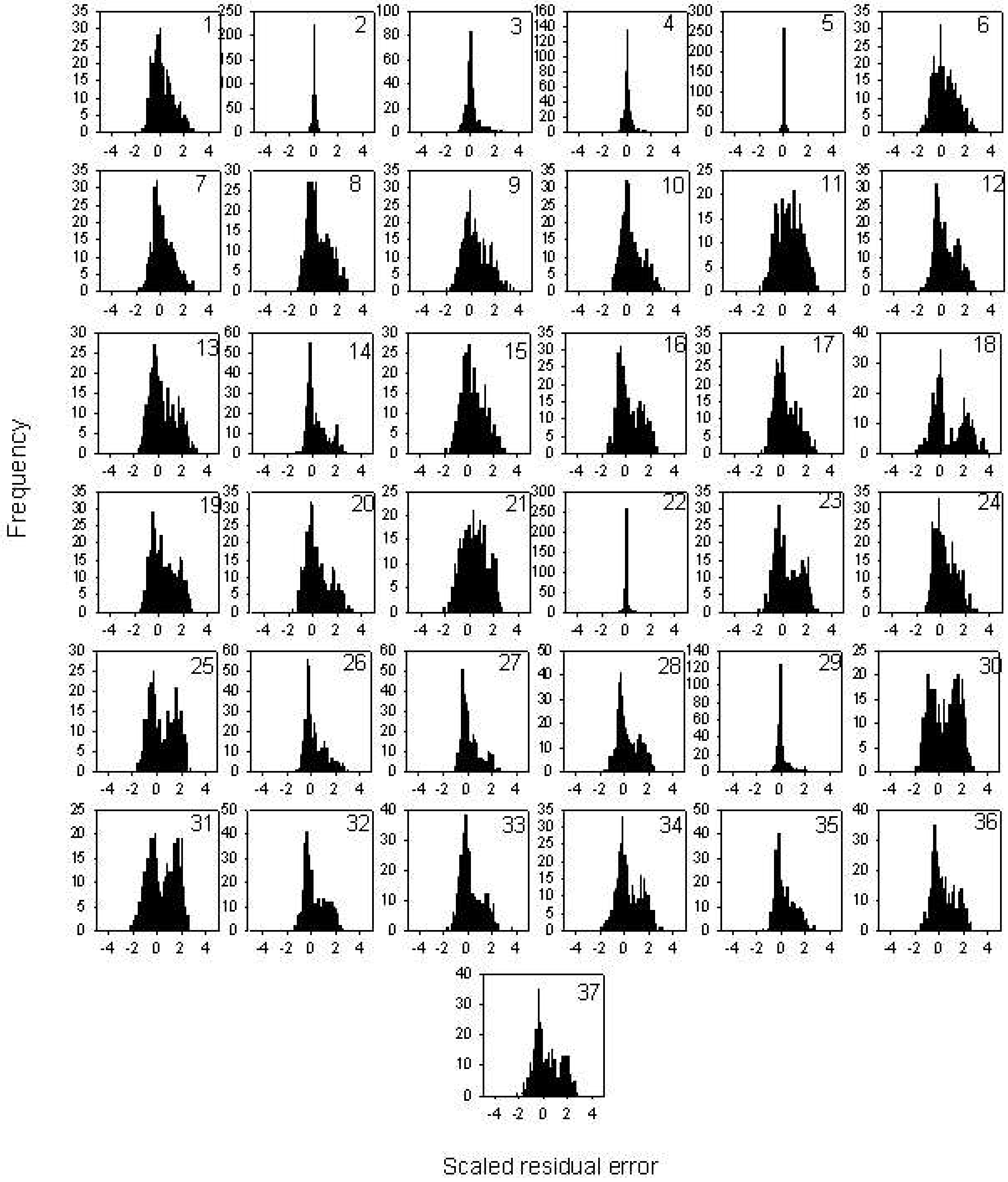


Figure 32. Scaled residuals for each of the 37 input classes.

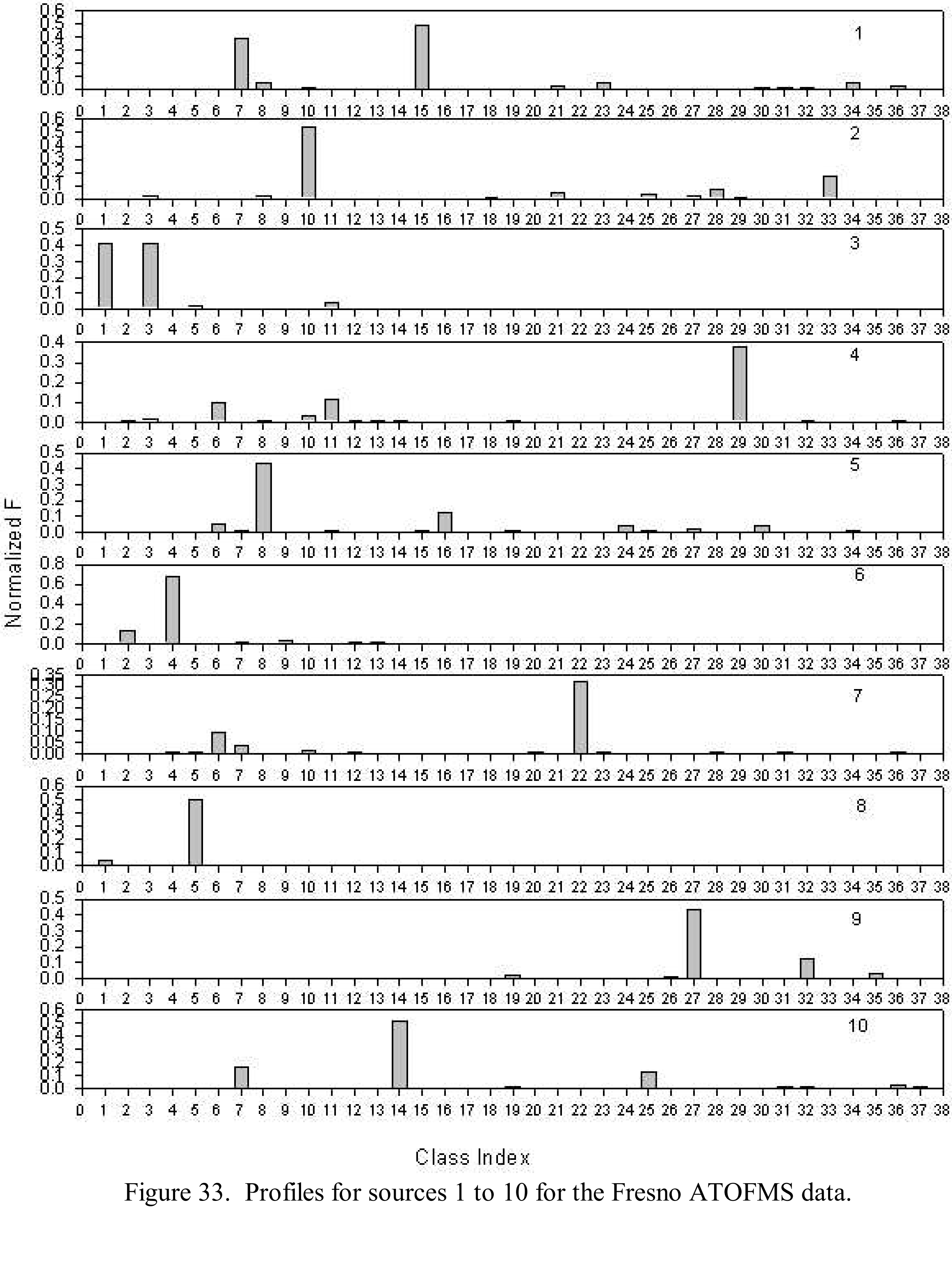


Figure 33. Profiles for sources 1 to 10 for the Fresno ATOFMS data.

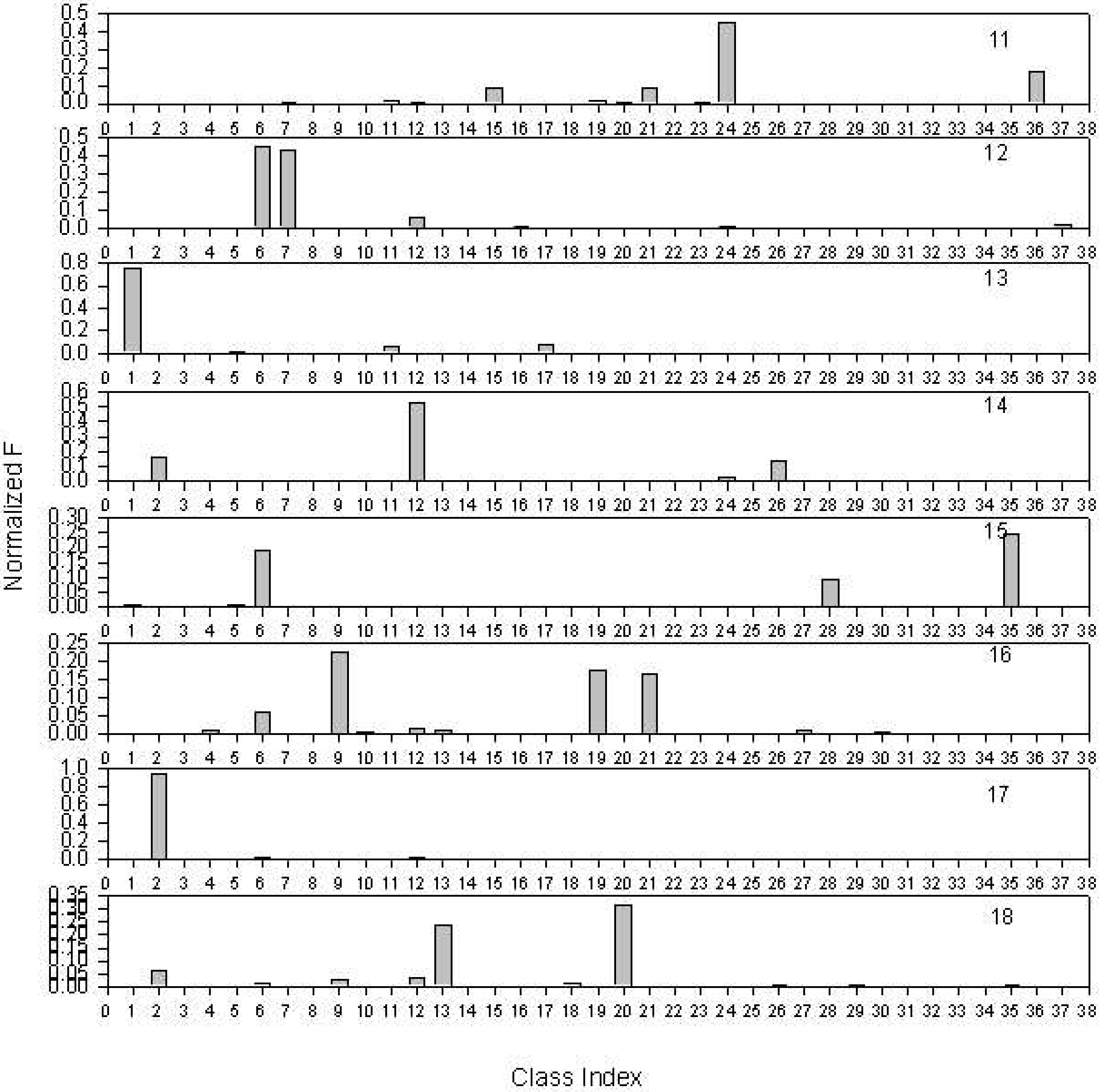


Figure 34. Profiles for sources 11 - 18 for the Fresno ATOFMS data.

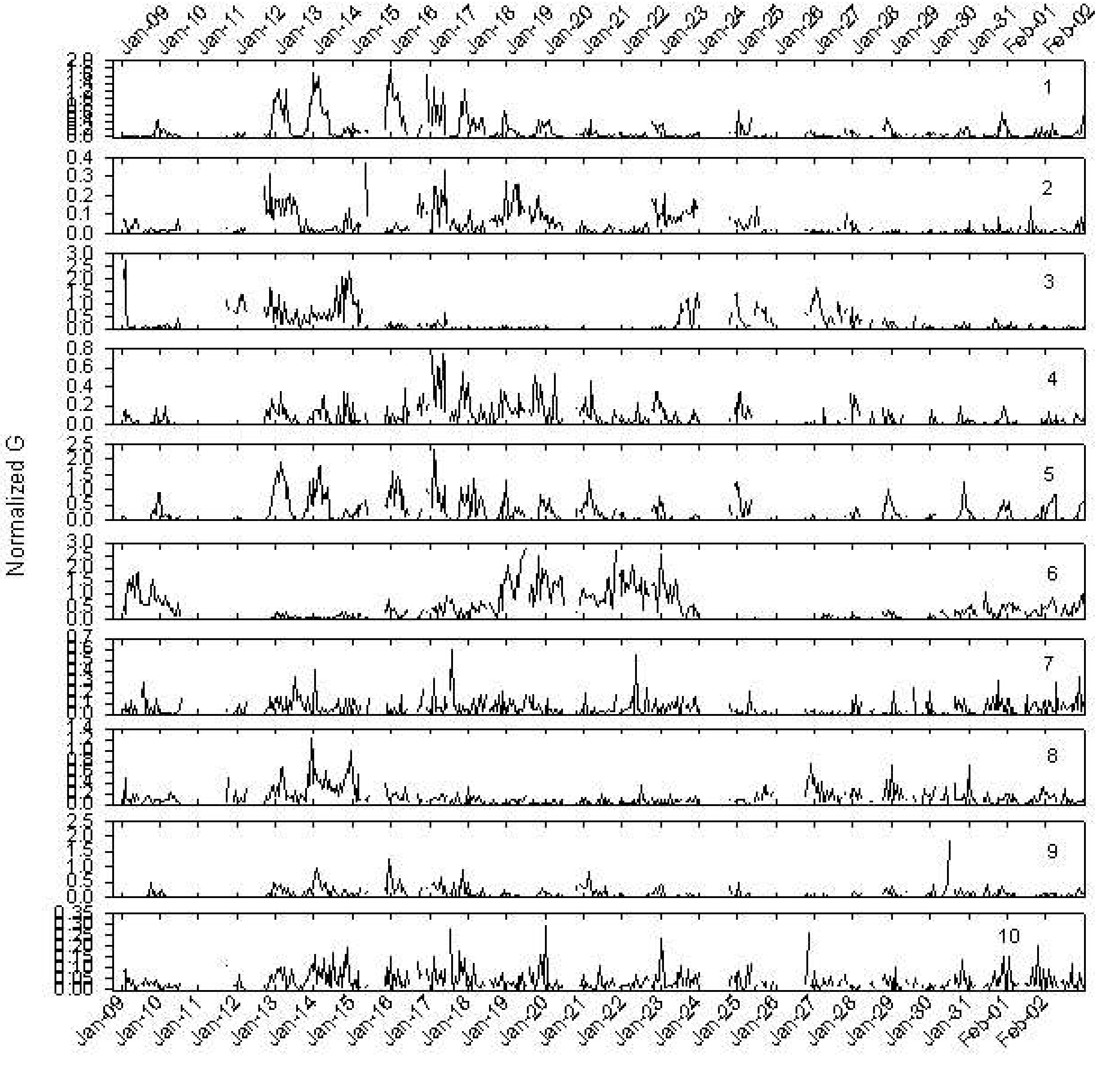


Figure 35. Time series of contributions for sources 1 to 10 for the Fresno ATOFMS data.

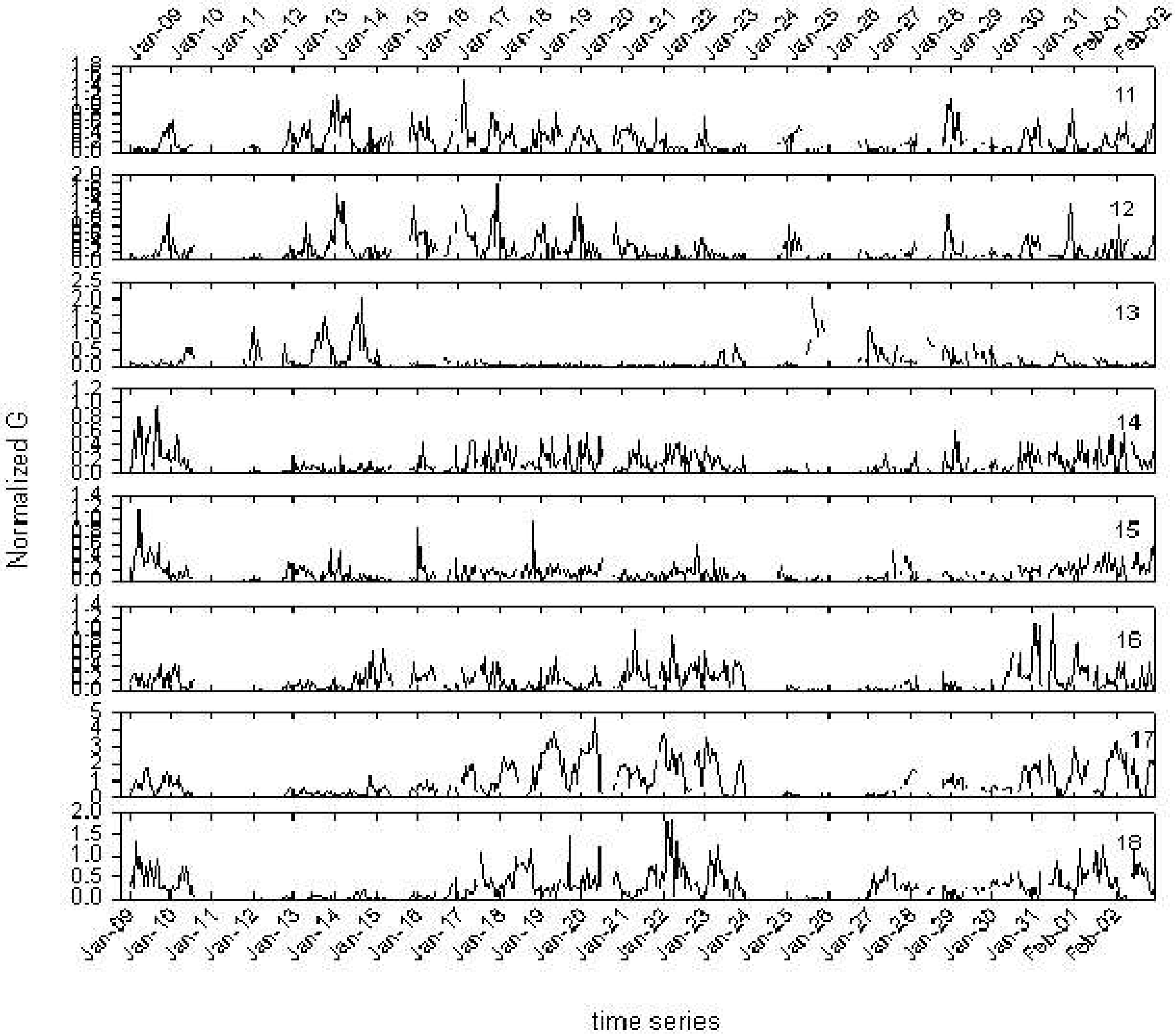


Figure 36. Time series of contributions for sources 11to 18 for Fresno ATOFMS data.

REFERENCES

Allen, J. O., Fergenson, D. P., Gard, E. E., Hughes, L. S., Morrical, B. D., Kleeman, M. J., Gross, D. S., Gälli, M. E., Prather, K. A., Cass, G. R., 2000. Particle Detection Efficiencies of Aerosol Time of Flight Mass Spectrometers under Ambient Sampling Conditions, *Environ. Sci. Technol.* 34, 211-217.

Andreae, M.O. (1990) Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. *Marine Chemistry* 30,1-29.

Anttila, P., P. Paatero, U. Tapper, and O. Järvinen (1995) Application of Positive Matrix Factorization to Source ApportionmentResults of a Study of Bulk Deposition Chemistry in Finland, *Atmospheric Environ.* 291705-1718.

Bates, T.S., Lamb, B.K., Guenther, A., Dignon, J., Stoiber, R.E. (1992) Sulfur emissions to the atmosphere from natural sources. *J. Atmospheric Chem.* 14, 315-337.

Berresheim, H., Wine, P.H., Davis, D.D. (1995) Sulfur in the atmosphere. In: Singh, H.B. (Ed.), *Composition, Chemistry and Climate of the Atmosphere*. Van Nostrand Reinhold,New York, ISBN 0-442-01264-0, pp. 251-302.

Carpenter, G. A., Grossberg, S., Rosen, D. B., 1991. ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition, *Neural Networks* 4, 493-504.

Charlson, R.J., Lovelock, J.E., Andreae, M.O., Warren, S.G. (1987) Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate, *Nature* 326, 655-661.

Chen, L.-L.; Carmichael, G. R.; Hong, M.-S.; Ueda, H.; Shim, S.; Song,C. H.; Kim, Y. P.; Arimoto, R.; Prospero, J.; Savoie, D.; Murano, K.; Park,J. K.; Lee, H.; Kang, C. (1997) Influence of continental outflow events on the aerosol composition at Cheju Island, South Korea, *J. Geophys. Res.*102(D23), 28,551-28,574.

Chow, J.C., Waston, J.G., Lowenthal, D.H., Solomon, P.A., Magliano, K.L., Ziman, S.D, Richards, L.W. (1992) PM_{10} source apportionment in California's San Joaquin Valley. *Atmospheric Environment* 26A, 3335–3354.

Chow, J.C.; Watson, J.G.; Pritchett, L.C.; Pierson, W.R.; Frazier, C.A.; Purcell, R.G. (1993) The DRI Thermal/Otical Reflectance Carbon Analysis System: Description, Evaluation and Applications in U.S. Air Quality Studies; *Atmospheric Environment* 27A, 1185-1201.

Chueinta, W., Hopke, P.K., Paatero, P. (2000) Investigation of sources of atmospheric aerosol at urban and suburban residential areas in Thailand by positive matrix factorization. *Atmospheric Environment* 34, 3319–3329.

Cohen, D.D. (1999) Accelerator based ion beam techniques for trace element aerosol analysis, in *Advance in Environmental, Industrial and Process Control Technologies, Volume 1: Elemental Analysis of Airborne Particles*, S. Landsberger and M. Creatchman(ed.), Gordon and Breach Science Publishers, Tennessee.

CRPAQS, 2003: http://www.arb.ca.gov/airways/crpaqs/overview.htm

Draxler, R.R. and G.D. Hess, 1998. An Overview of the Hysplit_4 Modeling System for Trajectories, Dispersion, and Deposition, *Aust. Met. Mag.* 47, 295-308.

Efron, B., 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics 38, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Fergenson, D. P., Song, X., Ramadan, Z., Allen, J. O., Hughes, L. S., Cass, G. R., Hopke, P. K., Prather, K. A., 2001. Quantification of ATOFMS data by multivariate methods. *Anal. Chem.* 73. 3535-3541.

Finlayson-Pitts, B.J., Pitts, Jr., J.N., 1986. *Atmospheric Chemistry Fundamentals and Experimental Techniques*, Wiley, New York, 1098 pp.

Harrison, R. M., and Perry, R., 1986. *Handbook of Air Pollution Analysis - Second Edition*. Chapman and Hall.

Hopke, P.K., Lamb, R.E., Natusch, D.F.S., 1980. Multielemental Characterization of Urban Roadway Dust, *Environ. Sci. Technol.* 14, 164-172.

Hoskuldsson, A., 1988. PLS regression methods, *J. Chemometrics* 2, 211-228.

Huber, P.J. (1981) *Robust Statistics*, Wiley, New York, 320 pp.

Hughes, L. S., Allen, J. O., Kleeman, M. J., Johnson, R. J., Cass, G. R., Gross, D. S., Gard, E. E., Gälli, M. E, Morrical, B. DFergenson, D. P., Dienes, T., Noble, C. A., Liu, D.-Y., Silva, P. J., Prather, K. A., 1999. Size and Composition Distribution of Atmospheric Particles in Southern California, *Environ. Sci. Technol.* 33, 3506-3515.

Husar, R.B., Tratt, D.M., Schichtel, B.A., Falke, S.R., Li, F., Jaffe, D., Gasso, S., Gill, T., Laulainen, N.S., Lu, F., Reheis, M.C., Chun, Y., Westphal, D., Holben, B.N., Gueymard, C., McKendry, I., Kuring, N., Feldman, G.C., McClain, C., Frouin, R.J., Merrill, J., DuBois, D., Vignola, F., Murayama, T., Nickovic, S., Wilson, W.E., Sassen, K., Sugimoto, N., Malm, W.C. (2001) Asian dust events of April 1998, *J. Geophys. Res.* 106 (D16), 18317-18330.

Juntto, S. and P. Paatero (1994) Analysis of daily precipitation data by positive matrix factorization, *Environmetrics* 5, 127-144.

Kim, E. and P.K. Hopke, 2004. Source apportionment of fine particles at Washington, DC utilizing temperature resolved carbon fractions, *Journal of Air and Waste Management Association,* In Press

Kim, K.-H., Choi, G.-H., Kang, C.-H., Lee, J.-H., Kim, J.Y., Youn, Y.H., Lee, S.R. (2003) The chemical composition of fine and coarse particles in relation with the Asian Dust events. *Atmospheric Environment* 37, 753–765.

Kim, E., P.K. Hopke, and E. Edgerton, 2003a. Source Identification of Atlanta Aerosol by Positive Matrix Factorization, *J. Air Waste Mange. Assoc.* 53, 731-739.

Kim, E., Larson, T.V., Hopke, P.K., Slaughter, C., Sheppard, L.E., Claiborne, C., 2003b. Source Identification of PM2.5 in an Arid Northwest U.S. City by Positive Matrix Factorization, *Atmospheric Res*., 66, 291–305.

Kim, E., P.K. Hopke, and E.S. Edgerton, 2004. Improving source identification of Atlanta aerosol using temperature resolved carbon fractions in Positive Matrix Factorization, *Atmospheric Environment,* In Press.

Koutrakis, P., Spengler, J.D. (1987) Source apportionment of ambient particles in Steubenville, OH using specific rotation factor analysis, *Atmospheric Environment* 21,1511–1519.

Kowalczyk, G.S., Gordon, G.E., Rheingrover, S.W. (1982) Identification of atmospheric particulate sources in Washington, D.C. using chemical element balances, *Environ. Sci. Technol.* 16, 79–90.

Lee, J.H., Y. Yoshida, B.J. Turpin, P.K. Hopke, and R.L. Poirot (2002) Identification of Sources Contributing to the Mid-Atlantic Regional Aerosol, *J. Air Waste Manage. Assoc.* 52, 1186-1205 (2002).

Liu, W., Hopke, P.K., Vancuren, R.A., 2003. Origins of fine aerosol mass in the western United States using positive matrix factorization, *J. Geophys. Res.* 108(D23), Art. No. 4716.

Malm, W.C., Sisler, J.F., Huffman, D., Eldred, R.A., Cahill, T.A., 1994. Spatial and seasonal trends in particle concentration and optical extinction in the United States, *J. Geophys. Res.* 99, 1347-1370, 1994.

Mason, B., 1966. *Principles of Geochemistry, Third Edition*. J. Wiley & Sons, Inc., New York.

Maykut, N.N., J. Lewtas, E. Kim, and T.V. Larson, 2003. Source apportionment of $PM_{2.5}$ at an urban IMPROVE site in Seattle, WA., *Environ. Sci. Technol.* 37, 5135-5142.

Nishikawa, M., Hao, Q., Morita, M., 2000. Preparation and evaluation of certified reference materials for Asian mineral dust. *Global Environ. Res.* 4, 103-113

Paatero P., Tapper, U., 1993. Analysis of different modes of factor analysis as least squares fit problems, *Chemom. Intell. Lab. Syst.* 18, 183-194.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* 5, 111-126.

Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. *Chemom. Intell. Lab. Syst.* 37, 15-35.

Paterson, K.G. , Sagady, J.L., Hooper, D.L., Bertman, S.B., Carroll , M.A., Shepson, P.B., 1999. Analysis of Air Quality Data Using Positive Matrix Factorization, *Environ. Sci. Technol.* 33635-641

Phares, D.J., Rhoads, K.P., Wexler, A.S., Kane, D.B., Johnston, M.V., 2001. Application of the ART-2a algorithm to laser ablation aerosol mass spectrometry of particle standards, *Anal. Chem.* 73. 2338-2344.

Polissar, A.V., P.K. Hopke, W.C. Malm, J.F. Sisler (1996) The Ratio of Aerosol Optical Absorption Coefficients to Sulfur Concentrations, as an Indicator of Smoke from Forest Fires when Sampling in Polar Regions, *Atmospheric Environ.* 301147-1157.

Polissar, A.V., P.K. Hopke, W.C. Malm, J.F. Sisler (1998) Atmospheric Aerosol over Alaska2. Elemental Composition and Sources, *J. Geophys. Res.* 10319,045-19,057.

Polissar, A.V. , P.K. Hopke, and R.L. Poirot (2001) Atmospheric Aerosol over VermontChemical Composition and Sources, *Environ. Sci. Technol.* 354604-4621

Prather, K. A., Nordmeyer, T., Salt, K., 1994. Real-Time Characterization of Individual Aerosol Particles Using Time-of-Flight Mass Spectrometry, *Anal. Chem.* 66, 1403-1407.

Schauer, J.J., Rogge, W.F., Hildemann, L.M., Mazurek, M.A., Cass, G.R., Simoneit, B.R.T., 1996. Source Apportionment of Airborne Particulate Matter Using Organic Compounds as Tracers, *Atmospheric Environ.* 30, 3837-3855.

Schauer, J. J., Kleeman, M. J., Cass, G. R., and Simoneit, B. R. T., 1999. Measurement of Emissions from Air Pollution Sources. 1. C1 through C29 Organic Compounds from Meat Charbroiling, *Environ. Sci. Technol.* 33, 1566-1577.

Schauer, J. J., Kleeman, M. J., Cass, G. R., and Simoneit, B. R. T., 2001. Measurement of Emissions from Air Pollution Sources. 3. C1-C29 Organic Compounds from Fireplace Combustion of Wood, *Environ. Sci. Technol.* 35, 1716-1728.

Schauer, J. J., Kleeman, M. J., Cass, G. R., and Simoneit, B. R. T., 2002. Measurement of Emissions from Air Pollution Sources. 5. C1-C32 Organic Compounds from Gasoline-Powered Motor Vehicles, *Environ. Sci. Technol.* 36, 1169-1180.

Seinfeld, J.H., Pandis, S.N., 1998. *Atmospheric Chemistry and Physics*. Wiley, New York, pp. 533.

Song, X.-H., Polissar, A.V., Hopke, P.K. (2001a) Sources of fine particle composition in the northeastern US, *Atmospheric Environ.* 35, 5277–5286.

Song, X.-H., Faber, N.M., Hopke, P.K., Suess, D.T., Prather, K.A., Schauer, J.J., Cass, G.R., 2001b. Source Apportionment of Gasoline and Diesel by Multivariate Calibration Based on Single Particle Mass Spectral Data, *Anal. Chim. Acta* 446, 329-343.

Song, X., Hopke, P.K., Fergenson, D.P., Prather K.A., 1999.Classification of Single Particles Analyzed by ATOFMS Using an Artificial Neural Network, ART-2A. *Anal. Chem.* 71, 860-865.

Spiro, P.A., Jacob, D.J., Logan, J.A. (1992) Global inventory of sulfur emissions with 1^0x1^0 resolution, *J. Geophys. Res.* 97: 6023-6036.

VanCuren, R.A.; Cahill, T.A. (2002) Asian aerosols in North America: Frequency and concentration of fine dust, J. Geophys. Res. 107(D24), Art. No. 4804.

Vedal, S., (1997) Critical Review: Ambient Particles and Health – Lines that Divide, *J. Air Waste*

Manage. Assoc. 47, 551-581.

Watson, J.G., Chow, J.C., Lowenthal, D.H., Pritchett, L.C., Frazier, C.A. (1994) Differences in the carbon composition of source profiles for diesel and gasoline powered vehicles, Atmospheric Environment, 28: 2493-2505.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58, 109-130.

Xiao, H.; Carmichael, G. R.; Durchenwald, J.; Thornton, D.; Bandy, A., 1997. Long-range transport of SOx and dust in East Asia during the PEM B experiment, J. Geophys. Res.102, 28,589-28,612.

Xie, Y. Hopke, P.K., and Wienke, D., 1994. Airborne Particle Classification With A Combination Of Chemical Composition And Shape Index Utilizing An Adaptive Resonance Artificial Neural Network. Environ. Sci. Technol. 28, 1921-1928.

Xie, Y. L., Hopke, P., 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