Assessing Ozone Networks Using Positive Matrix Factorization

Michael J. Rizzo[a] and Peter A. Scheff[a,b]

[a] Air Monitoring Section, Air and Radiation Division, U.S. Environmental Protection Agency Region 5, Chicago, IL

[b] Environmental and Occupational Health Sciences, University of Illinois at Chicago, School of Public Health, Chicago, IL 60612; pscheff@uic.edu (for correspondence)

Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/ep.10018

In 2001, the United States Environmental Protection Agency (USEPA) began the process of examining the national monitoring networks to assess the contribution of individual monitoring sites in providing useful information to the public and regulatory agencies. One of the first networks to be examined was ozone, with the assessment being initially completed on a national level and then further refined on a regional basis. The goal of the regional analysis was to determine which monitors may be providing redundant information and could, therefore, be removed or relocated to another area in need of additional monitoring data.

One technique which was used in the regional analysis of the ozone network was positive matrix factorization (PMF). This technique is similar to classical factor analysis, which allows for a series of related variables to be grouped into a smaller set of independent factors that represent combinations of the original variables. In addition to grouping the data into factors, this novel approach also provides predicted values of the analysis variable. Comparison of the predicted to the actual values not only gave an indication of how well the model fitted the ozone concentrations, but also aided in the determination of the information value of individual monitors.

Hourly ozone data were polled from the USEPA's national data archive for a total of 24 states for the prime ozone formation months of May through September for 1996 to 2000. Daily maximum 8-hour concentrations were calculated for each site according to the methods contained in 40 CFR Part 50 Appendix H. Because PMF requires a complete data record across all sites for all days analyzed, sites that were missing data were interpolated linearly over time.

The results of the PMF analysis contained 10 factors representing various areas of the country including the Lake Michigan, Atlantic Coast, North Carolina, St. Louis/Indianapolis, Upper New York State, Ohio, Pennsylvania, Kansas/Southeast Missouri/Arkansas, Minnesota/Northwest Wisconsin, and Kentucky/Tennessee areas. Actual to predicted ratios were calculated for each day at each site and the coefficients of variation (CVs) of the individual ratio distributions were utilized as a metric to determine which sites were consistently being predicted well by PMF. Sites with low CVs were interpreted as being well predicted and considered not to be providing ambient ozone information as valuable as that provided by monitors that were poorly predicted by the model. © 2004 American Institute of Chemical Engineers Environ Prog, 23: 110–119, 2004

INTRODUCTION

Ozone is a photochemical pollutant that forms from atmospheric reactions of oxides of nitrogen, reactive hydrocarbons, and UV radiation. The ultimate control strategy for ozone will involve source reductions and will impact all industrial, commercial, mobile, and residential sources of hydrocarbons and nitrogen oxides. In 2001, the U.S. Environmental Protection Agency (USEPA) began the process of examining the national monitoring networks to assess the contribution of individual monitoring sites in providing useful information to the public and regulatory agencies. This process was necessary to begin evaluating state resource and program commitments in order to assess the capability of assuming other national monitoring programs such as air toxics. As a part of the assessment, Region 5 conducted a variety of spatial analyses on the existing ozone data to determine whether any sites were reporting redundant information. These analyses included the use of positive matrix factorization (PMF) [1, 2]. The PMF technique has been used extensively for source apportionment work [3–10]. PMF has also recently been used to assess the spatial distribution of $PM_{2.5}$ in the

© 2004 American Institute of Chemical Engineers

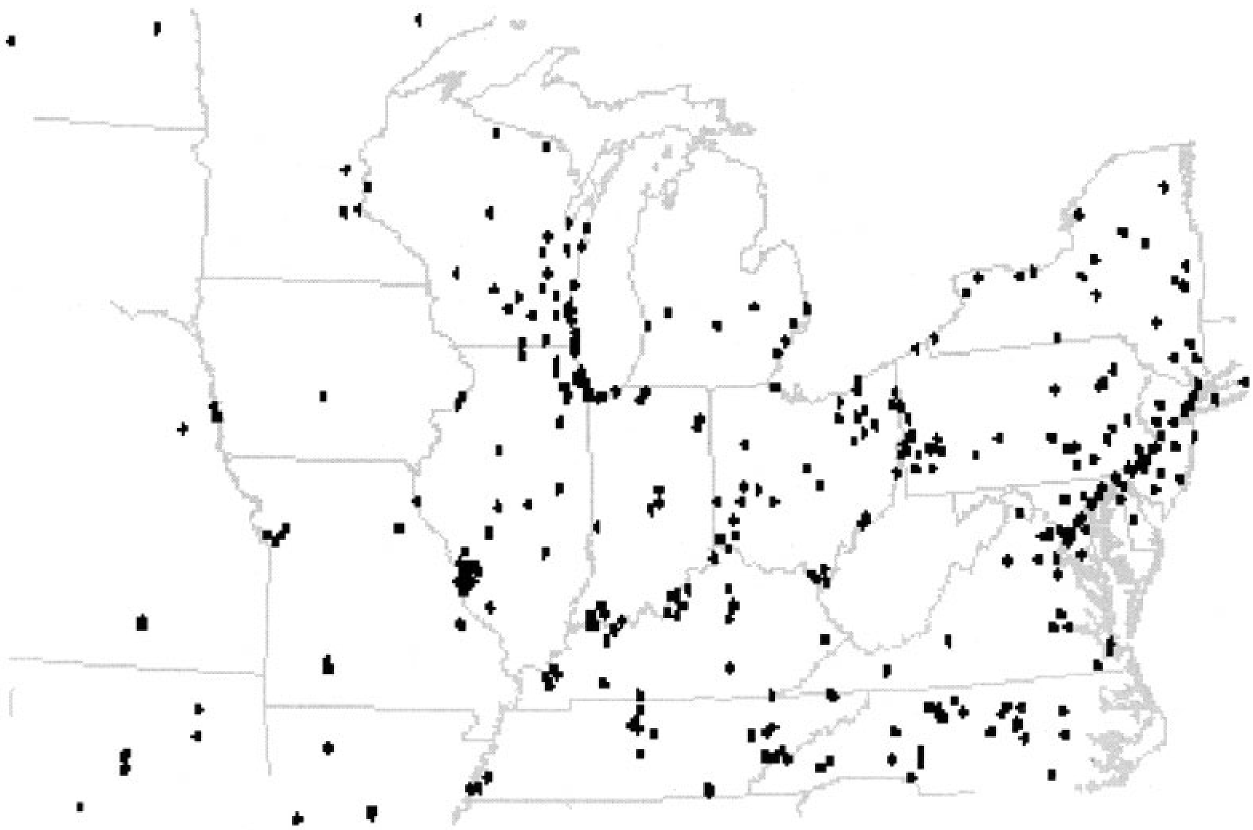

Figure 1. Ozone monitoring sites supplying data used in PMF analysis.

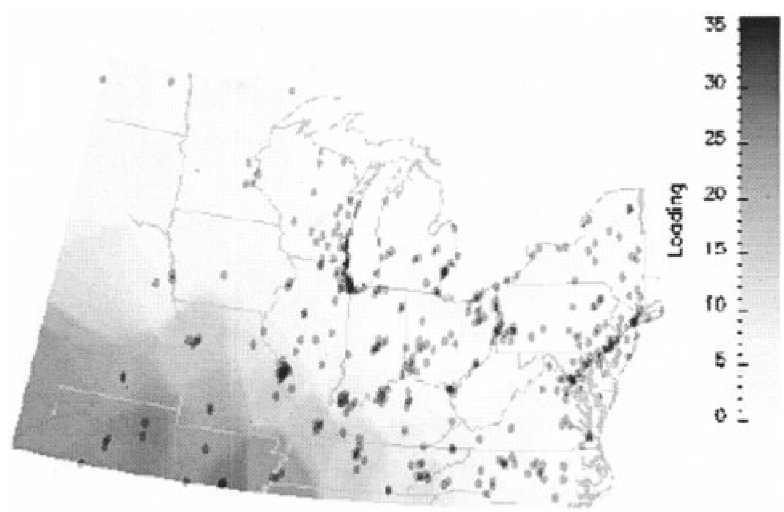

Figure 2. MO, AR, KS, and OK factor.

eastern United States [11]. This is an example of using PMF to group monitors with similar ozone concentrations.

METHODOLOGY

PMF is a technique that solves a linear model to determine the contribution a particular factor makes to an individual observation. The primary use of PMF to date has been source apportionment. This is the first attempt to use PMF in an ozone monitoring network review. The general equation displaying the relationship between the concentrations at individual ozone sites and the respective factors can be written as

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

where $\mathbf{X}$ is the $n \times m$ matrix of ozone concentrations

Figure 3. Central TN/KY factor.

Figure 4. Upper NY factor.

Figure 5. East Coast factor.

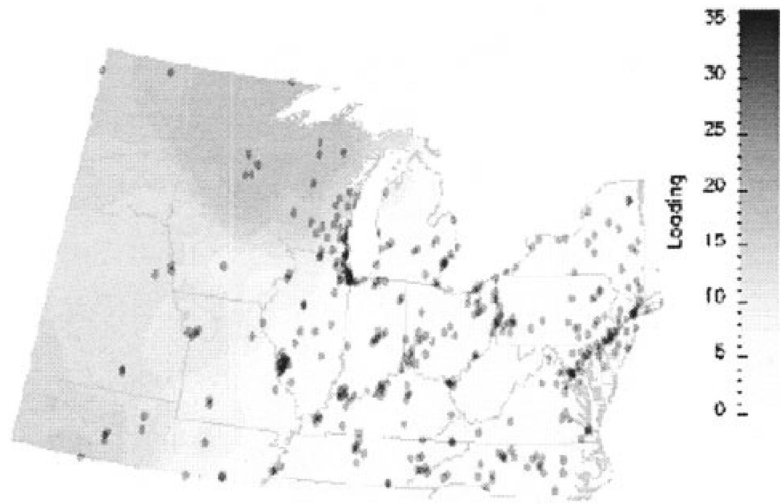

Figure 6. Upper MN/WI factor.

Figure 7. Ohio area factor.

by day and site; **G** is the $n \times p$ matrix of the factor strengths for each observation at time t; **F** is the $p \times m$ matrix of the individual loadings for each factor; and **E** is the $n \times m$ error matrix.

In this case, n is the total number of observations, m is the total number of sites, and p is the number of factors. Values for the **G** and **F** matrices are constrained positive, which allows the contribution from each factor for each time t to be physically realistic (i.e., no factor contributes negatively to a station's ozone concentration).

Equation 1 is solved using an iterative least squares approach with Q representing the weighted sum of the squared errors,

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{E_{ij}^2}{\sigma_{ij}^2}, \tag{2}$$

where Q is the sum of squared error divided by the uncertainty for each observation; E is the residual for the observation made at time i and site j; and σ is the uncertainty associated with the measurement at time i and site j.

For the purposes of this analysis, the uncertainty for each observation was determined using the following method. If the measurement was not missing, 10% of the measured value was used unless the resulting uncertainty was less than a critical value, at which point the critical value was used as the uncertainty. If the measurement was missing, 20% of the interpolated value was used unless the resulting uncertainty was less than 2 times the critical value, at which point two times the critical value was used. The critical value chosen for this analysis was the same limit as used for the rounding convention for ozone measurements for the National Ambient Air Quality Standard (NAAQS). For reporting ozone concentrations in parts per million, the digit in the third decimal place is rounded up if the value is greater than 5 or rounded down if less than 5

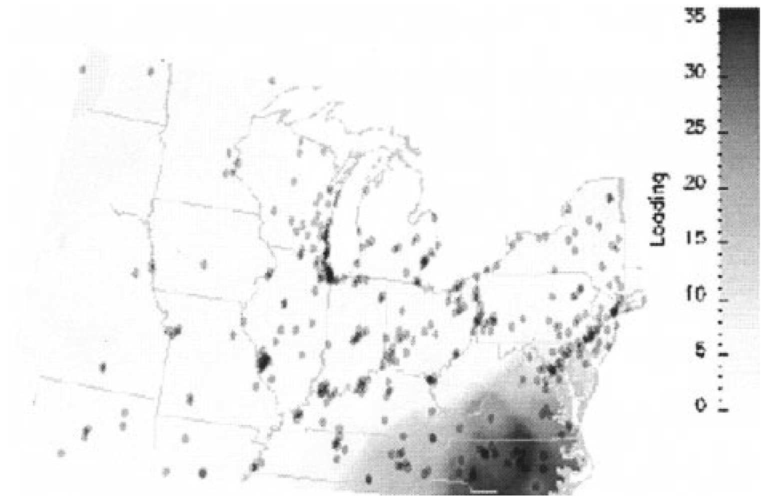

Figure 8. North Carolina factor.

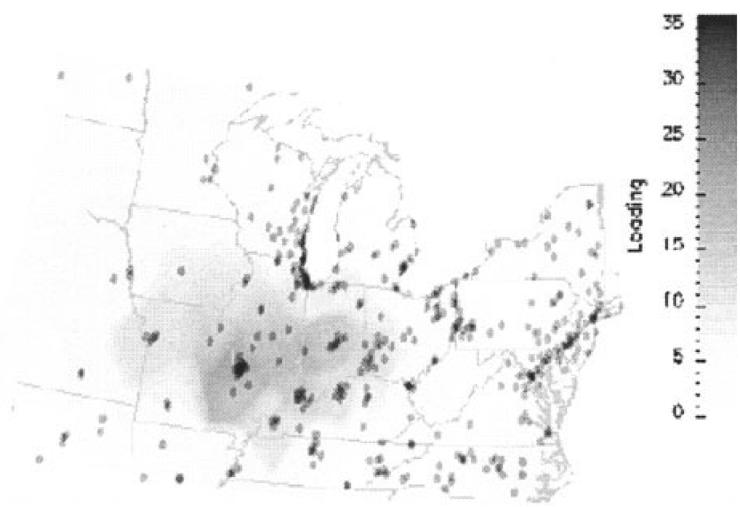

Figure 9. St. Louis/Indianapolis factor.

as listed in 40 CFR Part 50 Appendix I. This essentially says that the ozone value is to be rounded to the nearest 10 ppb. For this analysis, if 10% of the measured value was less than 5 ppb, 5 ppb was used as the uncertainty. If the value was missing and 20% of the interpolated value was less than 10 ppb (2 × 5 ppb), then 10 ppb was used as the uncertainty for that particular observation.

In order to determine the optimum number of factors to represent our domain, the FACTOR procedure in SAS was used. The value for the percentage variability explained by each successive factor decreases as the number of factors explaining the independent variable increases. Thus, each successive factor accounts for less of the variability in the analysis variable. The results of this analysis suggested that a total of 10 factors explained a total of 78% of the variability in the ozone measurements across the domain. Therefore, PMF was instructed to confine its calculations to a 10-factor solution.

The predicted values from the model were compared to the actual observed ozone concentrations to determine how well the model fitted the data. Ratios of the actual to predicted concentrations were calculated for each day. Ratios close to 1 meant that the model did a very good job of predicting the ozone concentration at the respective site, while values much less or much greater than 1 signified that the concentration at the site was not predicted very well. Because the average ratios for most sites were very close to 1, the coefficient of variation of the ratio at each site was calculated and normalized so the resulting values would be within 0 and 1. The resulting value was used as a "uniqueness" criterion to categorize each monitor. Sites with a uniqueness value close to 0 were sites where the day-to-day variation in the observed to predicted ratios was small and not considered to contribute important information to the characterization of ozone concentrations. These sites could be possibly removed or relocated to another area, while sites with higher uniqueness values were considered to contribute information that PMF could not account for and, therefore, needed to be retained.

Figure 10. Pennsylvania area factor.

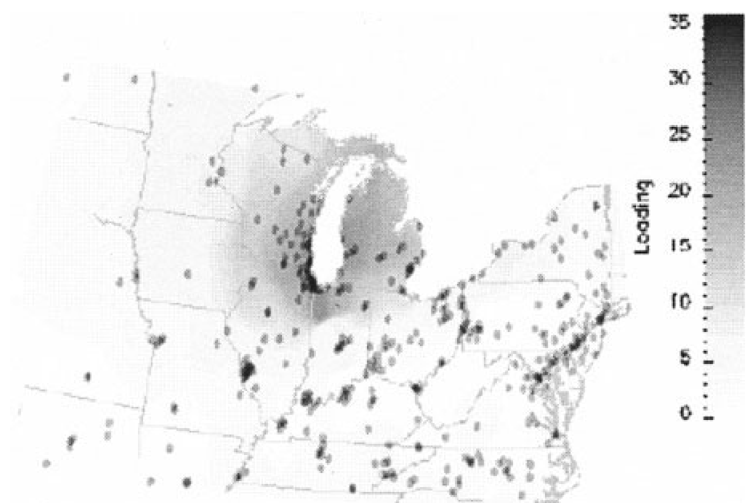

Figure 11. Lake Michigan area factor.

It is possible to further investigate the relationship between the individual factors by analyzing the scores found in the **G** matrix. The **G** matrix represents the predicted concentrations for each time period for each factor. A time series analysis was conducted examining the cross correlations of the *G* scores between the factors [12]. In this case, the cross correlation is the temporal or time-related relationship between the factor's *G* scores.

Data were collected from all sites shown in Figure 1 for 1996 through 2000 from the Aerometric Information Retrieval System (AIRS). Only those sites that were consistently measuring ozone between May and September of 1996 through 2000 were included in the analysis. Since the emphasis was placed on assessing the Region 5 ozone network, which is contained within the states of Illinois, Indiana, Ohio, Wisconsin, Minnesota, and Michigan, an effort was made to also include those sates which surround the Region 5 area. This amounted to a total of 375 sites and 765 daily observations across the 24-state domain. The PMF model has been demonstrated to have an "edge" effect where monitors on the edge of the domain are not predicted very well. In order to account for this effect, the domain was extended as far east, west, north, and south of Region 5 as possible.

Hourly 8-hour averaged ozone concentrations were calculated for each site according to the methodology in 40 CFR Part 50, Appendix H. The maximum 8 hour average for each day was used to represent a 24-hour period. Because PMF requires that every site has a complete set of observations, missing data were linearly interpolated temporally. Without the interpolation, whole observations would have to be deleted or the monitor removed, thus reducing either the amount of data or the number of sites which could be included in the analysis.

RESULTS AND DISCUSSION

Figures 2 through 11 show the factor loadings, which are the values contained in the **F** matrix. Figure 12 shows the uniqueness values normalized between 0

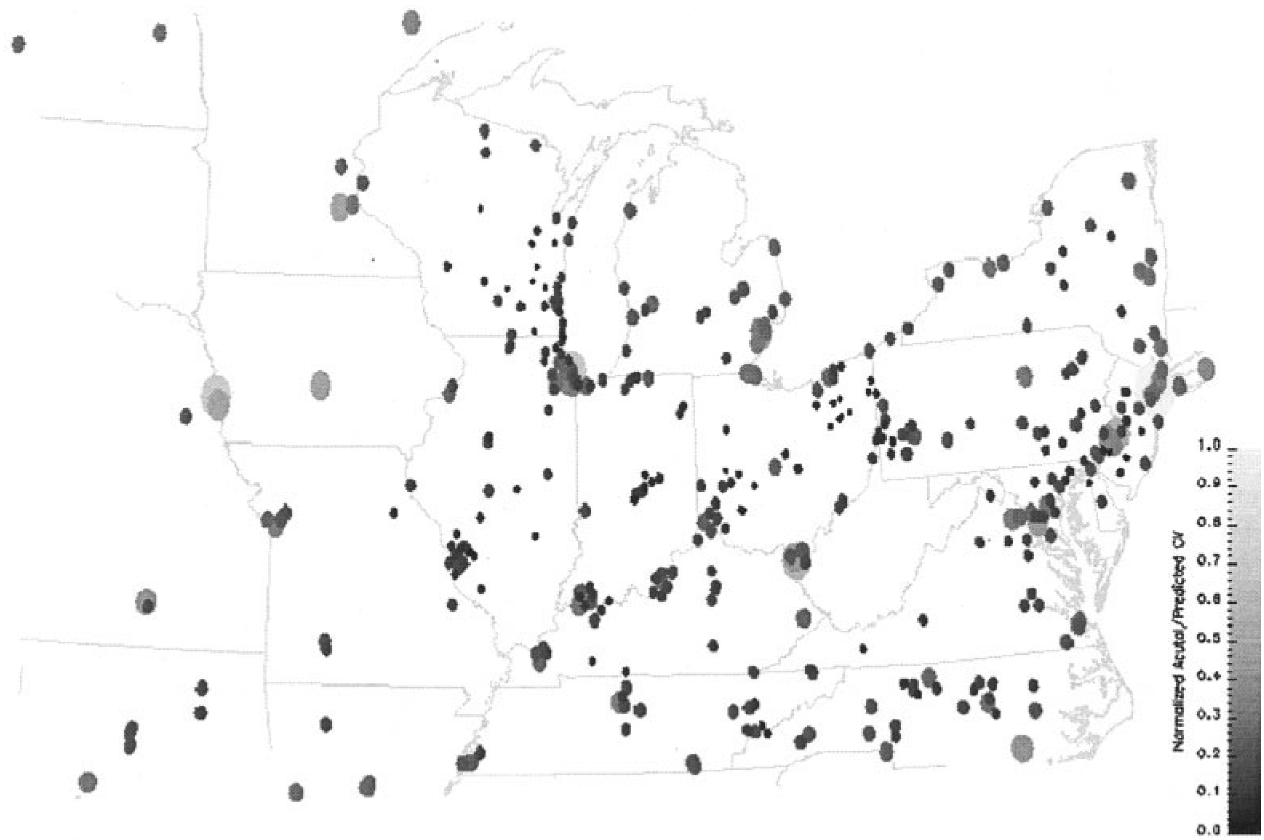

Figure 12. Uniqueness metric for 10-factor solution (summers from 1996 to 2000).

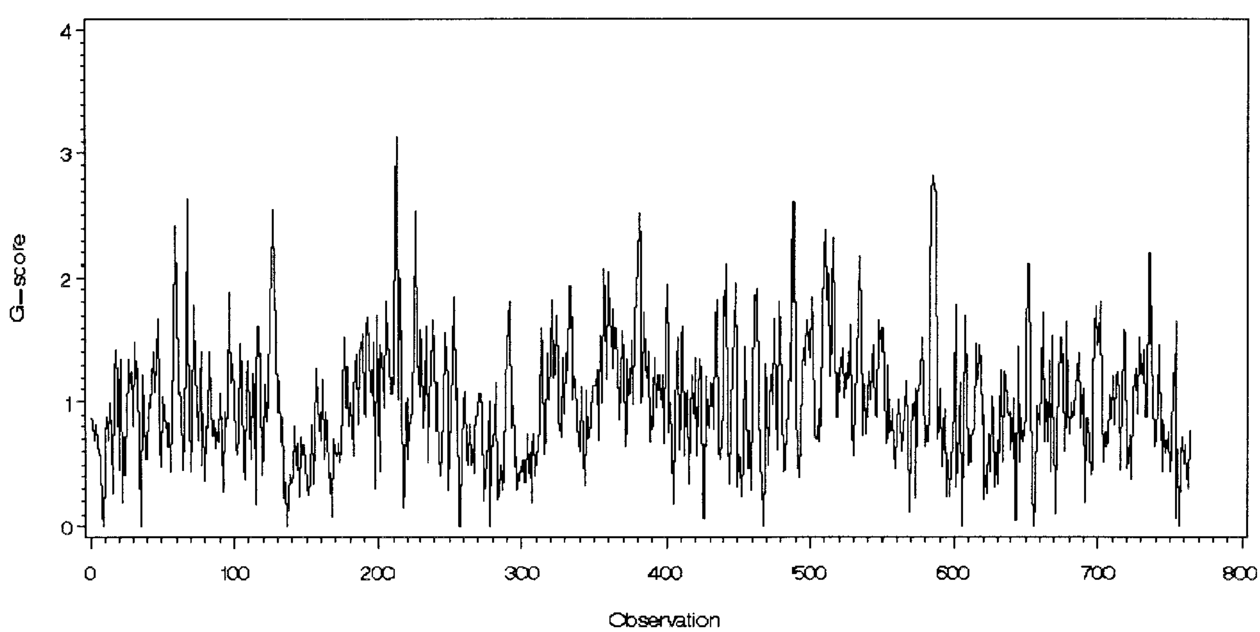

Figure 13. *G* score time series for the Lake Michigan factor.

and 1 for each monitor used in the analysis. Figure 13 shows an example of the time series of the individual factor scores created from the **G** matrix. Table 1 shows the results of the time series analysis on the factor scores from **G** matrix.

Figures 2 through 11 suggest strong regional relationships between monitoring locations. This is most likely caused by similar meteorological patterns in those areas. The individual factor loadings for PMF do not carry the same connotation as they do for tradi-

Table 1. Summary of cross-correlation results for 10-factor solution.

	Independent variable									
Dependent variable	**Lake Michigan**	**MO, AR, KS, OK**	**Central TN and KY**	**Upper NY**	**East Coast**	**Upper MN and WI**	**Ohio area**	**North Carolina**	**St. Louis/ Indianapolis**	**Pennsylvania**
Lake Michigan						1 day ago: 0.407			Same day: 0.195 1 day ago: 0.197	
MO, AR, KS, OK						Same day: −0.359				
Central TN and KY		Same day: −0.163 1 day ago: 0.213				Same day: −0.133			Same day: 0.149 1 day ago: 0.267	
Upper NY	1 day ago: 0.328	1 day ago: −0.077 2 days ago: −0.127	Same day: −0.149 1 day ago: 0.087 2 days ago: 0.127			2 days ago: 0.207	1 day ago: 0.288	Same day: −0.271	Same day: −0.232	
East Coast	1 day ago: 0.209		1 day ago: 0.126			Same day: −0.270 2 days ago: 0.127	Same day: 0.166 1 day ago: 0.277	1 day in future: 0.170 Same day: 0.101 1 day ago: 0.124		Same day: .248 1 day ago: .223
Upper MN and WI		Same day: −0.356	Same day: −0.142		Same day: −0.278					
Ohio area	Same day: 0.140 1 day ago: 0.334		Same day: 0.112 1 day ago: 0.186			Same day: −0.141 1 day ago: 0.105 2 days ago: 0.181			1 day ago: 0.318	
North Carolina			1 day ago: 0.276		1 day in future: 0.131 Same day: 0.111 1 day ago: 0.177	Same day: −0.178 1 day ago: −0.169		Same day: −0.273	Same day: −0.076 1 day ago: 0.100 2 days ago: 0.086	1 day in future: 0.109 1 day ago: 0.216
St. Louis/ Indianapolis		1 day ago: 0.188		Same day: −0.226		1 day in future: −0.137 1 day ago: 0.174				
Pennsylvania	1 day ago: 0.238		Same day: 0.179 1 day ago: 0.207			2 days ago: 0.134	1 day ago: 0.231		1 day ago: 0.201	

tional factor analysis. In traditional factor analysis, the loadings usually represent the correlation between the factor and the individual variables. For PMF, the loadings represent the weight each factor has on predicting the individual variable, which in this case represents the ozone concentrations at a particular monitoring location.

Figures 2 to 11 show the spatial uniformity of the 10 PMF factors. In these figures, high factor loadings (darker color) on a PMF factor represent monitors that have high underlying correlations. As can be seen from Figures 2 through 11, monitors within close proximity to each other usually load high on the same factor, thereby suggesting a relationship between those sites. For some sites, particularly on the western edge of the domain where the monitor density is low, PMF has a greater difficulty in determining which sites load high in a particular factor. This can be seen in the factor that covers Minnesota and western Wisconsin. There is a large overlap between the sites for this factor and those in the factor which includes the northern part of Oklahoma.

Figure 12 shows the normalized CV for each monitor. In this plot, a value of 0 represents no variation and a value of 1.0 represents the maximum observed day-to-day variation in the observed to predicted ratio. Figure 12 shows that the results are very complex. For example, the uniqueness results in Figure 12 show that PMF is able to predict ozone concentrations fairly well where the network is dense. This result is seen in areas such as Cleveland, Indianapolis, the Chicago metropolitan area, and the St. Louis metropolitan area. This suggests that monitors within these densely packed areas are related fairly well to one another and that it may be possible to reduce the number of monitoring sites within the area and still obtain the same amount of information that was originally present with the full complement of sites. Some caution, however, must be used when interpreting the uniqueness metric for the monitors near the edge of the domain. As mentioned earlier, the Western edge of the analysis domain has a sparse network density; PMF has a difficult time determining in which factor those sites have the greatest influence. This can also be seen in the uniqueness metric. PMF had a greater difficulty predicting the concentrations for these sites compared to sites located more in the center of the domain or sites that were a part of a large cluster of monitoring locations.

The results of the time series correlation analysis on the *G* scores show some interesting characteristics. The cross correlation table is not symmetric. The relationships between factors when taking into account the dimension of time are directional. For example, today's ozone in the Upper New York factor is dependent on the ozone concentrations from the day before from the Lake Michigan area. If Lake Michigan becomes the response (dependent) variable and Upper New York assumes the role of the predictor variable, the cross correlation becomes physically uninterruptible (non-significant). In that case, the significant cross correlation would have shown that tomorrow's ozone value from Upper New York affects today's ozone in the Lake Michigan area, which is physically impossible.

The cross correlations demonstrate transport from one particular location (factor) to another. Table 1 suggests a time period for air masses to move from one area to another. (Note that only statistically significant correlations are included in Table 1.). For example, the ozone concentration from the Upper Minnesota and Wisconsin area (independent variable) from 2 days ago is correlated to today's ozone concentration in Upper New York, suggesting that it takes approximately 2 days for the air mass from the Upper Minnesota and Wisconsin region to reach the Upper New York area.

CONCLUSIONS

1. PMF partitions the ozone monitoring sites according to geographic region and implies a relationship between the ozone concentrations for sites that load high in an individual factor.

2. The uniqueness metric provides a means for determining monitors that may be redundant and could be either discontinued or relocated to areas lacking proper monitoring coverage. If a group of monitors has high loadings on a common factor and uniqueness values close to 0, then the measurements must be highly correlated with an observed-to-predicted ratio that is always close to 1. These would be the monitors that are good candidates for removal. However, we understand that there may be political reasons to leave a monitor in the network even if it provides no new information.

3. Time series analysis of the *G* scores gives an indication of ozone transport from one region of the domain to another. The lag time associated with the relationship between one area and another gives an indication of the length of time it takes for an air mass to move from one region to another as well as the path the air mass may take. This is representative of the meteorology of the area. The cross correlation does not need to have a high correlation to be important. The fact that meteorology changes day to day insures that the cross correlations will not be very high. That is, not every high occurrence of ozone in one location will be transported to the same area on the next day. This variability insures that cross correlations will not be very high.

DISCLAIMER

The views expressed in this paper are those of the authors and not necessarily those of the USEPA.

LITERATURE CITED

1. 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.
2. Paatero, P. (1997). least squares formulation of robust non-negative factor analysis, Chemometrics and Intelligent Laboratory Systems, 37, 23–25.
3. Anderson, M.J., Miller, S.L., & Milford, J.B. (2001). Source apportionment of exposure to toxic volatile organic compounds using positive matrix factorization, Journal of Exposure Analysis and Environmental Epidemiology, 11(4), 295–307.

4. Ramadan, Z., Song, X.H., & Hopke, P.K. (2000). Identification of sources of phoenix aerosol by positive matrix factorization, Journal of the Air & Waste Management Association, 50(8), 1308–1320.
5. 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(20), 3319–3329.
6. Lee, E., Chan, C.K., & Paatero, P. (1999). Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, Atmospheric Environment, 33(19), 3201–3212.
7. Polissar, A.V., Hopke, P.K., Paatero, P., Kaufmann, Y.J., Hall, D.K., Bodhaine, B.A., Dutton, E.G., & Harris, J.M. (1999). The aerosol at Barrow, Alaska: long-term trends and source locations, Atmospheric Environment, 33(16), 2441–2458.
8. Xie, Y.-L., Hopke, P.K., Paatero, P., Barrie, L.A., & Li, S.-M. (1999). Identification of source nature and seasonal variations of arctic aerosol by the multilinear engine, Atmospheric Environment, 33(16), 2549–2562.
9. Prendes, P., Andrade, J.M., López-Mahía, P., & Prada, D. (1999). Source apportionment of inorganic ions in airborne urban particles from Coruña city (N. W. of Spain) using positive matrix factorization, Talanta, 49(1), 165–178.
10. Anttila, P., Paatero, P., Tapper, U., & Järvinen, O. (1995). Source identification of bulk wet deposition in Finland by positive matrix factorization, Atmospheric Environment, 29(14), 1705–1718.
11. Paatero, P., Hopke, P., Hoppenstock, J., & Eberly, S. (2003). Advanced factor analysis of spatial distributions of $PM_{2.5}$ in the eastern United States, Environmental Science and Technology, 37(11), 2460–2476.
12. Rizzo, M., Scheff, P., & Ramakrishnan, V. (2002). Defining the photochemical contribution to particulate matter in urban areas using time-series analysis, Journal of the Air & Waste Management Association, 52, 593–605.