Atmospheric Environment Vol. 21, No. 9, pp. 2005–2013, 1987.
Printed in Great Britain.

0004-6981/87 $3.00+0.00

APPLICATION OF THE FACTOR-ANALYSIS RECEPTOR MODEL TO SIMULATED URBAN- AND REGIONAL-SCALE DATA SETS

DOUGLAS H. LOWENTHAL and KENNETH A. RAHN

Center for Atmospheric Chemistry Studies, Graduate School of Oceanography, University of Rhode Island, Narragansett, RI 02882-1197, U.S.A.

(*First received* 21 *August* 1986 *and in final form* 4 *March* 1987)

Abstract—Factor analysis has been used extensively to model the sources of ambient aerosol. In this study, simple urban- and regional-scale simulations showed that factor analysis may not always produce reliable results. The accuracy of apportionments of total mass depended on the rotation scheme used to transform the factors. Varimax-rotated solutions were generally independent of the degree of random error in the data, but were sensitive to collinearities in profiles, correlations of source strengths, and magnitudes of source strengths. Target-transformation factor analysis was more successful than varimax rotation when targets were similar to true profiles.

In simple regional simulations, midwestern and northeastern sources were resolved qualitatively by varimax-rotation and quantitatively by target-transformation factor analysis. Signatures determined from principal-component analysis of ambient data in the northeastern U.S. and the Arctic resembled those determined independently.

Key word index: Factor, rotation, varimax, target, principal component, correlation, loadings, source strength, source profiles, receptor modeling.

INTRODUCTION

The fundamental mass balance for atmospheric aerosols is:

$$C_{ti} = \sum_{j=1}^{p} S_{tj}A_{ji} + e_{ti} \qquad (1)$$

where C_{ti}, the concentration of the i^{th} species in the t^{th} receptor sample, is the sum of the contributions of the i^{th} species from p sources, S_{tj} is the true mass per unit volume of air contributed by the j^{th} source to the t^{th} sample, A_{ji} is the fractional abundance of the i^{th} species in the j^{th} source profile, and e_{ti} is the residual for the i^{th} species in the t^{th} sample. Ideally, all possible sources are included and their profiles are correct.

The two basic statistical methods used to solve this equation, chemical mass balance and factor analysis, were reviewed by Henry *et al.* (1984). Chemical mass balance (CMB) uses measured signatures and ambient samples to solve for the source strengths by least-squares regression. Factor analysis involves a class of multivariate techniques which derive both the strengths and compositions of sources from ambient samples. Factor analysis was earlier described in depth by Harman (1967), and has subsequently been applied frequently to determining sources of aerosols (Hopke *et al.*, 1976; Heidam 1982, 1984, for example).

In classical, or '*R*-mode', analysis, the matrix of correlations or covariances of C_{ti} over the samples is decomposed into a smaller number of underlying sources of variation. In common-factor analysis, $S_{tj}A_{ji}$ represents only common, or shared, variation, while e_{ti} represents unique, or random, variation. Receptor modelers frequently use a principal-component model, where both $S_{tj}A_{ji}$ and e_{ti} include common and random variation. In a successful factor analysis, however, $S_{tj}A_{ji}$ should contain more common than random variance. The principal components are generally rotated to attempt to give them an environmentally plausible interpretation. For example, Thurston and Spengler (1985) used varimax rotation of the principal components of the correlation matrix while Hwang *et al.* (1984) rotated the principal components of the covariance matrix about the origin to targets by a modified least-squares procedure. Varimax rotation attempts to achieve a 'simple structure' of independent factors. In environmental terms, a group of sources which emit unique suites of chemical species might be modeled as a simple structure. Target-transformation factor analysis, or TTFA (Alpert and Hopke, 1981), is based on an oblique rotation to prospective profiles, or targets.

Henry (1985) discussed theoretical limitations of factor analysis for apportioning sources. For example: an infinite number of rotations will explain the data equally well; the constraint of independence imposed by varimax rotation may be unrealistic because of similarity between source profiles or correlation between strengths of different sources; target transformation may be quite sensitive to differences between target profiles and true profiles. Thus, neither rotation automatically provides an environmentally realistic solution.

In practice, factor analysis has also been of limited utility. Using TTFA, Alpert and Hopke (1981) could not distinguish between the similar profiles of coal and soil. Thurston and Spengler (1985), with *R*-mode analysis followed by varimax rotation, obtained a solution which was qualitatively plausible but which apportioned nearly 50% of marker elements such as V and Pb to sources other than oil and gasoline, respectively. By contrast, Alpert and Hopke (1981) and Kowalczyk *et al.* (1978) apportioned more than 90% of the marker elements to the correct sources, with TTFA and CMB, respectively.

To date, factor analysis has been used primarily for urban problems. Now that receptor modeling is being applied to regional-scale problems (Rahn and Lowenthal, 1984, 1985), the utility of regional-scale factor analysis should be investigated as well.

This study had two principal objectives. The first was to determine how sensitive varimax rotation and TTFA are to parameters such as: (1) random variation in source profiles and ambient measurements, (2) collinearity in source profiles, (3) magnitudes of source strengths, and (4) correlation of source strengths. The second objective was to investigate whether factor analysis can be used (1) to resolve regional sources in simulated aerosol data, and (2) to determine regional signatures from ambient samples.

METHODS

Synthetic data sets were generated from urban-scale source profiles used by the Quail Roost II workshop (Currie *et al.*, 1984) and from regional-scale signatures (Rahn and Lowenthal, 1984, 1985) by Monte Carlo methods outlined in Watson *et al.* (1984) and Currie *et al.* (1984). The 'true' source strengths [the S_{ij}'s of Equation (1)] were generated by randomly perturbing average values with a coefficient of variation of 50% and restricting random-normal deviates to absolute values of less than two. To test the important assumption of all factor models that the strengths of different sources are uncorrelated in time, some of the simulations included correlated source strengths. Synthetic urban data sets were composed of 100 samples, each with 20 or 21 chemical species from three sources; synthetic regional data sets included 100 samples with seven species from two sources. Random error in sources and samples was introduced to the data as follows:

$$C_{ii} = \sum_{j=1}^{p} [(A_{ji} + \varepsilon_{ji}\sigma_{A_{ji}})S_{ij}] + \varepsilon_{ii}\sigma_{C_{ii}} \quad (2)$$

where $\sigma_{A_{ji}}$ is the measurement uncertainty of the i^{th} species in the j^{th} signature, $\sigma_{C_{ii}}$ is the measurement uncertainty of the i^{th} species in the ι^{th} sample, and ε_{ji} and ε_{ii} are standardized random-normal deviates.

A computer program was written to generate the correlation matrix from the simulated samples, calculate the principal components, rotate the axes, and estimate source strengths. The number of rotated components was always the same as the number of sources used to create the data. The simple method of Lowenthal and Rahn (1985) was used to transform factor loadings into concentration profiles: rearranging Equation (20) of Henry *et al.* (1984) and using the notation of Equation (1) shows that:

$$A_{ji} = g_{ji} \times \left[\frac{\overline{X_i^2} - \overline{X}_i^2}{\overline{S_j^2} - \overline{S}_j^2}\right]^{1/2} \quad (3)$$

where g_{ji} is the factor loading (correlation of the i^{th} species with the j^{th} factor), the numerator in brackets is the variance of the i^{th} species multiplied by the number of observations, and the denominator is a constant related to the source strengths of the j^{th} source. Thus, one need only multiply an element's loading by its standard deviation over the samples to produce a relative concentration in a source profile. Because total mass can also be included in the factor analysis, whether it is a dependent or independent variable and if including it does not unduly affect the other loadings (Heidam, 1981), the fractional abundance of the i^{th} species in a factor is the ratio of the product of its standard deviation and factor loading to the product of the standard deviation and factor loading of aerosol mass.

The varimax-rotated loadings were transformed (destandardized) into source profiles A_{ji} and the source strengths S_{ij} were estimated by unconstrained, weighted least-squares. The solution for the source strengths is given in matrix notation as:

$$S = (A'WA)^{-1}A'WC \quad (4)$$

where W is a diagonal matrix of weights. The source profiles derived from varimax rotation were then rotated with unconstrained weighted least-squares using true and randomly perturbed source profiles as targets. The TTFA solution for the predicted profiles is given in matrix notation as:

$$A^* = A(A'WA)^{-1}A'WB \quad (5)$$

where A^* is the matrix of predicted profiles, W is the matrix of weights, and B is the target matrix. Source strengths were estimated from the predicted profiles by substituting A^* for A in Equation (4). The weights were the inverses of the sample variances when no random error was introduced to the data, or the inverses of the mean-squared errors of the ambient data otherwise.

The derived source strengths were compared to the true values by means of the average absolute percent error (*APE*):

$$APE_j = \sum_{\iota=1}^{100} \left[\frac{|S_{ij} - \mathrm{True}_{ij}|}{\mathrm{True}_{ij}}\right] \times 100 \quad (6)$$

where True_{ij} is the true source strength for the j^{th} source in the ι^{th} sample.

RESULTS AND DISCUSSION

Simple urban case

Case 1. The first simulation used the Incinerator, Basalt and Road profiles (Currie *et al.*, 1984), chosen because they were found by Lowenthal *et al.* (1987) not to be seriously collinear, according to the diagnostic procedure of Belsley *et al.* (1980). The average strength of each source was made roughly 4000 ng m^{-3}. Measurement uncertainties of signatures and ambient samples were set to 10% and 10%, then to 20% and 10%, respectively. For TTFA, true source profiles and profiles generated by perturbing the latter randomly (element by element) by 20% and 50% were used.

Table 1a gives the average true and predicted source strengths and the *APE*s for each source and each rotation scheme as well as their average values over the three sources. The varimax and target-transformation rotations will be henceforth referred to as 'varimax' and 'target (%)', where the value in parentheses represents the degree of perturbation in the targets.

The *APE* for Basalt shows that varimax predicted the source strengths to within roughly a factor of two