Multivariate Receptor Modeling for Air Quality Data in Space and/or Time

Eun Sug Park[1], Peter Guttorp[2], and Ho Kim[3]

[1]Texas Transportation Institute, 3135 TAMU, College Station, TX 77843-3135, USA. E-mail: e-park@ttimail.tamu.edu. [2]National Research Center for Statistics and the Environment, University of Washington, Seattle, WA 98195, USA. [3]Biostatistics &Epidemiology, School of Public Health, Seoul National University, 28 Yunkeon-Dong, Chongro-Gu, Seoul 100-799, Korea

Introduction

Previous studies on Seoul air pollution focused on finding associations between daily mortality and ambient concentrations of pollutants such as NO_2, SO_2, CO, O_3, and PM_{10} (see Hong et al., 1999; Lee et al., 2000). There have not been any attempts for identifying pollution sources based on these air pollution data. Identifying important pollution sources that contribute to ambient concentrations of pollutants is essential for developing an effective air quality management plan. Multivariate receptor modeling aims to achieve this goal by unfolding the air pollution data into components associated with different sources. Traditionally, a multivariate receptor model has been applied to the measurements on multiple chemical species (say, p chemical species) collected at a receptor. See Henry (1991) for a comprehensive review of conventional multivariate receptor models such as principal component analysis, exploratory factor analysis, target transformation factor analysis, and self-modeling curve resolution. Assuming that there are q major pollution sources, a basic physical model in multivariate receptor modeling can be written as

$$y_t = \alpha_t P + \varepsilon_t \qquad (1)$$

where $y_t = (y_{t1} \quad \text{L} \quad y_{tp})$ is the measured concentration of p chemical species at time t, $\alpha_t = (\alpha_{t1} \quad \text{L} \quad \alpha_{tq})$ is a vector of source contributions at time t, $P = \left[P_1' \quad \Lambda \quad P_q'\right]'$, $k = 1, \Lambda, q$, $P_k = (p_{k1} \quad \Lambda \quad p_{kp})$ is the source composition profile (source fingerprint) for the kth source, and $\varepsilon_t = (\varepsilon_{t1} \quad \text{L} \quad \varepsilon_{tp})$ is a vector of measurement errors at time t. This might be viewed as a factor analysis model in that both α_t and P are unobservable, and in this case, P can be related to a loading matrix and α_t can be related to factor scores. The usual challenges in factor analysis models, unknown number of factors (q) and/or non-identifiability of parameters, are also encountered in multivariate receptor models. These issues were fully addressed in a number of recent studies, Park, Oh, and Guttorp (2001) and Park, Spiegelman, and Henry (2001).

Air pollution data are often obtained as concentrations of a single species (e.g., PM_{10}) measured from multiple monitoring sites. This can be considered as a multivariate dataset for which different monitoring sites are treated as different variables. Let p be the number of monitoring sites. Assuming that there are a few underlying spatial patterns of sources and meteorological conditions are fairly stable over the monitoring sites, the above basic model can still be applied with new interpretation for P (Park, Oh, and Guttorp 2001; Park, Spiegelman, and Henry 2001). Here, each row of P corresponds to a spatial profile for a source (source type), which represents the relative amounts of the pollutant conveyed to the p monitoring sites from the source. The closer the monitoring site is to the source, the higher the relative amount in the spatial profile is. Thus, spatial profiles can also play a role of source fingerprints in the sense that we can locate the sources (source types) based on them.

The gaseous species, NO_2, SO_2, CO, and O_3, have been monitored for several years (since 1992) in Seoul, and all have conformed to the standard. On the other hand, PM_{10} had not been regulated until 1997. The ambient concentration of PM_{10} is generally high (close to or above the standard) and there

does not seem to be any decreasing trend. Also, a number of recent studies including Hong et al. (1999) suggested that PM_{10} effect on all-cause mortality was significant. It has been conjectured that the major air pollution sources in Seoul are automobile exhaust and domestic heating. There are also some light industry areas in Seoul. It is of great interest to identify major sources of PM_{10} based on the ambient concentration data for an effective air quality management in Seoul. The primary goal of this paper is to show how multivariate receptor models can be used to understand relationships between source emissions and receptor concentrations on air pollutants, by locating the major PM_{10} source areas in Seoul.

Data

The dataset used in this paper consists of 24 hour average PM_{10} concentrations measured from 17 monitoring stations in Seoul, Korea, for 1/1/99~12/31/99. The analysis was confined to 1999 data to optimize the number of sites and the length of the data record. Monthly wind direction data for Seoul in 1999 was also obtained, which shows a different pattern for each season. Since spatial profiles of PM_{10} sources will clearly depend on wind pattern, we analyze the data separately for Winter period (1~3 and 10~12) and Summer period (4~9). Classifying the observations according to each period and deleting any rows (observations) with missing values leads to the final datasets of 91 observations for Winter and 68 observations for Summer on 17 variables (monitoring stations), respectively. At all monitoring stations, the observed PM_{10} level is much higher during Winter than during Summer in terms of both average concentration and variability.

Method

Note that in model (1), the number of factors (major pollution sources) q is unknown. Also, the parameters A and P in the model are not uniquely defined even under the assumption that q is known because $AP = ARR^{-1}P$ for any $q \times q$ nonsingular matrix R. This is a well-known non-identifiability problem in factor analysis and often referred to as "factor indeterminacy" or "factor rotations". Fortunately, under some constraints (identifiability conditions) on either A or P, the parameters are uniquely defined. It should be emphasized that identifiability conditions are additional assumptions on the parameters and so it is important to select conditions that are physically meaningful in a given context of the problem. Here, we employ one such type of identifiability conditions, prespecification of zero elements in P, which assume that there are at least $q-1$ zero elements in each row of P and the rows of q sub-matrices composed of the columns containing the assigned 0's in each row with those assigned 0's deleted are linearly independent. In terms of our data, these conditions imply that at some monitoring stations PM_{10} concentrations are not contributed by a particular source type, and no two source types share exactly the same set of zeros.

Note that estimation of our key parameter P depends on q and also on where to pre-assign zeros in P. In some cases, we may have some prior knowledge on this, i.e., the number of sources and the position of zeros can be assumed known (see, e.g., Park, Guttorp, and Henry 2001). More frequently, that information is lacking, and it becomes a main source of model uncertainty. Park, Oh, and Guttorp (2001) proposed a Bayesian approach that can simultaneously estimate such model uncertainty as well as the parameters in each model. The method computes the posterior probabilities using Markov chain Monte Carlo (MCMC) for a range of plausible models (rather than a single model) selected by varying the number of sources and zero elements in P. We employ their approach here since the information on the number of major sources and the sites with one source missing is not given *a priori* for Seoul PM10 data.

We fit model (1), separately, to Winter data and to Summer data. For q, we try the values q = 1, 2, 3, 4, for each season. As mentioned earlier, each possible combination of q and positions of zeros in the source composition matrix P yields a different model. Selection of candidate positions of zeros in P requires some strategy since there are too many possible combinations for each q. Although it is possible to go through an infinite number of models in principle, it is wiser to keep the number of models compared under some reasonable number (e.g., 20 or 30) due to computational cost. Often we can build up

plausible hypotheses based on exploratory analyses. To find out plausible sets of zeros for each q, we conduct such exploratory analyses using the least squares methods with no identification condition and using UNMIX (Henry, 2000). We then try the elements giving the low proportions in P obtained from those analyses as the candidate zeros (q-1 zeros for each row). It is also possible to try any other sets of zeros based on a physical reason such that industrial emission would not contribute to pollution at station 2.

Results

Sixteen candidate models are selected both for Summer and Winter. The posterior probabilities for those 16 models under the indifference prior are calculated for Summer and Winter, separately, though we do not report those values here for the space. Two-source model with the corresponding posterior probability of 0.99 and 3-source model with the corresponding posterior probability of 0.99 are selected as the best model for Summer and for Winter, respectively. For the best model selected for each season, we report the posterior mean for P based on MCMC samples in Tables 1-2. For ease of interpretation, source profiles are normalized to sum to 100.

Table 1. Posterior mean for the source profiles P for the best model, Summer

site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
S1	4.7	5.0	6.3	2.7	1.9	6.8	10.3	2.7	5.6	3.0	7.5	0	1.8	17.8	15.9	6.7	1.4
S2	6.2	7.6	5.2	5.8	4.7	5.5	4.8	6.1	6.4	4.7	4.7	10.5	8.1	0	2.3	8.1	9.5

Table 2. Posterior mean for the source profiles P for the best model, Winter

site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
S1	0	4.8	8.4	5.5	0	5.4	8.6	2.3	5.6	1.7	3.5	4.6	7.0	14.4	13.7	7.5	7.1
S2	7.2	9.1	2.9	5.6	9.8	6.3	5.4	7.6	4.3	5.7	6.8	7.8	7.2	0	0	6.8	7.8
S3	6.4	1.7	10.9	3.8	5.0	9.5	13.4	2.0	2.9	5.4	4.3	2.3	0	11.9	11.7	9.1	0

The spatial profiles of major sources in those tables show an interesting pattern. For both seasons, one of the source profiles show high proportions at sites, 7, 14, and 15, which are all close to light industry area located in the west part of Seoul. This is consistent to our prior expectation that industrial sources would be one of major PM10 contributors for Seoul throughout the year. The second source spatial profile for Summer and for Winter fairly spreads out over the entire region, which is suspected to be transportation sources. For Winter, we obtain one more source profile, suggesting that there is an additional source type specific to Winter. It is being investigated if it is related to heating sources in Winter or any other source types.

Conclusions

We have presented an approach for identifying the major source types of PM_{10} for Seoul based on the observed PM_{10} concentrations from multiple monitoring sites. Using a factor analytic model, we have attempted to explain the spatial correlations in the data by a set of common sources. To deal with the problem of unknown number of sources and identifiability conditions, a Bayesian approach calculating the posterior probability of each of candidate models based on a MCMC sample was employed. The resulting estimates of the source spatial profiles seemed to be consistent with our prior expectation about the PM_{10} sources in Seoul.

There are several possible directions for future work. Air pollution data often show temporal dependence when measurements are made hourly or at a shorter time intervals. Park, Guttorp, and Henry (2001) developed a multivariate receptor model for temporally correlated data with the assumption on known number of sources and identifiability conditions. When this prior knowledge is not available, how to determine the number of sources and identifiability conditions all together from the data is an open problem. We might still be able to use a Bayesian approach calculating posterior probability for each model, but the computational cost would be formidable due to complexity of each model. In this paper, we considered a

case where a single pollutant is measured over multiple monitoring sites. When multiple pollutants are measured from multiple monitoring sites, how to incorporate spatial variability as well as temporal variability in multivariate receptor modeling is a challenging problem. Even in the case of no temporal dependence, this problem remains unexplored.

References

Henry, R.C. (1991), Multivariate Receptor Models, In Receptor Modeling for Air Quality Management (ed. P. Hopke), pp. 117-147. Amsterdam: Elsevier.

Henry, R.C. (2000), *Personal communication.*

Hong, Y.C., Lee, J. H., Ha, E. H., and Christiani, D. C. (1999), PM_{10} exposure, gaseous pollutants, daily mortality in Inchon, South Korea. *Environ. Health Perspect.* 107, 873-878.

Lee, J.T., Kim, Ho, Hong, Y.C., Kwon, H.J. Schwartz, J., and Christiani, D.C., (2000), Air pollution and daily mortality in seven major cities of Korea, 1991-1997, *Environmental Research.*

Park, E.S, Guttorp, P., and Henry, R.C. (2001), Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC, to appear in *Journal of the American Statistical Association.*

Park, E.S., Oh, M.S., and Guttorp, P. (2001), Multivariate Receptor Models and Model Uncertainty, to appear in *Chemometrics and Intelligent Laboratory Systems.*

Park, E.S., Spiegelman, C.H., and Henry, R.C. (2001) Bilinear estimation of pollution source profiles and amounts by using the multivariate receptor models, to appear in *Environmentrics*.