RESEARCH REPORT

HEALTH EFFECTS INSTITUTE

Number 183
Parts 1 & 2
June 2015

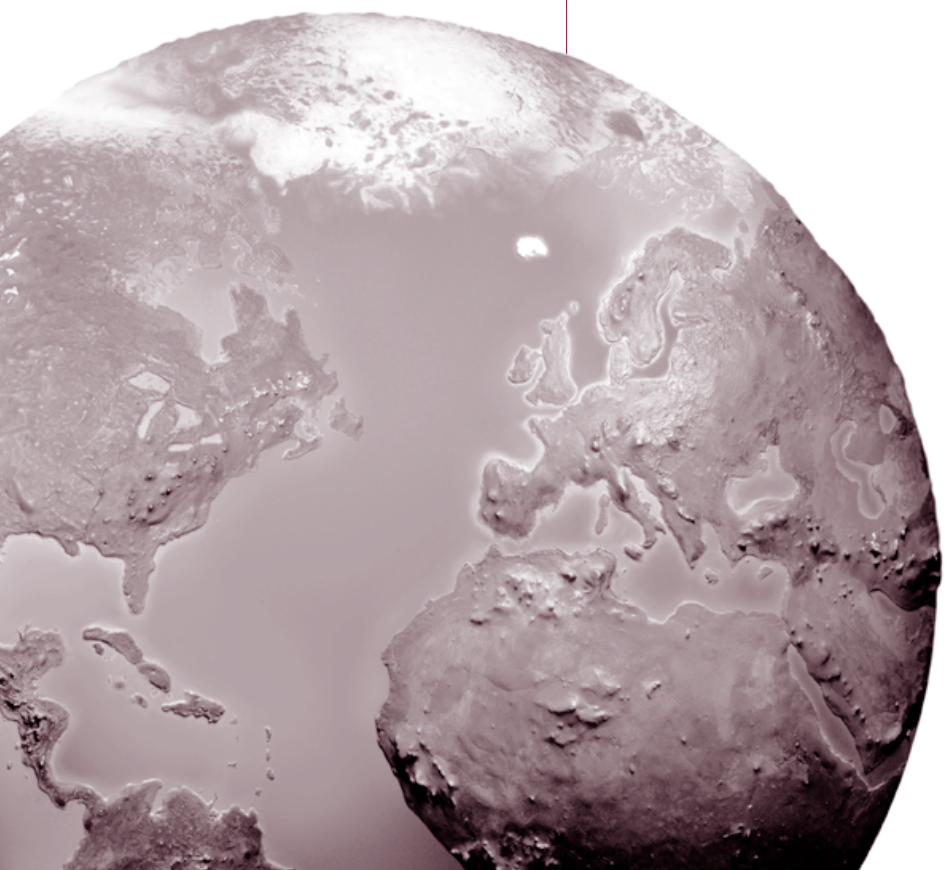
Development of Statistical Methods for Multipollutant Research

Part 1. Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents

Brent A. Coull, Jennifer F. Bobb, Gregory A. Wellenius, Marianthi-Anna Kioumourtzoglou, Murray A. Mittleman, Petros Koutrakis, and John J. Godleski

Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants

Eun Sug Park, Elaine Symanski, Daikwon Han, and Clifford Spiegelman



Development of Statistical Methods for Multipollutant Research

Part 1. Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents

Brent A. Coull, Jennifer F. Bobb, Gregory A. Wellenius, Marianthi-Anna Kioumourtzoglou, Murray A. Mittleman, Petros Koutrakis, and John J. Godleski

Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants

Eun Sug Park, Elaine Symanski, Daikwon Han, and Clifford Spiegelman

with a Critique by the HEI Health Review Committee

Research Report 183

Health Effects Institute

Boston, Massachusetts

Trusted Science • Cleaner Air • Better Health

Publishing history: This document was posted at *www.healtheffects.org* in June 2015.

Citation for Research Report 183 in its entirety:

Health Effects Institute. 2015. Development of Statistical Methods for Multipollutant Research. Research Report 183, Parts 1 & 2. Boston, MA:Health Effects Institute.

Citation for Part 1 only:

Coull BA, Bobb JF, Wellenius GA, Kioumourtzoglou M-A, Mittleman MA, Koutrakis P, et al. 2015. Part 1. Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents. In: Development of Statistical Methods for Multipollutant Research. Research Report 183. Boston, MA:Health Effects Institute.

Citation for Part 2 only:

Park ES, Symanski E, Han D, Spiegelman C. 2015. Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants. In: Development of Statistical Methods for Multipollutant Research. Research Report 183. Boston, MA:Health Effects Institute.

 Cameographics, Belfast, Me., Compositor. Printed by Recycled Paper Printing, Boston, Mass. Library of Congress Catalog Number for the HEI Report Series: WA 754 R432.

Cover paper: made with at least 55% recycled content, of which at least 30% is post-consumer waste; free of acid and elemental chlorine. Text paper: made with 100% post-consumer waste recycled content; acid free; no chlorine used in processing. The book is printed with soy-based inks and is of permanent archival quality.

CONTENTS

ABOUT HEI

The Health Effects Institute is a nonprofit corporation chartered in 1980 as an independent research organization to provide high-quality, impartial, and relevant science on the effects of air pollution on health. To accomplish its mission, the institute

- Identifies the highest-priority areas for health effects research;
- Competitively funds and oversees research projects;
- Provides intensive independent review of HEI-supported studies and related research;
- Integrates HEI's research results with those of other institutions into broader evaluations; and
- Communicates the results of HEI's research and analyses to public and private decision makers.

HEI typically receives half of its core funds from the U.S. Environmental Protection Agency and half from the worldwide motor vehicle industry. Frequently, other public and private organizations in the United States and around the world also support major projects or research programs. HEI has funded more than 330 research projects in North America, Europe, Asia, and Latin America, the results of which have informed decisions regarding carbon monoxide, air toxics, nitrogen oxides, diesel exhaust, ozone, particulate matter, and other pollutants. These results have appeared in more than 260 comprehensive reports published by HEI, as well as in more than 1000 articles in the peer-reviewed literature.

HEI's independent Board of Directors consists of leaders in science and policy who are committed to fostering the public–private partnership that is central to the organization. The Health Research Committee solicits input from HEI sponsors and other stakeholders and works with scientific staff to develop a Five-Year Strategic Plan, select research projects for funding, and oversee their conduct. The Health Review Committee, which has no role in selecting or overseeing studies, works with staff to evaluate and interpret the results of funded studies and related research.

All project results and accompanying comments by the Health Review Committee are widely disseminated through HEI's Web site (*www.healtheffects.org*), printed reports, newsletters and other publications, annual conferences, and presentations to legislative bodies and public agencies.

ABOUT THIS REPORT

Research Report 183, *Development of Statistical Methods for Multipollutant Research,* presents two studies funded by the Health Effects Institute. These studies were conducted by Drs. Brent Coull of the Harvard T.H. Chan School of Public Health, Eun Sug Park of the Texas A&M Transportation Institute, and their colleagues. The report contains the following main elements:

The HEI Statement, prepared by staff at HEI, is a brief, nontechnical summary of the two studies and their findings; it also briefly describes the Health Review Committee's comments on the studies.

The Investigators' Reports, prepared by the two investigative teams, describe the scientific background, aims, methods, results, and conclusions of the studies.

The Critique is prepared by members of the Health Review Committee with the assistance of HEI staff; it places the studies in a broader scientific context, points out their strengths and limitations, and discusses remaining uncertainties and implications of the studies' findings for public health and future research.

The two studies contained in Research Report 183 have gone through HEI's rigorous review process. When an HEI-funded study is completed, the investigators submit a draft final report presenting the background and results of the study. This draft report is first examined by outside technical reviewers and a biostatistician. The report and the reviewers' comments are then evaluated by members of the Health Review Committee, an independent panel of distinguished scientists who have no involvement in selecting or overseeing HEI studies. During the review process, the investigators have an opportunity to exchange comments with the Review Committee and, as necessary, to revise their reports. The Critique reflects the information provided in the final versions of both Investigators' Reports.

PREFACE

HEI's Research Program to Develop Methods for Analyzing Multiple Air Pollutants and Health Outcomes

INTRODUCTION

Air pollution is a complex mixture of gaseous, liquid, and solid components that varies greatly in composition and concentration across the United States and around the world owing to differences in sources, weather, and topography. Air pollution also varies from day to day and by season within a region. Although it is clear that people are exposed to complex mixtures of pollutants emitted by diverse sources, the U.S. Clean Air Act — and most existing air quality guidelines and standards to protect public health — focuses on controlling a common set of pollutants individually (called criteria pollutants in the United States). Given this regulatory approach, it is perhaps not surprising that the majority of data on ambient air pollution levels and on human exposures and their health effects have focused on individual pollutants.

Since the air we breathe is a mixture, the scientific community has considered the possibility that the observed adverse health effects associated with individual pollutants may be partly attributable to the combined effects of multiple pollutants. However, the challenges of determining whether effects are additive, synergistic, or less-than-additive, and of identifying possible effect modifiers in epidemiologic studies, are substantial (Mauderly and Samet 2009). Often, a high degree of correlation exists among levels of different pollutants emitted from similar sources or generated through similar atmospheric processes; and there may be nonlinear interactions among pollutants in relation to health outcomes. These issues complicate and may even preclude the use of conventional linear regression approaches. Exposure measurement and exposure modeling errors contribute additional complications; pollutants that are measured relatively easily (i.e., more frequently and accurately because their concentrations are well above detection levels) will tend to dominate the estimation, even if their effects are less strong than those of other pollutants.

HEI issued Request for Applications (RFA) 09-1, "Methods to Investigate the Effects of Multiple Air Pollution Constituents" in 2009 because it was clear that advancing scientific understanding would require improved statistical methods to determine how the health effects of a pollutant mixture as a whole differ from the effects of individual pollutants within the mixture.

GOALS OF THE RESEARCH PROGRAM

RFA 09-1 solicited research proposals that would address the methodologic difficulties associated with investigating the health effects of multiple pollutants through the development of innovative statistical methods. HEI primarily sought applications for research in which existing statistical approaches (including those from fields outside epidemiology) could be modified, extended, or combined, and then applied to a real-world exposure and health problem, rather than proposals for the development of purely theoretical statistical approaches. RFA 09-1 defined two specific objectives:

1. The research should support the development of innovative statistical methods for studying the combined effects of individual pollutants within complex pollutant mixtures. Analytic approaches could include improvements to existing multivariate methods and the development of strategies for their application or the proposal of new approaches. Of particular interest were multivariate methods adapted to studying highly correlated pollutants and methods to detect the presence of interactions between two or more pollutants and to evaluate their combined effects. All methods proposed were required to include validation of the approach either by using simulation studies or

by conducting a thorough sensitivity analysis with widely available data sets.

2. The research should support the development of innovative statistical methods for studying health effects of air pollution mixtures in animal models and human populations. Of particular interest were methods for characterizing mixtures emitted by specific pollutant sources or groups of sources.

The RFA welcomed proposals for methods that would explore how the effects of a pollutant mixture as a whole differ from the effects of individual pollutants within the mixture. Applicants were expected to employ methods that would be able to analyze both highly correlated pollutant concentration variables and assess the potential effects of measurement error within the chosen statistical framework.

BACKGROUND

At the time the RFA was issued, adequate statistical methods designed for analyzing the relationships among multiple pollutants and health effects were unavailable. In order to better understand the health effects of exposure to the mixture of air pollutants that people actually breathe, to delineate the contribution of individual pollutants or mixtures to adverse health effects, and to address emissions from the sources of those pollutants more cost-effectively, approaches that would go beyond the single-pollutant framework were clearly needed. A 2004 report from the National Research Council (NRC) Committee on Air Quality Management in the United States called for changing the entire air quality management system to a multipollutant approach. The report recommended that the U.S. Environmental Protection Agency (U.S. EPA) consider multiple pollutant scenarios in the National Ambient Air Quality Standards (NAAQS) review and standard setting process: "Although the committee does not believe that the science has evolved to a sufficient extent to permit the development of multipollutant NAAQS, it would be scientifically prudent to begin to review and develop NAAQS for related pollutants in parallel and simultaneously" (NRC 2004).

The U.S. EPA responded to the NRC report by undertaking a number of activities in support of multipollutant research and a NAAQS targeted specifically to multipollutant mixtures. In late 2006, the Agency hosted the first of several workshops on multipollutant research and commenced efforts to develop a multipollutant NAAQS in 2010 (U.S. EPA 2006, 2011). In 2007, the U.S. EPA also began development of its first two-pollutant Integrated Science Assessment for nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), which was finalized in December 2008.

Following the NRC recommendations, HEI also included multipollutant research as part of its research agenda, specifying in its Strategic Plan for 2005–2010 the health effects of air pollution mixtures as a priority research and review topic. Specifically, this plan called for HEI to "undertake targeted research programs on PM (particulate matter) and gases and on air toxics, two important mixtures within the broader air pollution mixture". Following the discussions about research needs at the U.S. EPA workshops, HEI issued RFA 09-1 in 2009.

At the time, some existing multipollutant modeling approaches were available to researchers in the fields of epidemiology and air pollution exposure. The process of attributing measured concentrations of multiple pollutants to the emissions from specific categories of sources, known as source apportionment, had been evolving and had become increasingly standardized during the early 2000s (Thurston et al. 2005). When statistically feasible, researchers also employed variations on linear regression, such as multivariate regression models, which simultaneously incorporated covariates for multiple pollutants. Both approaches are briefly described here.

SOURCE APPORTIONMENT

When strong correlations among pollutants in given mixtures preclude the use of multiple individual exposure variables in conventional health effects models, source apportionment is used to analyze the mixture of pollutants over time and space. It is a latent-variable method, usually applied in models that include multiple variables, at least one of which is unobserved (or latent). Factor analysis is a special type of latent-variable model used in source apportionment where the analysis assumes that multiple variables are linked together through their association with a small number of latent variables, called factors. Source apportionment is the process of attributing emission sources to factors based

on the composition of the factor. For example, a factor analysis of roadside particulate pollution data may yield a factor in which levels of copper and iron are high and vary together; in a source apportionment, this factor might be attributed to tire and brake wear given what we know about the composition of tires and brakes.

Using source apportionment to classify mixtures — based on source-specific markers in a mixture — can also link health effects with emissions from specific sources (such as facilities or activities). This approach uses the resulting quantification of components that comprise the different source mixtures in a given environment to evaluate their individual or combined contributions to health effects.

However, source-apportionment techniques are not capable of assessing the effects of interactions among the different source-apportioned mixtures, and they may not take into account the underlying biological plausibility of any given mixture to affect health. In addition, when HEI issued RFA 09-1, many researchers were using source-apportionment methods and multivariate-receptor models as "black box tools" and were not linking them sufficiently to rigorous statistical practice or demonstrating an understanding of method limitations. Moreover, the inherent uncertainty of variables generated through source apportionment, due in part to errors in the measurement of individual pollutant concentrations, was not reflected in the estimates of their associations with the health outcomes, thus rendering reproducibility and comparison among different studies difficult.

MULTIVARIATE REGRESSION MODELS

When data sets contain measurements of many constituents of air pollution obtained at different places and time points together with information about health outcomes, and when there is sufficient variability in these data, multivariate analyses of the association between constituents and health outcomes may be possible. Such analyses are aimed primarily at estimating the effects of specific constituents of interest while accounting for the potential effects of confounding. Moreover, multivariate regression models can be used to detect whether the effects of various pollutants are additive or not. However, there are limitations to the value of simply introducing a number of pollutant variables and interaction terms simultaneously into a regression analysis and carrying out multivariate rather than univariate regressions. For example, high degrees of correlation among covariates render the results statistically unstable and difficult to interpret, and stepwise methods are inadequate in the presence of strong collinearity.

STUDIES FUNDED UNDER RFA 09-1

The three studies funded under RFA 09-1 represent a variety of statistical approaches and of data sets used to test them. The studies by Dr. Brent Coull and Dr. Eun Sug Park and their colleagues are described in Parts 1 and 2 of this report (Research Report 183). The study by Dr. John Molitor and associates has been completed and is expected to be published in 2016. The studies are described briefly below.

Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents, Brent Coull, Harvard T.H. Chan School of Public Health (Principal Investigator)

Coull and colleagues developed a new analysis framework based on methods that simultaneously quantify variability in health outcomes and exposure data for multiple pollutants in order to identify the mixture profiles (groupings of pollutants and concentrations) most highly associated with the health outcomes. They developed and applied these methods using simulations, pollutant concentration and health outcomes data from the "Maintenance of Balance, Independent Living, Intellect, and Zest in the Elderly of Boston" (MOBILIZE) study cohort of senior citizens living in the Boston area, and toxicologic data from canine studies.

Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants, Eun Sug Park, Texas A&M Transportation Institute (Principle Investigator)

Park and colleagues developed enhanced statistical methods to jointly assess source factors and health effects using multivariate source-characterization and source-apportionment models together with a health outcomes analysis. The investigators' approach incorporated the uncertainty in the source apportionments into the estimation of the source-related health effects. They applied their methods to data sets for daily pollutant concentrations and acute health outcomes in

Phoenix, Arizona, and Houston, Texas, and compared the results with those obtained using conventional methods of estimation.

***Modeling of Multipollutant Profiles with Applications to RIOPA Study Data and to Indicators of Adverse Birth Outcomes Using Data from the UCLA Environment and Pregnancy Outcomes Study*, John Molitor, Oregon State University (Principal Investigator)**

Molitor and colleagues developed and applied statistical methods to examine associations among geographically based patterns of air pollutant concentrations, birth outcomes, and socioeconomic status. The investigators used a large data set of pollutant concentrations (for NO_2, PM ≤ 2.5 μm in aerodynamic diameter, and on-road and off-road diesel exhaust) and data on birth outcomes from Los Angeles County, California. They first used Bayesian statistical methods to identify clusters of specific mixtures of pollutants and pollutant concentrations frequently found together in census units, and then associated those pollutant profiles with data on socioeconomic status and health outcomes using regression methods.

REFERENCES

Mauderly JL, Samet JM. 2009. Is there evidence for synergy among air pollutants in causing health effects? Environ Health Perspect 117:1–6. doi:10.1289/ehp.11654.

National Research Council. 2004. Research Priorities for Airborne Particulate Matter. IV. Continuing Research Progress. Washington, DC:National Academy Press.

Thurston GD, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, et al. 2005. Work-group report: Workshop on source apportionment of particulate matter health effects — Intercomparison of results and implications. Environ Health Perspect 113:1768–177.

U.S. Environmental Protection Agency. 2006. Workshop on Interpretation of Epidemiologic Studies of Multipollutant Exposure and Health Effects. Federal Register/ Vol. 71, No. 225 / Wednesday, November 22, 2006. *http://docs.regulations.justia.com/entries/2006-11-22/E6-19806.pdf*.

U.S. Environmental Protection Agency. 2011. Overview of EPA Multipollutant Science and Risk Analysis Workshop: February 22–24, 2011. *www.epa.gov/ncer/publications/ workshop/04_07_2011/djohns_cac_110408.pdf* .

ABBREVIATIONS AND OTHER TERMS

NAAQS	National Ambient Air Quality Standard
NO_2	nitrogen dioxide
NRC	National Research Council
PM	particulate matter
RFA	request for applications
U.S. EPA	U.S. Environmental Protection Agency

HEI STATEMENT

Synopsis of Research Report 183, Parts 1 & 2

New Statistical Methods for Analyzing Multiple Pollutants, Sources, and Health Outcomes

BACKGROUND

The National Research Council recommended in 2004 that the U.S. Environmental Protection Agency take steps to address the presence of a complex, multipollutant atmosphere in the process for reviewing and setting the National Ambient Air Quality Standards, which are currently based on single pollutants. One of those steps included improving statistical methods to evaluate how simultaneous exposure to multiple ambient air pollutants affects human health. Conventional statistical methods are not well suited to deal with high correlations among pollutants, differences in the composition of pollutant mixtures over time and space, or differences in how accurately a person's actual exposures to individual pollutant concentrations have been estimated. These factors can lead to errors in the estimation of the health effects associated with individual or multiple pollutants and the emission sources with which they may be associated.

In response to these concerns, the Health Effects Institute issued request for applications 09-1, "Methods to Investigate the Effects of Multiple Air Pollution Constituents," to fund development of innovative statistical methods that could be applied to real-world exposures and health problems. HEI funded three studies: the two studies led by Dr. Brent Coull and Dr. Eun Sug Park are described in the current report, and a third study led by Dr. John Molitor is expected to be published in 2016. Both Coull and Park proposed the use of Bayesian statistical methods, which essentially allow for the integration of prior knowledge or data about a problem with new data in the same analysis, thus allowing for a more comprehensive evaluation of available information, including the characterization of uncertainty in the analytic process. Both investigative teams explored joint modeling of exposure and health outcomes in contrast to the conventional two-stage approach in which exposures are

What These Studies Add

- Both the Coull and Park studies have advanced the use of Bayesian statistical methods to address shortcomings in the ability of existing approaches to disentangle the roles of individual pollutants in short-term studies of complex multipollutant exposures and their health effects.
- Coull and associates developed methods to identify which key pollutants within a simple mixture are most closely associated with adverse health outcomes, to accommodate a variety of exposure–response relationships, and to characterize uncertainty in the estimated health effects more fully.
- Park and colleagues extended existing methods for characterizing relationships between emission sources and health by (1) allowing for the contributions from sources to be correlated and (2) making sure that the health effects estimates account for various uncertainties in estimating the source contributions. The team also developed enhanced models that could take into account correlations among pollutants from more than one monitoring location and estimate source contributions at locations of interest for which monitoring data may be lacking.

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Brent A. Coull of the Harvard T.H. Chan School of Public Health, Boston, MA, and by Eun Sug Park of the Texas A&M Transportation Institute, College Station, TX, and their colleagues. Research Report 183 contains both the detailed Investigators' Reports and a Critique of the study prepared by the Institute's Health Review Committee.

estimated first and then input to the health effects analysis. Each team developed and demonstrated their methods using data on health outcomes related to short-term changes in particulate matter composition and levels.

STUDY BY COULL AND COLLEAGUES: STATISTICAL LEARNING METHODS FOR THE EFFECTS OF MULTIPLE AIR POLLUTION CONSTITUENTS

Approach

Coull and his colleagues developed methods that simultaneously select pollutants to include in the models, provide flexible approaches to estimating exposure–response relationships (for example, allowing them to be nonlinear), identify interactions among pollutants (for example, additive or synergistic effects), and allow for the quantification of uncertainty (by using Bayesian kernel machine regression [BKMR] methods). These methods involve a joint-estimation approach in which the identification of important exposure variables is, in a sense, "supervised," or influenced by, the data on health outcomes.

The investigators formally developed and tested different features of their BKMR methods first in three simulation studies and then in two real-world health and exposure data sets. The simulation studies were designed to compare the performance of their methods with those of more conventional ones in a range of plausible scenarios, defined by the investigators, involving different numbers of important pollutants or sources, nonlinear as well as linear exposure–response relationships, and different kinds of interactions among the exposure constituents. An important feature of their simulations was that their air pollution data sets were generated from actual $PM_{2.5}$ constituent data measured at a Boston monitoring site, thereby retaining the realistic joint distributions and correlations among the multiple pollutants. They then applied their methods to data from two previously published Boston studies — an epidemiologic study that had evaluated changes in blood pressure after short-term exposure to constituents of $PM_{2.5}$ in patients 70 years of age and older, and a toxicologic study with laboratory dogs. These studies also relied on pollutant data from the same Boston monitoring site.

Results and Discussion

In its independent review of the study, the HEI Health Review Committee noted that the statistical approach developed by Coull and associates had carefully addressed a number of the challenges researchers face in dealing with multiple pollutants and sources when using more conventional statistical methods. Their methods also allowed the uncertainties associated with statistical modeling to be more fully reflected in the health effects estimates, providing useful insight into the degree of confidence in those estimates. The Committee thought the investigators had provided a strong theoretical basis for their approach and that their simulations and real-world applications were well chosen to demonstrate the practical use of their methods. A strength of those choices was that the simulated pollutant data sets were generated from the same pollutant data used in the two real-world studies and thus made the simulation results more relevant for comparisons with those of the previous epidemiologic and toxicologic analyses.

The methods worked as expected in the simulations but with some limitations in the analyses of the epidemiologic and toxicologic data sets. In the simulated data sets, the methods characterized exposure–response relationships in various forms and were more likely to correctly identify the pollutants used to predict the adverse health outcomes than more standard methods. However, as with conventional statistical methods, it remained challenging to identify correctly the relative importance of an individual pollutant's contribution to health outcomes when high degrees of correlation existed among the suite of pollutants. A limitation in the data sets for older adults and for dogs was that they did not have either the size or complexity to represent the kinds of interactions among pollutants or the nonlinearities in the exposure–response relationships that would be necessary to test those features of the methods.

The Committee thought it was likely that the methods developed by Coull and colleagues were more systematic and transparent in identifying the absence of interactions or nonlinearities than conventional data analysis approaches would be. In conventional analyses, investigators would ordinarily need to cycle through a series of models to

test whether interactions were present, a process that would require a number of analytic choices and raise issues of multiple testing and possibly false-positive findings. Methods such as the ones developed by Coull and associates could be helpful in minimizing the ad hoc nature of this process. However, as is the case in the development of any statistical approach, these methods need to be applied in a broader range of scenarios before their usefulness in real-world practice can be ascertained.

STUDY BY PARK AND COLLEAGUES: DEVELOPMENT OF ENHANCED STATISTICAL METHODS FOR ASSESSING HEALTH EFFECTS ASSOCIATED WITH AN UNKNOWN NUMBER OF MAJOR SOURCES OF MULTIPLE AIR POLLUTANTS

Approach

Park and her colleagues developed a set of methods to analyze daily variations in health and source-apportioned air pollution (time-series data). In their first specific aim, Park and associates developed a Bayesian modeling approach that estimated the number of sources and the contributions of each to exposures at the same time as it estimated the effects of those exposures on human health outcomes. Their joint modeling approach incorporated uncertainties in the source-apportionment process into the final estimates of uncertainty in the health effects estimates. More specifically, their analysis allowed them to examine the impact of correlations among source contributions to exposure as well as incorporating uncertainty from other modeling assumptions into their final estimates of health effects. In standard applications of source apportionment, the number of sources and how much each one contributes to exposure is assumed to be known without error; if this assumption is not true, it could lead to misspecification of how health effects are attributed to different categories of sources.

In their second specific aim, the team developed methods for modeling the contributions of multiple sources to exposures that could handle more complex data and model structures than conventional source-apportionment methods. That is, they designed methods to incorporate data from more than one monitoring location to account for spatial correlations among multiple pollutant measurements collected at several locations, and to estimate source contributions at locations where no data were available.

For each of the specific aims, the investigators evaluated their methods in two steps. First, they conducted simulation studies in which the characteristics of the data, sources, and health effects were specified by the investigators (that is, they did not draw directly from actual collected data as did Coull and associates). Second, they applied their methods to real-world data sets in order to gain a more practical perspective. To test their methods for estimating source-related health effects (Aim 1), the investigators studied the associations of daily $PM_{2.5}$ speciation data with respiratory mortality in Houston, Texas, and with cardiovascular mortality in Phoenix, Arizona. In both of these cases, $PM_{2.5}$ data had been collected at individual monitoring sites. To test their more complex source-apportionment models (Aim 2), they examined data on volatile organic compounds from nine monitoring sites in Harris County, Texas, near Houston.

Results and Discussion

In its independent review of the study, the HEI Health Review Committee concluded that Park and colleagues had tackled an extremely challenging technical problem and, in spite of its difficulty, conducted a high-quality study that has provided a meaningful extension of existing source-apportionment approaches. Useful innovations include the joint estimation of sources and health effects, while accounting for uncertainty in source-apportionment models, allowing for spatial correlations among data from multiple monitoring locations, and estimating source contributions to exposure at locations without monitoring data. However, implementing the joint models and the spatial multivariate receptor models is challenging because they require data to be in a specific form that is often not available in existing data sets.

The Committee thought that Park and associates had raised important scientific issues with existing methods and developed new approaches to address them. The investigators properly developed and tested their methods in both simulations and applications to real-world data sets. The simulations

performed well under the range of conditions evaluated. The applications of the methods to real-world data also provided evidence that the methods appear to work as intended in identifying sources and source-related health effects, albeit with differences from other published studies that need further investigation. Uncertainties in the health effects estimates tended to be larger, which reflected the more comprehensive accounting for uncertainty in the work. The Committee thought the enhanced modeling methods developed as part of the investigators' second aim appeared to be a useful innovation and were able to predict source contributions at unmonitored locations. However, for any of these methods to gain widespread scientific applicability, the Committee advised that they need to be applied in other settings, particularly ones that would allow comparisons with other studies that use more conventional approaches.

CONCLUSIONS

The HEI Health Review Committee concluded that each of the studies by Coull and Park and their colleagues addressed important but separate questions in multipollutant research. Both investigator teams followed logical steps in developing their methods from the conceptual underpinnings and then applying the methods to simulated and real-world data sets. Each team made considerable progress in demonstrating the feasibility and applicability of their approaches.

Challenges still remain, however, and further work is necessary to apply and evaluate the proposed methods. Although both sets of methods are already quite computationally demanding, they need to be evaluated in a broader range of real-world settings representing different levels of data complexity. They have also not yet been evaluated in studies of long-term exposure to air pollution. Where possible, these evaluations should include side-by-side comparisons of the new approaches against the more conventional two-stage approach. Such direct comparisons could help to determine whether the additional complexity of these new methods will lead to better understanding of how pollutant mixtures and their sources may contribute to effects on human health and, ultimately, to better decisions about how to control them.

INVESTIGATORS' REPORT

Part 1. Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents

Brent A. Coull, Jennifer F. Bobb, Gregory A. Wellenius, Marianthi-Anna Kioumourtzoglou, Murray A. Mittleman, Petros Koutrakis, and John J. Godleski

Department of Biostatistics (B.A.C., J.F.B.), Department of Environmental Health (M.-A.K., P.K., J.J.G.), and Department of Epidemiology (M.A.M.), Harvard T.H. Chan School of Public Health, Boston, Massachusetts; Center for Environmental Health and Technology, Brown University, Providence, Rhode Island (G.A.W.)

ABSTRACT

INTRODUCTION

The United States Environmental Protection Agency (U.S. EPA*) currently regulates individual air pollutants on a pollutant-by-pollutant basis, adjusted for other pollutants and potential confounders. However, the National Academies of Science concluded that a multipollutant regulatory approach that takes into account the joint effects of multiple constituents is likely to be more protective of human health. Unfortunately, the large majority of existing research had focused on health effects of air pollution for one pollutant or for one pollutant with control for the independent effects of a small number of copollutants. Limitations in existing statistical methods are at least partially responsible for this lack of information on joint effects. The goal of this project was to fill this gap by developing flexible statistical methods to estimate the joint effects of multiple pollutants, while allowing for potential nonlinear or nonadditive associations between a given pollutant and the health outcome of interest.

METHODS

We proposed Bayesian kernel machine regression (BKMR) methods as a way to simultaneously achieve the multifaceted goals of variable selection, flexible estimation of the exposure–response relationship, and inference on the strength of the association between individual pollutants and health outcomes in a health effects analysis of mixtures. We first developed a BKMR variable-selection approach, which we call component-wise variable selection, to make estimating such a potentially complex exposure–response function possible by effectively using two types of penalization (or regularization) of the multivariate exposure–response surface. Next we developed an extension of this first variable-selection approach that incorporates knowledge about how pollutants might group together, such as multiple constituents of particulate matter that might represent a common pollution source category. This second grouped, or hierarchical, variable-selection procedure is applicable when groups of highly correlated pollutants are being studied.

To investigate the properties of the proposed methods, we conducted three simulation studies designed to evaluate the ability of BKMR to estimate environmental mixtures responsible for health effects under potentially complex but plausible exposure–response relationships. An attractive feature of our simulation studies is that we used actual exposure data rather than simulated values. This real-data simulation approach allowed us to evaluate the performance of BKMR and several other models under realistic joint distributions of multipollutant exposure. The simulation studies compared the two proposed variable-selection

This Investigators' Report is one part of Health Effects Institute Research Report 183, which also includes a Critique by the Health Review Committee and an HEI Statement about the research project. Correspondence concerning the Investigators' Report may be addressed to Dr. Brent Coull, Department of Biostatistics, the Harvard T. H. Chan School of Public Health, Boston, MA 02115; e-mail: *bcoull@hsph.harvard.edu*.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83467701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

* A list of abbreviations and other terms appears at the end of the Investigators' Report.

approaches (component-wise and hierarchical variable selection) with each other and with existing frequentist treatments of kernel machine regression (KMR).

After the simulation studies, we applied the newly developed methods to an epidemiologic data set and to a toxicologic data set. To illustrate the applicability of the proposed methods to human epidemiologic data, we estimated associations between short-term exposures to fine particulate matter constituents and blood pressure in the Maintenance of Balance, Independent Living, Intellect, and Zest in the Elderly (MOBILIZE) Boston study, a prospective cohort study of elderly subjects. To illustrate the applicability of these methods to animal toxicologic studies, we analyzed data on the associations between both blood pressure and heart rate in canines exposed to a composition of concentrated ambient particles (CAPs) in a study conducted at the Harvard T. H. Chan School of Public Health (the Harvard Chan School; formerly Harvard School of Public Health; Bartoli et al. 2009).

RESULTS

We successfully developed the theory and computational tools required to apply the proposed methods to the motivating data sets. Collectively, the three simulation studies showed that component-wise variable selection can identify important pollutants within a mixture as long as the correlations among pollutant concentrations are low to moderate. The hierarchical variable-selection method was more effective in high-dimension, high-correlation settings. Variable selection in existing frequentist KMR models can incur inflated type I error rates, particularly when pollutants are highly correlated.

The analyses of the MOBILIZE data yielded evidence of a linear and additive association of black carbon (BC) or Cu exposure with standing diastolic blood pressure (DBP), and a linear association of S exposure with standing systolic blood pressure (SBP). Cu is thought to be a marker of urban road dust associated with traffic; and S is a marker of power plant emissions or regional long-range transported air pollution or both. Therefore, these analyses of the MOBILIZE data set suggest that emissions from these three source categories were most strongly associated with hemodynamic responses in this cohort.

In contrast, in the Harvard Chan School canine study, after controlling for an overall effect of CAPs exposure, we did not observe any associations between DBP or SBP and any elemental concentrations. Instead, we observed strong evidence of an association between Mn concentrations and heart rate in that heart rate increased linearly with increasing concentrations of Mn. According to the positive matrix factorization (PMF) source apportionment analyses of the multipollutant data set from the Harvard Chan School Boston Supersite, Mn loads on the two factors that represent the mobile and road dust source categories.

The results of the BKMR analyses in both the MOBILIZE and canine studies were similar to those from existing linear mixed model analyses of the same multipollutant data because the effects have linear and additive forms that could also have been detected using standard methods.

CONCLUSIONS

This work provides several contributions to the KMR literature. First, to our knowledge this is the first time KMR methods have been used to estimate the health effects of multipollutant mixtures. Second, we developed a novel hierarchical variable-selection approach within BKMR that is able to account for the structure of the mixture and systematically handle highly correlated exposures. The analyses of the epidemiologic and toxicologic data on associations between fine particulate matter constituents and blood pressure or heart rate demonstrated associations with constituents that are typically associated with traffic emissions, power plants, and long-range transported pollutants. The simulation studies showed that the BKMR methods proposed here work well for small to moderate data sets; more work is needed to develop computationally fast methods for large data sets. This will be a goal of future work.

INTRODUCTION

Air pollution is a modifiable risk factor shown to be associated with increased respiratory and cardiovascular morbidity and mortality. Air pollution is a complex mixture of gaseous and particulate constituents. The U.S. EPA currently regulates air pollutants individually on the basis of analyses adjusted for other pollutants and potential confounders. However, the National Academies of Science (National Research Council 2004) concluded that a multipollutant regulatory approach that takes into account the joint effects of multiple constituents would likely be more protective of human health. Specifically, the U.S. EPA recently stated, "In recent years, air pollution scientists and policy makers have recognized the potential benefits of adopting a multipollutant approach to evaluating health impacts of air pollution and management of air quality (National Research Council [2004]). . . . A single-pollutant approach fails to address unmeasured or infrequently measured pollutants that could have significant health effects. There can be important health consequences from exposure to the air pollution mixture as a whole (Brook et al. 2009)."

Several environmental health applications have shown that exposures may have nonlinear effects on a health outcome, and that pollutants may also interact to yield joint

effects on health. A good example of a nonlinear effect is the effect of climate or weather on mortality, which follows a U shape reflecting that more people tend to die in periods of extreme cold and extreme heat (Zanobetti and Schwartz 2008). In children's health, it has been shown that associations between biomarkers of Mn exposure and neurodevelopment can follow an inverse-U relationship reflecting that adverse effects are associated with low doses and high doses, but not moderate doses, possibly because Mn can biologically play the role of both a nutrient and a toxin (Claus Henn et al. 2010). Related work has also shown that multiple metals may interact in their effects on neurodevelopment (Claus Henn et al. 2012).

Unfortunately, in large part due to limitations in existing methods for assessing the health impacts of multipollutant mixtures, the majority of existing research on the effects of specific constituents has produced estimates from a model that assumes additivity and linearity of these effects (Dominici et al. 2010; Greenbaum and Shaikh 2010; Hidy and Pennell 2010; Vedal and Kaufman 2011; Mauderly et al. 2010). Despite recommendations by the National Academy of Sciences and the U.S. EPA that a multipollutant risk assessment should incorporate potential interactions among pollutants in their effects on health, to our knowledge very few epidemiologic studies have reported interactions between even the most often studied pollutants — ozone and particles. The only epidemiologic reports of interactions among environmental exposures relate to pollution and temperature (Pattenden et al. 2010; Qian et al. 2010; Ren et al. 2011). As a result, Mauderly and colleagues (2010) concluded that the ability of the air pollution health research community to support a multipollutant air quality management framework was limited. The goal of the current project was to address this limitation by developing flexible methods to estimate the joint effects (i.e., interactions) of multiple pollutants, while allowing for potential nonlinear associations between a given pollutant and the health outcome of interest.

The most common approach to assessing the joint health effects of multiple pollutants is to use a form of multiple regression that includes as predictors both the concentrations of the pollutants and the variables that confound the exposure–health relationship. Consider health outcome Y_i for observation i, and denote the mean of the health outcome as $\mu_i^Y = E(Y_i)$. The index i here may represent a subject in a cohort study or a day in a time-series study. The typical form of the model is

$$g\left(\mu_i^Y\right) = \boldsymbol{\beta}_0 + \mathbf{z}_i^\top \boldsymbol{\gamma} + \mathbf{x}_i^\top \boldsymbol{\beta}, \quad (1)$$

where $\mathbf{z}_i$ is an $M \times 1$ vector of pollutant concentrations corresponding to outcome i, $\mathbf{x}_i$ is a $q \times 1$ vector of variables containing information on potential confounders, Y_i has a distribution of natural exponential family form (although in this project we considered normally distributed, continuous outcomes), and g is a monotone link function that depends on the type of outcome analyzed. Here, the $M \times 1$ vector $\boldsymbol{\gamma}$ and the $q \times 1$ vector $\boldsymbol{\beta}$ represent the independent effects of the multiple pollutants and of the confounders, respectively. This formulation for the confounding variables $\mathbf{x}_i$ is general enough to accommodate complex forms for the effects of confounders, such as nonlinear natural spline terms. This approach assumes that the independent effect of each pollutant affects the mean outcome (which has been suitably transformed by the link function) in a linear and additive way (referred to as the additive linear model). Such an approach has several drawbacks when our interest focuses on the joint effects of a multipollutant mixture. For example, this approach without any form of variable selection can have problems associated with multicollinearity, which can cause the estimates of the independent effects γ_j, $j = 1,\ldots,M$, to be unstable.

A popular approach to dealing with the collinearity problems in a standard multipollutant regression analysis is to apply one of several variable-selection methods that retains only a subset of pollutants in the model. From simplest to most complex, these include (1) select a priori a subset of the pollutants thought to represent different pollution source categories or different atmospheric processes and thus not to be highly correlated; (2) orthogonalize the pollutant data by fitting some type of principal component analysis, factor analysis, source apportionment, or multivariate receptor model and plugging the resulting components, factors, or source category contributions as predictors into the model; and (3) apply more sophisticated variable-selection techniques such as the least absolute shrinkage and selection operator (LASSO; Tibshirani 1994) or the Bayesian stochastic search variable-selection method (George and McCulloch 1993); the latter method estimates the probability of including each predictor in the model.

Because all of these methods are based on the additive linear model, they still fall short of the ultimate goal of flexibly estimating joint effects of a complex multipollutant mixture. For example, the basic model (equation 1) to which these variable-selection procedures are applied assumes a lack of synergy (i.e., no interactions) among the exposure terms in the model, whether they be pollutants selected a priori, latent factors, or pollutants remaining after an advanced variable-selection algorithm is applied. One can manually build such interactions into the model, but that can result in an explosion of the number of terms

and thus create an even more difficult model-selection problem. Moreover, shrinkage and other regression methods that use the output of dimension reduction techniques as covariates in a regression framework typically make strong assumptions about the form of the exposure–response relationship. For instance, LASSO methods shrink individual regression coefficients back toward zero, but such shrinkage is typically based on a model assuming that all of the pollutants or source contributions have a linear relationship with the mean of the outcome. LASSO also presents challenges for formal inference because obtaining standard errors for non-zero coefficient estimates is not yet a straightforward process, although some progress has been made in this area recently (Chatterjee and Lahiri 2011).

More generally, Billionnet and associates (2012) provided a review of more refined, complex methods to analyze the health effects of exposure to multipollutant mixtures. The authors considered existing statistical methods within the broad categories of dimension reduction (principal component analyses, partial least squares, and sparse extensions thereof), hierarchical model formulations (two-stage analyses), classification methods that categorize patterns of exposure (K-means, hierarchical clustering, and self-organizing maps), and recently-developed statistical learning methods (classification and regression trees, random forest analysis, and logic regression). Most of these methods are useful in summarizing evidence addressing some, but typically not all, of the aims of this project.

For instance, regression tree (Hastie et al. 2009) and random forest (Brieman 2001) methods allow for nonlinear and nonadditive effects of individual pollutants. Regression trees repeatedly partition the predictor space at the cutpoint of a single predictor that gives maximal separation in the outcome of interest. Consider the simple setting with only two predictors, X_1 and X_2. Then suppose the first cutpoint is the value a of X_1. The algorithm then proceeds to find the next cutpoint, which could again be applied to X_1 or to the second exposure X_2. Suppose this next cutpoint is the value b of exposure X_2 for those observations having $X_1 < a$, and at value c of X_2 for those observations having $X_1 > a$. The algorithm is repeated until one of several possible stopping criteria is met, yielding a partition of the multipollutant predictor space in which each group of observations (or cluster) within the partition is estimated to have a constant outcome mean. However, standard regression tree analyses do not provide parsimonious descriptions of health effects, which can make it difficult to achieve meaningful quantification of health effects for risk assessment purposes.

Random forests (see also Liaw and Wiener 2002) partially address this shortcoming by producing "variable importance scores". The most basic form of random forests repeatedly performs regression tree analysis on randomly resampled observations from the original data set, otherwise known as bootstrapping a regression tree analysis. This approach improves upon regression trees in that it produces a single summary importance score for each pollutant while relaxing the simple assumptions of linearity and additivity. However, because these scores simply reflect the decrease in predictive performance, rather than direction or pattern of an association between a given pollutant or group of pollutants and health, it is difficult to summarize the nature of a given exposure–response relationship. Further, it is not immediately clear how to control for confounding within a random forest analysis, although we are aware of work in progress generalizing this class of models to adjust for confounding. Liaw and Wiener (2002) noted that the absolute magnitude of the variable importance scores can be sensitive to tuning parameters chosen as part of the analysis, such as the number of bootstrap resamples and the size of the subset of variables chosen in each resample, but that the rank ordering of these importance scores across the multiple pollutants is relatively stable. Accordingly, these random forest importance scores are most useful as a type of screening tool that ranks the pollutants in order of importance in predicting a health outcome under very general assumptions about the form of the relationship between the multivariate exposure and the outcome.

In this project we propose BKMR methods as a way to simultaneously achieve the multifaceted goals of a health effects analysis of mixtures. KMR methods have become a popular tool in statistical genomics, resulting in powerful tests of genetic effects on health endpoints (Liu et al. 2007, 2008) and providing flexible methods for risk prediction based on genomic inputs (Cai et al. 2011). In genomic applications, KMR methods have been applied primarily to test for the overall effect of a genetic pathway (Liu et al. 2007, 2008) or for the effect of a gene in the presence of a possible gene–gene or gene–environment interaction (Zou et al. 2010; Maity and Lin 2011). In the context of computer experiments, Linkletter and colleagues (2006) applied Gaussian process models (a special case of KMR) with variable selection to identify a subset of inputs with the largest impacts on the system being studied. Savitsky and associates (2011) considered a general framework for Gaussian process models with variable selection and evaluated their performance in terms of their predictive power and ability to correctly select relevant variables.

In contrast to this previous work that has focused on variable selection and prediction, the major goal in environmental health applications is estimating the exposure–response function. To our knowledge the BKMR approach, which can be considered as an analysis that is "supervised" by health data, has not been explored in environmental epidemiologic settings involving multipollutant exposures.

We adopted a Bayesian paradigm for several reasons. A Bayesian approach allows for simultaneous, as opposed to sequential, testing of the importance of individual pollutants. This feature provides a realistic assessment of the uncertainty associated with identifying the important pollutants in the mixture (which we refer to as variable selection) as opposed to calculating standard errors conditional on the set of pollutants included in the model. In the end, the strategy effectively estimates a high-dimensional, potentially nonlinear and nonadditive function of pollutant concentrations that reflects the joint association of the mixture with health outcomes.

The first Bayesian variable-selection scheme, which we refer to as component-wise variable selection, is similar in spirit to what has been proposed previously (Zou et al. 2010; Savitsky et al. 2011) and assesses the importance of each pollutant individually. This approach makes estimating a complex exposure–response function possible by effectively using two types of penalization, or regularization, of the multivariate exposure–response surface. The variable-selection scheme itself serves to remove, or effectively zero-out, pollutants for which the data do not provide evidence of an association with the outcome. This selection feature reduces the dimensionality of the multivariate exposure vector for which the form of the exposure–response relationship is estimated. Furthermore, this relationship is estimated while penalizing the complexity of the multivariate surface, much like existing mixed-model formulations of penalized-spline generalized additive models (GAMs; Wood 2006). Rather than being conducted in two separate steps, these two forms of regularization are used within the proposed Bayesian model-fitting algorithm, which provides a unified scheme for variable selection, model fitting, and inference.

Component-wise variable selection is not particularly effective when the pollutants are highly correlated, which is sometimes the case for environmental mixtures. Several preprocessing steps could be taken to deal with the correlation issue in an informal way. For instance, one could select a single pollutant to be a representative, or tracer, of a highly correlated group of pollutants; or one could use some other dimension-reduction technique, such as principal components analysis, to reduce the dimension of the multipollutant exposure.

In this project, we pursued an extension of the component-wise variable selection that introduces a hierarchical, or multistep, approach to variable selection. Suppose pollutants can be partitioned into groups. The groups may be defined by high correlations, by external knowledge such as the source category of each component, or by another common characteristic. In our initial explorations into this approach, we started by assuming a relatively simple grouping structure whereby group membership was known and each pollutant belonged to one and only one group. Once group membership was defined, we could include a large number of pollutants (correlated or not) in the kernel function; we then defined a hierarchical variable-selection strategy that first estimated the probability that each group of pollutants should be included in the model, and then assessed whether there was evidence in the data that one of the pollutants in a group drives the group's effect.

OUR ORIGINAL PROPOSAL: MODEL-BASED SUPERVISED CLUSTERING

The BKMR framework we developed in this project differs slightly from the model framework outlined in our original proposal — a supervised clustering approach for assessing the health effects of multiple pollutants. That approach is less restrictive than methods that hinge on the additive linear model (see equation 1) in that it does not make any assumptions about the synergy or lack of synergy among pollutants in inducing health effects. It assumes that each exposure is associated with a latent class (i.e., cluster) indicator $K_i \in \{1, 2,\ldots,K\}$, such that

$$\mathbf{Z}_i \mid K_i = k \sim \mathrm{N}\left(\mu_k^Z, \Sigma_k\right). \quad (2)$$

Conditional on an exposure $\mathbf{Z}_i$ falling within cluster k, this approach assumes that the health outcome depends on the cluster to which a given exposure occasion belongs:

$$g\left(\mu_i^Y\right) = \alpha + \gamma_k + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}. \quad (3)$$

Fitting the cluster model (equation 2) to the exposure data and fitting the regression model (equation 3) to the health outcome can be performed either sequentially (referred to as unsupervised clustering of the exposure data) or jointly (referred to as supervised clustering). In our original formulation of the problem we proposed to fit the models jointly and supervise the clustering of the exposure data by a given health outcome. However, since we submitted our proposal, the initial work on this project and parallel work on other projects motivated us to set

aside the supervised clustering approach originally proposed and adopt the BKMR approach we report here.

Specifically, in other work, members of this research team explored a framework for unsupervised clustering of multipollutant data based on K-means clustering (Austin et al. 2012). That work showed that it is possible to use diagnostic ratios of pollutants and back-trajectory analysis of daily air masses to attach physical meaning to the resulting clusters to identify the types of pollutant sources that may contribute to measurements at a monitoring site and the meteorologic conditions that govern the transport and transformation of the emitted pollutants. Therefore, this unsupervised approach can identify distinct pollution patterns at a given site, and the days with common physicochemical properties and meteorologic conditions can then be separately described and investigated. This approach appears to work well in health effects studies involving very large administrative databases and in large cohort studies in which the sample size is sufficiently large to produce large clusters (Zanobetti et al. 2014). In small studies, such as toxicologic studies, because the clustering model effectively categorizes the multivariate exposure distribution, a clustering approach suffers from a loss of power if there is, in fact, an association between a health outcome and pollutant concentrations. This loss of power is analogous to what would occur if one were to discretize a univariate exposure in the presence of an exposure–response relationship. Such loss of power is greater when significant within-cluster variability in the pollutant concentrations exists; thus one is effectively throwing away that variability in exposure when entering cluster into the health model as a predictor.

Early work in this project showed such within-cluster variability to be present in the CAPs exposures conducted as part of the toxicologic study of canines (described later). Figure 1 shows results from an unsupervised clustering of PM constituent data recorded over a period of 100 days by the Harvard Ambient Particle Concentrator at the Harvard Chan School (Bartoli et al. 2009). The data show significant within-cluster variability for many of the constituents. Finally, assessing the importance of any given pollutant in explaining differences in a health outcome across clusters is typically informal; investigators describe the characteristics of the pollution profiles in the cluster that is identified as being associated with a particular health outcome.

Therefore, we ultimately developed an alternative supervised modeling approach that, after adjusting for relevant confounding factors, assumes that the outcome varies continuously as a function of pollution composition. This approach simultaneously achieved all of the goals we described for a method: that is, a method that (1) avoids stringent assumptions of linearity and additivity for the functional form of the exposure–response relationship, as accomplished by regression trees, random forests, and other clustering-based approaches; (2) helps identify the pollutants within the mixture that are responsible for the observed health effects of the overall mixture; (3) estimates the functional form of the (potentially multidimensional) exposure–response curve; and (4) instead of simply presenting a partition of the predictor space corresponding to extreme values of the health endpoints, allows us to quantify the strength of any associations (i.e., perform inference) observed between individual pollutants and health.

KERNEL MACHINE REGRESSION

For the remainder of this work, for concreteness we assume that the health outcome Y_i is a normally distributed continuous random variable. We first outline the general KMR framework, and then consider a Bayesian treatment of the KMR model that includes a variable-selection procedure that yields inferences (the evidence provided by the data) that a given pollutant plays a role in the overall health effect of the mixture.

MODEL DEFINITION

For each subject $i = 1,\ldots,n$, we assume

$$Y_i = h\left(z_{i1},\ldots,z_{iM}\right) + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, \tag{4}$$

where Y_i is a health endpoint, $\mathbf{z}_i = (z_{i1},\ldots,z_{iM})^{\mathsf{T}}$ is a vector of M pollutant concentrations, and $\mathbf{x}_i$ contains a set of potential confounders. In this formulation, $h(\cdot)$ is an unknown function to be estimated either parametrically or nonparametrically, $\boldsymbol{\beta}$ contains the unknown but estimable effects of the confounders, and $\varepsilon_i \overset{iid}{\sim} \mathrm{N}\left(0,\sigma^2\right)$.

A kernel machine representation of the h term reflecting the association between the pollution mixture and health assumes that h: $\mathbf{R}^M \rightarrow \mathbf{R}$ resides in a function space H_K with a positive semidefinite reproducing kernel K: $\mathbf{R}^M \times \mathbf{R}^M \rightarrow \mathbf{R}$. A kernel function $K(\mathbf{z},\mathbf{z}')$ has two arguments: $\mathbf{z}$, which represents the $M \times 1$ covariate vector for one subject, and $\mathbf{z}'$, which represents the $M \times 1$ covariate vector for a second subject. In essence, $K(\mathbf{z},\mathbf{z}')$ measures the similarity, or distance, between two subjects' exposure covariate vectors.

There are two ways to characterize h. One can use a basis-function representation within regression equation (4), termed the primal form, with $h(\mathbf{z}) = \sum_{l=1}^{L}\phi_l(\mathbf{z})\eta_l$, for some set of basis functions $\{\boldsymbol{\phi}_l\}_{l=1}^{L}$ and coefficients $\{\eta_l\}_{l=1}^{L}$. Alternatively, one can represent h using a positive

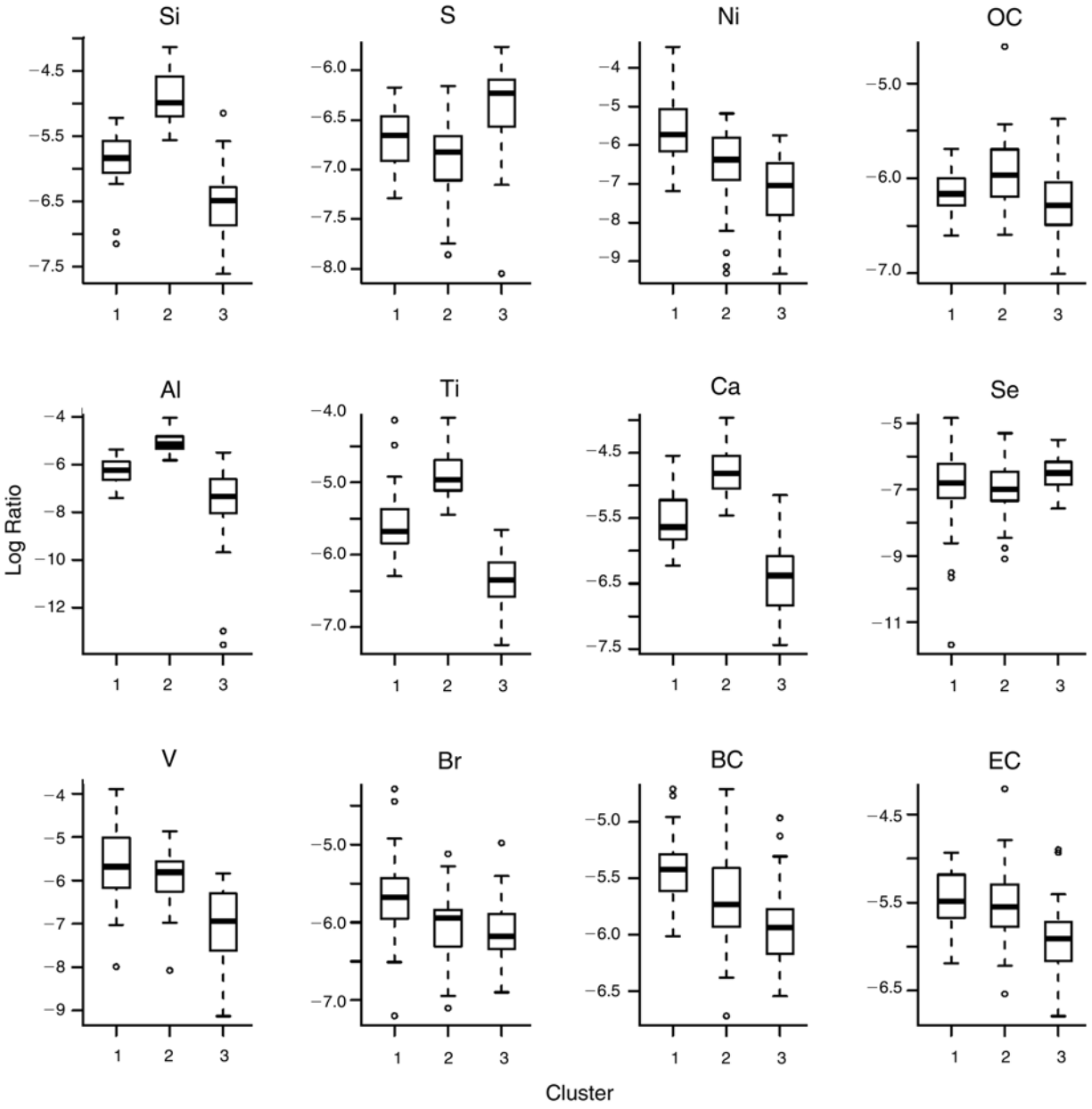


Figure 1. Between- and within-cluster variability of the Harvard Chan School CAPs composition data.

definite kernel function $K(\cdot,\cdot)$, termed the dual form, with $h(\mathbf{z}) = \sum_{i=1}^{n} K(\mathbf{z}_i, \mathbf{z})\alpha_i$ for some set of coefficients $\{\alpha_i\}_{i=1}^{n}$. The Mercer theorem (Cristianini and Shawe-Taylor 2000) established that a kernel function $K(\cdot,\cdot)$ used in the dual form for h implicitly specifies a unique function spanned by a particular set of orthogonal basis functions in the primal representation of h. Examples of this correspondence include:

- Linear kernel: $K(\mathbf{z},\mathbf{z}') = 1 + z_1 z_1' + \ldots + z_M z_M'$.
  - Basis representation: $\phi(\mathbf{z}) = [z_1, z_2, \ldots, z_M]$.
- Quadratic kernel: $K(\mathbf{z},\mathbf{z}') = \left(1 + z_1 z_1' + \ldots + z_M z_M'\right)^2$
  - Basis representation: $\phi(\mathbf{z}) = \left[z_1,\ldots,z_M, z_1^2,\ldots,z_M^2, z_1 z_2,\ldots, z_{M-1} z_M\right]$.

- Gaussian kernel: $K(\mathbf{z},\mathbf{z}') = \exp\left\{-\frac{1}{\rho}\sum_{m=1}^{M}(z_m - z'_m)^2\right\}$.
  - Basis representation: $\phi(\mathbf{z})$ is space spanned by radial basis functions.

Operationally, Liu and colleagues (2007) showed that one can fit the kernel machine regression to data using the mixed-model representation

$$y_i = h_i + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1,\ldots,n, \tag{5}$$

where $\mathbf{h} = (h_1,\ldots,h_n)^{\mathsf{T}} \sim MVN(\mathbf{0}, \tau\mathbf{K})$, $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)^{\mathsf{T}} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$, and $\mathbf{h} \perp \varepsilon$. The kernel matrix $\mathbf{K}$ has (i,j)-element $K(\mathbf{z}_i,\mathbf{z}_j)$. The majority of work on this mixed model has focused on fitting it using frequentist methods, which allows one to conduct tests of the overall effect of the mixture, H_0: $h = 0$. Maity and Lin (2011) generalized this approach to test whether a given pollutant of the $\mathbf{z}_i$ vector contributes to the overall effect.

Clearly, fitting a KMR model relies on specifying which kernel function to use. In this project we mainly focus on the Gaussian kernel, which flexibly captures a wide range of underlying functional forms for $h(\cdot)$, though our approach for estimating the health effects of environmental mixtures is applicable to a broad choice of kernels. To provide some intuition for KMR using the Gaussian kernel, consider the effect on health of exposure to the profile $\mathbf{z}_i$ for the ith person, given by $h_i = h(\mathbf{z}_i)$. Under model (5), we assume $\text{cor}(h_i,h_j) = \exp\left\{-\frac{1}{\rho}\sum_{m=1}^{M}\left(z_{im} - z_{jm}\right)^2\right\}$, which implies that two subjects with similar exposures ($\mathbf{z}_i$ is close to $\mathbf{z}_j$) will have more similar risks (h_i will be close to h_j).

COMPONENT-WISE VARIABLE SELECTION

To allow for variable selection within a Bayesian paradigm for KMR, we expand the formulation in the previous section to include an indicator variable (δ_m) for each pollutant concentration z_m. More specifically, we define the augmented Gaussian kernel function to be

$$K(\mathbf{z},\mathbf{z}';\mathbf{r}) = \exp\left\{-\sum_{m=1}^{M} r_m(z_m - z'_m)^2\right\}, \tag{6}$$

where $\mathbf{r} = (r_1,\ldots,r_M)^{\mathsf{T}}$, and we define $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ to be the $n \times n$ matrix with (i,j)-element equal to $K(\mathbf{z}_i,\mathbf{z}_j;\mathbf{r})$. We assume a slab and spike prior on the auxiliary parameters

$$r_m \mid \delta_m \sim \delta_m\,\text{Gamma}(a_r,b_r) + (1 - \delta_m)P_0,\ m = 1,\ldots,M, \text{ and}$$
$$\delta_m \sim \text{Bernoulli}(\pi), \tag{7}$$

where P_0 denotes the density with point mass at 0. This mixture representation is analogous to the Bayesian variable-selection model for multiple regression problems (George and McCulloch 1993) and has been applied in Gaussian process models (Linkletter et al. 2006; Savitsky et al. 2011). The posterior mean of the indicator δ_m given the data, $\pi_m = Pr(\delta_m = 1 \mid \mathbf{y}, \mathbf{z}, \mathbf{x})$, has the natural interpretation as the posterior probability that pollutant m is an important component of the mixture, also referred to as the posterior inclusion probability (PIP) of pollutant m. Other kernel functions may be augmented in a similar way. For example, the quadratic kernel may be expanded as

$$K(\mathbf{z},\mathbf{z}';\mathbf{r}) = \left(1 + r_1 z_1 z'_1 + \ldots + r_m z_M z'_M\right)^2.$$

In this framework we view a given inclusion probability as an indicator of the importance of a given pollutant within the multipollutant mixture. Our experience with many data sets suggests that, like the variable-importance scores provided by a random forest analysis, the absolute magnitudes of these inclusion probabilities depend on the tuning parameters chosen for a given analysis, which in this setting include the prior parameters a_r and b_r in the mixture prior shown in equation (7). (Later, we provide an example of this sensitivity in the analysis of exposure to $PM_{2.5}$ with different compositions and blood pressure and heart rate.) Therefore, we take the strategy analogous to a random forest analysis in which we use the inclusion probabilities to order the importance of each pollutant in predicting health outcomes, and then follow up by evaluating the exposure–response relationships for those pollutants ranked most important.

HIERARCHICAL VARIABLE SELECTION

In situations where concentrations of the pollutants in the mixture are highly correlated, the above formulation that treats pollutants exchangeably may fail because the data may not be able to distinguish among the correlated pollutants. We therefore also propose a hierarchical variable-selection approach that incorporates knowledge of the structure of the mixture into the model.

Suppose the pollutants $z_1,\ldots,z_M$ can be partitioned, using prior knowledge, into groups S_g ($g = 1,\ldots,G$). For example, a wealth of information about air pollution source categories allows for pollution constituents to be grouped such that within-group correlation is high and across-group correlation is moderate to low. We then assume that the indicator variables from the slab and spike prior in equation (7) are distributed as

$$\boldsymbol{\delta}_{S_g} \mid \omega_g \sim \text{Multinomial}\left(\omega_g, \boldsymbol{\pi}_{S_g}\right), \quad g = 1,\ldots,G, \text{ and} \tag{8}$$
$$\omega_g \sim \text{Bernoulli}(\pi),$$

where $\boldsymbol{\delta}_{S_g} = (\delta_m)_{z_m \in S_g}$ is the vector of indicator variables and $\boldsymbol{\pi}_{S_g}$ is the corresponding vector of prior probabilities for the pollutants z_m in group S_g. This approach allows, at most, a single pollutant from a group of highly correlated

pollutants to enter into the model at a time. Although this assumes that two pollutants from the same group do not have independent or interactive effects on the health outcome, in the setting of high within-group correlation such effects would not be identifiable by any model.

PRIOR SPECIFICATION

To complete the model specification, we must specify prior distributions for the regression parameters $\boldsymbol{\beta}$ and σ^2, the variance component τ, and the prior inclusion probability π, and we must select values for the hyperparameters a_r and b_r of the variable selection prior in equation (7).

For the regression parameters, we assume $\boldsymbol{\beta} \sim 1$ (flat prior) and $\sigma^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$, where we set the shape parameter a_σ and scale parameter b_σ to each be 0.001. It is convenient to parameterize BKMR by $\lambda \equiv \tau\sigma^{-2}$, and we assume a Gamma prior distribution for λ with mean and variance each set to 100 (let a_λ, b_λ denote corresponding shape and rate parameters). We assume the prior probability (π) that a pollutant is included in the model comes from a beta distribution; that is, $\pi \sim \text{Beta}(a_\pi, b_\pi)$. In our implementations, we set $a_\pi = b_\pi = 1$ (Uniform).

Before discussing the choice of values for a_r and b_r, first note that without variable selection, KMR is analogous to a spatial regression model in which $\mathbf{z}_i$ corresponds to spatial locations, and the parameter ρ in the Gaussian kernel corresponds to the range parameter (Mikl et al. 2008). It is well known that the spatial range parameter is only weakly identified in this setting (Banerjee et al. 2008) and so it is necessary to specify an informative prior distribution for ρ (and hence for $1/\rho$). In addition, Bayesian variable selection can be highly sensitive to the specification of the mixture prior. Thus, care must be taken to select the hyperparameters a_r and b_r in equation (7).

To select these hyperparameters, it is helpful to further consider the interpretation of the r_m in our augmented Gaussian kernel (equation 6). As noted earlier, the model assumes that for each pollutant z_m, two individuals with more similar exposure levels will have more highly correlated health outcomes, and the correlation will decay as the difference between exposure levels increases. The parameter r_m controls the rate of decay in correlation with this difference: smaller values of r_m correspond to a slower decay, which results in a greater degree of smoothing in the estimated h as a function of z_m, whereas larger values of r_m correspond to a faster decay resulting in a more oscillatory (i.e., less smooth) h function of z_m (see Figure 2). For automatic hyperparameter selection, we therefore recommend selecting a prior for r_m that captures a range of values of the anticipated smoothness of h as a function of each z_m based on expert knowledge. For example, in Figure 2 the smallest value of r shown (0.01) corresponds to a very smooth exposure–response function h, and the largest value of r (0.54) corresponds to an exposure–response function that is more oscillatory than those expected in many environmental health studies. To translate this knowledge into a prior distribution, for example, we might select values of a_r and b_r such that $\Pr(r < 0.01) = \Pr(r > 0.54) = \alpha/2$ for α small. In some cases, one might have prior evidence that the given health effects of a mixture are linear in the pollutant concentrations. In such a case, one might want to set the hyperparameters to induce more smoothing, which corresponds to smaller values of a_r and b_r. In such cases, one can obtain a preliminary estimate of ρ from the fit of the frequentist model of Liu and associates (2007) and take as the value of these hyperparameters a fraction of this parameter estimate. (We take this approach in our analysis of the Harvard Chan School canine toxicologic study data shown later.) Note that the data should be transformed so that each z_m is on the same scale to ensure that the r_m values correspond to the same degree of smoothness in h as a function of each z_m. Finally, for the hierarchical variable-selection approach, for $\boldsymbol{\pi}_{S_g}$ we assumed that each pollutant of the same group was equally likely to be included in the model.

SIMULATION STUDIES

To investigate the properties of the proposed methods, we conducted three simulation studies designed to evaluate the ability of BKMR to estimate environmental mixtures responsible for health effects under potentially complex but plausible exposure–response relationships. An attractive feature of our simulation studies is that we used actual exposure data rather than simulated values, meaning that we sampled directly from empirical data to generate data sets for the simulations. This approach has the advantage that it evaluates the performance of BKMR and several other models under realistic joint distributions of the multipollutant exposure.

Our primary data set was from the Harvard Chan School Boston Supersite for monitoring multiple air pollutants (located at the Francis A. Countway Library of Medicine on the Harvard Medical School campus). This is a unique data set in that PM composition was recorded most days from 1998 through 2011, resulting in data for 4,853 days. Also, because the same multipollutant data were used for the MOBILIZE analyses, the simulation studies are as relevant as possible to the epidemiologic analyses conducted in that project.

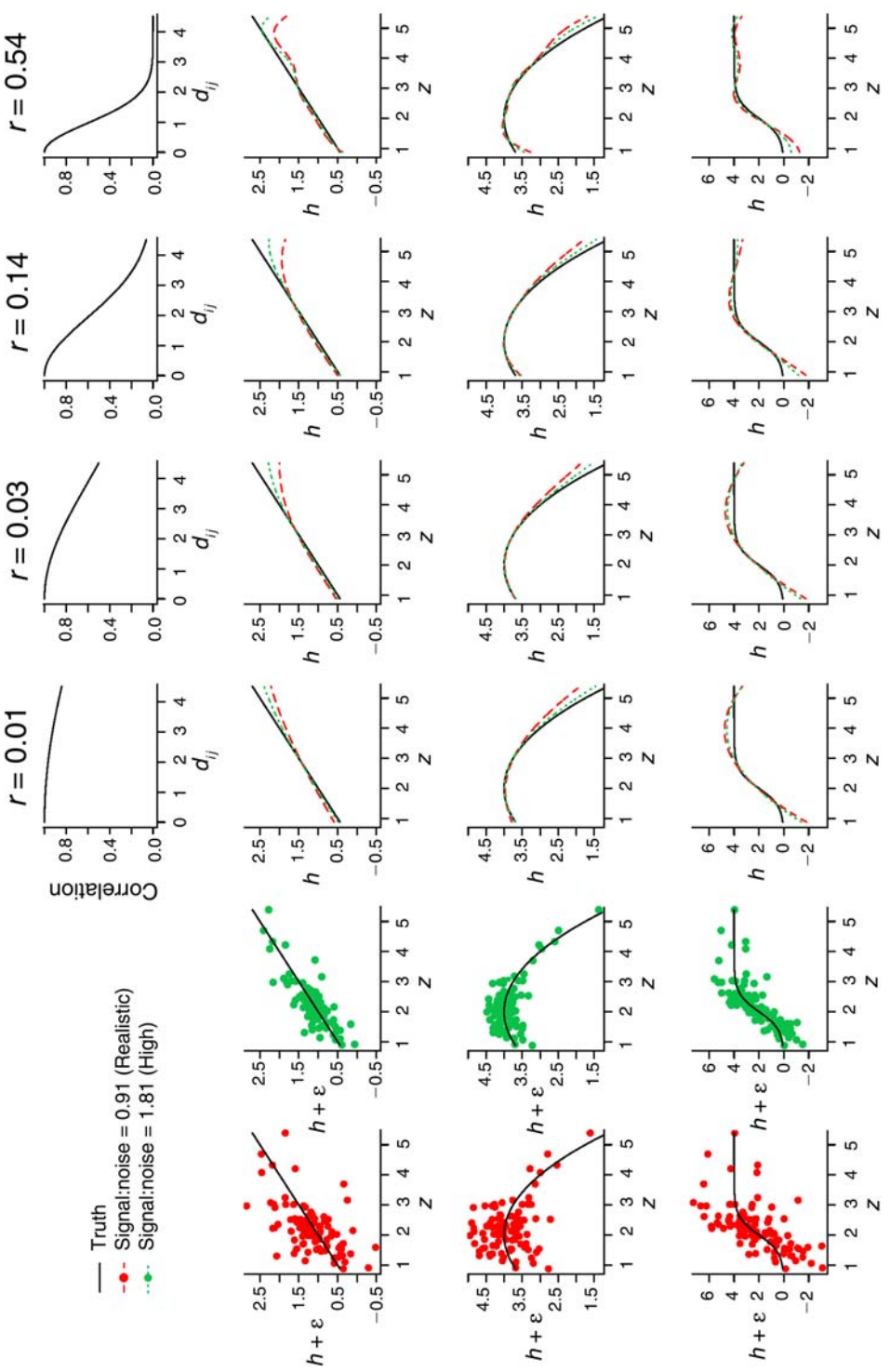


Figure 2. Interpretation of the variables r_m in the augmented Gaussian kernel for variable selection (equation 6). The top row of panels shows the correlation r in the health outcomes in two individuals as a function of their difference in exposure to pollutant z (taken to be sampled values of Al from the 1998–2011 Harvard Chan School Boston Supersite multipollutant data set), given by $\mathrm{cor}(h_i, h_j) \propto \exp\left\{-r \cdot d_{ij}^2\right\}$ for four different values of r, where $d_{ij} = |z_i - z_j|$. Rows 2 through 4 show, for different examples of a univariate $h(z)$, the estimated h from fitting the KMR in equation (5) with Gaussian kernel to two simulated data sets (realistic and high signal-to-noise ratios) for $\rho = 1/r$ fixed to each of the four different values for r. These values of r correspond to a decay in the correlation by 50% over different fractions, q, of the range of the data: $q = 2, 1, ½, ¼$. $h + \varepsilon$ represents the observed data as a function of z.

Simulation Study 1 assumed that one or more of the $PM_{2.5}$ constituents exhibits an association with health effects; this assumption reduced the dimensionality of the highly correlated x-ray fluorescence (XRF) elemental concentrations by selecting a subset of concentrations to represent a group of correlated constituents that are not themselves highly correlated. If we assume two pollutants are associated with effects on health, we assume interactions exist between the two pollutants.

Simulation Study 2 compared the performance of (a) BKMR using component-wise variable selection with (b) BKMR using hierarchical variable selection applied to a large number of pollutants, some highly correlated. In these first two simulation studies, we used each BKMR approach to test the performance of the frequentist approach to KMR variable selection proposed by Maity and Lin (2011).

In Simulation Study 3 we assumed that the constituent concentrations were generated from a source apportionment model and that pollutants from a particular source category were associated with adverse health effects.

SIMULATION STUDY 1. COMPONENT-WISE VARIABLE SELECTION FOR A MODERATE NUMBER OF POLLUTANTS

Simulation Setup

For each simulation scenario we generated 100 data sets of 100 observations each, $\{y_i, x_i, \mathbf{z}_i\}_{i=1}^{100}$, where $\mathbf{z}_i = (z_{i1},\ldots,z_{iM})$ represents concentrations of M pollutants; x_i is a confounder (associated with z_{i1}) generated as $x_i \sim \text{N}(3\cos z_{i1}, 2)$; and the health outcomes $y_i \sim \text{N}[\boldsymbol{\beta} x_i + h(z_{i1},\ldots,z_{iQ}), \boldsymbol{\sigma}^2]$. We assumed that a health outcome will depend on only a subset of $Q < M$ of the available exposure variables, and the value of Q will depend on the particular simulation scenario.

We analyzed three different h functions:

- h_1: a univariate nonlinear exposure–response relationship that depends only on z_{i1};
- h_2: a linear exposure–response function with main effects of z_{i1} and z_{i2} and an interaction between these two exposures; and
- h_3: a nonlinear exposure–response function of both z_{i1} and z_{i2} with a synergistic interaction between these two exposures.

For the exposure data, we chose two existing exposure data sets, one for $M = 3$ pollutants and another for $M = 9$ pollutants, and generated the simulated data sets by sampling from these exposure data. For the $M = 3$ simulation we generated data sets based on the empirical distributions of the concentrations of As, Mn, and Pb from a study on the neurotoxicity of in utero exposure to metal mixtures conducted in Bangladesh as part of the Harvard Superfund Research Program (Bobb et al. 2014). The concentrations were based on measurements of biomarkers collected in cord blood. Exposure to metal mixtures and neurotoxicity in children is a setting in which both nonlinearity and interaction in the exposure–response relationship have been reported (Claus Henn et al. 2012; Bobb et al. 2014). We constructed each exposure data set $\{\mathbf{z}_i\}_{i=1}^{100}$ by sampling 100 rows from the metal mixtures data set.

To evaluate the BKMR methods in air pollution epidemiologic settings, in which a large number of pollutants are analyzed, for the $M = 9$ simulation we used the Harvard Chan School Boston Supersite multipollutant data set that consisted of daily measurements of ambient air pollutants from 1998 through 2011.

XRF is used to measure the concentrations of many elements. However, it is well known that XRF measures some elements very well, others moderately well, and some with high levels of error. For instance, experience from our Harvard Clean Air Research Center is that Cu, Zn, Ni, V, S, Ti, Ca, Mg, K, Cl, Na, Al, and Si are measured well, and Br, Sr, Pb, and Mn are measured reasonably well. However, other elements, such as Cr, Se, As, Ba, Sn, Ce, Co, Cs, Sb, Zr, Ag, In, W, Hg, Hf, Sm, Eu, Y, Te, Tl, Cd, and other exotic elements are not measured well. Therefore, for the $M = 9$ simulation studies, we started by considering the 17 elements that are measured well or reasonably well by XRF. We also included BC, as measured by aethalometer, because it is a well-known marker of traffic-related ambient particles. Because high correlations can negatively affect the performance of any component-wise variable-selection procedure, in Simulation Study 1 we selected 9 of the 18 constituents that represent major pollution sources but are not too highly correlated: Al, S, Ni, BC, Cu, Zn, Mg, K, and Cl. The daily multipollutant data were standardized by subtracting the median and dividing by the interquartile range (IQR); days with outlier values (defined as values greater than 5 IQR away from the median) were removed.

Figure 3 presents the correlation matrix for the subset of constituents used in Simulation Study 1. For this $M = 9$ scenario we generated each exposure data set $\{\mathbf{z}_i\}_{i=1}^{100}$ by sampling (with replacement) 100 rows of the Harvard Chan School Boston Supersite multipollutant data set. In this way, the distributions of (and correlations among) the multiple pollutants in our simulation study preserved the underlying structure of realistic data settings.

To specify values for σ^2, we considered two signal-to-noise ratios. The first ratio we set to be equivalent to the signal-to-noise ratio (0.91) estimated from the metal mixtures data set (referred to as the realistic signal-to-noise scenario). In order to evaluate the performance of the methods in a more idealistic scenario in which we have a lot of information in the data, we set a second ratio to be twice (1.82) the ratio used in the realistic scenario (high signal-to-noise scenario). We tested the methods under the high signal-to-noise ratio as a type of positive control, in that the method should perform as expected when the exposure–response association is strong and there is little residual error in the response. To get a sense of the amount of noise implied by the realistic and high signal-to-noise ratios, data generated based on these ratios are shown for different univariate exposure–response functions h in Figure 2.

To each of the 100 simulated data sets under each of the 12 data-generating scenarios (two exposure data-generating mechanisms [$M = 3$, $M = 9$]; three exposure–response functions [h_1, h_2, h_3]; and two signal-to-noise ratios [realistic, high]), we fit five different models. To start, we fit a BKMR model that included all M of the available exposure variables (note that only one or two are truly associated with the outcome), both without variable selection (BK-MR) and with variable selection (BKMRvs). To specify a prior distribution for r_m, we followed the automatic selection procedure discussed in the Prior Specification section. Specifically, for the multipollutant exposure data ($M = 9$), we selected hyperparameters for the Gamma prior for r_m such that $\Pr(r_m < 0.01) = \Pr(r_m > 0.54) = 0.005$; namely, $a_r = 2.1$ and $b_r = 14.0$. The rationale for the bounds 0.01 and 0.54 are illustrated in Figure 2, which shows examples of simulated data sets based on Al pollution data and different univariate h functions. For the metal mixtures exposure data ($M = 3$), because all exposure data were standardized by subtracting the median and dividing by the IQR, the range of the metal concentrations was similar to the range of the $M = 9$ constituent concentrations. Therefore, we applied this same prior distribution to both the $M = 3$ and $M = 9$ data sets. We ran each Markov chain Monte Carlo (MCMC) for 10,000 iterations and kept the last 8,000 samples.

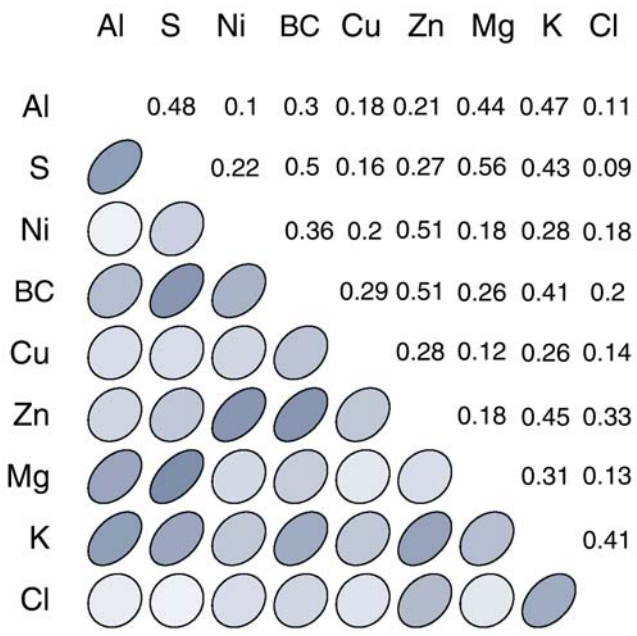


Figure 3. Correlation matrix of multipollutant data set used for the $M = 9$ constituents scenario in Simulation Study 1. The shading and shapes indicate the strength of the correlation between a pair of constituents, with darker and more oval shapes reflecting higher correlations. These provide a visual representation of the correlation structure present in the numerical values shown in the upper triangle of the matrix.

Next, we fit KMR using a frequentist approach (Liu et al. 2007), again without (KMR) and with (KMRvs) variable selection. The standard errors for the estimated exposure–response $\hat{h}$ from this approach were obtained using the best linear unbiased prediction representation of $\hat{h}$ in this model, as given in equation (15) in Liu and associates (2007). To conduct the variable selection, we applied the Garrote test from Maity and Lin (2011) to each pollutant $z_m\,(m = 1,\ldots,M)$ one at a time, and then re-fit the KMR including only those pollutants for which the Garrote test yielded $P < 0.05$. We evaluated the power and type I error rates for this frequentist approach.

Finally, to quantify the optimal performance achievable if the true form of the h function were known, we applied an oracle model, which is the model that is best suited for a given h function. For h_1 we fit a GAM including only z_1, modeled using penalized splines and a thin-plate regression basis (Wood 2006). For h_2, we fit a linear model with $z_1, z_2,$ and the interaction term $z_1 z_2$. For h_3 we fit a GAM including a bivariate function of z_1 and z_2.

Results for Estimating the Exposure–Response Function

We first evaluated the ability of each of the five approaches to estimate the subject-specific mixture effects $h_i = h(\mathbf{z}_i)$ for each of the simulated data sets by regressing the estimated $\hat{h}_i$ on the true h_i and reporting the average intercept, slope, and R^2 from this regression across simulation repetitions.

Under the high signal-to-noise ratio (Table 1), both approaches with variable selection (KMRvs and BKMRvs) performed comparably to the oracle model (and outperformed the corresponding models without variable selection) in estimating the h_i across the three true $h(\cdot)$ and for both the metals ($M = 3$) and multipollutant ($M = 9$) exposure data.

Table 1. Performance of Estimated $h_i = h(\mathbf{z}_i)$ Across 100 Simulated Datasets Under a High Signal-to-Noise Ratio

	Regression of $\hat{h}$ on h			Uncertainty[a]		
	Intercept	Slope	R^2	SD ($\hat{h} \mid h$)	SE	Coverage
***M* = 3 Metals**						
$h_1(\mathbf{z})$						
Oracle	0.02	0.98	0.98	0.19	0.17	0.93
KMR	0.07	0.93	0.95	0.30	0.24	0.88
KMRvs	0.02	0.97	0.98	0.19	0.17	0.91
BKMR	0.07	0.93	0.95	0.29	0.26	0.91
BKMRvs	0.02	0.98	0.99	0.19	0.17	0.92
$h_2(\mathbf{z})$						
Oracle	−0.01	1.00	0.99	0.07	0.06	0.94
KMR	0.02	0.98	0.98	0.09	0.08	0.93
KMRvs	0.01	0.99	0.99	0.08	0.07	0.93
BKMR	0.02	0.98	0.98	0.09	0.09	0.95
BKMRvs	0.01	0.99	0.98	0.09	0.09	0.95
$h_3(\mathbf{z})$						
Oracle	0.02	0.95	0.96	0.17	0.17	0.96
KMR	0.03	0.93	0.95	0.19	0.16	0.92
KMRvs	0.03	0.95	0.96	0.17	0.14	0.91
BKMR	0.03	0.94	0.95	0.19	0.17	0.94
BKMRvs	0.02	0.95	0.97	0.16	0.15	0.94
***M* = 9 Constituents**						
$h_1(\mathbf{z})$						
Oracle	0.02	0.99	0.99	0.19	0.17	0.92
KMR	0.26	0.84	0.89	0.46	0.34	0.82
KMRvs	0.04	0.97	0.97	0.25	0.17	0.85
BKMR	0.23	0.85	0.90	0.45	0.38	0.89
BKMRvs	0.03	0.98	0.98	0.21	0.18	0.91
$h_2(\mathbf{z})$						
Oracle	−0.01	1.01	0.99	0.11	0.10	0.94
KMR	0.05	0.97	0.96	0.19	0.18	0.94
KMRvs	0.01	0.99	0.98	0.15	0.11	0.92
BKMR	0.05	0.97	0.96	0.20	0.20	0.96
BKMRvs	0.01	0.99	0.98	0.15	0.14	0.96
$h_3(\mathbf{z})$						
Oracle	0.02	0.98	0.97	0.25	0.23	0.95
KMR	0.09	0.91	0.92	0.39	0.33	0.90
KMRvs	0.03	0.97	0.96	0.29	0.21	0.86
BKMR	0.07	0.92	0.92	0.39	0.36	0.93
BKMRvs	0.02	0.98	0.97	0.27	0.24	0.94

[a] SD denotes the empirical standard deviation of the estimated $\hat{h}$; SE denotes the estimated standard error, which in a Bayesian analysis is the posterior standard deviation of the $\hat{h}_i$; and coverage is the proportion of times that the 95% confidence intervals (for oracle, KMR, and KMRvs) or posterior credible intervals (for BKMR and BKMRvs) covered the true h_i.

Table 2. Performance of Estimated $h_i = h(\mathbf{z}_i)$ Across 100 Simulated Datasets Under a Realistic Signal-to-Noise Ratio

	Regression of $\hat{h}$ on h			Uncertainty[a]		
	Intercept	Slope	R^2	SD ($\hat{h} \mid h$)	SE	Coverage
***M* = 3 Metals**						
$h_1(\mathbf{z})$						
Oracle	0.04	0.95	0.95	0.35	0.31	0.92
KMR	0.17	0.82	0.88	0.50	0.40	0.87
KMRvs	0.06	0.92	0.95	0.35	0.30	0.91
BKMR	0.15	0.85	0.89	0.47	0.44	0.93
BKMRvs	0.05	0.94	0.96	0.33	0.34	0.95
$h_2(\mathbf{z})$						
Oracle	−0.02	1.01	0.97	0.14	0.13	0.94
KMR	0.08	0.93	0.95	0.16	0.15	0.94
KMRvs	0.07	0.94	0.94	0.16	0.13	0.91
BKMR	0.05	0.95	0.94	0.17	0.17	0.96
BKMRvs	0.04	0.95	0.93	0.17	0.17	0.95
$h_3(\mathbf{z})$						
Oracle	0.05	0.90	0.89	0.29	0.28	0.95
KMR	0.10	0.81	0.84	0.33	0.28	0.91
KMRvs	0.10	0.80	0.81	0.35	0.23	0.85
BKMR	0.08	0.84	0.86	0.32	0.30	0.95
BKMRvs	0.07	0.87	0.88	0.29	0.28	0.95
***M* = 9 Constituents**						
$h_1(\mathbf{z})$						
Oracle	0.07	0.96	0.95	0.36	0.31	0.91
KMR	0.49	0.70	0.77	0.69	0.51	0.82
KMRvs	0.12	0.93	0.93	0.41	0.30	0.86
BKMR	0.41	0.74	0.78	0.67	0.60	0.92
BKMRvs	0.07	0.96	0.96	0.35	0.36	0.95
$h_2(\mathbf{z})$						
Oracle	−0.03	1.01	0.97	0.22	0.20	0.94
KMR	0.16	0.90	0.90	0.33	0.31	0.95
KMRvs	0.07	0.96	0.94	0.28	0.20	0.91
BKMR	0.13	0.92	0.89	0.35	0.36	0.97
BKMRvs	0.12	0.92	0.90	0.33	0.31	0.95
$h_3(\mathbf{z})$						
Oracle	0.03	0.95	0.92	0.43	0.38	0.93
KMR	0.17	0.82	0.84	0.58	0.50	0.91
KMRvs	0.12	0.87	0.85	0.56	0.32	0.80
BKMR	0.13	0.85	0.84	0.58	0.59	0.96
BKMRvs	0.06	0.92	0.89	0.49	0.48	0.95

[a] SD denotes the empirical standard deviation of the estimated $\hat{h}$; SE denotes the estimated standard error, which in a Bayesian analysis is the posterior standard deviation of the $\hat{h}_i$; and coverage is the proportion of times that the 95% confidence intervals (for oracle, KMR, and KMRvs) or posterior credible intervals (for BKMR and BKMRvs) covered the true h_i.

Compared with the high signal-to-noise scenario (Table 1), under the realistic scenario (Table 2) all approaches performed worse than the oracle models at estimating $h(\cdot)$, although the gains achieved by incorporating variable selection were larger. For example, for $h_1(\cdot)$ and the multipollutant exposure data ($M = 9$), the R^2 increased from 0.78 under BKMR to 0.96 under BKMRvs for the realistic signal-to-noise ratio (Table 2), compared with an increase from 0.90 to 0.98 under the high signal-to-noise ratio (Table 1). The difference between BKMRvs and frequentist KMRvs results also became more apparent under the realistic signal-to-noise scenario (Table 2). For example, the less smooth exposure–response functions $h_1(\cdot)$ and $h_3(\cdot)$ were better estimated by the BKMRvs approach [for $M = 9$, $R^2 = 0.96$ for BKMRvs and $R^2 = 0.93$ for KMRvs for $h_1(\cdot)$, and $R^2 = 0.89$ for BKMRvs and $R^2 = 0.85$ for KMRvs for $h_3(\cdot)$]; and the smoothest (linear) function $h_2(\cdot)$ was better estimated by the frequentist KMR approach (for $M = 9$, $R^2 = 0.94$ for KMRvs and 0.90 for BKMRvs). In this realistic signal-to-noise setting, by comparing R^2 values for BKMRvs for h_1 to those for h_2 and h_3, we also noted a suggestion that the BKMRvs method appeared to perform slightly better when there were fewer pollutants assumed to be causal ($Q = 2$ versus $Q = 1$).

Across all scenarios, the Bayesian approaches were better able to capture the uncertainty in the $\hat{h}_i$ than the corresponding frequentist KMR methods, achieving posterior SD estimates that were close to the empirical SEs and interval coverage closest to the nominal (95%) level (Tables 1 and 2).

Results for Identifying Important Pollutants

We next evaluated whether the proposed KMRvs and BKMRvs approaches could correctly identify which pollutant or pollutants were predictive of the health outcome [i.e., included in $h(\cdot)$]. Figures 4 and 5 show (1) the boxplots for the PIPs under BKMRvs, and (2) the proportion of iterations for which each pollutant was identified as statistically significant under the Maity and Lin Garrote KMR test (2011; values shown beneath the x axes) for the high (Figure 4) and realistic (Figure 5) signal-to-noise scenarios. For the high signal-to-noise ratio, across $h(\cdot)$ functions and for both the metals and multipollutant exposure data sets, BKMRvs correctly assigned high posterior support to the pollutants that were truly predictive of health outcomes and low posterior support to the pollutants that were not (Figure 4). As expected, the ability of BKMRvs to identify the correct pollutants was diminished under the more realistic signal-to-noise scenario (Figure 5), in which more individual PIPs were below 0.6, although the PIPs were generally still distinguishable between causal and non-causal pollutants. Under both high and realistic signal-to-noise scenarios, the frequentist KMRvs demonstrates reasonable power and type I error rates that are generally close to the target 0.05 level, although this error rate is inflated for a few elements in a few h functions (e.g., Cu at 9% and Mg at 12% for high signal-to-noise under h_2 [Figure 4]).

A major factor affecting the ability to correctly identify important pollutants is the correlation structure of the exposure data. For the bivariate functions $h(\cdot)$ and multipollutant exposure data ($M = 9$), because Al and S (the correct pollutants) were moderately correlated ($r = 0.48$), once one of these pollutants was selected to be in the model, insufficient residual information remained in the data to identify the second pollutant for a portion of the simulation iterations. If, instead of Al and S being the correct pollutants, we replaced S with Ni (whose correlation with Al was only 0.1), then the ability of the approach to detect both important pollutants was substantially improved (see Figure 6). On the other hand, for the metals exposure data ($M = 3$; Figure 5), the important pollutants (Pb and Mn) were uncorrelated and so both were able to be identified by the model for most of the simulation iterations.

Similarly, having an unimportant pollutant that is moderately or highly correlated with an important pollutant also challenges variable selection. For example, for the metals exposure data (Figure 5), high posterior support was assigned to As for a portion of simulation iterations, though it was not truly predictive of health. To investigate whether this was due to correlated exposures or to some other feature of the approach, we repeated the simulation and replaced As with a sham exposure variable that had the same mean and standard deviation as As but was independent of Pb and Mn. We found that the median (and IQR) PIP declined from 0.06 (0.10) for As to 0.05 (0.07) for the sham exposure under $h_1(\cdot)$, from 0.17 (0.20) to 0.08 (0.15) under $h_2(\cdot)$, and from 0.16 (0.26) to 0.08 (0.14) under $h_3(\cdot)$, suggesting that the correlated exposures are at least partly responsible for the inflated inclusion probabilities of As.

At the suggestion of the HEI Review Committee, in addition to summarizing the distribution of the PIPs across simulations for the h_2 and h_3 functions in which two pollutants affected health, we also summarized how often these procedures correctly selected none, one, or two of the important pollutants. In order to "select" a pollutant, one must set a threshold value such that the pollutant is selected if its PIP in a given analysis is above that threshold and not selected if not. We summarized the proportion of times important variables were selected for two threshold

choices: either PIP > 0.5 or PIP > 0.8. Table 3 shows the results, which suggest that the BKMR approach does very well in identifying both important pollutants if there are a small number of pollutants (M = 3 metals) under either threshold value. For the more complicated simulation (M = 9 constituents), the methods are generally able to identify at least one of the two pollutants, with the probability of identifying both important pollutants at 70% to 80% if one uses the less stringent threshold of PIP > 0.5.

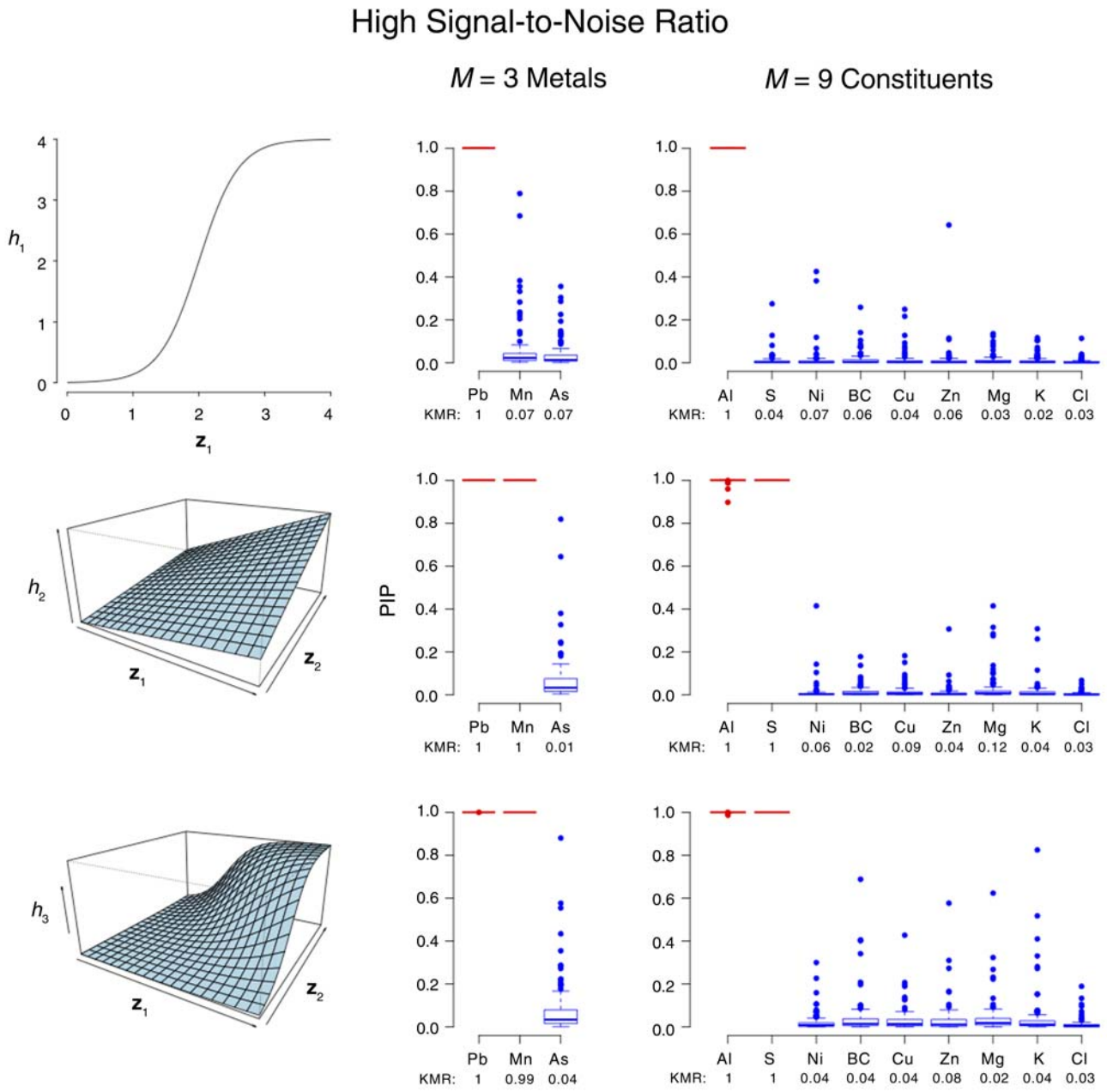


Figure 4. Median (25%, 75%) of the PIPs from BKMRvs under a high signal-to-noise ratio. Calculated across 100 simulated data sets, for each of three true $h(\mathbf{z})$ functions. The vector of exposure data $\mathbf{z}$ was generated based on either M = 3 metals (Pb, Mn, As) or on M = 9 air pollution constituents (Al, S, Ni, BC, Cu, Zn, Mg, K, Cl). The proportion of simulation iterations for which each pollutant had $P < 0.05$ under the frequentist KMR approach of Maity and Lin (2011) is printed below the x axis.

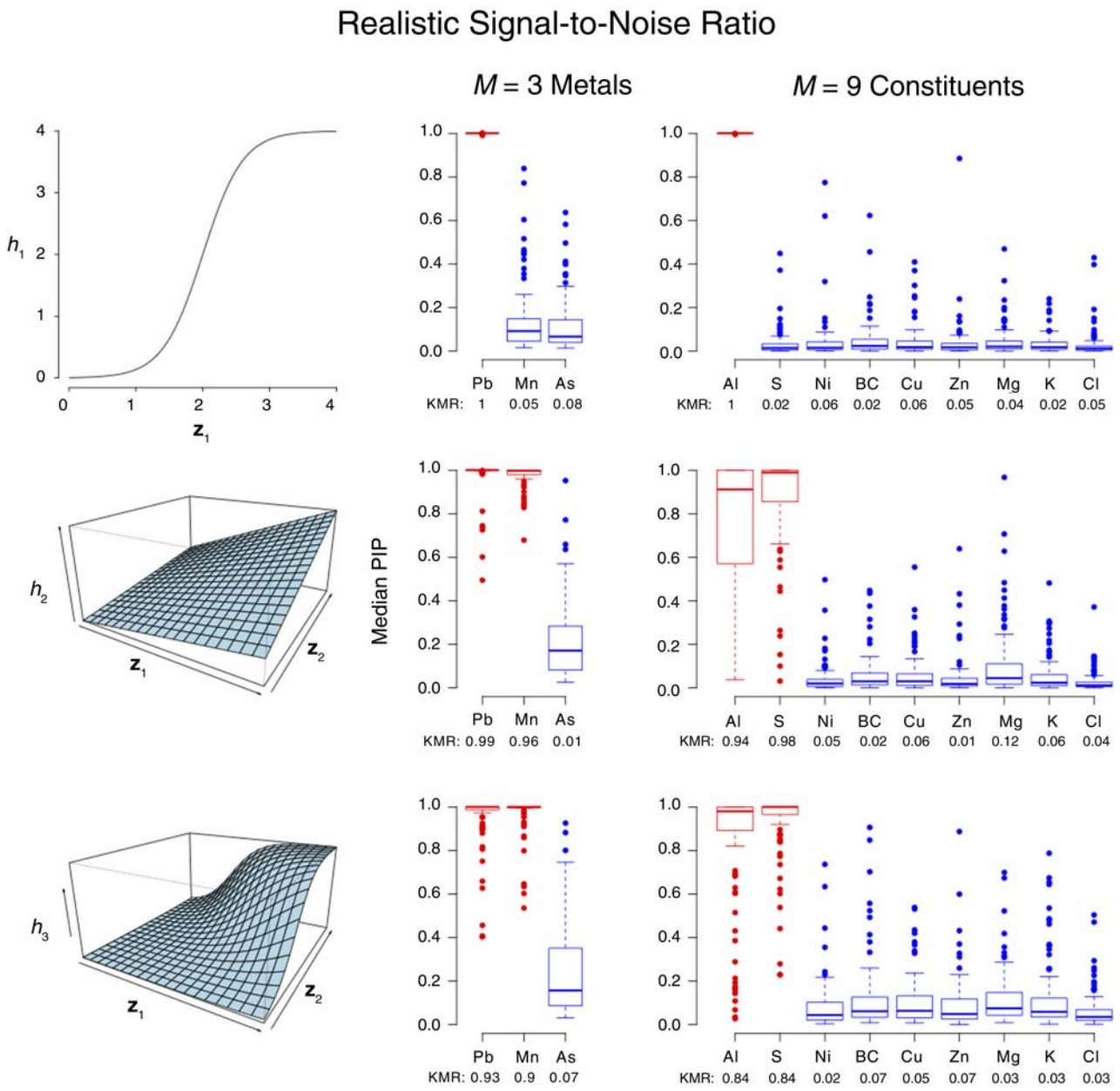


Figure 5. Median (25%, 75%) of the PIPs from BKMRvs under a realistic signal-to-noise ratio. Calculated across 100 simulated data sets, for each of three true $h(\mathbf{z})$ functions. The vector of exposure data $\mathbf{z}$ was generated based on either $M = 3$ metals (Pb, Mn, As) or on $M = 9$ air pollution constituents (Al, S, Ni, BC, Cu, Zn, Mg, K, Cl). The proportion of simulation iterations for which each pollutant had $P < 0.05$ under the frequentist KMR approach of Maity and Lin (2011) is printed below the x axis.

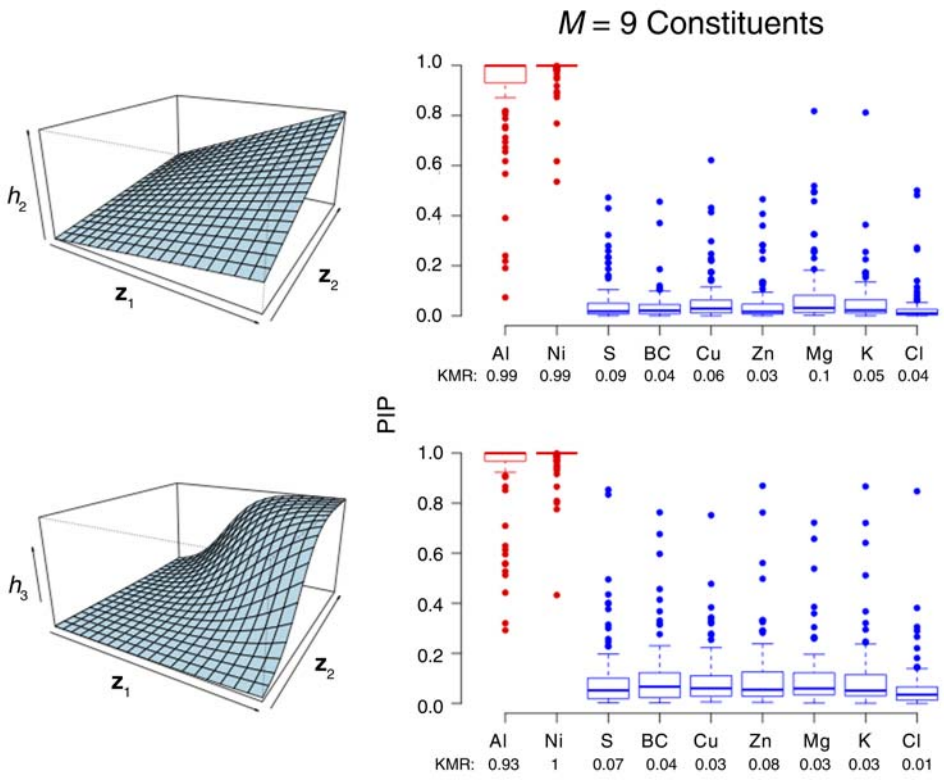


Figure 6. Median (25%, 75%) of the PIPs from BKMRvs under a realistic signal-to-noise ratio, with non-null effects for Al and Ni. Calculated across 100 simulated data sets, for each of the two bivariate $h(\mathbf{z})$ functions. The vector of exposure data $\mathbf{z}$ was generated based on $M = 9$ air pollution constituents (Al, Ni, S, BC, Cu, Zn, Mg, K, Cl). The only pollutants truly associated with the outcome were Al and Ni. Compare these results with Figure 5 in which Al and S were the associated pollutants for $h(\cdot)$. The proportion of simulation iterations for which each pollutant had $P < 0.05$ under the frequentist KMR approach of Maity and Lin (2011) is printed below the x axis.

Table 3. Ability of BKMRvs to Correctly Select None, One, or Two of the Important Pollutants in h_2 and h_3 Bivariate Exposure Scenarios (Simulation Study 1)[a]

			Number of Simulated Datasets			
Exposure	Function	PIP Threshold	Total	Both Identified	One Identified	Neither Identified
$M = 3$ Metals	h_2	PIP > 0.5	100	99	1	0
	h_2	PIP > 0.8	100	94	6	0
	h_3	PIP > 0.5	100	97	3	0
	h_3	PIP > 0.8	100	90	9	1
$M = 9$ Constituents	h_2	PIP > 0.5	99	71	28	0
	h_2	PIP > 0.8	99	46	50	3
	h_3	PIP > 0.5	97	80	17	0
	h_3	PIP > 0.8	97	65	32	0

[a] Results are shown for both $M = 3$ metals and $M = 9$ constituents for two h functions and two PIP thresholds [PIP > 0.5 or > 0.8] for determining that an important pollutant had been selected.

SIMULATION STUDY 2. COMPONENT-WISE VERSUS HIERARCHICAL VARIABLE SELECTION IN HIGH-CORRELATION SETTINGS

In Simulation Study 2, we generated data sets in the same way as for Simulation Study 1 using the Harvard Chan School Boston Supersite multipollutant data set, but using $M = 13$ PM constituents (Al, Si, Ti, Ca, Ni, V, Zn, S, BC, Cu, K, Cl, and Mn), some of which were highly correlated with one another. Figure 7 presents the correlation matrix for these 13 constituents. The daily data were again standardized by subtracting the median and dividing by the IQR, and days with outlier values (greater than 5 IQR away from the median) were removed. We set σ^2 to correspond to the realistic signal-to-noise ratio from Simulation Study 1.

Models

We again fit the KMR using a frequentist approach (Liu et al. 2007), both without (KMR) and with (KMRvs) variable selection. Second, we fit three BKMR models that included all 13 of the available exposure variables: a model without variable selection (BKMR), a model with component-wise variable selection (BKMRvs), and a model with hierarchical variable selection (BKMRhvs). For BKMRhvs we defined the pollutant groups $S_1,\ldots,S_8$ based on a combination of knowledge of Boston air pollution source categories (Clarke et al. 2000; Kioumourtzoglou et al. 2014; Nikolov et al. 2008) and the empirical correlations among the different pollutants: S_1 = Al, Si, Ti, and Ca; S_2 = Ni, V, and Zn; S_3 = S; S_4 = BC; S_5 = Cu; S_6 = K; S_7 = Cl; and S_8 = Mn. The within-group correlations ranged from 0.68 to 0.87 in group S_1 and from 0.45 to 0.8 in group S_2. We again also fit the oracle models to each data set simulated under each h_1, h_2, and h_3 function.

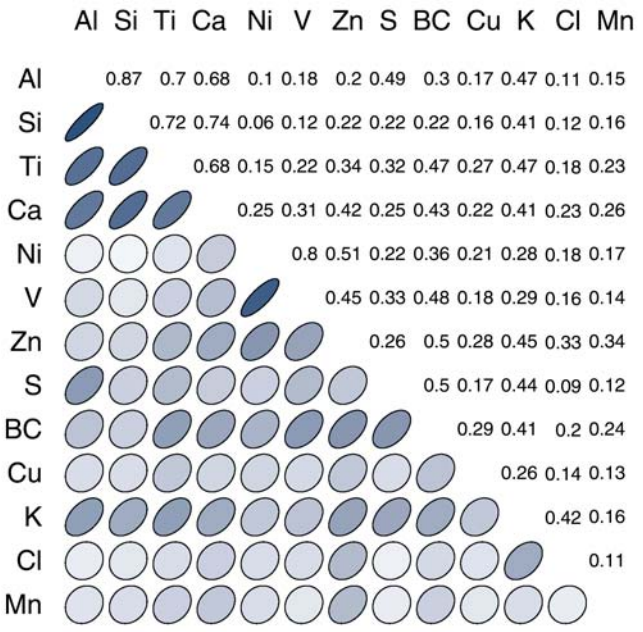


Figure 7. Correlation matrix of multipollutant mixture data set used for the $M = 13$ scenario in Simulation Study 2 (measured daily during 1998–2011 at the Harvard Chan School Boston Supersite). The shading and shapes indicate the strength of the correlation between a pair of constituents, with darker and more oval shapes reflecting higher correlations. These provide a visual representation of the correlation structure present in the numerical values shown in the upper triangle of the matrix.

Results for Identifying Important Pollutants

BKMRvs and BKMRhvs performed almost identically in terms of their ability to estimate h; both methods outperformed the frequentist KMRvs approach. We therefore do not show the results for KMRvs here; the full details are reported elsewhere (Bobb et al. 2014).

Here we present the ability of three methods (BKMRvs, BKMRhvs, and the Maity and Lin [2011] Garrote KMR test) to correctly identify which mixture or which individual pollutants were predictive of the health outcome [i.e., included in $h(\cdot)$]. Figure 8 shows the boxplots of the PIPs under BKMRvs, as well as the proportion of iterations for which each pollutant was identified as statistically significant by the Maity and Lin Garrote KMR test (2011; values shown beneath the x axes). Both BKMRvs and the Garrotte KMR test were able to identify Cu, a pollutant whose correlation with the other pollutants ranged from 0.13 to 0.29, as important in the functions h_2 and h_3, in which Cu was assumed to be truly associated with the health outcome. On the other hand, for Al, a pollutant highly correlated with several others (correlation of 0.87 with Si, 0.70 with Ti, and 0.68 with Ca), the Garrote KMR test had lower power and had inflated type I errors with its correlated exposures, especially Si. For BKMRvs, although the PIPs remained higher for Al than for its correlated constituents, Si also had slightly higher PIPs compared with the other non-causal pollutants.

Figure 9 shows the PIPs for each group (i.e., the posterior mean of the group indicators ω_g), as well as the conditional PIPs for the pollutants of group S_1 (Al, Si, Ti, and Ca; i.e., the posterior mean of $\boldsymbol{\delta}_{S_1} \mid \omega_1 = 1$) under BKMRhvs. The results show a clear separation between the PIPs for the groups that included one of the pollutants that was truly predictive of health (S_1 under h_1; and S_1 and S_5 [Cu] under h_2 and h_3) and those that did not. Under BKMRhvs, the PIPs for group S_1, which includes the pollutant Al that was truly predictive of the outcome in $h(\cdot)$, were higher than for

any of the individual pollutants of S_1, suggesting that BKMRhvs is more likely to detect important pollutants in the high-correlation setting. Within group S_1, BKMRhvs was better able to distinguish between the important pollutant (Al) and the unimportant pollutants (Si, Ti, Ca) compared with BKMRvs (Figure 8), suggesting that there is added value in using the hierarchical formulation compared with the component-wise variable selection in high-correlation settings.

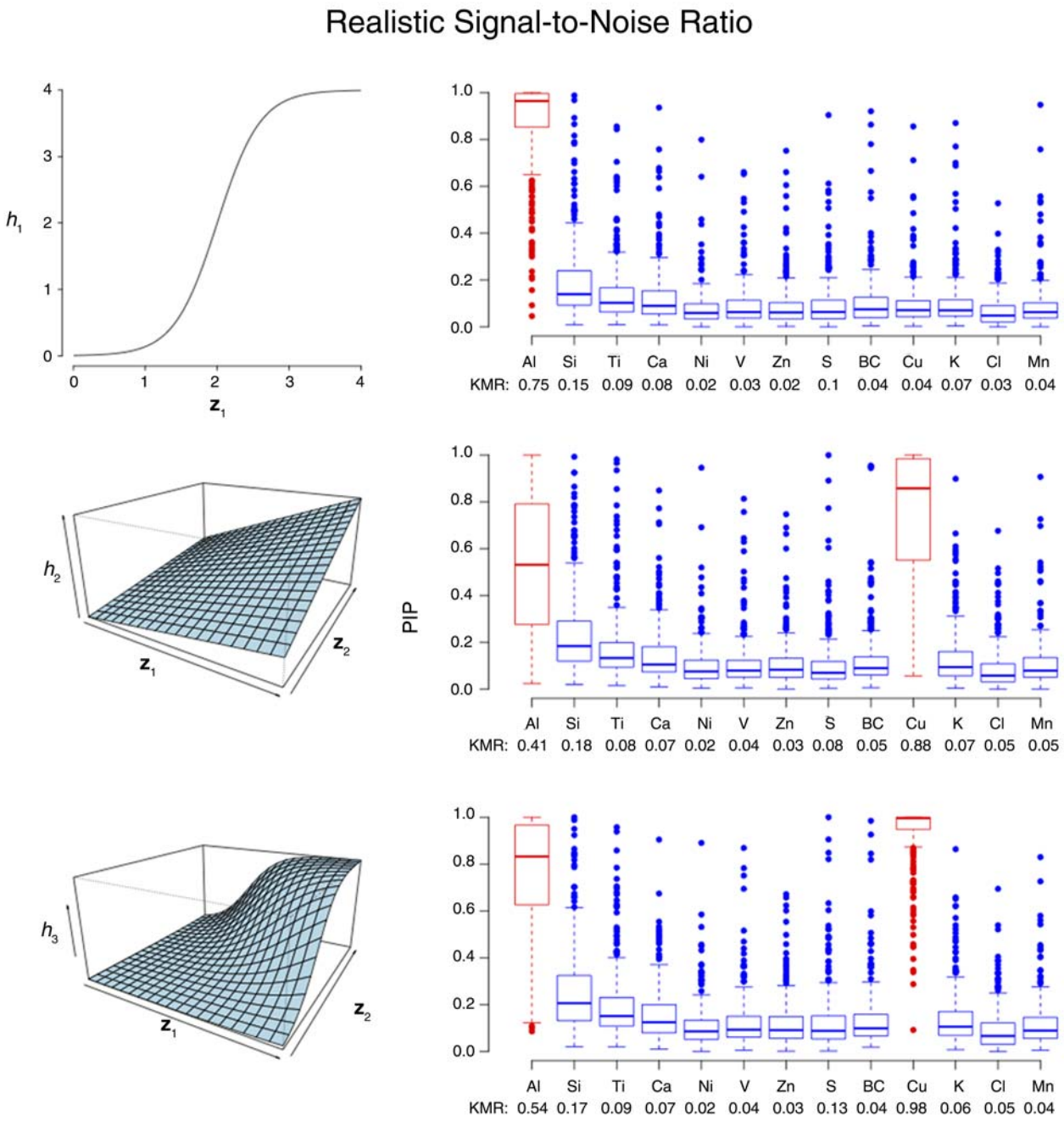


Figure 8. Median (25%, 75%) of the PIPs from BKMRvs under a realistic signal-to-noise ratio. Calculated across 100 simulated data sets, for each of three true $h(\mathbf{z})$ functions. The vector of exposure data $\mathbf{z}$ was generated from the Harvard Chan School Boston Supersite multipollutant data set with $M = 13$ air pollution constituents, in which the truly associated pollutants were Al for the first h function, and Al and Cu for the second and third h functions. The proportion of simulation iterations for which each pollutant had $P < 0.05$ under the frequentist KMR approach of Maity and Lin (2011) is printed below the x axis.

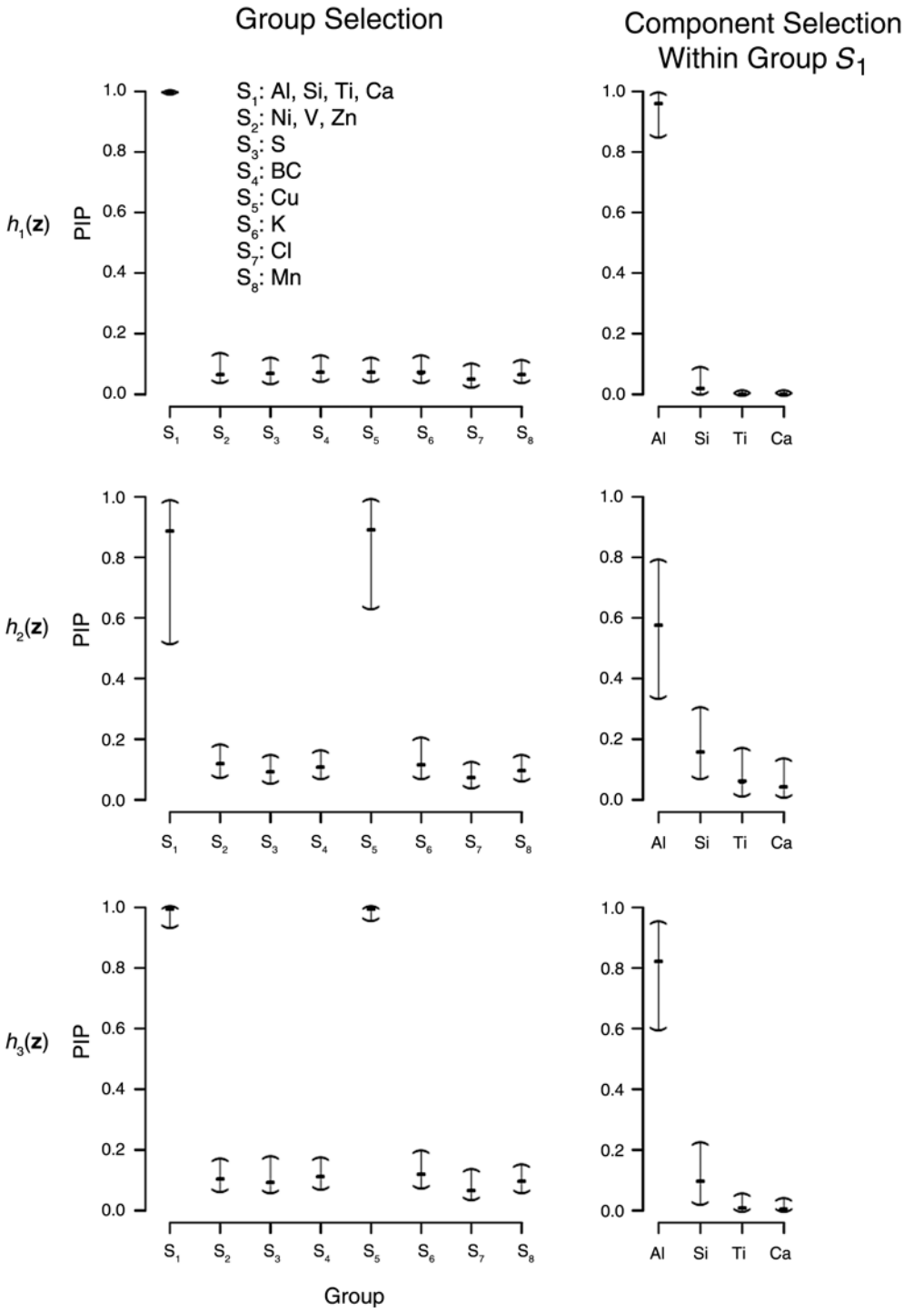


Figure 9. Median (25%, 75%) of the PIPs from BKMRhvs. Calculated for each of three true $h(\mathbf{z})$ functions. Exposure data $\mathbf{z}$ were generated from the Harvard Chan School Boston Supersite multipollutant data set for $M = 13$ air pollution constituents in eight pollutant groups. The truly associated pollutants were Al (one of four pollutants in group S_1) for the h_1 function, and Al and Cu (sole pollutant in group S_5) for the h_2 and h_3 functions. Plots on the left show the PIPs for each group; plots on the right show the conditional PIPs for the pollutants in group S_1, given that group S_1 was included in the model.

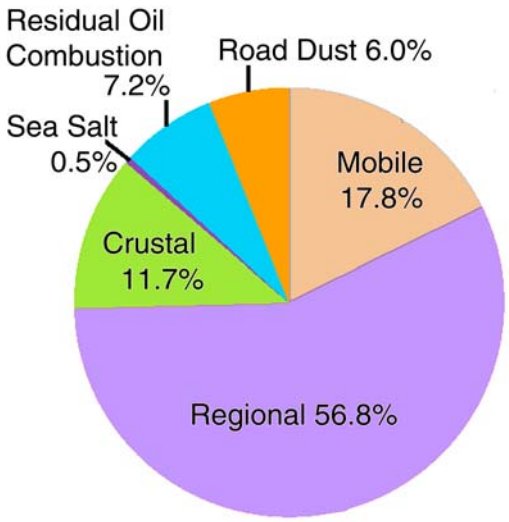


Figure 10. Source apportionment of $PM_{2.5}$ mass from PMF analysis that generated the source contribution estimates for use in the source category–specific health effects analysis in Simulation Study 3.

SIMULATION STUDY 3. SOURCE CATEGORY–SPECIFIC HEALTH EFFECTS

Simulation Setup

Source categories of $PM_{2.5}$ and its constituents were identified using the Harvard Chan School Boston Supersite multipollutant data set described earlier and a standard source apportionment method: the U.S. EPA PMF 3.09 (Paatero and Tapper 1994). This method was selected because it has been extensively used in fine-particulate source apportionment in the past (Ito et al. 2006; Lall et al. 2011). We used the source apportionment results described by Kioumourtzoglou and associates (2014), in which PMF was applied to $PM_{2.5}$ mass and 19 of its constituents. Table 4 shows the constituents included in the PMF analysis. PMF identified six factors consistent with the source categories of regional, mobile, crustal, residual oil combustion, road dust, and sea salt. Figure 10 shows the source category–specific apportionment of the mass across the entire period of 1998

Table 4. Estimated Factor Loadings from the PMF Analysis Used to Identify Six Pollution Source Categories (Simulation Study 3)[a]

Constituent	Mobile	Regional	Crustal	Sea Salt	Residual Oil Combustion	Road Dust
$PM_{2.5}$	1601.30	5122.40	1057.80	43.08	646.27	544.23
BC	436.50	129.91	6.08	3.596	29.94	14.66
Na	7.03	90.73	29.63	20.12	22.95	2.91
Al	3.24	9.90	33.01	0.267	1.24	0.00
Si	5.27	0.95	61.88	0.000	0.61	1.44
S	0.00	814.35	79.16	0.532	69.07	20.41
Cl	0.00	0.00	0.00	14.249	0.00	0.00
Ca	9.76	0.25	14.35	0.719	1.71	2.22
Ti	1.29	0.15	1.69	0.004	0.00	0.10
V	0.27	0.17	0.00	0.000	2.88	0.00
Cr	0.21	0.04	0.10	0.002	0.00	0.04
Mn	0.31	0.00	0.04	0.006	0.00	0.25
Fe	33.46	0.75	19.67	0.142	1.40	5.34
Ni	0.00	0.00	0.00	0.001	2.38	0.29
Cu	1.53	0.28	0.50	0.048	0.06	0.53
Zn	1.36	0.78	0.00	0.000	0.00	8.99
Se	0.02	0.06	0.00	0.000	0.00	0.02
Br	0.19	0.25	0.08	0.026	0.05	0.17
Ba	3.57	0.85	1.85	0.120	0.27	0.53
Pb	1.99	0.70	0.97	0.037	0.44	0.73

[a] Values are ng/m^3. Data analyzed were from the Harvard Chan School Boston Supersite; source apportionment results are described in Kioumourtzoglou and associates 2014.

through 2011. The correlations across factors in the PMF solution were small to moderate; maximal correlation was found between road dust and mobile source categories (r = 0.45). We then randomly generated data sets containing source category–specific exposures by sampling days during the period 1998–2011; and, using the estimated source contributions from the PMF analysis, we subsequently generated health outcomes for a simulated sample assuming that only one source category was associated with adverse health effects. Specifically, for the kth source on the ith day (W_{ki}), we assumed $Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 W_{ki} + \varepsilon_i$. We set $\boldsymbol{\beta}_1$ to be relatively large, such that if we were to fit a regression model using the source contributions used to simulate the responses, the power to detect the association with the correct source category was 100%. We then ran a BKMR analysis using as pollutants 14 constituent concentrations (Al, S, Ni, BC, Na, Cu, Zn, V, Ti, Ca, Mg, K, Cl, and Si) and checked to see which pollutants were identified, using the PIPs, as being associated with the health outcome. Our a priori hypothesis was that one or more of the pollutants serving as tracers for a given source category, as characterized by high factor loadings for that constituent on that category, would be identified as "included" in the model by the proposed BKMR.

Results

For each of the six simulation scenarios corresponding to health effects associated with a given pollution source category (mobile, regional, crustal, sea salt, residual oil combustion, and road dust), Figure 11 shows the distribution of the PIPs for the 14 constituents included in the analyses. Overall, the distributions of the inclusion probabilities for constituents thought to be tracers for a particular source category assumed to generate health effects were shifted relative to the other constituents in the model. For instance, for the scenario in which the mobile source category was assumed to generate health effects, the median of the PIP distribution for BC was > 0.9, whereas the medians of the distributions of PIPs for all other constituents were < 0.5. For the scenario in which the source category of regional pollution was assumed to generate health effects, the median of the PIP distribution for S was > 0.9, whereas the medians of the PIP distributions for all other constituents were < 0.2. For the scenario in which the sea salt source category was assumed to generate health effects, the median of the PIP distribution for Cl was > 0.5, whereas the medians of the PIP distributions for all other constituents were < 0.10. For these three scenarios, each of the three constituents has been identified as a strong tracer for the corresponding source category (Clarke et al. 2000; Nikolov et al. 2008).

For the scenarios with the crustal and residual oil combustion source categories, the two highly correlated constituents identified are both tracers for that source category; the BKMR inclusion probabilities tended to split the weight of evidence across the two relevant tracers. For instance, for the scenario in which the crustal source category was assumed to generate health effects, the distributions of the PIPs for both Al and Si were shifted relative to those for the other constituents included in the analysis; the median of Al was 0.25 and of Si was 0.95; whereas the medians of the PIP distributions for the other constituents were all < 0.10. Similarly, for the scenario in which the residual oil combustion source category was assumed to generate health effects, the distributions of the PIPs for both V and Ni were shifted relative to those for the other constituents.

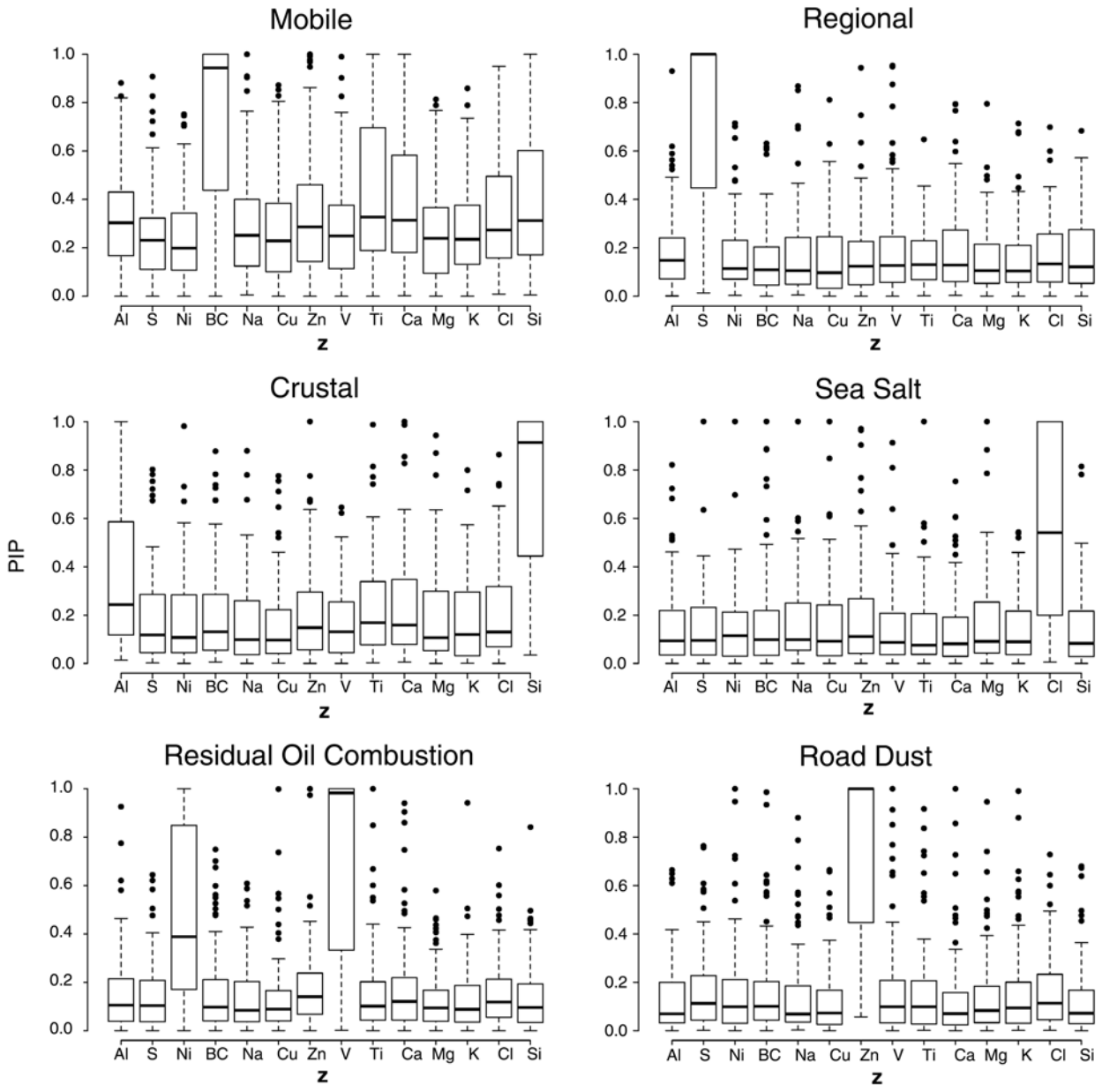


Figure 11. Histogram of the PIPs for each pollutant across 100 simulated data sets for six source categories (resulting from source apportionment analyses) when health effects were generated from source contributions estimated from the Harvard Chan School Boston Supersite multipollutant data set. The vector of exposure data **z** represents a set of 14 air pollution constituents (Al, S, Ni, BC, Na, Cu, Zn, V, Ti, Ca, Mg, K, Cl, Si).

PM COMPOSITION AND BLOOD PRESSURE IN THE MOBILIZE STUDY

We used the BKMR model to evaluate the association between short-term changes in $PM_{2.5}$ composition and dynamic BP responses to orthostatic challenge within the context of the MOBILIZE Boston study, a prospective, community-based cohort study of healthy aging subjects. Previous work has described the design of the study and associations between BP outcomes and short-term changes in $PM_{2.5}$ levels (Wellenius et al. 2012).

STUDY DESIGN

Briefly, between 2005 and 2008, the MOBILIZE study recruited 748 men and women age 70 years or older who were not institutionalized, were able to communicate in English, resided within 5 miles (8.0 km) of the study clinic at Hebrew SeniorLife, and were able to walk 20 feet (6.1 m) without assistance. Individuals not planning to reside in the study area for 2 or more years and those with terminal diseases, severe vision or hearing impairment, or cognitive impairment (defined by a Mini-Mental State Examination score of < 18) were excluded. Wellenius and associates (2012) have reported in detail the characteristics of the study participants. All subjects provided written informed consent upon enrollment. This analysis was approved by the Institutional Review Boards at Hebrew SeniorLife and Brown University.

A description of the BP measurement procedure is described in detail elsewhere (Wellenius et al. 2012). Participants were asked to stand and, with the BP cuff kept at heart level, measurements (both diastolic and systolic) were repeated 1 and 3 minutes after both feet touched the floor. Data from both a baseline and a follow-up visit were used, resulting in a total of 1362 subject observations, including two repeated measurements for most subjects (82%). In this analysis we focused on outcomes defined as DBP and SBP measurements 1 minute after standing.

We used ambient measurements of $PM_{2.5}$ and its constituents from the Harvard Chan School Boston Supersite multipollutant data set. This monitoring station is located less than 10 km from the study clinic site and less than 20 km from the residential address of any study participant. We obtained hourly meteorologic data from the National Weather Service station at Boston Logan Airport.

Wellenius and coworkers (2012) reported that standing BP measurements were associated with $PM_{2.5}$ mass averaged over the previous 7 and previous 14 days before the BP measurement. We applied the BKMR method to analyze the association between standing DBP or SBP and the 7-day moving averages of $PM_{2.5}$ constituent concentrations. Because some days were missing $PM_{2.5}$ composition data, the analytic data set contained 1050 observations for 681 subjects.

STATISTICAL ANALYSIS

We used linear mixed models and a mixed-model extension of the BKMR model (equation 4) that handles longitudinal data to evaluate the association of DBP and SBP each with $PM_{2.5}$ constituent concentrations. Subject-specific random intercepts were included in all models to account for the within-subject correlation among repeated measures (baseline and follow-up visits) taken on the same subject. Based on our previous work selecting relevant confounders for inclusion in the model (Wellenius et al. 2012), we controlled for age (natural cubic spline with 3 degrees of freedom), sex, race (white versus other), smoking status (never, former, or current), hypertension status (normotension, controlled hypertension, or uncontrolled hypertension), diabetes mellitus, body mass index (natural cubic spline with 3 degrees of freedom), visit number, day of week, ambient and dew-point temperatures (natural cubic splines with 3 degrees of freedom each), season (sine and cosine of each calendar day), and long-term temporal trends (calendar day as a linear continuous variable).

Let BP_{it} denote the standing blood pressure (either diastolic or systolic) value for subject i at measurement occasion t. We fit the model

$$BP_{it} = h\left(\mathrm{Ni}_{it}, \mathrm{Cu}_{it}, \mathrm{Zn}_{it}, \mathrm{S}_{it}, \mathrm{Ti}_{it}, \mathrm{Mn}_{it}, \mathrm{BC}_{it}\right) + \mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + b_i + \varepsilon_{it}, \quad (9)$$

where Ni_{it}, Cu_{it}, Zn_{it}, S_{it}, Ti_{it}, Mn_{it}, and BC_{it} represent the 7-day moving averages of Ni, Cu, Zn, S, Ti, Mn, and BC, and $\mathbf{x}_{it}$ contains variables that represent the confounders listed above, for the tth measurement occasion for the ith subject. The subject-specific terms $b_i \overset{iid}{\sim} \mathrm{N}\left(0, \sigma_b^2\right)$ are independent from ε_{it}. In the conventional frequentist KMR approach to estimating the parameters in equation (4), we would refer to b_i as the random subject-specific intercepts. In the BKMR, all unknown parameters are treated as random variables, so the distinction between fixed and random effects is not applicable. To complete the prior specification for this longitudinal extension of the model, for computational convenience we reparameterized the model as $\lambda_b = \sigma_b^2 / \sigma^2$ and assigned a prior on λ_b of Gamma(100, 100). In contrast to some of the other existing approaches to assessing the effects of mixtures, the ease with which BKMR can be extended to popular study designs that generate correlated data is a strength of the method.

RESULTS

Based on the results from Simulation Study 1 suggesting that one must choose thoughtfully the pollutants entered into the kernel (because inclusion of two highly correlated pollutants tends to decrease the inclusion probabilities of both pollutants simultaneously), of the nine pollutants used in Simulation Study 1, we removed K, Al, and Mg due to their correlations with the remaining six pollutants, removed Cl since it is not associated with a major pollution source, and added Ti because it is thought to be a component of road dust. In addition, we included Mn since exploratory analyses of the animal data from the Harvard Chan School toxicologic BP canine study (reported later) suggested it could be important, and a goal of this project was to use the same set of $PM_{2.5}$ constituents in analyses of both the MOBILIZE and canine data.

In exploratory analyses of the linear associations between standing BP and each of these seven pollutants (Ni, Cu, Zn, S, Ti, Mn, and BC), we fit both one-pollutant and seven-pollutant linear mixed-effects models, adjusted for the confounders listed in the previous section. Table 5 presents the results for DBP and SBP. For DBP, results suggest that Cu, Ti, and BC have the strongest univariate associations with the outcome. As one might expect in a multipollutant model with seven pollutants, some of which are moderately correlated with one another, the strength of these associations weakens. In the seven-pollutant model, Cu and BC are the only two pollutants for which any evidence of an association remains, and the estimated slopes for these two pollutants decrease by 20% to 30%.

Also for DBP, Figure 12 displays the PIPs for the seven pollutants from the BKMR fit. Results show BC and Cu to be the two constituents with the highest PIPs. Accordingly, Figure 13 presents the estimated bivariate exposure–response surface $\hat{h}$, as a function of BC and Cu, evaluated at the median levels of the other five pollutants. The left panel of this figure presents an image plot of $\hat{h}(Ni_{50}, Cu, Zn_{50}, S_{50}, Ti_{50}, Mn_{50}, BC)$, and the right panel plots three cross-sections of this surface across the range of BC concentrations, taken at the 10th, 50th, and 90th percentiles of the observed Cu distribution. The line styles of the cross-sections in the image plot correspond to the exposure–response curves depicted in the right panel. This figure plots the exposure–response relationship only for points within the range of the data in order to avoid extrapolating the estimated relationships beyond that range. Figure 14 displays univariate associations between BC and DBP at fixed levels of Cu, but with point-wise 95% credible intervals added for each exposure–response association. Figures 15 and 16 depict the same image plot and univariate associations, but this time reversing the role of Cu and BC; results show the association between DBP and Cu at increasing levels of BC. Taken in total, the BKMR analyses provided moderate evidence of an association between standing DBP and 7-day moving averages of Cu and BC concentrations.

For SBP, Figure 17 displays the PIPs for the seven pollutants from the BKMR fit. Results show S to have the highest inclusion probability; those for Ni and BC are essentially tied for second highest. Accordingly, Figure 18 presents the estimated bivariate exposure–response surface $\hat{h}$, as a function of S and Ni evaluated at the median levels of the other five pollutants, again only for combinations of S and Ni within the range of the observed data. Figure 19 adds 95% credible intervals to the plots of the estimated SBP–S associations. Cross-sections and associated 95% credible intervals in the Ni direction at increasing levels of S showed no association between SBP and Ni (not shown). Because the inclusion probabilities for Ni and BC were essentially tied, we also inspected the resulting image plots as a function of S and BC, evaluated at the median values of the other five pollutants (Figures 20 and 21). These plots reinforce the evidence of a linear, additive effect of S, and cross-sections in the BC direction showed no association between SBP and BC (not shown).

Taken all together, these analyses suggest that the associations of DBP with Cu and BC and of SBP with S are linear and additive. Therefore, the full generality of the Gaussian kernel for the multidimensional exposure–response surface was not required in this case; however, we would not have known this without fitting the more general model. One advantage of using the BKMRvs approach when the effects are linear and additive is that the variable-selection component of the method allows us to detect these linear and additive effects even when we enter a relatively large number of pollutants in the model. That is, even in cases of linear and additive effects, it appears that the BKMR model yields more efficient estimates of the linear Cu exposure–response function compared with the full seven-pollutant linear mixed model (Table 5); this is likely due to the variable-selection feature of BKMRvs that serves to minimize the impact of the inclusion of unimportant pollutants in the kernel function. This selection serves to decrease the effective dimension of the multipollutant covariate vector, leading to a more parsimonious model and therefore more efficient effect estimates.

Table 5. Estimated Regression Coefficients, Standard Errors, and *P* Values from Linear Mixed Models Applied to DBP and SBP Data from the MOBILIZE Cohort[a]

	Single-Pollutant Model			Seven-Pollutant Model		
Pollutant	$\hat{\beta}$	SE $(\hat{\beta})$	*P* Value	$\hat{\beta}$	SE $(\hat{\beta})$	*P* Value
DBP						
Ni	0.06	0.82	0.94	−0.79	0.91	0.39
Cu	0.63	0.29	0.03	0.45	0.33	0.17
Zn	0.61	0.46	0.19	0.19	0.61	0.76
S	0.71	0.49	0.14	−0.11	0.63	0.85
Ti	0.69	0.39	0.08	0.38	0.46	0.41
Mn	0.03	0.50	0.96	−0.53	0.57	0.36
BC	0.99	0.46	0.03	0.84	0.63	0.18
SBP						
Ni	1.23	1.53	0.42	1.02	1.70	0.55
Cu	0.87	0.53	0.11	0.60	0.62	0.33
Zn	−0.43	0.87	0.62	−1.83	1.15	0.11
S	1.80	0.91	0.05	1.63	1.17	0.17
Ti	0.23	0.74	0.75	−0.28	0.87	0.75
Mn	−0.62	0.95	0.51	−0.68	1.07	0.53
BC	1.14	0.86	0.19	1.11	1.18	0.35

[a] The single-pollutant models used the 7-day moving average of each constituent's concentration; the seven-pollutant model included all seven pollutants simultaneously.

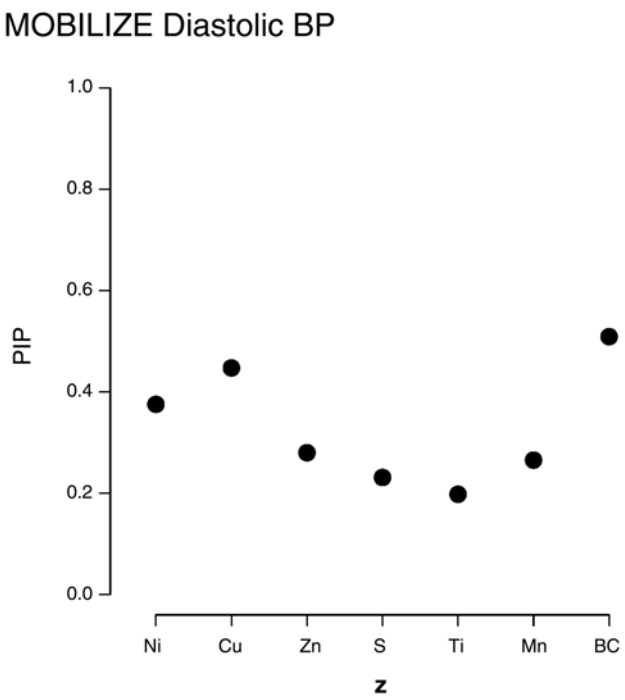


Figure 12. PIPs from BKMR analysis of DBP in the MOBILIZE data set, using prior hyperparameters equal to those from a frequentist KMR fit.

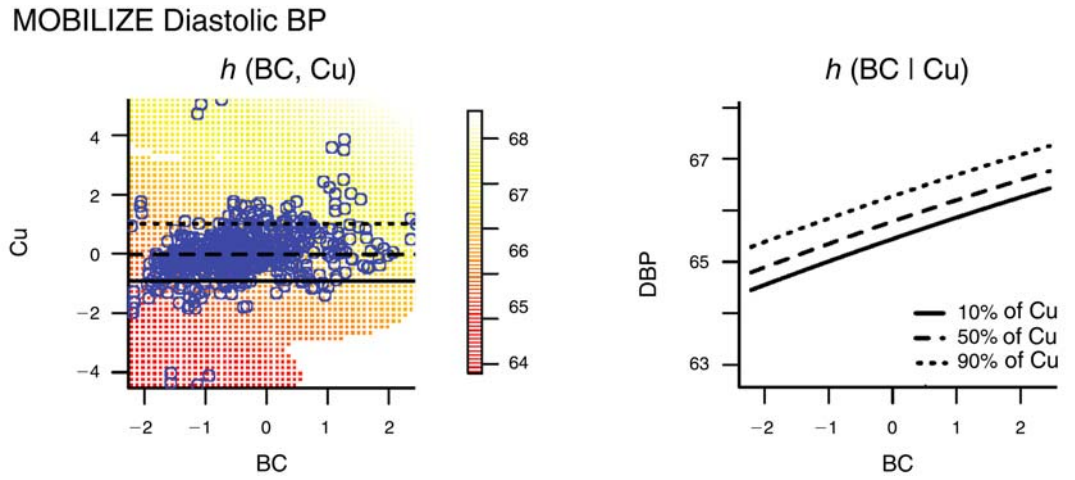


Figure 13. Image plot of $\hat{h}$ as a function of BC and Cu concentrations for DBP, evaluated at the median concentrations of the other five pollutants in the model (Ni, Zn, S, Ti, Mn).

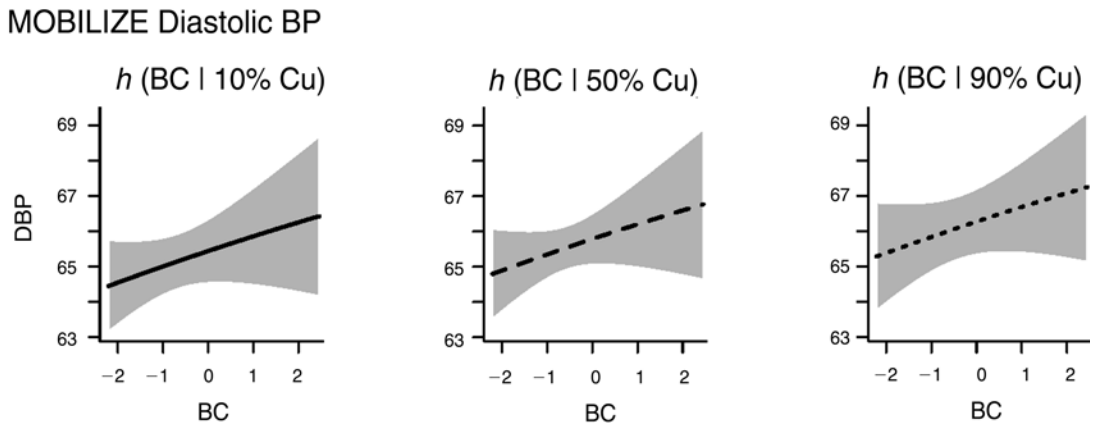


Figure 14. Plot of $\hat{h}$ as a function of BC for DBP, and the associated 95% pointwise credible intervals, evaluated at increasing concentrations of Cu and the median concentrations of the other five pollutants in the model (Ni, Zn, S, Ti, Mn).

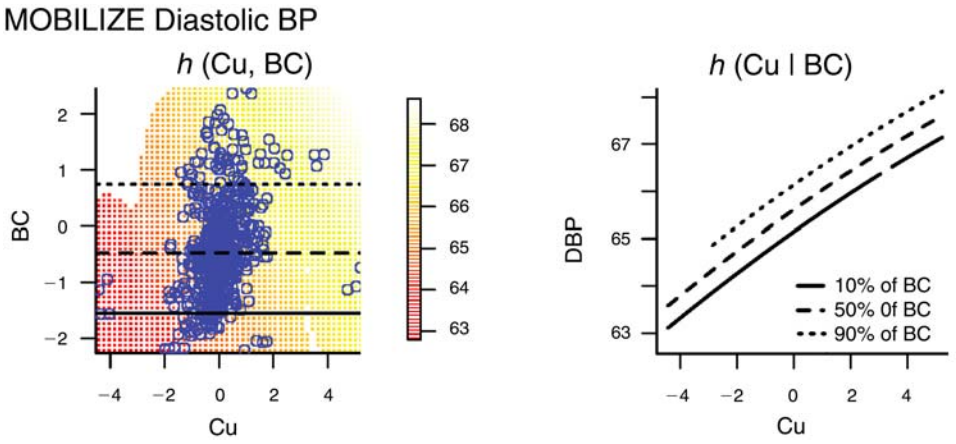


Figure 15. Image plot of $\hat{h}$ as a function of Cu and BC concentrations for DBP, evaluated at the median concentrations of the other five pollutants in the model (Ni, Zn, S, Ti, Mn).

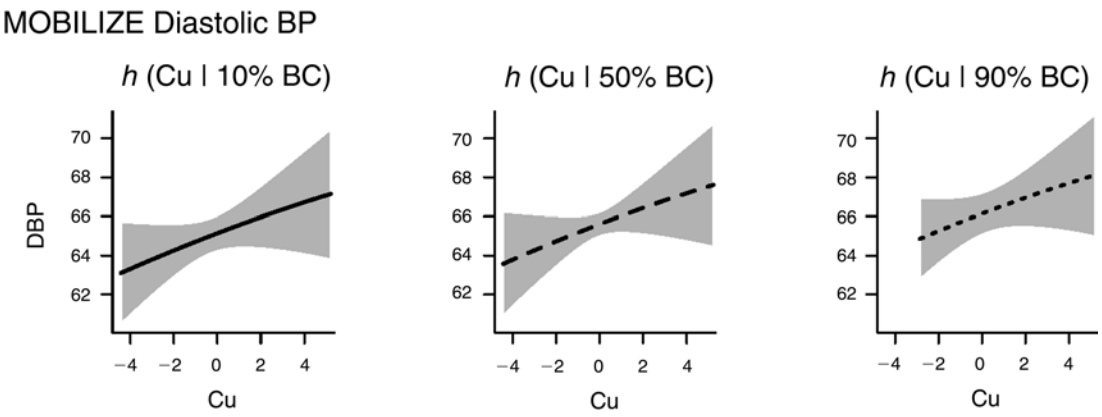


Figure 16. Plot of $\hat{h}$ as a function of Cu for DBP, and the associated 95% pointwise credible intervals, evaluated at increasing concentrations of BC and the median concentrations of the other five pollutants in the model (Ni, Zn, S, Ti, Mn).

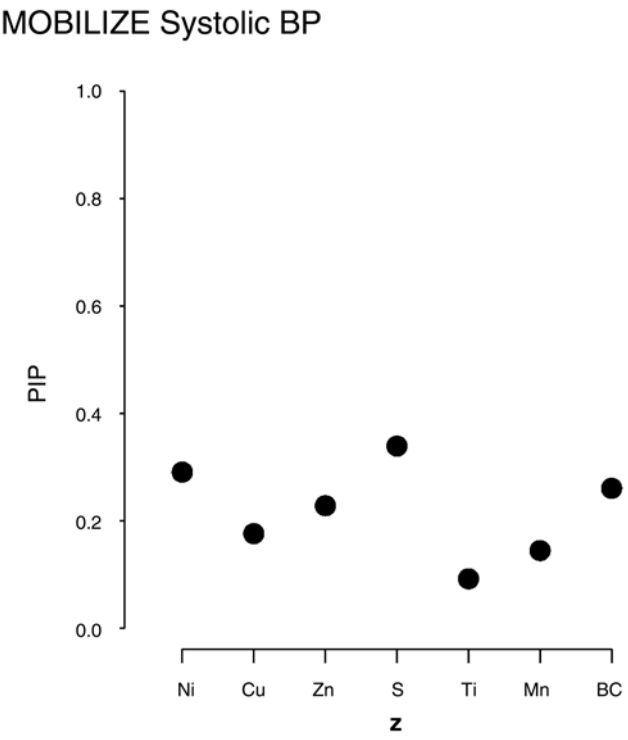


Figure 17. PIPs from BKMR analysis of SBP in the MOBILIZE data set, using prior hyperparameters equal to those from a frequentist KMR fit.

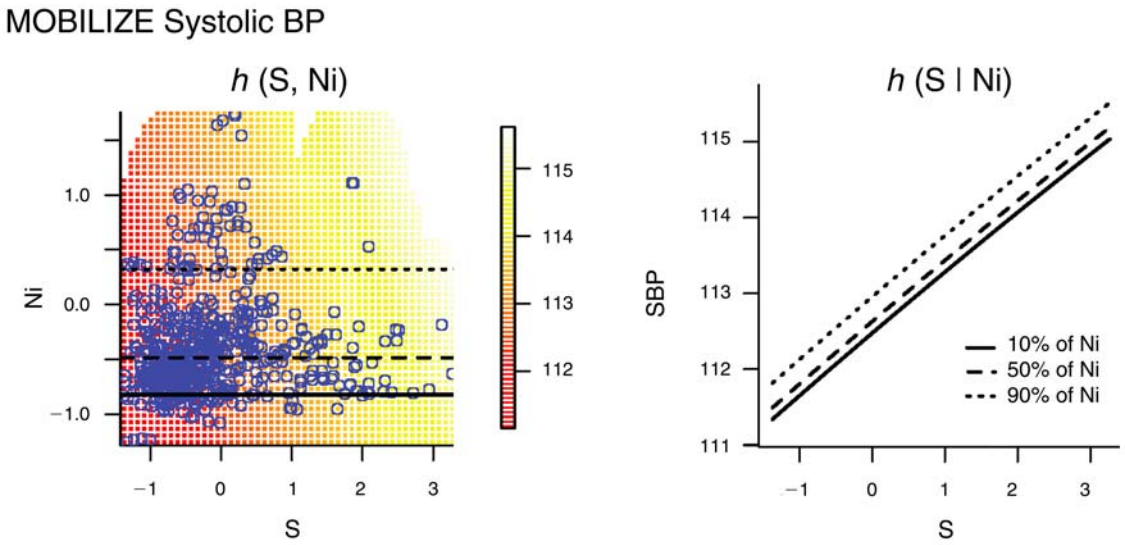


Figure 18. Image plot of $\hat{h}$ as a function of S and Ni concentrations for SBP, evaluated at the median concentrations of the other five pollutants in the model (Cu, Zn, Ti, Mn, BC).

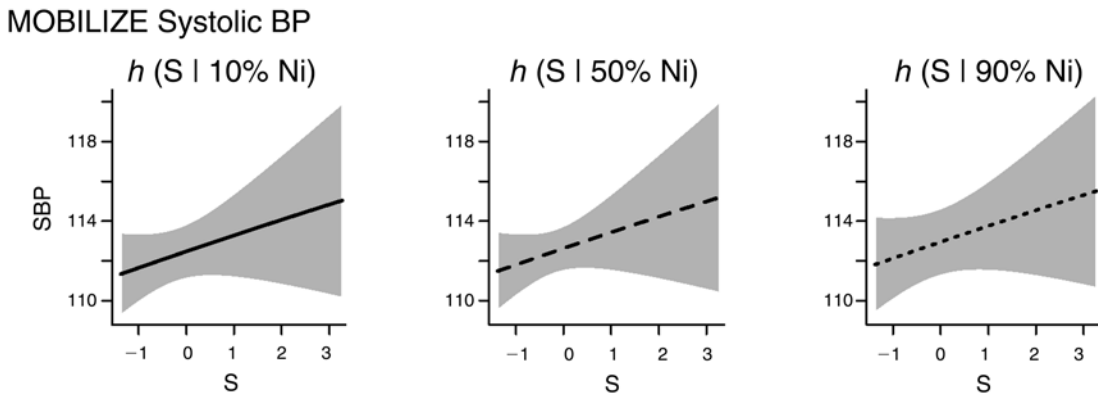


Figure 19. Plot of $\hat{h}$ as a function of S for SBP, and the associated 95% pointwise credible intervals, evaluated at increasing concentrations of Ni and the median concentrations of the other five pollutants in the model (Cu, Mn, Ti, Zn, BC).

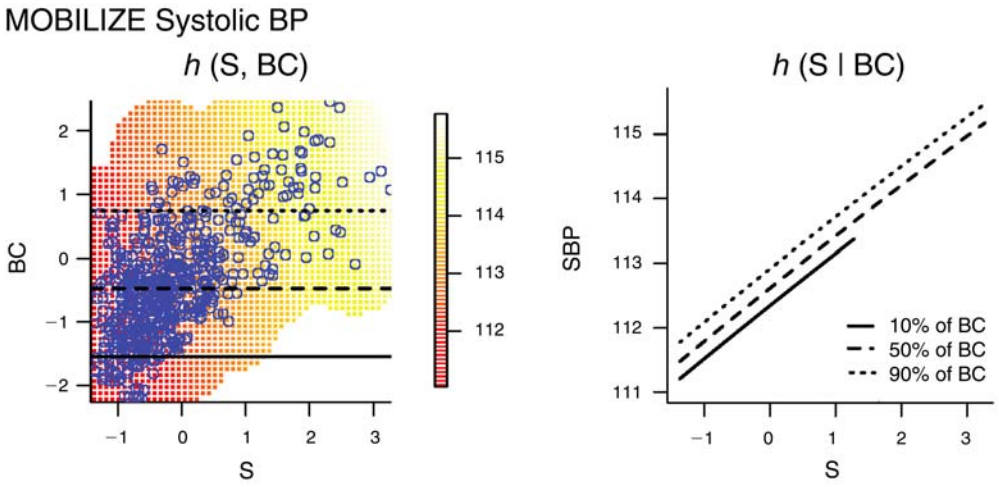


Figure 20. Image plot of $\hat{h}$ as a function of S and BC concentrations for SBP, evaluated at the median concentrations of the other five pollutants in the model (Cu, Mn, Ti, Zn, Ni).

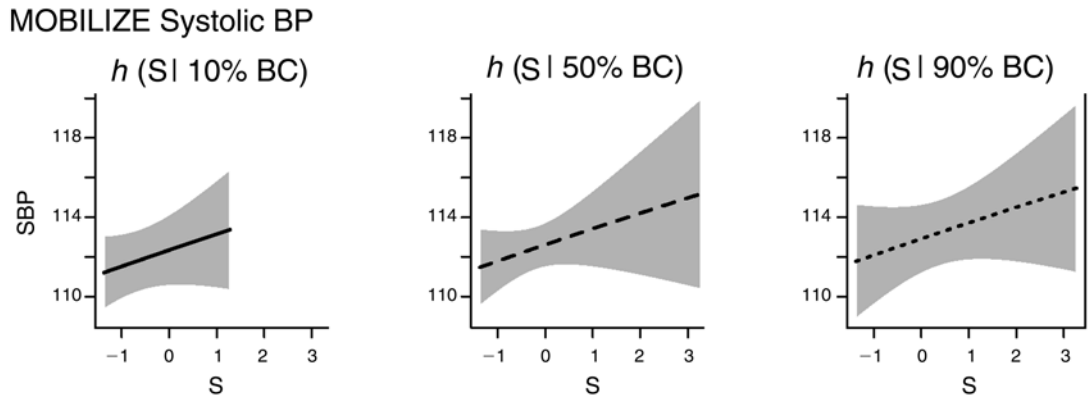


Figure 21. Plot of $\hat{h}$ as a function of S for SBP, and the associated 95% pointwise credible intervals, evaluated at increasing concentrations of BC and the median concentrations of the other five pollutants in the model (Cu, Mn, Ti, Zn, Ni).

PM COMPOSITION AND BLOOD PRESSURE IN THE HARVARD T.H. CHAN SCHOOL CANINE STUDY

To illustrate the applicability of the BKMRvs and BKMRhvs methods to animal toxicologic studies, we analyzed hemodynamic data from the CAPs toxicologic study conducted in Dr. John Godleski's laboratory. Bartoli and colleagues (2009) reported effects of CAPs exposure, as well as associations of continuous measures of $PM_{2.5}$ mass, BC concentrations, and particle number, with several hemodynamic health outcomes, as detailed below. In this report we follow up these results by applying BKMRvs and BKMRhvs to assess the effects of CAPs composition on the same outcomes.

STUDY DESIGN

The protocol for the study that generated these data has been described in detail previously (Bartoli et al. 2009). To evaluate the acute effects of ambient PM on arterial blood pressure, thirteen female mixed-breed dogs were repeatedly exposed for 5 hours to either CAPs or filtered air in a crossover protocol. For convenience, animals were exposed in pairs in which one animal was assigned to CAPs exposure and the other to filtered air. In most instances, exposure days were separated by at least 7 days during which no exposures took place. Blood pressure data were available from these thirteen animals exposed to filtered air on 63 days and to CAPs on 55 days. The range of repeated exposures per dog was unbalanced across the 13 animals and ranged from 4 to 22.

Additional experiments were conducted with prazosin, an α-adrenergic antagonist, with 8 of the 13 animals (n = 15 filtered air exposures, 16 CAPs exposures; the specific procedures for administration are outlined in Bartoli et al. 2009). Furthermore, 11 out of 13 dogs had some blood pressure readings after an occlusion induced by a balloon occluder implanted in the left anterior descending coronary artery (Wellenius et al. 2003, Bartoli et al. 2009). These occlusions had been induced to produce reversible ischemia, and BP and heart rate measurements had been recorded before and after the occlusions. Thus, in our analyses, we designated whether an exposure occasion within a given dog was under baseline, post-occlusion, or post-prazosin conditions.

Arterial BP was monitored and recorded continuously throughout exposures (DSI Dataquest ART 3.1; Data Sciences). SBP, DBP, mean pressure, pulse arterial pressure, and heart rate were derived from the arterial BP recordings. Rate–pressure product (R × P), a standard index of myocardial metabolic demand, was calculated as the product of heart rate and SBP (Rooke and Feigl 1982). Bartoli and colleagues (2009) reported statistically significant effects of exposure to CAPs, continuous concentrations of $PM_{2.5}$ mass, BC, and particle number on multiple hemodynamic outcomes. Specifically, after controlling for animal, week, and time within a dog–exposure sequence, they reported effect estimates (SE) for CAPs exposure (yes vs. no) with each of the following: SBP = 2.7 (1.0) mmHg; DBP = 4.1 (0.8) mmHg; mean pressure = 3.7 (0.8) mmHg; pulse pressure = −1.7 (0.7) mmHg; heart rate = 1.6 (0.5) bpm; and R × P (bpm × mmHg) = 539 (110).

EXPOSURE TECHNOLOGY AND CHARACTERIZATION

The characteristics of the Harvard Ambient Particle Concentrator (HAPC) and exposure chamber are well documented (Godleski et al. 2000; Sioutas et al. 1995). The HAPC concentrates ambient fine PM with an aerodynamic diameter between 0.15 and 2.5 μm to approximately 30 times ambient levels with minimal effects on the particle-size distribution or chemical composition. Particles with diameter > 2.5 μm are removed upstream of the HAPC, whereas particles with diameter < 0.15 μm and ambient gases are neither enriched nor excluded. CAPs mass concentration was measured continuously using a tapered element oscillating microbalance (TEOM Series 1400a; Rupprecht & Patashnick, East Greenbush, NY); BC concentration was measured by Aethalometer (Model AE-9; Magee Scientific, Berkeley, CA); and CAPs particle number concentration was measured using a condensation particle counter (CPC Model 3022A; TSI, Shoreview, MN), as previously described (Godleski et al. 2000). In addition, each CAPs exposure was measured for sulfate (SO_4^{2-}) via ion chromatography; elemental carbon (EC) and organic carbon (OC) determined with a thermal and optical reflectance method; and elemental concentrations (in $\mu g/m^3$) were collected via XRF for Al, As, Ba, Br, Ca, Cd, Cl, Cr, Cu, Fe, K, Mn, Ni, Na, Pb, S, Se, Si, Ti, V, and Zn.

STATISTICAL ANALYSIS

For the analyses conducted as part of this project, for each outcome (DBP, SBP, mean pressure, pulse pressure, heart rate, and R × P) we averaged all 5-minute averages from a given dog–exposure to obtain a single value for that dog for that exposure. Constituent concentrations for all filtered air exposures were assigned a value of zero. After removing several outliers (defined as being more than six SDs above the mean value) in the constituent concentrations, the data set consisted of n = 142 dog–exposures for 13 animals.

Because of the longitudinal crossover design of the study, we used the same mixed-model extension of the BKMRvs (outlined in the analysis of the MOBILIZE data) that included random effects. In order to make qualitative comparisons to the epidemiologic findings from the MOBILIZE analyses, we fit a model using the same pollutants as in that

analysis: Ni, Cu, Zn, S, Ti, Mn, and BC ($M = 7$). Specifically, let Y_{it} denote a given outcome (DBP, SBP, pulse, heart rate, or R × P) for animal i averaged over the 5-hour exposure at exposure occasion t. For each outcome separately, we fitted the model

$$Y_{it} = h\left(\mathrm{Ni}_{it}, \mathrm{Cu}_{it}, \mathrm{Zn}_{it}, \mathrm{S}_{it}, \mathrm{Ti}_{it}, \mathrm{Mn}_{it}, \mathrm{BC}_{it}\right) + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + b_i + \varepsilon_{it}, \quad (10)$$

where, again, the value of any given constituent in a filtered-air exposure combination i and t was equal to zero; and $\mathbf{x}_{it} = [\mathbf{I}(\text{post-occlusion})_{it}, \mathbf{I}(\text{post-prazosin})_{it}, \mathbf{I}(\text{CAPs})_{it}]^{\top}$ represents a 3 × 1 vector of dummy variables to indicate whether at exposure occasion t, animal i was exposed post-occlusion or post-prazosin administration and whether exposure was to CAPs or to filtered air.

In addition, in previous work (Bartoli et al. 2009) all outcomes exhibited an effect of CAPs (binary yes/no). An issue that arises in such cases is that, because all filtered air exposures are assigned concentration values of zero for all constituents, most $PM_{2.5}$ constituents exhibit associations with each outcome given the high collinearity between CAPs and each constituent concentration. That is, it is possible that differences between exposure groups, and not an exposure–response between concentration and outcome within an exposure group, drive an association between an outcome and a concentration. (See Coull et al. 2011 for a detailed discussion of this point.) Accordingly, we also controlled for an overall effect of CAPs exposure while estimating the association between an outcome and constituent concentrations among animals in the exposed (CAPs) group.

This is not unlike well-established methods to estimate the effects of smoking (among never, former, and current smokers) in epidemiologic analyses. Such methods typically include both an indicator term for former and current smokers (which reflects that some baseline effect of any exposure exists) and the number of cigarettes smoked by a current smoker (which reflects the effect of the amount of current exposure). As in the MOBILIZE analyses, we assumed $b_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_b^2)$, and the subject-specific intercepts b_i were independent of residual errors ε_{it}.

We also directly compared the BKMRvs and BKMRhvs approaches when applied to a large number of correlated pollutants. Specifically, we fit the model

$$Y_{it} = h\left(\mathbf{z}_{it}\right) + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + b_i + \varepsilon_{it}, \quad (11)$$

where $\mathbf{z}_{it}$ contained the $M = 13$ constituents used in Simulation Study 2 (i.e., $\mathbf{z}_{it} = [\mathrm{Al}_{it}, \mathrm{Si}_{it}, \mathrm{Ti}_{it}, \mathrm{Ca}_{it}, \mathrm{K}_{it}, \mathrm{Cu}_{it}, \mathrm{Mn}_{it}, \mathrm{Ni}_{it}, \mathrm{V}_{it}, \mathrm{Zn}_{it}, \mathrm{S}_{it}, \mathrm{Cl}_{it}, \mathrm{BC}_{it}]$) and $\mathbf{x}_{it} = [\mathbf{I}(\text{post-occlusion})_{it}, \mathbf{I}(\text{post-prazosin})_{it}, \mathbf{I}(\text{CAPs})_{it}]^{\top}$ contained the same variables used in the $M = 7$ analysis of the canine data. We began analyses with the Simulation Study 2 pollutant groups of the 13 constituents: S_1 = Al, Si, Ti, and Ca; S_2 = Ni, V, and Zn; S_3 = S; S_4 = BC; S_5 = Cu; S_6 = K; S_7 = Cl; and S_8 = Mn. However, because in this small subsample of days, K, Cu, and Mn were also highly correlated with Al, Si, Ti, and Ca, we merged groups S_1, S_5, S_6, and S_8 to form a new group S_1 (all pair-wise correlations in this new S_1 were > 0.76). This S_1 group of elements has been documented to collectively represent pollution from road dust and crustal source categories in the Boston area (Clarke et al. 2000). Therefore, the final groups for this analysis were S_1 = Al, Si, Ti, Ca, K, Cu, and Mn; S_2 = Ni, V, and Zn; S_3 = S; S_4 = BC; S_5 = Cl.

RESULTS

Component-Wise Variable Selection on a Subset of Pollutants

As done for the analysis of the MOBILIZE data, we first present results from standard mixed-model analyses. That is, we fit one-pollutant versions of model (10) and multipollutant models containing linear functions of all seven chosen constituents; then we applied the BKMRvs and BKMRhvs models. In conducting all three steps of this strategy, after controlling for an overall CAPs effect, pre-vs-post occlusion, and prazosin exposure, no evidence was found of effects of $PM_{2.5}$ composition on SBP, DBP, mean pressure, pulse pressure, or R × P (data not shown). We did find, however, strong evidence of an effect of the multipollutant exposure on heart rate. We therefore focus on the results for this outcome.

Table 6 presents the results from the single- and seven-pollutant linear mixed models. Results from both analyses provide strong evidence of an exposure–response association between Mn concentration and heart rate. Because these models assume linear and additive associations between the response and the individual constituent concentrations, we checked this assumption in two ways. First, we fit model (10) without the exposure term $h(\mathrm{Ni}_{it}, \mathrm{Cu}_{it}, \mathrm{Zn}_{it}, \mathrm{S}_{it}, \mathrm{Ti}_{it}, \mathrm{Mn}_{it}, \mathrm{BC}_{it})$ and plotted the residuals against Mn concentration. The resulting residual plot (shown in Figure 22 with a smoothed local regression [LOESS] fit added to reflect the degree of linearity) suggests that the linearity assumption is plausible for this heart rate–Mn association.

Second, we fit the mixed BKMRvs model in equation (10). Figure 23 shows the PIPs of variables for the seven constituents. These probabilities were calculated while taking as the hyperparameters of the Gamma component of the mixture prior for r_m values estimated by the frequentist KMR approach of Liu and colleagues (2007) divided by 10. Figure 24

shows how these inclusion probabilities change if one uses hyperparameter values equal to half those suggested by the frequentist KMR approach. A comparison of the two sets of estimates (Figures 23 and 24) shows that the relative rankings of the inclusion probabilities are insensitive to this change, although the absolute magnitude of the probabilities are increased as the hyperparameter values are decreased. The results of this analysis confirm that, under very general assumptions for the effect of the multipollutant mixture, the evidence is strongest for a Mn effect, and Zn is estimated to have a slightly lower probability of inclusion.

Figure 25 shows the bivariate exposure–response relationship; points were deleted from the image if they were beyond the range of the data. The results suggest a linear

Table 6. Estimated Regression Coefficients, Standard Errors, and *P* Values from Linear Mixed Models Applied to Heart Rate Data from the Harvard Chan School Canine Study[a]

	Single-Pollutant Model			Seven-Pollutant Model		
Pollutant	$\hat{\beta}$	SE $(\hat{\beta})$	*P* Value	$\hat{\beta}$	SE $(\hat{\beta})$	*P* Value
Ni	−0.17	0.87	0.84	−1.25	1.01	0.22
Cu	2.12	0.99	0.03	0.88	1.74	0.62
Zn	0.53	2.08	0.80	−3.41	2.58	0.19
S	−0.35	0.65	0.59	−0.13	0.83	0.88
Ti	1.82	0.83	0.03	−0.11	1.29	0.93
Mn	2.77	0.99	0.006	3.59	1.80	0.05
BC	0.43	0.89	0.63	0.27	1.16	0.81

[a] The single-pollutant models used the 7-day moving average of each constituent's concentration; the seven-pollutant model included all seven pollutants simultaneously.

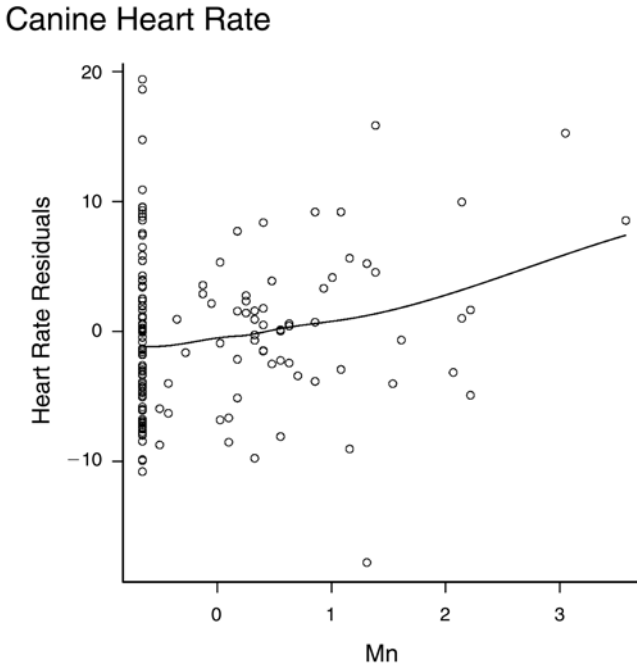


Figure 22. Residuals from the linear mixed model that contained the confounders and random effects only (no exposure terms) plotted as a function of Mn, as applied to heart rate in the Harvard Chan School canine data.

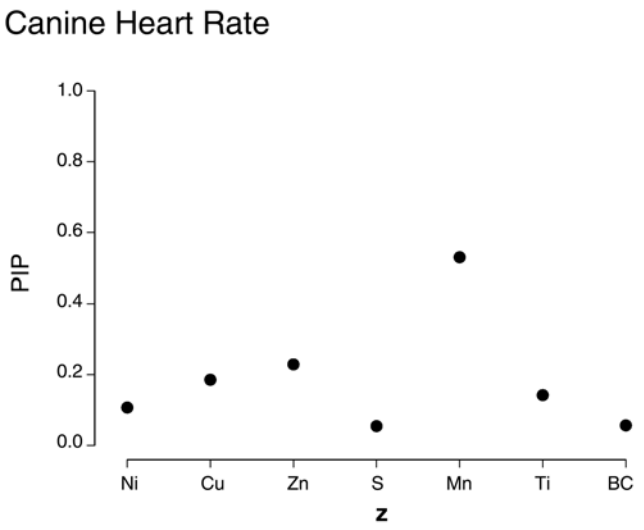


Figure 23. PIPs for the pollutants in the z vector from BKMR analysis of the Harvard Chan School canine data for the CAPs effects on heart rate, using prior hyperparameters equal to those from a frequentist KMR fit divided by 10.

exposure–response relationship between Mn concentrations and heart rate at all levels of Zn. Due to the smaller number of data available in the canine study compared with the MOBILIZE data set, Figure 26 shows cross-sections of the image in Figure 25 calculated at the 25th, 50th, and 75th percentiles (compared with 10th, 50th, and 90th used with the MOBILIZE data) of Zn; corresponding point-wise 95% credible intervals around these conditional exposure–response relationships (conditional at the other six constituent concentrations being set at specific values) are shown; the values of Zn are demarcated by the horizontal lines in the image plot shown in Figure 25. Taken together, these analyses provide strong evidence of an effect of Mn concentration on heart rate and the effect appears to be linear with the exposure concentration and relatively constant across concentrations of the other constituents.

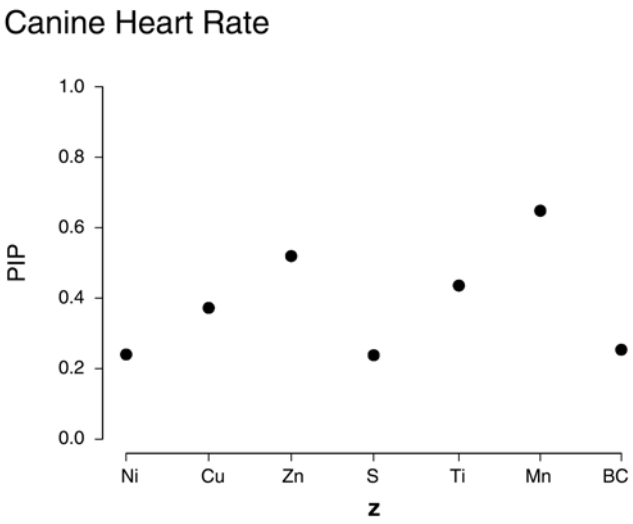


Figure 24. PIPs for the pollutants in the z vector from BKMR analysis of the Harvard Chan School canine data for the CAPs effects on heart rate, using prior hyperparameters equal to half those used for Figure 23.

Component-Wise Versus Hierarchical Variable Selection

In the analyses that used a larger number of pollutants ($M = 13$), we found that, although each pollutant had a PIP of < 0.4 under the BKMRvs approach, group S_1 had a PIP of 0.79 under the BKMRhvs approach (Figure 27). Given the strong correlations among pollutants in group S_1 (Al, Si, Ti, Ca, K, Cu, and Mn), the data did not strongly favor one constituent over the others as driving the observed association between heart rate and this group of constituents (the conditional inclusion probabilities ranged from 0.04 for Cu to 0.36 for Si). In this case, our strong preference is the two-level BKMRhvs approach, because it accurately conveys that a group of constituents is associated with the outcome of interest, but that the data cannot definitively identify a single constituent as driving this association. In contrast, a single-level variable-selection approach (BKMRvs) gives the mistaken impression that no association exists between the outcome and the mixture.

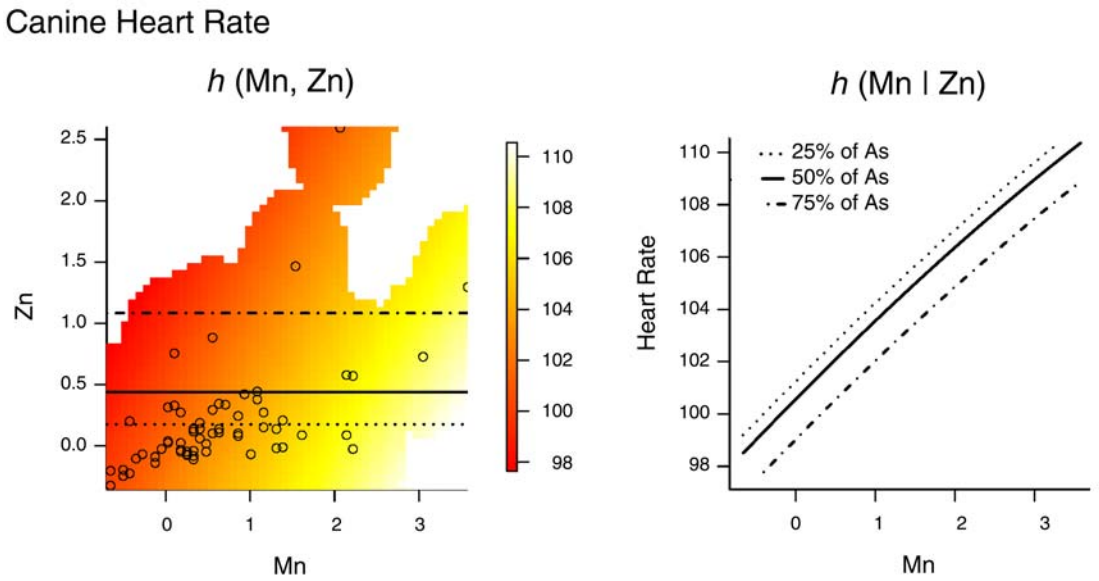


Figure 25. Plot of $\hat{h}$ as a function of Mn and Zn for heart rate, evaluated at the median concentrations of the other five pollutants in the model (Ni, Cu, S, Ti, BC).

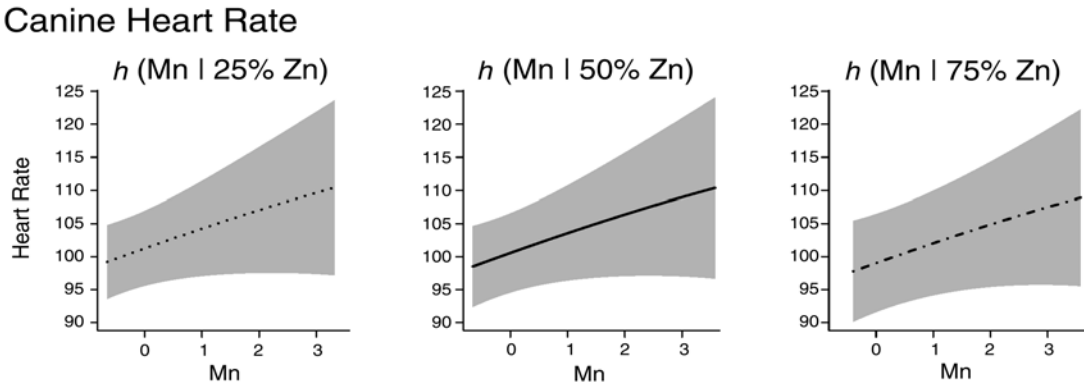


Figure 26. Plot of $\hat{h}$ as a function of Mn for heart rate, and the associated 95% pointwise credible intervals, evaluated at the 25th, 50th, and 75th percentiles of Zn and the median concentrations of the other five pollutants in the model (Ni, Cu, S, Ti, BC).

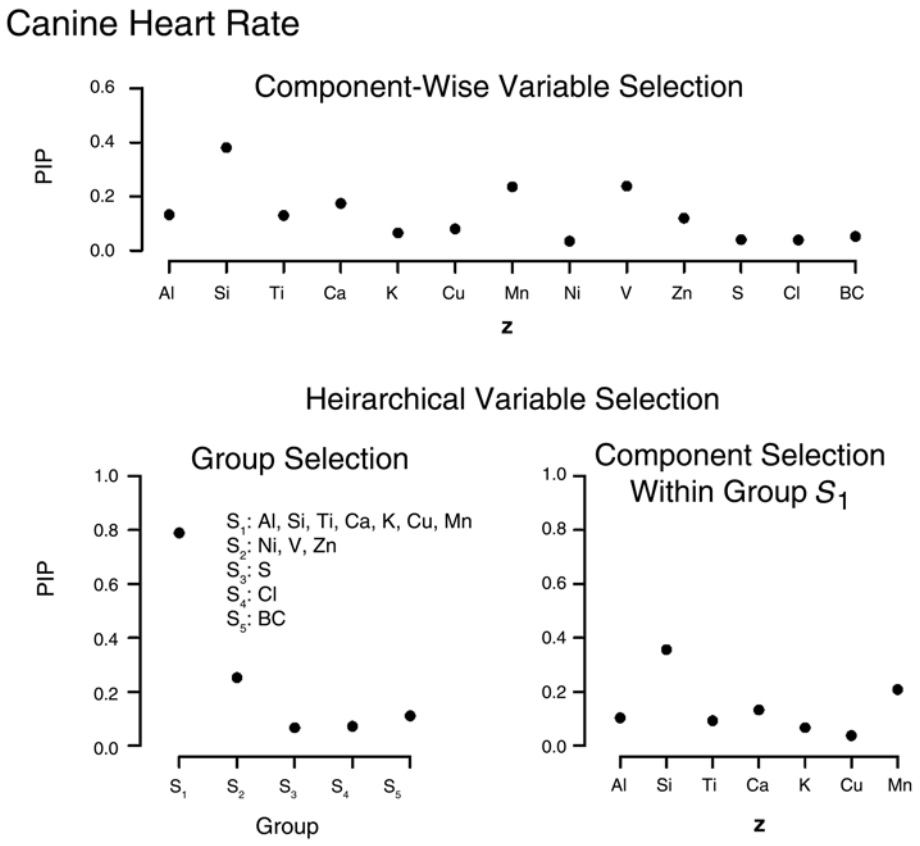


Figure 27. PIPs for the toxicologic canine data estimated from BKMRvs (top) and BKMRhvs (bottom). The upper panel shows the pollutant-specific PIPs. The lower left panel shows the source category–specific PIPs; the lower right panel shows conditional PIPs for the pollutants in group S_1. Note that the scales on the y axes differ.

DISCUSSION

We have proposed a new approach to estimating the health effects of multipollutant mixtures while simultaneously identifying which of the exposures are driving the association. We conducted three simulation studies to demonstrate the operating characteristics of BKMR, both in terms of identifying important pollutants and in its ability to estimate the form of the exposure–response relationship. Previously, KMR has been applied to genetic data, in which a Garrote kernel has been used to identify particular genes associated with a health outcome while allowing for gene–gene interaction within a frequentist KMR treatment of the model (Maity and Lin 2011). This approach could be applied to environmental mixtures; however, because it tests each pollutant sequentially, to estimate the health effects of the mixture one would then need a second-stage model that includes only the pollutants found to be important (e.g., pollutants with P value < 0.05). Our simulations suggest that, particularly in settings in which pollutant concentrations are highly correlated, the BKMR did a better job of fully reflecting the uncertainty associated with identifying the important pollutants within the mixture and propagating this uncertainty through the estimation of health effects. By considering a Bayesian paradigm, we were able to perform variable selection and estimation in one analytic stage, and could therefore accurately capture the uncertainty associated with the final health effect estimates.

This work provides several contributions to the kernel regression literature. First, to our knowledge this is the first time KMR methods have been used for estimating the health effects of multipollutant mixtures. Unlike previous studies that focused mostly on variable selection and prediction, this project's major goal was to estimate the exposure–response function. Second, we developed a novel hierarchical variable-selection approach within BKMR that is able to account for the structure among pollutants in the mixture and systematically handle highly correlated exposures. Third, we conducted simulation studies based on real multipollutant data sets, which allowed for a tailored evaluation of the performance of BKMR in realistic scenarios with complex correlation structures. Finally, our work adds to the literature more generally in terms of variable-selection methods because the proposed methods are a viable alternative for variable selection in longitudinal and other correlated data settings, for which few methods are readily available (the R package *lmmlasso* being one notable exception).

One of the main goals of this project was to analyze both epidemiologic and toxicologic data on the health effects of air pollution mixtures focusing on a common mechanism (hemodynamics) and using a common statistical method for data obtained in the same geographic location. We analyzed BP endpoints in the data from the Boston-based MOBILIZE study and from the Harvard Chan School canine study. The results from two sets of analyses did not show a common set of effects across the same endpoints.

The MOBILIZE analyses yielded evidence of a linear and additive association between BC and Cu exposures for standing DBP, and a linear association of S with standing SBP. Cu and BC are used as markers of traffic contributions to air pollution and S is used as a marker of power plant emissions or regional (or long-range transported) air pollution. Therefore, these analyses suggest that emissions from these three source categories were most strongly associated with hemodynamic responses in this cohort.

In contrast, in the Harvard Chan School canine study, we did not observe any associations between DBP or SBP and any elemental concentrations (after controlling for an overall effect of CAPs exposure). Instead, we observed strong evidence of an association between Mn and heart rate in which heart rate increased linearly with increasing concentrations of Mn. According to the PMF source apportionment analyses of the XRF data from the Harvard Chan School Boston Supersite (Table 4), which is located next to the Harvard Chan School animal exposure facilities, Mn loads on the factors that represent the mobile and road dust source categories, both of which are also related to traffic. The results of the BKMR analyses were similar to those from existing linear mixed-model analyses of these data in that the effect had a linear and additive form that is straightforward to detect with standard statistical methods.

There are several possible reasons why the data analyses did not provide any evidence of nonlinearities or interactions among PM constituents in these studies. The signals may have been too small and the data sufficiently noisy that it was difficult to pick up any nonlinearities or interactions that were in fact present. Or it may be that for the MOBILIZE and CAPs toxicologic applications considered here, the true exposure–response function is linear and additive. We note that the BKMR method has been able to detect nonlinearities and interactions in other multipollutant settings, in particular that of metal mixtures and neurodevelopment (as measured in cord blood for infants; Bobb et al. 2014).

We also note that BKMR could be applied to data for exposures to other air pollution mixtures, such as particles and gases, that may also interact with temperature or other climatic factors. We therefore believe there is value in formulating methods that may allow for general exposure–response

associations that may arise in several environmental settings where mixtures of pollutants exist. We could then assess whether evidence exists in a particular data set to support these associations. We may not find such evidence in any given application, but we feel this is preferable to an approach that assumes a priori that such general exposure–response associations are not present.

We have R scripts available for running the proposed BKMR analyses, and are currently building an R package that will make the methods widely available. The factor driving the computational complexity of the proposed approaches is the number of observations in the analysis, since the kernel matrix is $n \times n$ (where n is the number of observations). In this work we were able to easily fit the small data set from the toxicologic study; and although it took a couple of days, we were able to apply the model to the larger MOBILIZE study data set with our current computing resources. Therefore, for large cohorts or large time-series studies involving tens to hundreds of thousands of observations, computation based on the model-fitting algorithms as currently developed is not feasible. The development of computationally fast methods for big data sets is an area we are actively pursuing. It will also be of interest to extend the Bayesian fitting algorithms to the case of non-normal outcomes, such as binary, count, and time-to-event endpoints. We will include the resulting algorithms into the R package that we are building as part of this project as they become available.

Beyond the benefits of the Bayesian paradigm for fully accounting for the uncertainty in estimating the health effects of the mixture, BKMR has several features that may be particularly appealing for analyzing multipollutant mixtures, which will be topics of future research by our group. One feature, although not given as a specific aim of this research project, is that the method could be extended to quantify evidence of (1) an overall effect of the mixture, or (2) interactions among the pollutants in the mixture. That is, it is of interest to compare the general model (equation 4), the additive model $h(z_{i1},\ldots,z_{iM}) = h_1(z_{i1}) + h_2(z_{i2}) + \ldots + h_M(z_{iM})$, and the null model $h(z_{i1},\ldots,z_{iM}) = 0$. At the current time, one could implement existing frequentist solutions developed in genomic applications to test the null hypothesis $h = 0$ (Liu et al. 2007) against the general model (equation 4), but to our knowledge comparison between the general model and the additive model is an open problem.

Another extremely useful extension of the BKMR model would be to have it accommodate exposure covariates that involve measurement error. Such error could arise in several contexts, including but not limited to (1) known measurement error in the measured concentrations, and (2) the use of location-specific exposure measures that are themselves estimates from a spatiotemporal exposure model. (Such models are designed to address the fact that the measured concentrations are recorded at air monitoring locations that are different from the locations at which the health endpoints for study subjects are measured [Gryparis et al. 2009; Szpiro et al. 2011]). One approach to accommodating measurement errors that arise from such scenarios could treat each constituent concentration z_{im} as a random variable in the Bayesian analysis, and specify a prior distribution for each measured value with standard deviation set as equal to the measurement uncertainty reported for that measurement. In cases in which the exposure estimates are outputs from a spatial exposure model, the prior distribution for each of the exposure variables (z_{i1}, $z_{i2},\ldots,z_{iM}$) for subject i can be taken to be the interim multivariate posterior distribution of these missing exposures, given the observed data (Gryparis et al. 2009).

ACKNOWLEDGMENTS

The authors acknowledge the funding support from the Health Effects Institute that made this research possible, the project scientists Kate Adams and Katherine Walker for their work on the project, and to Virgi Hepner and her team for their editing and production work on the final Research Report.

REFERENCES

Austin E, Coull B, Thomas D, Koutrakis P. 2012. A framework for identifying distinct multipollutant profiles in air pollution data. Environ Int 45:112–121.

Banerjee S, Gelfand A, Finley A, Sang H. 2008. Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B Stat 70:825–848.

Bartoli CR, Wellenius GA, Coull BA, Akiyama I, Diaz EA, Lawrence J, et al. 2009. Concentrated ambient particles alter myocardial blood flow during acute ischemia in conscious dogs. Environ Health Perspect 117:333–337.

Billionnet C, Sherrill D, Annesi-Maesano I. 2012. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol 22:126–141.

Bobb JF, Valeri L, Claus-Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. 2014. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics doi:10.1093/biostatistics/kxu058.

Breiman L. 2001. Random forests. Machine Learning 45:5–32.

Brook JR, Demerjian KL, Hidy G, Molina LT, Pennel WT, Scheffe R. 2009. New directions: results-oriented multi-pollutant air quality management. Atmos Environ 43:2091–2093.

Cai T, Tonini G, Lin X. 2011. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67:975–986.

Chatterjee A, Lahiri SN. 2011. Bootstrapping LASSO estimators. J Am Stat Assoc 494:608–625.

Clarke RW, Coull BA, Reinisch U, Catalano PJ, Killingsworth CR, Koutrakis P, et al. 2000. Inhaled concentrated ambient particles are associated with hematologic and bronchoalveolar lavage changes in canines. Environ Health Perspect 108(12):1179–1187.

Claus Henn B, Ettinger AS, Schwartz J, Tellez-Rojo MM, Lamadrid-Figueroa H, Hernandez-Avila M, et al. 2010. Early postnatal blood manganese levels and children's neurodevelopment. Epidemiology 21:433–439.

Claus Henn B, Schnaas L, Ettinger AS, Schwartz J, Lamadrid-Figueroa H, Hernandez-Avila M, et al. 2012. Associations of early childhood manganese and lead coexposure with neurodevelopment. Environ Health Perspect 120:126–131.

Coull BA, Wellenius GA, Gonzalez-Flecha B, Diaz EA, Koutrakis P, Godleski JJ. 2011. The toxicologic evaluation of realistic emissions of source aerosols (TERESA): statistical methods. Inhal Toxicol 23:31–41.

Cristianini N, Shawe-Taylor J. 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK:Cambridge University Press.

Dominici F, Peng RD, Barr CD, Bell ML. 2010. Protecting human health from air pollution shifting from a single-pollutant to a multipollutant approach. Epidemiology 21:187–194.

George EI, McCulloch RE. 1993. Variable selection via Gibbs sampling. J Am Stat Assoc 88:881–889.

Godleski JJ, Verrier RL, Koutrakis P, Catalano P. 2000. Mechanisms of Morbidity and Mortality from Exposure to Ambient Air Particles. Research Report 91. Cambridge, MA:Health Effects Institute.

Greenbaum D, Shaikh R. 2010. First steps toward multi-pollutant science for air quality decisions. Epidemiology 21:195–197.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10:258–274.

Hastie T, Tibshirani R, Friedman JH. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed. New York:Springer.

Hidy GM, Pennell WT. 2010. Multipollutant air quality management. J Air Waste Manage Assoc 60:645–654.

Ito K, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, et al. 2006. PM source apportionment and health effects: 2. an investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. J Expo Sci Environ Epidemiol 16:300–310.

Kioumourtzoglou MA, Coull BA, Dominici F, Koutrakis P, Schwartz J, Suh HH. 2014. The impact of source contribution uncertainty on the effects of source-specific $PM_{2.5}$ on hospital admissions: a case study in Boston MA. J Expo Sci Environ Epidemiol 24:365–371.

Lall R, Ito K, Thurston GD. 2011. Distributed lag analyses of daily hospital admissions and source apportioned fine particle air pollution. Environ Health Perspect 119:455–460.

Liaw A, Wiener M. 2002. Classification and regression by random forest. Available at *http://CRAN.R-project.org/doc/Rnews/*. R News 2:18–22.

Linkletter C, Bingham D, Hengartner N, Higdon D, Kenny QY. 2006. Variable selection for Gaussian process models in computer experiments. Technometrics 48:478–90.

Liu D, Ghosh D, Lin X. 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9:292.

Liu D, Lin X, Ghosh D. 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63:1079–1088.

Maity A, Lin X. 2011. Powerful tests for detecting a gene effect in the presence of possible gene–gene interactions using Garrote kernel machines. Biometrics 67:1271–1284.

Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet J, Stieb DM, et al. 2010. Is the air pollution health

research community prepared to support a multipollutant air quality management framework? Inhal Toxicol 22:1–19.

Mikl M, Marecek R, Hlustik P, Pavlicova M, Drastich A, Chlebus P, Brazdil M, Krupa P. 2008. Effects of spatial smoothing on fMRI group inferences. Magn Reson Imaging 26:490–503.

National Research Council. 2004. Research Priorities for Airborne Particulate Matter. IV. Washington, DC:National Academy Press.

Nikolov MC, Coull BA, Catalano PJ, Diaz E, Godleski JJ. 2008. Statistical methods to evaluate health effects associated with major sources of air pollution: A case study of breathing patterns during exposure to concentrated Boston air particles. J R Stat Soc Ser C 57:357–378.

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.

Pattenden S, Armstrong B, Milojevic A, Heal MR, Chalabi Z, Doherty R, et al. 2010. Ozone, heat and mortality: acute effects in 15 British conurbations. Occup Environ Med 67:699–707.

Qian Z, He Q, Lin H-M, Kong L, Zhou D, Liang S, et al. 2010. Part 2. Association of daily mortality with ambient air pollution, and effect modification by extremely high temperature in Wuhan, China. In: Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Four Cities. Research Report 154. Boston, MA:Health Effects Institute.

Ren C, O'Neill MS, Park SK, Sparrow D, Vokonas P, Schwartz J. 2011. Ambient temperature, air pollution, and heart rate variability in an aging population. Am J Epidemiol 173:1013–1021.

Rooke GA, Feigl EO. 1982. Work as a correlate of canine left ventricular oxygen consumption, and the problem of catecholamine oxygen wasting. Circ Res 50(2):273–286.

Savitsky T, Vannucci M, Sha N. 2011. Variable selection for nonparametric Gaussian process priors: models and computational strategies. Stat Sci 26:130–149.

Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, Davies N, et al. 2004. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. Biometrics 60:812–819.

Sioutas C, Koutrakis P, Burton RM. 1995. A technique to expose animals to concentrated fine ambient aerosols. Environ Health Perspect 103:172–177.

Szpiro AA, Sheppart L, Lumley T. 2011. Efficient measurement error correction with spatially misaligned data. Biostatistics 12:610–623.

Tibshirani R. 1994. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288.

Vedal S, Kaufman J. 2011. What does multi-pollutant air pollution research mean? Am J Respir Crit Care Med 183:4–6.

Wellenius GA, Coull BA, Godleski JJ, Koutrakis P, Okabe K, Savage ST, et al. 2003. Inhalation of concentrated ambient air particles exacerbates myocardial ischemia in conscious dogs. Environ Health Perspect 111:402–408.

Wellenius GA, Wilhelm-Benartzi CS, Wilker EH, Coull BA, Suh HH, Koutrakis P, et al. 2012. Ambient particulate matter and the response to orthostatic challenge in the elderly: the Maintenance of Balance, Independent Living, Intellect, and Zest in the Elderly (MOBILIZE) of Boston study. Hypertension 59:558–563.

Wood S. 2006. An Introduction to Generalized Additive Models: An Introduction with R. London:Chapman & Hall/CRC Press.

Zanobetti A, Austin E, Coull BA, Schwartz J, Koutrakis P. 2014. Health effects of multi-pollutant profiles. Technical Report, 2013. Environ Int 71:13–19.

Zanobetti A, Schwartz J. 2008. Temperature and mortality in nine US cities. Epidemiology 19:563–570.

Zou F, Huang H, Lee S, Hoeschele I. 2010. Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene–environment interaction. Genetics 186:385–394.

HEI QUALITY ASSURANCE STATEMENT

The conduct of this research project was subjected to independent quality assurance (QA) oversight by Abt Associates. The audits were led by Dr. Sue Greco, who has over 15 years of experience in human health risk assessment of $PM_{2.5}$. The QA oversight consisted of a First QA Audit (focused on organizational structure; quality of air pollution, toxicologic, and human health data; and protection of animals and human subjects) and a Final QA Audit (focused on the Investigators' Final Report). The dates of the QA audits and activities are summarized below. Both audits were conducted in Boston, MA, at the Harvard Chan School and the Beth Israel Deaconess Medical Center (BIDMC), with follow up questions by telephone and email.

January 17–18, 2013. First QA Audit conducted on-site at the Harvard Chan School and BIDMC

This "readiness review" audit was intended to review the standard operating procedures and data management practices used in the research to ensure that these procedures were followed consistently by all members of the research team. The auditors met with Dr. Coull and team members (Mittleman, Godleski, Wellenius, Bobb, and Diaz) at the Harvard Chan School or BIDMC. All data for this simulation study had been collected previously under other grants. The auditors observed relevant IRB documents from the Harvard Chan School, BIDMC, and Brown University (Dr. Wellenius' affiliation) for the MOBILIZE human blood pressure data. The animal protocol was closed at the start of 2013, but the researchers could still use the previously collected dog blood pressure data in the analyses. The auditors found that all study team members were well-qualified to conduct the research and that sufficient levels of oversight had been implemented in the study.

May 23, 2014. Final QA Audit conducted on-site at the Harvard Chan School

In the Final QA Audit, the researchers were asked to indicate how approximately 12 tables and figures from the August 2013 version of the Investigators' Final Report were generated, starting with the raw data. (The tables and figures were preselected.) The researchers who created the tables and figures were Drs. Coull, Bobb, and Kioumourtzoglou. Since Dr. Kioumourtzoglou was out of the country, the auditors followed up with her separately for one table and figure related to source apportionment (see June 19 entry). For the other tables and figures, Drs. Coull and Bobb indicated where the raw data were located, how the data were processed, and how the final tables and figures were generated. The auditors visually inspected the model output to ensure that it matched what was in the Investigator's Final Report; no issues impacting the study conclusions were identified. The auditors recommended that all files pertaining to this research effort be archived under one HEI project folder once the Investigators' Final Report was finalized.

June 19, 2014. Final QA Audit completed by teleconference

A telephone meeting was held with the auditors and Dr. Kioumourtzoglou to discuss the source apportionment table and figure she had generated for the Investigators' Final Report. No issues were identified with the generation of the table and figure.

Overall, the auditors found the researchers to be well-organized and cooperative during the audits. The study procedures, analysis steps, and data storage were systematic, consistent, and well-designed to manage the various data and analytical streams necessary to complete the study.

Sue Greco, Sc.D.

APPENDIX A. ESTIMATION AND PREDICTION

Here we detail the MCMC sampler used to fit BKMR with variable selection.

In order to apply a standard Gibbs sampler in which samples are generated from the full conditional distributions of each of the parameters, the augmented kernel matrix $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ (equation 6 in the main report) must be inverted at each iteration of the sampler, which can lead to numeric instability if the kernel is nearly singular (see discussion in Zou et al. 2010). This problem can be avoided by integrating out $\mathbf{h}$, and obtaining posterior samples from the marginal posterior distribution of the remaining parameters.

SUMMARY OF BKMR AND VARIABLE SELECTION

Model Specification

Likelihood (Kernel Machine Regression section):

$$\begin{cases} \mathbf{y} \mid \mathbf{h},\boldsymbol{\beta},\sigma^2,\mathbf{X} \sim \mathrm{N}\left(\mathbf{h}+\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n\right) \\ \mathbf{h} \mid \tau,\mathbf{r},\mathbf{Z} \sim \mathrm{N}\left(\mathbf{0},\tau\mathbf{K}_{\mathbf{Z},\mathbf{r}}\right) \end{cases}$$

Variable Selection (Component-Wise Variable Selection section):

$$\begin{cases} r_m \mid \delta_m \sim \delta_m \mathrm{Gamma}\left(a_r,b_r\right)+\left(1-\delta_m\right)P_0 \\ \delta_m \mid \pi \sim \mathrm{Bernoulli}(\pi) \end{cases}$$

Priors (Prior Specification section):

$$\begin{cases} \boldsymbol{\beta} \sim 1 \\ \sigma^{-2} \sim \mathrm{Gamma}(a_\sigma,b_\sigma) \\ \lambda \equiv \tau\sigma^{-2} \sim \mathrm{Gamma}(a_\lambda,b_\lambda) \\ \pi \sim \mathrm{Beta}(a_\pi,b_\pi) \end{cases}$$

Notation

Indices:
- $i = 1,\ldots,n$ — Subjects
- $m = 1,\ldots,M$ — Pollutants

Data:
- $\mathbf{y} = (y_1,\ldots,y_n)^\mathsf{T}$ — Health outcomes
- $\mathbf{X}$ — Covariate design matrix with rows $\mathbf{x}_i^\mathsf{T}$
- $\mathbf{Z}$ — Exposure design matrix with rows $\mathbf{z}_i^\mathsf{T} = \left(z_{i1},\ldots,z_{iM}\right)$

Parameters:
- $\mathbf{h} = (h_1,\ldots,h_n)^\mathsf{T}$ — Subject-specific health effects $h_i = h(\mathbf{z}_i)$
- $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ — $n \times n$ kernel matrix for variable selection with (i,j)-element $\exp\left\{-\sum_{m=1}^{M} r_m(z_{im}-z_{jm})^2\right\}$
- $\mathbf{r} = (r_1,\ldots,r_M)^\mathsf{T}$ — Augmented variables in kernel matrix for variable selection, which controls smoothness of $h(\cdot)$
- $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_M)^\mathsf{T}$ — Inclusion indicators for pollutants

MCMC SAMPLER

Integrating over π (which is not of interest) and $\mathbf{h}$ and applying the prior distributions specified in the Prior Specification section under Kernel Machine Regression in the main report, the posterior is given by

$$\begin{aligned} & f\left(\boldsymbol{\beta},\sigma^2,\lambda,\mathbf{r},\boldsymbol{\delta}\mid\mathbf{y},\mathbf{X},\mathbf{Z}\right) \\ & \propto \Gamma\left(\sum_m \delta_m + a_\pi\right)\Gamma\left(M-\sum_m \delta_m + b_\pi\right) \\ & \quad \times \mathrm{N}\left(\mathbf{y}\mid\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}\right) \\ & \quad \times \left\{\prod_{m=1}^{M} f\left(r_k\mid\delta_k\right)\right\}\mathrm{Gamma}\left(\sigma^{-2}\mid a_\sigma,b_\sigma\right) \\ & \quad \times \mathrm{Gamma}\left(\lambda\mid a_\lambda,b_\lambda\right), \end{aligned} \tag{12}$$

where $\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}} = \mathbf{I}_n + \lambda\mathbf{K}_{\mathbf{Z},\mathbf{r}}$, and $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ is the augmented kernel matrix for variable selection (which depends on $\mathbf{Z}$ and $\mathbf{r}$) defined in equation (6) in the main report.

We updated $\boldsymbol{\beta}$ and σ^2 using separate Gibbs steps, with full conditionals given by

$$\boldsymbol{\beta}\mid\sigma^2,\lambda,\mathbf{r},\boldsymbol{\delta},\mathbf{y},\mathbf{X},\mathbf{Z} \sim \mathrm{N}\left(\boldsymbol{\beta}\mid\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}^{\mathsf{T}}\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}\mathbf{y},\mathbf{V}_{\boldsymbol{\beta}}\right),$$

$$\mathbf{V}_{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}\mathbf{X}\right)^{-1}, \text{ and}$$

$$\sigma^{-2}\mid\boldsymbol{\beta},\lambda,\mathbf{r},\boldsymbol{\delta},\mathbf{y},\mathbf{X},\mathbf{Z} \sim \mathrm{Gamma}\left\{a_\sigma + n/2, b_\sigma + WSS_{\boldsymbol{\beta},\lambda,\mathbf{r}}/2\right\},$$

where $WSS_{\boldsymbol{\beta},\lambda,\mathbf{r}}$ is the weighted sum of squares

$$\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right).$$

We updated λ using a Metropolis-Hastings step, where the full conditional is given by

$$\begin{aligned} & f\left(\lambda\mid\boldsymbol{\beta},\mathbf{r},\boldsymbol{\delta},\mathbf{y},\mathbf{X},\mathbf{Z}\right) \\ & \propto \left(\det\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}\right)^{-1/2}\exp\left\{-WSS_{\boldsymbol{\beta},\lambda,\mathbf{r}}/\left(2\sigma^2\right)\right\}. \end{aligned}$$

We used a Gamma proposal distribution with mean set to the value of λ from the previous iteration and variance tuned to produce a good acceptance rate.

Because sampling individually from the full conditionals of $\mathbf{r}$ and $\boldsymbol{\delta}$ leads to a reducible Markov chain, we instead sampled $(\mathbf{r}, \boldsymbol{\delta})$ jointly by adapting the Metropolis-Hastings algorithm from Sha and colleagues (2004). To obtain a sample at the sth iteration of the MCMC, this procedure generated a proposal $(\mathbf{r}^*, \boldsymbol{\delta}^*)$ by randomly selecting one of the following moves:

1. Randomly select $m \in \{1,\ldots,M\}$ and set $\delta_m^* = 1 - \delta_m^{(s-1)}$. If $\delta_m^* = 0$, set $r_m^* = 0$; else, generate the proposal r_m^* from a proposal distribution with density $q_1(\cdot)$.

2. Among the components of $\boldsymbol{\delta}^{(s-1)} = 1$, randomly choose one (say δ_m^*) and generate the corresponding r_m^* from a proposal distribution with density $q_2\left(\cdot\mid r_m^{(s-1)}\right)$.

We considered q_1 to be a Gamma density with mean 1 and SD 2, and q_2 to be a Gamma density with mean set to the value at the previous iteration, $r_m^{(s-1)}$, and variance tuned to have a good acceptance rate for those iterations where move 2 (above) was selected.

ESTIMATING SUBJECT-SPECIFIC HEALTH EFFECTS

To obtain posterior samples of h_i, which represents the subject-specific association between exposure to the environmental mixture and health, first note that the posterior density $f(\mathbf{h}, \boldsymbol{\beta}, \sigma^2, \lambda, \mathbf{r}, \boldsymbol{\delta} \mid \mathbf{y}, \mathbf{X}, \mathbf{Z})$ can be decomposed in the usual way as $f(\mathbf{h} \mid \boldsymbol{\beta}, \sigma^2, \lambda, \mathbf{r}, \boldsymbol{\delta}, \mathbf{y}, \mathbf{X}, \mathbf{Z}) \times f(\boldsymbol{\beta}, \sigma^2, \lambda, \mathbf{r}, \boldsymbol{\delta} \mid \mathbf{y}, \mathbf{X}, \mathbf{Z})$, where the conditional distribution of $\mathbf{h}$ is given by

$$\begin{aligned} & \mathbf{h}\mid\boldsymbol{\beta},\sigma^2,\lambda,\mathbf{r},\boldsymbol{\delta},\mathbf{y},\mathbf{X},\mathbf{Z} \\ & \sim \mathrm{N}\left[\lambda\mathbf{K}_{\mathbf{Z},\mathbf{r}}\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right),\sigma^2\lambda\mathbf{K}_{\mathbf{Z},\mathbf{r}}\mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}\right]. \end{aligned} \tag{13}$$

Therefore for each sample $[\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}, \lambda^{(s)}, \mathbf{r}^{(s)}, \boldsymbol{\delta}^{(s)}]$ generated from the marginal posterior in equation (12) with our MCMC sampling algorithm, we generated a sample $\mathbf{h}^{(s)}$ from its full conditional equation (13).

PREDICTING HEALTH EFFECTS AT NEW EXPOSURE PROFILES

A critical aim in analyzing the health effects of environmental mixtures is to estimate (and make visible) the exposure–response surface. This entails not only estimating $h_i = h(\mathbf{z}_i)$ at the observed data points, but also predicting h at a collection of unobserved exposure profiles, $\mathbf{z}_1^{new},\ldots,\mathbf{z}_P^{new}$. Let $\mathbf{Z}_{new}$ be the $P \times M$ design matrix (with rows) $\mathbf{z}_p^{new}$ of new exposure profiles, and let $\mathbf{h}_{new}^{\mathsf{T}} = \left(h_1^{new},\ldots,h_P^{new}\right)$ denote the desired predictions. In the mixed-model representation of KMR, we can consider the joint distribution of the observed and new exposure profiles as

$$\begin{pmatrix}\mathbf{h}\\ \mathbf{h}_{new}\end{pmatrix} \sim \mathrm{N}\left\{\mathbf{0},\tau\begin{pmatrix}\mathbf{K}_{\mathbf{Z},\mathbf{r}} & \mathbf{K}_{\mathbf{Z},\mathbf{Z}_{new},\mathbf{r}}\\ \mathbf{K}_{\mathbf{Z},\mathbf{Z}_{new},\mathbf{r}}^{\mathsf{T}} & \mathbf{K}_{\mathbf{Z}_{new},\mathbf{r}}\end{pmatrix}\right\},$$

where $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ denotes the augmented kernel matrix defined in equation (6), $\mathbf{K}_{\mathbf{Z},\mathbf{Z}_{new},\mathbf{r}}$ is the $n \times P$ matrix with (i,j)-element $\exp\left\{-\sum_{m=1}^{M} r_m\left(z_{im} - z_{jm}^{new}\right)^2\right\}$, and $\mathbf{K}_{\mathbf{Z}_{new},\mathbf{r}}$ is the $P \times P$

matrix with (i,j)-element $\exp\left\{-\sum_{m=1}^{M} r_m\left(z_{im}^{new} - z_{jm}^{new}\right)^2\right\}$.

Following routine calculations, the conditional posterior distribution of $\mathbf{h}_{new}$ is given by

$$\mathbf{h}^{new} \mid \boldsymbol{\beta}, \sigma^2, \lambda, \mathbf{r}, \boldsymbol{\delta}, \mathbf{y}, \mathbf{X}, \mathbf{Z} \sim \mathrm{N}\left\{\begin{array}{l}\lambda \mathbf{K}_{\mathbf{Z},\mathbf{Z}_{new},\mathbf{r}}^{\mathsf{T}} \mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\ \sigma^2\lambda\left(\mathbf{K}_{\mathbf{Z}_{new},\mathbf{r}} - \lambda \mathbf{K}_{\mathbf{Z},\mathbf{Z}_{new},\mathbf{r}}^{\mathsf{T}} \mathbf{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1} \mathbf{K}_{\mathbf{Z},\mathbf{Z}_{new},\mathbf{r}}\right)\end{array}\right\}. \quad (14)$$

In theory, we could obtain predictions by generating $\mathbf{h}^{new}$ from its conditional distribution and computing the mean and variance of the posterior samples. However, because, in practice, a large number of predictions are typically needed (e.g., to plot cross-sections of the estimated exposure–response surface on a grid of points), this posterior simulation can be very computationally expensive in that it requires repeated simulations from a high-dimensional multivariate normal distribution. Therefore, we propose to approximate the posterior mean (variance) of $\mathbf{h}^{new}$ as its conditional posterior mean (variance) evaluated at the posterior mean of the other parameters.

ABOUT THE AUTHORS

Brent Coull is professor of biostatistics and codirector of the Environmental Statistics Program at the Harvard T. H. Chan School of Public Health. He is a biostatistician with 17 years of experience in a wide range of environmental health research, including work assessing the neurodevelopmental effects of environmental exposures and the development and application of statistical methods for the assessment of the health effects of air pollution. He is principal investigator for the Environmental Statistics and Bioinformatics Core of the Harvard National Institute for Environmental Health Sciences (NIEHS) Center and for the Biostatistics Core of the Harvard EPA Clean Air Research Center. His methodologic research has focused on analytic methods for multiple outcomes, functional data and distributed lag models, health effects of complex environmental mixtures, structural equation models, pathway analyses, and exposure measurement error.

Jennifer F. Bobb is a research associate in the Harvard T. H. Chan School Department of Biostatistics with expertise in statistical methods for estimating the health effects of simultaneous exposure to temperature and ambient air pollution as well as complex environmental mixtures, and in estimating the public health impact of climate change. Bobb earned her Ph.D. in biostatistics from the Johns Hopkins Bloomberg School of Public Health and has been a postdoctoral and research fellow at the Harvard T.H. Chan School since December 2011.

Gregory A. Wellenius is associate professor of epidemiology in the Center of Environmental Health and Technology at Brown University. He has studied environmental determinants of cardiovascular disease, primarily focusing on the effects of ambient air pollution on the risk of cardiovascular events and its effects on cardiovascular physiology. He currently serves as principal investigator of grants funded by the National Institutes of Health (NIH) focusing on associations between ambient air pollution and incident stroke and on the association between residential air pollution and risk of preeclampsia.

Marianthi-Anna Kioumourtzoglou is a research fellow in the Department of Environmental Health at the Harvard T. H. Chan School. Kioumourtzoglou earned her Sc.D. in environmental health from Harvard in 2013. She has expertise in statistical issues related to air pollution epidemiology, such as measurement error induced by use of surrogate instead of true exposures to fine particles; methods to deal with multipollutant exposures, such as hierarchical models, source apportionment, and clustering; and the impact of failure to account for uncertainty associated with estimation of source contributions in analyses that use the output from a source apportionment model as exposures in a health effects analysis.

Murray A. Mittleman is associate professor of epidemiology in the Department of Epidemiology at the Harvard T. H. Chan School, as well as the director of the Cardiovascular Epidemiology Research Unit at Beth Israel Deaconess Medical Center, and associate professor of medicine at Harvard Medical School. Mittleman's group has made important contributions to understanding the role of particulate air pollution in the onset of cardiovascular events. He served as the principal investigator of an NIH-funded project investigating health effects of particle exposures on stroke onset, and is currently principal investigator of a project in the Harvard EPA Clean Air Research Center focusing on health effects of air pollution mixtures on cardiovascular endpoints in cohorts followed by the Framingham Heart Study.

Petros Koutrakis is professor of environmental sciences in the Department of Environmental Health at the Harvard T. H. Chan School, and director of the Harvard EPA Clean Air Research Center. His research activities focus on the development of human exposure measurement techniques and the investigation of sources, transport, and fate of air pollutants. Koutrakis has conducted a number of comprehensive

air pollution studies in the United States, Canada, Spain, Chile, Kuwait, Cyprus, and Greece that investigated the extent of human exposures to gaseous and particulate air pollutants. Other research interests include the assessment of particulate matter exposures and their effects on cardiac and pulmonary health.

John J. Godleski is associate professor of pathology in the Department of Environmental Health at the Harvard T. H. Chan School and at Harvard Medical School. His research focuses on the pulmonary and systemic responses to inhaled ambient particles. His studies use cardiac and pulmonary mechanical measurements as well as cell and molecular biologic approaches with inhalation exposure to concentrated ambient particles. Godleski has served as deputy director of the Harvard EPA Clean Air Research Center for the past 8 years. He has headed the ambient particle research Core in the Harvard NIEHS Center and was principal investigator for a productive NIEHS program project grant.

OTHER PUBLICATIONS RESULTING FROM THIS RESEARCH

Bobb JF, Valeri L, Claus-Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. 2014. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics doi:10.1093/biostatistics/kxu058.

ABBREVIATIONS AND OTHER TERMS

BC	black carbon
BIDMC	Beth Israel Deaconess Medical Center
BKMR	Bayesian kernel machine regression
BKMRvs	Bayesian kernel machine regression with variable selection
BKMRhvs	Bayesian kernel machine regression with hierarchical variable selection
BP	blood pressure
CAPs	concentrated ambient particles
DBP	diastolic blood pressure
EC	elemental carbon
GAM	generalized additive model
HAPC	Harvard Ambient Particle Concentrator
iid	independent and identically distributed
IQR	interquartile range
KMR	kernel machine regression
KMRvs	kernel machine regression with variable selection
LASSO	least absolute shrinkage and selection operator
MCMC	Markov chain Monte Carlo
MOBILIZE	Maintenance of Balance, Independent Living, Intellect, and Zest in the Elderly of Boston study
NIH	National Institutes of Health
OC	organic carbon
PIP	posterior inclusion probability
PM	particulate matter
$PM_{2.5}$	particulate matter $\leq$ 2.5 μm in aerodynamic diameter
PMF	positive matrix factorization
RFA	Request for Applications
R × P	rate–pressure product
SBP	systolic blood pressure
SSVS	stochastic search variable selection
U.S. EPA	United States Environmental Protection Agency
XRF	x-ray fluorescence

ELEMENTS

Ag	silver
Al	aluminum
As	arsenic
Ba	barium
Br	bromine
Ca	calcium
Cd	cadmium
Ce	cerium
Cl	chlorine
Co	cobalt
Cr	chromium
Cs	cesium
Cu	copper
Eu	europium
Hf	hafnium
Hg	mercury
In	indium
K	potassium
Mg	magnesium

Mn	manganese
Na	sodium
Ni	nickel
Pb	lead
S	sulfur
Sb	antimony
Se	selenium
Si	silicon
Sm	samarium
Sn	tin
Sr	strontium
Te	tellurium
Ti	titanium
Tl	thallium
V	vanadium
W	tungsten
Y	yttrium
Zn	zinc
Zr	zirconium

STATISTICAL NOTATION

Y_i	health outcome
μ_i^Y	mean of health outcome Y_i
$\mathbf{z}_i$	an $M \times 1$ vector of pollutant concentrations corresponding to observation i
$\mathbf{x}_i$	a $q \times 1$ vector of variables containing information on potential confounders
$h(\cdot)$	unknown multivariate exposure–response function
$K(\mathbf{z},\mathbf{z}')$	kernel function quantifying the difference between $\mathbf{z}$ and $\mathbf{z}'$
n	number of observations
M	number of pollutants in model
Q	number of pollutants associated with the health outcome
$\boldsymbol{\beta}$	regression coefficients for confounders
b_i	subject-specific random effects in longitudinal mixed effects models
σ_b^2	subject-specific random effects variance
σ^2	residual variance
ρ	spatial correlation parameter in Gaussian kernel function
τ	variance of the kernel-based observation-specific random effects
r_m	pollutant-specific weight in the BKMRvs model
δ_m	indicator variable reflecting whether pollutant m is included in the model
π_m	posterior probability (given the data) that pollutant m is included in the health effects model
S_g	group g of pollutants formed for the BKMRhvs model
ω_g	indicator variable reflecting whether group g of pollutants is included in the BKMRhvs model
$\boldsymbol{\delta}_{S_g}$	vector of conditional indicator variables reflecting which component within group g should be included in the model, conditional on group g being included in the model
$\mathbf{K}_{\mathbf{Z},\mathbf{r}}$	$n \times n$ kernel matrix for variable selection with (i,j)-element $\exp\left\{-\sum_{m=1}^{M} r_m\left(z_{im}-z_{jm}\right)^2\right\}$
a_r, b_r	user-defined hyperparameters for the positive, Gamma portion of the mixture variable selection prior for r_m
π	prior probability of variable inclusion, common to all pollutants
a_λ, b_λ	user-defined hyperparameters for Gamma distribution prior for $\lambda \equiv \tau\sigma^{-2}$
a_σ, b_σ	user-defined hyperparameters for Gamma distribution prior for σ^{-2}
a_π, b_π	user-defined hyperparameters for beta distribution prior for π

INVESTIGATORS' REPORT

Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants

Eun Sug Park, Elaine Symanski, Daikwon Han, and Clifford Spiegelman

Texas A&M Transportation Institute, College Station, Texas, (E.S.P.); University of Texas School of Public Health, Houston, Texas, (E.S.); Texas A&M University, College Station, Texas, (C.S., D.H.)

ABSTRACT

A major difficulty with assessing source-specific health effects is that source-specific exposures cannot be measured directly; rather, they need to be estimated by a source-apportionment method such as multivariate receptor modeling. The uncertainty in source apportionment (uncertainty in source-specific exposure estimates and model uncertainty due to the unknown number of sources and identifiability conditions) has been largely ignored in previous studies. Also, spatial dependence of multipollutant data collected from multiple monitoring sites has not yet been incorporated into multivariate receptor modeling. The objectives of this project are (1) to develop a multipollutant approach that incorporates both sources of uncertainty in source-apportionment into the assessment of source-specific health effects and (2) to develop enhanced multivariate receptor models that can account for spatial correlations in the multipollutant data collected from multiple sites.

We employed a Bayesian hierarchical modeling framework consisting of multivariate receptor models, health-effects models, and a hierarchical model on latent source contributions. For the health model, we focused on the time-series design in this project. Each combination of number of sources and identifiability conditions (additional constraints on model parameters) defines a different model. We built a set of plausible models with extensive exploratory data analyses and with information from previous studies, and then computed posterior model probability to estimate model uncertainty. Parameter estimation and model uncertainty estimation were implemented simultaneously by Markov chain Monte Carlo (MCMC*) methods. We validated the methods using simulated data. We illustrated the methods using $PM_{2.5}$ (particulate matter $\leq$ 2.5 μm in aerodynamic diameter) speciation data and mortality data from Phoenix, Arizona, and Houston, Texas. The Phoenix data included counts of cardiovascular deaths and daily $PM_{2.5}$ speciation data from 1995–1997. The Houston data included respiratory mortality data and 24-hour $PM_{2.5}$ speciation data sampled every six days from a region near the Houston Ship Channel in years 2002–2005. We also developed a Bayesian spatial multivariate receptor modeling approach that, while simultaneously dealing with the unknown number of sources and identifiability conditions, incorporated spatial correlations in the multipollutant data collected from multiple sites into the estimation of source profiles and contributions based on the discrete process convolution model for multivariate spatial processes. This new modeling approach was applied to 24-hour ambient air concentrations of 17 volatile organic compounds (VOCs) measured at nine monitoring sites in Harris County, Texas, during years 2000 to 2005.

Simulation results indicated that our methods were accurate in identifying the true model and estimated parameters were close to the true values. The results from our methods agreed in general with previous studies on the source apportionment of the Phoenix data in terms of

This Investigators' Report is one part of Health Effects Institute Research Report 183, which also includes a Critique by the Health Review Committee and an HEI Statement about the research project. Correspondence concerning the Investigators' Report may be addressed to Dr. Eun Sug Park, Texas A&M Transportation Institute, The Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135; *e-park@tamu.edu*.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR-83467701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

* A list of abbreviations and other terms appears at the end of the Investigators' Report.

estimated source profiles and contributions. However, we had a greater number of statistically insignificant findings, which was likely a natural consequence of incorporating uncertainty in the estimated source contributions into the health-effects parameter estimation. For the Houston data, a model with five sources (that seemed to be Sulfate-Rich Secondary Aerosol, Motor Vehicles, Industrial Combustion, Soil/Crustal Matter, and Sea Salt) showed the highest posterior model probability among the candidate models considered when fitted simultaneously to the $PM_{2.5}$ and mortality data. There was a statistically significant positive association between respiratory mortality and same-day $PM_{2.5}$ concentrations attributed to one of the sources (probably industrial combustion). The Bayesian spatial multivariate receptor modeling approach applied to the VOC data led to a highest posterior model probability for a model with five sources (that seemed to be refinery, petrochemical production, gasoline evaporation, natural gas, and vehicular exhaust) among several candidate models, with the number of sources varying between three and seven and with different identifiability conditions.

Our multipollutant approach assessing source-specific health effects is more advantageous than a single-pollutant approach in that it can estimate total health effects from multiple pollutants and can also identify emission sources that are responsible for adverse health effects. Our Bayesian approach can incorporate not only uncertainty in the estimated source contributions, but also model uncertainty that has not been addressed in previous studies on assessing source-specific health effects. The new Bayesian spatial multivariate receptor modeling approach enables predictions of source contributions at unmonitored sites, minimizing exposure misclassification and providing improved exposure estimates along with their uncertainty estimates, as well as accounting for uncertainty in the number of sources and identifiability conditions.

INTRODUCTION

There has been growing interest in assessing health effects of air pollution based on multiple pollutants (Dominici et al. 2010). High correlations often exist among multiple pollutants measured in ambient air (due to common sources and meteorology), and these high correlations lead to an estimation problem such as collinearity when multiple pollutants are included as covariates in the health-effects regression model. Therefore, a straightforward extension of the existing single-pollutant health-effects models is not appropriate.

Considering source-specific exposures to quantify exposure to multiple air pollutants addresses the aforementioned problem. Air pollution is generated from several sources and each source simultaneously emits many pollutants. While high interpollutant correlations are problematic and lead to a collinearity problem in the multivariate regression models of pollutant-specific health effects, it is not a problem in estimating source-specific health effects. As a matter of fact, high interpollutant correlations make it possible to effectively characterize complex air pollution mixtures by a few common underlying source types using multivariate receptor models (Park et al. 2001, 2002a,b).

More importantly, from a regulation standpoint, assessing the health effects of specific sources or group of sources (i.e., source-specific health effects) may be more advantageous than assessing the health effects of individual pollutants themselves (i.e., pollutant-specific health effects), in that a more targeted and stringent enforcement strategy can be developed based on the sources that emit pollutants associated with increased risks for adverse health outcomes. Another advantage of assessing source-specific health effects is that the combined effects of exposures to multipollutants in ambient air (e.g., various VOCs or specific metal constituents of $PM_{2.5}$) can be evaluated. For example, fine particles that originate from specific sources (e.g., diesel and gasoline exhaust) could be more toxic than particulates from other sources, and thus may have more significant adverse health effects (Seagrave et al. 2006).

A major difficulty in achieving the goal of assessing source-specific health effects, however, is that in most cases the sources (source profiles) of ambient air pollutants are not known and source-specific exposures cannot be measured directly; rather they need to be estimated by decomposing ambient measurements of multiple air pollutants. While it is well recognized that multiple sources contribute to the concentrations of air pollutants measured at ambient monitoring stations, what is not well understood is how many sources are involved and what their relative contributions are to the mixture that is present in ambient air. To adopt a source-specific approach when evaluating the health effects associated with air pollution, the observed mixtures of air pollutants would first need to be decomposed into contributions from several major sources, using a source identification and apportionment method. Multivariate receptor models can resolve the measured mixture of pollutants into the contributions from the individual source types.

MULTIVARIATE RECEPTOR MODELING

Multivariate receptor modeling is a collection of methods for identifying major pollution sources and estimating the contribution of each source based on ambient measurements of air pollutants obtained at a given monitoring site, or receptor. A comprehensive review of the field of receptor modeling can be found in Hopke (1991, 2003, 2010). Traditionally, multivariate receptor models have been used to resolve the observed air pollutant mixtures into contributions from individual sources (or source types) based on time series of multiple (or multivariate) air pollutants — such as VOCs or specific metal constituents of fine particulate matter (i.e., $PM_{2.5}$) — at a receptor site (see e.g., Heaton et al. 2010; Henry 1997a; Hopke 1985, 2003; Park et al. 2001; Wolbers and Stahel 2005).

A basic multivariate receptor model takes the form of:

$$X_{tj} = \sum_{k=1}^{q} A_{tk} P_{kj} + E_{tj},$$

where X_{tj} is the mass concentration of pollutant j ($j = 1, \cdots, J$) measured at time t ($t = 1, \cdots, T$), A_{tk} is the mass concentration (contribution) of source k ($k = 1, \cdots, q$) at time t, P_{kj} is the relative concentration of pollutant j in source k, E_{tj} is the error associated with the jth pollutant concentration measured at time t, and q is the total number of major contributing sources.

In matrix terms, the above model can be written as:

$$\mathbf{X} = \mathbf{AP} + \mathbf{E},$$

where $\mathbf{X}$ is an n by J data matrix containing n concentrations of J pollutants at a receptor, $\mathbf{A}$ is the T by q source contribution matrix, $\mathbf{P}$ is the q by J source-composition matrix (where each row, a source-composition profile, can be considered as a chemical fingerprint for a source), and $\mathbf{E}$ is a T by J error matrix. The elements of $\mathbf{P}$ are assumed to be non-negative to be physically meaningful. In relation to statistical models, this may be viewed as a factor analysis model or latent variable model (apart from the non-negativity constraint on the elements of $\mathbf{P}$) in the sense that $\mathbf{X}$ is the only observable quantity whereas q (number of factors), $\mathbf{P}$ (factor loading matrix), and $\mathbf{A}$ (factor score matrix) are all unknown quantities that need to be estimated (or predicted). The usual challenges in factor analysis such as the unknown number of factors (sources) and non-identifiability of parameters (i.e., parameters $\mathbf{A}$ and $\mathbf{P}$ are not uniquely defined even when q is known) are also encountered in multivariate receptor models. As a matter of fact, the estimation of parameters $\mathbf{A}$ and $\mathbf{P}$ depends heavily on q and the identifiability conditions employed (additional constraints on the parameters needed to remove non-identifiability), and these could be a major source of uncertainty in factor analysis models and multivariate receptor models. In most cases, the number of factors and identifiability conditions were either assumed to be known or were chosen in advance, thus ignoring uncertainty involved in the selection of q and identifiability conditions. This issue of model identifiability and identifiability conditions will be discussed later.

Various forms of factor analysis or principal component analysis methods have been applied in multivariate receptor modeling for more than three decades. Among several methods, positive matrix factorization (PMF) (Paatero 1997; Paatero and Tapper 1994) and Unmix (*www.epa.gov/heasd/research/unmix.html*) (Henry and Kim 1990; Kim and Henry 1999, 2000) gained the most popularity among environmental engineers and scientists and have been widely used in practice. Until recently, statisticians have made relatively few contributions to the field of multivariate receptor modeling. See Pollice (2009) for a review of multivariate receptor modeling from a statistical perspective. Park and colleagues (2001) proposed time-series extensions of multivariate receptor models to account for temporal correlation in air pollution data in parameter estimation under a confirmatory factor analysis model. Billheimer (2001) developed compositional receptor modeling, which assumes that the source contributions and the errors are logistic normally distributed. Christensen and Sain (2002) developed a different approach to account for temporal dependence in multivariate receptor modeling, a nested block bootstrap method. Park and colleagues (2002a) proposed new sets of realistic identifiability conditions for multivariate receptor models and a constrained nonlinear least squares (CNLS) approach for parameter estimation. Gajeswski and Spiegelman (2004) developed estimators that are robust to outliers. In Park and colleagues (2002b, 2004), the unknown number of pollution sources and unknown identifiability conditions have been taken into account in the form of model uncertainty using a Bayesian approach for the conventional multivariate receptor modeling data, that is, for multiple-pollutant data measured at a single monitoring site or single-pollutant data collected from multiple monitoring sites. Wolbers and Stahel (2005) proposed the log-normal structural mixing model, which assumes a multiplicative error structure. Christensen and colleagues (2006) developed an iterated confirmatory factor analysis approach to source apportionment. Spiegelman and Park (2007) performed a jackknife evaluation of the uncertainty of the estimates of the source contribution and source-composition matrices as a way of incorporating dependence in

air pollution data into the estimation of parameters. Lingwall and colleagues (2008) developed Dirichlet-based Bayesian multivariate receptor modeling, and Heaton and colleagues (2010) proposed a Dirichlet process model to incorporate time-varying source profiles in multivariate receptor models. Nikolov and colleagues (2011) extended the multiplicative factor analysis model proposed by Wolbers and Stahel (2005) by imposing mixed models on the latent source contributions to include the covariate effects and to adjust for temporal correlation in the source contribution.

In all of the previous approaches, multivariate receptor models were applied to multiple air pollutant data measured at a single monitoring site or to single-pollutant data (e.g., nonspeciated $PM_{2.5}$) collected from multiple monitoring sites (see e.g., Henry 1997b; Park et al. 2002a,b, 2004). Even for the multipollutant data collected from multiple monitoring sites, most studies on source identification and apportionment employed a conventional multivariate receptor modeling approach to analyze the multipollutant data at each site separately (e.g., Buzcu and Fraser 2006) and ignored spatial correlations in the data. Incorporating spatial correlations in the multipollutant data collected from multiple monitoring sites into multivariate receptor modeling has been an open problem for many years (Park et al. 2001, 2004; Pollice 2009). Recently, Jun and Park (2013) proposed a spatial statistics extension of multivariate receptor models by modeling unobserved source contributions as a multivariate spatial process with the multivariate Matérn covariance model (Gneiting et al. 2010), under the assumption of the known number of sources and model identifiability conditions, using maximum likelihood estimation. However, accounting for uncertainty in the number of sources and identifiability conditions in spatial multivariate receptor modeling remains unexplored.

EVALUATING SOURCE-SPECIFIC HEALTH EFFECTS

The estimated source contributions (i.e., the amount of pollution from each source) from source-apportionment or multivariate receptor modeling can be viewed as source-specific exposures. This estimation process is challenging because of the aforementioned unknown number of sources and unknown identifiability conditions that constitute model uncertainty in multivariate receptor modeling.

While many studies have investigated myriad health effects of individual pollutants (including particulate matter [PM]), fewer have controlled for confounding due to other pollutants, and even fewer investigations have focused on source-specific health effects. Most of these investigations, but not all, have examined daily mortality or hospital admissions. Appendix A (available on the HEI Web site) contains a review of studies that evaluated the associations between the short-term effects of PM and the specific cardiovascular or respiratory causes of mortality, as well as studies that evaluated the associations between VOC exposures and adverse health effects. Appendix B (available on the HEI Web site) contains a summary of studies that have evaluated the health risks (mortality or morbidity) associated with source-apportioned PM.

Laden and colleagues (2000) used factor analysis to evaluate risks for total nonaccidental mortality, from 1979 to 1988, that were associated with source-apportioned $PM_{2.5}$ in six cities in the United States (Watertown, MA; Kingston-Harriman, TN; St. Louis, MO; Steubenville, OH; Portage, WI; Topeka, KS). Associations with cause-specific mortality due to ischemic heart disease, pneumonia, and chronic obstructive pulmonary disease were investigated as well. Findings suggested associations between daily mortality and fine particles from mobile- and coal-combustion sources. Mar and colleagues (2000) used chemical composition data from PM of varying sizes (PM_{10}, $PM_{10–2.5}$ [coarse fraction], $PM_{2.5}$) from a single monitoring station in the center of Phoenix, Arizona, in their study of total nonaccidental mortality and cardiovascular mortality among individuals 65 years and older. Findings indicated that factors related to motor vehicle exhaust and resuspended road dust; vegetative burning; and sulfate were associated with mortality from cardiovascular causes. Ostro and colleagues (2011) used PMF to examine sources of particulates ($PM_{2.5}$ and PM_{10}) and daily mortality from all causes and cardiovascular diseases over a five-year period (2003 to 2007) in a study conducted in Barcelona, Spain. They reported statistically significant associations for sources of $PM_{2.5}$ from vehicle exhaust, fuel oil combustion, secondary nitrate and organics, minerals, secondary sulfate and organics, and road dust with both all-cause and cardiovascular mortality. In a study conducted in Copenhagen, Denmark, Andersen and colleagues (2007) applied source apportionment of PM_{10} and total suspended particulate samples to study hospital admissions in elderly people (65 years and older) and children (years 5 to 18). Hospital admissions for the following specific diseases were evaluated among elderly people: specific cardiovascular disease (i.e., angina pectoris, acute and subsequent myocardial infarction, acute ischemic heart disease, chronic ischemic heart disease, pulmonary embolism, cardiac arrest, cardiac arrhythmias, and heart failure) and respiratory disease (i.e., chronic bronchitis, emphysema, other chronic obstructive pulmonary diseases, asthma, and status asthmaticus). For children, admissions due to pediatric asthma (asthma and status asthmaticus) were

examined. Statistically significant associations with hospital admissions for cardiovascular diseases were detected for PM_{10} from biomass, secondary oil, and crustal sources. Statistically significant associations with hospital admissions for respiratory diseases were detected for PM_{10} from biomass and secondary sources. Associations were also detected for PM_{10} from vehicular emissions and hospital admissions for asthma among children, although they did not reach statistical significance. Lall and colleagues (2011) conducted a study in Manhattan, New York, during the period from 2001 to 2002 to evaluate associations between source-apportioned fine particles (apportioned using PMF) and daily hospital admissions for respiratory and cardiovascular outcomes among individuals 65 years and older. Associations between sources of fine particles from steel metal works were significantly associated with respiratory hospital admissions (including pneumonia and asthma when evaluated separately); associations between sources of fine particles from traffic were associated with cardiovascular hospital admissions (including stroke and heart failure when evaluated separately).

A series of publications, which emanated from a workshop sponsored by the U.S. Environmental Protection Agency (U.S. EPA) in 2003, reports on results generated by investigators from seven different institutions across the United States who applied different source-apportionment techniques to $PM_{2.5}$ compositional data collected during 1995–1997 from Washington, D.C., and from Phoenix, Arizona. Methods that were used for source-apportionment analysis included targeted rotated principal component analysis, absolute principal component analysis, Unmix, and PMF. Source-apportioned results were then used to assess associations between $PM_{2.5}$ source contributions and daily mortality due to total nonaccidental causes, cardiovascular diseases, and cardiovascular plus respiratory diseases. A common Poisson regression model was applied to facilitate comparisons among source-apportionment methods. Thurston and colleagues (2005) provides an overall summary of the results of the 2003 U.S. EPA workshop, whereas Ito and colleagues (2006) and Mar and colleagues (2006) provide more detailed findings of differences or similarities in associations between source-apportioned fine particles and daily mortality in Washington, D.C., and Phoenix, Arizona, respectively. Generally, there was consistency across methods in identifying the major sources of fine particles, especially for sources from soil-, sulfate-, residual oil- and salt-associated fine particle mass, but to a lesser extent for sources from vegetative burning and traffic (Thurston et al. 2005). When comparing results from the Poisson regression models, sources associated with sulfate were most consistently significant across all source-apportionment methods. Also, the degree to which estimated values of the relative risk estimates varied among sources was significantly greater than the variation among research groups (Ito et al. 2006; Mar et al. 2006; Thurston et al. 2005).

In addition to studies of daily mortality or hospital admissions, a few more recent investigations have examined associations between sources of PM and other health endpoints. In a study conducted by Gent and colleagues (2009), investigators examined associations between source-apportioned PM and symptoms (wheeze, persistent cough, shortness of breath, chest tightness) and medication use in 149 children who had asthma and lived in New Haven County, Connecticut. Findings indicated associations between increased odds of symptoms or inhaler use and increases in same-day and 3-day (same day, and previous 2 days) averaged exposures to PM from traffic sources and road dust. For example, odds ratios (95% confidence intervals) for wheeze were 1.10 (1.01–1.19) and 1.26 (1.05–1.51) for a 5-$\mu g/m^3$ increase in particles averaged over three days from motor vehicles and road dust, respectively. Bell and colleagues (2010) used PMF to examine sources of $PM_{2.5}$ and two birth outcomes among infants born in four counties in Connecticut ($n = 3$) and Massachusetts ($n = 1$), that is, birth weight and small-at-term births. Findings indicated inverse associations for birth weight and fine particles from road dust, oil combustion, and motor vehicles; associations were also detected between fine particles from road dust and increased prevalence of full-term infants born small for their gestational age.

In addition to epidemiologic studies, toxicologic investigations have employed source receptor modeling as well. For example, Seagrave and colleagues (2006) applied a chemical mass balance receptor model to investigate the toxic effects of $PM_{2.5}$ (administered via intratracheal instillation) in rats. $PM_{2.5}$ samples had been collected during summer or winter from four different sites that had differing source profiles: two urban sites (Birmingham, Alabama, site — located in an urban area in close proximity to traffic and industry, including a coke production facility; Atlanta, Georgia, site — Jefferson Street); a mixed urban and residential site near the Gulf of Mexico (Pensacola, Florida); and a rural site (Centreville, Alabama). Projection-to-latent-surfaces techniques were used to examine relationships between source-apportioned fine particles and cytotoxic and inflammatory endpoints. Results from the source apportionment suggested differences by season (e.g., more wood smoke and secondary nitrates in winter months) and by site (e.g., the diesel exhaust component was a large contributor in the urban sites). Overall, toxicity was greatest for the $PM_{2.5}$ samples from the two urban

sources, with significant contributions from vehicular emissions. Nikolov and colleagues (2007, 2008) developed a Bayesian structural equation model to assess source-specific health effects and compared their new method via simulation techniques to traditional approaches (i.e., a tracer approach and a two-stage approach). They further illustrated the application of their method in a study to evaluate the association between source-apportioned $PM_{2.5}$ (collected in Boston that is believed to have four major sources: road dust, power plants, oil combustion, and motor vehicles) and myocardial ischemia in dogs (using ST-segment as the endpoint).

In most of the aforementioned studies that have employed a source-specific approach, the estimated source contributions were used as if they were the true source-specific exposures (thus ignoring the uncertainty associated with estimated source contributions) in the health-effects models. Also, the model uncertainty due to unknown numbers of sources and identifiability conditions was not taken into account in the assessment of source-specific health effects.

As is well known in the measurement error model literature (e.g., Carroll et al. 1995), ignoring uncertainty in exposure estimation results in a bias in the estimated health-effects regression coefficients. Notable exceptions are studies by Nikolov and colleagues (2007, 2008) who proposed a structural equation framework to assess source-specific health effects by fitting a receptor model and the health outcome model jointly to account for the uncertainty associated with the estimated source contributions in the health-effects estimates. More importantly, however, the number of major pollution sources that drives the number of regression terms in the health-effects model was assumed to be known (or treated as fixed once it was estimated) in all of the previous studies. The same is also true for model identifiability conditions. While identifiability conditions that are useful in multivariate receptor modeling have been proposed (Park et al. 2001, 2002a) and also utilized in recent source-specific health-effects studies (Nikolov et al. 2007, 2008), those conditions were assumed to be known. The model uncertainty due to the unknown number of sources and identifiability conditions has never been taken into account in the assessment of source-specific health effects. In this project, we developed a method that accounts for both model uncertainty and parameter uncertainty (uncertainty in estimated source-specific exposures) in the assessment of source-specific health effects based on time-series data. We related the daily mortality data of a population to daily fluctuations in source contributions estimated from multipollutant data measured at a single monitoring location for the region.

Another problem that we addressed is the extension of multivariate receptor modeling to spatial multivariate receptor modeling. Despite the growing availability of multipollutant data collected from multiple monitoring sites, there has been just one attempt to incorporate spatial dependence in such data into multivariate receptor modeling. Jun and Park (2013) proposed a spatial statistics extension of multivariate receptor modeling under the assumption of a known number of sources and model identifiability conditions. That research was produced as a by-product of the present project. When the number of sources and model identifiability conditions are unknown, taking into account such model uncertainty in multivariate receptor modeling is a challenging problem. Accounting for uncertainty in the number of sources and identifiability conditions in spatial multivariate receptor modeling has never been explored. In this project, we developed a Bayesian spatial multivariate receptor modeling approach that can incorporate spatial dependence into an estimation of source profiles and contributions and also effectively deal with the unknown number of pollution sources and identifiability conditions. Accounting for spatial dependence of multivariate air pollution data in source identification and apportionment not only leads to more efficient estimation of source profiles and contributions, but also enables prediction of pollutant concentration and source contributions at locations other than the monitoring sites. Spatial prediction of source contributions can minimize exposure misclassification and allows inference about the source-specific exposures in areas that do not have any monitoring stations. Bayesian spatial multivariate receptor models developed in this project can also provide uncertainty estimates of the predicted source contributions, which was not previously possible.

SPECIFIC AIMS

The overall goal of this study was to develop enhanced statistical methods for the assessment of source-specific health effects. The specific aims were as follows: (1) to develop a multipollutant approach that accounts for both model uncertainty in multivariate receptor models and uncertainty in estimated source-specific exposures in the assessment of source-specific health effects, and (2) to develop enhanced spatial multivariate receptor models that can account for spatial correlations in the multipollutant data collected from multiple monitoring stations while accounting for model uncertainty caused by the unknown number of major sources and identifiability conditions.

METHODS

APPROACH TO ASSESSING HEALTH EFFECTS ASSOCIATED WITH AN UNKNOWN NUMBER OF MAJOR SOURCES OF MULTIPLE AIR POLLUTANTS

Basic Modeling Framework

We employ a Bayesian hierarchical modeling framework to incorporate multiple data sources (ambient air pollution data and health outcome data) into a single coherent statistical model. Our model consists of two main parts: the receptor model and the health model. In this project, we focus on the time-series design for the health model. An additional hierarchical model on latent source contributions and distributional assumptions about errors can also be added, which is introduced in the next section. A basic model form can be given as:

Receptor model: $X_t = A_t\mathbf{P} + E_t, \quad t = 1,\cdots,T,$ (1)

Health model:

$$g\left(E\left(y_t\right)\right) = \lambda + A_t\beta + Z_t\eta = \lambda + \sum_{k=1}^{q}\beta_k A_{tk} + \sum_{i=1}^{I}\eta_i Z_{ti}, \text{ where} \quad (2)$$

$X_t = (X_{t1}, X_{t2},\cdots, X_{tJ})$: concentrations of J pollutants (chemical species) measured at time t at a receptor,

T: number of observations (number of days),

q: number of major pollution sources (unknown),

$\mathbf{P}$: $q \times J$ source-composition matrix of which rows are the source-composition profiles (P_k, $k = 1,\cdots, q$),

$P_k = (P_{k1}, P_{k2},\cdots, P_{kJ})$: kth source-composition profile consisting of the fractional amount of each chemical species in the emissions from the kth source,

$A_t = (A_{t1}, A_{t2},\cdots, A_{tq})$: source-contribution vector in time t where A_{tk} is the contribution from the kth source,

$E_t = (E_{t1}, E_{t2},\cdots, E_{tJ})$: measurement error in pollutant concentrations at time t,

y_t: health outcome at time t,

λ: overall baseline risk of death,

$\beta = (\beta_1,\cdots, \beta_q)'$: parameter describing the influence of each source-specific exposure on mortality rate,

$Z_t = (Z_{t1},\cdots, Z_{tI})$: transformations of confounding variables such as temperature, humidity, the day of the week, etc.,

$\eta = (\eta_1,\cdots, \eta_I)'$: parameter describing the influence of confounding variables on mortality.

The link function g can be changed depending on the type of the health outcome variable. For example, it can be the identity function for a continuous health outcome variable such as lung function, or the log function for a discrete health outcome variable such as daily mortality or morbidity count. We will assume that the measured pollutant concentrations and the health outcomes are conditionally independent given the unobserved source contributions and other covariates in the model, which seems to be a reasonable assumption.

Note that Equation 2 represents an individual-lag model. Without loss of generality, we present the model (and the method) using lag 0 contributions.

Other individual-lag l models can be expressed as:

Receptor model: $X_{t-l} = A_{t-l}\mathbf{P} + E_{t-l},$

Health model:

$$g\left(E\left(y_t\right)\right) = \lambda + A_{t-l}\beta + Z_t\eta = \lambda + \sum_{k=1}^{q}\beta_k A_{t-l,k} + \sum_{i=1}^{I}\eta_i Z_{ti}.$$

Our main goal is to estimate parameters $\mathbf{A}$, $\mathbf{P}$, and β. (A is a $T \times q$ source contribution matrix of which rows are A_t, $t = 1,\cdots, T$; γ and η are nuisance parameters), along with their uncertainties and model uncertainties. The parameters $\beta = (\beta_1,\cdots, \beta_q)'$ quantify the q source-specific health effects. A major source of model uncertainty in the model defined by Equations 1 and 2 is the unknown number of major pollution sources, q, and identifiability conditions.

Model Identifiability in Multivariate Receptor Models

It is well known in multivariate receptor modeling as well as in factor analysis that the receptor model in Equation 1 is not identifiable, even under the assumption that q is known (see Park et al. 2002a,b), without imposing additional constraints on the parameters. Since both $\mathbf{A}$ and $\mathbf{P}$ are unknown, the parameterization for the mean is not unique, i.e., $E(\mathbf{X}) = \mathbf{AP} = \mathbf{A}^*\mathbf{P}^*$. Under the additional assumption that $rank(\mathbf{A}) = q$ and $rank(\mathbf{P}) = q$, it can be shown that $\mathbf{AP} = \mathbf{A}^*\mathbf{P}^*$ always implies that $\mathbf{A}^* = \mathbf{AR}$ and $\mathbf{P}^* = \mathbf{R}^{-1}\mathbf{P}$ for a $q \times q$ nonsingular matrix $\mathbf{R}$. Thus, with the additional rank assumption, non-identifiability of the receptor model in Equation 1 can be reduced to the so-called *factor indeterminacy* problem in factor analysis. Fortunately, under some constraints (called identifiability conditions) on either $\mathbf{A}$ or $\mathbf{P}$, the parameters can be uniquely defined (see Park et al. 2002a). Since there are q^2 elements in the matrix $\mathbf{R}$, we need to impose q^2 independent conditions on $\mathbf{P}$ or $\mathbf{A}$

to rule out this indeterminacy. There could, in principle, be infinitely many different identifiability conditions that are each sufficient but not necessary. Because identifiability conditions are additional assumptions about the parameters, it is important to select conditions that are physically meaningful in the given context of the problem (though there could be many other purely mathematical identifiability conditions). For this reason, we restrict the type of identifiability conditions to be compared to those that are often reasonable and make sense in the context of receptor modeling or source apportionment. One set of such conditions is prespecification of zero elements in the source-composition matrix **P**:

C1. There are at least $q-1$ zero elements in each row of **P**;

C2. For each $k = 1,\cdots, q$, the rank of $\mathbf{P}^{(k)}$ is $q-1$, where $\mathbf{P}^{(k)}$ is the matrix composed of the columns containing the assigned zeros in the kth row with those assigned zeros deleted (i.e., the kth row deleted);

C3-1. $P_{kj} = 1$ (or any positive constant c_k) for some j ($j = 1,\cdots, J$) for each $k = 1,\cdots, q$; or

C3-2. $\sum_{j=1}^{J} P_{kj} = 1$ for each $k = 1,\cdots, q$.

The conditions C1–C2 imply that some pollutants (corresponding to zeros in **P**) are not contributed by a particular source type (i.e., the kth source does not affect the jth pollutant), and no two sources share the exactly same set of zeros. These are the same conditions as those used in confirmatory factor analysis to remove the factor indeterminacy problem (see, for example, Anderson [1984], Chapter 14.2.2). With regard to the condition C2 and its practical implementation, it needs to be ensured that $\mathbf{P}^{(k)}$ is not close to singular (in some cases, $\mathbf{P}^{(k)}$ may be close to a singular matrix although it may still have the rank $q-1$). The condition number of $\mathbf{P}^{(k)}$ (the ratio of the largest singular value of $\mathbf{P}^{(k)}$ to the smallest) can be examined to prevent $\mathbf{P}^{(k)}$ from being close to a singular matrix. A normalization constraint C3-1 or C3-2 (only one of them is needed) is enforced during estimation to remove indeterminacy resulting from the multiplication of a row of **P** by a scale constant. Note that the normalization constraint is somewhat arbitrary and does not recover the absolute values in **P**. The constraint $\sum_{j=1}^{J} P_{kj} = 1$ indicates that only the relative amount of species for each source profile is of interest. As long as the relative amounts of species in a source profile are given, the source can be identified. Note that any change in the scale of **P** can be absorbed into **A** and vice versa. Another way of eliminating scale invariance of factors by a constant multiplication is to assume that the source contributions have unit standard deviations or to assume an orthogonal factor model as a prior distribution for A_t (instead of C3-1 or C3-2). In practice, the source profiles and contributions are often rescaled to be $\sum_{j=1}^{J} P_{kj} = 1$ for ease of interpretation after estimation, even in the case that C3-1 or an orthogonal factor model for a prior for A_t is used during the estimation.

The related but much stronger set of identifiability conditions is:

D1. There are at least q columns in **P** with each of q columns containing only one nonzero element;

D2. Same as C2;

D3. Same as C3-1.

These conditions correspond to the assumption of having at least one *tracer* element for each source, an assumption that has been used for a long time for identifiability in the receptor modeling community. A tracer element is a single species that is contributed by only one pollution source type. Note that C1 is automatically satisfied if we have a tracer element for each source, but not vice versa. If columns of **P** (along with the columns of the data matrix) are reordered, it can be easily shown that the D1–D3 conditions are equivalent to specification of the leading $q \times q$ submatrix of **P** as $\mathbf{I_q}$ (the $q \times q$ identity matrix), which is another mathematically convenient set of identifiability conditions presented in Anderson (1984).

Both sets of identifiability conditions (C1–C3 and D1–D3) require the investigator to have some prior knowledge about likely sources, which might be obtained from previous studies or exploratory analyses. In practice, the preassigned zero elements are rarely actually zeros but they are small enough to be considered zero (i.e., minor compounds). As demonstrated by Nikolov and colleagues (2007) in a simulation study designed to investigate the impact of choosing incorrect identifiability constraints, and also by our simulation study described later, the results are not sensitive to errors of assuming a zero where the truth is nonzero, as long as the preassigned zero element is not a major constituent of a source profile. For example, when Road Dust is considered as one of the likely sources for PM speciation data, we may preassign zero for S in the candidate profile using prior knowledge that S is not a major constituent of the Road Dust profile. If we do not have any priori information on the position of zeros in **P**, then we may start with several candidate

positions for zeros proposed by exploratory data analysis (PMF or Unmix can be utilized in this step).

Instead of prespecification of zero elements in **P**, we may also consider preassigning zeros in the source contribution matrix **A**, which implies that each source is missing on some days (see Park et al. 2002a), or the combination of prespecification of zeros in **P** and **A**. We did not use constraints on **A** in this project because putting constraints on **A** would hinder the stochastic modeling of **A** and it would be hard to generate samples from the distribution of **A** in the current Bayesian modeling. The constraints on **A** have been utilized in the previous study (Park et al. 2002a) when implementing a frequentist approach to estimate **A**. Recall that there could be many different identifiability conditions that are sufficient, but there have been no identifiability conditions that are both sufficient and necessary. As long as at least one set of identifiability conditions are satisfied for the data at hand, the model should be identifiable.

As mentioned earlier, in most of the previous approaches to evaluating source-specific health effects, the estimated source contributions were used as if they were true source-specific exposures or, at least, the number of major pollution sources (that drives the number of regression terms in the health effects model) and identifiability conditions were assumed to be known. As a matter of fact, estimation of parameters **A**, **P**, and β, heavily depends on q and also on the identifiability conditions employed (e.g., where to preassign zeros in **P**), and these could be a major source of uncertainty in the estimated health effects. In some cases, we may have some prior knowledge about this, that is, the number of sources and the position of zeros can be assumed known (see, e.g., Park et al. 2001). More frequently, that information is lacking, and it becomes a main source of model uncertainty.

Park and colleagues (2002b) proposed a Bayesian approach that can simultaneously estimate such model uncertainty as well as the parameters in each model. The method computes the marginal likelihoods, the posterior model probabilities, or both using MCMC for a range of plausible models (rather than a single model) that are selected by varying the number of sources and zero elements in **P**. In this project, we aimed to quantify the uncertainties in q and identifiability conditions along with other parameter uncertainties so that the inherent variability of the source apportionment can be taken into account in the assessment of source-specific health effects.

To develop a method that can account for model uncertainty in the assessment of source-specific health effects, we build on the Bayesian method developed by Park and colleagues (2002b) that computes marginal likelihoods and posterior model probabilities for a range of plausible models (with different q and identifiability conditions) by MCMC.

We extend the method developed in Park et al. (2002b) in two aspects:

1. by adapting enhanced multivariate receptor models explicitly incorporating correlated source contributions, that is, assuming a priori correlated source contributions (an oblique factor model),
2. by including health models.

Estimation of Parameter Uncertainties and Model Uncertainties under Enhanced Multivariate Receptor Models and Health Models

The method of Park and colleagues (2002b) was developed by assuming an orthogonal factor model as a prior distribution for the source contributions, that is, assuming a priori uncorrelated centered source contributions,

$$\gamma_t = A_t - \xi \sim N_q\left(\mathbf{0}, \mathbf{I_q}\right), \tag{3}$$

where $\xi = \xi_1, \xi_2, \xi_3, \ldots, \xi_q$ is the mean of A_t, $\mathbf{I_q}$ is the $q \times q$ identity matrix, and $N_q(\cdot,\cdot)$ represents a q-variate normal distribution. A normal distribution was chosen as a prior for γ_t for mathematical convenience because γ_t can be considered as latent variables, and the form of the prior distribution of the latent variables is essentially arbitrary and largely a matter of convention (Bartholomew and Knot 1999). Note that the receptor model defined by Equation 1 can be reparameterized as:

$$X_t = \mu + \gamma_t \mathbf{P} + E_t, \quad t = 1, \cdots, T, \tag{4}$$

where $\mu = \xi \mathbf{P}$.

Although the method of Park and colleagues (2002b) was shown to be robust to violations of the prior assumption on the correlation structure of source contributions, in this project the method is generalized to formally account for correlated source contributions, that is, assuming a priori correlated source contributions (an oblique factor model) to see if it improves the prediction of source contributions. We assume an oblique factor model as a prior distribution for source contributions as follows:

$$\gamma_t \sim N_q(\mathbf{0}, \Omega) \tag{5}$$

or

$$A_t \sim N_q(\xi, \Omega) I\left(A_t \geq \mathbf{0}\right), \tag{6}$$

where Ω is a general covariance matrix. (Note that the models defined by Equations 1 and 2 can be reparameterized in terms of centered source contributions γ_t without loss of generality. In particular, our key parameter β is not affected whether or not the source contributions are centered.) Originally, we explored both priors in Equations 5 and 6, but it turned out that, for the purpose of computing marginal likelihoods, reparameterization of the Equation 1 and 2 models using centered source contributions, γ_t, is more convenient because we can more effectively cope with the issue of scale invariance of factors by a constant multiplication. Although we need a normalization constraint in estimation, such as $P_{kj} = 1$ of C3-1, the condition is enforced only to remove indeterminacy of factors by a constant multiplication in estimation; it does not mean that the true absolute value (there is no such thing, in fact) of P_{kj} is 1. In practice, the estimated source-composition profiles and the source contributions are usually renormalized to bring them onto the preferred scale, for example, see Ramadan and colleagues (2000, 2003). As a result, the estimated mean parameter, ξ, of source contributions under the identifiability condition, C3-1, is not to be interpreted as an estimate for the true mean source contribution unless it is renormalized. The same comments can also be applied to Ω that is, the estimated diagonal elements of Ω are not to be interpreted as an estimate for the true variance of source contributions. However, the estimated off-diagonal elements (relative to diagonal elements) of Ω can be used to estimate correlations among source contributions, which is scale-free and meaningful even without renormalization.

Note that the model for the centered source contributions in Equation 5 does not have a nonnegativity constraint. Nonnegativity constraints are placed on the elements of $\mathbf{P}$, however. As a matter of fact, the nonnegativity constraints on $\mathbf{P}$ are much more important for the resulting source profiles to be physically meaningful, and it is crucial to enforce those constraints in the estimation. Although we do not use the nonnegativity constraints on A_t in estimation (because we decided to work with the centered source contributions γ_t for the reasons stated above), the estimates of A_t (that can be obtained by adding the estimated mean contribution $\hat{\xi} = \hat{\mu}\hat{\mathbf{P}}\left(\hat{\mathbf{P}}\hat{\mathbf{P}}'\right)^{-1}$ to estimated γ_t) are usually nonnegative unless the true source contributions are negligible or close to zero.

Using the centered source contributions, we can rewrite the models in Equations 1 and 2 as follows:

$$\text{Receptor model: } X_t = \mu + \gamma_t \mathbf{P} + E_t,\ t = 1,\cdots,T, \tag{7}$$

$$\text{Health model: } g\left(E\left(y_t\right)\right) = \alpha + \gamma_t\beta + Z_t\eta = \alpha + \sum_{k=1}^{q}\beta_k\gamma_{t,k} + \sum_{i=1}^{I}\eta_i Z_{ti}, \tag{8}$$

where $\mu = \xi\mathbf{P},\ \alpha = \lambda + \xi\beta$.

Continuous Health Outcome When y_t is a continuous health outcome variable such as some biological parameter of pulmonary or cardiac function (e.g., peak ST-segment elevation in Nikolov et al. 2006, 2007) or daily mortality (or morbidity) count with a large enough mean, y_t may be assumed to follow a normal distribution, and the link function g in Equation 2 becomes the identity function. Using the centered source contribution and the identity link function for the health model, the model for source-specific health effects can be written as follows:

$$\text{Receptor model: } X_t = \mu + \gamma_t \mathbf{P} + E_t,\ t = 1,\cdots,T, \tag{9}$$

$$\text{Health model: } y_t = \alpha + \gamma_t\beta + Z_t\eta + \varepsilon_t = \alpha + \sum_{k=1}^{q}\beta_k\gamma_{tk} + \sum_{i=1}^{I}\eta_i Z_{ti} + \varepsilon_t, \tag{10}$$

where $\mu = \xi\mathbf{P},\ \alpha = \lambda + \xi\beta$.

We make the following assumptions about errors E_t and ε_t:

$E_t \sim N_J(\mathbf{0}, \Sigma)$, where $\Sigma = diag\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_J^2\right)$, and $\varepsilon_t \sim N\left(0, \sigma_y^2\right)$.

To complete a Bayesian model specification, the prior distributions for the unknown parameters, $\Gamma = \{\gamma_t,\ t = 1, \cdots, T\}$, Ω, $\mathbf{P}$, Σ, μ, α, β, η, and σ_y^2 are required. We assume independence among $\{\gamma_t,\ t = 1, \cdots, T\}$ given the hyperparameter Ω and the parameters $\mathbf{P}$, Σ, Ω, α, μ, β, η, and σ_y^2 as follows:

$$\begin{aligned} p\left(\Gamma, \mathbf{P}, \Sigma, \Omega, \mu, \alpha, \beta, \eta, \sigma_y^2\right) &= p(\Gamma|\Omega)\,p(\mathbf{P})\,p(\Sigma)\,p(\Omega)\,p(\mu)\,p(\alpha) \\ &\quad \times\ p(\beta)\,p(\eta)\,p\left(\sigma_y^2\right) \\ &= \left\{\prod_{t=1}^{n} p\left(\gamma_t|\Omega\right)\right\} \\ &\quad \times\ p(\mathbf{P})\,p(\Sigma)\,p(\Omega)\,p(\mu)\,p(\alpha)\,p(\beta)\,p(\eta)\,p\left(\sigma_y^2\right). \end{aligned}$$

As noted earlier, the prior distribution for the centered source contribution $\{\gamma_t\}$ is assumed to be $\gamma_t \sim N_q(\mathbf{0}, \Omega)$.

For a prior distribution for **P**, we assume a point mass at zero for $q(q-1)$ elements of **P** preselected for identifiability conditions. For the free elements of **P**, we use the truncated normal distribution, $vecP^+ \sim N_{Jq-q(q-1)}(c_0, C_0)\, \mathbf{I}\,(vecP^+ \geq \mathbf{0})$, where $vecP^+$ denotes the $Jq-q(q-1)$-dimensional vector of free elements of **P** stacked column-wise, to incorporate the nonnegativity constraints (which is critical for the estimated source profiles to make physical sense) while facilitating computation. Other choices for prior distributions for **P** have been used in the literature. For example, Lingwall and colleagues (2008) use a generalized Dirichlet distribution and Heaton and colleagues (2010) use a Dirichlet distribution to constrain profile elements to be small when accompanied by one or more dominant elements of the same profile. Note that a Dirichlet distribution might be more appropriate when the sum-to-one constraint for each row of **P** (C3-2) is used instead of C3-1. Nikolov and colleagues (2011) use a log-normal distribution for the free elements in **P**. As noted, however, defining a vague log-normal prior is not straightforward. In practice, the information on the possible range of values in **P** is often lacking and a vague prior for **P** is preferred. For the prior distributions of Σ, α, μ, Ω, β, η, and σ_y^2, we assume:

$$\sigma_j^{-2} \sim Gamma\left(a_0, b_{0j}\right),\ j = 1, \cdots, J,$$
$$\mu \sim N_J\left(m_0, M_0\right),$$
$$\alpha \sim N\left(\alpha_0, U_0\right),$$
$$\beta \sim N_q\left(\beta_0, B_0\right),$$
$$\eta \sim N_I\left(\eta_0, \Psi_0\right),$$
$$\sigma_y^{-2} \sim Gamma\left(a_0^y, b_0^y\right), \text{ and}$$
$$\Omega \sim \mathrm{IW}\left(R_0, r_0\right),$$

where IW refers to inverted Wishart and the density of Ω is given as:

$$f_{\mathrm{IW}}\left(\Omega \middle| R_0, r_0\right) = \frac{\left|R_0\right|^{\frac{1}{2}r_0} \left|\Omega\right|^{-\frac{1}{2}(r_0+q+1)} \exp\left\{-\frac{1}{2} tr\left(R_0 \Omega^{-1}\right)\right\}}{2^{\frac{1}{2}r_0 q}\, \Gamma_q\left(\frac{1}{2}r_0\right)},$$

where $\Gamma_q\left(\frac{1}{2}r_0\right)$ is the multivariate gamma function defined by $\Gamma_J(t) = \pi^{J(J-1)/4} \prod_{i=1}^{J} \Gamma\left[t - \frac{1}{2}(i-1)\right]$. Then the joint posterior distribution for $\left(\Gamma, \mathbf{P}, \Sigma, \Omega, \mu, \alpha, \beta, \eta, \sigma_y^2\right)$ is given by Equation 11 (see next page):

$$
\begin{aligned}
&\pi\left(\Gamma,\mathbf{P},\Sigma,\Omega,\mu,\alpha,\beta,\eta,\sigma_y^2\,|X,Y,Z\right) \qquad (11)\\
&\propto f\left(X,Y\left|\Gamma,\Omega,\mathbf{P},\Sigma,\mu,\alpha,\beta,\eta,\sigma_y^2\right.\right)p(\Gamma|\Omega)p(\mathbf{P})p(\Sigma)p(\Omega)p(\mu)p(\alpha)p(\beta)p(\eta)p\left(\sigma_y^2\right)\\
&=\prod_{t=1}^{T}\left[f\left(X_t\,|\mu,\gamma_t,\mathbf{P},\Sigma\right)p\left(\gamma_t\,|\Omega\right)f\left(y_t\left|\alpha,\beta,\gamma_t,\eta,\sigma_y^2\right.\right)\right]\\
&\quad\times\ p(\mathbf{P})p(\Sigma)p(\Omega)p(\mu)p(\alpha)p(\beta)p(\eta)p\left(\sigma_y^2\right)\\
&=\prod_{t=1}^{T}\left[\left|2\pi\Sigma\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(X_t-\mu-\gamma_t\mathbf{P}\right)\Sigma^{-1}\left(X_t-\mu-\gamma_t\mathbf{P}\right)'\right\}\right.\\
&\quad\times\ \left|2\pi\Omega\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\gamma_t\Omega^{-1}\gamma_t'\right)\times\left(2\pi\sigma_y^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_y^2}\left(y_t-\alpha-\gamma_t\beta-Z_t\eta\right)^2\right\}\Bigg]\\
&\quad\times\ \left[\prod_{j=1}^{Jq-q(q-1)}\Phi_Z\left(c_0^j\Big/\sqrt{C_0^{j,j}}\right)\right]^{-1}\left|2\pi C_0\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{\left(vecP^+-c_0\right)'C_0^{-1}\left(vecP^+-c_0\right)\right\}\right]\mathbf{I}\left(vecP^+\geq\mathbf{0}\right)\\
&\quad\times\ \prod_{j=1}^{J}\left[\frac{b_{0j}^{a_0}}{\Gamma\left(a_0\right)}\left(\frac{1}{\sigma_j^2}\right)^{a_0+1}\exp\left(-\frac{b_{0j}}{\sigma_j^2}\right)\right]\times\frac{\left|R_0\right|^{\frac{1}{2}r_0}\left|\Omega\right|^{-\frac{1}{2}(r_0+q+1)}\exp\left\{-\frac{1}{2}tr\left(R_0\Omega^{-1}\right)\right\}}{2^{\frac{1}{2}r_0q}\Gamma_q\left(\frac{1}{2}r_0\right)}\\
&\quad\times\ \left|2\pi M_0\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{\left(\mu-m_0\right)M_0^{-1}\left(\mu-m_0\right)'\right\}\right]\times\left(2\pi U_0\right)^{-\frac{1}{2}}\exp\left[-\frac{1}{2U_0}\left(\alpha-\alpha_0\right)^2\right]\\
&\quad\times\ \left|2\pi\mathrm{B}_0\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{\left(\beta-\beta_0\right)'\mathrm{B}_0^{-1}\left(\beta-\beta_0\right)\right\}\right]\times\left|2\pi\Psi_0\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{\left(\eta-\eta_0\right)'\Psi_0^{-1}\left(\eta-\eta_0\right)\right\}\right]\\
&\quad\times\ \frac{\left(b_0^y\right)^{a_0^y}}{\Gamma\left(a_0^y\right)}\left(\frac{1}{\sigma_y^2}\right)^{a_0^y+1}\exp\left(-\frac{b_0^y}{\sigma_y^2}\right),
\end{aligned}
$$

where Φ_z is the cdf (cumulative distribution function) of the standard normal distribution.

Because of the complexity of the joint posterior distribution in Equation 11, MCMC methods are employed for the estimation of parameters. In the MCMC sampling algorithm employed here, one sweep consists of nine updating procedures: (i) updating Γ, (ii) updating $\mathbf{P}$, (iii) updating Σ, (iv) updating Ω, (v) updating μ, (vi) updating α, (vii) updating β, (viii) updating η, and (ix) updating σ_y^2. The full conditional distributions are given in Appendix C (available on the HEI Web site).

As discussed earlier, each combination of q and identifiability conditions (here, position of prespecified zeros) leads to a different model. Assume that there are G candidate models associated with different q and identifiability conditions, $M_1, \cdots, M_G$. Typical Bayesian model comparison is based on posterior model probabilities:

$$
p\left(M_g\,|X,y\right)=\frac{p\left(M_g\right)l\left(X,y\left|M_g\right.\right)}{\sum_{k=1}^{G}\left[p\left(M_k\right)l\left(X,y|M_k\right)\right]},
$$

where $p(M_g)$ is the prior model probability and $l(X,y\,|\,M_g)$ is the marginal likelihood for model M_g, respectively. Note that under the indifference prior model probabilities, the posterior model probability is proportional to the marginal likelihood and the above equation becomes:

$$
p\left(M_g\,|X,y\right)=\frac{l\left(X,y\left|M_g\right.\right)}{\sum_{k=1}^{G}l\left(X,y|M_k\right)}.
$$

Thus, we only need to calculate the marginal likelihood of each model for model comparison. Note that $l(X,y \mid M_g)$ can be estimated by:

$$\hat{l}\left(X,y\middle|M_g\right)=\frac{l\left(X,y\middle|\theta_g^c,M_g\right)p\left(\theta_g^c\middle|M_g\right)}{\hat{\pi}\left(\theta_g^c\middle|X,y,M_g\right)},$$

where $l(X,y \mid \theta_g, M_g)$ is the likelihood of θ_g under model M_g, $p(\theta_g \mid M_g)$ is the prior of θ_g under model M_g, θ_g^c is a single point of $\theta_g=\left(\Gamma_g,\mathbf{P}_g,\Sigma_g,\Omega_g,\mu_g,\alpha_g,\beta_g,\eta_g,\sigma_{y,g}^2\right)$ under model M_g, and $\hat{\pi}\left(\theta_g^c\middle|X,y,M_g\right)$ is the estimated posterior density function of $\pi\left(\theta_g^c\middle|X,y,M_g\right)$. For simplicity of notation, we suppress the index g for the rest of the section. Using the same algorithm of Oh (1999), we have:

$$\begin{aligned}
&\pi\left(\theta^c\middle|X,y,M\right) \qquad (12)\\
&\quad=E\Big[\pi\left(\Gamma^c\middle|\mathbf{P}^c,\Sigma^c,\Omega^c,\mu^c,\alpha^c,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\mathbf{P}^c\middle|\Gamma,\Sigma^c,\Omega^c,\mu^c,\alpha^c,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\Sigma^c\middle|\Gamma,\mathbf{P},\Omega^c,\mu^c,\alpha^c,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\Omega^c\middle|\Gamma,\mathbf{P},\Sigma,\mu^c,\alpha^c,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\mu^c\middle|\Gamma,\mathbf{P},\Sigma,\Omega,\alpha^c,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\alpha^c\middle|\Gamma,\mathbf{P},\Sigma,\Omega,\mu,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\beta^c\middle|\Gamma,\mathbf{P},\Sigma,\Omega,\mu,\alpha,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\eta^c\middle|\Gamma,\mathbf{P},\Sigma,\Omega,\mu,\alpha,\beta,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\sigma_y^{2c}\middle|\Gamma,\mathbf{P},\Sigma,\Omega,\mu,\alpha,\beta,\eta\right)\Big]\\
&\quad=E\Big[\pi\left(\Gamma^c\middle|\mathbf{P}^c,\Sigma^c,\Omega^c,\mu^c,\alpha^c,\beta^c,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\mathbf{P}^c\middle|\Gamma,\Sigma^c,\mu^c\right)\\
&\qquad\times\pi\left(\Sigma^c\middle|\Gamma,\mathbf{P},\mu^c\right)\times\pi\left(\Omega^c\middle|\Gamma\right)\times\pi\left(\mu^c\middle|\Gamma,\mathbf{P},\Sigma\right)\\
&\qquad\times\pi\left(\alpha^c\middle|\Gamma,\beta^c,\eta^c,\sigma_y^{2c}\right)\times\pi\left(\beta^c\middle|\Gamma,\alpha,\eta^c,\sigma_y^{2c}\right)\\
&\qquad\times\pi\left(\eta^c\middle|\Gamma,\alpha,\beta,\sigma_y^{2c}\right)\times\pi\left(\sigma_y^{2c}\middle|\Gamma,\alpha,\beta,\eta\right)\Big].
\end{aligned}$$

Because the full conditional posterior density functions are known, $\pi\left(\hat{\theta}^c\middle|X,y,M\right)$ can be estimated as the sample average of the product of the full conditional posterior density functions using the posterior sample of θ under model M. Although in theory θ^c can be an arbitrary point in the parameter space, for efficiency it needs to be chosen from the region with high posterior density. An approximate posterior mode of θ^c, based on a preliminary MCMC run, would be a reasonable choice for θ^c.

Discrete Health Outcome When the health outcome in Equation 8 is a discrete variable such as a daily mortality count, assuming a normal linear model would not be appropriate unless the mean mortality count is large enough so that normal approximation is satisfied. For cause-specific mortality count data, the mean is typically small, and a Poisson model with a log link function is typically assumed for such data, for example,

Health model: $y_t \sim Poisson\left(E\left(y_t\right)\right)$,

$$\begin{aligned}
\log\left(E\left(y_t\right)\right)&=\alpha+\gamma_t\beta+Z_t\eta \qquad (13)\\
&=\alpha+\sum_{k=1}^{q}\beta_k\gamma_{tk}+\sum_{i=1}^{I}\eta_i Z_{ti}.
\end{aligned}$$

Note that Equation 13 represents an individual-lag model. Without loss of generality, we present the model (and the method) using lag 0 contributions. We assume that errors E_t follow a multivariate normal distribution with a mean vector $\mathbf{0}$ and the diagonal covariance matrix:

$$\Sigma=diag\left(\sigma_1^2,\sigma_2^2,\cdots,\sigma_J^2\right), \text{ i.e., } E_t\sim N_J\left(\mathbf{0},\Sigma\right).$$

To complete a Bayesian model specification, the prior distributions for the unknown parameters, $\Gamma=\{\gamma_t, t=1,\cdots,T\}$, Ω, $\mathbf{P}$, Σ, μ, α, β, and η are required. We assume independence among $\{\gamma_t, t=1,\cdots,T\}$ given the hyperparameter Ω and also among the parameters Ω, $\mathbf{P}$, Σ, α, μ, β, and η as follows:

$$\begin{aligned}
&p\left(\Gamma,\Omega,\mathbf{P},\Sigma,\mu,\alpha,\beta,\eta\right)\\
&\quad=p\left(\Gamma\mid\Omega\right)p(\Omega)p(\mathbf{P})p(\Sigma)p(\mu)p(\alpha)p(\beta)p(\eta)\\
&\quad=\left\{\prod_{t=1}^{T}p\left(\gamma_t\middle|\Omega\right)\right\}p(\Omega)p(\mathbf{P})p(\Sigma)p(\mu)p(\alpha)p(\beta)p(\eta).
\end{aligned}$$

It turned out that for discrete outcome data, the extension of the method in Park and colleagues (2002b) is much more difficult due to the complexity of full conditional distributions.

For the prior distribution $\{\gamma_t\}$ given the hyperparameter Ω, we assume $\gamma_t \sim N_q(\mathbf{0},\Omega)$, and for the prior distributions of $\mathbf{P}$, Σ, α, μ, Ω, β, and η, we assume:

$$vecP^+ \sim N_{p^+}\left(c_0,C_0\right)\mathbf{I}\left(vecP^+ \geq \mathbf{0}\right),$$

where $p^+ = Jq - q(q-1) =$ the number of free elements (that are not preassigned zeros) in $\mathbf{P}$,

$\sigma_j^{-2} \sim Gamma\left(a_0, b_{0j}\right), j = 1,\cdots,J,$

$\mu \sim N_J\left(m_0, M_0\right),$

$\Omega \sim \mathrm{IW}\left(R_0, r_0\right),$

$\alpha \sim N\left(\alpha_0, U_0\right),$

$\beta \sim N_q\left(\beta_0, \mathrm{B}_0\right),$ and

$\eta \sim N_I\left(\eta_0, \Psi_0\right).$

Then the joint posterior distribution for (Γ, Ω, $\mathbf{P}$, Σ, μ, α, β, η) is given by:

$$\begin{aligned}
&\pi\left(\Gamma,\Omega,\mathbf{P},\Sigma,\mu,\alpha,\beta,\eta|X,Y,Z\right) \qquad (14)\\
&\quad\propto f\left(X,Y|\Gamma,\Omega,\mathbf{P},\Sigma,\mu,\alpha,\beta,\eta\right)p\left(\Gamma|\Omega\right)p(\Omega)p(\mathbf{P})p(\Sigma)p(\mu)p(\alpha)p(\beta)p(\eta)\\
&\quad=\prod_{t=1}^{T}\left[f\left(X_t|\mu,\gamma_t,\mathbf{P},\Sigma\right)p\left(\gamma_t|\Omega\right)f\left(y_t|\alpha,\beta,\gamma_t,\eta\right)\right]p(\Omega)p(\mathbf{P})p(\Sigma)p(\mu)p(\alpha)p(\beta)p(\eta)\\
&\quad=\prod_{t=1}^{T}\left[|2\pi\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(X_t-\mu-\gamma_t\mathbf{P}\right)(\Sigma)^{-1}\left(X_t-\mu-\gamma_t\mathbf{P}\right)'\right\}\right.\\
&\qquad\times|2\pi\Omega|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\gamma_t\Omega^{-1}\gamma_t'\right)\\
&\qquad\left.\times\exp\left\{-\exp\left(\alpha+\gamma_t\beta+Z_t\eta\right)\right\}\left\{\exp\left(\alpha+\gamma_t\beta+Z_t\eta\right)\right\}^{y_t}\Big/y_t!\right]\\
&\qquad\times\frac{|R_0|^{\frac{1}{2}r_0}|\Omega|^{-\frac{1}{2}(r_0+q+1)}\exp\left\{-\frac{1}{2}tr\left(R_0\Omega^{-1}\right)\right\}}{2^{\frac{1}{2}r_0q}\Gamma_q\left(\frac{1}{2}r_0\right)}\\
&\qquad\times\left[\prod_{j=1}^{p^+}\Phi_Z\left(c_0^j\Big/\sqrt{C_0^{j,j}}\right)\right]^{-1}|2\pi C_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{\left(vecP^+-c_0\right)'C_0^{-1}\left(vecP^+-c_0\right)\right\}\right]\mathbf{I}\left(vecP^+\geq\mathbf{0}\right)\\
&\qquad\times\prod_{j=1}^{J}\frac{b_0^{a_0}}{\Gamma(a_0)}\left(\frac{1}{\sigma_j^2}\right)^{a_0+1}\exp\left(-\frac{b_0}{\sigma_j^2}\right)\\
&\qquad\times|2\pi M_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{(\mu-m_0)M_0^{-1}(\mu-m_0)'\right\}\right]\\
&\qquad\times\left(2\pi U_0\right)^{-\frac{1}{2}}\exp\left[-\frac{1}{2U_0}\left(\alpha-\alpha_0\right)^2\right]\\
&\qquad\times|2\pi\mathrm{B}_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{(\beta-\beta_0)\mathrm{B}_0^{-1}(\beta-\beta_0)'\right\}\right]\times|2\pi\Psi_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{(\eta-\eta_0)\Psi_0^{-1}(\eta-\eta_0)'\right\}\right].
\end{aligned}$$

It is clear from Equation 14 that, due to the presence of nonconjugacy for some of the model parameters, the full conditional posterior distributions for them are very complex and cannot be given in closed forms.

Our approach addresses this nonconjugate problem by introducing normal auxiliary variables into the model, following the ideas of Oh and Park (2002). They facilitated implementation of MCMC and computation of marginal posterior density functions based on the Gibbs outputs by introducing auxiliary variables into random effects generalized linear models (GLM) for count data.

Given $\gamma_{t,}$ α, β, and η, let $\delta_t = \alpha + \gamma_t\beta + Z_t\eta$ and introduce a latent variable W_t following $N(\delta_t, 1)$ distribution. Define

$$Y_t = y_t \text{ if } \Phi^{-1}[F(y_t - 1 \mid \delta_t)] < W_t - \delta_t \leq \Phi^{-1}[F(y_t \mid \delta_t)],$$

where Φ is the standard normal cdf and F is the cdf of Y_t which depends on δ_t.

The joint density function of (Y_t, W_t) given δ_t is:

$$\begin{aligned} f(y_t, W_t \mid \delta_t) = \phi(W_t \mid \delta_t, 1) \\ \times I(\Phi^{-1}[F(y_t - 1 \mid \delta_t)] < W_t - \delta_t \\ \leq \Phi^{-1}[F(y_t \mid \delta_t)]), \end{aligned}$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the density function of $N(\mu, \sigma^2)$, and I is the indicator function. It can be easily shown that the marginal density function of Y_t, derived from the joint density $f(y_t, W_t \mid \delta_t)$, correctly gives the density $f(y_t \mid \delta_t)$ corresponding to the cdf $F(y_t \mid \delta_t)$.

With the extra auxiliary variables W_t in the model, we may consider $\theta = (\{\gamma_t\}, \Omega, \mathbf{P}, \Sigma, \mu, \alpha, \beta, \eta, \{W_t\})$ as new unknown parameters in the model. Now, the posterior density function of θ is given as:

$$\begin{aligned} &\pi(\theta|X,Y,Z) \qquad (15) \\ &\propto \prod_{t=1}^{T}\left[f\left(X_t|\gamma_t,\mathbf{P},\Sigma\right)p\left(\gamma_t|\Omega\right)f\left(y_t,W_t|\alpha,\beta,\gamma_t,\eta\right)\right] \times p(\Omega)p(\mathbf{P})p(\Sigma)p(\mu)p(\alpha)p(\beta)p(\eta) \\ &= \prod_{t=1}^{T}\left[|2\pi\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(X_t-\mu-\gamma_t\mathbf{P}\right)\Sigma^{-1}\left(X_t-\mu-\gamma_t\mathbf{P}\right)'\right\}\right. \\ &\quad \times |2\pi\Omega|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\gamma_t\Omega^{-1}\gamma_t{}'\right\} \\ &\quad \left.\times (2\pi)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(W_t-\alpha-\gamma_t\beta-Z_t\eta\right)^2\right\}\right]\mathbf{I}\left[H_t\left(W_t,\delta_t,y_t\right)\right] \\ &\quad \times \frac{|R_0|^{\frac{1}{2}r_0}|\Omega|^{-\frac{1}{2}(r_0+q+1)}\exp\left\{-\frac{1}{2}tr\left(R_0\Omega^{-1}\right)\right\}}{2^{\frac{1}{2}r_0q}\Gamma_q\left(\frac{1}{2}r_0\right)} \\ &\quad \times \left[\prod_{j=1}^{p^+}\Phi\left(c_0^j\Big/\sqrt{C_0^{j,j}}\right)\right]^{-1}|2\pi C_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{\left(vecP^+-c_0\right)'C_0^{-1}\left(vecP^+-c_0\right)\right\}\right]\mathbf{I}\left(vecP^+\geq\mathbf{0}\right) \\ &\quad \times \prod_{j=1}^{J}\frac{b_0^{a_0}}{\Gamma(a_0)}\left(\frac{1}{\sigma_j^2}\right)^{a_0+1}\exp\left(-\frac{b_0}{\sigma_j^2}\right) \times |2\pi M_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{(\mu-\mu_0)M_0^{-1}(\mu-\mu_0)'\right\}\right] \\ &\quad \times \left(2\pi U_0\right)^{-\frac{1}{2}}\exp\left[-\frac{1}{2U_0}\left(\alpha-\alpha_0\right)^2\right] \\ &\quad \times |2\pi\mathrm{B}_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{(\beta-\beta_0)\mathrm{B}_0^{-1}(\beta-\beta_0)'\right\}\right] \times |2\pi\Psi_0|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left\{(\eta-\eta_0)\Psi_0^{-1}(\eta-\eta_0)'\right\}\right], \end{aligned}$$

where

$$H_t\left(W_t,\delta_t,y_t\right) = \left\{\left(W_t,\delta_t,y_t\right);\Phi^{-1}F\left(y_t-1|\delta_t\right)<W_t-\delta_t\leq\Phi^{-1}F\left(y_t|\delta_t\right)\right\}. \qquad (16)$$

From the above posterior kernel, it can be easily seen that the full conditional density of vecP^+, μ, Ω, and Σ are given as a multivariate truncated normal, a multivariate normal, an IW, and an inverse gamma density, respectively. In addition, if we ignore the restriction in the indicator function of H_t, then the log posterior is a quadratic function of α, γ_t, β, η, W_t. Hence, the full conditional posterior distributions of α, γ_t, β, η, W_t are all restricted normal distributions. See Appendix C for the full conditional distributions for the parameters.

In the MCMC sampling algorithm employed here, one sweep consists of nine updating procedures: (i) updating W, (ii) updating Γ, (iii) updating α, (iv) updating β, (v) updating η, (vi) updating Ω, (vii) updating μ, (viii) updating $\mathbf{P}$, and (ix) updating Σ. The sample generation of Σ, μ, and Ω is straightforward because there is no restriction in the full conditional posterior distribution of those. The full conditional posterior distribution of P_j^+ is a multivariate normal distribution restricted to nonnegative numbers, from which sample generation is also relatively easy.

On the other hand, the full conditional posterior distributions of γ_t, β, η, α, and W_t are restricted multivariate normals (or univariate normals for α and W_t) with highly complex forms of restrictions, and sample generations are not easy. Appendix C also contains the algorithm for sample generation of those parameters.

The marginal likelihood for each model can then be estimated by:

$$\hat{l}(X,y \mid M) = \frac{l\left(X \middle| \theta^C, M\right) p\left(\theta^C \mid M\right)}{\hat{\pi}\left(\theta^C \mid X, y, M\right)}, \tag{17}$$

where θ^C is a single point of $\theta = (\{\gamma_t\}, \Omega, \mathbf{P}, \Sigma, \mu, \alpha, \beta, \eta, \{W_t\})$ under model M.

Using the same algorithm of Oh (1999), in Equation 18 we have:

$$\begin{aligned}
&\pi\left(\theta^C \middle| X, y, M\right) \qquad (18)\\
&= E\Big[\pi\left(W^C \middle| \Gamma^C, \alpha^C, \beta^C, \eta^C, \Omega^C, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\Gamma^C \middle| W, \alpha^C, \beta^C, \eta^C, \Omega^C, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\alpha^C \middle| W, \Gamma, \beta^C, \eta^C, \Omega^C, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\beta^C \middle| W, \Gamma, \alpha, \eta^C, \Omega^C, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\eta^C \middle| W, \Gamma, \alpha, \beta, \Omega^C, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\Omega^C \middle| W, \Gamma, \alpha, \beta, \eta, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\mu^C \middle| W, \Gamma, \alpha, \beta, \eta, \Omega, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\mathbf{P}^C \middle| W, \Gamma, \alpha, \beta, \eta, \Omega, \mu, \Sigma^C\right)\\
&\quad \times \pi\left(\Sigma^C \middle| W, \Gamma, \alpha, \beta, \eta, \Omega, \mu, \mathbf{P}\right)\Big]\\
&= E\Big[\pi\left(W^C \middle| \Gamma^C, \alpha^C, \beta^C, \eta^C\right)\\
&\quad \times \pi\left(\Gamma^C \middle| W, \alpha^C, \beta^C, \eta^C, \Omega^C, \mu^C, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\alpha^C \middle| W, \Gamma, \beta^C, \eta^C\right)\\
&\quad \times \pi\left(\beta^C \middle| W, \Gamma, \alpha, \eta^C\right)\\
&\quad \times \pi\left(\eta^C \middle| W, \Gamma, \alpha, \beta\right)\\
&\quad \times \pi\left(\Omega^C \middle| \Gamma\right) \times \pi\left(\mu^C \middle| \Gamma, \mathbf{P}^C, \Sigma^C\right)\\
&\quad \times \pi\left(\mathbf{P}^C \middle| \Gamma, \mu, \Sigma^C\right) \times \pi\left(\Sigma^C \middle| \Gamma, \mu, \mathbf{P}\right)\Big],
\end{aligned}$$

and $\hat{\pi}\left(\theta^C \middle| X, y, M\right)$ can be estimated as the sample average of the product of the full conditional posterior density functions using the posterior sample of θ under model M.

A MATLAB program implementing MCMC and computing marginal likelihoods using the above method was developed. Coding of the algorithm turned out to be a formidable task and took a considerable amount of time and effort.

Evaluation by Simulation

Simulation Studies Under the Normal Health Outcome Model We conducted three simulation studies to assess the performance of the new method that incorporates parameter uncertainty in source contributions into the estimation of source-specific health effects, assuming the normal health outcome models, while coping with uncertainty in both the number of sources and identifiability conditions in multivariate receptor models.

Simulation 1 The data were generated using the same parameter values for $\mathbf{P}$, Σ, and Ω used in the simulation study of Nikolov and colleagues (2006). They based their simulation on the known sources of Boston PM pollution: Road Dust (P_1), Power Plants (P_2), Oil Combustion (P_3), and Motor Vehicles (P_4), and obtained realistic parameter settings for $\mathbf{P}$, Σ, and Ω from a confirmatory factor analysis on the complete aggregated exposure data. They constrained one tracer element for each of the four sources according to the D1–D3 identifiability conditions. The exposure data were simulated from the model in Equation 9 with the following parameter values. For the true source-composition matrix $\mathbf{P}$,

$$\mathbf{P} = \begin{bmatrix} & \text{Si} & \text{S} & \text{Ni} & \text{OC} & \text{Al} & \text{Ti} & \text{Ca} & \text{Sulfate} & \text{Se} & \text{V} & \text{Br} & \text{BC} & \text{EC} \\ P_1 & 1 & 0 & 0 & 0 & 0.88 & 0.83 & 0.91 & 0.00 & 0.02 & 0.16 & 0.18 & 0.17 & 0.13 \\ P_2 & 0 & 1 & 0 & 0 & 0.00 & 0.08 & 0.02 & 0.95 & 0.65 & 0.04 & 0.58 & 0.41 & 0.27 \\ P_3 & 0 & 0 & 1 & 0 & 0.01 & 0.34 & 0.31 & 0.00 & 0.05 & 1.02 & 0.26 & 0.44 & 0.51 \\ P_4 & 0 & 0 & 0 & 1 & 0.00 & 0.09 & 0.17 & 0.01 & 0.26 & 0.03 & 0.43 & 0.65 & 0.81 \end{bmatrix},$$

$$diag(\Sigma) = (0.08, 0.05, 0.22, 0.45, 0.05, 0.28, 0.35, 0.05, 0.31, 0.05, 0.31, 0.11, 0.10),$$

$$\mu = (7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7),$$

$\gamma_t \sim N_4(\mathbf{0}, \Omega)$, and

$$\Omega = \begin{bmatrix} & P_1 & P_2 & P_3 & P_4 \\ P_1 & 2.36 & 0 & 0 & 0 \\ P_2 & 0 & 1.60 & 0 & 0 \\ P_3 & 0 & 0 & 1.49 & 0 \\ P_4 & 0 & 0 & 0 & 1.62 \end{bmatrix}.$$

In Nikolov and colleagues (2006), the health outcome data y_t were generated from the following normal linear model without considering the weather variables:

$Y_t = \alpha^N + \gamma_t \beta^N + \varepsilon_t$,

where

$\varepsilon_t \sim N\left(0, \sigma_y^{2N}\right)$, $\alpha^N = 86$, $\sigma_y^{2N} = 64$, and $\beta^N = (2\ \ 0\ \ 0\ \ 0)'$.

In our simulation, the health outcome data were generated from a more general model, Equation 10, incorporating the weather variables and the following parameter values: $\beta = (0.5\ \ 0\ \ 1\ \ 0.5)'$, $\eta = (1\ \ 0.5)'$, $\alpha = 3$, and $\sigma_y^2 = 1$. The weather data were generated from the lognormal distribution as follows:

$$Log(Z_{ti}) \sim N(0,0.5), \quad t = 1,\cdots,T, \quad i = 1,2.$$

Recall that in real applications, we do not model weather variables; we just use the measured weather data (or transformations of them) as covariates.

The sample size T was taken to be 100. With the parameter values given above, the normal health outcome model leads to data (y) in a simulated data set with an average number of counts per day (e.g., average mortality count per day) and a 95% interval (2.5th to 97.5th percentile) of 4.6 and (0.65–8.52), respectively.

As opposed to assuming the known number of sources ($q_0 = 4$) and identifiability conditions, we defined the candidate models by varying the number of sources ($q = 1, 2, 3, 4, 5$) along with identifiability conditions D1–D3 and estimated the parameters under each model as well as computed marginal likelihoods. Although the diagonal covariance matrix was used for Ω to generate the simulated data (as in Nikolov et al. 2006), we treated Ω as an unknown general covariance matrix in estimation, allowing estimation of correlations among the source contributions.

Recall that under the indifference prior model probabilities, the posterior model probability is proportional to the marginal likelihood. Thus, we only need to calculate the marginal likelihood of each model for model comparison. The simulation was repeated 200 times. Throughout the simulation the true parameter values for $\mathbf{P}$, β, Σ, Ω, and μ remain the same as given above, and only the errors (for X and y) are regenerated to obtain the data at each simulation. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,13$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $r_0 = q$, $R_0 = \mathbf{I_q} \times (r_0 + q + 1)$, $\alpha_0 = 0$, $U_0 = 100$, $\beta_0 = 0_q$, $B_0 = 100 \times \mathbf{I_q}$, $\eta_0 = 0_I$, $\Psi_0 = 100 \times \mathbf{I_I}$, $a_0^y = 0.01$, and $b_0^y = 0.01$. The estimated marginal likelihoods for each q-source model are reported in Table 1 on a log scale as logMD (log of marginal likelihood) (only 10 cases are shown for illustration). The selected model is the one having the maximum logMD for each dataset. The true model (with $q = 4$) was selected for 199 out of 200 simulations, that is, for 99.5% of times. (For only one simulation out of 200, a model with $q = 5$ was selected.)

We also monitored the R^2 values among the true source-composition profiles and the estimated source-composition

Table 1. Log of Marginal Likelihood (logMD) for q-Source Models ($q = 1, 2, 3, 4, 5$)

	Number of Sources (q)[a]				
Dataset	1	2	3	4	5
1	−2419.1	−2258.9	−2069.6	−2001.3	−2086.9
2	−2413.6	−2268.8	−2063.9	−2003.7	−2089.2
3	−2428.4	−2271.3	−2097.6	−2000.3	−2069.0
4	−2389.0	−2242.9	−2042.1	−1960.5	−2028.9
5	−2421.0	−2264.6	−2078.1	−2004.9	−2068.4
6	−2416.9	−2245.7	−2059.0	−1963.5	−2040.8
7	−2375.7	−2234.9	−2063.4	−2007.6	−2076.5
8	−2421.1	−2260.4	−2066.4	−1990.5	−2075.1
9	−2387.1	−2257.1	−2051.3	−1982.9	−2070.0
10	−2441.7	−2296.3	−2073.5	−1995.8	−2080.3

[a] The model having the maximum logMD was selected for each dataset. $q = 4$ was selected for all 10 of these datasets.

profiles as well as among the true source contributions and the estimated source contributions for $q = 4$. Throughout the simulation, R^2 values were all greater than 0.94, which indicates that the estimated source profiles and contributions agree well with the true source profiles and contributions.

Figure 1 presents the time-series plots of the true centered source contributions and the estimated centered source contributions (based on one of the simulated datasets), which again shows that the estimated source contributions are very close to the true source contributions. Note that

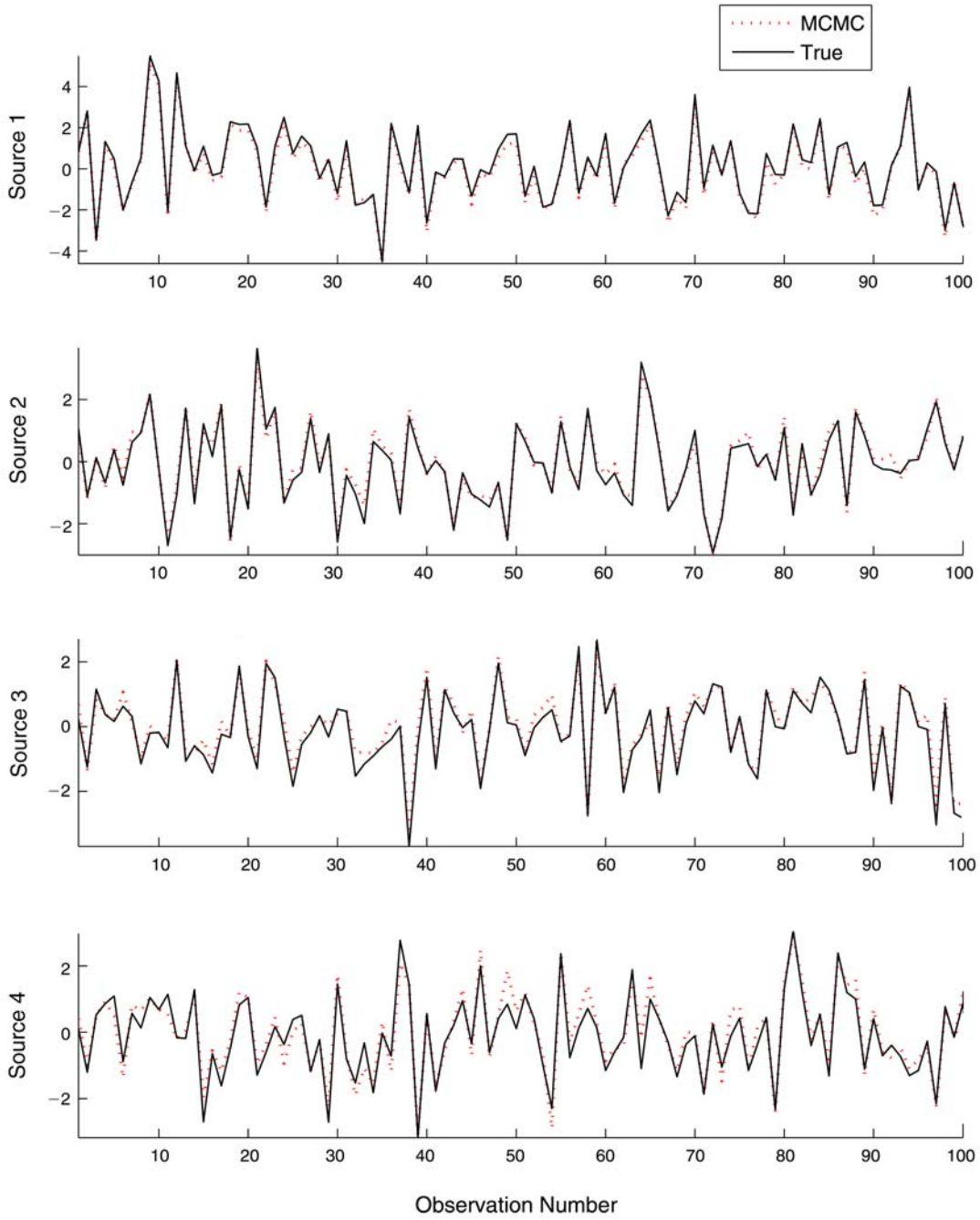


Figure 1. True and estimated centered source contributions from a simulated dataset under the normal health outcome model.

estimates for noncentered source contributions can be obtained by adding the estimated mean contribution $\hat{\xi} = \hat{\mu}\hat{\mathbf{P}}\left(\hat{\mathbf{P}}\hat{\mathbf{P}}'\right)^{-1}$ to the estimated centered source contributions. The mean squared errors for the source compositions and contributions under the true model were less than 0.01 and 0.1, respectively.

The estimated source-specific health-effects parameter β was very close to the true value throughout the simulation. The 95% posterior intervals were computed in each simulation. Overall, the posterior interval for each element of β contained the true value approximately 96% of the time (there were 32 instances out of 800 when the posterior interval for an element of β did not contain the true value). The average widths of posterior intervals for β parameters for each of the four sources over 200 simulations were 0.31, 0.38, 0.42, and 0.39, respectively.

Simulation 2 Next, we performed a simulation to compare the performances of two methods. Method 1 (an oblique factor model) assumed a priori correlated source contributions and Method 2 (an orthogonal factor model) assumed a priori uncorrelated source contributions. The data were generated using the same parameter values given above except for Ω. We used the Ω matrix allowing correlations among different source contributions for this simulation:

$$\Omega = \begin{bmatrix} & P_1 & P_2 & P_3 & P_4 \\ P_1 & 2.36 & 0.8 & 0.8 & 0.8 \\ P_2 & 0.8 & 1.60 & 0.8 & 0.8 \\ P_3 & 0.8 & 0.8 & 1.49 & 0.8 \\ P_4 & 0.8 & 0.8 & 0.8 & 1.62 \end{bmatrix}.$$

The sample size T was again taken to be 100. Recall that Method 2 does not need a separate normalization constraint, D3, because indeterminacy resulting from the multiplication of a row of $\mathbf{P}$ by a scale constant is removed by assuming a priori the unit standard deviations of the source contributions. We defined the candidate models by varying the number of sources (q = 1, 2, 3, 4, 5) along with identifiability conditions D1–D3 for Method 1 and D1–D2 for Method 2.

The simulation was repeated 50 times for each of Methods 1 and 2. Throughout the simulation the true parameter values for $\mathbf{P}$, β, Σ, Ω, and μ remain the same as given above, and only the errors (for X and y) are regenerated to obtain the data at each simulation. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,13$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $r_0 = q$, $R_0 = \mathbf{I_q} \times (r_0 + q + 1)$, $\alpha_0 = 0$, $U_0 = 100$, $\beta_0 = 0_q$, $B_0 = 100 \times \mathbf{I_q}$, $\eta_0 = 0_I$, $\Psi_0 = 100 \times \mathbf{I_I}$, $a_0^Y = 0.01$, and $b_0^Y = 0.01$. The estimated marginal likelihoods for each q-source model from Method 2 are reported in Table 2 on a log scale as logMD (only 5 cases are shown for illustration). The selected model is the one having the maximum logMD for each dataset. For both Methods 1 and 2, the true model (with q = 4) was selected for all 50 of the simulations.

We also monitored the R^2 values among the true source-composition profiles and the estimated source-composition profiles as well as among the true source contributions and the estimated source contributions for q = 4. Throughout the simulation, R^2 values were greater than 0.95 for both Methods 1 and 2, which indicates that the estimated source profiles and contributions agreed well with the true source profiles and contributions. Figure 2

Table 2. Log of Marginal Likelihood (logMD) for q-source Models from Method 2 (q = 1, 2, 3, 4, 5)

	Number of Sources (q)[a]				
Dataset	1	2	3	4	5
1	−1972.2	−1789.8	−1696.4	−1569.6	−1720.4
2	−1957.2	−1754.6	−1664.7	−1574.1	−1746.3
3	−1936.2	−1767.1	−1662.2	−1558.1	−1707.7
4	−1972.0	−1807.0	−1668.2	−1563.0	−1692.2
5	−1947.7	−1783.8	−1670.1	−1571.3	−1726.5

[a] The model having the maximum logMD was selected for each dataset. q = 4 was selected for all 5 of these datasets.

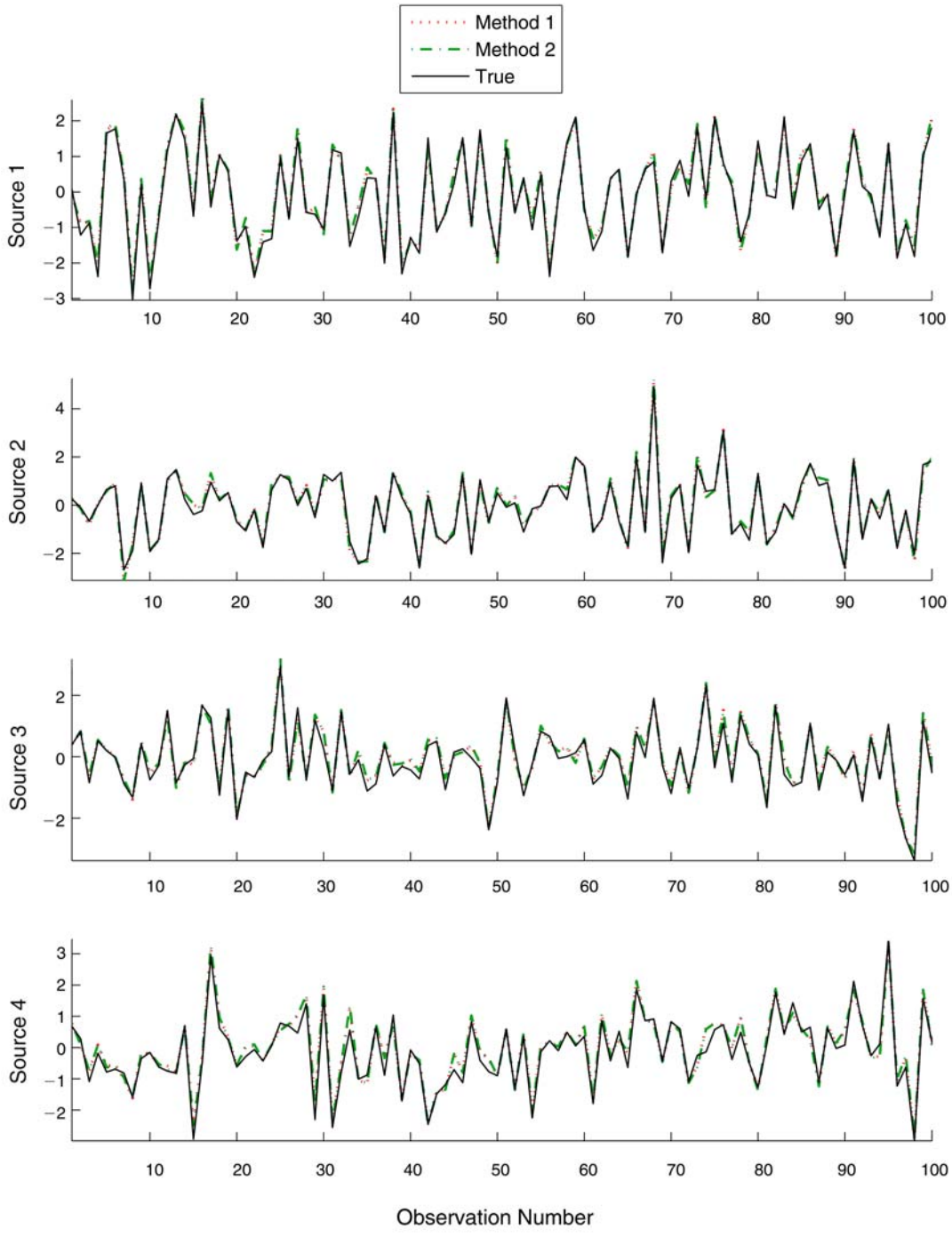


Figure 2. True and estimated centered source contributions by Methods 1 and 2 from a simulated dataset under the normal health outcome model.

presents the time-series plots of the true centered source contributions and the estimated centered source contributions (based on one of the simulated datasets) from Methods 1 and 2 (after rescaling the estimated source contributions appropriately, i.e., by the diagonal elements of the leading $q \times q$ matrix of **P** in this case), which shows that the estimated source contributions from Methods 1 and 2 are very close to the true source contributions. For both Methods 1 and 2, the mean squared errors for the source compositions and contributions under the true model were less than 0.01 and 0.1, respectively.

The estimated source-specific health-effects parameter β was close to the true value for both methods throughout the simulation. The 95% posterior intervals were computed in each simulation. For Method 1, the posterior interval for each element of β contained the true value approximately 98% of the time (there were 4 instances out of 200 when the posterior interval for an element of β did not contain the true value). The average widths of posterior intervals for β parameters for each of the four sources over 50 simulations were 0.31, 0.43, 0.44, and 0.41, respectively. For Method 2, the posterior interval for each element of β contained the true value approximately 98% of the time (there were also 4 instances out of 200 when the posterior interval for an element of β did not contain the true value). The average widths of posterior intervals for β parameters for each of the four sources over 50 simulations were 0.37, 0.44, 0.53, and 0.44, respectively. Our limited simulation suggests that there is not a real advantage of modeling γ to be correlated a priori in terms of parameter estimation. We could achieve approximately the same accuracy in estimating the key parameters, **P**, γ, and β, with assuming a priori an orthogonal factor model even when γ's were actually generated under an oblique factor model in simulation. We conjecture that it is because the form of the prior distribution of the latent variables, γ_t, is essentially arbitrary and largely a matter of convention as mentioned before.

Simulation 3 We performed an additional simulation to investigate the impact of preassigning zeros in **P** for identifiability conditions when the corresponding elements in true **P** are not zeros. This time we generated the data with a different true **P** that does not assume the existence of a tracer element for each source type and Ω matrix allowing correlations among different source contributions:

$$\mathbf{P} = \begin{bmatrix} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 & s_9 & s_{10} & s_{11} & s_{12} & s_{13} \\ P_1 & 1 & 0 & 0 & 0.2 & 0.88 & 0.83 & 0.91 & 0.20 & 0 & 0.76 & 0.18 & 0.17 & 0.13 \\ P_2 & 0 & 1 & 0.3 & 0 & 0.30 & 0.08 & 0.02 & 0.65 & 0.65 & 0 & 0.58 & 0.41 & 0.27 \\ P_3 & 0 & 0 & 1 & 0.6 & 0.20 & 0.34 & 0.80 & 0 & 0.05 & 1.02 & 0.26 & 0.44 & 0.51 \\ P_4 & 0.2 & 0 & 0 & 1 & 0.10 & 0.09 & 0.95 & 0.50 & 0.90 & 0 & 0.43 & 0.65 & 0.81 \end{bmatrix},$$

and

$$\Omega = \begin{bmatrix} & P_1 & P_2 & P_3 & P_4 \\ P_1 & 2.36 & 0.8 & 0.8 & 0.8 \\ P_2 & 0.8 & 1.60 & 0.8 & 0.8 \\ P_3 & 0.8 & 0.8 & 1.49 & 0.8 \\ P_4 & 0.8 & 0.8 & 0.8 & 1.62 \end{bmatrix}.$$

The other parameters are the same as those of Simulation 1. To examine the effect of preassigning zeros to nonzero elements on estimation, we included multiple 4-source candidate models with different sets of prespecified zeros in model comparison (note that true q is 4). In Model 2, zeros are preassigned to truly zero elements, in Model 3 zeros are preassigned to some minor nonzero elements (s_6 for Source 2, s_9 for Source 3, and s_6 for Source 4), in Model 4 a zero is preassigned to a major nonzero element (s_7 for Source 3). We also included models with an incorrect number of sources (Models 1 and 5). Table 3 gives the candidate models compared in Simulation 3.

The simulation was repeated 50 times with applying the method a priori assuming an orthogonal factor model (Method 2) with the preassigned zeros given in Table 3. We observed in this case that assuming an orthogonal factor models for γ is more beneficial in terms of more easily satisfying the constraint C2 and avoiding numerical problems in MCMC implementation. Throughout the simulation the true parameter values for $\mathbf{P}$, β, Σ, Ω, and μ remain the same, and only the errors (for X and y) are regenerated to obtain the data at each simulation. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,13$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $r_0 = q$, $R_0 = \mathbf{I_q} \times (r_0 + q + 1)$, $\alpha_0 = 0$, $U_0 = 100$, $\beta_0 = 0_q$, $B_0 = 100 \times \mathbf{I_q}$, $\eta_0 = 0_I$, $\Psi_0 = 100 \times \mathbf{I_I}$, $a_0^y = 0.01$, and $b_0^y = 0.01$. The estimated marginal likelihoods (on a log scale as logMD) as well as the corresponding posterior probabilities for candidate models considered are reported in Table 4 (only 5 cases are shown for illustration).

The simulation results suggest that the results are not sensitive to errors of assuming a zero where the truth is nonzero as long as the preassigned zero element is not actually large. Model 2 and Model 3 resulted in the highest marginal likelihood (or posterior model probability) for 44

Table 3. Candidate Models in Simulation 3

Model Number	q	Source	Prespecified Position of Zeros in P
1	3	1	s_2, s_3
		2	s_1, s_{10}
		3	s_1, s_2
2	4	1	s_2, s_3, s_9
		2	s_1, s_4, s_{10}
		3	s_1, s_2, s_8
		4	s_2, s_3, s_{10}
3	4	1	s_2, s_3, s_9
		2	s_1, s_4, s_6
		3	s_1, s_2, s_9
		4	s_2, s_3, s_6
4	4	1	s_2, s_3, s_9
		2	s_1, s_4, s_{10}
		3	s_1, s_2, s_7
		4	s_2, s_3, s_{10}
5	5	1	s_2, s_3, s_9, s_{13}
		2	s_1, s_4, s_6, s_7
		3	s_1, s_2, s_8, s_9
		4	s_2, s_3, s_6, s_{10}
		5	s_1, s_2, s_3, s_4

Table 4. Marginal Likelihoods for Candidate Models in Simulation 3[a]

Dataset / Model Number (q)	LogMD ($\times 10^4$)	PostP
1		
1 (3)	−1.7120	0.0000
2 (4)	−1.6312	1.0000
3 (4)	−1.6475	0.0000
4 (4)	−1.7052	0.0000
5 (5)	−1.6858	0.0000
2		
1 (3)	−1.6713	0.0000
2 (4)	−1.5924	0.9874
3 (4)	−1.5968	0.0126
4 (4)	−1.6645	0.0000
5 (5)	−1.6412	0.0000
3		
1 (3)	−1.6982	0.0000
2 (4)	−1.6195	0.9982
3 (4)	−1.6258	0.0018
4 (4)	−1.6814	0.0000
5 (5)	−1.6744	0.0000
4		
1 (3)	−1.6839	0.0000
2 (4)	−1.6080	0.6161
3 (4)	−1.6084	0.3839
4 (4)	−1.6738	0.0000
5 (5)	−1.6712	0.0000
5		
1 (3)	−1.6627	0.0000
2 (4)	−1.5767	0.9160
3 (4)	−1.5790	0.0840
4 (4)	−1.6200	0.0000
5 (5)	−1.7041	0.0000

[a] LogMD and PostP denote the Log of Marginal Likelihood and Posterior Model Probability, respectively.

and 6 out of 50 simulations, respectively. A candidate model incorrectly preassigning a zero to a major element (Model 4) as well as models with incorrect q (Model 1 or Model 5) results in smaller marginal likelihoods than models preassigning zeros to true zero elements (Model 2) or to minor nonzero elements (Model 3) almost all the time. As can be observed from Table 4, posterior probabilities for Models 2 and 3 are both greater than zero for datasets 2–5 and are sometimes comparable because the two models are very similar in nature. (Model 3 preassigning zeros to some minor elements in **P** is not materially different from Model 2.) The estimated **P** under both models are very close to each other (R^2 values were greater than 0.96) when the corresponding posterior probabilities are comparable. Figure 3 presents the time-series plots of the true centered source contributions and the estimated centered source

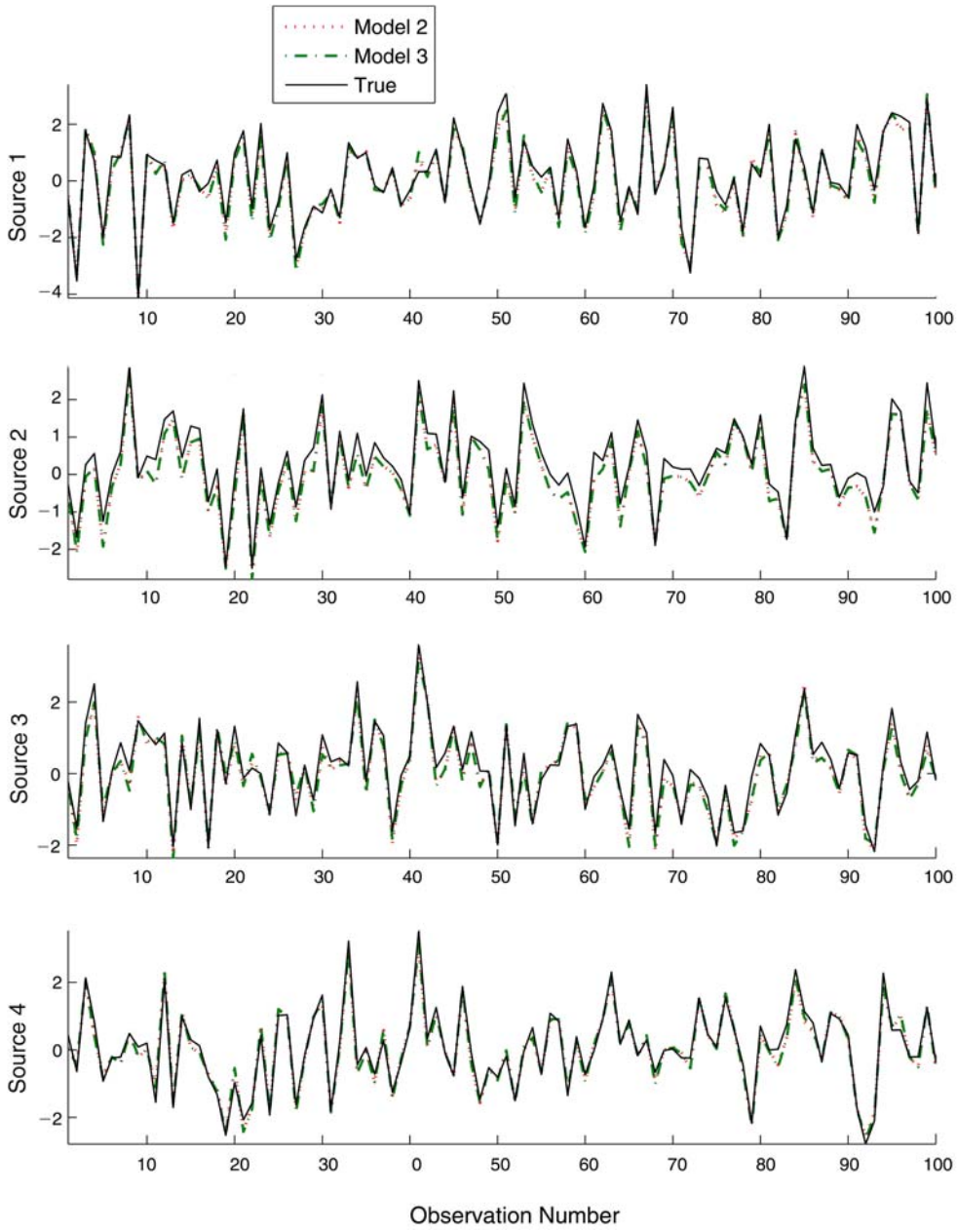


Figure 3. True and estimated centered source contributions under Models 2 and 3 based on Dataset 5 of Table 4.

contributions (based on one of the simulated datasets, Dataset 5 in Table 4) from Models 2 and 3, which shows that the estimated source contributions from Models 2 and 3 are both very close to the true source contributions and not really distinguishable.

The source-specific health-effects parameter β estimated under Models 2 and 3 are also close to each other when the corresponding posterior probabilities are comparable. The 95% posterior intervals were computed in each simulation. For Model 2, the posterior interval for each element of β contained the true value approximately 95% of the time (there were 11 instances out of 200 when the posterior interval for an element of β did not contain the true value). The average widths of posterior intervals for β parameters for each of the four sources over 50 simulations were 0.37, 0.41, 0.55, and 0.44, respectively.

It is worth noting that Bayesian model averaging (BMA) can be applied in the cases that we compare the models (such as Models 2 and 3) with different identifiability conditions but with the same number of sources and same source types. (Model averaging does not make sense when the number of sources or the interpretation of source profiles changes.) As a matter of fact, we may apply BMA to parameters between Model 2 and Model 3 although the actual consequence of BMA would not be much different from selecting one of Model 2 or Model 3 in this case (again because those two models are very similar in nature). This simulation suggests that the estimation results are not sensitive to errors of assuming a zero where the truth is nonzero as long as the preassigned zero element is not a major constituent of a source profile.

Simulation Study Under the Poisson Health Outcome Model We conducted a simulation study to assess the performance of the new method that incorporates parameter uncertainty in source contributions into the estimation of source-specific health effects, assuming the Poisson health outcome models, while coping with uncertainty in both the number of sources and identifiability conditions in multivariate receptor models.

The air pollution data were generated from the model in Equation 9 using the same parameter values for $\mathbf{P}$, Σ, and Ω used in the simulation study of Nikolov and colleagues (2006) presented in the previous section. The health outcome data y_t were generated from the following Poisson model:

$$y_t \sim Poisson(\lambda_t),$$

$$\log(\lambda_t) = \alpha + \gamma_t\beta + Z_t\eta = \alpha + \sum_{k=1}^{q} \beta_k\gamma_{tk} + \sum_{i=1}^{I} \eta_i Z_{ti},$$

where $\alpha = -0.1$, $\beta = (0 \quad 0.1 \quad 0.2 \quad 0)'$, $\eta = (0.2 \quad 0.1)'$, and $Log(Z_{ti}) \sim N(0,0.5)$, $t = 1,\cdots,T$, $i = 1,2$.

The sample size T was taken to be 200. Note that we tried to simulate the least favorable case (in terms of the sample size) for which our method still works. For the Poisson health outcome model, we needed a larger sample size than normal for a reasonable estimation of β. With the parameter values given above, the Poisson health outcome model leads to data (y) in a simulated dataset with an average number of counts per day (e.g., average mortality count per day) and a 95% interval (2.5th to 97.5th percentile) of 1.28 and (0–4), respectively.

Again, as opposed to assuming the known number of sources ($q_0 = 4$) and identifiability conditions, we defined the candidate models by varying the number of sources ($q = 3, 4, 5$) along with identifiability conditions D1–D3 and estimated the parameters under each model as well as computed marginal likelihoods.

It needs to be noted that implementing the Poisson health outcome model by MCMC took considerably more time (approximately 50 times longer) than implementing the normal health outcome model. For this reason, the simulation could be repeated only 30 times rather than 200 times as in the normal health outcome model. Throughout the simulation the true parameter values for $\mathbf{P}$, Σ, Ω, μ, β, η, and α remain the same as given above, and only the errors (for X and y) are regenerated to obtain the data at each simulation. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,13$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $r_0 = q$, $R_0 = \mathbf{I_q} \times (r_0 + q + 1)$, $\alpha_0 = 0$, $U_0 = 10$, $\beta_0 = 0_q$, $B_0 = 10 \times \mathbf{I_q}$, $\eta_0 = 0_I$, and $\Psi_0 = 10 \times \mathbf{I_I}$. The estimated marginal likelihoods for each q-source model are reported in Table 5 on a log scale (only 10 cases are shown for illustration). The selected model is the one having the maximum logMD. The true model (with $q = 4$) was selected for all of 30 simulations.

We also monitored the R^2 values among the true source-composition profiles and the estimated source-composition profiles as well as among the true source contributions and the estimated source contributions for $q = 4$. Throughout the simulation, R^2 values were all greater than 0.99, which indicates that the estimated source profiles and contributions agree well with the true source profiles and contributions. Figure 4 presents the time-series plots of the true centered source contributions and the estimated centered source contributions (based on one of the simulated datasets). Note that estimates for noncentered source contributions can be obtained by adding the estimated mean contribution

Table 5. Log of Marginal Likelihood (logMD) for q-Source Models (q = 3, 4, 5)

	Number of Sources (q)[a]		
Dataset	3	4	5
1	−3518.7	−3289.2	−3379.1
2	−3664.0	−3232.9	−3327.9
3	−3604.1	−3315.0	−3400.6
4	−3615.8	−3334.0	−3406.7
5	−3649.1	−3204.0	−3299.1
6	−3460.0	−3179.2	−3284.7
7	−3531.0	−3230.5	−3363.1
8	−3473.6	−3197.5	−3294.2
9	−3719.8	−3275.6	−3368.8
10	−3603.1	−3329.3	−3428.1

[a] The model having the maximum logMD was selected for each dataset. q = 4 was selected for all 10 of these datasets.

$\hat{\xi} = \hat{\mu}\hat{\mathbf{P}}\left(\hat{\mathbf{P}}\hat{\mathbf{P}}'\right)^{-1}$ to the estimated centered source contributions. The mean squared errors for the source compositions and contributions under the true model were less than 0.01 and 0.1, respectively.

The 95% posterior intervals were computed in each simulation. The posterior interval for each element of β contained the true value approximately 96% of the time (there were 5 instances out of 120 when the posterior interval for an element of β did not contain the true value).

APPROACH TO INCORPORATING SPATIAL DEPENDENCE IN MULTIPOLLUTANT DATA FROM MULTIPLE MONITORING SITES INTO MULTIVARIATE RECEPTOR MODELING: BAYESIAN SPATIAL MULTIVARIATE RECEPTOR MODELS

In this portion of the project, we developed a Bayesian spatial multivariate receptor modeling approach that can incorporate spatial dependence into the estimation of source profiles and contributions as well as cope with the unknown number of pollution sources and identifiability conditions. Space–time variation in source contributions and their associated levels of uncertainty can be examined using this approach. These enhanced multivariate receptor models enable prediction of unobserved source contributions (and pollutant concentrations) at locations other than the monitoring sites, which allows inference about the source-specific exposures and associated health effects in the study areas that do not have any monitoring stations.

Spatially Extended Multivariate Receptor Models

Let N be the number of receptors. As given in Equation 1 of the earlier section concerning a single receptor site, the standard model for the rth receptor at time t is:

$$X_t^r = A_t^r \mathbf{P} + E_t^r, \quad t = 1,\cdots,T, \quad r = 1,\cdots,N, \tag{19}$$

where $\mathbf{P}$ is a $q \times J$ source-composition matrix, $X_t^r = \left(X_{t1}^r,\cdots,X_{tJ}^r\right)$ is a vector of observed concentrations on J pollutants at receptor r at time t, $A_t^r = \left(A_{t1}^r,\cdots,A_{tq}^r\right)$ is a vector of contributions from q sources at receptor r at time t, and $E_t^r = \left(E_{t1}^r,\cdots,E_{tJ}^r\right)$ is a J-dimensional vector of errors associated with each observation at the rth receptor and time t. The elements of $\mathbf{P}$ are constrained to be nonnegative.

We model these multivariate spatial temporal data X_t^r by adapting the dynamic factor process convolution models (a version of multivariate spatial temporal process convolution models) introduced by Calder (2003, 2007), although we focus on spatial modeling and do not pursue the dynamic nature of the model over time in this project. Unlike traditional geostatistical models, the dynamic factor process convolution models do not require defining the cross-covariance function directly. Dynamic factor process convolution models are a generalization of the discrete process convolution approach to modeling spatial data proposed by Higdon (1998) for which the spatial process is expressed as a sum of the discrete underlying (latent) process defined on L locations on a coarse grid $\{\omega_1, \omega_2,\cdots, \omega_L\}$, covering the spatial domain, smoothed by the kernel κ. It is known that the covariance function of the spatial process obtained by this approach is guaranteed to be valid

(positive definite) and features such as nonstationarity and anisotropy can also be easily incorporated into the model (see Calder 2003, 2007). The dynamic factor process convolution models are constructed by inserting the factor model into the multivariate dynamic process convolution model to specify the covariance among the columns of the data at time t. This approach has several advantages, compared to traditional geostatistical models, in that it can cope with nonseparable covariance functions, potential asymmetry in the cross-covariance function, and with misaligned and missing data.

In Calder (2003, 2007), the number of factors (q) was assumed either known or chosen by a rule-of-thumb method without considering uncertainty in q. The model identifiability conditions were also assumed known. The conditions used in Calder (2003, 2007) assumes zeros for all elements of the lower triangular matrix of **P** (coupled with an assumption of orthogonal factor models). The assumption of the lower triangular matrix of **P** makes the order of

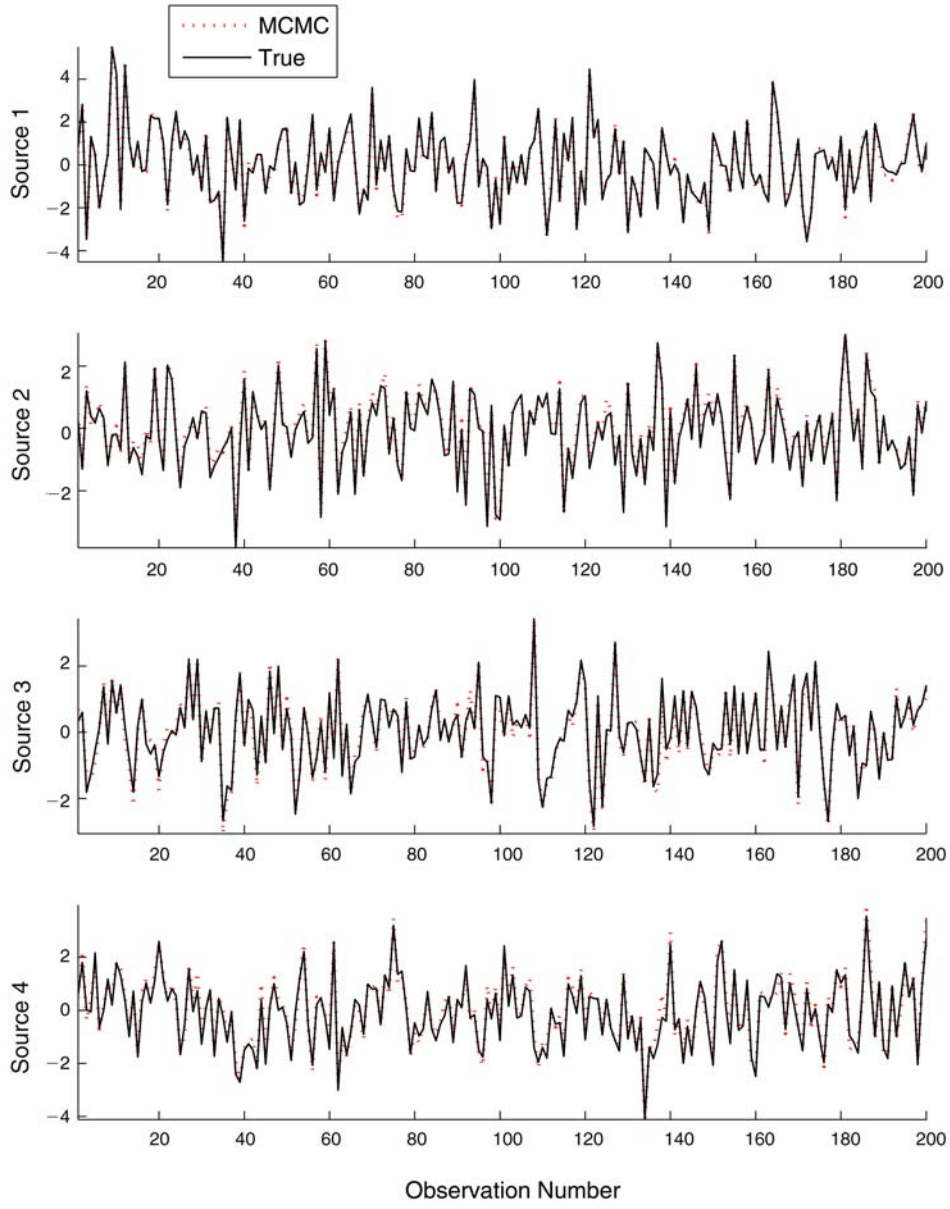


Figure 4. True and estimated centered source contributions from a simulated dataset under the Poisson health outcome model.

the pollutant series in **X** matter (because such constraint in **P** implies that the first factor is completely determined by the first series [pollutant] and the second factor is completely determined by the first and second series [pollutants], and so on), and the estimates of source-composition matrix and source contributions may be affected significantly depending on the order of columns in the data. Our Bayesian spatial multivariate receptor modeling relaxes the assumption of the known number of sources and identifiability conditions and also incorporates physically meaningful non-negativity constraints (that were not enforced in Calder 2007) for the source-composition matrix **P** into the estimation.

To extend multivariate receptor models to incorporate spatial dependence in multipollutant data obtained from multiple monitoring sites, we consider the following model for the multivariate data $\{\mathbf{X}(s_r, t), t = 1,\cdots,T\}$ collected at N spatial sites $\{s_1, s_2,\cdots, s_N\}$ over T time points:

$$\mathbf{X}(s_r, t) = \mathbf{K}(r)\mathbf{G_t}\mathbf{P} + \mu + \mathbf{E}(s_r, t), \quad (20)$$

where s_r is the spatial location of the r th receptor ($r = 1, \cdots, N$), $\mathbf{G_t}$ represents q underlying processes located at L spatial locations $\{\omega_1, \omega_2, \cdots, \omega_L\}$, $\mathbf{K}(r) = [\kappa(\omega_1 - s_r),\cdots,\kappa(\omega_L - s_r)]$, κ is a smoothing kernel, $\mu = (\mu_1,\cdots,\mu_J)$ is the mean of $\mathbf{X}(s_r, t)$, and $\mathbf{E}(s_r, t)$ is an independent and identically distributed, mean zero, Gaussian process on (s_r, t) with variance $\Sigma = diag\left(\sigma_1^2,\cdots,\sigma_J^2\right)$. Note that the source contributions at any location can be estimated as $\mathbf{K}(r)\mathbf{G_t}$ by plugging in the estimates for $\mathbf{G_t}$ and the corresponding values for $\mathbf{K}(r)$.

In matrix terms, the above model can be written as

$$\mathbf{X}_t = \mathbf{K}\mathbf{G_t}\mathbf{P} + \boldsymbol{\mu} + \mathbf{E}_t, \quad (21)$$

where

$$\mathbf{X_t} = \begin{bmatrix} X_{t1}^1 & X_{t2}^1 & \cdots & X_{tJ}^1 \\ X_{t1}^2 & X_{t2}^2 & \cdots & X_{tJ}^2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{t1}^N & X_{t2}^N & \cdots & X_{tJ}^N \end{bmatrix}_{N\times J},$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_J \\ \mu_1 & \mu_2 & \cdots & \mu_J \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1 & \mu_2 & \cdots & \mu_J \end{bmatrix}_{N\times J},$$

$$\mathbf{G_t} = \begin{bmatrix} G_{t1}^1 & G_{t2}^1 & \cdots & G_{tq}^1 \\ G_{t1}^2 & G_{t2}^2 & \cdots & G_{tq}^2 \\ \vdots & \vdots & \ddots & \vdots \\ G_{t1}^L & G_{t2}^L & \cdots & G_{tq}^L \end{bmatrix}_{L\times q},$$

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}(1) \\ \mathbf{K}(2) \\ \vdots \\ \mathbf{K}(N) \end{bmatrix}_{N\times L} = \begin{bmatrix} \kappa(\omega_1 - s_1) & \kappa(\omega_2 - s_1) & \cdots & \kappa(\omega_L - s_1) \\ \kappa(\omega_1 - s_2) & \kappa(\omega_2 - s_2) & \cdots & \kappa(\omega_L - s_2) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\omega_1 - s_N) & \kappa(\omega_2 - s_N) & \cdots & \kappa(\omega_L - s_N) \end{bmatrix}_{N\times L}, \text{ and}$$

$$\mathbf{E_t} = \begin{bmatrix} E_{t1}^1 & E_{t2}^1 & \cdots & E_{tJ}^1 \\ E_{t1}^2 & E_{t2}^2 & \cdots & E_{tJ}^2 \\ \vdots & \vdots & \ddots & \vdots \\ E_{t1}^N & E_{t2}^N & \cdots & E_{tJ}^N \end{bmatrix}_{N\times J}.$$

We assume that $\mathbf{G_t}$ and $\mathbf{E_t}$ are independent across location and time and have matrix normal distributions (see Dawid 1981) as follows:

$$\mathbf{G_t} \sim N(\mathbf{0},\mathbf{I_L},\Omega) \text{ and} \tag{22}$$
$$\mathbf{E_t} \sim N(\mathbf{0},\mathbf{I_N},\Sigma). \tag{23}$$

Using *vec*(·) operator stacking the columns of a matrix, the model in Equations 21–23 can be written as:

$$vec\left(\mathbf{X_t}'\right) = (\mathbf{K}\otimes\mathbf{P}')vec\left(\mathbf{G_t}'\right) + vec(\boldsymbol{\mu}') + vec\left(\mathbf{E_t}'\right), \tag{24}$$

$$vec\left(\mathbf{G_t}'\right) \sim N_{Lq}\left(\mathbf{0}_{Lq\times 1},\mathbf{I_L}\otimes\Omega\right), \text{ and} \tag{25}$$

$$vec\left(\mathbf{E_t}'\right) \sim N_{NJ}\left(\mathbf{0}_{NJ\times 1},\mathbf{I_N}\otimes\Sigma\right). \tag{26}$$

Let $\mathbf{X} = \begin{bmatrix}\mathbf{X_1}\\ \vdots \\ \mathbf{X_T}\end{bmatrix}$, $\mathbf{G} = \begin{bmatrix}\mathbf{G_1}\\ \vdots \\ \mathbf{G_T}\end{bmatrix}$, $\boldsymbol{\Gamma} = \begin{bmatrix}\boldsymbol{\Gamma_1}\\ \vdots \\ \boldsymbol{\Gamma_T}\end{bmatrix}$, and

$\boldsymbol{\Gamma_t} = \mathbf{K}\mathbf{G_t}$.

For prior distributions of parameters, we assume independent priors $p(\mathbf{G},\mathbf{P},\Sigma,\mu,\Omega) = p(\mathbf{G}\mid\Omega)p(\mathbf{P})p(\Sigma)p(\mu)p(\Omega)$ with multivariate truncated normal for free elements of $\mathbf{P}$ inverse gamma for diagonal elements of Σ, multivariate normal for μ, and IW distribution for Ω, respectively. That is,

$$vecP^{+} \sim N_{p^{+}}(c_0,C_0)\mathbf{I}\left(vecP^{+} \geq \mathbf{0}\right),$$
$$\left(\sigma_j\right)^{-2} \sim Gamma\left(\alpha_0,\beta_{0j}\right),\ j = 1,\cdots,J,$$
$$\mu \sim N_J\left(m_0,M_0\right), \text{ and}$$
$$\Omega \sim \mathrm{IW}\left(R_0,r_0\right).$$

We used the Gibbs sampling algorithm in implementation of MCMC for parameter estimation. One sweep consists of five updating procedures: (i) updating $\mathbf{G}$, (ii) updating $\mathbf{P}$, (iii) updating Σ, (iv) updating Ω, and (v) updating μ.

The full conditional distributions of $\mathbf{G}$, Σ, μ, Ω, and $\mathbf{P}$ are given as follows:

$$vec(G_t)|\cdots \sim N_{Lq}\left(m_{G_t},V_G\right),$$

where

$$V_G = \left\{(\mathbf{I_L}\otimes\Omega)^{-1} + (\mathbf{K}\otimes\mathbf{P}')'(\mathbf{I_N}\otimes\Sigma)^{-1}(\mathbf{K}\otimes\mathbf{P}')\right\}^{-1}, \text{ and}$$
$$m_G = V_G\left\{(\mathbf{K}\otimes\mathbf{P}')'(\mathbf{I_N}\otimes\Sigma)^{-1}\left(vec\left(\mathbf{X_t}'\right) - vec(\boldsymbol{\mu}')\right)\right\};$$
$$\left(\sigma_j\right)^{-2}\Big|\cdots \sim Gamma\left(a_0 + \tfrac{1}{2}TN, b_{0j} + \tfrac{1}{2}d_j\right),\ j = 1,\cdots,J,$$

where d_j is the jth diagonal element of

$$d = \left(\mathbf{X} - 1_{TN}\otimes\mu - \boldsymbol{\Gamma}\mathbf{P}\right)'\left(\mathbf{X} - 1_{TN}\otimes\mu - \boldsymbol{\Gamma}\mathbf{P}\right);$$
$$\mu|\cdots \sim N_J\left(m_\mu,V_\mu\right),$$
where
$$m_\mu = \left\{TN\left(\bar{X} - \bar{\gamma}\mathbf{P}\right)\Sigma^{-1} + m_0M_0^{-1}\right\}V_\mu, \text{ and}$$
$$V_\mu = \left(TN\Sigma^{-1} + M_0^{-1}\right)^{-1};$$
and $\Omega \sim \mathrm{IW}\left(R,TN + r_0\right)$, where $R = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} + R_0$.

For the columns of $\mathbf{P}$ with no preassigned (zero or one) elements,

$$P_j\Big|\cdots \sim N_q\left(c_j,C_j\right)\mathbf{I}\left(P_{kj} \geq 0, k = 1,\cdots,q\right),\ j = 1,\cdots,J,$$
where
$$c_j = C_j\left\{\left(\sigma_j\right)^{-2}\boldsymbol{\Gamma}'\left(X_j - \mu_j 1_{TN}\right) + C_{0j}^{-1}c_{0j}\right\},$$
$$C_j = \left(\left(\sigma_j\right)^{-2}\boldsymbol{\Gamma}'\boldsymbol{\Gamma} + C_{0j}^{-1}\right)^{-1},$$

c_{0j} is a q-dimensional prior mean vector of P_j, C_{0j} is a corresponding submatrix of C_0, X_j is the jth column of $\mathbf{X}$, and 1_{TN} is the TN-dimensional column vector consisting of 1's. For the columns of $\mathbf{P}$ containing preassigned (zero or one) elements,

$$P_j^{+}\Big|\cdots \sim N_{q+}\left(c_j^{+},C_j^{+}\right)\mathbf{I}\left(P_{kj} \geq 0, k = 1,\cdots,q_j^{+}\right),$$

where q_j^{+} is the number of free elements in the jth column of $\mathbf{P}$; and

$$c_j^{+} = C_j^{+}\left\{\left(\sigma_j\right)^{-2}\boldsymbol{\Gamma}_j^{+\prime}\left(X_j - \mu_j 1_{TN}\right) + \left(C_{0j}^{+}\right)^{-1}c_{0j}^{+}\right\}, \text{ and}$$
$$C_j^{+} = \left\{\left(\sigma_j\right)^{-2}\boldsymbol{\Gamma}_j^{+\prime}\boldsymbol{\Gamma}_j^{+} + \left(C_{0j}^{+}\right)^{-1}\right\}^{-1},$$

where c_{0j}^{+} is a q_j^{+}-dimensional prior mean vector of P_j^{+}, C_{0j}^{+} is a corresponding submatrix of C_0, and $\boldsymbol{\Gamma}_j^{+}$ consists of the columns of $\boldsymbol{\Gamma}$ corresponding to q_j^{+}-elements of the jth column of $\mathbf{P}$.

We also assessed uncertainty in the unknown number of sources and identifiability conditions by computing the marginal likelihood and posterior model probability of each model in the set of candidate models (with different q and prespecification of zeros in $\mathbf{P}$).

The marginal likelihood of model M can be estimated by

$$\hat{l}(\mathbf{X}|M) = \frac{l\left(X\left|\theta^c, M\right.\right) p\left(\theta^c|M\right)}{\hat{\pi}\left(\theta^c|\mathbf{X}, M\right)},$$

where θ^c is a single point of $\theta = (\mathbf{G}, \mathbf{P}, \Sigma, \Omega, \mu)$ under model M and $\hat{\pi}\left(\theta^c|\mathbf{X}, M\right)$ is the estimated posterior density function of $\pi\left(\theta^c|\mathbf{X}, M\right)$. Using the same algorithm of Oh (1999), we have:

$$\begin{aligned}\pi\left(\theta^c \mid \mathbf{X}, M\right) &= E\Big[\pi\left(\mathbf{G}^c \mid \mathbf{P}^c, \Sigma^c, \Omega^c, \mu^c\right) \\ &\quad \times \pi\left(\mathbf{P}^c \mid \mathbf{G}, \Sigma^c, \Omega^c, \mu^c\right) \times \pi\left(\Sigma^c \mid \mathbf{G}, \mathbf{P}, \Omega^c, \mu^c\right) \\ &\quad \times \pi\left(\Omega^c \mid \mathbf{G}, \mathbf{P}, \Sigma, \mu^c\right) \times \pi\left(\mu^c \mid \mathbf{G}, \mathbf{P}, \Sigma, \Omega\right)\Big] \\ &= \pi\left(\mathbf{G}^c \mid \mathbf{P}^c, \Sigma^c, \Omega^c, \mu^c\right) \times E\Big[\pi\left(\mathbf{P}^c \mid \mathbf{G}, \Sigma^c, \Omega^c, \mu^c\right) \\ &\quad \times \pi\left(\Sigma^c \mid \mathbf{G}, \mathbf{P}, \mu^c\right) \times \pi\left(\Omega^c \mid \mathbf{G}\right) \times \pi\left(\mu^c \mid \mathbf{G}, \mathbf{P}, \Sigma\right)\Big].\end{aligned}$$

Because the full conditional posterior density functions are known, $\pi\left(\hat{\theta}^c|\mathbf{X}, M\right)$ can be estimated as the sample average of the product of the full conditional posterior density functions using the posterior sample of θ under model M. We chose θ^c as an approximate posterior model based on a preliminary MCMC run. Note that when $\Omega = \mathbf{I_q}$ is assumed a priori to eliminate the multiplication of a row by a scale constant in $\mathbf{P}$, the step involving Ω in sample generation and marginal likelihood computation can be omitted.

Evaluation of Spatially-Enhanced Bayesian Multivariate Receptor Models by Simulation

We conducted a simulation study to assess the performance of the new enhanced multivariate receptor models accounting for spatial dependence in the data as well as uncertainty in both the number of sources and identifiability conditions. The spatial locations $(s_1, s_2, \cdots, s_N)$ for the data are generated randomly from a uniform distribution over a unit square with $N = 9$. The spatial locations $\{\omega_1, \omega_2, \cdots, \omega_L\}$ where the latent processes are generated were chosen so that the distance between adjacent locations (that is equal to standard deviation of the kernel if the Gaussian convolution kernel is used) is $\sigma_\kappa = 0.3$, which results in $L = 9$. The true source-composition matrix $\mathbf{P}$ and mean μ are set as:

$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.05 & 0.25 & 0.1 & 0 & 0 & 0.3 & 0.1 & 0.1 \\ 0 & 0.4 & 0 & 0.1 & 0.1 & 0.05 & 0.1 & 0.05 & 0.2 \\ 0.1 & 0 & 0.05 & 0 & 0.1 & 0.4 & 0.05 & 0.2 & 0.1 \end{bmatrix}$$

and $\mu = \xi\mathbf{P}$ where $\xi = (4, 6, 10)$, respectively.

The latent space–time process $\mathbf{G_t}$ and error process $\mathbf{E_t}$ were generated from the following matrix normal distributions:

$$\mathbf{G_t} \sim N(\mathbf{0}, \mathbf{I_9}, \Omega), \quad \Omega = \begin{bmatrix} 1 & 0.1 & 0.5 \\ 0.1 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix},$$

$$\mathbf{E_t} \sim N(\mathbf{0}, \mathbf{I_9}, \Sigma), \text{ and } \Sigma = diag(0.03, 0.02, 0.03, 0.02, 0.01, 0.04, 0.02, 0.03, 0.03).$$

The data were generated by

$$\mathbf{X_t} = \mathbf{K}\mathbf{G_t}\mathbf{P} + \boldsymbol{\mu} + \mathbf{E_t}, \text{ with}$$

$$\mathbf{K} = \begin{bmatrix} \exp\left(-\|\omega_1 - s_1\|^2 \Big/ \left(2\sigma_\kappa^2\right)\right) & \cdots & \exp\left(-\|\omega_L - s_1\|^2 \Big/ \left(2\sigma_\kappa^2\right)\right) \\ \exp\left(-\|\omega_1 - s_2\|^2 \Big/ \left(2\sigma_\kappa^2\right)\right) & \cdots & \exp\left(-\|\omega_L - s_2\|^2 \Big/ \left(2\sigma_\kappa^2\right)\right) \\ \vdots & \vdots & \vdots \\ \exp\left(-\|\omega_1 - s_N\|^2 \Big/ \left(2\sigma_\kappa^2\right)\right) & \cdots & \exp\left(-\|\omega_L - s_N\|^2 \Big/ \left(2\sigma_\kappa^2\right)\right) \end{bmatrix}_{N \times L}, \quad \text{where } \sigma_\kappa = 0.3$$

and $\boldsymbol{\mu} = 1_N \otimes \mu$.

To assess the prediction performance of the spatial multivariate receptor models for the source contributions at an unmonitored site, we used the data from eight locations for model fitting and computing marginal likelihood, and then predicted source contributions at the ninth location under the true model for model validation. The sample size T at each site was taken to be 100.

As opposed to assuming the known number of sources ($q_0 = 3$) and identifiability conditions, we defined the candidate models by varying the number of sources ($q = 2, 3, 4, 5$) and position of zeros in identifiability conditions (C1–C3) and estimated the parameters under each model as well as computed marginal likelihoods.

The simulation was repeated 50 times. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,9$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $r_0 = q$, and $R_0 = \mathbf{I_q} \times (r_0 + q + 1)$. The estimated marginal likelihoods for each q-source model are reported in Table 6 on a log scale (only 5 cases are shown for illustration). The selected model is the one having the maximum logMD. The true model (with $q = 3$) was selected for all 50 of the simulations.

We also monitored the sample correlations among the true source-composition profiles and the estimated source-composition profiles as well as among the true source contributions and the estimated source contributions for $q = 3$. Throughout the simulation, the correlations for source-composition profiles were all greater than 0.92, which indicates that the estimated source profiles agree well with the true source profiles. The sample correlations for source contributions at monitored sites were greater than 0.91. Figure 5 presents the time-series plots of the true centered source contributions and the estimated centered source contributions (based on one of the simulated datasets) at one of the monitored sites (site 1). Note that estimates for noncentered source contributions can be obtained by adding the estimated mean contribution $\hat{\xi} = \hat{\mu}\hat{\mathbf{P}}\left(\hat{\mathbf{P}}\hat{\mathbf{P}}'\right)^{-1}$ to the estimated centered source contributions. The mean squared errors for the source compositions and contributions under the true model were less than 0.2 and 0.4, respectively.

Next, we predicted the source contribution at an unmonitored site (site 9). Figure 6 contains the time-series plots of the true centered source contributions and predicted centered source contributions (based on one of the simulated datasets) at the unmonitored site. The sample correlations for source contributions at the unmonitored site were greater than 0.91, and the mean square error of predicted source contributions at the unmonitored site under the true model was less than 0.5.

Then, we performed a simulation assuming $\Omega = \mathbf{I_q}$ a priori to eliminate the multiplication of a row by a scale constant in $\mathbf{P}$. The simulation was repeated 50 times. The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,9$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, and $M_0 = 100 \times \mathbf{I_J}$. As noted in an earlier section, assuming an orthogonal factor model for γ (with $\Omega = \mathbf{I_q}$) is more beneficial than assuming an oblique factor model (unless $\mathbf{P}$ can be assumed to have a special structure such as $\mathbf{I_q}$ for the leading $q \times q$ submatrix of $\mathbf{P}$ for identifiably conditions) in terms of more easily satisfying the constraint C2 and avoiding numerical problems in MCMC implementation. The estimated marginal likelihoods for

Table 6. Log of Marginal Likelihood (logMD) for q-Source Models Assuming a priori an Oblique Factor Model ($q = 2, 3, 4, 5$)

	Number of Sources (q)[a]			
Dataset	2	3	4	5
1	−1292.6	−802.4	−1201.1	−1301.2
2	−1668.4	−875.6	−1337.4	−1310.8
3	−1447.4	−980.8	−1414.7	−1413.1
4	−1624.8	−714.9	−1242.2	−1396.8
5	−1686.9	−690.7	−1634.1	−1112.8

[a] The model having the maximum logMD was selected for each dataset. $q = 3$ was selected for all 5 of these datasets.

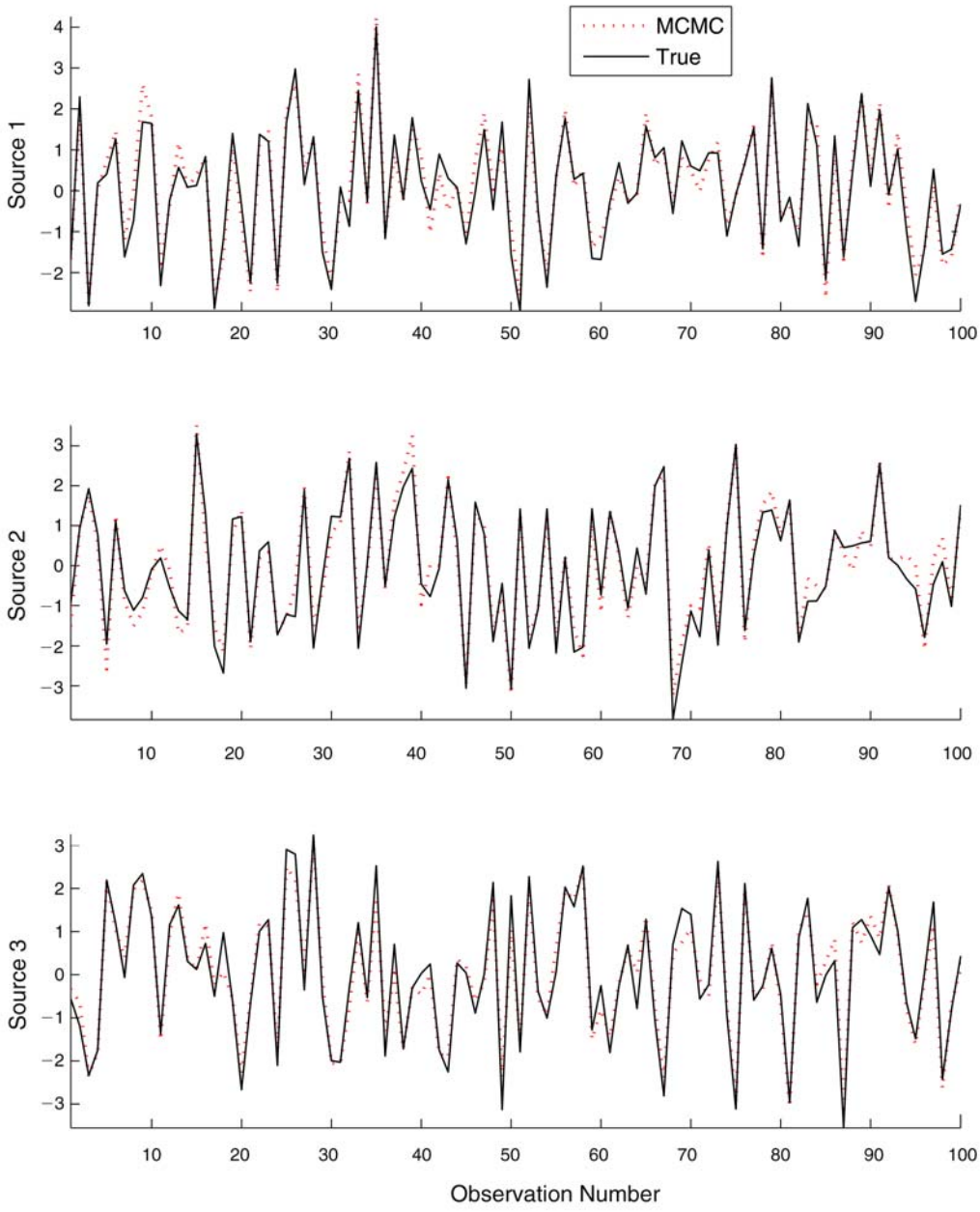


Figure 5. True and estimated centered source contributions at one of the monitoring sites in a simulated dataset assuming a priori an oblique factor model.

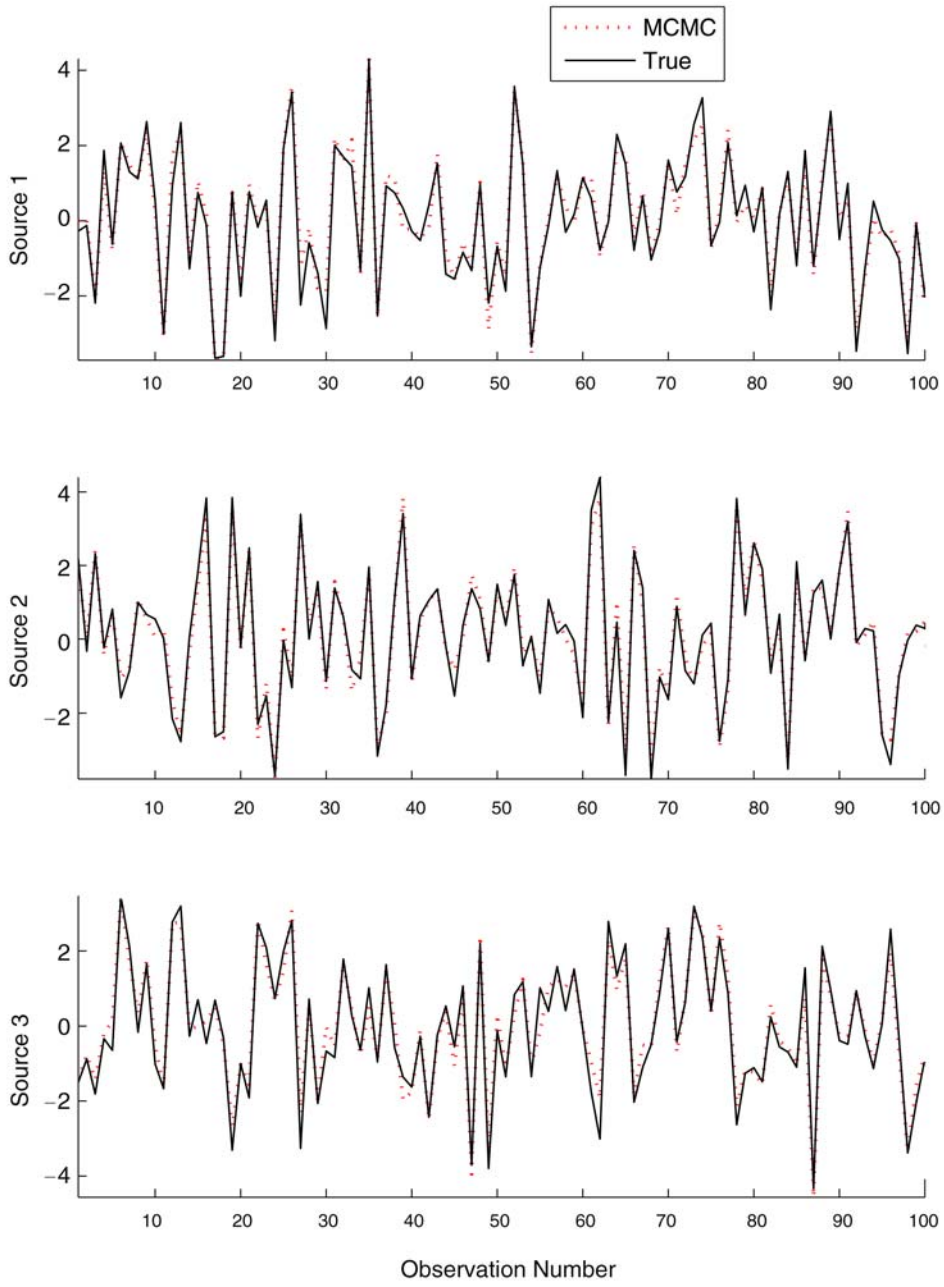


Figure 6. True and predicted centered source contributions at an unmonitored site in a simulated dataset assuming a priori an oblique factor model.

each q-source model are reported in Table 7 on a log scale (only 5 cases are shown for illustration). The selected model is the one having the maximum logMD. The true model (with $q = 3$) was selected for all 50 of the simulations.

We also monitored the sample correlations among the true source-composition profiles and the estimated source-composition profiles as well as among the true source contributions and the estimated source contributions for $q = 3$. Throughout the simulation, the correlations for source-composition profiles were all greater than 0.93, which indicates that the estimated source profiles agree well with the true source profiles. The sample correlations for source contributions at monitored sites were greater than 0.91. Figure 7 presents the time-series plots of the true centered source contributions and the estimated centered source contributions (based on one of the simulated datasets) at one of the monitored sites (site 1). Note that estimates for noncentered source contributions can be obtained by adding the estimated mean contribution $\hat{\xi} = \hat{\mu}\hat{\mathbf{P}}\left(\hat{\mathbf{P}}\hat{\mathbf{P}}'\right)^{-1}$ to the estimated centered source contributions. The mean squared errors for the source compositions and contributions under the true model were less than 0.05 and 0.2, respectively.

Next, we predicted the source contribution at an unmonitored site (site 9). Figure 8 contains the time-series plots of the true centered source contributions and predicted centered source contributions (based on one of the simulated datasets) at the site. The sample correlations for source contributions at the unmonitored site were greater than 0.92, and the mean square error of predicted source contributions at the site under the true model was less than 0.3.

RESULTS

PHOENIX DATA ANALYSIS

The method developed in the previous section has been applied to the Phoenix $PM_{2.5}$ speciation data along with temperature and relative humidity data and daily cardiovascular mortality data (Hopke et al. 2006; Mar et al. 2006). The same data were used in the 2003 workshop on the Source Apportionment of PM Health Effects, sponsored by the U.S. EPA PM centers (Hopke et al. 2006; Mar et al. 2006; Thurston et al. 2005).

The daily 24-hour $PM_{2.5}$ speciation data for Phoenix were obtained from Dr. Philip Hopke. The original $PM_{2.5}$ data consisted of 981 samples, collected over a 1208-day period (3/11/1995–6/30/1998), with measured concentrations for 46 chemical elements: Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Rb, Sr, Y, Zr, Mo, Rh, Pd, Ag, Cd, Sn, Sb, Te, I, Cs, Ba, La, W, Au, Hg, Pb, organic carbon (OC), and elemental carbon (EC). We also received the Phoenix mortality data (i.e., daily numbers of deaths due to cardiovascular causes and due to all nonaccidental causes) for residents $\geq$ 65 years of age at the time of death along with the corresponding weather data (temperature and relative humidity) from Dr. Teresa Mar. These data were collected over a 1057-day period (2/9/1995–12/31/1997). The overlap in collection dates for the Phoenix $PM_{2.5}$ speciation data and the mortality data is a 1027-day period (3/11/1995–12/31/1997).

The original Phoenix $PM_{2.5}$ speciation data with 981 observations on 46 chemical species contains many negative values. For some species, more than half of the measurements have negative values. Species with more than 200 negative values are: Sc (662), V (281), Cr (305), Co (513), Ni (203), Ga (552), Ge (450), Se (226), Y (601), Zr (527),

Table 7. Log of Marginal Likelihood (logMD) for q-Source Models Assuming a priori an Orthogonal Factor Model ($q = 2, 3, 4, 5$)

	Number of Sources (q)[a]			
Dataset	2	3	4	5
1	748.6	1223.6	1160.5	1137.4
2	393.0	1330.1	656.7	884.4
3	708.0	1229.7	1187.6	1158.3
4	639.0	1144.3	1102.7	1051.7
5	675.8	1205.8	1148.2	1123.8

[a] The model having the maximum logMD was selected for each dataset. $q = 3$ was selected for all 5 of these datasets.

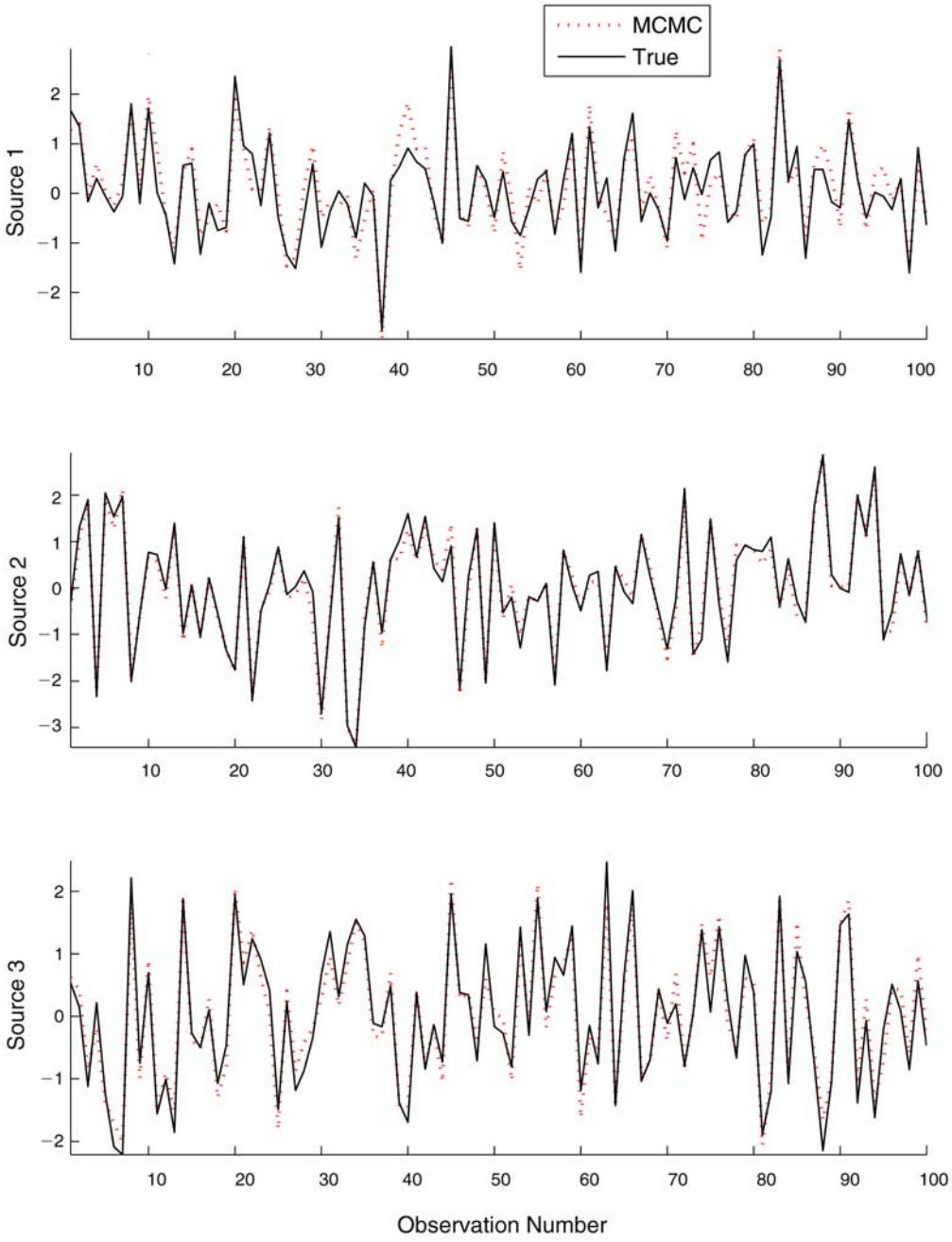


Figure 7. True and estimated centered source contributions at one of the monitoring sites in a simulated dataset assuming a priori an orthogonal factor model.

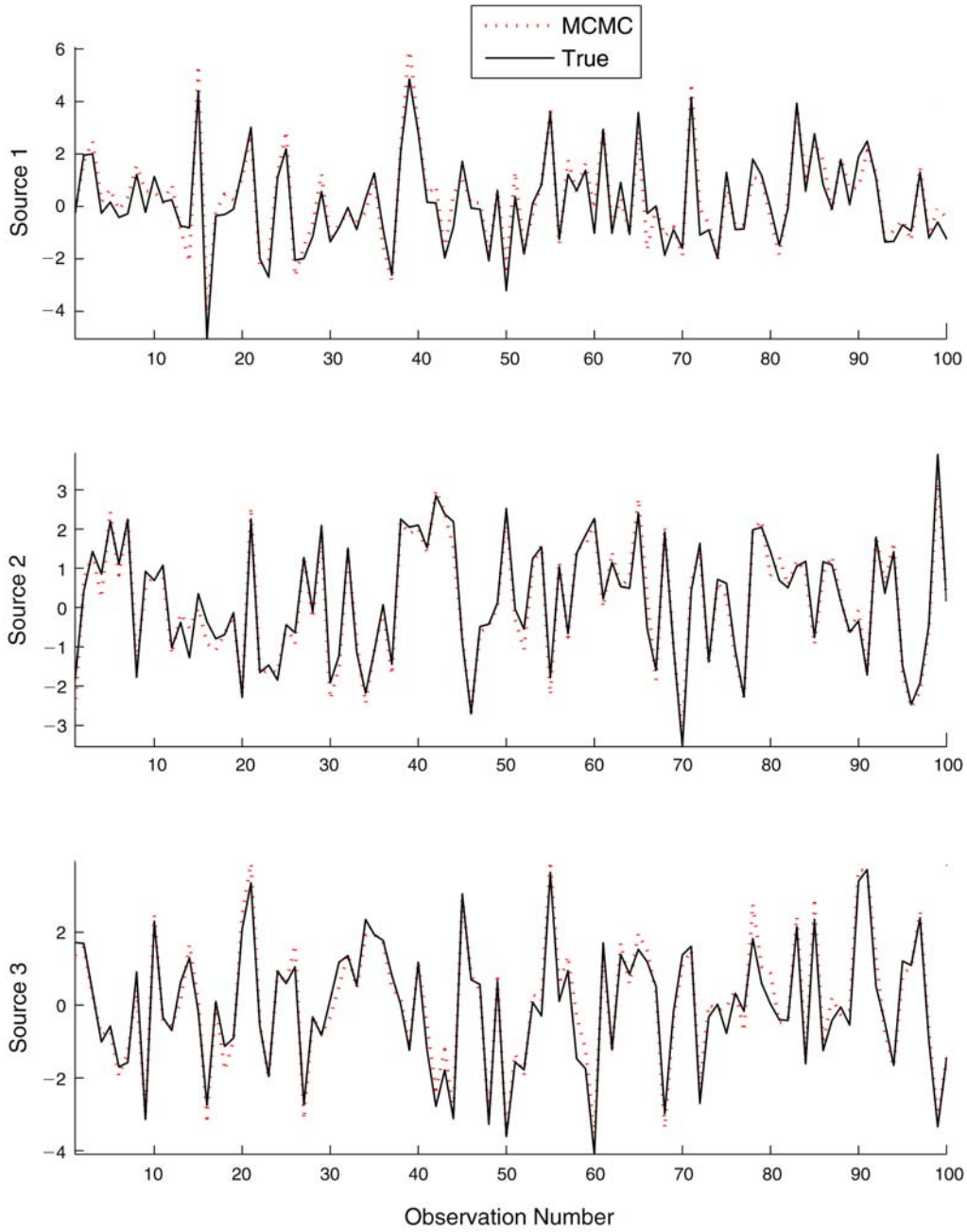


Figure 8. True and predicted centered source contributions at an unmonitored site in a simulated dataset assuming a priori an orthogonal factor model.

Mo (486), Rh (596), Pd (561), Ag (539), Cd (470), Sn (349), Sb (349), Te (446), I (246), Cs (277), La (257), W (515), Au (517), and Hg (524) where the number in parentheses represents the number of negative measurements. Another problem with the $PM_{2.5}$ data was that some species were missing a large number of values (e.g., Na and Mg). The pattern of the missing values in the original Phoenix data is given in Appendix D (available on the HEI Web site).

The first important step in multivariate receptor modeling is to select an appropriate subset of species for the analysis. Prior knowledge about the potential sources for the region of the study (e.g., Lewis et al. 2003; Ramadan et al. 2003), as well as an expert's opinion (Dr. Hopke), was utilized in the initial selection of species for fitting. None of the species listed above (with > 200 negative values) were deemed to be associated with major pollution sources.

It needs to be emphasized that although we removed species that had a large number of negative or missing values, it does not mean that the effect of sources containing the removed species cannot be incorporated into examining source-specific health effects. For example, copper smelters are one of the main sources of Co and although Co is not selected as one of the fitting species in this case, copper smelters may still be identified by other species in the data; their contributions, as well as their health effects, can still be estimated. This is one of the major advantages of a source-specific approach over a pollutant-specific approach in assessing associations between air pollutants and health outcomes.

In a pollutant-specific approach, once Co is removed from the data, the effect of Co on a health outcome (e.g., respiratory disease) cannot be assessed. On the other hand, in a source-specific approach, we are concerned about the effect of a combination of species (realized as source contributions), and even the effect of missing species can be incorporated into the health-effects analysis of sources emitting that species. Note that individual species affect only the estimation of source-composition profiles. The source contributions are the same for all species constituting source-composition profiles and do not depend on individual species.

The only exception to incorporating the effect of a removed species into the analysis is when the source of that species cannot be identified (and so the corresponding contributions cannot be estimated), that is, when the removed species is a tracer element or a key species for a non-negligible source. In such cases, although the species have many missing values, they may still need to be retained in the data (with help of a good imputation procedure) rather than be deleted.

As a matter of fact, we had such an incidence with Na. Although Na is one of the key species for Sea Salt identified in previous studies for the region along with Cl, about 75% of Na values were missing. Without Na, however, Sea Salt did not appear to be able to be successfully identified in our analyses. Thus, we finally decided to include Na (after imputation) in the fitting species.

The final selection of species used in the model included the following 15 species:

Na (645 missing), Al (19), Si (0), S (0), Cl (60), K (0), Ca (0), Mn (1), Fe (0), Cu (6), Zn (0), Br (2), Pb (6), OC (28 missing), and EC (28 missing),

where the number in parentheses represents the number of negative measurements or missing values in the data for 868 out of 1027 days for which $PM_{2.5}$ speciation measurements existed.

There were no missing values for the mortality data for this 1027-day period. On the other hand, the $PM_{2.5}$ data were available for only 868 of these days at most as mentioned above (223 days for Na, 840 days for OC and EC, 868 days for the remaining 12 species). The weather data were also missing for some of 1027 days (temperature was missing for 56 days and relative humidity was missing for 76 days). Although one option for evaluating the source-specific health effects would be to include only the 868 days that are common for all of $PM_{2.5}$ data, weather data, and the mortality data, it would not allow for exploration of a different lag structure of the source-specific effects on mortality other than a lag 0 effect. Thus, we decided to impute the missing values (in $PM_{2.5}$ data and weather data) prior to applying the new method so that they could be aligned with the mortality data for all 1027 days. Appendix D contains the plots showing the structure of missing values in two species (OC and K) when all 1027 days are considered. Other species also show a similar structure because most of the missing values come from the days when no measurements on $PM_{2.5}$ species were available.

The SPSS procedure based on the EM algorithm was used for imputation of missing values. The EM algorithm uses a linear model for the variables that are present to estimate the missing data in an iterative manner. It uses maximum likelihood estimates for the missing data (M step), then plugs those estimates (E step) into the data and iterates until convergence (see Dempster et al. 1977). In cases where all the $PM_{2.5}$ species are missing, we still had temperature and relative humidity to use for imputations. We used a *t*-distribution with 7 degrees of freedom (*df*) to capture the heavy-tailed (outlier prone) distribution of our data. Thus, the EM method uses a *t*-distribution with 7 *df* as the likelihood to maximize in the M step of the EM algorithm.

To provide a consistent basis for comparison with other results from Mar and colleagues (2006), we used the same mortality model as they did, controlling for confounding by including an indicator variable for extreme temperatures, a day-of-week variable, and smoothing terms for time trends, temperature, and relative humidity (namely, natural spline smoothers with 12 *df* for the smoothing of time trend, 5 *df* for the smoothing of temperature with 2 days lag, and 2 *df* for the smoothing of relative humidity with 0 days lag).

We constructed a range of different receptor models (resulting from each combination of numbers of sources (q) and identifiability conditions) to be compared with each other for the Phoenix data. Based on several previous studies on the Phoenix $PM_{2.5}$ data (Hopke et al. 2006; Lewis et al. 2003; Ramadan et al. 2000; Ramadan et al. 2003) and the NUMFACT procedure (Henry et al. 1999; Park et al. 2000), we presumed that the number of major sources is between three and eight. For candidate positions of zeros in **P** under each q-source model, we also used the information on the major sources, such as the information on minor species (absent or low in concentration) for each potential source type, obtained from previous studies on Phoenix $PM_{2.5}$. For example, we could use the information that the element Al is typically not present in emissions from Motor Vehicles (Traffic) and then prespecify zero to Al. In practice, the elements that are preassigned zero elements are rarely actual zeros but are small enough to be considered zero (i.e., minor compounds). We preassign zeros to absent species for each source profile to indicate that the true concentrations for the corresponding elements are relatively small, if not zeros. For example, in Table 8, Source 1 under Model 1 has Al and S absent from its profile; in practice it could be any source profile (such as Biomass Burning, Traffic, or Sea Salt) for which those species are not major constituents. Source 2 with Cl and Fe absent from the profile (or low in concentration) may represent Smelter, Biomass Burning, or Secondary Sulfate), and Source 3 with OC and EC absent (or low) in the profile could be a profile not having those species as major constituents (e.g., Soil or Sea Salt). Note that we use this type of information from previous studies only to find out the plausible sets of identifiability conditions (positions of zeros) under each q-source model. Other than that, the candidate models do not depend on the results from those previous studies. The ten candidate models for comparisons with different numbers of sources (q = 3, 4, 5, 6, 7, 8) and different prespecification of identifiability conditions (zeros in **P**) are listed in Table 8.

The $PM_{2.5}$ data and cardiovascular mortality data were simultaneously fitted to estimate source-composition profiles, contributions, and health-effects parameters as well as marginal likelihood under each model in Table 8 at lag 0–5 days. Because concentrations of $PM_{2.5}$ species differed by two or three orders of magnitude, each element was

Table 8. Candidate Models for Phoenix $PM_{2.5}$ Speciation Data

Model Number	q	Source	Prespecified Position of Zeros in P
1	3	1	Al, S
		2	Cl, Fe
		3	OC, EC
2	4	1	Al, Si, K
		2	Al, Cl, Fe
		3	Cl, OC, EC
		4	Al, Si, Ca
3	5	1	Al, Si, S, K
		2	Al, Si, Cl, Fe
		3	Cl, Cu, OC, EC
		4	K, Ca, Br, EC
		5	Al, Si, OC, EC
4	5	1	Al, Si, S, K
		2	Al, Si, Cl, Fe
		3	Cl, Cu, OC, EC
		4	Al, Si, Ca, Fe
		5	K, Ca, Br, EC
5	5	1	Al, S, Cl, Fe
		2	Cl, Fe, Cu, Zn
		3	Cl, Cu, OC, EC
		4	Al, Si, Ca, Cu
		5	Al, Si, K, Ca
6	6	1	Al, Si, S, Cl, Ca
		2	Al, Si, Cl, K, Fe
		3	Cl, Cu, Pb, OC, EC
		4	Al, Si, Cl, Ca, Fe
		5	Al, Si, Cl, K, Ca
		6	Al, Cl, K, Ca, EC
7	6	1	Al, Si, S, Cl, K
		2	Cl, Ca, Mn, Br, EC
		3	Cl, Cu, Pb, OC, EC
		4	Al, Si, Cl, Ca, Fe
		5	Cl, Fe, Cu, Zn, Pb
		6	Al, K, Pb, OC, EC
8	6	1	Al, Si, S, Cl, K
		2	Al, Si, Cl, K, Fe
		3	Cl, Cu, Pb, OC, EC
		4	Al, Si, Cl, Ca, Fe
		5	Al, Cl, Mn, Br, EC
		6	Al, K, Pb, OC, EC
9	7	1	Al, Si, S, Cl, K, Fe
		2	Al, Cl, Ca, Mn, Br, EC
		3	Cl, Cu, Br, Pb, OC, EC
		4	Al, Si, Cl, Ca, Fe, Zn
		5	Na, Al, Si, Cl, K, Ca
		6	Na, Cl, Fe, Cu, Zn, Pb
		7	Al, K, Cu, Pb, OC, EC
10	8	1	Al, S, Ca, Mn, Zn, Br, Pb
		2	Al, Cl, Mn, Fe, Cu, Zn, Pb
		3	Cl, Mn, Cu, Br, Pb, OC, EC
		4	Al, Si, Cl, Ca, Mn, Fe, Zn
		5	Al, Si, Cl, K, Ca, Mn, Br
		6	Al, Cl, Mn, Zn, Br, Pb, EC
		7	Al, Cu, Zn, Br, Pb, OC, EC
		8	S, K, Fe, Cu, Pb, OC, EC

scaled by its sample standard deviation before running MCMC. It is known that convergence problems are common when elemental concentrations are on widely different scales (Nikolov et al. 2007).

However, after the run, the individual elements of the estimated source profiles were multiplied by the corresponding sample standard deviations to bring them back to the original scale, so that the relative amounts of species in each profile are physically interpretable. The estimated source contributions were also rescaled by multiplying the sum of elements of the corresponding source-composition profile in the original scale. It needs to be noted that although the PM data were scaled by the standard deviations at the beginning, this does not actually affect the estimation of **P** or **A**. It only changes the scales in the source-composition matrix during the MCMC implementation. By rescaling back to the original scale at the end, however, the relative amounts of species in each source profile are preserved.

The following hyperparameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,15$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $\alpha_0 = 0$, $U_0 = 100$, $\beta_0 = 0_q$, $B_0 = 100 \times \mathbf{I_q}$, $\eta_0 = 0_I$, $\Psi_0 = 100 \times \mathbf{I_I}$, $a_0^y = 0.01$, and $b_0^y = 0.01$. Also, an orthogonal factor model assuming $\Omega = \mathbf{I_q}$ a priori was employed for γ_t. Note that from a Bayesian standpoint, Ω can be viewed as a hyperparameter of the prior distribution for γ, and as shown in Park et al. (2002b) the correlation structure in γ can still be uncovered by the sample correlations of the estimated γ's even in the case that $\Omega = \mathbf{I_q}$ is misspecified a priori.

For each model, an approximate posterior mode is obtained from a preliminary MCMC run, and this is used for $\theta^c = \left(\mu^c, \Gamma^c, \mathbf{P}^c, \Sigma^c, \alpha^c, \beta^c, \eta^c, \sigma_y^{2c}\right)$, at which the marginal likelihood is calculated. An approximate posterior mode is obtained by evaluating the joint posterior density for 100,000 iterations after the first 100,000 draws are discarded. A main MCMC run is then started from $\theta^c = \left(\mu^c, \Gamma^c, \mathbf{P}^c, \Sigma^c, \alpha^c, \beta^c, \eta^c, \sigma_y^{2c}\right)$, and the samples are collected for 200,000 iterations, subsampling every 10th value (resulting in 20,000 samples), without additional burn-in (the practice of discarding some iterations at the beginning of an MCMC run). The marginal likelihood for each model can be computed in sample generation without storing the samples. Table 9 gives the estimated marginal likelihoods (in log) for each model of Table 8, jointly modeling the $PM_{2.5}$ and cardiovascular mortality data at lag day 0. The posterior probability of each model under the indifference prior is also provided in Table 9. Model 7 with six sources is selected as the best model because the posterior probability for Model 7 is almost 1. For other lag days (lag days 1–5) Model 7 also led to the highest posterior model probability that is very close to 1. This is consistent with the observation from Mar and colleagues (2006), who noted that there are six source components most consistently reported for the Phoenix data by the various investigators and methods.

Table 9. Marginal Likelihoods for Candidate Models for Phoenix $PM_{2.5}$ Speciation Data and Cardiovascular Mortality at Lag 0 Days[a]

Model Number	Number of Sources (q)	LogMD ($\times 10^4$)	PostP
1	3	−1.5761	0.0000
2	4	−1.5580	0.0000
3	5	−1.5598	0.0000
4	5	−1.5219	0.0000
5	5	−1.5549	0.0000
6	6	−1.5392	0.0000
7	6	−1.5153	1.0000
8	6	−1.5316	0.0000
9	7	−1.5440	0.0000
10	8	−1.5849	0.0000

[a] LogMD and PostP denote the Log of Marginal Likelihood and Posterior Model Probability, respectively.

The estimated source profiles under Model 7 are given in Table 10. The estimated source profiles and contributions based on the $PM_{2.5}$ and cardiovascular mortality data with other lags are materially the same as those in Table 10. Major species in the estimated source profiles of Table 10 are consistent with main elements of major $PM_{2.5}$ sources for Phoenix identified by several previous studies (Hopke et al. 2006; Lewis et al. 2003; Ramadan et al. 2000; Ramadan et al. 2003). For example, Source 3, which is characterized by Al, Si, K, Ca, and Fe, and Source 4, which is characterized by high percentages of OC and EC, and some K appear to correspond to Soil/Crustal Matter and Biomass/Wood Combustion, respectively. Source 6 is characterized by high percentages of Na and Cl and likely represents Aged Sea Salt. Source 2 and Source 5 are both characterized by high percentages of S and OC. Source 2 is also characterized by Cu, Zn, and Pb, while percentages of those elements are forced to be zeros in the Source 5 profile. It is conjectured that Source 2 represents a Smelter and Source 5 represents Secondary Sulfate.

Source 1, characterized by high percentages of OC and EC, seems to represent Traffic, as defined by Mar and colleagues (2006). The Heavy Duty Diesel source that had been identified in some of the previous studies (Lewis et al. 2003; Ramadan et al. 2000; Ramadan et al. 2003) could not be separated from Motor Vehicles. The non-negligible

Table 10. Estimated Source Composition Profiles of 15 Species Under Model 7[a]

Species	Source 1 *Traffic*	Source 2 *Smelter*	Source 3 *Soil/Crustal*	Source 4 *Biomass / Wood Combustion*	Source 5 *Secondary Sulfate*	Source 6 *Sea Salt*
Source Composition						
Na	0.00	0.04	1.11	0.03	7.81	34.90
Al	**0**	0.23	17.04	**0**	0.39	**0**
Si	**0**	2.58	44.61	**0**	0.45	8.58
S	**0**	64.33	1.85	0.20	28.82	1.36
Cl	**0**	**0**	**0**	**0**	**0**	41.36
K	**0**	1.58	6.70	3.15	2.34	**0**
Ca	0.53	**0**	15.25	**0**	0.13	6.32
Mn	0.11	**0**	0.34	0.00	0.01	0.32
Fe	2.61	0.87	12.73	**0**	**0**	5.62
Cu	0.11	1.14	**0**	0.04	**0**	0.06
Zn	0.38	1.46	0.29	0.05	**0**	1.26
Br	0.04	**0**	0.07	0.04	0.14	0.20
Pb	0.16	1.73	**0**	0.13	**0**	**0**
OC	67.51	26.05	**0**	76.54	58.35	**0**
EC	28.55	**0**	**0**	19.82	1.57	**0**

[a] Source compositions are normalized to Sum = 100%. **Bolding** gives the position of pre-assigned zeros. Source names in italics are conjectures based on the source compositions of previous studies.

percentage of Fe in the Source 1 profile seems to imply that Source 1 might be a mixture of Motor Vehicles and Heavy Duty Diesel. In their intercomparison of source apportionment studies, Hopke and colleagues (2006) reported that some investigators identified and quantified Diesel separately from Motor Vehicles, and some did not. As a result, in the study by Mar and colleagues (2006) that assessed the health effects of $PM_{2.5}$ sources based on the source-apportionment results of Hopke and colleagues (2006), Diesel and Motor Vehicles were combined and named Traffic.

To obtain the corresponding source contributions that are scaled appropriately by the normalizing constants of the source profiles, S and OC in Table 10 needed to be rescaled because all of the S would be present as sulfate, which has three times the mass of S. OC includes only the carbon in organic compounds and does not include the unmeasured H, O, and N that will also be in the organic species. Since ammonium is not included in the profile, S was multiplied by 4.125 to be converted to $(NH_4)_2SO_4$. Also, OC was multiplied by 1.5 to be converted to organic matter (OM) that includes H, O, and N. The rescaled source-composition profiles along with their uncertainty estimates (95% posterior intervals) and the estimates of the mean, standard deviations, and the 5th-to-95th percentiles of source contributions are presented in Table 11.

Figure 9 contains the time-series plots of the estimated source contributions (in $\mu g/m^3$) for the 1027-day study period along with the 95% posterior intervals. In general, the daily patterns of estimated source contributions of Figure 3 are similar to those of figure 1 in Mar and colleagues (2006) and those of figure 2 in Ramadan and colleagues (2003).

The plots of predicted versus measured concentrations for species used in model fitting, as well as the plot of the sum of estimated source contributions versus measured total $PM_{2.5}$ mass concentration (which was not used in model fitting), are also provided in Appendix D. The R^2 values between the measured and predicted values were greater than 0.7 for all but two minor species (Zn and Br). The R^2 values between the sum of the estimated source contributions and measured total $PM_{2.5}$ mass concentration was 0.93.

Table 12 presents source-specific effects on cardiovascular mortality at lag days 0–5. Only the source-specific effects due to Source 2 (that appears to be Smelter) at lag day 0 and Source 6 (that appears to be Sea Salt) at lag day 5 were statistically significant (i.e., a 95% posterior interval does not contain 0). In Mar and colleagues (2006), the effects of Sulfate (lag 0), Traffic (lag 1), Smelter (lag 0), and Sea Salt (lag 5) on cardiovascular mortality were found to

Table 11. Rescaled Source Composition Profiles and Contributions of 15 Species Under Model 7[a]

Species	Source 1 *Traffic*	Source 2 *Smelter*	Source 3 *Soil/Crustal*	Source 4 *Biomass / Wood Combustion*	Source 5 *Secondary Sulfate*	Source 6 *Sea Salt*
Source Composition: % (95% Posterior Intervals)[b]						
Na	0.00 (0.00–0.01)	0.01 (0.00–0.05)	1.05 (0.76–1.33)	0.02 (0.00–0.08)	3.56 (3.14–4.07)	33.48 (29.23–37.30)
Al	**0**	0.07 (0.00–0.25)	16.11 (14.87–17.28)	**0**	0.18 (0.02–0.37)	**0**
Si	**0**	0.82 (0.33–1.35)	42.17 (38.96–45.20)	**0**	0.20 (0.01–0.57)	8.24 (4.61–11.47)
$(NH_4)_2SO_4$	**0**	84.49 (72.97–95.59)	7.23 (0.75–14.08)	0.60 (0.02–2.00)	54.22 (47.82–61.59)	5.37 (0.15–16.70)
Cl	**0**	**0**	**0**	**0**	**0**	39.68 (34.42–44.13)
K	**0**	0.50 (0.04–1.07)	6.33 (5.68–7.01)	2.27 (2.08–2.47)	1.07 (0.71–1.42)	**0**
Ca	0.40 (0.35–0.44)	**0**	14.42 (13.29–15.49)	**0**	0.06 (0.00–0.18)	6.07 (4.27–7.78)
Mn	0.08 (0.08–0.09)	**0**	0.32 (0.29–0.36)	0.00 (0.00–0.00)	0.00 (0.00–0.01)	0.31 (0.16–0.47)
Fe	1.95 (1.89–2.03)	0.28 (0.04–0.56)	12.03 (11.01–13.01)	**0**	**0**	5.39 (3.41–7.23)
Cu	0.08 (0.08–0.09)	0.36 (0.31–0.42)	**0**	0.03 (0.02–0.04)	**0**	0.06 (0.00–0.17)
Zn	0.28 (0.26–0.31)	0.46 (0.34–0.60)	0.27 (0.15–0.40)	0.04 (0.00–0.08)	**0**	1.21 (0.47–1.96)
Br	0.03 (0.03–0.03)	**0**	0.06 (0.05–0.08)	0.03 (0.02–0.03)	0.07 (0.05–0.08)	0.19 (0.09–0.31)
Pb	0.12 (0.11–0.13)	0.55 (0.47–0.64)	**0**	0.09 (0.07–0.11)	**0**	**0**
OM	75.71 (75.04–76.34)	12.44 (1.19–24.04)	**0**	82.65 (81.14–83.96)	39.93 (32.44–46.10)	**0**
EC	21.34 (20.76–21.96)	**0**	**0**	14.27 (13.19–13.36)	0.71 (0.02–2.18)	**0**
Source Contribution ($\mu g/m^3$)						
Mean	4.37	0.31	0.83	1.92	2.34	0.03
Standard deviation	3.62	0.65	0.64	1.85	0.97	0.10
5th-to-95th increment	11.81	1.92	1.91	4.97	2.97	0.14

[a] Source names in *italics* are conjectures based on the source compositions of previous studies (cited in the text).

[b] Source composition percentages are normalized to Sum = 100%. **Bolding** gives the position of preassigned zeros.

be statistically significant. The effects of the fine particle soil and biomass burning factors were not significant at any lag days in Mar and colleagues (2006) or in our analysis.

Overall, the effects of Smelter, Sea Salt, Soil, and Biomass Burning on cardiovascular mortality seemed to be consistent between our analysis and that of Mar and colleagues (2006). However, Secondary Sulfate at lag day 0 and Traffic at lag day 1 that were statistically significant in Mar and colleagues (2006) were not statistically significant in our analysis. Recall that the uncertainties in the estimated source contributions were not accounted for in the estimation of the health effects parameters in Mar and colleagues (2006), which may have introduced the potential bias as noted in their study. On the other hand, our

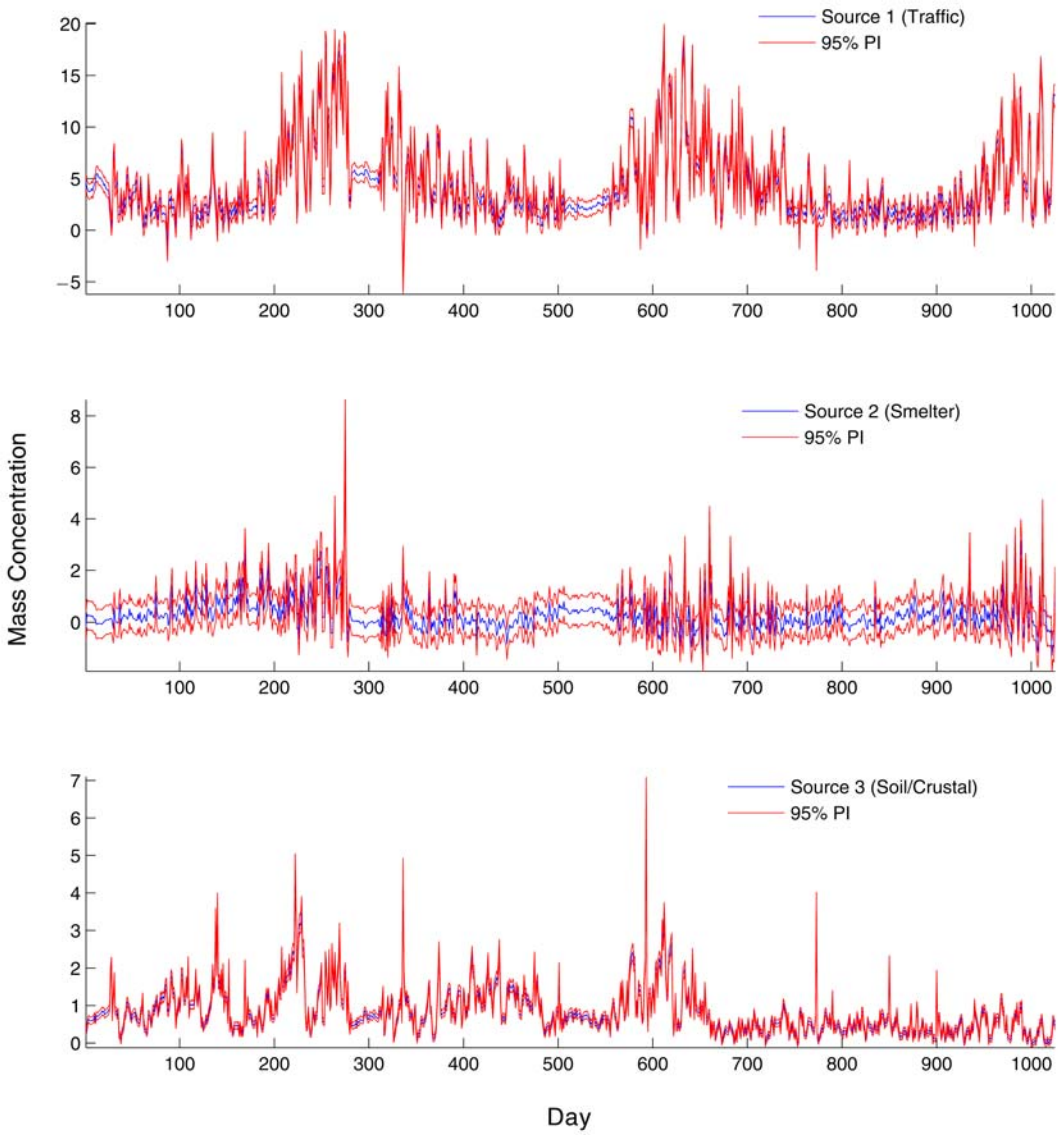


Figure 9. Time-series plots of the estimated source contributions (in μg/m^3) for 1027 days along with their uncertainty estimates (95% posterior intervals). (*Continues next page.*)

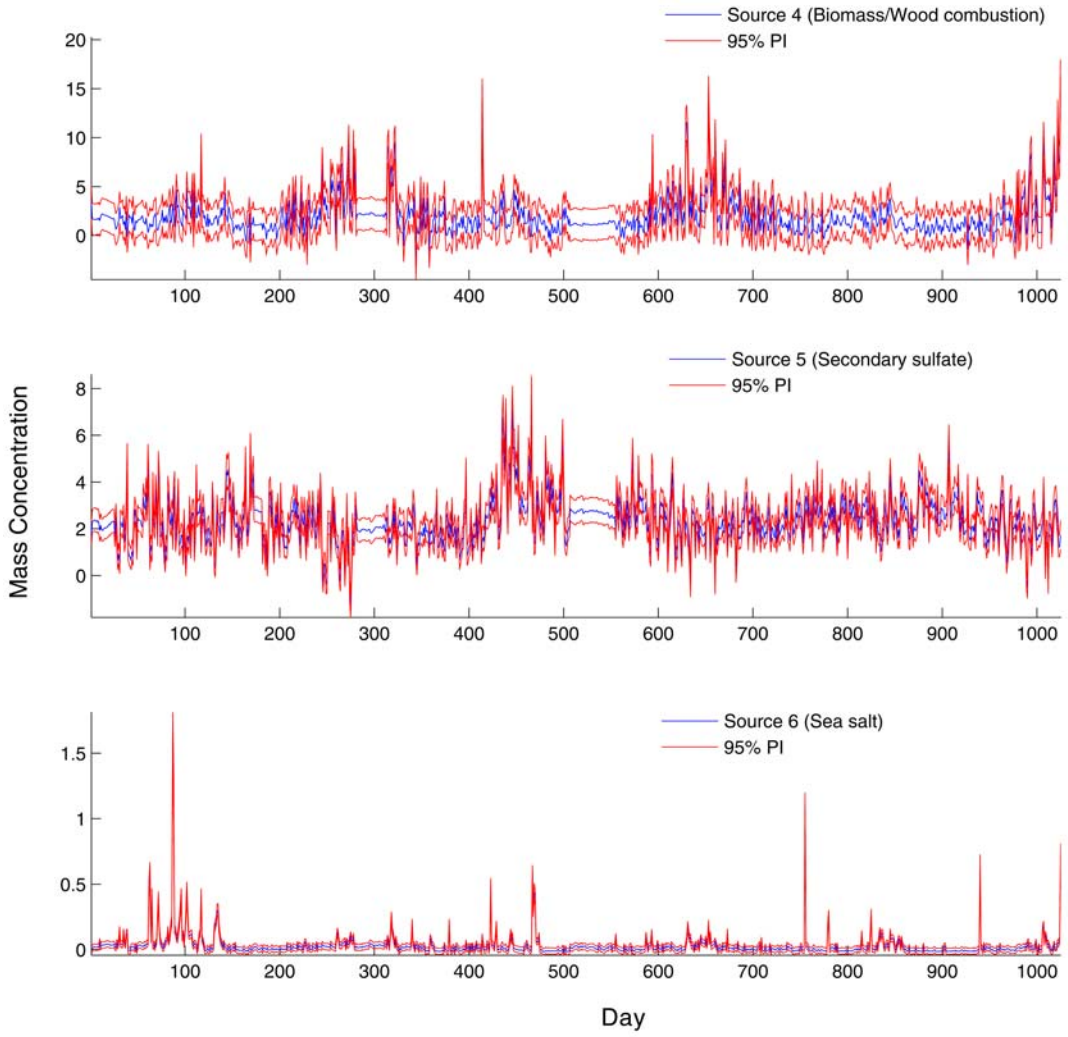


Figure 9 (*Continued*).

Table 12. Source-Specific Effects on Cardiovascular Mortality[a]

	Source 1 *Traffic*	Source 2 *Smelter*	Source 3 *Soil/Crustal*	Source 4 *Biomass / Wood Combustion*	Source 5 *Secondary Sulfate*	Source 6 *Sea Salt*
β (lag 0)	−0.37 (−0.99 to 0.26)	**0.46** (0.02 to 0.91)	0.08 (−0.48 to 0.65)	0.13 (−0.28 to 0.55)	0.28 (−0.19 to 0.74)	−0.11 (−0.31 to 0.09)
β (lag 1)	0.29 (−0.34 to 0.88)	−0.03 (−0.47 to 0.42)	−0.16 (−0.71 to 0.38)	0.24 (−0.18 to 0.66)	0.07 (−0.39 to 0.54)	−0.14 (−0.34 to 0.05)
β (lag 2)	0.00 (−0.64 to 0.60)	0.08 (−0.42 to 0.47)	−0.23 (−0.75 to 0.29)	0.19 (−0.20 to 0.59)	0.08 (−0.36 to 0.70)	−0.03 (−0.25 to 0.18)
β (lag 3)	0.08 (−0.54 to 0.69)	−0.03 (−0.48 to 0.41)	−0.34 (−0.86 to 0.17)	0.27 (−0.13 to 0.67)	0.16 (−0.28 to 0.61)	0.17 (−0.03 to 0.38)
β (lag 4)	0.31 (−0.32 to 0.92)	0.15 (−0.29 to 0.59)	−0.45 (−0.96 to 0.07)	−0.21 (−0.62 to 0.20)	0.45 (−0.00 to 0.90)	0.12 (−0.08 to 0.32)
β (lag 5)	−0.25 (−0.87 to 0.37)	0.01 (−0.42 to 0.43)	−0.03 (−0.55 to 0.48)	0.10 (−0.29 to 0.49)	−0.27 (−0.73 to 0.19)	**0.39** (0.19 to 0.59)

[a] The β coefficient of $PM_{2.5}$ contributions from each source type represents the estimated average increase in daily mortality counts per 5th-to-95th percentile increment of estimated $PM_{2.5}$ source contribution (μg/m^3) while controlling for other variables in the model; significant effects are denoted in **bold**; 95% credible intervals are given in parentheses. Source names in italics are conjectures based on the source compositions of previous studies.

approach does account for the uncertainty in the estimated source contributions in the estimation of the health effects parameters. Statistically insignificant estimates for Secondary Sulfate (lag 0) and Traffic (lag 1) might have been a consequence of incorporating the uncertainty that had not previously been addressed.

HOUSTON DATA ANALYSIS

We obtained data on mortality, weather, and air pollution for the Houston, Texas, area near Clinton Drive (which is close to the Houston Ship Channel) shown in Figure 10, as well as for the entire Harris County (including the Clinton Drive area). Appendix E (available on the HEI Web site) contains detailed explanations on database development and summary statistics.

For the assessment of source-specific health effects in the Clinton Drive region, we used the $PM_{2.5}$ speciation data for which prior information on potential source types around the area were available from previous studies (e.g., Sullivan 2007). Because there are no $PM_{2.5}$ speciation data available at the Clinton Drive monitoring site, the $PM_{2.5}$ data measured (every 6th day from January 2002 to August 2005) at the Houston East monitoring station on Mae Drive were used for the analysis. This site is closest to Clinton Drive (three miles northeast of Clinton), and the data from this site has been used in other source-apportionment analyses of Clinton Drive (see Sullivan 2007). The original $PM_{2.5}$ speciation data consist of measurements on 77 species. Summary statistics for the original 77 $PM_{2.5}$ species measured at the Houston East monitoring station are provided in Appendix Table E.15. As noted earlier, the first important step in multivariate receptor modeling is to select an appropriate subset of species for the analysis.

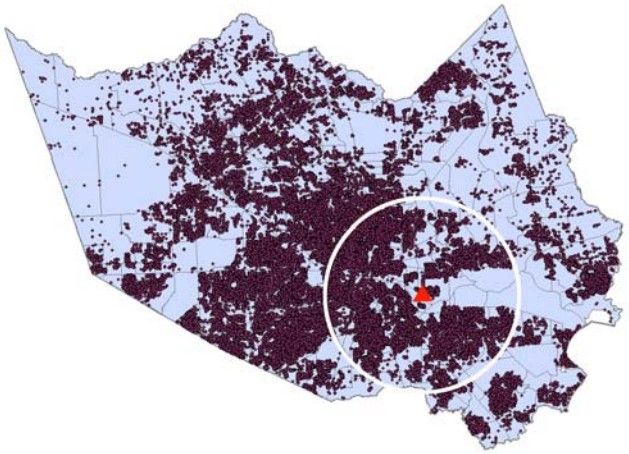


Figure 10. Map of Harris County, TX (2000–2005), showing a 10-mile buffer surrounding the Clinton Drive monitoring site (red triangle) with geocoded residences of decedents (purple dots).

Inclusion of unhelpful species (e.g., species of which uncertainty values exceed the measurement value) in fitting may hinder receptor modeling. The same set of 17 species used in Sullivan (2007) that incorporated both prior knowledge about potential sources for the region and signal-to-noise ratio for the selection of species was also selected for our model fitting. Table 13 contains the list of selected species as well as their summary statistics. Recall that in a source-specific approach, we are concerned about the effect of a combination of species (realized as source contributions) and even the effect of excluded species can be incorporated into the health-effects analysis of sources emitting that species because the source contributions do not depend on individual species.

The method developed to analyze the discrete health outcome data with a low mean has been applied to daily counts of deaths due to respiratory causes along with the $PM_{2.5}$ speciation data collected from the Houston East monitoring site. Summary statistics for all nonaccidental causes and specific-cause mortality by area (10-mile buffer region, Harris County) for decedents of all ages and decedents 65 and older at the time of death are shown in Table 14. Our approach will be primarily illustrated with respiratory mortality data because daily mean counts are small with mean values of 1.0 and 3.1 (for decedents 65 and older) for the 10-mile buffer region near Clinton drive and for the entire Harris County, respectively.

The weather data from the Clinton Drive monitoring station had too many missing values (4% of temperature data and 80% for relative humidity data), so temperature and dew point temperature data collected from the Hobby airport were used. The airport is close to Clinton Drive (see Appendix Figure E.6).

Previously, as many as seven sources were identified as the sources affecting the Houston East monitoring site by PMF analysis: Sulfate-Rich Secondary Aerosol, Motor Vehicles, Industrial Combustion, Biomass Burning, Soil/Crustal Matter, Sea Salt, and Nitrate-Rich Aerosol (see Sullivan 2007). We constructed a set of plausible candidate models (resulting from each combination of numbers

Table 13. Summary Statistics for 17 $PM_{2.5}$ Chemical Species Measured at Houston East[a]

Species Number	$PM_{2.5}$ Species[b]	Number of Nonmissing Values	Average	SD[b]	Minimum	Maximum
1	Aluminum	217	0.073	0.191	0	1.410
2	Calcium	217	0.099	0.066	0	0.395
3	Chromium	217	0.001	0.003	0	0.035
4	Chlorine	217	0.070	0.184	0	1.440
5	Iron	217	0.111	0.114	0	0.877
6	Nickel	217	0.002	0.002	0	0.008
7	Titanium	217	0.008	0.013	0	0.113
8	Vanadium	217	0.005	0.005	0	0.029
9	Silicon	217	0.283	0.426	0	3.220
10	Zinc	217	0.016	0.021	0	0.201
11	Potassium	217	0.075	0.059	0	0.359
12	Ammonium ion	217	1.251	0.884	0	5.690
13	Sodium	217	0.158	0.236	0	1.490
14	OC CSN unadjusted	224	3.464	1.690	0.416	9.610
15	EC CSN	224	0.676	0.348	0	2.050
16	Nonvolatile nitrate	217	0.345	0.438	0.017	3.630
17	Sulfate	217	3.761	2.279	0.020	14.800

[a] All units are in $\mu g/m^3$.

[b] CSN denotes chemical speciation network; SD denotes standard deviation.

of sources and identifiability conditions) to be compared based on previous studies of the region and exploratory analyses using PMF and the NUMFACT procedure. Table 15 contains six candidate models corresponding to the numbers of sources (q = 4, 5, 6, 7) and identifiability conditions (prespecification of zeros in **P**) that we compared.

The $PM_{2.5}$ speciation data, respiratory mortality data, and weather data were simultaneously fitted to estimate source-composition profiles, contributions, and health-effects parameters as well as marginal likelihood under each model in Table 15 at lag days 0–2. We first considered the respiratory mortality data for decedents (of all ages) whose residences at the time of death belonged to the 10-mile buffer region surrounding the Clinton Drive monitoring site. We constructed the base health effects model as a function of time trends and seasonality and of weather variability in Poisson GLMs. The model includes smoothing terms for calendar time, temperature, and dew point temperature. Using natural spline smoothers, the base model was constructed to adjust for long-term time trends and other unmeasured seasonal confounders as well as potential nonlinearity in the relationship between weather conditions and mortality (Bell et al. 2004). The degrees of freedom for the natural splines were selected to minimize autocorrelation in the residuals and the Akaike information criterion (AIC) values. Indicator variables for the day of the week and extreme temperature were also included in the model. Extreme temperature values (< 5th or > 95th percentiles) were coded as 1, all other temperature values were coded as 0. The base model was compared with and without including extreme temperature in the model as a part of sensitivity analyses. We also compared a day-of-week variable

Table 14. Summary Statistics for All Nonaccidental Mortality and Specific Causes of Cardiovascular and Respiratory Mortality for the 10-Mile Buffer Region Surrounding the Clinton Drive Monitoring Site and for all of Harris County, Texas, 2000–2005

	Area			
	10-Mile Buffer Region		Harris County[a]	
Population / Mortality Cause[b]	Number	Mean Daily Number of Deaths	Number	Mean Daily Number of Deaths
All Ages				
Cardiovascular (I00-I99)	15,316	6.7	41,708	19.0
IHD (I20-I25)	7,384	3.4	20,370	9.3
Acute MI (I21)	2,910	1.3	7,786	3.6
Heart failure (I50)	1,088	0.5	2,884	1.3
Stroke (I60-69)	2,806	1.3	8,129	3.7
Respiratory (J00-J98)	2,818	1.3	8,478	3.9
COPD (J40-44)	1,317	0.6	4,124	1.9
Pneumonia (J12-18)	825	0.4	2,335	1.1
Nonaccidental causes (A00-R99)	38,610	17.6	106,772	48.7
≥65 Years				
Cardiovascular (I00-I99)	11,287	5.1	31,425	14.3
IHD (I20-I25)	5,478	2.5	15,370	7.0
Acute MI (I21)	2,216	1.0	5,990	2.7
Heart failure (I50)	976	0.4	2,606	1.2
Stroke (I60-69)	2,216	1.0	6,694	3.1
Respiratory (J00-J98)	2,250	1.0	6,930	3.1
COPD (J40-44)	1,111	0.5	3,537	1.6
Pneumonia (J12-18)	652	0.3	1,902	0.9
Nonaccidental causes (A00-R99)	25,494	11.6	71,986	32.8

[a] Includes the 10-mile buffer region.

[b] IHD denotes ischemic heart disease, MI denotes myocardial infarction, and COPD denotes chronic obstructive pulmonary disease.

Table 15. Candidate Models for Houston $PM_{2.5}$ Speciation Data

Model Number	q	Source	Prespecified Position of Zeros in P (Species Number[a] with Preassigned Zeros)
1	4	1	1, 6, 16
		2	12, 13, 16
		3	6, 10, 12
		4	7, 10, 11
2	5	1	1, 4, 6, 16
		2	3, 7, 10, 11
		3	3, 5, 6, 10
		4	4, 6, 10, 12
		5	4, 7, 10, 11
3	5	1	1, 4, 6, 16
		2	3, 7, 13, 16
		3	3, 5, 6, 10
		4	4, 6, 10, 12
		5	3, 7, 10, 11
4	6	1	1, 4, 5, 6, 16
		2	4, 9, 12, 13, 16
		3	3, 5, 6, 8, 10
		4	4, 6, 10, 12, 13
		5	1, 4, 7, 10, 11
		6	1, 3, 4, 5, 17
5	6	1	1, 4, 5, 6, 16
		2	4, 9, 12, 13, 16
		3	3, 5, 6, 8, 10
		4	4, 6, 10, 12, 13
		5	1, 4, 7, 10, 11
		6	4, 6, 7, 9, 10
6	7	1	1, 4, 5, 6, 8, 16
		2	1, 3, 4, 5, 7, 17
		3	4, 8, 9, 12, 13, 16
		4	3, 5, 6, 7, 8, 10
		5	3, 4, 6, 10, 12, 13
		6	1, 2, 4, 7, 10, 11
		7	4, 6, 7, 8, 9, 10

[a] Species numbers are defined in Table 13.

and workday variable (i.e., 1 for weekday and 0 for weekend) in the base model, and found almost no difference in AIC. A range of alternative lags and degrees of freedom were explored. We explored 2–6 *df* for the smoothing of temperature and dew point temperature with 0–2 lags, but there was no big difference in AIC. We ended up using the base model with extreme temperature, workday, 4 *df* per year for the smoothing of calendar time, 4 *df* for the smoothing of temperature, and 3 *df* for dew point temperature, with 2 lag days for temperature and 0 lag days for dew point temperature, respectively.

In implementing MCMC, we used the following hyperparameter values for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,17$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I_{p+}}$, $m_0 = \bar{X}$, $M_0 = 100 \times \mathbf{I_J}$, $\alpha_0 = 0$, $U_0 = 10$, $\beta_0 = 0_q$, $B_0 = 10 \times \mathbf{I_q}$, $\eta_0 = 0_I$, $\Psi_0 = 10 \times \mathbf{I_I}$, $r_0 = q$, and $R_0 = \mathbf{I_q} \times (r_0 + q + 1)$. For each model, an approximate posterior mode is obtained from a preliminary MCMC run, and this is used for $\theta^C = \mu^C, \Gamma^C, \Omega^C, \mathbf{P}^C, \Sigma^C, \alpha^C, \beta^C, \eta^C, W^C$, at which the marginal likelihood is calculated. An approximate posterior mode is obtained by evaluating the joint posterior density for 30,000 iterations after the first 30,000 draws are discarded. A main MCMC run is then started from $\theta^C = \mu^C, \Gamma^C, \Omega^C, \mathbf{P}^C, \Sigma^C, \alpha^C, \beta^C, \eta^C, W^C$, and the samples are collected for 30,000 iterations, subsampling every 10th value (resulting in 3,000 samples), without additional burn-in. The marginal likelihood for each model was computed in the main run. Table 16 gives the estimated marginal likelihoods (in logs) for each model of Table 15, jointly modeling the $PM_{2.5}$ and respiratory mortality data at 0 lag days. The posterior probability of each model under the indifference prior is also provided in Table 16. Model 2 with five sources led to the highest posterior probability (that is almost 1) among the models considered. For other lag days (lag days 1–2) also, Model 2 led to the highest posterior model probability that is very close to 1.

The estimated source-composition profiles, along with their uncertainty estimates (95% posterior intervals) under Model 2 are given in Table 17. The estimated source-composition profiles and contributions based on the $PM_{2.5}$ and respiratory mortality data with other lags are materially the same as those in Table 17. Major species in the estimated

Table 16. Marginal Likelihoods for Candidate Models for Houston East $PM_{2.5}$ Speciation Data and Respiratory Mortality at Lag 0 Days[a]

Model Number	Number of Sources (q)	LogMD	PostP
1	4	−5276.3	0.0000
2	5	−5116.0	1.0000
3	5	−5298.6	0.0000
4	6	−5276.9	0.0000
5	6	−5422.6	0.0000
6	7	−5639.6	0.0000

[a] LogMD and PostP denote the Log of Marginal Likelihood and Posterior Model Probability, respectively.

Table 17. Estimated Source Composition Profiles for 17 Species Under Model 2 for Houston East[a]

Species Number	$PM_{2.5}$ Species[b]	Source 1 *Motor Vehicles*	Source 2 *Industrial Combustion*	Source 3 *Sea Salt*	Source 4 *Soil/Crustal Matter*	Source 5 *Sulfate-Rich Secondary Aerosol*
Source Composition: % (95% Posterior Intervals)						
1	Aluminum	**0**	0.17 (0.01 to 0.52)	4.16 (1.78 to 6.56)	18.11 (15.52 to 20.62)	0.22 (0.01 to 0.55)
2	Calcium	1.09 (0.76 to 1.43)	0.23 (0.03 to 0.47)	3.52 (2.49 to 4.77)	4.95 (4.12 to 5.81)	0.06 (0.00 to 0.17)
3	Chromium	0.01 (0.00 to 0.03)	**0**	**0**	0.16 (0.12 to 0.20)	0.01 (0.00 to 0.01)
4	Chlorine	**0**	0.58 (0.06 to 1.25)	22.85 (17.60 to 28.62)	**0**	**0**
5	Iron	1.22 (0.88 to 1.61)	0.36 (0.10 to 0.62)	**0**	11.67 (10.08 to 13.20)	0.03 (0.00 to 0.10)
6	Nickel	**0**	0.05 (0.05 to 0.06)	**0**	**0**	0.00 (0.00 to 0.00)
7	Titanium	0.00 (0.00 to 0.01)	**0**	0.07 (0.01 to 0.16)	1.43 (1.23 to 1.61)	**0**
8	Vanadium	0.00 (0.00 to 0.01)	0.17 (0.15 to 0.19)	0.03 (0.00 to 0.07)	0.04 (0.00 to 0.08)	0.01 (0.00 to 0.02)
9	Silicon	0.11 (0.00 to 0.54)	0.92 (0.07 to 2.01)	8.79 (2.21 to 14.68)	44.76 (38.76 to 50.34)	0.07 (0.00 to 0.30)
10	Zinc	0.43 (0.29 to 0.59)	**0**	**0**	**0**	**0**
11	Potassium	0.85 (0.49 to 1.21)	**0**	3.68 (2.53 to 4.96)	3.16 (2.40 to 3.98)	**0**
12	Ammonium ion	1.78 (0.10 to 4.38)	9.85 (7.03 to 12.61)	1.00 (0.02 to 3.61)	**0**	21.48 (19.84 to 28.28)
13	Sodium	0.21 (0.01 to 0.70)	0.28 (0.01 to 0.93)	31.59 (25.29 to 38.22)	1.35 (0.05 to 3.65)	0.32 (0.01 to 0.87)
14	OC CSN unadjusted	71.46 (63.33 to 80.64)	42.81 (36.78 to 49.95)	3.71 (0.09 to 12.67)	4.07 (0.13 to 13.10)	20.95 (16.10 to 25.31)
15	EC CSN	11.26 (8.95 to 13.70	4.89 (3.19 to 6.55)	0.88 (0.02 to 3.16)	0.76 (0.01 to 2.48)	0.43 (0.01 to 1.25)
16	Nonvolatile nitrate	**0**	4.31 (1.97 to 6.61)	12.58 (3.58 to 20.90)	3.04 (0.16 to 7.69)	0.68 (0.02 to 1.93)
17	Sulfate	11.56 (1.89 to 20.21)	35.39 (29.52 to 40.83)	7.16 (0.17 to 22.84)	6.53 (0.27 to 16.45)	55.75 (52.05 to 59.88)
	Source Contribution ($\mu g/m^3$)					
	Mean	3.32	2.75	0.41	0.52	3.72
	Standard deviation	1.94	2.47	0.59	0.91	3.42
	5th-to-95th increment	5.92	7.31	1.50	2.25	9.46

[a] Source composition percentages are normalized to Sum = 100%. **Bolding** gives the position of preassigned zeros. Source names in *italics* are conjectures based on the source compositions of previous studies (cited in the text).

[b] CSN denotes chemical speciation network.

source-composition profiles of Table 17 were consistent with main elements of Sulfate-Rich Secondary Aerosol, Motor Vehicles, Industrial Combustion, Soil/ Crustal Matter, and Sea Salt that were also identified in previous studies (Sullivan 2007). For example, Source 1, which is characterized by high percentages of OC and EC, seems to represent Motor Vehicles (or Traffic). Source 3, which is characterized by high chlorine and sodium, and Source 4, which is characterized by aluminum, iron, and silicon, appear to correspond to Sea Salt and Soil/Crustal Matter, respectively. Source 2 and Source 5 both have ammonium, OC, and sulfate as main species, which likely represent Industrial Combustion and Sulfate-Rich Secondary Aerosol, respectively.

Figure 11 contains the time-series plots of the estimated source contributions (in μg/m^3) for 215 days (1/2/2002–8/26/2005) along with their uncertainty estimates (95% posterior intervals), which play a role of source-specific exposures in health effects estimation.

Table 18 presents source-specific effects on respiratory mortality at 0–2 lag days. Only the effect due to Source 2

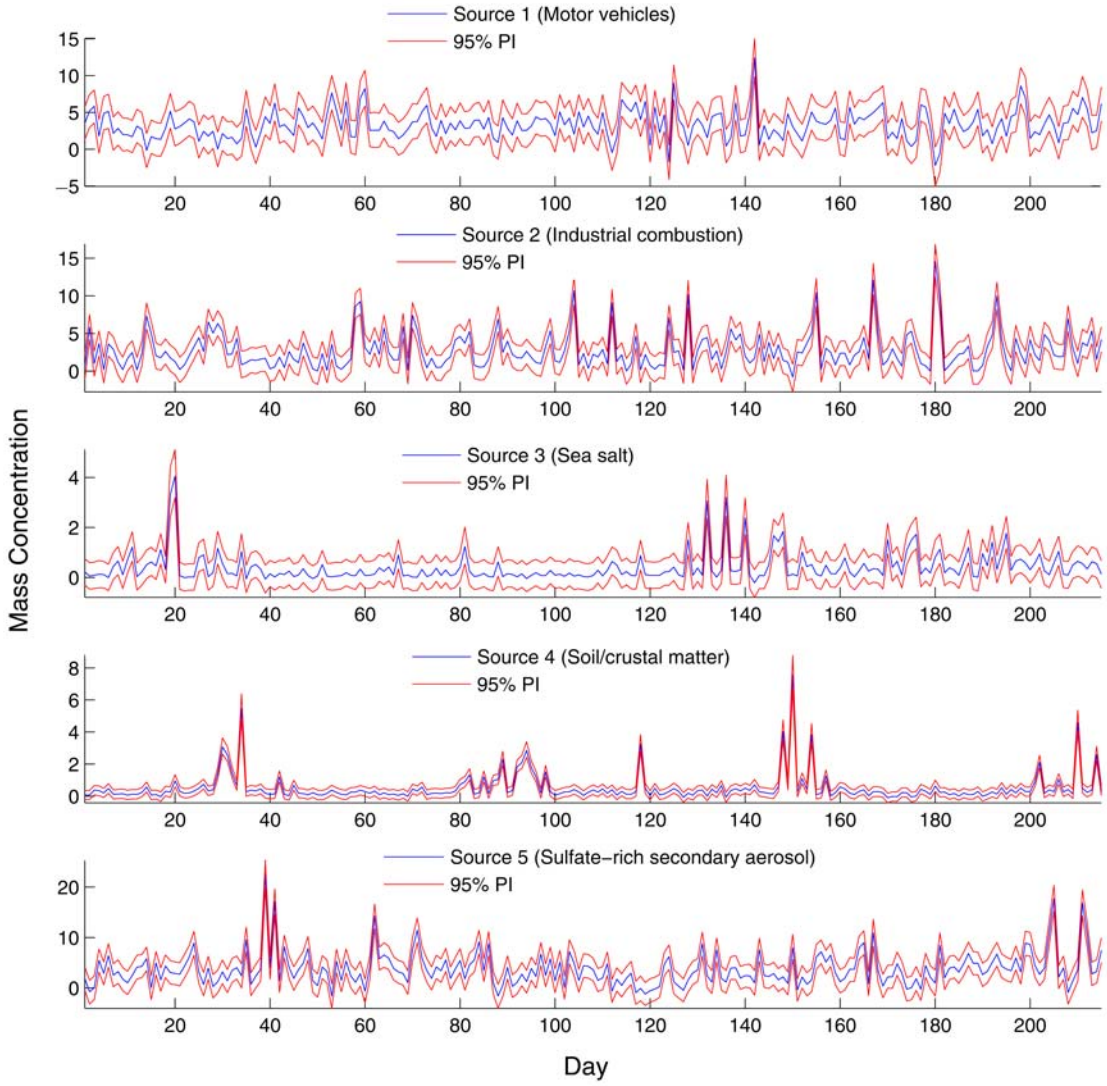


Figure 11. Time-series plots of the estimated source contributions (in μg/m^3) for 215 days along with their uncertainty estimates (95% posterior intervals).

(that appears to be Industrial Combustion) at 0 lag days was statistically significant (i.e., a 95% posterior interval does not contain 0).

Next, we considered decedents ≥65 years of age whose residences at the time of death due to respiratory disease were within the 10-mile buffer region surrounding the Clinton Drive monitoring site to determine whether effects were different among an older population. Model 2 again led to the highest posterior probability that is close to 1, and estimated source-composition profiles and source contributions did not change materially from those of Table 17. The estimated health effects parameters at 0–2 lag days are provided in Table 19. As in the case of Table 18, only the health effects due to Source 2 (that appears to be Industrial Combustion) at 0 lag days was statistically significant (i.e., a 95% posterior interval does not contain 0). There appeared to be slightly stronger effects of Source 2 for decedents 65 and older, compared to decedents of all ages. Note that the length of time series is very short for these data (only 215 days), which might have contributed to the insignificance and frequent sign changes of most of the health effects. The 95% posterior intervals given in Tables 18–19 generally appear to be wide, perhaps because of the small number of observations, but also because uncertainty in the estimated source contributions was incorporated into these intervals.

We also analyzed the daily COPD mortality counts (for decedents ≥ 65 years of age whose residences at the time of death were within the 10-mile buffer region surrounding the Clinton Drive monitoring site) along with the same $PM_{2.5}$ speciation data and weather data as above. There was no noticeable change in the estimated source-composition profiles and contributions as well as in the number of sources. The estimated $PM_{2.5}$ source-specific effects associated with COPD at 0–2 lag days are provided in Table 20. No significant effects were observed.

Table 18. $PM_{2.5}$ Source-Specific Effects on Respiratory Mortality for Decedents (Regardless of Age) Who Resided at the Time of Death in the 10-Mile Buffer Region[a]

	Source 1	Source 2	Source 3	Source 4	Source 5
β (lag 0)	−0.06	**0.44**	0.14	0.17	0.23
	(−0.48 to 0.39)	**(0.07 to 0.88)**	(−0.29 to 0.55)	(−0.19 to 0.48)	(−0.24 to 0.64)
β (lag 1)	0.02	0.16	0.12	0.22	0.10
	(−0.42 to 0.47)	(−0.25 to 0.58)	(−0.29 to 0.50)	(−0.13 to 0.53)	(−0.35 to 0.53)
β (lag 2)	−0.13	0.17	0.27	−0.32	0.00
	(−0.59 to 0.32)	(−0.23 to 0.57)	(−0.09 to 0.63)	(−0.86 to 0.13)	(−0.44 to 0.44)

[a] The β coefficient of $PM_{2.5}$ contributions from each source type represents the estimated log-relative risk per 5th-to-95th percentile increment of estimated $PM_{2.5}$ source contribution (μg/m³); significant effects are denoted in **bold**; 95% posterior intervals are given in parentheses.

Table 19. $PM_{2.5}$ Source-Specific Effects on Respiratory Mortality for Residents ≥ 65 Years at the Time of Death Who Resided in the 10-Mile Buffer Region[a]

	Source 1	Source 2	Source 3	Source 4	Source 5
β (lag 0)	0.13	**0.54**	−0.04	0.14	0.20
	(−0.39 to 0.70)	(0.09 to 0.98)	(−0.57 to 0.46)	(−0.30 to 0.52)	(−0.28 to 0.65)
β (lag 1)	0.25	0.13	0.18	0.19	0.12
	(−0.31 to 0.83)	(−0.33 to 0.60)	(−0.28 to 0.62)	(−0.20 to 0.57)	(−0.41 to 0.64)
β (lag 2)	0.00	0.26	0.30	−0.52	−0.08
	(−0.47 to 0.49)	(−0.15 to 0.67)	(−0.11 to 0.71)	(−1.23 to 0.04)	(−0.57 to 0.40)

[a] The β coefficient represents the estimated log-relative risk per 5th-to-95th percentile increment of estimated $PM_{2.5}$ source contribution (μg/m³); significant effects are denoted in **bold**; 95% posterior intervals are given in parentheses.

Lastly, we analyzed the respiratory mortality counts and the $PM_{2.5}$ speciation data (controlling for weather variables as before) for Harris County for residents ≥ 65 years of age at the time of death to determine whether observed associations are modified when the study area is increased to include decedents whose residences are distant from the monitoring site where the source contributions were estimated. There was no noticeable change in the estimated source-composition profiles and contributions as well as in the number of sources. However, regression coefficients were generally attenuated and no sources were statistically significant. Even the effects of Source 2 (at 0 lag days) diminished greatly in magnitude and lost statistical significance compared with the significant effects for the 10-mile buffer zone (see Tables 19 and 21). This may indicate that source-specific effects from estimated source contributions that rely on monitoring data from a single monitoring site are attenuated when the study area expands in size.

ANALYSIS OF HARRIS COUNTY VOC DATA COLLECTED FROM MULTIPLE MONITORING SITES

The spatially-enhanced receptor models developed in this project, Bayesian spatial multivariate receptor models, have been applied to the 24-hour VOC data collected every six days from nine monitoring sites in Harris County during January 1, 2000–December 30, 2005. Figure 12 shows the locations of the nine monitoring sites.

There were a total of 421 days when VOC measurements were made for at least one of the nine monitoring sites. Occasionally, observations were made less than 6 days apart at some of the locations, which led to 56 extra days in addition to those sampled every 6th day. The number of missing observations (days with no VOC measurements) at each site ranges from 34 to 128. We imputed the missing observations by *k*-nearest neighbor imputation (Little and Rubin 1987), namely, using the spatial average of pollutants from three nearest neighboring sites for each day. The patterns of missing data are given in Appendix Tables E.16 and E.17.

Table 20. $PM_{2.5}$ Source-Specific Effects on Mortality from COPD for Residents ≥65 Years at the Time of Death Who Resided in the 10-Mile Buffer Region[a]

	Source 1	Source 2	Source 3	Source 4	Source 5
β (lag 0)	−0.15	0.37	0.03	0.22	0.27
β (lag 1)	0.24	0.25	−0.04	−0.35	−0.38
β (lag 2)	−0.33	0.30	0.38	−0.45	0.10

[a] The β coefficient represents the estimated log-relative risk per 5th-to-95th percentile increment of estimated $PM_{2.5}$ source contribution (μg/m^3); significant effects are denoted in **bold**.

Table 21. $PM_{2.5}$ Source-Specific Effects on Mortality from Respiratory Causes for Residents ≥65 Years at the Time of Death Who Resided in Harris County[a]

	Source 1	Source 2	Source 3	Source 4	Source 5
β (lag 0)	−0.07	−0.06	0.05	0.01	0.19
β (lag 1)	0.04	0.00	0.18	0.01	−0.04
β (lag 2)	0.06	−0.08	−0.04	0.03	0.07

[a] The β coefficient represents the estimated log-relative risk per 5th-to-95th percentile increment of estimated $PM_{2.5}$ source contribution (μg/m^3); significant effects are denoted in **bold.**

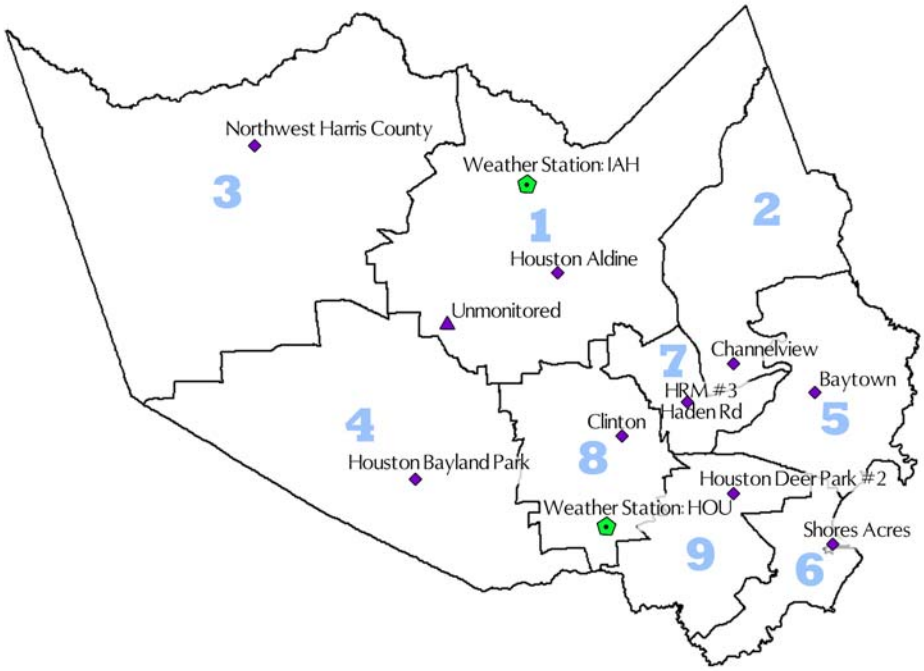


Figure 12. Map of nine monitoring sites (diamonds) in Harris County, TX. Unmonitored site (triangle): latitude: 29.856, longitude: −95.452.

Recall that the first important step in multivariate receptor modeling is to select an appropriate subset of species for an analysis; inclusion of noisy or unhelpful species could hinder source apportionment (Park et al. 2001). It is better to remove any species with many near-zero values, unless it is a tracer element of a major source. Also, reactive species are typically excluded from the analysis because those species do not satisfy the basic underlying assumptions in multivariate receptor modeling. It needs to be noted that the number of sources that can be identified also depends on which species were included in the analysis. It is possible to identify more sources with the inclusion of more species. Thus, it is important not to omit key species for potential sources for the region while excluding noisy species. Based on the previous studies on the region (e.g., Buzcu and Fraser 2006), refineries, petrochemical production facilities, unburned gasoline (liquid or evaporated), natural gas, vehicle exhaust, and aromatics were presumed to be potential candidate sources affecting the region. This prior knowledge was utilized in selecting an appropriate subset of the species that are contributed by those sources as well as in prespecification of zeros in the source-composition profile matrix to later achieve model identifiability. Table 22 gives the major species for each of the candidate source types. We selected 17 species (presented in Table 24) that seem to be important at the sites considered (in Figure 12) from the 107 VOC species originally measured. Spatial correlations over nine monitoring sites in Figure 12 for some of the VOCs included in Table 24 are given in Appendix E.

We constructed a range of different models (resulting from each combination of different number of sources and identifiability conditions) to be compared for the Harris County VOC data. Based on previous studies on source identification and apportionment of VOCs for the region and the NUMFACT procedure, we presumed that the number of major sources was between four and seven. For candidate positions of zeros in $\mathbf{P}$ under each q-source model, we again used the information on the major sources from previous studies and also conducted exploratory data analyses using PMF. For example, we can use the information that ethane is typically not present in emissions from Gasoline Evaporation and can be prespecified to be zero. Note that we use this type of information from previous studies only to find out the plausible sets of identifiability conditions (positions of zeros) under each q-source model. Other than that, the candidate models do not depend on the results from those previous studies. We compared seven candidate models with different numbers of sources (q = 4, 5, 6, 7) and different prespecification of identifiability conditions (zeros in $\mathbf{P}$) in Table 23.

Table 22. Major Species for Candidate Sources Considered in the Analysis

Candidate Sources	Major Species
Refinery	Propane, ethane, *n*-butane, isobutane
Petrochemical production	Ethylene, propylene
Unburned gasoline	*n*-Butane, isopentane, isobutane, *n*-pentane, tolune
Natural gas	Ethane, propane
Vehicle exhaust	Tolune, xylenes, acetylene, ethylene, isobutane, propylene
Aromatric	Tolune, xylenes

We fitted Bayesian spatial multivariate receptor models to the data consisting of 17 VOC species and estimated source-composition profiles and other model parameters along with marginal likelihood under each model. The following hyper-parameter values were used for generating MCMC samples: $a_0 = 0.01$, $b_{0j} = 0.01$ ($j = 1,\cdots,15$), $c_0 = 0.5 \times 1_{p+}$, $C_0 = 100 \times \mathbf{I}_{\mathbf{p+}}$, $m_0 = \bar{X}$, and $M_0 = 100 \times \mathbf{I}_{\mathbf{J}}$. Also, we set $\Omega_q = \mathbf{I}_{\mathbf{q}}$ as a way to get around a scale invariance problem for these data.

For each model, an approximate posterior mode is obtained from a preliminary MCMC run, and this is used for $\theta^c = \left(\mu^c, G^c, \mathbf{P}^c, \Sigma^c\right)$, at which the marginal likelihood is calculated. An approximate posterior mode is obtained by evaluating the joint posterior density for 50,000 iterations after the first 50,000 draws are discarded. A main MCMC run is then started from $\theta^c = \left(\mu^c, G^c, \mathbf{P}^c, \Sigma^c\right)$, and the samples are collected for 50,000 iterations, subsampling every 10th value (resulting in 5,000 samples), without additional burn-in. The marginal likelihood for each model can be computed in sample generation without storing the samples. The estimated marginal likelihood (in logs) for each model is also provided in Table 23 as well as the posterior probability under the indifference prior. Model 2 with 5 sources is selected as the best model because the posterior probability for Model 2 is the highest (almost 1) among the candidate models considered.

Table 24 gives the estimated source-composition profiles along with their uncertainty estimates (95% posterior intervals) and the estimated mean source contributions under

Table 23. Candidate Models for Harris County VOC Data[a]

Model Number	q	Source	Prespecified Position of Zeros in **P** (Species Number[b] with Preassigned Zeros)	LogMD ($\times 10^4$)	PostP
1	4	1	8, 12, 16	−8.8043	0.0000
		2	10, 16, 17		
		3	6, 8, 12		
		4	8, 12, 15		
2	5	1	2, 9, 12, 16	−8.7913	1.0000
		2	6, 11, 14, 16		
		3	4, 6, 12, 17		
		4	1, 3, 12, 17		
		5	1, 3, 14, 17		
3	5	1	4, 8, 12, 15	−8.8059	0.0000
		2	3, 7, 8, 12		
		3	4, 6, 7, 8		
		4	6, 8, 11, 15		
		5	1, 11, 14, 17		
4	5	1	8, 12, 14, 15	−8.8083	0.0000
		2	3, 7, 8, 12		
		3	4, 6, 7, 8		
		4	4, 10, 11, 14		
		5	1, 10, 14, 17		
5	6	1	4, 8, 12, 16, 17	−8.8381	0.0000
		2	4, 7, 10, 14, 17		
		3	1, 4, 6, 7, 8		
		4	4, 8, 9, 10, 12		
		5	1, 10, 11, 14, 16		
		6	8, 9, 10, 11, 14		
6	6	1	1, 3, 7, 10, 12	−8.8230	0.0000
		2	4, 7, 10, 14, 17		
		3	1, 4, 6, 7, 8		
		4	2, 4, 8, 10, 12		
		5	1, 10, 11, 14, 16		
		6	8, 9, 10, 11, 14		
7	7	1	2, 4, 8, 12, 16, 17	−8.8440	0.0000
		2	4, 7, 10, 14, 16, 17		
		3	1, 2, 4, 6, 7, 8		
		4	4, 8, 9, 10, 12, 16		
		5	1, 2, 10, 11, 14, 16		
		6	8, 9, 10, 11, 14, 16		
		7	4, 6, 8, 10, 12, 16		

[a] LogMD and PostP denote the Log of Marginal Likelihood and Posterior Model Probability, respectively.

[b] Species numbers are defined in Table 24.

Table 24. Estimated Source Composition Profiles Under Model 2[a]

Species Number	Species Name	Source 1 *Refinery*	Source 2 *Petrochem*	Source 3 *Gasoline*	Source 4 *Natural Gas*	Source 5 *Vehicle Exhaust*
Source Composition: % (95% Posterior Intervals)						
1	1,2,4-Trimethylbenzene	**0**	**0**	0.31 (0.19 to 0.47)	0.45 (0.34 to 0.60)	1.04 (0.51 to 1.77)
2	1,3-Butadiene	0.33 (0.24 to 0.42)	0.51 (0.18 to 0.85)	0.89 (0.58 to 1.25)	**0**	1.75 (0.54 to 3.33)
3	2,2,4-Trimethylpentane	**0**	**0**	1.27 (0.96 to 1.68)	0.43 (0.16 to 0.76)	0.89 (0.08 to 2.03)
4	Acetylene	2.46 (0.13 to 2.78)	0.63 (0.03 to 1.58)	**0**	4.80 (3.61 to 6.09)	20.84 (13.54 to 31.95)
5	Benzene	1.33 (1.11 to 1.56)	1.53 (0.28 to 2.41)	2.67 (1.94 to 3.51)	0.38 (0.02 to 0.95)	2.57 (0.26 to 5.79)
6	Ethane	26.58 (24.50 to 28.63)	25.83 (19.11 to 31.52)	**0**	42.06 (37.74 to 46.16)	**0**
7	Ethylbenzene	0.11 (0.05 to 0.16)	0.07 (0.00 to 0.19)	0.34 (0.15 to 0.55)	0.47 (0.30 to 0.66)	0.93 (0.21 to 1.85)
8	Ethylene	6.91 (6.09 to 7.74)	12.61 (9.75 to 15.89)	10.32 (7.64 to 13.25)	3.41 (0.96 to 5.80)	26.31 (17.27 to 36.58)
9	Isobutane	16.91 (14.43 to 19.45)	1.08 (0.03 to 3.84)	29.64 (21.76 to 36.11)	**0**	5.09 (0.15 to 16.69)
10	Isopentane	6.61 (5.95 to 7.28)	7.13 (4.80 to 9.57)	15.60 (13.33 to 18.67)	2.13 (0.59 to 3.74)	4.34 (0.25 to 10.11)
11	Propane	19.87 (18.57 to 21.24)	17.87 (13.24 to 21.85)	5.38 (1.24 to 8.85)	27.92 (24.82 to 30.82)	**0**
12	Propylene	**0**	25.51 (19.51 to 32.59)	**0**	**0**	23.00 (2.30 to 40.59)
13	Toluene	0.45 (0.02 to 1.10)	1.40 (0.07 to 3.48)	3.91 (1.71 to 6.55)	4.94 (2.64 to 7.53)	5.10 (0.24 to 13.10)
14	*n*-Butane	14.16 (13.08 to 15.23)	**0**	20.32 (17.27 to 23.37)	10.10 (6.96 to 13.15)	**0**
15	*n*-Hexane	1.01 (0.87 to 1.15)	1.93 (1.36 to 2.61)	2.89 (2.34 to 3.61)	0.17 (0.01 to 0.49)	2.07 (0.39 to 4.06)
16	*n*-Pentane	3.26 (2.96 to 3.57)	3.91 (2.71 to 5.14)	6.48 (5.38 to 7.82)	**0**	**0**
17	*m*- & *p*-Xylenes	**0**	**0**	**0**	2.74 (1.32 to 4.21)	6.07 (0.93 to 12.71)
	Mean source contribution (ppbC)	21.61	10.79	2.28	6.49	2.42

[a] Source composition percentages are normalized to Sum = 100%. **Bolding** gives the position of preassigned zeros. Source names in *italics* are conjectures based on the source compositions of previous studies (cited in the text). ppbC denotes parts per billion carbon.

Model 2. Major species in the estimated source-composition profiles of Table 24 are consistent with main elements of major VOC sources for Harris County identified by previous studies, namely, Refinery, Petrochemical Production, Gasoline Evaporation, Natural Gas, and Vehicle Exhaust.

The estimated mean source contributions indicate that overall refineries and petrochemical production facilities play a major role in VOC emissions for the region; this agrees with result in Buzcu and Fraser (2006). We illustrated an application of estimated source contributions from spatially-enhanced multivariate receptor models in health effects modeling in Appendix F (available on the HEI Web site).

Note that spatial multivariate receptor models result in more precise estimates of source profiles compared to an approach that did not account for spatial dependence, as

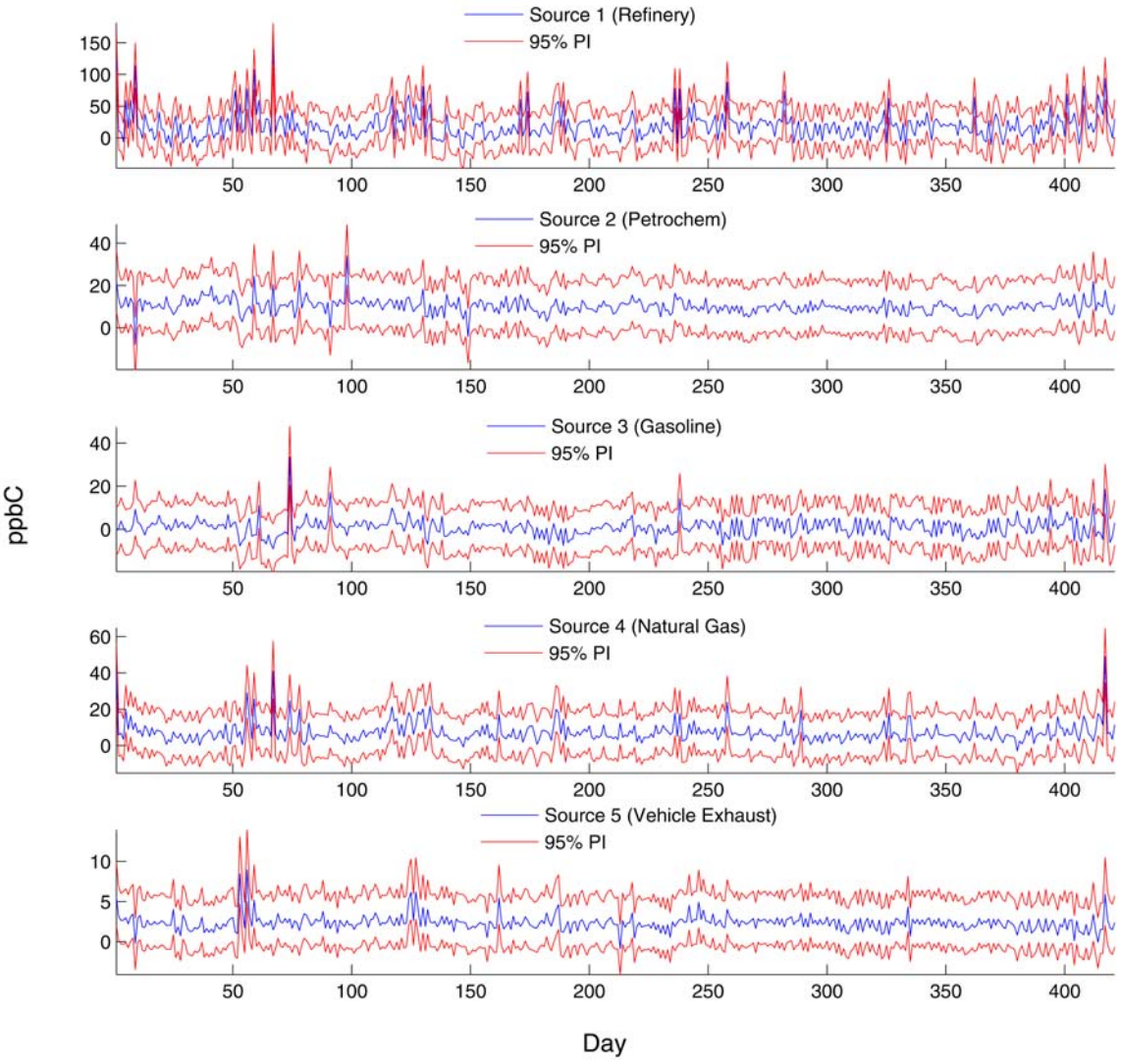


Figure 13. Time-series plots of the predicted source contributions along with uncertainty estimates (95% posterior intervals) at the unmonitored site (see Figure 12). ppbC denotes parts per billion carbon.

demonstrated in Jun and Park (2013). In addition, the new Bayesian spatial multivariate receptor models allow us to predict source contributions at any site (not just at monitoring sites) along with their uncertainty estimates, providing improved exposure estimates. Figure 13 contains the time-series plots of the predicted source contributions, along with their uncertainty estimates (95% posterior intervals), at an unmonitored site (denoted by a triangle) in Figure 12. This location is in a highly populated residential area with no available monitoring sites nearby. Air pollution epidemiologists or policy makers who develop air quality management plans may desire to know contributions of sources at such a location.

SUMMARY AND DISCUSSION

We have developed new statistical approaches to evaluate source-specific health effects associated with an unknown number of major sources of multiple air pollutants. The proposed methods effectively deal with model uncertainty in source apportionment by providing posterior model probabilities for a range of candidate models resulting from different numbers of sources and identifiability conditions, while accounting for parameter uncertainty that has been largely ignored in the previous assessments of source-specific health effects. The estimated posterior model probabilities can also be used as a basis for BMA, which is often used as a way of accounting for model uncertainty when competing models are nested (e.g., in variable selection in regression). We did not pursue BMA in this project because, in our case, models with different numbers of sources are not nested. For example, the source-composition and contribution parameters as well as the health-effects parameters under a 2-source model and a 3-source model do not have the same physical interpretations. Although we can estimate the posterior model probabilities for models with different numbers of sources, we cannot take an average over source-composition profiles, contributions, or the corresponding health-effects parameters estimated with different numbers of sources (i.e., BMA is not possible for models with different numbers of sources).

We discovered that the posterior model probabilities for competing models were comparable only if the models were very similar in nature, (i.e., having the same number of sources and choosing one set of prespecified zeros is not materially different from choosing another set). Although we selected a single model with the highest posterior model probability from the candidate models in the examples we considered in the project (because the models we compare are usually distinctive and posterior model probabilities were incomparable), it needs to be remembered that BMA is also possible when comparing similar models (i.e., models with the same number of sources and same source types).

The approach assuming normally distributed health outcome data (that may be used for continuous health outcome data or the count data with a large enough mean) was illustrated with $PM_{2.5}$ speciation data and cardiovascular mortality data from Phoenix. The results from our methods agreed in general with those from the previously conducted workshops and studies on PM source-apportionment and health effects for the Phoenix data, in terms of the number of major contributing sources, as well as estimated source profiles and contributions.

For the health effects of specific sources, there were similarities and dissimilarities. The health effects of Soil/Crustal Matter and Biomass/Wood Combustion were statistically insignificant both in Mar and colleagues (2006) and in our analysis. However, while Mar and colleagues (2006) identified adverse health effects for four source types (Sulfate at lag 0, Traffic at lag 1, Smelter at lag 0, and Sea Salt at lag 5, our analysis identified only two to be statistically significant (Smelter at lag 0 and Sea Salt at lag 5), which seems to be a natural consequence of incorporating uncertainty in the estimated source contributions into the health effects parameter estimation.

Computation of marginal likelihoods for Poisson models imposes enormous computational demands. Because of these computational demands, the approach assuming Poisson models for the health outcome data will be most beneficial in comparing only a few plausible models, rather than comparing numerous models, when the mean of the count data is very small, say less than 2, for which a normal approximation may not be well-justified.

Although not included in the report, our limited simulation indicated that even for a Poisson health outcome variable with a low mean (such as 1 or 2), the approach developed assuming the normal health outcome variable is also accurate in selecting the true model. We conjectured that it is because the likelihood of the air pollution data contributes much more to marginal likelihood (in comparing models with different numbers of sources and identifiability conditions for source apportionment) than does the likelihood of the health outcome data. (We speculate that the reason that the air pollution likelihood contributed more to the inference is because the number of pollution sources (q) is typically greater than one, so we expect that the health effects contributed $1/(q+1)$th of the modeling variation. This follows because the rank of the observed pollution data matrix of dimension T by J has rank q. The health model adds one more independent equation to the likelihood model.) It might not be the case if we also compared models with different baseline models (varying the forms of weather variables and long-term trends) related to health-outcome variables through marginal likelihoods (not just by sensitivity analysis), although we did not attempt that in this project.

Because there is no explicit relationship between the parameters for Poisson regression and those for normal linear regression, we did not pursue the approach for the normal health outcome variable further for the discrete health outcome data with a low sample mean. However, when the length of the time series is large, it may be computationally more advantageous to first use the approach for the normal health outcome variable to select a few of

the most plausible models from a rather large number of models. Then the method for the Poisson health outcome variable may be applied to select among those few competing models, while simultaneously obtaining health effects parameter estimates under Poisson models. Our limited observation was that estimation of the parameters for multivariate receptor models was minimally affected whether normal health outcome models or Poisson health outcome models were used.

The approach assuming Poisson models was applied to the $PM_{2.5}$ speciation data and respiratory mortality data from the 10-mile buffer region surrounding Clinton Drive and from Harris County. When the $PM_{2.5}$ data were fitted jointly with the health outcome data for the Clinton Drive area, five sources — that might represent Sulfate-Rich Secondary Aerosol, Motor Vehicles, Industrial Combustion, Soil/Crustal Matter, and Sea Salt — were identified as important sources associated with respiratory mortality. There was a statistically significant positive association only between same-day $PM_{2.5}$ concentrations attributed to one of the sources (that appears to be either Sulfate-Rich Secondary Aerosol or Industrial Combustion) and respiratory mortality. No other statistically significant association was observed for other source types or lags. Sign changes of the health-effects estimates across lags were also observed for some source types. Note that the length of the time series was very short for these data (only 215 days), and this might have contributed to insignificance of the health effects and their sign changes. Again, it needs to be emphasized that our approach to the evaluation of source-specific health effects incorporates uncertainty in the estimated source contributions into the estimation of source-specific health effects, while coping with model uncertainty that has not been addressed in previous studies on assessing source-specific health effects.

We have also developed a Bayesian spatial multivariate receptor modeling approach that can incorporate spatial dependence in the multivariate pollutant data collected from multiple monitoring sites into an estimation of source-composition profiles and a prediction of source contributions. This new approach can also take into account model uncertainty caused by the unknown number of sources and identifiability conditions, which has never been accounted for in previous spatial multivariate receptor modeling.

The spatially-enhanced multivariate receptor models enable predictions of source contributions at any site (whether monitored or unmonitored). These predicted source contributions, along with their uncertainty estimates, can greatly enhance air pollution epidemiologic studies and facilitate development of an effective air quality management plan by quantifying environmental impacts of pollution sources where no monitoring stations are available. Although we illustrated (in Appendix F) the use of estimated source contributions from spatially-enhanced receptor models to evaluate associations between source-apportioned pollutants and health outcomes based on subregion-level aggregated mortality data, our method can also be used to improve exposure estimates based on individual-level data because spatially-enhanced receptor models can predict source contributions (exposure estimates) at any location. For instance, our approach may be of great significance in air pollution epidemiology studies of exposure–health outcomes for a target population in a relatively small geographic area, especially when there is a strong spatial gradient in concentrations of pollutants from the source, as is the situation for many mobile-source traffic-related pollutants (Briggs 2005). We also illustrated an application using a GLM procedure in R as a two-stage approach rather than MCMC implementation, due to computational burdens of MCMC implementation that increases with the size of the data and the number of models to fit. However, in principle, estimation of spatially-enhanced receptor models and health effects parameters can be simultaneously performed as in the analyses near one monitoring site.

Extending our Bayesian spatial multivariate receptor models developed in this project to account for missing values in the air pollution data is ongoing. Also, in our spatial modeling, we assumed the independence of observations over time, which is typically satisfied when the data are measured at longer time intervals, such as every three or six days. When pollutants are measured at hourly intervals, temporal correlation often exists in the data. The spatial modeling approach presented in this project can be further generalized to account for spatiotemporal correlation in the data by considering temporal dependence structures for errors as in Park and colleagues (2001) or by including a temporal evolution equation for $\mathbf{G_t}$ such as $\mathbf{G_t} = \mathbf{G_{t-1}\Phi} + \mathbf{u_t}$ as in Calder (2003, 2007).

It needs to be kept in mind that as with all other model comparison approaches, there should be at least one good model (useful model if not the true model) in the set of models to be compared. It would not be meaningful to select the best model if none of the models compared are reasonable or realistic. The results are conditional on the set of models considered, and that is why exploratory analysis or good prior knowledge about the problem is so important, especially in selecting candidate models to be compared. As illustrated in this report, prior knowledge about likely sources or source types for the region, that can be obtained from previous studies or exploratory analysis,

needs to be utilized as much as possible in selecting a range of the number of sources and the positions of preassigned zeros.

While we did, for the first time, incorporate model uncertainty in source apportionment into source-specific health-effects evaluations, we did not address some other potentially important uncertainties such as uncertainty due to the choice of the confounder model or uncertainty due to data imputation. Although our method of computing marginal likelihoods and posterior probabilities can in principle be used to address uncertainty for the choice of the confounder model, we did not pursue it, as it is beyond the scope and the budget of the project. Note that the number of β parameters is not affected by the choice of the confounder model as opposed to being affected by the number of sources included, and BMA based on posterior probabilities (that can be computed by our method) for different confounder models can be used for the β parameter of interest to address uncertainty with the choice of the confounder model. We are continuing research on extending the Bayesian spatial multivariate receptor models developed in this project to account for missing values in the air pollution data.

Finally, it will be important to have rich interdisciplinary input into future implementation of the methods proposed in this research. It is reasonable to treat the examples in this report as preliminary (since they focused on demonstrating the statistical methods rather than ensuring scientific credibility) and recognize that actual applications will need meaningful collaboration from scientists in other disciplines, such as atmospheric science, to assure that the assumptions and results are scientifically credible.

ACKNOWLEDGMENTS

Other people who have contributed to this project through research collaboration or advice include Dr. Man-Suk Oh of the Department of Statistics at Ehwa Women's University, Dr. Philp K. Hopke of the Center for Air Resources Engineering and Science at Clarkson University, Dr. Inyoung Kim of the Department of Statistics at Virginia Tech University, and Dr. Jim Price of the Texas Commission on Environmental Quality. We gratefully acknowledge their help. We also thank Ms. Melanie Hotchkiss of the Texas Commission on Environmental Quality for her help with the acquisition of the Harris County air pollution data.

REFERENCES

Andersen ZJ, Wahlin P, Raaschou-Nielsen O, Scheike T, Loft S. 2007. Ambient particle source apportionment and daily hospital admissions among children and elderly in Copenhagen. J Expo Sci Environ Epidemiol 7:625–636.

Anderson TW. 1984. An Introduction to Multivariate Statistical Analysis (2nd ed.). New York:Wiley.

Bartholomew DJ, Knott M. 1999. Latent Variable Models and Factor Analysis. 2nd ed. New York:Oxford University Press.

Bell ML, Belanger K, Ebisu K, Gent JF, Lee HJ, Koutrakis P, et al. 2010. Prenatal exposure to fine particulate matter and birth weight: variations by particulate constituents and sources. Epidemiology 21(6):884–891.

Bell ML, Samet JM, Dominici F. 2004. Time-series studies of particulate matter. Annu Rev Public Health 25:247–280.

Billheimer D. 2001. Compositional receptor modeling. Environmetrics 12:451–467.

Briggs D. 2005. The role of GIS: coping with space (and time) in air pollution exposure assessment. J Toxicol Environ Health A. 68(13–14):1243–1261.

Buzcu B, Fraser MP. 2006. Source identification and apportionment of volatile organic compounds in Houston, TX. Atmos Environ 40:2385–2400.

Calder CA. 2003. Exploring latent structure in spatial temporal processes using process convolution, PhD Thesis, Duke University, Institute of Statistics and Decision Sciences (*https://stat.duke.edu/people/theses/CalderK.html*).

Calder CA. 2007. A dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. Environ Ecol Stat 14:229–247.

Carroll RJ, Ruppert D, Stefanski LA. 1995. Measurement error in nonlinear models. London:Chapman & Hall.

Christensen WF, Sain SR. 2002. Accounting for dependence in a flexible multivariate receptor model. Technometrics 44:328–337.

Christensen WF, Schauer JJ, Lingwall JW. 2006. Iterated confirmatory factor analysis for pollution source apportionment. Environmetrics 17:663–681.

Dawid AP. 1981. Some matrix-variate distribution theory: notational considerations and a Bayesian application. Biometrika 68: 265–274.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM Algorithm. J R Stat Soc Ser B 39:1–38.

Dominici F, Peng RD, Barr CD, Belle ML. 2010. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. Epidemiology 21(2):187–194.

Gajewski BJ, Spiegelman CH. 2004. Correspondence estimation of the source profiles in receptor modeling. Environmetrics 15:613–634.

Gent JF, Koutrakis P, Belanger K, Triche E, Holford TR, Bracken MB, et al. 2009. Symptoms and medication use in children with asthma and traffic-related sources of fine particle pollution. Environ Health Persp 117(7):1168–1174.

Gneiting T, Kleiber W, Schlather M. 2010. Matérn cross-covariance functions for multivariate random fields. J Amer Stat Assoc 105:1167–1177.

Heaton MJ, Reese CS, Christensen WF. 2010. Incorporating time-dependent source profiles using the Dirichlet distribution in multivariate receptor models. Technometrics 52:67–79.

Henry RC. 1997a. History and fundamentals of multivariate air quality receptor models, Chemometr Intell Lab Syst 37:37–42.

Henry RC. 1997b. Receptor model applied to patterns in space (RMAPS) part ii: apportionment of airborne particulate sulfur from Project MOHAVE. J Air Waste Manage Assoc 47:220–225

Henry RC, Kim BM. 1990. Extension of self-modeling curve resolution to mixtures of more than three components. Part I: finding the basic feasible region. Chemom Intel Lab Syst 8:205–216.

Henry RC, Park ES, Spiegelman CH. 1999. Comparing a new algorithm with the classic methods for estimating the number of factors. Chemometr Intell Lab Syst 48:91–97.

Higdon D. 1998. A process-convolution approach to modeling temperatures in the North Atlantic Ocean. Environ Ecol Stat 5:173–190.

Hopke PK. 1985. Receptor Modeling in Environmental Chemistry. New York:Wiley.

Hopke PK. 1991. An introduction to receptor modeling. Chemometr Intell Lab Syst 10:21–43.

Hopke PK. 2003. Recent developments in receptor modeling. J Chemom 17:255–265.

Hopke PK. 2010. The Application of Receptor Modeling to Air Quality Data. Pollution Atmosphérique Special Issue, Sept:91–109.

Hopke PK, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, et al. 2006. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. J Expo Sci Environ Epidemiol 16(3):275–286.

Ito K, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, et al. 2006. PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. J Expo Sci Environ Epidemiol 16:300–310.

Jun M, Park ES. 2013. Multivariate receptor models for spatially correlated multi-pollutant data. Technometrics 55:309–320.

Kim BM, Henry RC. 1999. Extension of self-modeling curve resolution to mixtures of more than three components. Part II: finding the complete solution. Chemom Intel Lab Syst 49:67–77.

Kim BM, Henry RC. 2000. Extension of self-modeling curve resolution to mixtures of more than three components. Part III: atmospheric aerosol data simulation study. Chemom Intel Lab Syst 52:145–154.

Laden F, Neas LM, Dockery DW, Schwartz J. 2000. Association of fine particulate matter from different sources with daily mortality in six U.S. cities. Environ Health Persp 108:941–947.

Lall R, Ito K, Thurston GD. 2011. Distributed lag analyses of daily hospital admissions and sourceapportioned fine particle air pollution. Environ Health Persp 119:455–460.

Lewis CW, Norris GA, Conner TL, Henry RC. 2003. Source apportionment of Phoenix $PM_{2.5}$ aerosol with the UNMIX receptor model. J Air Waste Manag Assoc 53:325–338.

Lingwall JW, Christensen WF, Reese CS. 2008. Dirichlet based Bayesian multivariate receptor modeling. Environmetrics 19:618–629.

Little RJA, Rubin DB. 1987. Statistical analysis with missing data. 2nd ed. New Jersey:Wiley.

Mar TF, Ito K, Koenig JQ, Larson TV, Eatough DJ, Henry RC, et al. 2006. PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of PM(2.5) and daily mortality in Phoenix, AZ. J Expo Anal Environ Epidemiol 16(4):311–320.

Mar TF, Norris GA, Koenig JQ, Larson TV. 2000. Associations between aire pollution and mortality in Phoenix, 1995–1997. Environ Health Persp 108:347–353.

Nikolov MC, Coull BA, Catalano PJ, Diaz E, Godleski JJ. 2008. Statistical methods to evaluate health effects associated with major sources of air pollution: a case-study of breathing patterns during exposure to concentrated Boston air particles. J R Stat Soc Ser C Appl Stat 57:357–378.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. 2006. An informative Bayesian structural equation model to assess source-specific health effects of air pollution. Harvard University Biostatistics Working Paper Series 46.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. 2007. An informative Bayesian structural equation model to assess source specific health effects of air pollution. Biostatistics 8:609–624.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. 2011. Multiplicative factor analysis with a latent mixed model structure for air pollution exposure assessment. Environmetrics 22:165–178.

Oh MS. 1999. Estimation of posterior density functions from a posterior sample. Comput Stat Data Anal. 29:411–427.

Oh MS, Park TS. 2002. Bayesian parameter estimation and variable selection in random effects GLM for count data. J Korean Stat Soc 31:93–107.

Ostro B, Tobias A, Querol X, Alastuey A, Amato F, Pey J, et al. 2011. The effects of particulate matter sources on daily mortality: a case-crossover study of Barcelona, Spain. Environ Health Persp 119:1781–1787.

Paatero P. 1997. Least squares formulation of robust, nonnegative factor analysis. Chemom Intell Lab Syst 37:23–35.

Paatero P, Tapper U. 1994. Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics 5:111–126.

Park ES, Guttorp P, Henry RC. 2001. Multivariate receptor modeling for temporally correlated data by using MCMC. J Am Stat Assoc 96:1171–1183.

Park ES, Guttorp P, Kim, H. 2004. Locating major PM_{10} source areas in Seoul using multivariate receptor modeling. Environ Ecol Stat 11:9–19.

Park ES, Henry RC, Spiegelman CH. 2000. Estimating the number of factors to include in a high-dimensional multivariate bilinear model. Commun Stat Simul Comput 29:723–746.

Park ES, Oh MS, Guttorp P. 2002b. Multivariate receptor models and model uncertainty. Chemometr Intell Lab Syst 60:49–67.

Park ES, Spiegelman CH, Henry RC. 2002a. Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models. Environmetrics 13:775–798.

Pollice A. 2009. Recent statistical issues in multivariate receptor models. Environmetrics; doi 10.1002/env.1021 (Early View).

Ramadan Z, Eickhout B, Song XH, Buydens LMC, Hopke PK. 2003. Comparison of positive matrix factorization and multilinear engine for the source apportionment of particulate pollutants. Chemom Intel Lab Syst 66:15–28.

Ramadan Z, Song XH, Hopke PK. 2000. Identification of sources of Phoenix aerosol by positive matrix factorization. J Air Waste Manag Assoc 50:1308–1320.

Seagrave J, McDonald JD, Bedrick E, Edgerton ES, Gigliotti AP, Jansen JJ, et al. 2006. Lung toxicity of ambient particulate matter from southeastern U.S. sites with different contributing sources: relationships between composition and effects. Environ Health Perspect 114:1387–1393.

Spiegelman CH, Park ES. 2007. A computation saving jackknife approach to receptor model uncertainty statements for serially correlated data. Chemom Intel Lab Syst 88:170–182.

Sullivan D. 2007. Research on $PM_{2.5}$ concentrations at Houston Clinton Drive CAMS 403. Final report submitted to the Texas Commission on Environmental Quality, The University of Texas at Austin (*www.h-gac.com/taq/airquality/raqpac/documents/2013/PM%20Advance%20Meetings/March%2011,%202013/Work%20order%206%20final%20report.pdf*).

Thurston GD, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, et al. 2005. Workgroup report: workshop on source apportionment of particulate matter health effects: intercomparison of results and implications. Environ Health Perspect 113:1768–1774.

Wolbers M, Stahel W. 2005. Linear unmixing of multivariate observations: a structural model. J Amer Stat Assoc 100(472):1328–1342.

HEI QUALITY ASSURANCE STATEMENT

The conduct of this research project was subject to independent quality assurance (QA) oversight by Abt Associates. The audit was led by Mr. Jose Vallarino, who has overseen QA programs for the past 20 years. The QA oversight consisted of a First QA Audit (focused on organizational structure, data quality of air pollution and health data, and human subjects' protection) and a Final QA Audit (focused on the Investigators' Final Report). The dates of the QA audits and activities are summarized below. Both audits were conducted on site at the Texas Transportation Institute (TTI) in College Station, Texas.

August 27–28, 2012. First QA Audit conducted on-site at TTI.

This "readiness review" audit was intended to review the standard operating procedures and data management practices used in the research to ensure that these procedures were followed consistently by all members of the research team. The auditors met with Dr. Park and all team members. The auditors observed the relevant Institutional Review Board documents from the University of Texas Health Science Center and Texas A&M University. The auditors made minor recommendations related to documentation of hand-checking procedures that had been performed.

May 15–16, 2014. Final QA Audit conducted on-site at TTI.

Dr. Park showed the auditors how approximately 15 tables and figures from the December 2013 version of the Investigators' Final Report were generated, starting with the raw data. (The tables and figures were preselected.) The auditors identified two minor discrepancies that could be readily addressed and did not impact the overall data quality or the findings of the report.

Overall, the auditors found the researchers to be well organized and cooperative during the audits. The study procedures, analysis steps, and data storage were systematic, consistent, and well designed for managing the various data and analytical streams that were necessary to complete the study.

Jose Vallarino, M.Sc.

APPENDICES AVAILABLE ON THE WEB

Appendices A–F contain supplemental material not included in the printed report. They are available on the HEI Web site *http://pubs.healtheffects.org*.

Appendix A: Literature Review on Studies that Evaluated the Effects of Short-Term Exposures to Air Pollution ($PM_{2.5}$ or VOCs) on Mortality (or morbidity) from Specific Cardiovascular Diseases and Respiratory Diseases

Appendix B: Summary of Studies Evaluating Health Effects Associated with Source-Apportioned Particulate Matter (PM)

Appendix C: Full Conditional Distributions for Parameters

Appendix D: Phoenix, Arizona Data

Appendix E: Database Development and Summary Statistics for Harris County Data

Appendix F: Application to the Harris County Mortality Data from Multiple Subregions

ABOUT THE AUTHORS

Eun Sug Park is a research scientist at the Texas A&M Transportation Institute (TTI). Her research spans statistics and environmental and transportation engineering. It focuses on modeling and analysis of various transportation and air pollution data, including multivariate receptor modeling, the health effects of multiple air pollutants, and the estimation of traffic-related air pollution. Prior to joining TTI, she was a research associate at the University of Washington's National Research Center for Statistics and the Environment. She holds a Ph.D. in statistics from Texas A&M University, an M.S. in statistics, and a B.S. in computer science and statistics, both from Seoul National University.

Elaine Symanski is an associate professor in the Division of Epidemiology, Human Genetics and Environmental Sciences at the School of Public Health, University of Texas. She is also director of the Southwest Center for Occupational and Environmental Health. With training in environmental sciences and epidemiology, her research expertise is in developing methods for assessing workplace and environmental exposures, especially as they relate to their application in health effects studies of vulnerable populations. Dr. Symanski earned an M.S.P.H. and Ph.D. from the School of Public Health at the University of North Carolina at Chapel Hill.

Daikwon Han is an associate professor of epidemiology at Texas A&M University. His research interests include the development of geospatial analysis and exposure assessment

methods in environmental health studies, social–environmental determinants of health such as issues of environmental justice and disparity, and life course influences on chronic health outcomes. He also serves as the director for the program for Geographic Information Systems and spatial statistics, which focuses on the development and application of spatial epidemiology methods in public health research and practice.

Clifford Spiegelman is a distinguished professor of statistics at Texas A&M University, where he has been since 1987. His interests include statistical applications to environmental models, transportation, forensics, and chemometrics. He is the chemometrics section associate editor for the Encyclopedia of Environmetrics.

OTHER PUBLICATIONS RESULTING FROM THIS RESEARCH

Park ES, Hopke PK, Oh MS, Symanski E, Han D, Spiegelman CH. 2014. Assessment of source-specific health effects associated with an unknown number of major sources of multiple air pollutants: a unified Bayesian approach. Biostatistics 15:484–497.

Jun M, Park ES. 2013. Multivariate Receptor Models for Spatially Correlated Multi-Pollutant Data. Technometrics 55:309–320.

ABBREVIATIONS AND OTHER TERMS

AIC	Akaike information criterion
BC	black carbon
BMA	Bayesian model averaging
cdf	cumulative distribution function
CNLS	constrained nonlinear least squares
COPD	chronic obstructive pulmonary disease
df	degrees of freedom
EC	elemental carbon
EM	algorithm with an E-step and an M-step
GLM	generalized linear models
IW	inverted Wishart
logMD	log of marginal likelihood
MCMC	Markov chain Monte Carlo
OC	organic carbon
OM	organic matter
PM	particulate matter
$PM_{2.5}$	PM ≤ 2.5 μm in aerodynamic diameter
PM_{10}	PM ≤ 10 μm in aerodynamic diameter
PMF	positive matrix factorization
PostP	posterior model probability
QA	quality assurance
RFA	request for applications
TTI	Texas A&M Transportation Institute
U.S. EPA	U.S. Environmental Protection Agency
VOC	volatile organic compound

Chemical Elements

Ag	silver
Al	aluminum
As	arsenic
Au	gold
Ba	barium
Br	bromine
Ca	calcium
Cd	cadmium
Cl	chlorine
Co	cobalt
Cr	chromium
Cs	cesium
Cu	copper
Fe	iron
Ga	gallium
Ge	germanium
Hg	mercury
I	iodine
K	potassium
La	lanthanum
Mg	magnesium
Mn	manganese
Mo	molybdenum
Na	sodium
Ni	nickel
P	phosphorus
Pb	lead
Pd	palladium
Rb	rubidium
Rh	rhodium
S	sulfur
Sb	antimony
Sc	scandium

Se	selenium
Si	silicon
Sn	tin
Sr	strontium
Te	tellurium
Ti	titanium
V	vanadium
W	tungsten
Y	yttrium
Zn	zinc
Zr	zirconium

CRITIQUE

Health Review Committee

HEI

Research Report 183, Parts 1 & 2, *Development of Statistical Methods for Multipollutant Research,* B.A. Coull et al. and E.S. Park et al.

INTRODUCTION

Air pollution is a complex mixture of gaseous, liquid, and solid components that varies greatly in composition and concentration across the United States and around the world owing to differences in sources, weather, and topography. Yet, in part due to a regulatory history focused on a limited number of criteria air pollutants, the necessary air pollution monitoring networks, scientific study designs, and statistical methods have not evolved as fully as they could to study this multipollutant atmosphere. The scientific community has long considered the possibility that the observed adverse health effects associated with individual pollutants may be attributable, in part, to the combined effects of multiple pollutants. A 2004 report from the National Research Council's Committee on Air Quality Management in the United States recommended that the U.S. Environmental Protection Agency (U.S. EPA*) take steps to address the presence of a complex, multipollutant atmosphere in the process for reviewing and setting National Ambient Air Quality Standards (National Research Council 2004). The U.S. EPA undertook that challenge in a series of workshops beginning in 2006.

The scientific challenges of understanding the health effects of exposure to the mixture of air pollutants that people actually breathe, estimating better the contribution of individual pollutants or their mixtures, and ultimately, addressing more cost-effectively the sources of those pollutants are substantial (Mauderly and Samet 2009). Conventional statistical methods are not well suited to deal with high correlations among pollutants, differences in the composition of pollutant mixtures over time and space, or differences in how accurately or precisely a person's exposure[s] to individual pollutants have been estimated. These factors can lead to errors (e.g., bias, incomplete accounting for uncertainty, or both) in the estimation of the health effects associated with individual pollutants or the joint contributions of multiple pollutants and the sources with which they may be associated.

The limitations of existing methods in dealing with these complexities have made it clear that advancing scientific understanding of multipollutant mixtures would require improved statistical methods. In response, HEI issued request for applications (RFA) 09-1, "Methods to Investigate the Effects of Multiple Air Pollution Constituents," which solicited research proposals that would address these methodologic difficulties through the development of innovative statistical methods. The RFA primarily sought applications in which existing statistical approaches (including those from fields outside of epidemiology) could be modified, extended, or combined, and then applied to a real-world exposure and health problem, rather than the development of purely theoretical approaches. (See the Preface for more detail on the scientific background for the RFA's development.)

Three studies were funded under RFA 09-1 and represent a variety of statistical approaches and data sets necessary to test them. The studies by Dr. B. A. Coull and Dr. E. S. Park and their associates are described in Parts 1 and 2 of this report (Research Report 183). The study by Dr. J. Molitor and his colleagues has been completed and is expected to be published in 2016.

Development of methods must typically follow a series of important steps before they can enter general use (see sidebar — Process of Statistical Methods Development and Evaluation). Any new method must first have a strong conceptual basis before being applied in data sets whose properties are well known. Simulated data sets that are designed to have specific properties amenable to testing by the proposed methods are often a first step, followed by application of the method in a real-world data set that is either relatively simple or has been well studied. Use of

Dr. Brent A. Coull's 2-year study, "Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents," began in January 2010. Total expenditures were $257,361. The draft Investigators' Report from Dr. Coull and colleagues was received for review in September 2013. A revised report, received in June 2014, was accepted for publication in July 2014.

Dr. Eun Sug Park's 2-year study, "Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants," began in May 2010. Total expenditures were $251,811. The draft Investigators' Report from Dr. Park and colleagues was received for review in February 2013. A revised report, received in December 2013, was accepted for publication in February 2014.

During the review process, the HEI Health Review Committee and both teams of investigators had the opportunity to exchange comments and to clarify issues in the Investigators' Reports and in the Review Committee's Critique.

* A list of abbreviations and other terms appears at the end of each Investigators' Report.

Process for Statistical Methods Development and Evaluation

- Formulate problem
- Develop conceptual framework for statistical models and specify parameters of interest including, where appropriate, development of statistical theory
- Write software to estimate parameters based on the conceptual framework
- Conduct preliminary tests in simulated data sets with known attributes
- Conduct preliminary tests in a well-studied data set, if available
- Test in a simplified real-world setting
- Test in a complex real-world setting
- Other investigators apply methods in settings that differ from those with which the method was developed and tested

these simpler data sets, whether simulated or real, often requires that a number of simplifying statistical assumptions be made. At any step, iterations of the design or its assumptions might be necessary to adjust the method before going forward. Only when the method behaves as expected a priori in simple settings is it time to move in a careful and systematic way to test it in more complex settings where more unknown factors may come into play.

The considerable work conducted by Drs. Coull and Park and their colleagues addressed the first several steps in the methods development process. An overview of the two studies, their goals, their use of simulation studies and applications in real-world data sets is provided in Critique Table 1. In its independent review of the two studies, the HEI Health Review Committee focused on a critical evaluation of the progress made by each of the investigator teams toward developing their methods, the quality and the limitations of the work completed, and what next steps might be needed to extend their work to more complex multipollutant settings.

This critique provides the HEI Health Review Committee's evaluation of the reports by Coull and Park. It is intended to aid the sponsors of HEI and the public by highlighting both the strengths and limitations of the studies and by placing the Investigators' Reports into scientific and regulatory perspective.

PART 1. STUDY CONDUCTED BY COULL AND COLLEAGUES

SCIENTIFIC BACKGROUND AND METHODS

Dr. Coull and his colleagues proposed Bayesian kernel machine regression (BKMR) methods designed to simultaneously address a number of goals in evaluating multipollutant exposures and associated health effects. These goals include selection of exposure variables, nonparametric estimation of possibly nonlinear exposure–response relationships, identification of interactions among pollutants (e.g., additivity or synergism), and quantification of uncertainty through Bayesian posterior distributions. These methods model the relationship between the exposure and a health outcome in a flexible manner that draws strength from the idea that subjects with similar exposures should have similar risks. This approach naturally assumes, like most other approaches, that the relationship between the exposures and the outcome is smooth. The design for these steps to be done simultaneously means that the methods, in essence, let the data drive some decisions in the modeling process that might otherwise be made on a more ad hoc basis by the analysts.

One of the early challenges the investigators faced was how to select the pollutants for inclusion in their models (i.e., variable selection). They explored two Bayesian approaches: (1) an approach using component-wise variable selection in which pollutants are essentially selected based on the strength of their individual associations with the health outcome of interest but in concert with the overall model fitting and inference steps; and (2) a hierarchical or multistep approach that first determines likely groupings of pollutants based on their correlations or common sources and then selects important individual pollutants within the groups. The investigators added the second approach because variable-selection methods, including the first approach, tend to have difficulty in cases where the variables exhibit a high degree of correlation. By grouping the variables (in this case, pollutants) into mutually exclusive sets according to a priori knowledge about their degree of correlation in the data or the commonality of their sources, the investigators created sets of pollutants that are approximately uncorrelated. Their method was then adapted so that, conditional on a group being selected to be in the model, only one pollutant in that group ultimately enters the model. This approach avoids the difficult statistical situation of having highly correlated predictors in the same model.

Coull and colleagues formally developed and tested their methods in three simulation studies that were

designed to evaluate different features of their approach. The simulations involved scenarios using various pollutant-interaction structures, nonlinear exposure–response relationships, and variables of different importance in the model (details are given in Critique Table 1). An important feature of their simulations was that their air pollution data sets were generated from actual $PM_{2.5}$ (particulate matter ≤ 2.5 µm in aerodynamic diameter) constituent data measured at the Harvard T.H. Chan School of Public Health Boston Supersite, thereby retaining the concentration profiles and the realistic joint distributions and correlations among the multiple pollutants. They then tested how their methods performed in each of the scenarios and compared their results with those found using the frequentist kernel machine regression (KMR) approach that has been proposed by other authors (Maity and Lin 2011).

The investigators next tested the methods using data from two studies that relied on the same Boston monitoring site used to generate data sets for the simulation studies. They evaluated (1) associations between changes in blood pressure and 7-day exposures to constituents of $PM_{2.5}$ in an epidemiologic study of patients 70 years of age and older — the MOBILIZE study (the Maintenance of Balance, Independent Living, Intellect, and Zest in the Elderly of Boston study; Wellenius et al. 2012); and (2) associations between blood pressure and heart rate in a toxicologic study conducted with dogs exposed for 5-hour periods either to filtered air or to concentrated ambient particles with known chemical composition (Bartoli et al. 2009). They compared their findings using the new methods with those from previously published studies that had used more standard linear, mixed models to analyze the same MOBILIZE and canine data.

HEI HEALTH REVIEW COMMITTEE'S CRITIQUE OF THE STUDY BY COULL AND COLLEAGUES

In its independent review of the study, the HEI Health Review Committee thought that the approach taken by the Coull investigative team addressed key problems in multipollutant research — the need to analyze large numbers of exposures or predictors of health outcomes whose interaction structure or exposure–response relationships with the outcome may be only vaguely known. Their approach therefore included both variable-selection approaches that are useful for identifying a small subset of exposures that are considered important in relation to the health effects of interest, as well as flexible nonparametric methods that are useful for allowing a wide range of possible exposure–response functions beyond the linear functions more typically assumed.

Throughout the research and review phases of their project, the investigators demonstrated a willingness to change course and improve upon their analyses. The investigators initially proposed a supervised clustering model that groups pollutants into categories or clusters based on several factors in the data that relate them to one another. However, in discussions with the HEI Research Committee, they realized that such an approach was better suited to large population-based studies using administrative databases in which a sufficient number of clusters could be identified, but it did not work as well in smaller-scale studies with small sample sizes, like theirs. The investigators are to be commended for realizing this limitation early on and taking an alternate (and ultimately more successful) route. Furthermore, in response to the initial review of the draft final report by the HEI Health Review Committee, the investigators conducted several additional analyses and other revisions that substantially improved the research.

One of the benefits of the BKMR described by Coull and colleagues is that it removes some of the vagaries of conventional data analysis methods used in air pollution and health research. Although statistical methods for performing variable selection or for estimating nonlinear exposure–response relationships already exist, their application to problems involving multidimensional exposures are typically ad hoc and require that many choices be made by the investigator, some of which may go undocumented. Particularly in a multidimensional setting, it is essential to have a statistical method that can simultaneously address exposure variable selection and the potential for nonlinear exposure–response relationships. In addition, ad hoc applications of conventional approaches generally do not account for the total uncertainty introduced and probably underestimate the uncertainty in the final health effects estimate. The approach developed by the investigators carefully brings all of these challenges under a Bayesian umbrella so that the uncertainties associated with both exposure variable selection and exposure–response estimation can be properly propagated to the health effects estimates.

The Committee thought that the three simulation studies conducted by the investigators to test their methods were well-crafted to represent scenarios that were also reflected in the data sets for the MOBILIZE and canine studies. A notable strength of the simulation studies was that the exposure data sets were generated from observed data and therefore reflected the kinds of correlation structures among pollutants that have been measured in the real world.

In their analyses of both the MOBILIZE study and the canine study data sets, Coull and colleagues did not find substantial evidence of interactions among the pollutants

Critique Table 1. Overview of Coull and Park Statistical Methods Studies

Investigator / Project Goals	Simulations	Applied Data Sets	Differences Between New Methods and Conventional Methods
Coull			
To develop flexible statistical methods to estimate the joint effects of multiple pollutants, while allowing for potential nonlinear or nonadditive associations between a given pollutant and the health outcome of interest To apply variable selection using Bayesian kernal machine regression (BKMR): (a) component-wise variable selection, and (b) hierarchical variable selection To compare results with those of frequentist KMR variable-selection methods	Simulated exposure data sets by sampling directly from actual exposure data (the Harvard T.H. Chan School Boston Supersite multipollutant data set) rather than by a set of characteristics specified by investigators. Exposure–response relationships were assumed. **(1)** ***Developed component-wise variable selection methods for a moderate number of pollutants.*** Generated data sets for (a) three metals (As, Mn, and Pb) and (2) nine PM constituents (Al, S, Ni, BC, Cu, Zn, Mg, K, and Cl) where only one or two were assumed to be known as causal. Assumed three different exposure–response relationships. Assumed two signal-to-noise ratios (high, realistic). **(2)** ***Compared component-wise and hierarchical variable selection methods in high-correlation settings.*** Generated data sets for 13 PM constituents with moderate to high correlations (Al, Si, Ti, Ca, Ni, V, Zn, S, BC, Cu, K, Cl, and Mn). Assumed the same three exposure–response relationships and the realistic signal-to-noise ratio **(3)** ***Tested ability to identify source category–specific health effects.*** Based on prior source-apportionment analyses of the Harvard T.H. Chan School Boston Supersite multipollutant data set, modeled exposures to 14 constituents (Al, S, Ni, BC, Na, Cu, Zn, V, Ti, Ca, Mg, K, Cl, and Si) representing six source categories, only one of which was assumed to be associated with adverse health outcomes.	**(1)** ***MOBILIZE prospective cohort study.*** Estimated associations between short-term (7-day) exposures to seven $PM_{2.5}$ constituents (Ni, Cu, Zn, S, Ti, Mn, and BC) and blood pressure in healthy, aging, human subjects, 2005–2008. **(2)** ***Canine toxicologic studies at the Harvard T.H. Chan School.*** Estimated associations between exposure to concentrated ambient particles (CAPs) and blood pressure and heart rate in dogs.	• Implements automatic model selection • Accounts for uncertainty in model selection • Allows for nonlinear, nonadditive exposure–response relationships • Allows for complex interactions among pollutants • Identifies important pollutants within the mixture via inclusion probabilities • BKMR variable-selection approach allows for "supervision" of selection by health data; has not been explored in environmental health

Table continues next page

Critique Table 1 (*Continued*). Overview of Coull and Park Statistical Methods Studies

Investigator / Project Goals	Simulations	Applied Data Sets	Differences Between New Methods and Conventional Methods
Park			
To develop a multipollutant approach that accounts for both model uncertainty in multivariate receptor models and uncertainty in estimated source-specific exposures in assessing source-specific health effects	Generated synthetic data sets. **(1) *Assumed four sources that have known tracer elements.*** Evaluated ability of model to estimate health effects given uncertainty about the number of sources (up to five) and the correlations among source contributions. **(2) Extended previous simulation to include comparison of two methods.** Method 1 assumed a priori correlated source contributions; Method 2 assumed a priori uncorrelated source contributions. **(3) *Assumed sources that did not rely on tracer pollutants.*** Varied the number of sources and identifiability conditions. **(4) *Conducted simulations that assumed either normal- or Poisson-distributed health outcomes.***	**(1) *Phoenix, AZ.*** Time series of daily $PM_{2.5}$ speciated data and counts of cardiovascular disease mortality from 1995–1997. **(2) *Houston, TX.*** Time series of 24-hour $PM_{2.5}$ speciated data and counts of respiratory mortality from 2002–2005.	• Joint modeling estimates parameters in the multivariate receptor models at the same time as estimating health effects parameters. • Incorporates source apportionment into a time-series health effects analysis using a Bayesian hierarchical modeling framework. • Quantifies model uncertainty caused by the unknown number of sources and identifiability conditions in source apportionment. • Accounts for uncertainty in source-specific exposures in the health effects parameter estimates.
To develop enhanced spatial multivariate receptor models that can account for spatial correlations in the multipollutant data collected from multiple monitoring stations	***Assumed nine monitoring locations and three underlying sources.*** Eight locations were used for model fitting and one was used for model performance. Sources were assumed a priori to be correlated in one case, uncorrelated in another.	**(1) *Houston, TX.*** 24-hour ambient concentrations of 17 VOCs measured at nine monitoring sites from 2000–2005.	• Implements Bayesian multivariate receptor models. • Incorporates time-series data from multiple monitoring locations. • Evaluates impact of correlations among sources contributions. • Demonstrates that source contributions can be predicted at an unmonitored location.

or nonlinearity in the exposure–response relationships, either because they do not exist or because there was insufficient power in the studies to detect them. However, as the investigators noted, they did not know in advance whether nonlinearity or an interaction structure might exist. Conventional data analysis approaches would have needed to cycle through a series of models to test whether interactions were present, which would have raised issues of multiple testing and possibly false-positive findings. Although they might have led to the same results achieved by these investigators, the numerous steps might be difficult for others to reproduce. Ultimately, data analysis can never be fully automated, and sound judgment by investigators always plays an important role; as problems become more multidimensional, methods like the one the investigators propose are important tools for minimizing the ad hoc nature of the process.

Despite the lack of the hoped-for complexity of the exposure–response relationships in the MOBILIZE study data set and the canine study data set, the investigators analyses did demonstrate that the new methods were practical to apply to real data sets and that they produced results that were largely sensible. It is arguably unreasonable to expect that the first application of new statistical methods would necessarily identify groundbreaking scientific findings. Rather, in this context, it is important that the applications contribute to an understanding of the methods' theoretical properties and how they perform on real-world data sets. Coull and colleagues have amply achieved both of these goals and their findings suggest that their methods need to be applied in a greater range of scenarios before we can ultimately evaluate their usefulness in real-world practice.

When applying statistical methods, typically a tradeoff is made between the number of assumptions about the data and the computational complexity of the method. These investigators' approach is quite flexible and allows for a wide class of model structures, different statistical distributions, and choices of variables. However, the approach is quite a bit more computationally demanding than standard approaches. Coull and colleagues report that software to implement the method will soon become available; the Committee thought it will be a welcome addition to the statistical toolbox. The availability of user-friendly software that is efficiently implemented is a key component of the wider adoption of this approach beyond the statistical community.

PART 2. STUDY CONDUCTED BY PARK AND COLLEAGUES

SCIENTIFIC BACKGROUND

Dr. Park and her colleagues focused on improvements to multivariate receptor models, the class of models used to identify the number of pollution sources, to characterize the pollutant composition profiles associated with each source, and to estimate the contributions of each source to exposure experienced by the "receptor", which in practice is the monitoring location, but in concept represents exposed individuals. The results of the receptor models can then be input to epidemiologic models in which the contribution of different sources to health outcomes can be evaluated.

The Park study had two specific aims: (1) to develop a multipollutant approach that, in the estimation of source-specific health effects, accounts for both model uncertainty in multivariate receptor models and uncertainty in the source-specific exposures; and (2) to develop enhanced spatial multivariate receptor models that can account for spatial correlations in the multipollutant data collected from several monitoring stations. Typically, multivariate models are applied to data from one monitoring station or receptor and the output of the model (number of sources and strength of each source's contribution) are treated as certain. The investigators aimed to account fully for uncertainties in the numbers and contributions of the sources identified, and in the health effects associated with them, by addressing more explicitly the implications of key assumptions in the receptor modeling choices (i.e., model uncertainty). In contrast to Coull's approach, Park did not allow for flexibility in the shape of the exposure–response relationship in her models.

For each of the specific aims, Park and colleagues evaluated the methods they developed in two steps. First, they used simulation studies in which all of the distributions, data structures, and parameters to be estimated were specified by the investigators so the "truth" is known. This kind of evaluation provides an important first step in assuring that the methods are performing properly because the estimated parameters can be compared with their "true" values. Second, they applied their methods to three existing data sets, two to address Aim 1, and one to address Aim 2. Critique Table 1 gives a simple overview of the project and its comparison with the work by Coull and colleagues; details are provided below.

SPECIFIC AIM 1

Most time-series studies of health effects reported in the air pollution epidemiology literature used a two-stage approach in which exposures are estimated first and the results input to the health effects analysis. In their first specific aim, Park and colleagues extended existing methods by developing a joint modeling approach that estimated parameters in the multivariate receptor models at the same time as the health effects parameters. Specifically, they incorporated source apportionment into a time-series health effects analysis using a Bayesian hierarchical modeling framework. Their focus was on estimating the q source-specific health effects parameters, where q is the number of sources. In this Bayesian framework, prior distributions are specified for the parameters in the health model and for latent source contributions in the multivariate receptor models. Multivariate receptor models are used to decompose a time series of multipollutant measurements (e.g., matrix **X**) into a time series of mass contributions from multiple sources (e.g., matrix **A**) multiplied by a matrix of relative contributions of each pollutant to each source (e.g., matrix **P**) plus a matrix of errors (e.g., matrix **E**).

Multivariate receptor models are in general not "well identified", even when the number of sources is known. That is, there is not a unique set of values that make the equations in the model true. In the example given above, there are infinite combinations of the values of **A** (source contributions) and **P** (source composition profiles) that can satisfy the equality presented by the model. Thus, one of the key challenges is to impose constraints or assumptions (referred to in the report as identifiability conditions) that both make the problem analytically tractable and are scientifically meaningful. The identifiability conditions for sources include (1) specifying which pollutants do not contribute to a specific source, (2) assuming that this set of noncontributing pollutants is not identical for any two sources, and (3) assuming that the relative contributions across all pollutants sum to 1 for each source.

Plausible models for these identifiability conditions can be based on preliminary data analyses or estimates culled from the literature. Estimation depends upon the assumed number of sources and the identifiability conditions selected. One of the important contributions of Park's previous work (Park et al. 2002) was to account for the uncertainty in both of these assumptions through a Bayesian approach to modeling, allowing for straightforward comparisons between model fit under different assumptions through posterior probability estimates. (The choice of the number of sources and identifiability conditions is specified through prior distribution assumptions in Bayesian modeling.)

In this first aim, Park and colleagues extended their previous work to (1) explicitly incorporate an a priori assumption of correlated source contributions, and (2) expand the modeling framework to also estimate the health effects parameters while accounting for the uncertainty in estimating the source contributions. This approach was implemented using a Markov chain Monte Carlo procedure by evaluating the joint models for a specific number of sources and set of identifiability conditions and then comparing the resulting log of the marginal likelihoods (and posterior model probabilities).

Evaluation of Aim 1 Methods

In the first simulation Park and colleagues based their framework on a previously published simulation study that used four known sources of particulate matter pollution in Boston, with some limited modifications to better understand the health effects parameter estimation. By doing so, they provided a comparison between the results of their more complex method with those of a more conventional and simpler one (i.e., Nikolov et al. 2006, 2007). In this case, they assumed that a tracer pollutant is known to exist for each source (this use of tracer pollutants is an alternative approach to addressing the model identifiability challenge). The investigators then ran simulations with scenarios that allowed for one to five possible sources and allowed for correlations among the source contributions. They found that the estimated source-specific health effects estimates were consistent with the truth; that is, the posterior credible intervals for each parameter contained the true value approximately 96% of the time.

In the second simulation, they extended the first simulation to determine the performance of a model that assumes a priori that source contributions are uncorrelated when the underlying source contributions were actually correlated. They then compared these results with those of the first model that assumed a priori that source contributions are correlated. They found little difference in the results.

In a third simulation, Park and colleagues used a more complex model to identify the underlying sources that did not rely on tracer pollutants, and investigated the impact of misspecification of zeros in the source composition matrix. They considered several different combinations of the number of sources and assumed identifiability conditions; they then studied the performance of their models under those sets of assumptions. They found that their estimates of source-specific health effects were not very sensitive to the misspecification of pollutants as absent species (zeros) for particular sources as long as those pollutants were not actually important contributors to a given source. They also found that the 95% posterior credible

intervals for the estimated health effects parameters contained the true value 96% of the time. These findings held when the health outcomes were modeled using either continuous (normal) distributions or discrete (Poisson) distributions, although the investigators noted that the Poisson models took considerably longer to compute. Thus, based on the simulation studies, the investigators found that they could obtain reasonable results using the methods developed in the first aim.

Evaluation of Aim 1 Using Real-World Data Sets

To evaluate the methods developed in the first aim from a more practical perspective, Park and colleagues used data sets from Houston, Texas, and from Phoenix, Arizona. The Houston data included 24-hour $PM_{2.5}$ speciation data from 2002–2005 in a region near the Houston Ship Channel and actual counts of respiratory deaths in the region. The Phoenix data included daily $PM_{2.5}$ speciation data from 1995–1997 and daily counts of cardiovascular deaths. The Phoenix data had been used in studies previously published by others (Hopke et al. 2006; Mar et al. 2006; Thurston et al. 2005). For sake of brevity, we focus on the Phoenix results to illustrate the evaluation of her methods.

The Phoenix data illustrated some practical problems faced in these real-world applications, namely that many of the 46 pollutant species had negative concentration estimates, and data for various species in the data set were missing on several days. The investigators addressed these challenges by reducing the number of species considered in their models to 15 and by imputing missing measurements for species or confounding variables (temperature, relative humidity). They applied models with different lag days and numbers of sources and compared the results. The model selected had six sources and gave results that were consistent with observations reported by Mar and colleagues (2006). However, the uncertainties of the health effects estimates tended to be larger, which reflected the more comprehensive accounting for uncertainty in this work.

SPECIFIC AIM 2

In the second aim, the investigators' goal was to extend multivariate receptor models to incorporate data from multiple spatial locations and to allow for spatially dependent data among locations. This approach is in contrast to more conventional multivariate receptor model analyses with time-series data that typically do neither. It also allows source contributions to be predicted at locations where monitoring data are not collected.

In more specific terms, Park and colleagues extended existing methods through adapting dynamic factor process convolution models (which are versions of multivariate spatial temporal process convolution models that require fewer investigator assumptions about underlying spatial correlations than do conventional geostatistical models) to a Bayesian framework where assumptions about the number of sources and identifiability conditions were relaxed and the standard nonnegativity constraints commonly included in source apportionment were applied. Unlike the first aim, this aim focused entirely on source apportionment without also incorporating estimation of health effects. This extension of their method also has the advantage that it can handle more complex data and model structures than a conventional geostatistical modeling approach. As in the investigators' first aim, the uncertainty in the number of sources and identifiability conditions was addressed by comparing posterior model probabilities.

Evaluation of Aim 2 Using Simulations

Park and colleagues used two types of simulations to test the spatially dependent multivariate receptor modeling approach. For all simulations, they generated the data with three underlying sources (of which contributions are correlated with one another) from nine monitoring locations in a multipollutant framework. In the first simulation, they performed estimation assuming a priori that the contributions from three sources were correlated with one another, and in the second simulation they performed estimations assuming a priori that the contributions from three sources were uncorrelated. In both simulations, eight monitoring locations were used for model fitting and a ninth (the "unmonitored" site) was evaluated for model performance. For both methods, the estimated number of sources was the same as the true number of sources in every simulation, and the predicted source contributions at the "unmonitored" site were highly correlated with the true source contributions.

Evaluation of Aim 2 Using Real-World Data Sets

Park and colleagues evaluated the performance of their enhanced multivariate receptor model using monitoring data for volatile organic compounds (VOCs) measured every six days between 2000 and 2005 from nine monitoring sites in Harris County, near Houston, Texas. Based on knowledge of local sources from previous studies of the area, of the possible 107 VOCs measured they selected for analysis a subset of 17 as being relevant to the likely local sources. They constructed a range of candidate models to evaluate, varying the number of sources from four to seven and varying the assumptions about identifiability conditions. As in the first

specific aim, the investigators implemented the models using a Markov chain Monte Carlo procedure and then compared their resulting log of the marginal likelihoods and posterior model probabilities. For this data set, the model with the highest posterior probability was one with five VOC sources that were consistent with those found in earlier studies of the area. Park and colleagues then demonstrated how their spatial multivariate receptor model might be allowed to predict source contributions in a populated area for which a monitoring site was lacking.

HEI HEALTH REVIEW COMMITTEE'S CRITIQUE OF THE STUDY BY PARK AND COLLEAGUES

In its independent evaluation of the study, the HEI Health Review Committee thought that Park and her colleagues had tackled an extremely challenging technical problem yet produced work that represents a meaningful extension of source-apportionment estimation to jointly estimate health effects (Aim 1) and to predict source contributions at new locations (Aim 2). Throughout the review process, Park worked collaboratively with the Committee to make her work clearer and more accessible to a broader audience.

Estimating health effects from pollutant sources has been of great interest, and previous methods have been limited by their inability to account for uncertainties in the multivariate receptor modeling in the health effects estimation. The investigators' work advances the statistical methods for source apportionment in time-series health effects studies, by both (1) estimating source-related health effects in which the uncertainty in the sources and source contributions is properly accounted for in the models, and (2) extending multivariate receptor models to incorporate dependence among source contributions.

The Committee also concluded that the methods described in the report were properly developed and tested. The simulations demonstrated good performance under the range of conditions evaluated; that is, the correct number of sources was selected and the 95% posterior intervals for the health effects parameters were close to their nominal, or assumed, values. The caveat that needs to be stated for this work, as for all simulation studies, is that the simulations provide evaluations of only a defined number of conditions; thus generalizations about the performance of the methods under other conditions need to be made cautiously.

The applications of the methods to real-world data also lend credibility to the conclusions that the methods appear to work as the investigators intended. Their analyses to replicate the findings from Mar and colleagues (2006) generally succeeded in doing so, albeit with fewer statistically significant findings for the health effects. The Committee agreed with the investigators that fewer such findings would be expected from a method that incorporates uncertainty in the receptor model into the health effects estimates. Overall, the authors have reached conclusions from their data analysis that are appropriate within the context of an exercise in methods development.

One of the key challenges faced by the investigators was that the joint modeling framework was not as straightforward when discrete health outcome data were used, such as the daily mortality counts in time-series studies, as it was when health outcomes were modeled using continuous (normal) distributions for which they first developed their methods and software. When extending the methods to a discrete outcome modeled using Poisson distributions, they needed to cope with additional complexity because the conditional densities were not all conjugate (i.e., in the same family of distributions) and the resulting full conditional posterior distributions could not be written in closed form (and, for example, solved analytically). They solved this problem by introducing normal auxiliary variables into the model in order to obtain tractable full conditional densities that could be estimated using Markov chain Monte Carlo procedures. However, this entire approach was computationally very complex and, as their report states, "coding of the algorithm turned out to be a formidable task, and took a considerable amount of time and effort."

The Committee thought that the spatial multivariate receptor modeling method, developed in the second specific aim to handle time series of concentration data from multiple locations, appeared to be a useful innovation. From the simulation and from the application of the method to the VOC data, it was apparent that source contributions can be predicted at unmonitored locations. However, for this method to gain widespread scientific applicability, the Committee thought that future work should include an additional explanation for nonstatisticians on the use of the method to deal with correlation among source contributions at different locations.

The challenge for implementing both of these joint models for health effects estimation and the spatial multivariate receptor models for prediction of sources at unmonitored locations is that they require data in a specific form, which is often not available in existing data sets. For example, the spatial multivariate receptor model requires complete multipollutant and speciation data from multiple monitoring sites. In reality, relatively few locations exist where multiple sites within the same airshed collect detailed data on PM constituent species or on VOCs. In order to show proof-of-concept, a number of

imputation and data-massaging steps were necessary to get the data into the necessary form before analysis. In order to use the Phoenix data, for example, the investigators needed to reduce the number of pollutants they analyzed and impute missing data for some of the remaining species as well as for the weather variables. Such steps are not ideal, but are not uncommon for real-world data. However, the potential impacts of these assumptions on outcomes from the analysis were not studied and could be important.

A more complex question relates to the potential for feedback in joint receptor and disease modeling when either the health outcome or the exposure data are unbalanced or the models are poorly specified. That is, to the extent that the health outcome data influences the choice of sources, and the choice of sources affects the estimates of source-specific health effects, it is possible that biases can develop in the joint model estimates that are artifacts of the quality of the underlying data. A related conceptual issue is whether, when one is interested in multiple health outcomes, the source apportionment should be the same or different for each health outcome. A single joint model would likely produce slightly different source apportionments for each outcome, though some scientists might argue that they should be the same.

Problems with underlying data also affect estimation in more conventional two-stage models. However, in recent work in the air pollution literature dealing with error in exposure measurement, investigators have intentionally chosen to model exposure first, and then correct explicitly for the impact of the exposure modeling error on health effect inferences (see, for example, Gryparis et al. 2009; Szpiro et al. 2011; Szpiro and Paciorek 2013). The two-stage approach not only avoids the potential bias issues due to feedback but also reduces the computational challenges in joint approaches. More work needs to be done to explore the feedback question in joint modeling and to understand the larger question of how best to estimate regression parameters for inference about health in complex settings such as the one tackled in this research.

The test cases in Phoenix and Houston were an important first step but their results provide some indications that further development work is necessary. For example, the Committee noted that the Phoenix test case identified a preferred model (with posterior probability of 1) with seven sources that accounted for only about 80% of emissions and that appeared to be missing some important sources associated with the Phoenix area (e.g., secondary organic aerosols and secondary nitrates). The Houston test case identified only five sources within a complex area whereas other studies have more recently identified as many as ten (Sullivan et al. 2013). Although these differences suggest the need for more comparisons with other source-apportionment studies, they may also reflect differences in the questions that the studies' methods were designed to address; for example, if they were designed to identify all sources or just those sources associated with health effects.

In summary, the Committee thought the new methods from Park and associates were properly implemented; the Committee recognized the inherent validity in advancing statistical methods, even when the details of the application are simplified or in other ways altered to make the problem tractable. To this extent, the work led by Park has raised important scientific issues with existing methods and she and her team have developed new approaches to address them. The Committee did not expect that this work alone would resolve how well the methods will work in practice in more complex settings. An important philosophical question for future investigators considering joint estimation approaches is to how to evaluate and weigh carefully the scientific reasons for and against doing so.

SUMMARY AND CONCLUSIONS

HEI issued RFA 09-1 to encourage the development of new statistical methods to address the challenges investigators have long faced in characterizing health risks from multipollutant atmospheres. Conventional statistical methods are not well suited to deal with high degrees of correlation among pollutants, differences in the composition of pollutant mixtures over time and space, or differences among pollutants in how completely or representatively a person's exposure to individual pollutants has been measured. Consequently conventional methods can lead to errors (e.g., bias, incomplete accounting for uncertainty, or both) in the estimation of the health effects of individual pollutants or in the estimation of the joint contributions of multiple pollutants and the sources with which they may be associated.

The two studies discussed in this Critique each addressed important but separate questions in multipollutant research. The study by Coull and colleagues developed a BKMR method that simultaneously performs exposure variable selection, allows for nonparametric estimation of nonlinear exposure–response relationships, and allows for quantification of uncertainty through Bayesian posterior distributions. The study by Park and colleagues set out to develop a statistical approach to analyze multipollutant time-series data that incorporated different sources of uncertainty in source-apportionment models in estimating

source-specific health effects. The Committee recognized two important distinctions between the two studies. First, Coull and colleagues focused on dimension reduction (or reduction of the complexity of the exposure data to the most influential pollutants) to simplify the analysis, whereas Park and colleagues used all of the available exposure data, in some cases from multiple monitoring locations, to choose and characterize sources. Second, Coull and colleagues partitioned pollutant exposures into mutually exclusive groups, whereas Park and colleagues estimated sources while allowing any single pollutant to belong to multiple sources.

Despite the differences in the specific questions they set out to answer, the two projects share some similarities in their modeling approaches. Both teams used Bayesian statistical frameworks to estimate health effects using joint, or supervised, health and exposure models. The objective of such approaches is to evaluate the set of exposure conditions that are associated with the greatest health risk estimates at the same time. Both teams followed appropriate processes expected of statistical methods development (see sidebar), whereby each began by developing a solid conceptual basis for their methods, and then tested their methods first in simulation studies using data sets with known properties and next in real-world data sets that were either relatively simple or previously studied.

Both sets of methods are complex and computationally demanding. They may be even more demanding when tested in even more complex exposure environments (although the Houston data set used by Park and colleagues is already quite complex). Neither project could answer all the questions that RFA 09-1 posed; for example, they were unable to characterize more directly the interactions between individual pollutants or to evaluate nonlinearities in exposure–response relationships.

From the standpoint of methods development, further work is necessary to evaluate the methods proposed in these reports. They need to be applied in a broader range of settings representing different types of sources, components, and levels of data complexity. These newer models need to be compared side-by-side against the more conventional two-stage approach in air pollution epidemiology, in which exposure is modeled first and those results are used in a health model in which exposure measurement error is explicitly considered. Future work should investigate the potential for feedback in the joint modeling approaches proposed by both investigative teams. Such analyses could help to determine whether the additional complexity of these new methods will lead to better understanding of how pollutant mixtures and their sources may contribute to effects on human health and, ultimately, to better decisions about how to control them. In any future extension or evaluation of these methods, the active involvement of subject-matter experts will be important to keep the statistical methods well-tuned to scientific needs and to realistic interpretation of the results.

Ultimately, the 2006 U.S. EPA workshops on multipollutant science recognized that statistical methods alone would likely not be sufficient to address the many issues in characterizing multipollutant exposures and health effects. "We can't model our way out of it; we can't measure our way out of it…." was the refrain repeated throughout the workshop. Further advances are necessary in monitoring and modeling the spatial and temporal variability in components of the air pollution mixture, in characterizing their related human exposures, and in statistical methods to deal with the inevitable and ongoing uncertainties that remain.

ACKNOWLEDGMENTS

The Health Review Committee thanks the ad hoc reviewers for their help in evaluating the scientific merit of the Investigators' Reports. The Committee is also grateful to Drs. Kate Adams and Sumi Mehta for their oversight of the studies, to Drs. Katherine Walker and Aaron Cohen for their assistance in preparing its Critique, to Virgi Hepner and Carol Moyer for science editing of the Investigators' Reports and the Critique, and to Hope Green, Fred Howe, and Ruth Shaw for their roles in preparing this Research Report for publication.

REFERENCES

Bartoli CR, Wellenius GA, Coull BA, Akiyama I, Diaz EA, Lawrence J, et al. 2009. Concentrated ambient particles alter myocardial blood flow during acute ischemia in conscious canines. Environ Health Perspect 117:333–337.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10: 258–274.

Hopke PK, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, et al. 2006. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. J Expo Sci Environ Epidemiol 16:275–286.

Maity A, Lin X. 2011. Powerful tests for detecting a gene effect in the presence of possible gene–gene interactions using Garrote kernel machines. Biometrics 67:1271–1284.

Mar TF, Ito K, Koenig JQ, Larson TV, Eatough DJ, Henry RC, et al. 2006. PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily mortality in Phoenix, AZ. J Expo Sci Environ Epidemiol 16:311–320.

Mauderly JL, Samet JM. 2009. Is there evidence for synergy among air pollutants in causing health effects? Environ Health Perspect 117:1–6. doi:10.1289/ehp.11654.

National Research Council. 2004. Research Priorities for Airborne Particulate Matter. IV. Continuing Research Progress. Washington, DC:National Academy Press.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. 2006. An informative Bayesian structural equation model to assess source-specific health effects of air pollution. Harvard University Biostatistics Working Paper Series 46.

Nikolov MC, Coull BA, Catalano PJ, Godleski JJ. 2007. An informative Bayesian structural equation model to assess source specific health effects of air pollution, Biostatistics 8:609–624.

Park ES, Oh MS, Guttorp P. 2002. Multivariate receptor models and model uncertainty. Chemom Intell Lab Syst 60:49–67.

Sullivan DW, Price JH, Lambeth B, Sheedy KA, Savanich K, Tropp RJ. 2013. Field study and source attribution for $PM_{2.5}$ and PM_{10} with resulting reduction in concentrations in the neighborhood north of the Houston Ship Channel based on voluntary efforts. Air Waste Manag Assoc 63:1070–1082.

Szpiro AA, Paciorek CJ. 2013. Measurement error in two-stage analyeses, with application to air pollution epidemiology. Environmetrics 24:501–517.

Szpiro AA, Sheppard L, Lumley T. 2011. Efficient measurement error correction with spatially misaligned data. Biostatistics 12:610–623.

Thurston GD, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, et al. 2005. Workgroup report: workshop on source apportionment of particulate matter health effects: intercomparison of results and implications. Environ Health Perspect 113:1768–1774.

U.S. Environmental Protection Agency. 2006. Workshop on Interpretation of Epidemiologic Studies of Multipollutant Exposure and Health Effects. Federal Register/ Vol. 71, No. 225 / Wednesday, November 22, 2006. *http://docs.regulations.justia.com/entries/2006-11-22/E6-19806.pdf*.

Wellenius GA, Wilhelm-Benartzi CS, Wilker EH, Coull BA, Suh HH, Koutrakis P, et al. 2012. Ambient particulate matter and the response to orthostatic challenge in the elderly: the Maintenance of Balance, Independent Living, Intellect, and Zest in the Elderly (MOBILIZE) of Boston study. Hypertension 59:558–563.

RELATED HEI PUBLICATIONS: MULTIPLE POLLUTANTS, SHORT-TERM STUDIES, AND METHODS

Number	Title	Principal Investigator	Date*
Research Reports			
178	National Particle Component Toxicity (NPACT) Initiative Report on Cardiovascular Effects	S. Vedal	2013
177	National Particle Component Toxicity (NPACT) Initiative: Integrated Epidemiologic and Toxicologic Studies of the Health Effects of Particulate Matter Components	M. Lippmann	2013
176	Effect of Air Pollution Control on Mortality and Hospital Admissions in Ireland	D.W. Dockery	2013
175	New Statistical Approaches to Semiparametric Regression with Application to Air Pollution Research	J.M. Robins	2013
171	Multicity Study of Air Pollution and Mortality in Latin America (The ESCALA Study)	I. Romieu	2012
170	Impact of the 1990 Hong Kong Legislation for Restriction on Sulfur Content in Fuel	C-M. Wong	2012
169	Effects of Short-Term Exposure to Air Pollution on Hospital Admissions of Young Children for Acute Lower Respiratory Infections in Ho Chi Minh City, Vietnam	HEI Collaborative Working Group	2012
161	Assessment of the Health Impacts of Particulate Matter Characteristics	M.L. Bell	2012
157	Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Two Indian Cities		2011
	Part 1. Short-Term Effects of Air Pollution on Mortality: Results from a Time-Series Analysis in Chennai, India	K. Balakrishnan	
	Part 2. Time-Series Study on Air Pollution and Mortality in Delhi	U. Rajarathnam	
154	Public Health and Air Pollution in Asia (PAPA): Coordinated Studies of Short-Term Exposure to Air Pollution and Daily Mortality in Four Cities		2010
	Part 1. A Time-Series Study of Ambient Air Pollution and Daily Mortality in Shanghai, China	H. Kan	
	Part 2. Association of Daily Mortality with Ambient Air Pollution, and Effect Modification by Extremely High Temperature in Wuhan, China	Z. Qian	
	Part 3. Estimating the Effects of Air Pollution on Mortality in Bangkok, Thailand	N. Vichit-Vadakan	
	Part 4. Interaction Between Air Pollution and Respiratory Viruses: Time-Series Study of Daily Mortality and Hospital Admissions in Hong Kong	C-M. Wong	
	Part 5. Public Health and Air Pollution in Asia (PAPA): A Combined Analysis of Four Studies of Air Pollution and Mortality	C-M. Wong	
152	Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis	J.I. Levy	2010

Continued

Copies of these reports can be obtained from HEI; pdf's are available for free downloading at http://pubs.healtheffects.org.

RELATED HEI PUBLICATIONS: MULTIPLE POLLUTANTS, SHORT-TERM STUDIES, AND METHODS

Number	Title	Principal Investigator	Date*
148	Impact of Improved Air Quality During the 1996 Summer Olympic Games in Atlanta on Multiple Cardiovascular and Respiratory Outcomes	J.L. Peel	2010
142	Air Pollution and Health: A European and North American Approach (APHENA)	K. Katsouyanni and J.M. Samet	2009
127	Personal, Indoor, and Outdoor Exposures to $PM_{2.5}$ and Its Components for Groups of Cardiovascular Patients in Amsterdam and Helsinki	B. Brunekreef	2005
123	Time-Series Analysis of Air Pollution and Mortality: A Statistical Review	F. Dominici	2004
Special Reports			
18	Outdoor Air Pollution and Health in the Developing Countries of Asia: A Comprehensive Review		2010
	Revised Analyses of Time-Series Studies of Air Pollution and Health		2003
HEI Communications			
12	Internet-Based Health and Air Pollution Surveillance System	S.L. Zeger	2006

Copies of these reports can be obtained from HEI; pdf's are available for free downloading at http://pubs.healtheffects.org.

HEI BOARD, COMMITTEES, and STAFF

Board of Directors

Richard F. Celeste, Chair *President Emeritus, Colorado College*

Sherwood Boehlert *Of Counsel, Accord Group; Former Chair, U.S. House of Representatives Science Committee*

Enriqueta Bond *President Emerita, Burroughs Wellcome Fund*

Purnell W. Choppin *President Emeritus, Howard Hughes Medical Institute*

Michael T. Clegg *Professor of Biological Sciences, University of California–Irvine*

Jared L. Cohon *President Emeritus and Professor, Civil and Environmental Engineering and Engineering and Public Policy, Carnegie Mellon University*

Stephen Corman *President, Corman Enterprises*

Linda Rosenstock *Dean Emerita and Professor of Health Policy and Management, Environmental Health Sciences and Medicine, University of California–Los Angeles*

Henry Schacht *Managing Director, Warburg Pincus; Former Chairman and Chief Executive Officer, Lucent Technologies*

Warren M. Washington *Senior Scientist, National Center for Atmospheric Research; Former Chair, National Science Board*

Archibald Cox, Founding Chair *1980–2001*

Donald Kennedy, Vice Chair Emeritus *Editor-in-Chief Emeritus,* Science*; President Emeritus and Bing Professor of Biological Sciences, Stanford University*

Health Research Committee

David L. Eaton, Chair *Dean and Vice Provost of the Graduate School, University of Washington–Seattle*

David Christiani *Elkan Blout Professor of Environmental Genetics, Harvard T.H. Chan School of Public Health*

Francesca Dominici *Professor of Biostatistics and Senior Associate Dean for Research, Harvard T.H. Chan School of Public Health*

David E. Foster *Phil and Jean Myers Professor Emeritus, Department of Mechanical Engineering, Engine Research Center, University of Wisconsin–Madison*

Uwe Heinrich *Professor, Hannover Medical School; Executive Director, Fraunhofer Institute for Toxicology and Experimental Medicine, Hanover, Germany*

Barbara Hoffmann *Professor of Environmental Epidemiology and Head of Environmental Epidemiology of Aging, IUF-Leibniz Research Institute for Environmental Medicine, and Professor, Medical Faculty, Heinrich Heine University of Düsseldorf, Germany*

Allen L. Robinson *Raymond J. Lane Distinguished Professor and Head, Department of Mechanical Engineering, and Professor, Department of Engineering and Public Policy, Carnegie Mellon University*

Richard L. Smith *Director, Statistical and Applied Mathematical Sciences Institute, University of North Carolina–Chapel Hill*

HEI BOARD, COMMITTEES, and STAFF

Health Review Committee

James A. Merchant, Chair *Professor and Founding Dean, College of Public Health, University of Iowa*

Michael Brauer *Professor, School of Environmental Health, University of British Columbia, Canada*

Bert Brunekreef *Professor of Environmental Epidemiology, Institute of Risk Assessment Sciences, University of Utrecht, the Netherlands*

Mark W. Frampton *Professor of Medicine and Environmental Medicine, University of Rochester Medical Center*

Stephanie London *Senior Investigator, Epidemiology Branch, National Institute of Environmental Health Sciences*

Roger D. Peng *Associate Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

Armistead Russell *Howard T. Tellepsen Chair of Civil and Environmental Engineering, School of Civil and Environmental Engineering, Georgia Institute of Technology*

Lianne Sheppard *Professor of Biostatistics, School of Public Health, University of Washington–Seattle*

Officers and Staff

Daniel S. Greenbaum *President*

Robert M. O'Keefe *Vice President*

Rashid Shaikh *Director of Science*

Barbara Gale *Director of Publications*

Jacqueline C. Rutledge *Director of Finance and Administration*

April Rieger *Corporate Secretary*

Zachary Abbott *Research Assistant*

Kate Adams *Senior Scientist*

Sarah Benckart, *Science Administration Assistant*

Hanna Boogaard *Staff Scientist*

Adam Cervenka *Research Assistant*

Aaron J. Cohen *Principal Scientist*

Maria G. Costantini *Principal Scientist*

Philip J. DeMarco *Compliance Manager*

Hope Green *Publications Associate*

L. Virgi Hepner *Senior Science Editor*

Anny Luu *Executive Assistant*

Nicholas Moustakas *Policy Associate*

Hilary Selby Polk *Senior Science Editor*

Evan Rosenberg *Staff Accountant*

Robert A. Shavers *Operations Manager*

Geoffrey H. Sunshine *Senior Scientist*

Annemoon M.M. van Erp *Managing Scientist*

Donna J. Vorhees *Senior Scientist*

Katherine Walker *Senior Scientist*

HEALTH
EFFECTS
INSTITUTE

101 Federal Street, Suite 500
Boston, MA 02110, USA
+1-617-488-2300
www.healtheffects.org

RESEARCH
REPORT

Number 183 Parts 1 & 2
June 2015