24

Statistical Methods for Environmental Epidemiology

Francesca Dominici
Harvard University, Cambridge, Massachusetts

Ander Wilson
Colorado State University, Fort Collins, Colorado

CONTENTS

24.1 Introduction .......................................................... 547
  24.1.1 Data Characteristics ........................................... 548
  24.1.2 Sources of Confounding Bias ............................... 549
24.2 Epidemiological Designs .......................................... 550
  24.2.1 Multi-Site Time Series Studies ............................. 551
    24.2.1.1 Distributed Lag Models .............................. 552
  24.2.2 Cohort Studies ................................................ 553
  24.2.3 Intervention Studies ......................................... 554
  24.2.4 Spatial Misalignment ......................................... 556
  24.2.5 Exposure Prediction Modeling .............................. 558
24.3 Estimating the Exposure-Response Relationship .............. 559
  24.3.1 Generalized Linear Models ................................. 559
  24.3.2 Semi-Parametric Approaches ............................... 560
  24.3.3 Model Uncertainty in the Shape of the Exposure-Response 560
24.4 Confounding Adjustment ...................................... 561
  24.4.1 Bayesian Adjustment for Confounding .................... 562
  24.4.2 Relation to BMA .............................................. 564
  24.4.3 Air Pollution Example ....................................... 564
  24.4.4 Concluding Remarks ......................................... 565
24.5 Estimation of Health Effects From Simultaneous Exposure to Multiple Pollutants ......................... 567
  24.5.1 Multi-Pollutant Profile Clustering and Effect Estimation .... 568
  24.5.2 High-Dimensional Exposure-Response Function Estimation .... 569
  24.5.3 Confounding Adjustment in Multiple Pollutant Models .... 569
Bibliography ................................................................. 571

24.1 Introduction

Estimated health risks from short and long term exposure to air pollution from epidemiological studies have provided the necessary evidence base for setting more stringent National Ambient Air Quality Standards (NAAQS) in the United States and around the world [161, 162, 173, 174]. These estimated health risks have carried an enormous weight in set-
ting more stringent NAAQS. The annual cost of implementation and compliance with these NAAQS is now reaching astronomical numbers: billion of dollars in the US alone. Given the need to justify these costs, statistical analyses aimed at addressing questions in environmental epidemiology have been subjected to immense scrutiny both from industry and regulatory agencies. Such statistical analyses need to overcome many data challenges. First, a critical and timely question is whether current levels of air pollution, which in the United States are low and close to background levels, are harmful. This calls for statistical methods that estimate the exposure-response relationship and can reliably estimate health risks at low levels of exposure. Second, health outcomes, air pollution levels and potential confounders vary in time and space, and because of the observational nature of the data, there are many potential measured and unmeasured confounders. Third, exposure is often measured with error. Fourth, often the signal-to-noise ratio in the data is low and the target of inference is often the estimation of a health risk that is small and hard to detect. Fifth, often we are interested in estimating health effects associated with simultaneous exposure to multiple pollutants, thus raising the need for a rigorous definition of health risk associated with simultaneous exposure to multiple pollutants, and also the need for a new framework for adjusting for confounding in the context of multiple exposures. Sixth, air pollution data are incomplete and their missing data pattern might not be at random. For example, often ambient levels of ozone are not measured in the winter season. These and many other challenges have led to the development of sophisticated statistical methods with the goal of providing estimates of health effects that are unbiased and precise.

In this chapter, we first introduce the data setting and provide a high level overview of the concept of confounding in air pollution epidemiology. In Section 24.2, we describe the three most common study designs for estimating health effects of short term exposures, long term exposures, and air quality interventions: time series, cohort, and the intervention studies. In Section 24.3, we provide a general overview of methods for estimating the exposure-response relationships in the context of time series and cohort studies. In Section 24.4, we summarize some recent methodological developments for adjustment for confounding that identify key confounders and adjust for model uncertainty using model averaging. Finally, in Section 24.5 we summarize recent developments for estimating the health effects of exposure to multiple pollutants. Please note that, this chapter does not attempt to review the whole body of the literature on statistical contributions in air pollution epidemiology, which is massive. Our goal is to provide a useful summary of the challenges and opportunities in this field.

### 24.1.1 Data Characteristics

Data in air pollution tend to have some unique features. For example, outdoor levels of air pollution are measured by monitoring stations maintained by government agencies [160]. The levels of these ambient air pollutants are continuous (e.g. particulate matter and ozone) and vary across time and space and for the most part are available at daily levels. Although there are several approaches for measuring exposure (e.g. exposure assessment at individual level using personal monitors), here we will present ideas in the context of exposure to outdoor levels from monitoring stations maintained by the US Environmental Protection Agency (EPA) or other government agencies. In this setting, the most common approach is to assign to each individual or to each geographical area the ambient level from the closest monitoring station or group of stations. More specifically, in the context of time series studies, which will be described more in detail below, the daily average level of air pollution in a geographical area (such as a county or a city) is calculated by taking an average measure of the daily levels from all the monitoring stations that are located in that geographical location [89, 139, 140, 178]. In the context of cohort studies, long term exposure to air pollution at the residential addresses of the members in the cohort
need to be calculated. In the last few years there have been numerous contributions in the context of developing spatio-temporal models that use monitoring data, land use regression variables, atmospheric models, and satellite data to estimate these residential exposures at the highest level of accuracy as possible [54, 98, 142, 168, 180, 181]. The development and application of these models for exposure prediction is changing the landscape of air pollution epidemiology. With these approaches, it is now becoming possible to estimate exposure to air pollution at the residential address of each participant in the cohort with a high level of accuracy in terms of spatial resolution (at the address) and temporal resolution (daily). These predictions are cross-validated with ambient levels from monitoring stations and have a high level of prediction accuracy (see Section 24.2.5 for additional details).

Data on health outcomes are often obtained from cohorts [14, 42, 47, 55, 68, 75, 90, 117, 126, 129, 170, 172] and administrative data, such as Medicare claims data [29, 61, 147, 169, 186, 189]. For example, the Harvard Six Cities Study (SCS) and the American Cancer Society (ACS) study are the two landmark epidemiological cohort studies that had an enormous impact on our understanding of the health effects of air pollution [55, 126]. These studies enroll subjects at the baseline, and at baseline collected an extensive list of individual level characteristics (e.g. body mass index, smoking, occupation, income). In these cohort studies, individuals are followed over time to ascertain time to an event (e.g. a heart attack, hospitalization, or death). In the SCS and ACS, average exposure aggregated at the county level or at the larger metropolitan area level was assigned to each individual. This is a fairly coarse description of a subject’s exposure to air pollution that lends itself to exposure measurement error. As mentioned above, other more recent cohort studies estimate individual level exposure at the subject’s residential address using prediction models that can reduce measurement error [42, 64, 146].

Traditional cohort studies, such as the SCS and the ACS have the advantage of capturing extensive individual level information on the potential confounders but they are limited by the fact that they are “closed” cohort studies in the sense that they do not allow enrollment of new individuals into the cohort. As such, these cohorts cannot be used to estimate the health effects of air pollution using more recent data, nor can be used to track health effects in the future as air pollution levels continue to decline. To overcome this challenge, more recent epidemiological studies have leveraged “open” cohort data, such as Medicare claims, to estimate the effects of long-term exposure to air pollution [72, 94, 153, 189]. Open cohort data permits new enrollees to enter into the cohort every year, thus allowing us to routinely estimate health effects over time as air pollution levels continue to decline. A limitation of the administrative cohorts is that they tend to measure a limited set of potential measured confounders compared to the traditional cohorts. On the other hand, they include large and often nationally representative study populations.

### 24.1.2 Sources of Confounding Bias

A key concern in environmental epidemiology studies is confounding. The source and degree of confounding bias and the type of statistical methods needed to adjust for confounding depend on the study design. In the context of randomized experiments, because assignment to treatment (that is exposure to high versus low levels of air pollution) is at random, characteristics of the individuals that are “exposed” and “not exposed” to air pollution are automatically balanced between the two groups because of the randomization. In the context of time series studies, the goal is to estimate the association between day-to-day changes in air pollution levels and day-to-day changes in the daily number of deaths (or hospitalizations) within a city and then combine information regarding these estimated associations across cities (or other geographical locations, such as counties or metropolitan areas). Because of the nature of a time series design, city-specific characteristics that do
not vary from day to day (e.g. smoking prevalence or average income in that city) are not confounders [122]. In cohort studies, the goal is to estimate the association between long term exposure to air pollution and time to an event for individuals that live in different geographical locations with different levels of air pollution. Therefore individual level and/or area level characteristics that are associated with the exposure and also predictive of the outcome (e.g. smoking, body mass index, occupation, income) can be confounders [111, 115].

In the context of intervention studies we are interested in estimating the effect of a specific intervention on both: 1) air pollution levels and 2) health outcomes. For example, let’s consider the intervention designed to reduce traffic congestion during the 1996 Summer Olympic Games in Atlanta, Georgia, and the effect of this intervention on ambient ozone levels [119]. In this context, a potential confounder of the effect of the intervention on ambient ozone levels is time trend. It is possible the ambient air pollution levels could have been declining over time regardless of the intervention. A key questions is, therefore, whether the observed degree of reduction in ambient levels of ozone can be “truly” attributed to the intervention or would this decline have occurred anyway even absent the intervention [10, 58, 119, 192].

In addition, when the goal is estimation of the health effects of a single and pre-specified environmental agent (e.g. ozone), environmental epidemiology studies are particularly prone to confounding bias because of the many co-exposure and covariates that are correlated with the exposure and also influence health. For example, temperature is highly correlated with ozone because ozone formation occurs on warm sunny days. Temperature also affects health and therefore can confound the estimation of health risks associated with short term exposure to ozone [15, 17, 20, 22, 24, 84]. In the context of chronic effect studies, where most of the information comes from variation across geographical locations in long-term exposure to air pollution, an important potential confounder that must be consider is average income. For example, geographical locations with lower average income also have higher levels of air pollution and average income affects health outcomes.

An increasing number of papers in the air pollution epidemiological literature use the term “causal effects” in the abstract and in the summary [74, 147, 169, 193]. In the context where observational data is analyzed to assess the “effect” of an action (such as an intervention, an exposure, or a treatment) on a given outcome, the scientific community is increasingly relying on statistical methods being labeled as “causal,” and such methods are frequently contrasted against more traditional methods labeled as “associational” and therefore non-causal. In the specific context of air pollution epidemiology, the promise of identifying causal relationships between, for example, air pollution exposure and human health, has rightly generated ample enthusiasm.

We have argued that causal inference methods can provide a rigorous framework to use data to learn about consequences of specific actions [194]. More specifically, causal inference methods are tools for: 1) formalizing thinking about what data can tell about cause and effect; 2) forcing explicit definitions of familiar notions; 3) clarifying common threats to validity in epidemiological studies, while at the same time providing a remedy. In this chapter, we will not formalize the presentation of the statistical methods using the the formal notation of potential outcomes in causal inference, but we will provide references of recent contribution on this area.

## 24.2 Epidemiological Designs

In this section, we briefly review three popular designs in environmental epidemiology: multi-site time series studies, cohort studies, and intervention studies. Time series studies
have been a popular choice for air pollution epidemiology to estimate acute health effects associated with short term exposure. Cohort studies are a popular choice to study the chronic health effects associated with long term exposure. Intervention studies are very useful to assess the consequences of a specific air quality intervention on both the ambient levels of air pollution and health outcomes. It is important to recognize that several other epidemiological study designs can be used to estimate the health effects of environmental exposures in addition to the those discussed here, other popular choices include case-crossover and panel studies (see for example [21, 51, 87, 96]). We conclude this section by discussing spatial misalignment and exposure prediction modeling as they relate to environmental epidemiology studies.

24.2.1 Multi-Site Time Series Studies

Multi-site time series studies of air pollution and health outcomes (i.e. mortality and morbidity) have provided evidence that daily variation in air pollution levels is associated with daily variation in mortality and morbidity counts [3, 18, 67, 128]. These findings have served as key epidemiological evidence for several reviews of the NAAQS for particulate matter and ozone (for example [20, 61, 138, 144, 149, 187]). Considering the large policy impact of epidemiological studies of short term effects of air pollution, critics have raised concerns about the adequacy of current model formulations. In this section, we briefly review two important methodological contributions for the analyses of multi-site time series studies: 1) model choice to adjust for unmeasured confounding and 2) accounting for the delayed effect of a few days by the development and application of distributed lag models.

Time series studies estimate the association between day-to-day changes in air pollution levels and day-to-day changes in the daily number of events (deaths or hospitalizations) that have occurred within a given geographical location, such as a city or a county [23, 63].

Generalized linear models (GLM) with parametric splines (e.g., natural cubic splines) [113] or generalized additive models (GAM) with nonparametric smoothers (e.g., smoothing splines or loess smoothers) [79], are generally used to estimate effects associated with exposure to air pollution while accounting for smooth fluctuations in mortality that confound estimates of the pollution effect [60].

Time series data on pollution and mortality are generally analyzed by using log-linear, Poisson regression models for over-dispersed counts with the daily number of deaths (or hospitalizations) as outcome, the (possibly lagged) daily level of pollution as a linear predictor and smooth functions of weather variables and calendar time used to adjust for time-varying confounders. Investigators around the world have used different approaches to adjust for confounding, making it difficult to compare results across studies (for example, [89, 141]).

To date, the statistical properties of these different approaches have not been comprehensively compared. Peng et al conducted an extensive simulation study aimed at characterizing model uncertainty and model choice in adjusting for seasonal and long-term trends in time series models of air pollution and mortality (unmeasured confounders) [122]. The authors generated data under several confounding scenarios and systematically compare the performance of the various methods with respect to the mean squared error of the estimated air pollution coefficient. They found that the bias in the estimates generally decreases with more aggressive smoothing and that model selection methods which optimize prediction, which have been the standard approach, may not be suitable for obtaining an estimate with small bias. In another paper, Dominici et al developed a bandwidth selection method to reduce confounding bias in the health risk estimate due to unmeasured time-varying factors, such as season and influenza epidemics [59]. More specifically, in this paper the authors calculate in closed form the asymptotic bias and variance of the air pollution risk estimate as they vary the degree of adjustment for unmeasured confounding factors. The authors then
show that confounding bias can be removed by including in the Poisson regression model
smooth functions of time and temperature that are sufficiently flexible to predict pollution
(instead of tuning to predict the outcome). These asymptotic calculations show that in
most situations it is possible to effectively reduce unmeasured confounding bias by estimat-
ing the number of degrees of freedom in the smooth functions of time and temperature that
best predict pollution levels. Controlling for the potential confounding effects of “measured
confounders” (such as weather variables) is a better-identified problem than controlling for
“unmeasured confounders.” More details on how to adjust for measured confounders and at
the same time account for model uncertainty in the confounding adjustment are discussed
later in this chapter (see Section 24.4).

24.2.1.1 Distributed Lag Models

Time series studies have provided strong evidence of an association between increased levels
of ambient air pollution and increased hospitalizations, typically at a single lag of 0, 1 or
2 days after an air pollution episode. An important scientific objective is to estimate the
cumulative risk of hospitalization of an air pollution episode over a few days after the air
pollution event. More specifically, if in a given day the level of air pollution increases from
$x_t$ to $x_{t+\delta}$, our goal is to estimate the cumulative risk associated with a potential increase
in the adverse health outcome on the same day $t$ but also on the following $L$ days (e.g.
$y_t, y_{t+1}, \ldots, y_{t+L}$).

A distributed lag model (DLM) is a regression model that includes lagged exposure vari-
ables as covariates; its corresponding distributed lag (DL) function describes the relationship
between the lag and the coefficient of the lagged exposure variable. DLMs have recently been
used in environmental epidemiology for quantifying the cumulative effects of weather and
air pollution on mortality and morbidity (for example [66, 80, 81, 123, 171, 184]).

Standard methods for formulating DLMs include unconstrained, polynomial, and penal-
ized spline DLMs [4, 84, 107, 134, 150, 151, 154, 184, 188]. Standard methods for fitting
DL functions may fail to take full advantage of prior information about the shape of the
DL function for environmental exposures, or for any other exposure with effects that are
believed to smoothly approach zero as lag increases, and are therefore at risk of producing
suboptimal estimates. In a paper by Welty et al, the authors propose a Bayesian DLM
(BDlagM) that incorporates prior knowledge about the shape of the DL function and also
allows the degree of smoothness of the DL function to be estimated from the data [171]. In a
simulation study, they compare the proposed Bayesian approach with alternative methods
that use unconstrained, polynomial, and penalized spline DLagMs. They also illustrate the
connection between BDlagMs and penalized spline DLagMs.

In a subsequent paper, the same set of authors extended the BDlagMs to multi-site time
series studies [123]. Specifically these authors introduce a Bayesian hierarchical distributed
lag model (BHDLM) for estimating a national average distributed lag function relating
particulate matter (PM) air pollution exposure to hospitalizations for cardiovascular and
respiratory diseases. The BHDLM builds on earlier work [171, 184] by smoothing DL func-
tion estimates across lags and by providing a method for combining these functions across
locations. The specific prior allows for more flexibility in the shape of the DL function at
the shorter lags but less flexibility in the shape of the DL function at the longer lags. In
addition, the hierarchical model lets us examine the range of shapes in the county-specific
distributed lag functions. The authors have established that this proposed methodology is
related to penalized spline modeling with a special type of penalty. This connection, along
with evidence from simulation studies that were conducted by Welty et al [171], creates a
basis for understanding the statistical properties of the approach.
24.2.2 Cohort Studies

Air pollution cohort studies associate long-term exposure with health outcomes. Either a prospective or retrospective design is possible. In a prospective design, participants complete a questionnaire at entry into the study to elicit information about age, sex, weight, education, smoking history, and other subject-specific characteristics. They are followed over time for mortality or other health events. A measure of cumulative air pollution is often used as the exposure variable. A key design consideration for air pollution cohort studies is identifying a cohort with sufficient exposure variation. Individuals from multiple geographic locations must be studied in order to assure sufficient variation in cumulative exposure, particularly when ambient air pollution measurements are used. However, by maximizing the geographical variability of exposure, the relative risk estimates from cohort studies are likely to be confounded by area-specific characteristics.

Survival analysis tools can evaluate the association between air pollution and mortality. Typically the Cox proportional-hazards model [40, 44] is used to estimate mortality rate ratios for airborne pollutants while adjusting for potential confounding variables (for example [6]). Relative risk is estimated as the ratio of hazards for an exposed relative to an unexposed or reference group.

The epidemiological evidence on the long-term effects of air pollution on health has been reviewed by Pope [124]. The Harvard SCS and the ACS [55, 126] are among the largest air pollution prospective cohort studies. In the SCS [55, 105, 106] a random sample of 8111 adults who resided in one of the six US communities at the time of the enrollment was followed for 14 to 16 years. An analysis of all-cause mortality revealed an increased risk of death associated with increases in particulate matter and sulfate air pollution after adjusting for individual-level confounders. Because of the small number of locations, findings of this study cannot be generalized easily.

The ACS study [125, 126, 127] evaluated effects of pollution on mortality using data from a large cohort drawn from 151 metropolitan areas. Ambient air pollution from these areas was linked with individual risk factors for 552,138 adult residents. The ACS study covered a larger number of areas, however, the subjects were not randomly sampled as in the SCS. Both studies reported similar results: the relative risk of all-cause mortality was 1.26 (95% CI 1.08, 1.47) for an 18.6 \( \mu g/m^3 \) change in PM less than 2.5 \( \mu m \) in aerodynamic diameter (PM\(_{2.5}\)) in the SCS and 1.17 (95% CI 1.09, 1.26) for a 24.5 \( \mu g/m^3 \) change in PM\(_{2.5}\) in the ACS study. A detailed reanalysis of these two studies [102, 103] and a new ACS study incorporating data for a longer period of time [127] replicated and extended these results by incorporating a number of new ecological covariates and applying several models for spatial autocorrelation.

More recently, estimates of the long term effects of air pollution on health have been obtained with statistical methods that resemble the ones of multi-site time series studies [72, 86, 94, 157]. In the papers by Janes et al (2007) and Greven et al (2011), the authors aim at addressing a very specific question: whether there is an association between month-to-month variations in mortality rates and month-to-month variations in mortality rates in PM\(_{2.5}\) for the previous 12 months (global effect) in the Medicare population [72, 86]. In these studies the authors decompose the global effects into two parts: 1) the association between the national average trend (NAT) in the monthly PM\(_{2.5}\) levels averaged over the previous 12 months and the NAT in monthly mortality rates (national effect); and 2) the association between the deviation of the community-specific trend from the NAT of PM\(_{2.5}\) and the deviation of the community-specific trend from the NAT of mortality rates (local effect). They decompose the global effect into a national effect plus a local effect because they hypothesize that the national effect is more likely to be affected by unmeasured confounding than the local effect. They also argue that if there are large differences between the local and the national
effects then the global effect should not be reported without a more in depth investigation of confounding. They acknowledge in the papers that differences between the local and national effects might be due to measurement error and not necessarily unmeasured confounding.

Kioumourtzoglou et al (2015) introduced a similar modeling approach where the authors specify time-varying Cox proportional hazards models separately in each city [94]. More specifically, by analyzing data within each city, they first assess whether year-to-year fluctuations in PM$_{2.5}$ concentrations, around their long-term trends, are associated with year-to-year survival variations within cities. City-specific analyses eliminated confounding by factors that do not vary across time but vary across cities. They specify a multi-stage model. First, they fitted separate Cox proportional hazards models in each city, stratifying by age (5-year categories), gender, race (white, black, other) and follow-up time, with follow-up beginning on January 1st after entry in the cohort. They used annual and 2-year total PM$_{2.5}$ mass concentrations, separately, as time-varying exposures. They used the counting process extension of the proportional hazards model by Andersen and Gill [5] and created multiple observations for each subject, with each observation representing a person-year of follow-up. In addition, they adjusted linearly for calendar year, hence controlling for long-term trends and focusing the analysis on variations in exposure around its time trend. In the second stage, they combined the city-specific health effect estimates, using a random effects meta-analysis. With this approach, they eliminated all confounding by covariates that vary across cities, since this is a city-specific analysis, and by covariates whose long-term trends coincide with trends in PM$_{2.5}$ within cities, since those trends were removed.

24.2.3 Intervention Studies

The regulatory and policy environment surrounding air quality management warrants new types of epidemiological evidence above and beyond the one provided by estimating the exposure-response function from time series and cohort studies (often called observational studies). Whereas air pollution epidemiology has typically informed policies with estimates of exposure-response relationships between pollution and health outcomes, new types of evidence can inform current debates about the actual health impacts of air quality regulations. Directly evaluating specific regulatory strategies is distinct from and complements estimating exposure-response relationships and puts increased emphasis on assessing the effectiveness of well-defined regulatory interventions [192, 193]. In this section we want to sharpen the analytic distinctions between studies that directly evaluated the effectiveness of specific policies and those that estimated exposure-response relationships between pollution and health from observational studies. A recent report by Zigler et al published by the Health Effects Institute [194] provides a comprehensive overview of statistical methods for causal inferences to assess the health impacts of air quality regulations.

Randomized control trials would be the best way to measure the health benefits of PM reductions, but for obvious reasons, true experiments are generally not feasible. One exception is chamber studies of controlled exposure, but they rely on healthy subjects and focus only on limited subclinical endpoints. An observational study of the health effects of particulates boils down to a comparison of health outcomes across space and/or time among places with differing levels of air pollution (see above the sections on multi-site time series studies and cohort studies). For the cohort studies, one challenge is that the people who live in the more polluted places frequently have differing initial levels of health (e.g., due to differences in smoking rates, diet, or socioeconomic status) from the levels of people who live in the less polluted places. Another challenge is that there may be locational determinants of health (e.g., hospital quality or water pollution) that differ across the places and are correlated with air pollution levels. Further, people may sort to locations based on their (likely unobserved) susceptibility to pollution and other related health problems.
and/or they may spend greater resources on self-protection in polluted locations in ways that are not measured in available datasets. Statistical methods, mostly based on regression approaches, aim to “adjust” for observed confounders, by including the available measures of behavioral, socioeconomic, and locational differences as covariates in the regression model, but many of the determinants of health are unobserved and can lead to biased estimates of the relationship between health and particulates [58].

Thanks to the rigorous statistical methods that have been developed and applied to the assembled data, and to the enormous effort of government agencies and specific investigators in conducting independent re-analyses (see for example Krewski et al [104]) analyses of observational data have had a large impact on air quality regulations and on the supporting analyses of their accompanying benefits. Nonetheless, legitimate concerns remain. While important progress has been made in adjusting for measured and unmeasured confounding in observational studies, it remains true that there may be unobserved differences across the populations and locations, and the measurable differences may not have been adjusted for sufficiently.

One type of intervention study is the quasi-experimental (QE) studies. In a QE evaluation, the researcher compares outcomes between a treatment group and a control group, just as in a classical experiment; but, treatment status is determined by politics, an accident, or most importantly a regulatory action. The key difference with an observational study in this setting is that the QE approach is devoted to identifying treatment-induced variation in particulates that plausibly mitigates confounding or omitted variables bias in the estimated relationship between human health and particulates, rather than relying on the variation presented by nature and optimizing agents. Despite the “nonrandom” assignment of treatment status, it is possible to draw causal inferences from the differences in outcomes (here by outcomes we refer to both air pollution levels and health risks) between the treatment and control groups in a quasi- or natural experiment, provided certain assumptions are met. This approach has been used extensively in recent years and has permitted more credible inferences about the impacts of a wide range of relationships. In fact, there is an emerging QE literature on the human health effects of air pollution that relies on designs where an “action” has affected the ambient levels and the chemical composition of air pollution. Some of the most well-known examples are: 1) the ban of coal sales in Dublin [39]; 2) the differential reduction in total suspended particulates (TSPs) across the country as a consequence of the 1981-1982 recession [35]; 3) the air pollution reduction interventions before, during, and after the Beijing Olympic games [133]; 4) a steel plant strike [131]; 5) features of the Clean Air Act [53, 193] and, 6) the Chinese policy that provided free coal for heating in cities north of the Huai River [36]. See also [1, 69, 114, 159] for detailed reviews.

The Huai River study illustrates some of the appealing features of QE designs and more specifically of regression discontinuity (RD) [108]. It exploits a Chinese policy that provided free coal for winter heating in areas north of the Huai River and denied coal-based heating to the south of the river. The idea is to compare locations just north and south of the river. In this setting, the RD design relies on the assumption that any confounders (both observed and unobserved) vary smoothly with latitude as one crosses the Huai River, except for the availability of coal-based indoor heating. The authors controlled for these potential confounders through adjustment for a flexible polynomial in distance to the river, measured as degrees latitude that each location is north of the Huai River. The authors find that north of the river the policy led to discrete increases in TSPs and discrete decreases in life expectancy (derived from age-specific mortality rates). Importantly, the effect of TSPs on life expectancy is largely insensitive to whether observable covariates are included in the specification, which would be the case in a randomized control trial.

While QE approaches promise more credible estimates, they are not without limitations. It is important that QE designs are able to demonstrate that observable covariates
are balanced by the treatment and credibly explain why unobserved ones are likely to be balanced too. In cases where the covariates are imbalanced and/or the unobserved ones are unlikely to be balanced, quasi-experimental estimates are unlikely to be more credible than associational estimates. Further, QE approaches can often be demanding of the data and lack statistical power. As is the case with associational estimates, applying QE estimates to other settings (e.g., places, periods, and demographic groups) requires careful consideration and in some cases may be inappropriate. This challenge can be greater with QE approaches where the selection of the study population is dictated by the available treatment (e.g., the Chinese policy) and therefore is beyond the researcher’s controls.

24.2.4 Spatial Misalignment

Estimating the health risks of environmental exposures often involves examining data at different levels of spatial resolution. This mismatch between data measured at different resolutions results in spatial misalignment [1], which can bias the health risk estimates. Spatial misalignment in environmental health studies is very common because the air pollution data and the data on the health outcomes often come from different and unrelated sources. For example, data on ambient air pollution levels often are based on a network of monitors operated by the US EPA where each monitor measures daily levels of ambient air pollution at a specific point location. On the other hand, data on health outcomes, such as the numbers of hospital admissions for cardiovascular disease and number of deaths, generally are obtained from Centers for Medicare and Medicaid Services (CMS) and other governmental agencies, such as the National Center of Health Statistics (NCHS). Because health and exposure data are often collected independently of each other, they are rarely spatially aligned. Hence, a direct comparison of the exposure and health outcome is not possible without a model (or an assumption) to align the two sources of information both spatially and temporally. In a time series study of air pollution and health, one is interested in estimating associations between daily changes in county-wide hospital admissions or mortality counts and daily changes in county-wide average levels of a specific pollutant. The problem is that we do not directly observe county-wide average pollutant levels. Rather, we have measurements taken at a handful of monitors (sometimes only one) located somewhere inside the county.

For a spatially homogeneous pollutant, the value of the pollutant at a single monitor may be representative of the county-wide average ambient level of that pollutant. Some pollutants, particularly some gases such as ozone, are reasonably spatially homogeneous across the area of a county. The total mass of PM$_{2.5}$, whose health risks have been examined extensively, is fairly spatially homogeneous, and monitor measurements of PM$_{2.5}$ in counties with multiple monitors tend to be highly correlated across both time and space [16, 19, 121].

With a pollutant such as PM$_{2.5}$, the misalignment between the continuous nature of the pollutant process and the aggregated nature of the health data does not typically pose as serious a problem as some other pollutants. In this situation, current approaches for data analysis may provide reasonable estimates of risk. There are many sources of measurement error in the analysis of air pollution and health data, and much previous work has focused on the mismatch between personal and ambient exposures to an airborne pollutant and how to adjust for measurement error in health effect models [56, 95, 152, 190].

This is indeed an important problem, but it is typically not one that can be dealt with using the types of data that are routinely available. Given an aggregated health outcome, the ideal exposure is the average “personal” exposure over the target population [190]. Because it is unrealistic to measure this quantity repeatedly over long periods of time, population studies must estimate this value [31, 34] or, more commonly, resort to suitable proxies such as the average “ambient” concentration. A key assumption made in previous time series analyses of air pollution and health data has been that the pollutant of interest is spatially
homogeneous and that the monitor value on a given day (or the average of a few monitors) is approximately equal to the true ambient average concentration over the study population area. Any difference between the monitor value and the true ambient average concentration is what we call “spatial misalignment error.”

In past analyses, the assumption that this error was zero may have been reasonable given that most previous analyses focused on pollutants such as the total mass of PM, ozone, and other pollutants that have been shown to be fairly spatially homogeneous over relatively long distances [140]. Recently, data have become available from the US EPA’s Chemical Speciation Trends Network (STN) as well as state and local air monitoring stations which provide daily mass concentrations of approximately 60 different chemical elements of PM$_{2.5}$. These data are monitored in over 200 locations around the United States starting from the year 2000. Although the data are promising and are the subject of intense interest and new research, they also raise new statistical challenges. In particular, the usual assumption that the monitor value is approximately equal to the ambient average is less tenable when examining certain components of PM$_{2.5}$ [16]. For example, when estimating the correlations between pairs of monitors in the STN for 7 chemical components of PM$_{2.5}$ mass as a function of the distance between the monitors, these authors found that some of the chemical components of PM$_{2.5}$ are spatially homogeneous with high correlations over long distances (> 50 km), while other components are spatially heterogeneous and exhibit practically no correlation beyond short distances (< 20 km). Much of the spatial heterogeneity in the chemical components measured by the STN can be explained by the nature of the sources of the various components. For example, elemental carbon (EC) and organic carbon matter (OCM) tend to be emitted primarily from vehicle or other mobile sources, and thus their spatial distribution can depend on the localized nature of those sources. Secondary pollutants such as sulfate and nitrate are created in the air by the chemical and physical transformation of other pollutants and tend to be more regional in nature. Hence, for spatially heterogeneous pollutants such as EC or OCM, the daily level of those pollutants at a single monitor may be a poor surrogate for the daily county-wide average ambient level of that pollutant.

In a paper by Peng and Bell the authors describe a general method for estimating health risks associated with PM components from time series models while adjusting for potential spatial misalignment error [120]. They first develop a spatial-temporal model for the exposure of interest and estimate the degree of spatial misalignment error for each component in a location. Then they apply two methods—a regression calibration procedure and a two-stage Bayesian model—to estimate the health risks associated with these components and compare two results to standard approaches. Their findings indicate that the effect of spatial misalignment depends on monitor coverage within a county and the spatial variability of the pollutant of interest.

Although here we are discussing this issue mainly on time series studies, spatial misalignment can also induce error in cross-sectional studies of air pollution and health. Gryparis et al demonstrate in detail how to handle the misalignment errors in these types of studies and compare the performance of a number of different statistical approaches [73]. An alternate modeling approach has been proposed by Fuentes et al for estimating the spatial association between spectated fine particles and mortality [65]. Both approaches introduce a spatial model for the monitored pollutant concentrations and either predict pollutant values at unobserved locations or compute area averages over counties to link with county-level health data.

Ultimately, the best way to address the problem of spatial misalignment might be to move away from the county-based summaries of the outcome of interest when possible and begin using summaries with finer spatial resolution, such as zip codes or census tracts. In general, a decrease in the area covered per monitor is associated with lower spatial mis-
alignment error, and this effect is far more pronounced for pollutants that are inherently heterogeneous such as sodium ion, silicon, and EC. This requires the availability of exposure and health data for the smaller areal units. Exposure data from sparse monitoring networks may not be available for smaller areal units; however, this can be overcome by exposure prediction modeling as discussed in Section 24.2.5. Unfortunately, many types of health data are simply not available at finer spatial resolution, and we often must accept what is available. Furthermore, due to activity patterns, a high spatial resolution does not necessarily better capture personal exposure than a larger area when individuals move between areas (e.g. live in one zip code, but work in another). Thus, there is a strong need for methods that address spatial misalignment of air pollutant concentrations used in health studies.

24.2.5 Exposure Prediction Modeling

More recently there have been several contributions in the literature that estimates exposure to PM$_{2.5}$ at small spatial resolution and in geographical areas that are far from monitoring stations. These methods include interpolation or Kriging methods [12] and land-use regression models [64]. Recent studies have focused on improving prediction by combining sophisticated statistical models with multiple data sources. For example, several recent papers have used sophisticated spatiotemporal models and supplemented existing monitoring networks with other data sources including land use characteristics, chemical transport models, and/or supplemental monitoring data [13, 38, 82, 91, 93, 109, 116, 163, 168, 183]. These approaches, in combination of improved validation procedures [167], have proven successful in improving prediction accuracy of PM$_{2.5}$, PM components, and other ambient pollutants. For example, one of these hybrid approaches has focused on use of satellite-retrieved aerosol optical depth (AOD), in addition to monitor and land use data, to predict ground-level PM$_{2.5}$ concentrations [2, 85, 98]. Kloog et al have developed hybrid prediction models that use satellite-based AOD data in conjunction with daily calibration and spatiotemporal statistical models to predict daily PM$_{2.5}$ mass concentrations at 1km $\times$ 1km grid cells in New England, Mid-Atlantic States, and the Southeastern US. This provides a major advance in the field of particle exposure assessment and epidemiology because it predicts exposure levels in both urban and rural environments. Using these data, they simultaneously estimate chronic and acute effects of PM$_{2.5}$ on cardiovascular hospital admissions, mortality, and other specific outcomes, including examining the effect at low PM$_{2.5}$ concentrations [97, 99, 100]. Even with improved prediction methods, there is still potential for bias due to measurement error and several recent studies have proposed additional adjustment approaches [11, 77, 158].

In a recent paper, Cefalu and Dominici pointed out that in environmental epidemiology, we are often faced with two challenges [32]. First, an exposure prediction model is needed to estimate the exposure to an agent of interest, ideally at the individual level (see above several references). Second, when estimating the health effect associated with the exposure, confounding adjustment is needed in the health-effects regression model (see for example the next section on Bayesian Adjustment for Confounding). The current literature addresses these two challenges separately. That is, methods that account for measurement error in the predicted exposure often fail to acknowledge the possibility of confounding, whereas methods designed to control confounding often fail to acknowledge that the exposure has been predicted.

Recent work by these authors and by Szpiro et al [32, 156] consider exposure prediction and confounding adjustment in a health-effects regression model simultaneously. Using theoretical arguments and simulation studies, they show that the bias of a health-effect estimate is influenced by the exposure prediction model, the type of confounding adjustment used
in the health-effects regression model, and the relationship between these two. Moreover, the authors argue that even with a health-effects regression model that properly adjusts for confounding, the use of a predicted exposure can bias the health-effect estimate unless all confounders included in the health-effects regression model are also included in the exposure prediction model. Bergen et al illustrate the importance of accounting for exposure model characteristics when estimating health effects using predicted exposure through a case study using the Multi-Ethnic Study of Atherosclerosis and Air Pollution study [26].

24.3 Estimating the Exposure-Response Relationship

To protect public health and welfare against the dangers of air pollution, the US EPA establishes NAAQS. In 2012, in response to mounting evidence demonstrating the harmful effects of exposure to fine particulate matter, the EPA enacted more stringent NAAQS for PM$_{2.5}$. However, as air pollution levels continue to decrease, regulatory actions are becoming increasingly expensive to maintain. The annual cost of implementation and compliance with the NAAQS has reached 350 million dollars. Given the need to justify these costs, research examining the public health benefits of cleaner air will be subject to immense scrutiny. Yet significant gaps in knowledge remain, particularly with regard to the health effects of long-term exposure to lower levels of air pollution.

Despite a substantial amount of epidemiological literature on the health effects of both short-term and long-term exposure to air pollution, few studies have characterized the health effects of air pollution at levels in accordance with or lower than the most recent NAAQS for PM$_{2.5}$ (now set at 12 µg/m$^3$ for annual mean PM$_{2.5}$). As air pollution levels decrease, studies are needed to determine if further reductions will lead to substantial improvements in health. In this setting, estimation of an exposure-response function to estimate the health effects at various exposure levels (say lower than the current standards) is needed.

24.3.1 Generalized Linear Models

The most common approach to estimate the exposure-response relationship is a GLM with linear exposure-response relationship. For observation $i$, the model is

$$g(E(Y_i|X_i, Z_i)) = \beta_0 + X_i\beta + \sum_{j=1}^{p} Z_{ij}\gamma_j,$$  \hspace{1cm} (24.1)

where $Y_i$ is the health outcome, $X_i$ is the exposure of interest, $Z_{i1}, \ldots, Z_{ip}$ are a set of subject specific covariates or confounding variables. However, the linear exposure-response function does not answer these important questions about health effects specifically at low (or high) concentrations. In this section we will introduce approaches to address this problem.

When the goal is estimation of the health effects of exposure to PM$_{2.5}$ at low concentrations, a common and intuitive approach is to restrict the analysis to days and/or geographical locations that have low exposure. For example, a recent paper by Shi et al [153] used these types of restrictions to study the effect of long-term exposure to low concentrations of PM$_{2.5}$ on mortality in a New England time-series study. In this setting, (24.1) took the form a Poisson regression with a log link, $g(\mu) = \log(\mu)$, and the exposure was average PM$_{2.5}$ over the previous 365 days. The analysis was conducted on all individuals and restricted to those that were exposed to a average PM$_{2.5}$ less than 10 µg/m$^3$. The authors found a 7.52% (95% CI: 1.95, 13.40%) increase in morality associated with a 10 µg/m$^3$ in PM$_{2.5}$ in
the complete data analysis but a 9.28% (95% CI: 0.76, 18.52%) increase in mortality in the restricted analysis. In this case restriction is able to tell us two important things about the exposure-response relationship. First, comparing the restricted analysis to the unrestricted analysis is suggestive (although not statistically significant) of a larger exposure effect at lower concentration levels. Second, the presence of an exposure effect of long-term exposure to PM$_{2.5}$ below 10 µg/m$^3$ provides a key piece evidence that additional reduction in PM$_{2.5}$ exposure below 10 µg/m$^3$ could reduce the mortality burden of PM$_{2.5}$ exposure.

A key aspect that makes restriction appealing is that it requires no statistical methods beyond GLM, yet by carefully selecting the exposure range for restriction it can estimate a different exposure effects over different concentration ranges. In practice, it can be difficult to choose the cut points. One approach is to estimate the model with several cut points. Then, use an information criterion such as AIC or BIC to find the best fitting model.

### 24.3.2 Semi-Parametric Approaches

A flexible alternative to a linear exposure-response function or restriction is a nonlinear exposure-response function. Several approaches have been proposed, for example, natural cubic splines [22, 49, 57], penalized splines [148, 153], B-splines [155], or loess [145]. Here, we consider using a spline basis function to estimate a nonlinear exposure-response function $f$. The regression model then takes the form

$$
g(E(Y_i | X_i, Z_i)) = \beta_0 + f(X_i; \mathbf{v}) + \sum_{j=1}^{p} Z_{ij} \gamma_j \quad (24.2)$$

where $f(X; \mathbf{v})$ is modeled as a spline basis with $k$ knots at locations $\mathbf{v} = (v_1, \ldots, v_k)$. The number of knots is generally specified a priori based on prior subject knowledge or chosen to maximized model fit as measured by AIC BIC, or GCV. However, this results in inference that is conditional upon the number and location of knots.

To account for uncertainty in knot selection, Dominici et al [57] estimated the number and location of knots from the data using reversible jump Markov chain Monte Carlo (RJMCMC) [70] in an analysis of the 88 largest US cities. This approach, allows the model to change both the number and location of knots. In addition, RJMCMC accounts for model uncertainty in knot selection. An alternative approach to account for model uncertainty in knot selection is to average the estimates over several candidate models with different number and location of knots [148] using Bayesian model averaging (BMA) [83, 130].

### 24.3.3 Model Uncertainty in the Shape of the Exposure-Response

Accounting for model uncertainty when estimating the shape of the exposure-response is a critically important aspect of studies of air pollution and health. Bobb et al have proposed a flexible class of time series models to estimate the relative risk of mortality associated with heat waves and conduct BMA to account for the multiplicity of potential models for the exposure-response function [27]. The proposed approach overcomes many of the challenges in estimating the adverse health effects of heat-wave events. More specifically, within each city, the authors specify a semi-parametric model to flexibly capture the nonlinear relation between several weather variables and mortality. The model makes as few assumptions as possible about the shape of the temperature-mortality function and does not require the cities to have the same model or even to include the same temperature predictors (e.g. one city could us maximum daily temperature and another could use mean daily temperature). This allows for heterogeneity of the temperature-mortality association across cities, in accordance with findings in prior studies that the shape of the temperature-mortality curve
varies by US region [48]. The authors also incorporate model uncertainty in the specification of the temperature-mortality exposure-response function by conducting BMA. They found that for some cities the estimation of health risks associated with heat waves is sensitive to the specification of the exposure-response function (e.g. specification of the relationship between weather variables and mortality counts).

An important element of this analysis was the comparison of the posterior variance of the log relative risk under BMA to the variance under a single model which emphasized a twofold benefit of model averaging. Under the model averaging approach, each candidate model is weighted by its posterior probability and the uncertainty of estimates from less plausible models (those with low posterior model probability) do not contribute to the model-averaged uncertainty estimate. However, when multiple models are plausible, BMA incorporates the variability from each potential model. Thus conditioning inference on a single model obtained through a model selection procedure likely underestimates statistical uncertainty.

### 24.4 Confounding Adjustment

Estimating the effect of an exposure on an outcome, while properly adjusting for confounding factors, is a common and challenging goal in the analysis of observational studies. In the context of estimation of health effects of exposure to environmental contaminants, the choice of approaches for adjusting for confounding and approaches for selecting the key confounders among a large set is critical. Indeed health effect estimates can be sensitive to these choices [58].

A common practice is to select a statistical model for the estimation of the effect, and report effect estimates and confidence intervals that are conditional on that model being correct. This does not account for “adjustment uncertainty,” that is uncertainty about which variables should be included in the model to properly adjust for confounding. It is possible to effectively convey this uncertainty by sensitivity analysis, showing the variation of the effect estimate and its interval over a range of plausible choices of confounders [59, 122].

BMA has been suggested as a more formal tool to account for model uncertainty [83, 130]. The BMA approach used here is based on augmenting the model with indicator variables of whether each predictor is included in the model as an unknown nuisance parameters. This results in a weighted average of predictions whose weights depend on the support that each selection receives from the data. This principled approach enjoys a number of desirable properties from a frequentist point of view as well, and has performed competitively in out-of-sample prediction comparisons [37, 182]. The conceptual simplicity and solid logic behind treating the unknown confounder subset as an unknown parameter is attractive in adjustment uncertainty as well. BMA estimates the exposure effect as a weighted average of model-specific effect estimates, again using the model’s posterior probabilities as weights. Viallefort et al [164] applied this method to estimate an exposure’s odds ratio in case-control studies. Other applications include air pollution research [41, 101].

However, while effective in some cases, traditional implementations of BMA, with an uninformative prior on the inclusion indicator of each potential confounder in the regression model can face severe limitations [45, 166]. Most of these can be traced to the fundamental difficulty arising with the fact that the regression coefficient, representing the effect of exposure to air pollution on a health outcome, may have a different interpretation across models that include a different set of covariates to adjust for confounding [43]. Crainiceanu et al [45] noted that model uncertainty methods useful in prediction may not generally
They introduced a two-step approach (CDP) to estimate an exposure effect accounting for adjustment uncertainty. In the first step, this approach regresses exposure on a large set of potential confounders and selects confounders that are associated with exposure. In the second step, it regresses outcome on exposure, after including the confounders identified in the first step. Compared to this approach, traditional BMA with non-informative priors on whether or not include a covariate into the regression model to adjust for confounding did not perform well. This is because the posterior model probabilities obtained from non-informative priors will assign high weights to models that include covariates that are strong predictors of the outcome only, and will assign low weights to models that include strong predictors of the exposure and weak predictors of the outcome. It is well known that when we omit these variables in the regression model, the health effect estimates will be affected by confounding bias (see [166] for a compelling example that illustrates this point).

Wang et al [165, 166] have developed a novel Bayesian approach which we call Bayesian Adjustment for Confounding (BAC) to adjust for confounding and account for uncertainty in the choice of confounders in the context where we estimate the health effect of air pollution using a GLM and we adjust for confounding by including covariates into the regression model as a linear term. Here, we describe in detail the approach proposed by Wang et al [165, 166]. Other more sophisticated approaches that account for model uncertainty in the confounding adjustment have been developed by Zigler et al in the context of propensity score methods [195] and by Cefalu et al in the context of double robust estimation [33].

### 24.4.1 Bayesian Adjustment for Confounding

We start by introducing some notation. We denote by $X$ the exposure and $Y$ the outcome. We also assume that we have information on a set of $p$ potential confounders $Z = \{Z_1, \ldots, Z_p\}$ identified because they are likely to affect $Y$, though their effects could be weak. A priori, there may be uncertainty about whether potential confounders should be adjusted for in effect estimation.

Though many of our ideas are more general, we discuss our approach in the context of two linear regression models: one for exposure and one for outcome. In each equation, potential confounders are either included or excluded, depending on unknown vectors of indicators $\alpha^X \in \{0, 1\}^p$ and $\alpha^Y \in \{0, 1\}^p$. Here $\alpha^X_j = 1$ (or $\alpha^Y_j = 1$) whenever $Z_j$ is included in the exposure (or outcome) model. For brevity, we refer to the parameters $\alpha^X$ and $\alpha^Y$ as “models”. Conditional on unknown parameters, and confounders, the regression equations for exposure $X_i$ and outcome $Y_i$ are

$$E\{X_i\} = \sum_{j=1}^{p} \alpha^X_j \delta^X_j Z_{ij} \quad (24.3)$$

$$E\{Y_i|X_i\} = \beta^Y X_i + \sum_{j=1}^{p} \alpha^Y_j \delta^Y_j Z_{ij} \quad (24.4)$$

where $i$ indexes the sampling unit. For regression coefficients, $\beta$ and $\delta$, we use a notation that explicitly keeps track of the fact that those coefficients differ in meaning with $\alpha^X$ and $\alpha^Y$. This is especially important when one attempts to make inferences that involve estimates of the exposure effect obtained using different models. Intercept columns can be included among the $Z$s. Some $\alpha^Y_m$s can be fixed with value 1 if confounders are deemed required a priori.

In developing a model for effect estimation, when a true confounder is added or removed from the regression model, the interpretation of the exposure coefficient changes; however,
Statistical Methods for Environmental Epidemiology 563

when a model includes all true confounders, and one adds an additional variable that is not associated with X or that is neither associated with X nor Y, the interpretation of the exposure coefficient does not change. This is in contrast to prediction, where the predicted quantities typically maintain the same interpretation across models.

Thus, when studying confounding adjustment, it is useful to consider the smallest outcome model that includes all the necessary confounders. We denote it by \( \alpha^* \), and refer to it as the minimal model. The estimand of interest — the true effect of X on Y, is the coefficient of X in this model, or \( \beta_* = \beta^{\alpha^*} \). If there are interactions between exposure and confounders, the estimands are model coefficients of both the main effect and the interaction terms. See Wang et al (2015) \[165\] for details regarding a BMA approach to account for model uncertainty in both the selection of confounders and in the selection of interaction terms between exposure variable and each of the potential confounders.

We will focus on the situation where there are no interaction terms. Our goal is estimation of \( \beta_* \) when \( \alpha^*_Y \) is unknown. A key observation is that all models that contain at least the confounders as in the minimal model will provide estimates of the exposure effect that are also interpretable as estimates of \( \beta_* \). On the other hand, a model that does not include the minimal model, that is, a model that excludes at least one true confounder, will provide estimates of a parameter that are not the estimated parameter of interest. Hence, the resulting exposure effect estimate is confounded.

The importance of including in the outcome model all the potential confounders that belong to the minimal model suggests that an approach that acknowledges the fact that only a fraction of the models harbor the coefficient of interest with the correct interpretation, could be successful in addressing adjustment uncertainty from a Bayesian standpoint. BAC jointly considers the exposure and outcome models, as in equations (24.3) and (24.4), and includes unknown model selection parameters \( \alpha^X \) and \( \alpha^Y \).

The authors specify a prior distribution on \( \alpha^Y | \alpha^X \) such that

\[
\frac{P(\alpha^Y_j = 1 | \alpha^X_j = 1)}{P(\alpha^Y_j = 0 | \alpha^X_j = 1)} = \omega, \quad \frac{P(\alpha^Y_j = 1 | \alpha^X_j = 0)}{P(\alpha^Y_j = 0 | \alpha^X_j = 0)} = 1, \quad j = 1, \ldots, p, \tag{24.5}
\]

where \( \omega \in [1, \infty) \) is a dependence parameter denoting the prior odds of including \( Z_j \) into the outcome model when \( Z_j \) is included in the exposure model. When \( \omega = \infty \), the first equation in (24.5) becomes \( P(\alpha^Y_j = 1 | \alpha^X_j = 1) = 1 \), and requires that any \( Z_j \) for which \( \alpha^X_j = 1 \) is automatically included in the outcome model. When \( 1 < \omega < \infty \), our prior on \( \alpha^Y | \alpha^X \) provides a chance to rule out the predictors that are only associated with X but not associated with Y.

The conditional prior of \( \alpha^Y \) given \( \alpha^X \) in (24.5) plays a key role in approximating the marginal posterior distribution of the exposure coefficient under the minimal model, \( \beta_* \),

\[
P(\beta_*|D) = \sum_{\alpha^Y} P(\beta_*|\alpha^Y, D) P(\alpha^Y|D),
\]

where \( D = (X, Y) \) contains vectors of observed data for the exposure and the outcome. Our analysis is also conditional on observed data for potential confounders \( Z_1, \ldots, Z_j \), and they will not be noted in posteriors for simplicity of notation. When \( \omega \) is large, the conditional prior in (24.5) greatly increases the chance for predictors strongly correlated with X to be included in the outcome model. These predictors are confounders if they are also correlated with Y. Therefore, the prior leads to a posterior distribution of \( \alpha^Y \) \( (P(\alpha^Y|D)) \) that assigns mass mostly to models that are fully adjusted for confounding, that is, models containing the minimal model. For these models, \( \beta^{\alpha^Y} = \beta_* \) so that \( P(\beta_*|\alpha^Y, D) = P(\beta^{\alpha^Y}|\alpha^Y, D) \).
Therefore, approximately,

\[ P(\beta_*|D) \cong \sum_{\alpha^Y} P(\beta^{\alpha^Y}|\alpha^Y, D)P(\alpha^Y|D), \]  

(24.6)

where \( P(\beta^{\alpha^Y}|\alpha^Y, D) \) can be directly estimated from observed data. This approximation will be further discussed in Section 24.4.2.

Our goal is to calculate the posterior distribution of the parameters of interest \((\alpha^X, \alpha^Y, \beta_*)\) in equations (24.3) and (24.4). Details are in Wang et al 2012 [166].

### 24.4.2 Relation to BMA

In the context of effect estimation, several authors [83, 130] suggested to calculate the posterior distribution of the effect by taking an average over models, weighted by their posterior probabilities:

\[ \sum_{\alpha^Y} P(\beta^{\alpha^Y}|\alpha^Y, D)P(\alpha^Y|D). \]  

(24.7)

This corresponds to marginalization according to the law of total probabilities, but only if the parameters \( \beta^{\alpha^Y} \) have the same interpretation.

From the perspective of adjustment uncertainty, (24.7) can be decomposed into two parts: the sum over models that include the correct estimand, and the rest. That is

\[ \sum_{\alpha^Y \supseteq \alpha^Y_*} P(\beta_*|\alpha^Y, D)P(\alpha^Y|D) + \sum_{\alpha^Y \not\supseteq \alpha^Y_*} P(\beta^{\alpha^Y}|\alpha^Y, D)P(\alpha^Y|D). \]  

(24.8)

where \( \alpha^Y \supseteq \alpha^Y_* \) indicates that model \( \alpha^Y \) contains all the variables that are also contained in model \( \alpha^Y_* \). The second term of (24.8) averages across models that do not include \( \alpha^Y_* \), and therefore do not estimate the same effect.

In BMA one needs to be careful about not assigning large weights to the models in the second term of equation (24.8). A common practice in traditional implementations of BMA is to use uniform, or highly dispersed, priors on the \( \alpha^Y \)’s and often on the effect of interest as well. When the prior is the same for all models, the ratio of the weights given to models \( \alpha_1 \) and \( \alpha_2 \) is the Bayes Factor \( P(Y|\alpha_1)/P(Y|\alpha_2) \) [88] and the posterior model probabilities in BMA are driven by a model’s predictive ability, which may differ from its ability to properly adjust for confounding in effect estimation.

### 24.4.3 Air Pollution Example

In this section, we briefly summarize how the authors applied BAC to daily time series data for Nassau County, NY for the period 1999-2005. Although this data analysis was mainly used as an illustration of BAC, the results clearly illustrate the potential application and impact of BAC in epidemiology studies of observational data. The data include 1,532 daily records of emergency hospital admissions, weather variables, and PM\(_{2.5}\) levels. A more extensive description of this data set can be found in [61]. The goal was to estimate the increase in the rate of hospitalizations for cardiovascular disease (CVD) associated with a 10\(\mu\)g/m\(^3\) increase in PM\(_{2.5}\), while accounting for age-specific longer-term trends, weather and day of the week. The hospitalization rate was calculated separately for each age group (\(\geq 75\) or not) on each day. In the outcome model, to control for longer-term trends due, for example, to changes in medical practice patterns, seasonality and influenza epidemics, the authors included smooth functions of calendar time. They also included a smooth function to allow seasonal variations to be different in the two age groups. To control for the weather
Statistical Methods for Environmental Epidemiology

565

effect, they include smooth functions of temperature and dew point. The outcome model that includes all the necessary confounders can be defined below [59, 62, 122]

\[ Y_{at} = \beta \text{PM}_{2.5t} + \text{DOW} + \text{intercept for age group } a \]

\[ + ns(\text{Temp}_t, df_{\text{Temp}}) + ns(\text{Temp}_{t-1-3}, df_{\text{Temp}}) + ns(\text{Dew}, df_{\text{Dew}}) \]

\[ + ns(\text{Dew}_{t-1-3}, df_{\text{Dew}}) + ns(t, df_t) + ns(t, df_{at}) \times \text{age group} + \epsilon_t, \]

where the outcome \( Y_{at} = \sqrt{\text{CVD hospital admissions/size of population at risk for each age group } a \ (\geq 75 \text{ or not}) \text{ on day } t \ (t = 1, \ldots, 1532)}. \) PM\(_{2.5t}\) denotes the level of particulate matter having diameter less than 2.5 micrometer on day \( t \). DOW are indicator variables for the day of the week. \( \text{Temp}_t \) and \( \text{Temp}_{t-1-3} \) are the temperature on day \( t \) and the three-day running mean, respectively. \( \text{Dew}_t \) and \( \text{Dew}_{t-1-3} \) are the dew point on day \( t \) and the three-day running mean. The quantity \( ns(\cdot, df) \) is a natural cubic spline with \( df \) degrees of freedom. The authors included \( ns(t, df_t) \), \( ns(\text{Temp}_t, df_{\text{Temp}}) \), \( ns(\text{Temp}_{t-1-3}, df_{\text{Temp}}) \), \( ns(\text{Dew}, df_{\text{Dew}}) \) and \( ns(\text{Dew}_{t-1-3}, df_{\text{Dew}}) \) in the outcome model to adjust for the potential nonlinear confounding effects of seasonal variations, temperature and dew point. The quantity \( ns(t, df_{at}) \times \text{age group} \) is a natural cubic spline of \( t \) for the \( \geq 75 \) age group to allow its seasonal variation to be different from the other age group. Similar to [45], \( df_{\text{Temp}} \) was set to 12, \( df_{\text{Dew}} \) was set to 12, \( df_t \) was set to 16 per year, and \( df_{at} \) was set to 4. These degrees of freedom were considered sufficiently large for the full model to include all the potential confounders [45]. The residuals \( \epsilon_t \) were assumed to be independent and identically distributed with a normal \( N(0, \sigma^2) \) distribution. After dropping some potential confounders due to collinearity, the authors considered 164 potential confounders.

Several approaches were used to analyze the data: BAC, CDP [45], FBMA (BMA with the exposure variable always forced into the outcome model), NBMA (BMA with the exposure variable not forced into the outcome model), and stepwise. For BAC, the authors considered priors with \( \omega = 2, 4, 10 \text{ or } \infty \). The estimated PM\(_{2.5} \) effect (\( \times 10^4, 000 \)) denoted by \( \hat{\beta} \) is listed in Table 24.1: BAC (with \( \omega = \infty \)) and CDP provide estimates of the short-term effect of PM\(_{2.5} \) on CVD hospital admissions with 95% CIs that do not include 0. With \( \omega = \infty \), BAC provided similar estimates of the exposure effect as CDP. Moreover, all three methods provided smaller standard errors than the one obtained under the full model. In comparison, FBMA and NBMA provided a very different and not statistically significant estimate of the exposure effect. Some confounders known to be important, such as temperature and dew point, were down weighted in BMA. Both temperature and dew point were positively correlated with PM\(_{2.5} \) and negatively correlated with hospitalization rate. Failure to include them in the model diminished the PM\(_{2.5} \) effect. This example illustrated that in practical applications BMA and BAC can lead to different conclusions. The key difference lies in the linking strength between the exposure model and the outcome model. As the strength decreases, which corresponds to smaller value of \( \omega \), the estimates from BAC become closer to that from BMA.

### 24.4.4 Concluding Remarks

Estimating an exposure effect, while accounting for the uncertainty in the adjustment for confounding, is of essential importance in observational studies. Building upon earlier work [45, 59], in BAC the authors have developed Bayesian solutions to the estimation of the association between \( X \) and \( Y \) accounting for the uncertainty in the confounding adjustment. Given a set of potential confounders, the authors simultaneously address model selection for both the outcome and the exposure. While they presented their methods in the setting of linear models, BAC is a general concept and is not constrained to the linear case. These
TABLE 24.1
Comparison of estimates of PM$_{2.5}$ effect on CVD hospitalization rate based on BAC, CDP, FBMA, NBMA, stepwise, and the full model. (Table from Wang et al 2012)

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\beta}$</th>
<th>SE($\hat{\beta}$)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>0.291</td>
<td>0.092</td>
<td>(0.110, 0.471)</td>
</tr>
<tr>
<td>BAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega = \infty$</td>
<td>0.226</td>
<td>0.081</td>
<td>(0.067, 0.385)</td>
</tr>
<tr>
<td>$\omega = 10$</td>
<td>0.217</td>
<td>0.079</td>
<td>(0.060, 0.371)</td>
</tr>
<tr>
<td>$\omega = 4$</td>
<td>0.186</td>
<td>0.085</td>
<td>(0.019, 0.351)</td>
</tr>
<tr>
<td>$\omega = 2$</td>
<td>0.155</td>
<td>0.079</td>
<td>(0.007, 0.317)</td>
</tr>
<tr>
<td>CDP</td>
<td>0.221</td>
<td>0.089</td>
<td>(0.045, 0.396)</td>
</tr>
<tr>
<td>FBMA</td>
<td>0.140</td>
<td>0.077</td>
<td>(−0.008, 0.298)</td>
</tr>
<tr>
<td>NBMA</td>
<td>0.007</td>
<td>0.033</td>
<td>(0.000, 0.131)</td>
</tr>
<tr>
<td>Stepwise</td>
<td>0.106</td>
<td>0.066</td>
<td>(−0.023, 0.234)</td>
</tr>
</tbody>
</table>

approaches have been extended to generalized linear models using relatively well understood computational strategies [165].

Like BMA, BAC take a weighted average over models rather than making inference based on a single model. However, they attempt to provide an estimate of the exposure effect by combining information across regression models that include all the requisite confounders, to ensure that the regression coefficient of interest maintains the same interpretation across models. A nice feature of BMA that is retained by BAC is that the importance of confounders can be evaluated based on posterior inclusion probability. This information may reveal underlying connections between exposure and confounders, which may become of interest for future research. BAC is more computationally intensive than BMA.

Successful application of BAC relies on availability of all confounders. Scientific knowledge is required to ensure that these assumptions are valid. Statistical methods may also help to check whether there is evidence for the existence of unmeasured confounders. If there are no unmeasured confounders, the full model, that is, the model including all variables correlated with $X$ and $Y$, those correlated with $Y$ only, as well as potentially others that are not associated with either, will provide unbiased estimates of the exposure effect. However, using the full model will generally yield wider confidence intervals compared to BAC. By combining estimations from different smaller models, especially from models that only include requisite confounders but do not include many unnecessary variables, BAC can provide more precise inference than the full model.

In the propensity score literature, it is recommended to include variables that are strongly correlated with $Y$ but only weakly correlated with $X$ into the model for calculating the propensity score, as the bias resulting from their exclusion would dominate any loss of efficiency in modest or large studies [30, 137]. One of the strengths of the BAC method, shared by others such as doubly robust estimation [143], is that BAC can identify these in a data-based way, rather than having to rely on prior knowledge as required in propensity score adjustment.

Zigler et al introduced a new set of approaches to account for adjustment uncertainty when the adjustment for confounding is done by using propensity score (PS) methods [195]. In causal inference, typically, simple or ad hoc methods are employed to arrive at a single propensity score model, without acknowledging the uncertainty associated with the model selection. Zigler et al introduced three Bayesian methods for PS variable selection and model averaging that (a) select relevant variables from a set of candidate variables to include in
the PS model and (b) estimate causal treatment effects as weighted averages of estimates under different PS models. The associated weight for each PS model reflects the data-driven support for that model’s ability to adjust for the necessary variables [195]. One drawback of application of these approaches is that they assume that the treatment assignment or the exposure is binary, whereas the exposure to air pollution is a continuous variables. Extensions of these approaches in the context of generalized propensity score methods for multi-level treatments and/or continuous treatments [179, 191] are needed.

In this section, while we do not take a causal inference perspective, the methods illustrated here have several points of contacts with causal inference methodologies that are based on joint modeling of exposure and outcome as functions of confounders [135, 136] and with their Bayesian counterparts [112]. This literature strongly emphasizes, as we do, the critical role of model specification and the need for robustness to the choice of confounders [9, 71, 137]. From this perspective, the BAC approach and extensions of these ideas in the context of propensity score matching [195] and double robust estimation [33] achieves a combination of three desirable properties: effect estimation efficiency, via the exposure model; variable selection robustness, achieved by allowing the selection to be a random variable; and bias reduction, achieved by including prior information to favor predictors of exposure in the selection of variables for the outcome model.

24.5 Estimation of Health Effects From Simultaneous Exposure to Multiple Pollutants

The majority of environmental epidemiology has estimated the health effects of exposure to a single pollutant. However, populations are in reality exposed to a diverse mixture of pollutants. Recent research in environmental epidemiology has turned to estimate the health effects of mixtures of multiple pollutants.

In this setting, for each individual \( i \) we observe the exposure level for a vector of \( M \) pollutants \( X_i = (X_{i1}, \ldots, X_{iM})^T \). As \( M \) increases, the exposure-response function \( f(X) \), as in (24.2), becomes increasing more complex. In particular, there may be a nonlinear exposure-response relationship for each of the \( M \) pollutants as well as interactions between pollutants. In this setting, the curse of dimensionality makes estimating the exposure-response function a challenging statistical problem.

Several methods have been proposed to estimate the health effects of mixtures. In this section we highlight two recently developed methods that take different approaches to the multi-pollutant problem. The first approach was developed in the context of air pollution epidemiology and compares the health effect of ambient exposures on days with different “multi-pollutant profiles” [7, 185]. This can address questions such as are the health effects of PM\(_{2.5}\) different when the particulates come from fresh local traffic pollution verses regionally transported pollution, for example. The second approach takes a flexible machine learning approach to estimate a high-dimensional exposure-response surface [28]. This second approach can estimate the health effects at any level of exposure to a high-dimensional vector of pollutants. We conclude with a discussion of confounding adjustment in the context for multi-pollutant models.
24.5.1 Multi-Pollutant Profile Clustering and Effect Estimation

The multi-pollutant profiling approach to estimating the health effect of multi-pollutant mixtures is built on a two step procedure [7, 185]. The first step is to identify clusters of days or locations that have a similar multi-pollutant profile. For example, days with pollution profiles dominated by primary emissions from local traffic or those high in regional power-plant emissions. The second step is to estimate the health effects of exposure to each cluster and compare the relative toxicities of the clusters.

There are ample statistical methods for clustering [25, 52, 78, 92]. Here will will focus on the $k$-means algorithm as proposed for multi-pollutant clustering [7, 185]. The $k$-means algorithm was originally developed by Hartigan and Wang [78] and is appealing due to the computational efficiency. The proposed approach clusters days or locations based on the measured concentrations of $M$ pollutants. The general concept is to partition the $n$ observations of $M$ pollutants, $\{X_i = (X_{i1}, \ldots, X_{iM})^T\}_{i=1}^n$, into $k$ clusters. For each cluster there is a mean points ($k$ means in total) and each observations is nearest in euclidean distance to its clusters mean.

In the context of environmental epidemiology, $k$-means (or any clustering approach) can identify clusters of similar exposures. For example, Austin et al [7] found five clusters each consisting of days in Boston, MA, with one of five air pollution profiles: 1) “low particles–high ozone” contained lower concentrations of PM$_{2.5}$ and higher than average levels of O$_3$; 2) “crustal” which had higher levels of Si, Ca, Br, and Ti; 3) “winter primary” which contained higher levels of Ni, V, Zn, Br, CO, NO and SO$_2$; 4) “regional summer” which was heavy in regionally transported pollutants; and 5) “winter–low primary, higher O$_3$” which is similar to cluster 3 but with lower PM$_{2.5}$ and higher O$_3$ levels.

The challenge to $k$-means clustering is determining the optimal number of clusters. Ideally, the researcher would have prior knowledge about the number of clusters. However, this is often unknown and methods have been developed to select $k$. One such approach is the select the value of $k$ that minimizes the Davies-Bouldin (DB) index [50]. The DB index measures cluster compactness verse the distance between cluster centers. It is expressed as

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max \left\{ \frac{\sigma_i + \sigma_j}{\| \mu_i - \mu_j \|} \right\},$$

where $\mu_i$ is the center of cluster $i$, $\sigma_i$ is the average distance between $\mu_i$ and the observations in cluster $i$, and $n$ is the number of clusters. The choice of $k$ that minimizes (24.9) will be have compact clusters with well separated cluster means relative to other choices of $k$.

The second step of the multi-pollutant clustering and effect estimation approach is to estimate the relative toxicities of the clusters. Zanobetti et al [185] provide a nice illustration of this by interacting indicators of membership into five clusters with total PM$_{2.5}$ mass. Here $C_{tj} = 1$ if day $t$ had air pollution consistent with profile $j$ and $C_{tj} = 0$ otherwise. The proposed Poisson time-series model is then

$$g(E(Y_t|X_t, C_t, Z_t)) = \beta_0 + X_t\beta + \sum_{j=2}^{5} C_{tj}\gamma_j + \sum_{j=2}^{5} X_tC_{tj}\delta_j + \sum_{k=1}^{r} Z_{tk}\eta_j,$$

where cluster one is omitted as a reference category. The model can then be interpreted as the effect of total PM$_{2.5}$ mass ($X$ in this model) on mortality ($Y$ in this example) specifically on days with each of the five cluster profiles. Hence, we can estimate and test differences in the toxicity of PM$_{2.5}$ with various compositions.

This approach is appealing on several levels. First both the cluster and analysis stages can be completed with standard methods and existing software. Second, the approach is
easily interpretable because the potentially high-dimensional exposure profile is reduced to a categorical variable indicating cluster membership. However, the analysis is limited to the cluster-level and interpretability relies on identifying clusters that are scientifically meaningful.

24.5.2 High-Dimensional Exposure-Response Function Estimation

Another approach to directly estimate the multi-pollutant exposure-response functions. Under this approach, the regression model is

\[
g\{E(Y_i|X_i, Z_i)\} = f(X_i) + \sum_{k=1}^{r} Z_{ik}\gamma_j, \tag{24.11}\]

where \( f : \mathbb{R}^M \rightarrow \mathbb{R} \) is the multi-pollutant exposure-response function. Several approaches have been used to estimate nonlinear and non-additive multivariate exposure-response functions. These approaches include spline or polynomial expansions [36, 175], loess [132], and machine learning approaches [28]. Here, we focus on Bayesian kernel machine regression (BKMR) to estimate the multivariate-exposure-response surface as introduced by Bobb et al [28]. Compared to many alternative approaches, BKMR offers the advantage of easily scaling with the number of pollutants (\( M \)) while allowing for nonlinear effects and high order interactions.

Under the BKMR approach, we assume that the exposure-response function \( f \) resides in a function space with an associated positive semidefinite reproducing kernel \( K : \mathbb{R}^M \times \mathbb{R}^M \rightarrow \mathbb{R} \). The exposure-response function \( f \) can then be represented in either of two forms. First is the primal form where \( f \) can be represented with a basis expansions \( f(x) = \sum_{l=1}^{L_f} \phi_l(x)\theta_l \), where \( \{\phi_l(x)\}_{l=1}^{L_f} \) is a set of basis functions and \( \{\theta_l\}_{l=1}^{L_f} \) are regression coefficients.

The second form for representing \( f \), which is used here, is the dual form through the kernel \( K(\cdot,\cdot) \) as \( f(x) = \sum_{i=1}^{n} K(x_i,x)\alpha_i \), where \( \{\alpha_i\}_{i=1}^{n} \) are unknown regression coefficients. According to Mercer’s theorem [46] a kernel \( K(\cdot,\cdot) \) in the dual form corresponds to a unique set of orthogonal basis functions in the primal representation. Here, we consider the Gaussian kernel \( K(x,x') = \exp\{-\sum_{m=1}^{M} \rho_m (X_m - X'_m)^2\} \), where \( \{\rho_m\}_{m=1}^{M} \) are tuning parameters. The Gaussian kernel corresponds to a set of radial basis functions.

For the multi-pollutant normal linear regression, the dual representation can be represented as the mixed model [110]

\[
Y_i \sim N(f_i + Z_i r, \gamma, \sigma^2) \tag{24.12}
\]
\[
f \sim N(0, \tau K). \tag{24.13}
\]

In (24.13), \( f = (f_1, \ldots, f_n)^T \), \( K \) is a \( n \times n \) kernel matrix where the \((i,j)\) element is \( K_{i,j} = \exp\{-\sum_{m=1}^{M} \rho_m (X_{im} - X'_{jm})^2\} \), and \( \tau \) is a tuning parameter. The mixed model can then be solved with existing mixed model software [110] or through Bayesian methods [28] to estimate the multi-pollutant exposure-response function.

As illustrated in the original work [28], the dual form naturally scales as the number of exposures increases. At the same time, BKMR allows for a high-dimensional exposure-response function estimation. This exposure-response surface can include nonlinearities and complex interactions between two or more of the \( M \) exposures.

24.5.3 Confounding Adjustment in Multiple Pollutant Models

As with this single pollutant models discussed in Section 24.3, confounding is an important issue in multi-pollutant exposure-response modeling. Recent papers have addressed
confounding adjustment specifically in the multi-pollutant context [176, 177]. In particular, recent work highlights an important distinction between confounding adjustment when studying the health effects of a single pollutant and when studying the health effects of multiple pollutants [177]. When estimating the effects of multiple pollutants we are often interested in both main effects and interactions between pollutants. In order to properly adjust for confounding it is essential to not only consider confounders that are associated with any one pollutant and the outcome but to identify the set of confounders that are associated with the multivariate exposure (all main effects and interactions of interests) and outcome.

To address confounder adjustment when estimating the health effects of simultaneous exposure to multiple pollutants, a recent paper proposed using a Bayesian model averaging approach for estimating the health effects of simultaneous exposure to multiple pollutants (including main effects and interactions) that, at the same time, account for model uncertainty in the confounding adjustment [177]. More specifically, to adjust for confounding, the authors develop an informative prior on covariate inclusion. They apply the proposed method to the National Health And Nutrition Survey data (NHANES) data to estimate the effects of exposure to 132 nutrients and persistent pesticides grouped into 24 groups on lipid levels. When $p$ is close to $n$, multivariate exposure effect estimates with the proposed approach are fully adjusted for confounding and has smaller variance than the same estimate obtained under a model that includes all available covariates.

Importantly, their proposed method uses information from the relationship between the potential confounders and the multivariate exposure to construct the prior. This newly proposed method offers several advantages over previously developed confounder adjustment methods. First, this approach scales as the dimension of the multivariate exposure vector increases. Previously proposed confounder adjustment approaches have relied on exposure modeling [165, 166, 176] and would require an additional exposure model for each additional agent. This is impractical when the number of individual agents is large. For example, the one exposure group in above data analysis with 22 agents would require 22 separate exposure models using exposure modeling approaches [165, 166, 176], an infeasible number of exposure models. A second advantage of the proposed method is that it explicitly addresses confounding of the multivariate exposure effect. It is indeed important to point out that the confounders of the single agents (for example exposure to PM$_{2.5}$) are different from the confounders of the multivariate exposure effect (for example health effects for simultaneous exposure to PM$_{2.5}$, O$_3$, NO$_2$ and all their pairwise interactions). A third advantage is the specific guidance provided for tuning the strength of the prior on covariate inclusion (in Wang et al this parameter was denoted by $\omega$ as shown in (24.5)). In BAC, the authors need to specify a priori a value for $\omega$. There is a similar tuning parameter in Wilson et al [177]; however, the authors introduce an approach that allows identification of a suitable value of the tuning parameter that balances model parsimony and confounder adjustment.

With increased availability of high-dimensional exposure data (called exposome) there is growing interest in understanding the effect of the simultaneous exposure to multiple environmental agents on complex diseases. However, estimating a high-dimensional exposure-response function is statistically challenging and there is often uncertainty as to which covariates to include in the model to estimate the multivariate exposure effect. The methods described in this section fill methodological gaps in the area of estimating the health effects of simultaneous exposure to multiple pollutants. These methods are valuable tools that can reliably be used to estimate health effects of mixtures while simultaneously allowing a rigorous adjustment for confounding and guarantee model parsimony.
Bibliography


Michelle L. Bell, Keita Ebisu, Brian P. Leaderer, Janneane F. Gent, Hyung Joo Lee, Petros Koutrakis, Yun Wang, Francesca Dominici, and Roger D. Peng. Associations of PM2.5 Constituents and Sources with Hospital Admissions: Analysis of Four Counties in Connecticut and Massachusetts (USA) for Persons 65 Years of Age. Environmental Health Perspectives, 122(2):138–144, nov 2013.


[95] Marianthi-Anna Kioumourtzoglou, Donna Spiegelman, Adam A Szpiro, Lianne Sheppard, Joel D Kaufman, Jeff D Yanosky, Ronald Williams, Francine Laden, Biling


