

# An Approach to the Estimation of Chronic Air Pollution Effects Using Spatio-Temporal Information

Sonja GREVEN, Francesca DOMINICI, and Scott ZEGER

There is substantial observational evidence that long-term exposure to particulate air pollution is associated with premature death in urban populations. Estimates of the magnitude of these effects derive largely from cross-sectional comparisons of adjusted mortality rates among cities with varying pollution levels. Such estimates are potentially confounded by other differences among the populations correlated with air pollution, for example, socioeconomic factors. An alternative approach is to study covariation of particulate matter and mortality across time within a city, as has been done in investigations of short-term exposures. In either event, observational studies like these are subject to confounding by unmeasured variables. Therefore the ability to detect such confounding and to derive estimates less affected by confounding are a high priority.

In this article, we describe and apply a method of decomposing the exposure variable into components with variation at distinct temporal, spatial, and time by space scales, here focusing on the components involving time. Starting from a proportional hazard model, we derive a Poisson regression model and estimate two regression coefficients: the “global” coefficient that measures the association between national trends in pollution and mortality; and the “local” coefficient, derived from space by time variation, that measures the association between location-specific trends in pollution and mortality adjusted by the national trends. Absent unmeasured confounders and given valid model assumptions, the scale-specific coefficients should be similar; substantial differences in these coefficients constitute a basis for questioning the model.

We derive a backfitting algorithm to fit our model to very large spatio-temporal datasets. We apply our methods to the Medicare Cohort Air Pollution Study (MCAPS), which includes individual-level information on time of death and age on a population of 18.2 million for the period 2000–2006.

Results based on the global coefficient indicate a large increase in the national life expectancy for reductions in the yearly national average of  $PM_{2.5}$ . However, this coefficient based on national trends in  $PM_{2.5}$  and mortality is likely to be confounded by other variables trending on the national level. Confounding of the local coefficient by unmeasured factors is less likely, although it cannot be ruled out. Based on the local coefficient alone, we are not able to demonstrate any change in life expectancy for a reduction in  $PM_{2.5}$ . We use additional survey data available for a subset of the data to investigate sensitivity of results to the inclusion of additional covariates, but both coefficients remain largely unchanged.

**KEY WORDS:** Backfitting algorithm; Environmental epidemiology; Particulate matter; Spatio-temporal data; Specification test.

## 1. INTRODUCTION

The Clean Air Act (Environmental Protection Agency, last amended in 1990) requires the U.S. Environmental Protection Agency (EPA) to set National Ambient Air Quality Standards for seven pollutants considered harmful. Air quality standards for several air pollutants have since also been adopted by the European Union. Implementation of these standards led to decreases in air pollution concentrations in the United States (Bachmann 2008). From a public policy and public health perspective, it is of importance to assess whether these decreases have also led to an improvement in morbidity and mortality for the general population (Health Effects Institute 2003). Standards are reviewed periodically, with evidence from epidemiologic studies playing a large role in the public policy process (Kaiser 1997; Greenbaum et al. 2001; Samet et al. 2003). While there is substantial observational evidence that long-term

exposure to particulate air pollution is associated with premature death in urban populations, confounding by unmeasured variables remains a large concern in observational studies. The ability to detect such confounding and to derive estimates less affected by confounding thus are of great importance.

Evidence on the magnitude of the chronic effects of long-term exposure to air pollution on mortality stems mostly from cohort studies (see, e.g., Dockery et al. 1993; Pope et al. 2002; Laden et al. 2006; Eftim et al. 2008). These studies compare across locations long-term average air pollution concentrations and time-to-death in cohorts. Cohort studies allow the estimation of life expectancy lost due to air pollution (Künzli et al. 2001; Rabl 2003). They have been criticized (Moolgavkar 1994; Vedal 1997; Gamble 1998), due to the difficulty of fully accounting for all potential confounders, including individual risk factors and location-specific characteristics such as socioeconomic factors.

An alternative approach is to study covariation of particulate matter and mortality across time within a predefined geographical location (e.g., county or city), as has been done in investigations of health effects associated with short-term exposures. Time series studies (see, e.g., Schwartz and Dockery 1992; Spix et al. 1993; Kelsall et al. 1997) estimate acute effects of short-term exposure to air pollutants, comparing day-to-day variations in mortality with those in air pollution concentrations. Multisite time series studies (Katsouyanni et al. 1997;

Sonja Greven is Emmy Noether Junior Research Group Leader, Department of Statistics, Ludwig-Maximilians-Universität München, 80539 Munich, Germany (E-mail: [sonja.greven@stat.uni-muenchen.de](mailto:sonja.greven@stat.uni-muenchen.de)). Francesca Dominici is Professor, Department of Biostatistics, Harvard University, Boston, MA 02115. Scott Zeger is Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205. Funding was provided by the U.S. Environmental Protection Agency (EPA) (grants RD-83241701 and RD-83362201). Although the research described in this article has been funded wholly or in part by the U.S. EPA, it has not been subjected to the agency's required peer and policy review and does not necessarily reflect the views of the agency, and no official endorsement should be inferred. The first author was also funded by Emmy Noether grant GR 3793/1-1 from the German Research Foundation. The authors thank the referees, associate editor, and editor for helpful comments, and Aidan McDermott for help with the datasets.

Samet et al. 2000; Samoli et al. 2008; Wong et al. 2008) combine the evidence and statistical uncertainty across geographical locations (Dominici, Samet, and Zeger 2000; Dominici 2002). Due to the focus on short-term effects, time series studies do not allow an assessment of the years of life-time lost due to air pollution (Künzli et al. 2001). Potential confounders in time series studies are time-varying variables such as weather or seasonal effects, as well as slowly varying unmeasured factors. Typically, smooth functions of weather variables and calendar time are included in the regression model to account for temporal confounding. However, results have been found to be sensitive to the flexibility granted to these smooth functions (Samoli et al. 2001; Klemm and Mason 2003; Dominici, McDermott, and Hastie 2004; Peng, Dominici, and Louis 2006), and time series studies have also been criticized with regard to potential residual confounding (Vedal 1997; Lumley and Sheppard 2000; Moolgavkar 2005).

In this article, we develop a statistical approach for estimating chronic effects associated with long-term exposure to air pollution. We use available spatio-temporal information from large national databases to estimate two types of association between  $PM_{2.5}$  and mortality. The first type measures whether, on average across the nation, there is an association between the long-term trend in  $PM_{2.5}$  and the long-term trend in age-adjusted mortality rates (purely temporal association). The second type measures whether cities exhibiting a more rapid decline in  $PM_{2.5}$  also show a faster decline in mortality (residual spatio-temporal association, adjusting for purely temporal and spatial associations). We do not focus here on a third type used in cohort studies, measuring the association between average  $PM_{2.5}$  levels and average age-adjusted mortality rates across cities (purely spatial or cross-sectional association). We decompose the  $PM_{2.5}$  exposure variable into two components and estimate two regression coefficients. This decomposition allows us to assess whether the strength of the evidence on the association between  $PM_{2.5}$  and mortality is consistent across the time and space  $\times$  time scale in this large and complex dataset. In fact, absent confounding or other model misspecification, the two estimates should be similar, and large differences thus may indicate confounding of one or both estimates. This approach is related to specification tests in econometrics (Hausman 1978).

Starting from a proportional hazards model, we derive a Poisson regression model and estimate two regression coefficients. We derive a backfitting algorithm that makes use of the specific model structure to obtain an efficient implementation of our approach. This enables the fitting of our model to very large spatio-temporal datasets. We evaluate spatio-temporal correlation in the data and derive appropriate standard errors. We apply our methods to the Medicare Cohort Air Pollution Study (MCAPS), which includes individual-level information on time of death and age on a population of 18.2 million Medicare enrollees from 814 locations in the United States for the period 2000–2006. We use additional survey data available for a subset of the data to investigate sensitivity of the results to the inclusion of potential confounders in the model. Sizable differences between resulting estimated coefficients raise concerns about the presence of unmeasured confounding or other model failure. Persistence of differences, even after inclusion of measured time-varying confounders from the survey, indicates that

adjustment is not fully possible given the available information. While the estimate based on national trends is likely to be confounded by other variables trending on the national level, the estimate based on local trends is less likely to be confounded, although confounding cannot be ruled out.

We first introduce the data and our statistical model in Sections 2.1–2.3, proposing a decomposition of the spatio-temporal information to investigate confounding. The backfitting algorithm for fitting our proposed regression model is described in Section 2.4. In Sections 2.5 and 2.6, we explore the spatio-temporal correlation in the data and derive appropriate standard errors for estimates of regression coefficients and associated increases in life expectancy. Section 2.7, with a subset of the data where additional covariates on current smoking, body mass index, income, and race are available, details an analysis of the sensitivity of results to inclusion of these variables. In Section 3, we apply our methods to a population of 18.2 million Medicare enrollees from the MCAPS. Section 4 concludes with a discussion. Theoretical derivations are given in the Appendix. An online Appendix provides additional descriptions of the data and analysis results, the air pollution data, and all R code used to implement the methods and produce the results in this article.

## 2. METHODS

### 2.1 The Medicare Cohort Air Pollution Study Data

We construct a retrospective cohort study, by linking ambient levels of  $PM_{2.5}$  to mortality data by monitor during the period 2000–2006 (see also Zeger et al. 2008, for details).

Specifically, we obtain data from 1006  $PM_{2.5}$  monitors for the period 2000–2006 from the EPA monitoring network (<http://www.epa.gov/oar/data/>). In our analysis, we include data from 814 monitors in the continental United States. This subset was chosen on the basis of data availability. Each of the 814 monitors has measurements for at least four calendar years, with each of the four years including 10 or more months with at least four daily  $PM_{2.5}$  measurements. We divide the country into three geographical regions. These are the eastern region, the central region from the Mississippi River to the Sierra Nevada range, and the western United States (Zeger et al. 2008). Monitor locations and regional affiliation are depicted in Figure 1.

We define long-term exposure as the average of daily  $PM_{2.5}$  levels over the previous year. To take into account seasonality in the  $PM_{2.5}$  levels, and  $PM_{2.5}$  observations that are unevenly spread across the months, we calculate the yearly averages as follows. First, to fill small gaps in the data, we smooth the  $PM_{2.5}$  time series at each location using a linear regression with the daily  $PM_{2.5}$  values as the response, and with thin plate regression splines of time with four degrees of freedom per year (Janes, Dominici, and Zeger 2007) as the predictor. These correspond to an approximation to cubic smoothing splines, which is optimal in a certain sense (Wood 2003, 2006). For gaps longer than 90 days, we smooth the  $PM_{2.5}$  time series before and after the gap separately. Second, for each month, we calculate yearly averages of  $PM_{2.5}$  using the 365 predicted daily values from this model up to and including the respective month. In case of missing values, 350 days are deemed sufficient to compute the yearly average. The yearly averages thus obtained are

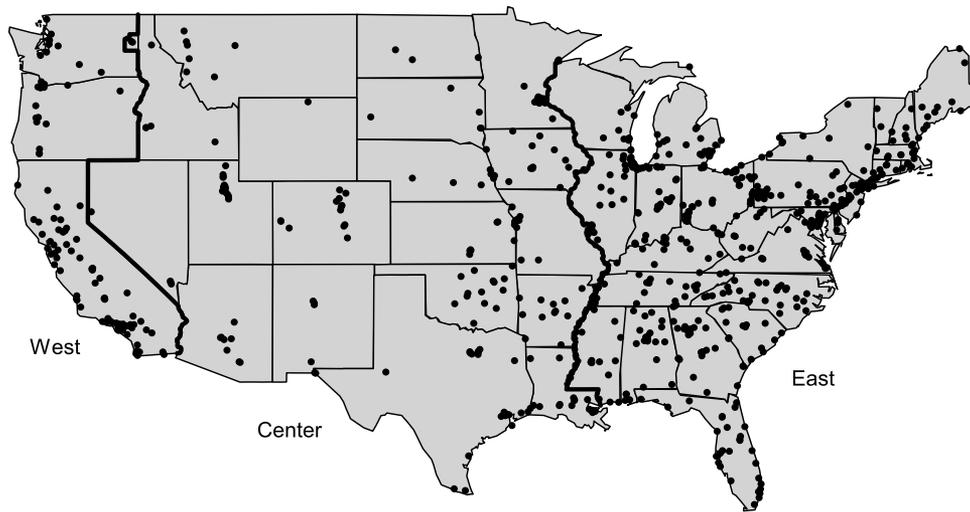


Figure 1. Locations of 814 EPA PM<sub>2.5</sub> monitoring sites in the continental United States used for the analysis. Boundaries of the three geographical regions are indicated by thicker lines.

not sensitive to the precise choice of smoothing procedure, and in fact the correlation with yearly averages from the raw data is 0.99. The 814 monitors provide up to 70 monthly measurements of yearly average PM<sub>2.5</sub> concentrations from December 2000 to September 2006. Summary statistics are given in Table 1.

We then link PM<sub>2.5</sub> data to the mortality data as follows: the same PM<sub>2.5</sub> exposure from a given monitoring site is assigned to all enrollees in the Medicare program, the U.S. health insurance program for those over 65, residing in a ZIP code (U.S. postal code) with a geographic centroid within a six mile radius from that site. As PM<sub>2.5</sub> is fairly spatially homogenous (Bell et al. 2007; Peng et al. 2008), the measurement error due to spatial variation in PM<sub>2.5</sub> should be of small order. The Medicare data provides demographic information (age, gender, race), and individual-level information on survival, with time of death or censoring precise up to the month. The dataset includes about 18.2 million enrollees and 3.2 million deaths in total, with an average of 10.4 million people enrolled in the cohort in any given month. Table 1 provides regional statistics on these enrollees.

The PM<sub>2.5</sub> data is publicly available, and we post the data in the online Appendix in the form used for the analysis. The Medicare mortality data used for this analysis is considered identifiable and the study was reviewed by the Centers for Medicare & Medicaid Services (CMS) Privacy Board. Requests for similar identifiable data files can be submitted to the CMS for review and approval. Nonidentifiable and limited

datasets are publicly available. More information can be found at <http://www.resdac.umn.edu/Medicare/>.

## 2.2 The Statistical Model

First, we specify the following proportional hazards model

$$h^c(a, t) = h^c(a) \exp(x_t^c \beta), \tag{1}$$

where  $h^c(a, t)$  denotes the hazard of dying at age  $a$  and time  $t$  for location  $c$ , and  $h^c(a)$  is a location-specific baseline hazard function. A location includes residents of ZIP codes with a geographic centroid within a six mile radius from the corresponding monitor.  $x_t^c$  is the average of the PM<sub>2.5</sub> levels at location  $c$  over the 12 months prior and including time  $t$ .

While the variables age  $a$  and time  $t$  are continuous variables in principle, the information in the Medicare data on time point of death or censoring is only precise up to the month. We thus discretize the time domain as follows. We measure  $t$  in monthly intervals, and denote the set of months with observations for location  $c$  by  $\mathcal{T}_c$ , where  $c = 1, \dots, C$ . Subject  $i$  contributes person-time, and, potentially, a death, to age interval  $a$  in a given month  $t$  if the subject turned 65 in month  $t - a$ . Monthly age is counted beginning at 65, since this is the cut-off for eligibility for Medicare. Assuming a constant hazard within each monthly age interval leads to a piecewise exponential survival model for life-tables (see Holford 1976) for each location.

With a study population of 18.2 million and 814 monitoring locations, naively fitting model (1) would require the handling

Table 1. Number of monitors, number of months with PM<sub>2.5</sub> data, average PM<sub>2.5</sub> level, number of Medicare enrollees and number of deaths among Medicare enrollees during the period December 2000 to September 2006. Values are medians among locations, with 25th and 75th percentile given in smaller print

Region	Monitoring stations	Months with available PM <sub>2.5</sub> data	Average PM <sub>2.5</sub> level [ $\mu\text{g}/\text{m}^3$ ]	Medicare enrollees	Deaths
West	96	597070	8.911.515.1	616413,28937,556	109930287270
Center	200	577070	9.510.612.0	743214,74629,073	81918083474
East	518	687070	12.613.915.1	702314,20727,688	140628745451
U.S.	814	627070	10.813.014.7	695714,50229,058	122025395120

of  $18.2 \text{ million} \times (814df + 2)$  matrices, where  $df$  are the degrees of freedom used in modeling  $h^c(a)$ . Computation proved to be infeasible due to memory restrictions, even on a cluster node with 64 GB of RAM. Instead, we use the log-linear regression model

$$\log E(Y_{at}^c) = \log(T_{at}^c) + \log(h^c(a)) + x_t^c \beta, \quad (2)$$

with the assumption that each  $Y_{at}^c$  is an independent (across calendar time, space, and age-months) Poisson variable, conditional on  $T_{at}^c$  and  $x_t^c$ . Here,  $Y_{at}^c$  is the number of deaths at age-month  $a$  in month  $t$  for location  $c$ , and  $T_{at}^c$  is the total time subjects of age  $a$  at location  $c$  were at risk of dying during month  $t$ . As the exact time of death or dropout during the month is not known, we approximate  $T_{at}^c$  by  $N_{at}^c$ , defined as the number of Medicare enrollees of age  $a$  with a ZIP code of residence in location  $c$  at the beginning of the month. Under the piecewise exponential survival model, models (1) and (2) are equivalent with regard to likelihood-based inference; please see the Appendix for a derivation using results by Holford (1980) and Laird and Olivier (1981). Independence assumptions that are made for this equivalence are independence between different locations and birth-month cohorts for model (1), and independence between locations  $c$ , months  $t$  and age-months  $a$  for model (2). We evaluate the justification of these independence assumptions in Section 2.5.

To make computation feasible and avoid excessive zero cell counts, we further assume a constant hazard of dying over one-year age intervals and after age 90. This allows us to collapse ages  $a$  into one-year intervals, and to combine all ages over 90 into one age group. Each of the resulting 1.4 million observations ( $Y_{at}^c, N_{at}^c, x_t^c$ ) then describes the mortality rate among people being  $a$  years of age at location  $c$  during month  $t$ , with average  $\text{PM}_{2.5}$  exposure  $x_t^c$  during the previous year. For each location  $c$ , we model the log-hazard function  $\log(h^c(a))$  in (2) using thin plate regression splines (Wood 2003, 2006) of age with three degrees of freedom, plus a location-specific intercept. To investigate sensitivity of results to this choice, we also repeat the analysis using five degrees of freedom. As the hazard of dying changes very slowly from one year of age to the next, results are practically identical, with changes in parameter estimates at most in the third significant figure. An additional indicator for ages over 90 allows for a potential discontinuity in the hazard function due to the mixture of hazard values in this last group. In model (2),  $\beta$  denotes the increase in the log-hazard of dying in a given month for an increase of  $1 \mu\text{g}/\text{m}^3$  in average  $\text{PM}_{2.5}$  concentrations during the previous year.

### 2.3 Using Spatio-Temporal Information to Investigate Confounding

Absent confounding and measurement error and given valid model assumptions, model (2) allows estimation of the effect of long-term exposure to  $\text{PM}_{2.5}$  on life expectancy. However, confounding is a common problem in air pollution studies.

To investigate the consistency of the evidence on  $\text{PM}_{2.5}$  and mortality, we propose to rewrite model (2) as follows (compare Janes, Dominici, and Zeger 2007).

$$\log E(Y_{at}^c) = \log(N_{at}^c) + \log(h^c(a)) + (x_t^c - \bar{x}_t - \bar{x}^c + \bar{x})\beta_1 + (\bar{x}_t - \bar{x})\beta_2, \quad (3)$$

where  $\bar{x}_t$  denotes the population-weighted average of the yearly  $\text{PM}_{2.5}$  averages in month  $t$  across locations,  $\bar{x}^c$  denotes the population-weighted average of the yearly  $\text{PM}_{2.5}$  averages for one location  $c$ , and  $\bar{x}$  denotes the overall population-weighted average. The Appendix gives details on these quantities, and shows approximate orthogonality of  $(x_t^c - \bar{x}_t - \bar{x}^c + \bar{x})$  and  $(\bar{x}_t - \bar{x})$  for the purpose of model (3). We thus decompose  $x_t^c$  into two orthogonal pieces of information. A third piece,  $(\bar{x}^c - \bar{x})$ , is absorbed by the location-specific log-hazard function  $\log(h^c(a))$ . Note that only changes over time in  $\text{PM}_{2.5}$  concentrations contribute to the estimation of  $\beta_1$  and  $\beta_2$ , avoiding cross-sectional confounding by individual-level risk factors or location-level characteristics.

In model (3),  $\beta_1$  and  $\beta_2$  both measure the strength of the association between  $\text{PM}_{2.5}$  and mortality, but use different sources of information. More specifically, the parameter  $\beta_2$  provides evidence as to whether nationally,  $\text{PM}_{2.5}$  and mortality rates are decreasing over time in parallel across the study period. The parameter  $\beta_2$  can be interpreted as the increase in the national log-mortality rate in a given month and age group, for an increase by  $1 \mu\text{g}/\text{m}^3$  in the national average  $\text{PM}_{2.5}$  concentration during the previous year. By contrast, the parameter  $\beta_1$  measures the strength of the evidence that mortality rates decline faster (slower) than the national average in locations where  $\text{PM}_{2.5}$  levels also decline faster (slower) than the national average.  $\beta_1$  can be interpreted as the additional increase in a local log-mortality rate for a  $1 \mu\text{g}/\text{m}^3$  increase in local  $\text{PM}_{2.5}$  concentrations over the national average level. The parameter  $\beta_1$  measures the association between local  $\text{PM}_{2.5}$  trends and local mortality trends, adjusting for the association between the national trends in  $\text{PM}_{2.5}$  and mortality rates.

An instructive parallel can be drawn to another approach that has been used to investigate long-term effects of air pollution on mortality using information over time. Pope, Ezzati, and Dockery (2009) plot differences in life expectancy  $\Delta\text{LE}_i$  between 1997–2001 and 1978–1983 for several U.S. counties against corresponding changes in  $\text{PM}_{2.5}$  concentrations  $\Delta\text{PM}_i$  in those counties. They obtain an effect estimate from the slope of a fitted linear regression line, regressing  $\Delta\text{LE}_i$  on  $\Delta\text{PM}_i$ . Note that such a plot yields two possible estimates of the  $\text{PM}_{2.5}$  effect: first, the slope of the estimated regression line, and second, the average increase in life expectancy over the mean reduction in  $\text{PM}_{2.5}$  across counties,  $\overline{\Delta\text{LE}}/\overline{\Delta\text{PM}}$ . The first coefficient has a similar interpretation to our  $\beta_1$ , the second coefficient corresponds to our  $\beta_2$ , even though our estimated coefficients compare more than two time points per location. We here investigate consistency of these two sources of information on the effects of long-term exposure to  $\text{PM}_{2.5}$  on mortality.

Estimation of the “global” parameter  $\beta_2$  and estimation of the “local” parameter  $\beta_1$  are differently affected by confounding. Consider the example of smoking, which has a large impact on mortality. Denote by  $z_t^c$  the proportion of smokers at location  $c$  at time  $t$ , and suppose that  $z_t^c$  has a large effect on mortality. Now, confounding can occur if smoking and  $\text{PM}_{2.5}$  are correlated. Even if smoking and  $\text{PM}_{2.5}$  concentrations are unrelated, confounding of  $\hat{\beta}_2$  can occur if  $(\bar{z}_t - \bar{z})$  and  $(\bar{x}_t - \bar{x})$  are not orthogonal. This could be due, for example, to downward trends in both  $\text{PM}_{2.5}$  and smoking rates on the national level over time, a scenario that is not unlikely to be true. As  $(\bar{z}_t - \bar{z})$

and  $(x_t^c - \bar{x}_t - \bar{x}^c + \bar{x})$  are approximately orthogonal by construction, confounding of  $\hat{\beta}_1$  can occur if  $(z_t^c - \bar{z}_t - \bar{z}^c + \bar{z})$  and  $(x_t^c - \bar{x}_t - \bar{x}^c + \bar{x})$  are not orthogonal. This would be true if communities which showed larger decreases in PM<sub>2.5</sub> than the national average also showed larger decreases in smoking rates than the national average, and vice versa. While this scenario is possible, it is less plausible than correlation in the national trends.

Absent confounding and given valid model assumptions,  $\beta_1$  and  $\beta_2$  are equal and can be collapsed into the single “overall” coefficient  $\beta$  in (2). Separate estimation of  $\beta_1$  and  $\beta_2$  allows us to diagnose unmeasured confounding, as large differences between  $\hat{\beta}_1$  and  $\hat{\beta}_2$  indicate that confounding is likely (see also Janes, Dominici, and Zeger 2007). For the inclusion of measured covariates, the approach can also aid in assessing whether adjustment for confounding is sufficient. This approach is related to specification tests in econometrics (Hausman 1978). An illustration of the interpretation of  $\beta_1$  and  $\beta_2$ , and of potential confounders, is given in the online Appendix.

### 2.4 Estimation Using the Backfitting Algorithm

Fitting model (3) directly is computationally very demanding. First, this is due to the high dimensionality of the dataset with 1.4 million observations  $(Y_{at}^c, N_{at}^c, x_t^c)$ , where  $a$  ranges through 26 (mostly yearly) age groups,  $t$  through the on average 65 months per location, and  $c$  through the 814 locations. Second, this is due to the complexity of the model, which specifies a log-hazard function  $\log(h^c(a))$  with 5 degrees of freedom for each of the 814 locations.

To reduce the dimensionality of the problem, we use a backfitting algorithm (Buja, Hastie, and Tibshirani 1989).

- Initialize  $\beta_1^{(0)} = \beta_2^{(0)} = 0$  and  $\log(h^c(a))^{(0)} \equiv 0, c = 1, \dots, C$ .
- *Step A:* For iteration  $j$ , set the offset to

$$\text{offset}_{at}^{c(j)} = \log(N_{at}^c) + \log(h^c(a))^{(j-1)}$$

for all  $a, t$ , and  $c$ . Fit the Poisson model

$$\log E(Y_{at}^c) = \text{offset}_{at}^{c(j)} + (x_t^c - \bar{x}_t - \bar{x}^c + \bar{x})\beta_1 + (\bar{x}_t - \bar{x})\beta_2$$

and set  $\beta_1^{(j)}$  and  $\beta_2^{(j)}$  to the estimated coefficients.

- *Step B:* For iteration  $j$ , set the offset to

$$\text{offset}_{at}^{c(j)} = \log(N_{at}^c) + (x_t^c - \bar{x}_t - \bar{x}^c + \bar{x})\beta_1^{(j)} + (\bar{x}_t - \bar{x})\beta_2^{(j)}$$

for all  $a, t$  and  $c$ . For  $c = 1, \dots, C$ , fit the Poisson model

$$\log E(Y_{at}^c) = \text{offset}_{at}^{c(j)} + \log(h^c(a))$$

to data from location  $c$ , and set  $\log(h^c(a))^{(j)}$  to the log-hazard function estimated from this model.

- While the change in  $\beta_1^{(j)}$  or  $\beta_2^{(j)}$  is larger than a certain stop criterion, repeat steps A and B. Conclude with step A.

The algorithm greatly reduces computational complexity by estimating the log-hazard function for each location separately. To investigate potential overdispersion, an overdispersion parameter  $\phi = \text{Var}(Y_{at}^c)/E(Y_{at}^c)$  can be included in the last Step A.

This backfitting algorithm is slightly different from the local scoring algorithm typically employed in estimation for generalized additive models (Hastie and Tibshirani 1990). There, one backfitting algorithm for additive models (inner loop) is carried out at each Newton–Raphson step (outer loop), and convergence results from the backfitting algorithm for additive models (Buja, Hastie, and Tibshirani 1989) carry over directly. Here, we carry out a full iteratively reweighted least squares algorithm (inner loop) for each step of the backfitting algorithm (outer loop). However, convergence of  $\beta_1^{(j)}$  and  $\beta_2^{(j)}$  to the unique maximum likelihood estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  is straightforward, and is shown in the Appendix.

We choose the stop criterion of the algorithm as a small relative change in the parameter estimates,  $\max\{|\beta_1^{(j)} - \beta_1^{(j-1)}|/\beta_1^{(j-1)}, |\beta_2^{(j)} - \beta_2^{(j-1)}|/\beta_2^{(j-1)}\} < 10^{-6}$ , which is reached within 4–8 iterations for the MCAPS data.

### 2.5 Variance Estimates and Spatio-Temporal Correlation

Variance estimators for  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$  that account for uncertainty in the estimation of the log-hazard functions  $\log(h^c(a))$  can be obtained from the likelihood using standard asymptotic theory; details are given in the Appendix. These model-based variance estimators are obtained under the assumption of independence across time, age groups, and geographical locations. In this section, we investigate the justification of this independence assumption.

We examine averaged empirical variograms over age, time, and space. More specifically, we define the empirical variogram over space, averaged over time and age, as follows. The averaged value for a bin of spatial distances  $(\delta_1, \delta_2]$  is given by

$$\frac{1}{N_{\delta_1, \delta_2}} \sum_{c=1}^C \sum_{u: \delta_{uc} \in (\delta_1, \delta_2]} \sum_{t \in \mathcal{T}_{cu}} \sum_{a=1}^A \frac{1}{2} (r_{at}^c - r_{at}^u)^2,$$

where  $r_{at}^c = (y_{at}^c - \hat{\mu}_{at}^c)/\sqrt{\hat{\mu}_{at}^c}$  is the Pearson residual and  $\hat{\mu}_{at}^c$  the fitted value from model (3) for location  $c$ , month  $t$ , and age  $a$ ,  $\mathcal{T}_{cu}$  is the set of months common to locations  $c$  and  $u$ ,  $\delta_{cu}$  is the spatial distance between locations  $c$  and  $u$ , and  $N_{\delta_1, \delta_2}$  is the number of terms in the sum for the interval  $(\delta_1, \delta_2]$ . As for the usual variogram over space (see, e.g., Diggle and Ribeiro 2007), the averaged variogram can be compared to the variance estimate  $\hat{\sigma}^2 = \sum_{a,t,c} r_{at}^c{}^2/N$ . Complete independence between spatial locations corresponds to variogram values close to  $\hat{\sigma}^2$  for all spatial distances. Averaged variograms over age and time are defined analogously.

### 2.6 Estimating Years of Life Gained

To translate the parameter estimates into values of relevance for public health, we estimate the years of life gained due to decreases in PM<sub>2.5</sub> exposure. For a known hazard function  $h(a)$ , we can calculate the life expectancy of a 65-year-old individual for a given exposure  $x$  and effect  $\beta$  as

$$\text{LE}(x, \beta) = \sum_a a [1 - \exp\{-h(a) \exp(x\beta)\}] \times \exp\left\{-\exp(x\beta) \sum_{b < a} h(b)\right\},$$

where  $a$  runs through the monthly ages starting with 65. We set the hazard function to a population-weighted average of the estimated hazard functions across all locations,  $h(a) \equiv \sum_{c=1}^C w_c \hat{h}^c(a)$ , where  $w_c$  weighs location  $c$  according to its average population over time,  $w_c = \bar{N}_c / \sum_c \bar{N}_c$  with  $\bar{N}_c = \sum_{a,t} N_{at}^c / |\mathcal{I}_c|$ .

To estimate the increase in life expectancy associated with a decrease in the annual average of  $PM_{2.5}$  by  $10 \mu\text{g}/\text{m}^3$ , we compute  $\Delta\text{LE}(\beta) = \text{LE}(x - 10, \beta) - \text{LE}(x, \beta)$ .  $\Delta\text{LE}(\beta)$  is the difference between the life expectancy assuming the personal exposure to be constant and equal to  $x$ , and the life expectancy assuming the exposure to be  $10 \mu\text{g}/\text{m}^3$  less than  $x$ . We choose  $x$  as the population-weighted average of the  $PM_{2.5}$  yearly average concentrations during the first year of the study period. Note that this approach estimates the increase in life expectancy after age 65, a lower bound for the overall increase in life expectancy.

We compute  $\Delta\text{LE}(\hat{\beta}_1)$  and  $\Delta\text{LE}(\hat{\beta}_2)$ , and their approximate standard errors using the Delta method. Details are given in the Appendix.

### 2.7 Sensitivity Analysis

To investigate the sensitivity of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  to the inclusion of potential confounders in the model, we use data from the Behavioral Risk Factor Surveillance System (BRFSS), available from the National Center for Chronic Disease Prevention and Health Promotion. The Selected Metropolitan/Micropolitan Area Risk Trends Survey (SMART) (<http://apps.nccd.cdc.gov/BRFSS-SMART/>) since 2002 provides data for selected counties in the United States on several risk factors for disease. The survey was designed specifically to look at trends in metropolitan and micropolitan areas, and provides information on several variables on the county level by month across several years, that is, on a similar level of spatio-temporal detail as for the  $PM_{2.5}$  and mortality variables. We use information available on current smoking, body mass index (BMI), income (in eight categories), and race (white/nonwhite) to construct monthly county averages, respectively proportions. More information on these variables  $z_t^c$  is given in the online Appendix. Appropriate weights are also available from the SMART website, and monthly county averages are based on on average 58 to 59 respondents, 51 for income. Covariate information is available only for a subset of locations and months. Still, the SMART survey seems the best available source of information on possible confounders on such a fine spatio-temporal level—illustrating also the difficulty to fully adjust for confounding in large observational studies. We use 173 locations with at least 80% of the maximum of 57 months of available information, corresponding to 17% of the original dataset. We repeat the analysis for this subset of data. We then additionally include the four variables into model (3), allowing (a) the same coefficient and (b) different coefficients for the components  $(\bar{z}_t - \bar{z})$  and  $(z_t^c - \bar{z}_t - \bar{z}^c + \bar{z})$ .

### 3. RESULTS FROM THE MCAPS STUDY

Yearly average  $PM_{2.5}$  concentrations have been decreasing during the study period in most of the study locations (Figure 2), with a pronounced drop in  $PM_{2.5}$  levels after September 2001. Population-weighted  $PM_{2.5}$  average levels in 2001 were highest in the west, intermediate in the east, and lowest in the central region. The west shows the strongest and most

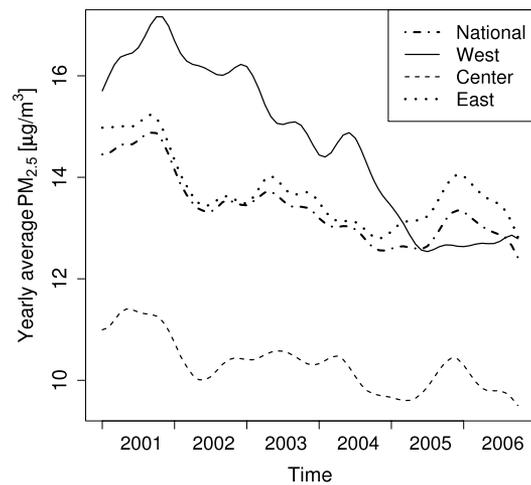


Figure 2. Average  $PM_{2.5}$  concentrations over the previous year for the months from December 2000 to September 2006. Depicted are both the population-weighted average across 531 monitors with complete time series in the continental United States (dotted-dashed line), as well as the population-weighted average by region: the west (solid line), the center (dashed line), and the east (dotted line).

consistent decline over time, which might reflect the stricter California ambient air quality standard of  $12 \mu\text{g}/\text{m}^3$  annual average  $PM_{2.5}$  that came into effect July 5, 2003 (California Environmental Protection Agency Air Resources Board, <http://www.arb.ca.gov/research/aaqs/aaqs.htm>). The decline in  $PM_{2.5}$  concentrations is less pronounced in the east and center, with higher average levels in 2005 after an initial decrease. The national average is dominated by values from the eastern region, which contributes 518 of the 814 monitors in this study.

Monthly age-standardized mortality rates have decreased over the same time period in all regions (Figure 3). Rates are comparable in the east and center, and lower in the west, with

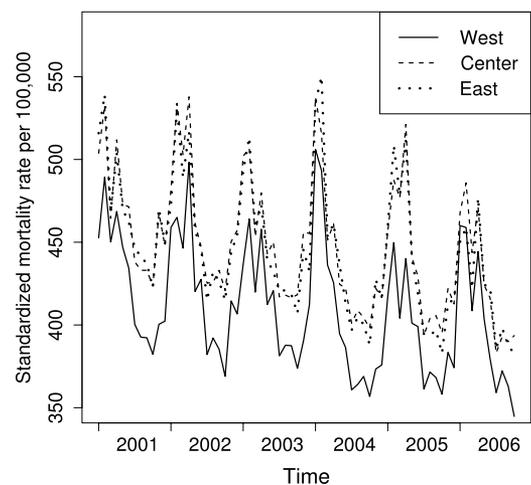


Figure 3. Monthly age-standardized mortality rates among Medicare enrollees from each of the three regions for December 2000 to September 2006. Mortality rates are standardized to the cohort study population in October 2003, the middle of the study period, i.e., the standardized rate  $r_t$  at time  $t$  is defined as  $r_t = \sum_a g_a Y_{at} / N_{at}$ , where  $Y_{at} = \sum_c Y_{at}^c$ ,  $N_{at} = \sum_c N_{at}^c$ , and  $g_a = N_{at_0} / \sum_a N_{at_0}$  indicates the age proportions in the population at  $t_0$ , October 2003.

Table 2. Estimated increase in the log-relative risk of dying in a given month per  $1\mu\text{g}/\text{m}^3$  increase in average  $\text{PM}_{2.5}$  concentrations during the previous year. The local coefficient  $\beta_1$  measures the association between local trends in  $\text{PM}_{2.5}$  and local trends in mortality rates, adjusting for the respective national trends. The global coefficient  $\beta_2$  measures the association between the  $\text{PM}_{2.5}$  national trend and the national trend in mortality rates. The overall coefficient  $\beta$  measures the association between local trends in  $\text{PM}_{2.5}$  and local trends in mortality, not adjusting for national trends

Region	Monitors	$100 \times \beta_1$ estimate(S.E.)	$100 \times \beta_2$ estimate(S.E.)	$100 \times \beta$ estimate(S.E.)
West	96	0.151(0.127)	2.021(0.100)	1.291(0.077)
Center	200	-0.110(0.193)	3.766(0.293)	1.080(0.159)
East	518	0.089(0.107)	3.795(0.108)	1.929(0.075)
U.S.	814	-0.061(0.064)	4.313(0.084)	1.562(0.051)

mortality rates peaking in the winter months. Maps of yearly average  $\text{PM}_{2.5}$  concentrations and age-standardized mortality rates by location are also given in the online Appendix.

In Table 2 and Figure 4, we report results from model (3) on the association between long-term exposure to  $\text{PM}_{2.5}$  and mortality. Table 2 gives estimated coefficients for the United States and each region, as well as their respective standard errors. We report estimated local and global coefficients  $\beta_1$  and  $\beta_2$  from model (3), and also the overall coefficient  $\beta$  from model (2). The corresponding relative risk estimates are depicted in Figure 4.

Estimated overdispersion parameter values for model (3) range from  $\hat{\phi} = 1.01$  to 1.02 across regions, and results shown are based on a Poisson model without overdispersion. Average empirical variograms (shown in the online Appendix) give no indication of correlation between observations over either space, time, or age. We therefore report model-based standard errors assuming independence across locations, months, and age-groups.

The estimate of the global coefficient  $\beta_2$  indicates that a  $10\mu\text{g}/\text{m}^3$  increase in the national average  $\text{PM}_{2.5}$  concentration over the previous year is associated with a significant 54%

increase in the risk of dying in a given month for our Medicare cohort. This estimate reflects that nationally, mortality rates are declining over the study period in parallel with  $\text{PM}_{2.5}$ . Estimates of the risk increase associated with a  $10\mu\text{g}/\text{m}^3$  increase in regional yearly average  $\text{PM}_{2.5}$  levels range from 22% to 46% across the three regions. The smallest value is estimated in the west, where the decrease in mortality rates is the smallest, but the decline in average  $\text{PM}_{2.5}$  concentrations is the largest (Figures 2 and 3). Note that the national  $\beta_2$  estimate is not a weighted average of the regional estimates and thus does not have to lie within their range.

Estimates of the local coefficient  $\beta_1$  are approximately zero and nonsignificant nationally and in all of the three regions. Estimates of  $\beta_1$  indicate that after adjusting for the association between national trends in mortality and  $\text{PM}_{2.5}$ , there is no significant association between an increase in the local yearly average  $\text{PM}_{2.5}$  concentration and the risk of dying in a given month for the local Medicare population.

Estimates of  $\beta$  lie between those of  $\beta_1$  and  $\beta_2$ , as they are a weighted average of these two estimates (Janes, Dominici, and Zeger 2007). However, the large differences between the estimated local and global coefficients  $\beta_1$  and  $\beta_2$  indicate that they cannot be combined into a single coefficient  $\beta$ . Indeed, a test for homogeneity  $H_0: \beta_1 = \beta_2$  results in  $p$ -values smaller than 0.0001 in all regions. Similar to a specification test, this result raises concerns about the model specification, with confounding the likely source for the discrepancy. Confounding of  $\beta_2$  by other variables trending on the national level is likely, while confounding of  $\beta_1$  by unmeasured factors, although less likely, cannot be ruled out.

The sensitivity of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  to unmeasured confounding is also investigated in the sensitivity analysis. We consider the following four county and month level variables: proportion of current smokers and of nonwhites, and mean income and body mass index. The national trend in  $\text{PM}_{2.5}$  shows population-weighted correlations of 0.08 to 0.22 in absolute value with the corresponding trends in these four variables; the respective correlations for the local deviations range from 0.00 to 0.04.

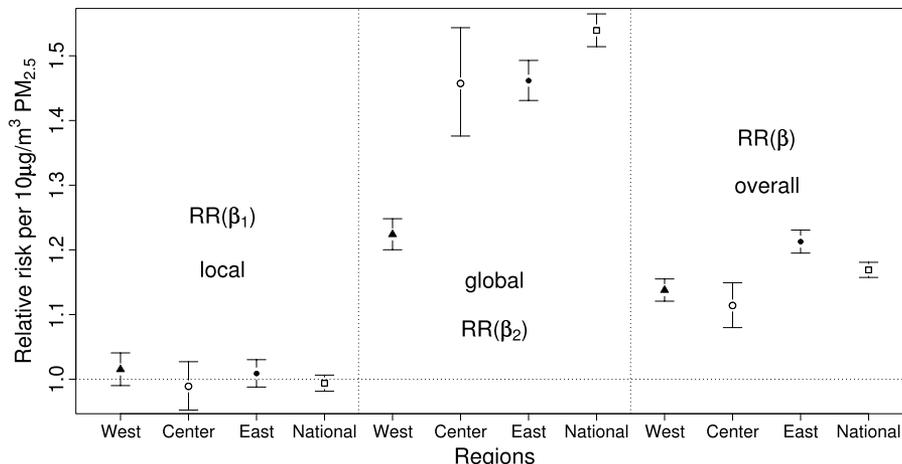


Figure 4. Estimated relative risk of dying in a given month per  $10\mu\text{g}/\text{m}^3$  increase in average  $\text{PM}_{2.5}$  concentrations during the previous year. Relative risk (RR) estimates based on the local coefficient  $\beta_1$ , the global coefficient  $\beta_2$ , and the overall coefficient  $\beta$  are shown with 95% confidence intervals.

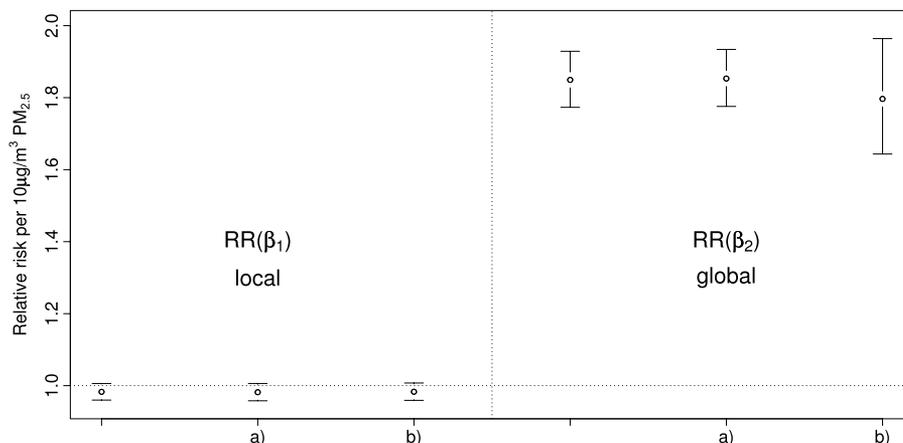


Figure 5. Sensitivity analysis using data on 173 locations with additional variables from the BRFSS-SMART survey. The left-most estimate shows estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  from model (3) for this subset of the data. a) indicates the analysis including additional variables on the level of the monitor’s county: the proportion of current smokers and of nonwhites, and the mean income and body mass index. b) gives the results for the same analysis allowing separate coefficients for the four variables’ global and local trends.

Figure 5 shows the results from the estimation of model (3) repeated on the subset of the data for which additional covariate information from the BRFSS-SMART survey is available. Estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are shown on the left of the respective panel. Both a) and b) indicate results including the additional covariates in the model, allowing for the same (a) or separate (b) coefficients for global and local trends, respectively. We find results not much altered by the inclusion of these variables; of the undecomposed variables (a) only race proved to be statistically significant. Some attenuation and widening of the confidence interval is observed for  $\hat{\beta}_2$  and b), as correlations between the variables’ national trends and the national trend in  $PM_{2.5}$  are largest. The full results of the sensitivity analysis are given in the online Appendix. Our decomposition and the remaining strong differences between  $\hat{\beta}_1$  and  $\hat{\beta}_2$  indicate that inclusion of these variables does not sufficiently adjust for confounding, illustrating the difficulties to fully adjust for confounding in large observational studies.

Table 3 translates parameter estimates into increases in life expectancy associated with a reduction in yearly average  $PM_{2.5}$  concentrations, giving estimates with 95% confidence intervals (CIs). Results based on the global coefficient indicate that a  $10 \mu g/m^3$  reduction in the yearly national average of  $PM_{2.5}$  is associated with an increase in life expectancy of 3.16 years (CI 3.05–3.26 years) in the Medicare population, although this estimate is likely confounded as discussed before. Results based

Table 3. Estimated increase in life expectancy  $\Delta LE$  in years for a  $10 \mu g/m^3$  reduction in average yearly  $PM_{2.5}$  exposure. Assumptions made in the calculation of  $\Delta LE$  are given in Section 2.6. Estimates are based on the local coefficient  $\beta_1$  or on the global coefficient  $\beta_2$ . Approximate 95% confidence intervals are derived from the standard errors for  $\hat{\beta}_1$  and  $\hat{\beta}_2$  using the Delta method

Region	Monitors	$\Delta LE(\hat{\beta}_1)$ (95%CI)	$\Delta LE(\hat{\beta}_2)$ (95%CI)
West	96	-0.08 0.13	1.43 1.57
Center	200	-0.40 -0.09	2.45 2.85
East	518	-0.10 0.07	2.68 2.81
U.S.	814	-0.16 -0.05	3.05 3.16

on local coefficients indicate no significant change in life expectancy for any reduction in  $PM_{2.5}$ .

#### 4. DISCUSSION

We have used spatio-temporal data from large national data bases to estimate the effect of long-term exposure to  $PM_{2.5}$  on life expectancy. We described an approach for decomposing the exposure variable into two components and for investigating differences between the corresponding estimated regression coefficients. Differences between the estimates of the two regression coefficients are indicative of unmeasured confounding or other model misspecification.

We have developed a statistical model and estimation procedure that allows the implementation of our approach for large spatio-temporal datasets. In the MCAPS study, the particulate air pollution application, sizable differences in effect estimates raise concerns about unmeasured confounding and the use of the aggregate data to draw conclusions on air pollution and mortality. While the coefficient based on national trends in  $PM_{2.5}$  and mortality is likely to be confounded by other variables trending on the national level, confounding of the local coefficient by unmeasured factors cannot be ruled out, although it is less likely. We used additional survey data available for a subset of the data to investigate sensitivity of results to the inclusion of additional covariates, but both coefficients remained largely unchanged. Measurement error for these additional covariates is likely to be large and available information is limited, illustrating the difficulty to comprehensively adjust for confounding in large observational studies. Our decomposition here can also help in assessing whether a given adjustment for confounding is sufficient. In our study, the remaining differences in effect estimates indicate that this cannot be achieved with the available data.

Few studies have investigated the association between particulate matter and mortality using temporal changes in long-term average  $PM_{2.5}$ . In a follow-up of the Six City Study, Laden et al. (2006) found a 1.14 (CI 1.06–1.22) mortality risk ratio for an increase by  $10 \mu g/m^3$  in the annual mean  $PM_{2.5}$  level in the year of death. However, the coefficient relies both on differences

in exposure between cities and within a city over time, and thus may be confounded by location-specific variables as well as variables trending on the national level. Pope, Ezzati, and Dockery (2009) regressed changes in life expectancy between 1978–1982 and 1997–2001 on changes in PM<sub>2.5</sub> concentrations between 1979–1983 and 1999–2000 in 51 metropolitan areas, thus also adjusting for location-specific confounders and trends on the national level, and obtaining estimates comparable in principle to our estimates  $\Delta LE(\hat{\beta}_1)$ . They found a 0.61 year (CI 0.22–1.0 year) increase in mean life expectancy associated with each 10  $\mu\text{g}/\text{m}^3$  PM<sub>2.5</sub> decrease, adjusting for changes in socioeconomic, demographic and smoking variables. Differences to our study that might explain the difference in results include (a) the earlier study period with the corresponding higher PM<sub>2.5</sub> levels, which could potentially have larger effects on mortality, (b) the longer study period resulting in larger differences over time and thus larger power to detect effects, (c) differences in the statistical model and estimation approach, (d) the population-based approach compared to the elderly population in our study, which yields only lower bounds on the overall increase in life expectancy, and (e) differences in the geographic locations included, with possible differences, for example, in the PM<sub>2.5</sub> composition. Future studies will investigate the reasons for the observed differences. Janes, Dominici, and Zeger (2007) in a previous analysis of a part of the Medicare cohort did not find evidence of an association between local trends in mortality and local trends in yearly average PM<sub>2.5</sub> after adjusting for the association between national trends and for location-specific differences, matching our own findings. A table with a full comparison of these four studies can be found in the online Appendix. Related work on PM<sub>10</sub> for the US (Zanobetti, Bind, and Schwartz 2008) and England (Janke, Propper, and Henderson 2009) has found significant associations with mortality, controlling for cross-sectional differences between locations and step-function or linear time trends. For earlier results on total suspended particulates in the 70s and 80s, see Chay, Dobkin, and Greenstone (2003), Chay and Greenstone (2003).

In our study, we used ambient PM<sub>2.5</sub> measurements from single monitors to measure PM<sub>2.5</sub> exposure. From a public policy perspective, decreases in ambient pollutant concentrations and associated decreases in mortality are of interest in assessing the impact of air quality regulations. Moreover, studies have shown that PM<sub>2.5</sub> is relatively homogeneous within a given county (Dominici et al. 2006; Janes, Dominici, and Zeger 2007), and ambient PM<sub>2.5</sub> is a strong proxy of personal PM<sub>2.5</sub> exposure (Sarnat et al. 2006).

To fully use the spatio-temporal variation in the data, we used the PM<sub>2.5</sub> average concentration over the last year as the relevant long-term exposure measurement. This approach could potentially miss longer-term effects or lag periods. However, the effects of long-term average PM<sub>2.5</sub> and PM<sub>2.5</sub> levels in the year of death have been found to be similar (Laden et al. 2006), which suggests reversibility of effects within about a year. This is plausible in light of the reversibility of the much larger increase in cardiovascular risk in smokers within about three years (Dobson et al. 1991; McElduff et al. 1998).

While we did not see any spatio-temporal correlation in the residuals in our analysis, it would be of interest in general to develop robust standard errors for the regression coefficients

that do not require an independence assumption across time and space. Other relevant extensions of our model are age-varying coefficients, to investigate potential differences in effects across age groups (see, e.g., Zeger et al. 2008), as well as time-varying coefficients. Of particular interest would be to include spatially varying coefficients for the air pollution effects, which would allow a more precise regional differentiation than the prespecified one considered here, and which could potentially discover differences in local effects due, for example, to differences in PM<sub>2.5</sub> composition. More work is needed to allow the fitting of these more complex models to large datasets such as the MCAPS data.

### APPENDIX: DERIVATIONS

#### Equivalence of Proportional Hazards Model and Poisson Model (Section 2.2)

The equivalence of the two models has been noted by Holford (1980) and Laird and Olivier (1981). First, consider the proportional hazards model (1) with month-wise constant hazard function for one location  $c$  and one birth-month cohort, which turns 65 in the same month  $t_0$ . For this cohort, the hazard of dying is constant in age-month interval  $a$ , and equal to  $h^c(a, t_0 + a) = h^c(a) \exp(x_{t_0+a}^c \beta)$ . The likelihood contribution from this cohort then is, analogous to Laird and Olivier (1981),

$$\prod_{a=1}^A [h^c(a, t_0 + a)^{Y_{a,t_0+a}^c} \exp(-h^c(a, t_0 + a)T_{a,t_0+a}^c)]^{\mathcal{I}(t_0+a \in \mathcal{T}_c)},$$

where  $Y_{at}^c$  is the number of deaths at age-month  $a$  in month  $t$  for location  $c$ ,  $T_{at}^c$  is the total time subjects of age  $a$  at location  $c$  were at risk of dying during month  $t$ ,  $\mathcal{T}_c$  is the location-specific set of observed months, and  $\mathcal{I}(\cdot)$  denotes the indicator function.

If we assume that observations which are from different locations or different birth-month cohorts are independent, the full likelihood can be written as

$$L_S(\beta; h^1(1), \dots, h^C(A)) = \prod_{c=1}^C \prod_{a=1}^A \prod_{t \in \mathcal{T}_c} (h^c(a) \exp(x_t^c \beta))^{Y_{at}^c} \exp(-h^c(a) \exp(x_t^c \beta) T_{at}^c).$$

For the log-linear Poisson model (2), under the assumption of independence between  $Y_{at}^c$  and  $Y_{\tilde{a}\tilde{t}}^{\tilde{c}}$  if  $(a, t, c) \neq (\tilde{a}, \tilde{t}, \tilde{c})$ , the likelihood is

$$L_P(\beta; h^1(1), \dots, h^C(A)) = \prod_{c=1}^C \prod_{a=1}^A \prod_{t \in \mathcal{T}_c} \frac{(h^c(a) T_{at}^c \exp(x_t^c \beta))^{Y_{at}^c} \exp(-T_{at}^c h^c(a) \exp(x_t^c \beta))}{Y_{at}^c!} \propto L_S(\beta; h^1(1), \dots, h^C(A)).$$

As the two likelihoods are proportional, the two models are equivalent with regard to likelihood-based inference.

#### Approximate Orthogonality of the Decomposition of PM<sub>2.5</sub> (Section 2.3)

Let  $N_t^c = \sum_{a=1}^A N_{at}^c$ , with  $N_t^c = 0$  if no observations are available at a given month  $t$  and location  $c$ . Define the population-weighted averages  $\bar{x}^c = (\sum_{t \in \mathcal{T}_c} N_t^c x_t^c) / (\sum_{t \in \mathcal{T}_c} N_t^c)$ , and  $(\bar{x}_t - \bar{x}) = \sum_{c=1}^C N_t^c (x_t^c - \bar{x}^c) / (\sum_{c=1}^C N_t^c)$ . Under model (3), under the assumption that  $\beta_1$  and  $\beta_2$  are small,  $h^c(a) \approx h(a)$  and  $N_{at}^c \approx N_t^c w_{at}$  for some weight  $w_{at}$ , we

have  $\mu_{at}^c = E(Y_{at}^c) \approx N_{at}^c w_{at} h(a)$ . Then,

$$\begin{aligned} & \sum_{c=1}^C \sum_{a=1}^A \sum_{t \in \mathcal{T}_c} \mu_{at}^c (\bar{x}_t - \bar{x})(x_t^c - \bar{x}^c - \bar{x}_t + \bar{x}) \\ & \approx \sum_{t=1}^T (\bar{x}_t - \bar{x}) \sum_{a=1}^A h(a) w_{at} \sum_{c=1}^C N_{at}^c (x_t^c - \bar{x}^c - \bar{x}_t + \bar{x}) = 0 \end{aligned}$$

by construction. The two variables thus are approximately orthogonal for the purpose of model (3) [compare the section below, Model-Based Variance Estimates (Section 2.5)].

**Convergence of the Backfitting Algorithm (Section 2.4)**

The log-likelihood  $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})$  is a function of  $\boldsymbol{\beta} = (\beta_1, \beta_2)$  and  $\boldsymbol{\gamma}$  of length  $5C$ , which contains an indicator, three spline basis functions in  $a$  and an indicator for ages over 90 for each location (see below, Model-Based Variance Estimates). The log-likelihood is based on an exponential family density and is strictly concave, as well as bounded above, with  $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) \rightarrow -\infty$  if one of the coordinates goes to  $\pm\infty$ . Thus, the maximum likelihood estimator of  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  exists and is unique, and there are no other local maximizers of the log-likelihood.

The backfitting algorithm alternates between maximizing  $\ell(\boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma})$  over  $\boldsymbol{\gamma}$  for fixed  $\boldsymbol{\beta}^{(j)}$ , and maximizing  $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(j)})$  over  $\boldsymbol{\beta}$  for fixed  $\boldsymbol{\gamma}^{(j)}$ . This corresponds to the Block Coordinate Descent/Ascent Method, which converges to  $\arg \max_{(\boldsymbol{\beta}, \boldsymbol{\gamma})} \ell(\boldsymbol{\beta}, \boldsymbol{\gamma})$ , as the log-likelihood is strictly concave and bounded above (Abatzoglou and O'Donnell 1982; Tseng 2001).

**Model-Based Variance Estimates (Section 2.5)**

The log-likelihood for model (3) can be defined as follows:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \sum_{c=1}^C \sum_{t \in \mathcal{T}_c} \sum_{a=1}^A \{Y_{at}^c (\mathbf{x}_t^{c'} \boldsymbol{\beta} + \mathbf{z}_a^{c'} \boldsymbol{\gamma}) - N_{at}^c \exp(\mathbf{x}_t^{c'} \boldsymbol{\beta} + \mathbf{z}_a^{c'} \boldsymbol{\gamma})\}.$$

Here,  $\mathbf{x}_t^c = \mathbf{x}_t^c$  of length 2 contains the  $PM_{2.5}$  variables for time  $t$  and location  $c$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2)$ , and  $\mathbf{z}_a^c \boldsymbol{\gamma}$  models the log-hazard functions  $\log(h^c(a))$ , where  $\mathbf{z}_a^c = \mathbf{z}_a^c$  of length  $5C$  contains an indicator, three spline basis functions in  $a$ , and an indicator for ages over 90 for each location  $\tilde{c}$ ,  $\tilde{c} = 1, \dots, C$ . This log-likelihood is based on the assumption of independence between all pairs  $Y_{at}^c$  and  $Y_{\tilde{a}\tilde{t}}^{\tilde{c}}$  for which  $(a, t, c) \neq (\tilde{a}, \tilde{t}, \tilde{c})$ .

The corresponding score equation is

$$\begin{aligned} S(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{c=1}^C \sum_{t \in \mathcal{T}_c} \sum_{a=1}^A (\mathbf{x}_t^{c'}, \mathbf{z}_a^{c'})' [Y_{at}^c - N_{at}^c \exp(\mathbf{x}_t^{c'} \boldsymbol{\beta} + \mathbf{z}_a^{c'} \boldsymbol{\gamma})] \\ &= (\mathbf{X} | \mathbf{Z})' (\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0}, \end{aligned}$$

where vectors  $\mathbf{Y}$  and  $\boldsymbol{\mu} = E(\mathbf{Y})$  of length  $N = A \sum_c \mathcal{T}_c$  contain entries  $Y_{at}^c$  and  $N_{at}^c \exp(\mathbf{x}_t^{c'} \boldsymbol{\beta} + \mathbf{z}_a^{c'} \boldsymbol{\gamma})$ , respectively, and  $(\mathbf{X} | \mathbf{Z})$  is the  $N \times (2 + 5C)$  matrix with rows  $(\mathbf{x}_t^{c'}, \mathbf{z}_a^{c'})$ ,  $a = 1, \dots, A$ ,  $t \in \mathcal{T}_c$ ,  $c = 1, \dots, C$ .

The model-based asymptotic covariance matrix for  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  then is (McCullagh and Nelder 1989)

$$\left[ -E \left( \frac{d}{d(\boldsymbol{\beta}, \boldsymbol{\gamma})} S(\boldsymbol{\beta}, \boldsymbol{\gamma}) \right) \right]^{-1} = ((\mathbf{X} | \mathbf{Z})' \text{diag}(\boldsymbol{\mu}) (\mathbf{X} | \mathbf{Z}))^{-1},$$

where  $\text{diag}(\boldsymbol{\mu})$  denotes the diagonal matrix with the entries in  $\boldsymbol{\mu}$  on the diagonal. Asymptotics here are for  $\sum_{t \in \mathcal{T}_c} N_{at}^c \rightarrow \infty$  for each  $a$  and  $c$ , while  $A$  and  $C$  are fixed, such that the number of parameters in  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  stays constant.

Note that the upper left corner of  $((\mathbf{X} | \mathbf{Z})' \text{diag}(\boldsymbol{\mu}) (\mathbf{X} | \mathbf{Z}))^{-1}$ , providing variance estimates for  $\hat{\boldsymbol{\beta}}$ , can be written as

$$\begin{aligned} & [\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{X}]^{-1} + [\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{X}]^{-1} [\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{Z}] \{ [\mathbf{Z}' \text{diag}(\boldsymbol{\mu}) \mathbf{Z}] \\ & - [\mathbf{Z}' \text{diag}(\boldsymbol{\mu}) \mathbf{X}] [\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{X}]^{-1} [\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{Z}] \}^{-1} \\ & \times [\mathbf{Z}' \text{diag}(\boldsymbol{\mu}) \mathbf{X}] [\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{X}]^{-1} \end{aligned}$$

using Schur complement and Woodbury formula. If  $\beta_1$  and  $\beta_2$  are small,  $h^c(a) \approx h(a)$  and  $N_{at}^c \approx N_{at}^c w_a$  for some weight  $w_a$ ,  $[\mathbf{X}' \text{diag}(\boldsymbol{\mu}) \mathbf{Z}]$  is close to zero analogous to the approximate orthogonality of the decomposition of  $PM_{2.5}$ . Variance estimates for  $\hat{\boldsymbol{\beta}}$  thus are little affected by estimation of even a large finite number of parameters  $\boldsymbol{\gamma}$ .

**Approximate Standard Errors for Estimated Years of Life Gained (Section 2.6)**

The quantity to be estimated is  $\Delta LE(\beta) = LE(x - 10, \beta) - LE(x, \beta) =: g(\beta)$ . An approximate standard error for  $g(\hat{\beta})$  using the Delta method is

$$\sigma(g(\hat{\beta})) \approx |g'(\hat{\beta})| \sigma(\hat{\beta}),$$

where  $\sigma(g(\hat{\beta}))$  and  $\sigma(\hat{\beta})$  are the standard errors of  $g(\hat{\beta})$  and  $\hat{\beta}$ , respectively,

$$\begin{aligned} g'(\beta) &= \frac{\partial}{\partial \beta} LE(x - 10, \beta) - \frac{\partial}{\partial \beta} LE(x, \beta) \quad \text{and} \\ \frac{\partial}{\partial \beta} LE(x, \beta) &= \sum_a ax \exp(x\beta) \exp \left\{ -\exp(x\beta) \sum_{b < a} h(b) \right\} \\ & \times \left[ \exp\{-h(a) \exp(x\beta)\} \sum_{c=1}^a h(c) - \sum_{c=1}^{a-1} h(c) \right]. \end{aligned}$$

These standard errors are for a given baseline hazard function  $h(a)$ , and do not account for uncertainty in estimating  $h(a)$ .

**SUPPLEMENTARY MATERIALS**

**Supplement:** *Web\_appendix.pdf* contains (1) illustrations of  $\beta_1$ ,  $\beta_2$  and potential confounders, (2) maps of yearly average  $PM_{2.5}$  concentrations and age-standardized mortality rates by location, (3) average empirical variograms as defined in Section 2.5, which illustrate the lack of correlation in the residuals over space, time or age for the MCAPS data, (4) sensitivity analyses using additional covariate information from the BRFSS-SMART survey, (5) a comparison of the current article with previous studies. *Rcode.R* provides an implementation of the methods in this article. *PM\_data.txt* contains the yearly average  $PM_{2.5}$  data used to produce the MCAPS results. Supplemental material is provided as a web appendix in a single zip file. (*Web\_appendix.zip*)

[Received June 2009. Revised January 2011.]

**REFERENCES**

Abatzoglou, T., and O'Donnell, B. (1982), "Minimization by Coordinate Descent," *Journal of Optimization Theory and Applications*, 36 (2), 163–174. [405]

Bachmann, J. (2008), "Air Pollution Forecasts and Results-Oriented Tracking," *Air Quality, Atmosphere & Health*, 1 (4), 203–207. [396]

Bell, M., Dominici, F., Ebisu, K., Zeger, S., and Samet, J. (2007), "Spatial and Temporal Variation in  $PM_{2.5}$  Chemical Composition in the United States for Health Effects Studies," *Environmental Health Perspectives*, 115 (7), 989–995. [398]

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17 (2), 453–510. [400]

Chay, K., and Greenstone, M. (2003), "The Impact of Air Pollution on Infant Mortality: Evidence From Geographic Variation in Pollution Shocks Induced by a Recession," *Quarterly Journal of Economics*, 118 (3), 1121–1167. [404]

Chay, K., Dobkin, C., and Greenstone, M. (2003), "The Clean Air Act of 1970 and Adult Mortality," *Journal of Risk and Uncertainty*, 27 (3), 279–300. [404]

Diggle, P., and Ribeiro, P. (2007), *Model-Based Geostatistics*, New York: Springer. [400]

- Dobson, A., Alexander, H., Heller, R., and Lloyd, D. (1991), "How Soon After Quitting Smoking Does Risk of Heart Attack Decline?" *Journal of Clinical Epidemiology*, 44 (11), 1247–1253. [404]
- Dockery, D. W., Pope, A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., and Speizer, F. E. (1993), "An Association Between Air Pollution and Mortality in Six U.S. Cities," *New England Journal of Medicine*, 329 (24), 1753–1759. [396]
- Dominici, F. (2002), "Invited Commentary: Air Pollution and Health—What Can We Learn From a Hierarchical Approach?" *American Journal of Epidemiology*, 155 (1), 11–15. [397]
- Dominici, F., McDermott, A., and Hastie, T. (2004), "Improved Semiparametric Time Series Models of Air Pollution and Mortality," *Journal of the American Statistical Association*, 99 (468), 938–948. [397]
- Dominici, F., Peng, R., Bell, M., Pham, L., McDermott, A., Zeger, S., and Samet, J. (2006), "Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases," *Journal of the American Medical Association*, 295 (10), 1127–1134. [404]
- Dominici, F., Samet, J., and Zeger, S. (2000), "Combining Evidence on Air Pollution and Daily Mortality from the 20 Largest US Cities: A Hierarchical Modelling Strategy," *Journal of the Royal Statistical Society, Ser. A*, 163 (3), 263–302. [397]
- Eftim, S., Samet, J., Janes, H., McDermott, A., and Dominici, F. (2008), "Fine Particulate Matter and Mortality: A Comparison of the Six Cities and American Cancer Society Cohorts With a Medicare Cohort," *Epidemiology*, 19 (2), 209–216. [396]
- Gamble, J. (1998), "PM<sub>2.5</sub> and Mortality in Long-Term Prospective Cohort Studies: Cause-Effect or Statistical Associations?" *Environmental Health Perspectives*, 106 (9), 535–549. [396]
- Greenbaum, D., Bachmann, J., Krewski, D., Samet, J., White, R., and Wyzga, R. (2001), "Particulate Air Pollution Standards and Morbidity and Mortality: Case Study," *American Journal of Epidemiology*, 154 (12), 78–90. [396]
- Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman & Hall. [400]
- Hausman, J. (1978), "Specification Tests in Econometrics," *Econometrica*, 46 (6), 1251–1271. [397,400]
- Health Effects Institute (2003), "Assessing Health Impact of Air Quality Regulations: Concepts and Methods for Accountability Research," Communication 11, available at <http://pubs.healtheffects.org/getfile.php?u=261>. [396]
- Holford, T. (1976), "Life Tables With Concomitant Information," *Biometrics*, 32 (3), 587–597. [398]
- (1980), "The Analysis of Rates and of Survivorship Using Log-Linear Models," *Biometrics*, 36 (2), 299–305. [399,404]
- Janes, H., Dominici, F., and Zeger, S. (2007), "Trends in Air Pollution and Mortality—An Approach to the Assessment of Unmeasured Confounding," *Epidemiology*, 18 (4), 416–423. [397,399,400,402,404]
- Janke, K., Propper, C., and Henderson, J. (2009), "Do Current Levels of Air Pollution Kill? The Impact of Air Pollution on Population Mortality in England," *Health Economics*, 18 (9), 1031–1055. [404]
- Kaiser, J. (1997), "Showdown Over Clean Air Science," *Science*, 277 (5325), 466–469. [396]
- Katsouyanni, K., Touloumi, G., Spix, C., Schwartz, J., Balducci, F., Medina, S., Rossi, G., Wojtyniak, B., Sunyer, J., Bacharova, L. et al. (1997), "Short Term Effects of Ambient Sulphur Dioxide and Particulate Matter on Mortality in 12 European Cities: Results From Time Series Data From the APHEA Project," *British Medical Journal*, 314 (7095), 1658–1662. [396]
- Kelsall, J., Samet, J., Zeger, S., and Xu, J. (1997), "Air Pollution and Mortality in Philadelphia, 1974–1988," *American Journal of Epidemiology*, 146 (9), 750–762. [396]
- Klemm, R., and Mason, R. (2003), "Replication of Reanalysis of Harvard Six-City Mortality Study," in *Revised Analyses of Time-Series of Air Pollution and Health. Special Report*, Boston, MA: Health Effects Institute, pp. 165–172. [397]
- Künzli, N., Medina, S., Kaiser, R., Quenel, P., Horak, F., and Studnicka, M. (2001), "Assessment of Deaths Attributable to Air Pollution: Should We Use Risk Estimates Based on Time Series or on Cohort Studies?" *American Journal of Epidemiology*, 153 (11), 1050–1055. [396,397]
- Laden, F., Schwartz, J., Speizer, F., and Dockery, D. (2006), "Reduction in Fine Particulate Air Pollution and Mortality: Extended Follow-Up of the Harvard Six Cities Study," *American Journal of Respiratory and Critical Care Medicine*, 173 (6), 667–672. [396,403,404]
- Laird, N., and Olivier, D. (1981), "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques," *Journal of the American Statistical Association*, 76 (374), 231–240. [399,404]
- Lumley, T., and Sheppard, L. (2000), "Assessing Seasonal Confounding and Model Selection Bias in Air Pollution Epidemiology Using Positive and Negative Control Analyses," *Environmetrics*, 11 (6), 705–717. [397]
- McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC. [405]
- McElduff, P., Dobson, A., Beaglehole, R., and Jackson, R. (1998), "Rapid Reduction in Coronary Risk for Those Who Quit Cigarette Smoking," *Australian and New Zealand Journal of Public Health*, 22 (7), 787–791. [404]
- Moolgavkar, S. (1994), "Air Pollution and Mortality," *New England Journal of Medicine*, 330 (17), 1237–1238. [396]
- (2005), "A Review and Critique of the EPAs Rationale for a Fine Particle Standard," *Regulatory Toxicology and Pharmacology*, 42 (1), 123–144. [397]
- Peng, R., Chang, H., Bell, M., McDermott, A., Zeger, S., Samet, J., and Dominici, F. (2008), "Coarse Particulate Matter Air Pollution and Hospital Admissions for Cardiovascular and Respiratory Diseases Among Medicare Patients," *Journal of the American Medical Association*, 299 (18), 2172. [398]
- Peng, R., Dominici, F., and Louis, T. (2006), "Model Choice in Time Series Studies of Air Pollution and Mortality," *Journal of the Royal Statistical Society, Ser. A*, 169 (2), 179–203. [397]
- Pope, C., Burnett, R., Thun, M., Calle, E., Krewski, D., Ito, K., and Thurston, G. (2002), "Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution," *Journal of the American Medical Association*, 287 (9), 1132–1141. [396]
- Pope, A., Ezzati, M., and Dockery, D. W. (2009), "Fine-Particulate Air Pollution and Life Expectancy in the United States," *New England Journal of Medicine*, 360 (4), 376–386. [399,404]
- Rabl, A. (2003), "Interpretation of Air Pollution Mortality: Number of Deaths or Years of Life Lost?" *Journal of the Air & Waste Management Association*, 53 (1), 41–50. [396]
- Samet, J., Dominici, F., Curriero, F., Coursac, I., and Zeger, S. (2000), "Fine Particulate Air Pollution and Mortality in 20 US Cities, 1987–1994," *New England Journal of Medicine*, 343 (24), 1742–1749. [397]
- Samet, J., Dominici, F., McDermott, A., and Zeger, S. (2003), "New Problems for an Old Design: Time Series Analyses of Air Pollution and Health," *Epidemiology*, 14 (1), 11–12. [396]
- Samoli, E., Peng, R., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., Burnett, R., Cohen, A., Krewski, D., Samet, J. et al. (2008), "Acute Effects of Ambient Particulate Matter on Mortality in Europe and North America: Results From the APHENA Study," *Environmental Health Perspectives*, 116 (11), 1480–1486. [397]
- Samoli, E., Schwartz, J., Wojtyniak, B., Touloumi, G., Spix, C., Balducci, F., Medina, S., Rossi, G., Sunyer, J., Bacharova, L. et al. (2001), "Investigating Regional Differences in Short-Term Effects of Air Pollution on Daily Mortality in the APHEA Project: A Sensitivity Analysis for Controlling Long-Term Trends and Seasonality," *Environmental Health Perspectives*, 109 (4), 349–353. [397]
- Sarnat, S., Coull, B., Schwartz, J., Gold, D., and Suh, H. (2006), "Factors Affecting the Association Between Ambient Concentrations and Personal Exposures to Particles and Gases," *Environmental Health Perspectives*, 114 (5), 649–654. [404]
- Schwartz, J., and Dockery, D. (1992), "Particulate Air Pollution and Daily Mortality in Steubenville, Ohio," *American Journal of Epidemiology*, 135 (1), 12–19. [396]
- Spix, C., Heinrich, J., Dockery, D., Schwartz, J., Völksch, G., Schwinkowski, K., Cöllen, C., and Wichmann, H. (1993), "Air Pollution and Daily Mortality in Erfurt, East Germany, 1980–1989," *Environmental Health Perspectives*, 101 (6), 518–526. [396]
- Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, 109 (3), 475–494. [405]
- Vedal, S. (1997), "Ambient Particles and Health: Lines That Divide," *Journal of the Air & Waste Management Association*, 47 (5), 551–581. [396,397]
- Wong, C.-M., Vichit-Vadakan, N., Kan, H., Qian, Z., and the PAPA Project Teams (2008), "Public Health and Air Pollution in Asia (PAPA): A Multi-city Study of Short-Term Effects of Air Pollution on Mortality," *Environmental Health Perspectives*, 116 (9), 1195–1202. [397]
- Wood, S. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society, Ser. B*, 65 (1), 95–114. [397,399]
- (2006), *Generalized Additive Models: An Introduction With R*, Boca Raton, FL: CRC Press. [397,399]
- Zanobetti, A., Bind, M., and Schwartz, J. (2008), "Particulate Air Pollution and Survival in a COPD Cohort," *Environmental Health*, 7 (1), 48. [404]
- Zeger, S., Dominici, F., McDermott, A., and Samet, J. (2008), "Mortality in the Medicare Population and Chronic Exposure to Fine Particulate Air Pollution in Urban Centers (2000–2005)," *Environmental Health Perspectives*, 116 (12), 1614–1619. [397,404]