MODELING THE HEALTH EFFECTS OF ENVIRONMENTAL EXPOSURES: PARTICULATE AIR POLLUTION AND MORTALITY IN URBAN AMERICA

Scott L. Zeger, Francesca Dominici, Aidan McDermott, Jonathan M. Samet

1 Introduction

Exposure to environmental pollutants likely contributes to morbidity and mortality around the world. An important question is how many persons in a given population suffer adverse health effects from environmental exposures. To address this question and others like it, United States government agencies, including the National Center for Health Statistics and Environmental Protection Agency, regularly collect monitoring data on health outcomes and pollutant levels. These data must be carefully analyzed to appropriately quantify health risks.

One example of a potentially important exposure in United States urban populations is particulate air pollution (Dockery et al., 1993; Pope et al., 1995a; American Thoracic Society, 1996; National Research Council, 1998; Pope and Dockery, 1999). The evidence for an association between particulate pollution concentrations and mortality and morbidity is strong (Lipfert and Wyzga, 1993; Pope et al., 1995a; Schwartz, 1995). Hence, public health and government officials seek to determine whether air pollution causes premature illness and death and, if so, to quantify the magnitude of the problem (Environmental Protection Agency, 1995; National Research Council, 1998).

In this chapter, we discuss statistical approaches for estimating the risk of death as a function of the concentration of particulate air pollution, a key component of risk analysis. We focus on multi-stage models of daily mortality data in the 88 largest cities in the U.S.A. to illustrate the main ideas. These models have been also used to quantify the risks of shorter-term exposure to particulate pollution and to address key causal questions (Samet et al., 2000a; Dominici et al., 2000a; Dominici et al., 2002a).

The air pollution mortality association has received substantial attention over the last decade since the publication of articles by Schwartz and Marcus (1990), Dockery et al. (1993), and Pope et al (1995b). That high levels of particulate pollution can cause illness and death has been generally accepted since the 1952 London smog caused thousands of excess deaths in a short time (Logan and Glasg, 1953; Bell and Davis, 2001). Governments in North America and Europe initiated large pollution-control programs in the 1970s as a result. However Schwartz and Marcus (1990) presented evidence from time series studies in selected U.S. cities to support the hypothesis that daily fluctuations in particulate levels within a population, well within current federal standards, continue to cause hospitalizations and premature deaths. Dockery, et al. (1993) and Pope et al (1995b) showed that persons living in cities with higher particulate levels suffer higher mortality risks than those living in less polluted cities, even after controlling for many important individual confounding variables such as smoking, socio-economic status and education.

The evidence for an association between particulate air pollution concentration and mortality derived from time series and from that derived from cohort studies is different. The time series studies (Schwartz and Marcus, 1990; Schwartz, 1995; Kelsall et al., 1997; Smith et al., 2000) rely on variations over time within a population, and compare the number of deaths on more-to-less polluted days. They depend on statistical control for within-population confounders including secular trends in mortality attributable to changing health practice and risk behaviors; effects of weather; and seasonal influenza epidemics. Even the multi-city studies (Katsouyanni et al., 1997; Samet et al., 2000c; Dominici et al., 2000a; Hwang and Chan, 2001; Dominici et al., 2002a) rely entirely on within-population comparisons; they increase power and possibly robustness by pooling relative risks across many cities. Time series models estimate the health risk associated with shorter-term exposures, on the order of days, weeks or at most months. Thus time series studies provide little evidence on the health effects of more chronic exposures. Kelsall et al. (1999), Zeger et al. (1999) , Schwartz (2000) and Dominici et al. (2002d) have discussed how to estimate pollution relative risks at different time scales using time series data.

The cohort studies (Dockery et al., 1993; Pope et al., 1995b; Krewski et al., 2000) compare mortality risks across distinct populations rather than over time for a population. They have shown that persons living in more-polluted cities are at higher risk for mortality than "otherwise similar" persons in less polluted cities. Here, "otherwise similar" refers to a population in which for age, gender, socioeconomic factors, smoking, and additional person-level confounding variables have been adjusted statistically. The cohort studies estimate the association of daily mortality with longer-term exposure. For example, Pope et al. (1995b) used average concentration for the past year as the predictor of mortality.

To date, the cohort studies have estimated larger relative risk than the time series studies. For example Pope et al. (1995b) in the American Cancer Society Study (ACS) report a 6.4 percentage increase in mortality per $10 \mu g/m^3$ increase in particulate matter less than 2.5 microns ($PM_{2.5}$) while Dominici et al. (2002a) in the NMMAPS multi-site time series study report a 0.41 percent per $10 \mu g/m^3$ increase in particulate matter less than 10 microns (PM_{10}). Recent re-analyses of the NMMAPS study have lowered this estimate to 0.21 percent per $10 \mu g/m^3$ increase in PM_{10} (Dominici et al., 2002d; Dominici et al., 2002b).

The difference between time series and cohort studies results likely reflects one or more factors, including the possibility that longer-term exposures have greater health consequences than do shorter-term exposures (Zeger et al., 1999; Schwartz, 2000; Dominici et al., 2002c), or that the cohort study estimates are confounded by unmeasured differences among the populations compared (Guttorp et al., 2001).

In this chapter, we present an overview of the time series approach to estimating the relative risk of mortality from particulate air pollution. We discuss how time series models might be used to partially address the possible difference in the mortality effects from shorter- to moderate-term variations of exposures.

In Section 2, we review the NMMAPS data base that comprises a time series of daily mortality, air pollution and weather variables between 1987 and 1994 for the 88 largest cities in the U.S. In section 3, we present several models for time series data. In Section 3.1, we describe the time series approach to estimating a pollution relative risk while controlling for likely confounders for a single city. In section 3.2, we summarize the models used to characterize the variability in relative risks among cities within and across regions and also to produce a pooled estimate across all 88 cities. We report the pooled estimate in section 3.3. In Section 4, we decompose the pollution time series into multiple predictors to estimate the relative risk of acute and more chronic exposures. We illustrate the difficulty of using time series models to estimate chronic health effects in 4 cities with daily particulate pollution data over 8 years. Finally, in Section 5 we offer an approach for future analyses that might combine cohort and time series information.

2 National Morbidity and Mortality Air Pollution Study Data

NMMAPS was a systematic investigation of the dependence of daily hospitalization and mortality counts on particulate and other air pollutants. Its database includes mortality, weather, and air pollution data for the 88 largest metropolitan areas in the U.S.A. for 1987 through 1994. Since its inception, the NMMAPS has contributed knowledge relevant to air quality policy, including:

- national, regional, and city estimates of the relative risk of mortality associated with concentrations of PM_{10} and other pollutants for the 88 urban centers (Samet et al., 2000a; Samet et al., 2000c; Samet et al., 2000b; Dominici et al., 2000a; Dominici et al., 2002a)
- a critical review of the impact of measurement error on time series estimates of relative risks and an estimate of the PM_{10} relative risk that is corrected for measurement error (Dominici et al., 2000b; Zeger et al., 2000);
- evidence that contradicts the "harvesting" hypothesis that the apparent association is attributable to the frail who die dying days earlier than they would have otherwise, absent air pollution (Zeger et al., 1999; Dominici et al., 2002c);
- evidence that contradicts the "threshold" hypothesis that particles and mortality except above a threshold concentration (Daniels et al., 2000; Schwartz, 2000; Dominici et al.,

2002a)

The daily time series of PM_{10}, O_3, mortality, temperature, and dew point for one NMMAPS urban center Pittsburgh Pennsylvania is illustrated in Figure 1. The data on mortality, air pollution, and weather, were drawn from public sources. The cause-specific mortality data at the county level were obtained from the National Center for Health Statistics. The focus was on daily death counts for each site, excluding accidental deaths and deaths of non-residents. Mortality information was available at the county level but not at a smaller geographic unit to protect confidentiality. All predictor variables were therefore also aggregated at the county level.

Hourly temperature and dew point data were obtained from the National Climatic Data Center, specifically the EarthInfo CD data base (`http://www.sni.net/earthinfo/cdroms/`). After extensive preliminary analyses that considered various daily summaries of temperature and dew point as predictors, we chose the 24-hour mean to characterize each day. If a county had more than one weather station, the average of the measurements from all stations was used.

The daily time series of PM_{10} and other pollutants for each city were obtained from the Aerometric Information Retrieval Service data base maintained by the US Environmental Protection Agency (`http://www.wpa.gov/airs/airs.html`). The pollutant data were also averaged over all monitors in a county. To protect against outlying observations and instrument drift, a 10% trimmed mean was used to average across monitors after correcting for each monitors yearly average. Information on several city-specific from the 1990 Census CD (`email:info@censuscd.com`) was also collected. A more detailed description of the data base is provided in NMMAPS reports (Samet et al., 2000c; Samet et al., 2000b).

3 Statistical Models for Time Series Data

3.1 Time Series Model for a Single Location

In this section, we specify the model for estimating air pollution-mortality relative risk separately for each location, accounting for age-specific longer-term trends, weather, and day of the week as potential confounding variables. The core analysis for each city is a log-linear generalized additive model (GAM) (Hastie and Tibshirani, 1990) with smoothing splines or a Poisson regression model with regression splines (GLM) (McCullagh and Nelder, 1989). Smoothing splines or regression splines, such as natural cubic splines, are included into the model to account for smooth fluctuations in mortality caused by non-pollution mechanisms that potentially confound estimates of the pollution effect, introduce autocorrelation in mortality series, or both.

We model daily expected deaths as a function of the pollution levels on the same or immediately preceding days, and not the average exposure for the preceding month, season, or year as might be done in a study of chronic effects. We use models that include smooth functions of time as predictors to control for possible confounding by seasonal influenza epidemics and by other longer-term changes in the population, demographic composition, personal behaviors (including smoking), and access to, and quality of, medical services.

To specify our approach, let y^c_{at} be the observed mortality for age group $a = (\leq 65, 65$ to $75, \geq 75$ years) on day t at location c, and $\boldsymbol{x}^c_{at}$ be a $p \times 1$ vector of air pollution variables. Let $\lambda^c_{at} = E(y^c_{at})$ be the expected number of deaths and $v^c_{at} = \text{var}(y^c_{at})$. We use a log-linear model

$$\begin{aligned} \log \lambda^c_{at} &= \boldsymbol{x}^{c\prime}_{at}\boldsymbol{\beta}^c + \text{confounders} \\ v^c_{at} &= \phi^c \lambda^c_{at},\ c = 1, \ldots, C \end{aligned} \qquad (1)$$

that allows the mortality counts to have variances v^c_{at} that may exceed their means (i.e., be overdispersed) with the overdispersion parameter ϕ^c also varying by location.

To protect the pollution relative risks $\boldsymbol{\beta}^c$ from confounding by longer-term trends, and to account for any additional temporal correlation in the count time-series, we estimate the pollution effect using only shorter-term variations in mortality and air pollution. To do so, we partial out the smooth, i.e. longer-term fluctuations in the mortality over time by including a smooth function (smoothing splines or natural cubic spline) of the calendar time $S^c(\text{time}, \nu)$ for each city. Here, ν is a smoothness parameter that we specify, from epidemiologic knowledge of the time scale of the major possible confounders to have 7 degrees of freedom per year of data so that little information from time-scales longer than approximately 2 months is included when $\boldsymbol{\beta}^c$ is estimated. We believe this choice substantially reduces confounding from seasonal influenza epidemics and from longer-term trends resulting from changing medical practice and health behaviors. We also control for age-specific longer-term and seasonal variations in mortality by adding a separate smooth function of time with 1 additional degree of freedom per year (8 total) for each age-group.

.To control for weather, we fit smooth functions of the same-day temperature (temp_0), average temperature for the 3 previous days (temp_{1-3}), each with 6 degrees of freedom, and the analogous functions for dew point ($\text{dew}_0, \text{dew}_{1-3}$), each with 3 degrees of freedom. In U.S. cities, mortality decreases smoothly with increasing temperature until reaching a relative minimum and then increases quite sharply at higher temperature (Curriero et al., 2002). We choose 6 degrees of freedom to capture this highly non-linear bend near the relative minimum as well as possible. Because there are missing values in air pollution concentration, and occasionally in other variables, we restricted analyses to days with no missing values across the full set of predictors.

In summary, we fit the following log-linear model to obtain the estimated pollution log-relative risk $\hat{\boldsymbol{\beta}}^c$ and the sample covariance matrix V^c at each location:

$$\begin{aligned} \log \lambda_{at}^c &= \boldsymbol{x}_{at}^{c\prime} \boldsymbol{\beta}^c + \gamma^c \text{DOW}_t + S_1^c(t, 7/\text{year}) + \\ &+ S_2^c(\text{temp}_{0t}, 6) + S_3^c(\text{temp}_{1-3t}, 6) + S_4^c(\text{dew}_{0t}, 3) + S_5^c(\text{dew}_{1-3t}, 3) \\ &+ \text{intercept for age group } a \\ &+ \text{separate smooth functions of time 1 df/year for age group } a \end{aligned} \tag{2}$$

where DOW$_t$ are indicator variables for day of week. Samet et al. (1999) and Kelsall et al. (1997) have given additional details about functions used to control for longer-term trends and weather. Alternative modeling approaches that consider different lag structures of the pollutants, and of the meteorological variables, have also been studied (Smith et al., 1997; Smith et al., 2000; Zanobetti et al., 2000) More general approaches that consider non-linear modeling of the pollutant variables have been discussed by Smith et al. (1997) and by Daniels at al. (2000) .

3.2 Model for Multi-Site Time Series Studies

In this section, we present a hierarchical regression model designed to estimate city-specific pollution relative risks, pooled national relative risk, and within- and between-region variances of the city-specific values. As we discussed in Section 2, even the pooled time series estimates derive entirely from within-population comparisons of shorter-term effects. We assume a three-stage hierarchical model with the following structure:

Stage I: City Given a time series of daily mortality counts at a given city (actually counties containing the city), the association between air pollution and health within the site is described using the regression model defined in Section 3.1. Among the output of the stage-I analysis are the point estimate $(\hat{\beta}_r^c)$ and the statistical variance (v_r^c) of the relative mortality risk (β_r^c) associated with each air pollutant within site c belonging to the geographical region r.

Stage II: Geographical region The information across cities within a region is combined by using a linear regression model in which the outcome variable is the log pollution relative risk for each city, and the explanatory variables (X_j^c) are site-specific characteristics, such as population density, yearly averages of the pollutants, and temperature. Formally:

$$\beta_r^c = \alpha_{0r} + \sum_{j=1}^{p} \alpha_j^c X_j^c + \text{error}_r.$$

If the predictors X_j^c are centered about their means, the intercept (α_{0r}) is interpreted as the pooled regional effect for a city with mean predictors. The regression coefficients (α_j^c) indicate the change in the relative risk of mortality associated with a unit change in the corresponding site-specific variable.

Stage III: Country The information across regions is combined by using a linear regression model in which the outcome variable is α_{0r} the true average regional relative mortality risk, and the explanatory variables (W_{jr}) are the region-specific characteristics (toxic composition of air pollution, climate variables). Formally:

$$\alpha_{0r} = \alpha_0 + \sum_{j=1}^{p} \alpha_j W_{jr} + \text{error}.$$

As in Stage II, if the predictors W_{jr} are centered about their means, the intercept α_0 is the pooled national effect for regions with mean values of the predictors. The regression coefficients (α_j) measure the change in true regional relative risk of mortality associated with a unit change in the corresponding region-specific variable.

The sources of variation in estimating of health effects of air pollution are specified by the levels of the hierarchical model. Under a three-stage model, the difference between the estimated site-specific relative risk ($\hat{\beta}^c_r$) and the true pooled relative risk (α_0) can be decomposed as:

$$(\hat{\beta}^c_r - \alpha_0) = (\hat{\beta}^c_r - \beta^c_r) + (\beta^c_r - \alpha_{0r}) + (\alpha_{0r} - \alpha_0).$$

The variation of $\hat{\beta}^c_r$ about β^c_r is measured by the within-site statistical variance (v^c_r) which depends on the number of days with air pollution data and on the predictive power of the site-specific regression model. The variation of β^c_r about α_{0r} is described by the between-site variance (τ^2) which measures the heterogeneity of the true air pollution effects across cities within a region. We assume τ^2 is constant across regions. Finally, the variance of α_{0r} about α_0 (σ^2) measures the heterogeneity of the true regional air pollution effects across regions.

A Bayesian hierarchical model is specified by selecting the prior distributions for the parameters at the top level of the hierarchy. If there is no desire to incorporate prior information into the analysis, then conjugate priors with large variances are often used. However, sensitivity of the substantive findings to the prior distributions should be investigated. Complex hierarchical models can be fit using simulation-based methods (Tierney, 1994; Gilks et al., 1996) that provide samples from the posterior distributions of all parameters of interest. Several software packages are now available (see for example `http://www.mrc-bsu.cam.ac.uk/bugs/`).

3.3 Results for National Mortality Morbidity Air Pollution Study

Dominici et al. (2002a) estimated city-specific air pollution effects by applying a GAM to each city, and approximated posterior distributions of regional, and national air pollution effects by applying a Bayesian three-stage hierarchical model to the NMMAPS data-base. Results are reported for 88 of the largest metropolitan areas in the U.S.A. from 1987 to 1994.

Recently, Dominici et al. (2002c) discovered that when GAM is applied to time-series analyses of contemporary data on air pollution and mortality, the defaults in the S-PLUS software (Version 3.4) package gam do not assure convergence of its iterative estimation procedure, and can provide biased estimates of the regression coefficients and standard errors. Thus the NMMAPS data base has been recently re-analyzed by using Poisson regression models with parametric non-linear adjustments for confounding factors (natural cubic splines) (Dominici et al., 2002b; Dominici et al., 2002d). The revised maximum likelihood estimates and 95% confidence intervals (CI) of the log-relative risks of mortality per 10 $\mu g/m^3$ increase in PM_{10} for each location are shown in Figure 2.

The estimate of the national pooled relative risk was a 0.22% increase in mortality per $10\mu g/m^3$ increase in PM_{10}, with a 95% posterior interval between 0.03 and 0.42. The pooled regional estimates of the PM_{10} effects vary across the regions; the estimated relative risk was greatest in the Northeast, with a value of 0.41% per $10\mu g/m^3$ (95% CI $0.12, 0.85$).

Figure 3 shows the maximum likelihood estimates (bottom) and the Bayesian estimates (top) of the relative risks of mortality for the largest 88 cities. Sizes of the circles are proportional to the statistical precisions. The Bayesian estimates of the city specific air pollution effects are shrunk toward the pooled estimate. The shrinkage factor is proportional to the statistical uncertainty of the maximum likelihood estimates, but inversely proportional to the degree of heterogeneity of the city-specific relative risks. In other words, imprecise maximum likelihood estimates (smaller dots) are shrunk toward the pooled estimate more heavily than precise maximum likelihood estimates (larger dots), and the shrinkage is more substantial when the between-city variance of the true relative risks is small.

4 Health Effects of Exposure over Different Time Periods

In previous section we discussed the use of time series data to quantify the association of mortality with short-term exposure to particulate pollution. As mentioned in the Introduction, short-term effects from time series studies appear to be smaller than long-term effects from chronic studies. In this section we discuss how within-city time series comparisons might be used to estimate the effects of exposures of different periods ranging from 1 day to 1 month. The effects of exposures on longer time scales are confounded in time series studies by other causes of seasonal and long-term fluctuations in mortality.

To start, we consider the log-linear distributed lags model for one city:

$$E(y_t) = \exp\left(\eta + \sum_{u=0}^{U} \theta_u x_{t-u}\right),$$

where x_{t-u} is the pollution concentration u days before day t (lag u). Under this model, a unit increase in exposure on day t produces an increase of θ_0 in the linear predictor; an increase of 1 unit for 3 consecutive days $t, t-1$, and $t-2$ causes an increase of $\theta_0+\theta_1+\theta_2$, and so forth. The lagged exposure series are highly co-linear, so it is desirable to constrain θ_u to a lower-dimensional sub-space. It is common to assume the θ_u form a polynomial or spline function of u (Judge, 1985; Davidson and MacKinnon, 1993; Zanobetti et al., 2000). More generally, we can let $\theta_u=\sum_{j=1}^{p} A_j(u)\beta_j$, so that the distributed lag models above become

$$\begin{aligned} Ey_t &= \exp(\eta+\textstyle\sum_{u=0}^{U}\theta_u x_{t-u}) \\ &= \exp\left\{\eta+\textstyle\sum_{j=0}^{p}\beta_j(\sum_{u=0}^{U}A_j(u)x_{t-u})\right\} \\ &= \exp\left\{\eta+\textstyle\sum_{j=1}^{p}\beta_j x_{jt}\right\} \end{aligned} \tag{3}$$

where $x_{jt}=\sum_{u=0}^{U}A_j(u)x_{t-u}$ is a linear combination of past exposures.

Kelsall et al. (1999), Zeger et al. (1999), Schwartz (2000), and Dominici et al. (2002d) defined the $A_j(u)$ to obtain a Fourier decomposition or moving average of the exposure time series, so that each $x_{jt},\ j=1,\ldots,p$ represents exposures at different time scales.

Here we define the x_{jt} as follows:

$$\begin{aligned} x_{1t} &= x_t & x_{4t} &= (x_{t-7}+\ldots+x_{t-29})/23 \\ x_{2t} &= (x_{t-1}+x_{t-2})/2 & x_{5t} &= (x_{t-30}+\ldots+x_{t-59})/30 \\ x_{3t} &= (x_{t-3}+\ldots+x_{t-6})/4 & & \end{aligned}$$

The β_1 is the log relative rate of mortality of current-day exposure; $\beta_1+\beta_2$ is the log relative rate of mortality of exposure over the last 3 days, and $\beta_1+\beta_2+\beta_3$ is the log relative rate of mortality for the past week exposure.

Figure 4 shows the decomposition of the PM_{10} and the temperature time series for Pittsburgh. The component series become smoother at the longer time scales, Also, the PM_{10} series at the longer time scales (30 to 60 days) and the temperature are highly correlated. This correlation reveals the potential for weather and or seasonality to confound the estimates of the chronic exposure effects.

Setting aside this concern temporarily, we estimated model 2 including all of the component PM_{10} series. Figure 5 displays the estimated relative risk of mortality associated with a 10 $\mu g/m^3$ increase in PM_{10}, for each of the 4 cities with daily PM_{10} data, for the time period indicated on the abscissa for each city (left top panel) and pooled across cities (right top panel). These estimates are just the cumulative sum of the regression coefficients for the components at time scales up to and including the indicated period.

Note that the Seattle results are distinct from those for Pittsburgh, Minneapolis, and Chicago. In Seattle, the relative risk depends on length of exposure: longer exposure has a greater estimated effect on mortality. Instead, in Chicago and Pittsburgh, longer exposure to increased levels of PM_{10} is associated with lower mortality.

These divergent patterns indicate that, at longer time scales, weather and/or seasonality might confound the air pollution mortality association. In Pittsburgh, Minneapolis, and Chicago, PM_{10} has a summer maximum, as does temperature. But mortality is highest in the winter because of weather and influenza epidemics. Hence, both temperature and the slowly varying components of PM_{10} are negatively correlated with mortality. Failure to completely control for weather, season, or both, would therefore bias the coefficients for longer-term components of PM_{10} to more negative values.

Because of use of wood stoves, in Seattle PM_{10} has a winter maximum instead. Hence, it has the opposite relationship to temperature. Here, failure to completely control for weather and season will tend to bias the coefficients at the longer time scales to more positive values. In summary, it is expected that confounding would produce the pattern seen here. In addition, the degree of separation in patterns between Seattle and the other cities changes with the numbers degrees of freedom in the spline function of time, further indicating that the pattern it is the result of confounding.

To simplify the problem of completely controlling for weather and seasonality, we repeated the analysis above restricting our attention to the non-winter days with temperatures in the temperate range 40-70 degrees Fahrenheit. Figure 5 shows the new estimates of the period-specific relative risks for the 4 cities and pooled across cities. The difference between Seattle and the other cities has been reduces but not entirely eliminated. This analysis demonstrates the difficulty of using time series data to estimate health effects of chronic exposures in the presence of other confounding variables at longer time scales.

5 Joint Analysis of Cohort and Time series Data

As we discussed in Section 4, the health effects of long-term exposure to air pollution cannot be estimated unless risks are compared across populations, in this case cities. This problem arises because long-term exposure varies a little within a city in long-term exposure and because of confounding by influenza and other trends represented by S^c(time) in model 2.

In this section, we briefly discuss a model for simultaneously estimating the effects of chronic and acute exposures on mortality. We envision a data set that consists of the mortality data, daily pollution concentrations, and personal risk factors for a population living in a large number of distinct geographic regions. Such data are available for the Medicare and veteran populations.

The basic model for λ_{it}^c, the risk for an individual i living in county c on day t, is

$$\lambda_{it}^c = \lambda_{0it}^c \exp\left(\sum_{j=0}^{p} x_{jt}^c \beta_j^c + z_{it}^c \alpha^c\right),$$

where λ_{0it}^c is the baseline risk without pollution exposure, x_{jt}^c is the j-th exposure variable in city

c on day t, as discussed in Section 4, and z_{it}^c is a vector of personal characteristics centered at their average value. To combine the cohort and time series analyses under a common umbrella, we define the first exposure $x_{0t}^c = \bar{x}^c$ to be the average exposure values for a prior extended period and center the remaining exposure variables about $\bar{x}^c$; that is. replace x_{jt}^c by $x_{jt}^c - \bar{x}^c$. To compare persons across cities, we must further assume that z_{it}^c includes all relevant persons and city-level confounders so that $\lambda_{0it} = \lambda_{0t}$. Then, we have

$$\lambda_{it}^c = \lambda_{0t} \exp\left(\bar{x}^c \beta_C + \sum_{j=1}^{p} x_{jt}^c \beta_{T_j}^c + z_{it}^c \alpha^c\right). \tag{4}$$

This model is the basis for simultaneous inference about both the "cohort" β_C and the "time series" $\beta_{T_j}^c$ relative risks. As discussed in Section 4, the x_{jt}^c that represent longer-term exposures are confounded by components in z_{it}^c at similar time scales (e.g., seasonality), necessitating comparisons across cities. We would again use $\beta_{T_j}^c$ in a two- or three-level hierarchical model to pool the time series relative risks for different exposure periods across cites and regions. By formulating the model to include both β_C and $\beta_{T_j}^c$, we can estimate and test the difference between β_C and the $\sum_{j=1}^{p} \beta_{T_j}^c$.

Estimating the parameters in Eq.4 is computationally intensive. At one extreme, this estimation can be treated as a very large parametric survival analysis with possibly millions of persons at risk and many time-varying predictor variables. But, the analyses can be simplified because z_{it}^c contains two kinds of predictor variables: those such as smoking or socioeconomic factors that vary across individuals but not across time (z_{1i}); and others such as temperature and season, that vary across time but not across individuals within a city (z_{2t}^c). We write $z_{it}^c \alpha^c = z_{1i}\alpha_1^c + z_{2t}^c \alpha_2^c$. Then, if we sum Eq.4 across individuals within a city, we have

$$\mathrm{E}(y_t^c) \;\simeq\; \exp\left(S^c(t) + \sum_{j=1}^{p} x_{jt}^c \beta_{T_j}^c + z_{2t}^c \alpha_2^c\right),$$

where $S^c(t) = \log \lambda_{0t} + \bar{x}^c \beta_C + \bar{z}_1 \alpha_1^c$. This is the usual time series model discussed in Section 3.

The likelihood function for $S^c(t)$, λ_{0t}, β_C, α_1^c, $\beta_{T_j}^c$, and α_2^c can be jointly maximized. Alternatively, a simpler algorithm might involve two steps: use standard semi-parametric log-linear regression to estimate $\beta_{T_j}^c$ and α_2^c, removing information about $S^c(t)$, and regressing $\hat{S}^c(t)$ on a smooth function of time common to all cities to estimate $\log \lambda_{0t}$, and on $\bar{x}^c$ and $\bar{z}_1$ to estimate β_C and α_1^c. Interesting methodological questions include how to formulate the joint estimation of β_C and $\beta_{T_j}^c$, in the hierarchical extension of Eq. 4; whether the simpler two-stage algorithm is as nearly efficient; and how to extend Eq. 4 to use cohort and time series information to estimate some of the $\beta_{T_j}^c$s for x_{jt}s that vary at longer time scales.

6 Discussion

In this chapter we illustrated the use of log-linear regression and hierarchical models to estimate the association of daily mortality with acute exposure to particulate air pollution. We used daily time series data for the 88 largest cities in the U.S.A. from 1987 to 1994 obtained from NMMAPS. Time series analyses such as those ones illustrated here rely entirely on comparisons of mortality across days within a population. They therefore have the advantage of avoiding confounding by unmeasured differences among populations, as may occur in cohort studies. They have the disadvantage of only providing evidence about the health effects of shorter-term acute exposures. The reason is that longer-term exposures vary a little within a population. Also processes such as influenza epidemics or changes in health behaviors are difficult to quantify, they occur at these longer time scales, and they will likely confound estimates of the health effects of chronic exposures. We chose to control for longer-term confounders by using models that partial out the variation in exposure and mortality at longer time scales. We achieved this control by including spline functions of time with 7 degrees of freedom per year and spline functions of temperature and dew point temperature to control for weather (smoothing splines or natural cubic splines). We checked whether using half or twice as many degrees of freedom quantitatively change the results and found that it does not. The possibility remains, however, that the effects of confounders are not totally controlled by these statistical adjustments.

We fit a separate log-linear regression for each of the 88 cities and conducted a second-stage analysis to estimate the variability in the pollution relative risk among cities within a region and among regions; to obtain empirical Bayes estimates of each city's relative risk, and to estimate the average relative risk across the 88 cities. As did others (Smith et al., 1997; Smith et al., 2000; Clyde, 2000), we found that the exact specification of the log-linear regression (e.g. the choice of the number of degrees of freedom in the smoothing splines or the number and location of knots in the regression splines), can substantially affect the estimated pollution relative risk for a particular city. However, we found that variations in the specification tend to have little effect on the pooled estimate or on the empirical Bayes estimates for that city. Hence, pooling information across a large number of cities provides a more robust model. The details of such investigations have been provided by Dominici et al. (2000a) .

We used a Bayesian hierarchical model to pool information across 88 cities. This model requires specifying prior distributions. Here, the key priors are for the within- and among-region variances of the true relative risks. We assumed half-normal priors with large variances instead of the more traditional conjugate inverse gamma prior. In our experience, the inverse gamma prior implicitly rules out the possibility of no variability in the relative risks and produces posterior distributions with more mass at larger values. The result is to borrow less strength across cities when estimating the risk for one city and wider confidence intervals for the estimated average coefficient across cities. The posterior distribution obtained, assuming a half-normal prior, is more similar to the likelihood function and hence we prefer it in this application. Sensitivity analyses of the posterior distribution of the pooled air pollution effect, under 4 specifications of the prior distributions for the within- and among-region variances of the true relative risks in alternative to the half-normal, have been discussed by Dominici et al.(2002a) .

In Section 4, we used the time series data from 4 cities with more complete particulate pollution records to estimate a distributed lags model. We then estimated the mortality relative risk associated with a 10 $\mu g/m^3$ increase in PM_{10} for 1, 3, 7, 30, and 60 days. We found evidence of possibly substantial confounding by seasonality and weather at the longer time scales. In fact, the longer-term variation in PM_{10} is highly correlated with temperature. To overcome this problem, we restricted the analyses to non-winter days with moderate temperature. In this subset of the data, we found that the mortality relative risk increases with the duration of exposure up to roughly 7 to 14 days but not beyond. The total effect sizes are still much smaller than those seen in the cohort studies. The evidence of confounding of the longer-term components of PM_{10} shows the limitations of time series models for estimating the health effects of chronic exposure to pollutants.

Finally, we briefly outlined an approach to jointly estimating the cohort and time series relative risks from a single, large database. Several such databases exist, including the Medicare Cohort from the National Claims History File. This database includes all mortality, morbidity, and basic demographic information on more than 2 million persons per year. In addition, the Medicare Current Beneficiary Survey provides detailed personal information on a representative sub-sample of roughly 13,000 persons per year. If these data were merged with weather and pollution time series for each person, chronic and acute effects could be estimated from a common data source. The differences in the current estimates of the particulate air pollution relative risks deriving from cohort and time series studies must be better understood. When these are obtained from different sources of data with different protocols, it is unclear whether the differences reflect greater effects of chronic exposure or biases from one or both of the studies. The proposed approach can address some of these issues using a refined version of the analyses discussed here.

References

American Thoracic Society, Bascom, R. (1996). Health effects of outdoor air pollution, Part 2. *American Journal of Respiratory and Critical Care Medicine* **153**, 477–498.

Bell, M. and Davis, D. (2001). Reassessment of the lethal london fog of 1952: Novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environmental Health Perspective* **109**, 389–394.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter environmetrics. *Environmetrics* **11**, 745–763.

Curriero, F., Heiner, K., Samet, J., Zeger, S., Strug, L., and Patz, J. (2002). Temperature and mortality in 11 cities of the eastern United States. *American Journal of Epidemiology* **155**, 80–87.

Daniels, M., Dominici, F., Samet, J. M., and Zeger, S. L. (2000). Estimating PM10-mortality dose-response curves and threshold levels: An analysis of daily time-series for the 20 largest US cities. *American Journal of Epidemiology* **152**, 397–412.

Davidson, R. and MacKinnon, J. (1993). *Estimation and Inference in Econometrics*. New York: Oxfors University.

Dockery, D., Pope, C. A., Xu, X., Spengler, J., Ware, J., Fay, M., Ferris, B., and Speizer, F. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* **329**, 1753–1759.

Dominici, F., Daniels, M., Zeger, S. L., and Samet, J. M. (2002a). Air pollution and mortality: Estimating regional and national dose-response relationships. *Journal of the American Statistical Association* **97**, 100–111.

Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. M. (2002b). 0n the use of Generalized Additive Models in Time Series Studies of Air Pollution and Health. *American Journal of Epidemiology* **156**, 1–11.

Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. M. (2002c). Airborne particulate matter and mortality: Time-scale effects in four US cities. *American Journal of Epidemiology (in press)* .

Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. M. (2002d). *A Report to the Health Effects Institute on Reanalyses of the NMMAPS Database*. The Health Effects Institute, Cambridge, MA.

Dominici, F., Samet, J. M., and Zeger, S. L. (2000a). Combining evidence on air pollution and daily mortality from the twenty largest US cities: A hierarchical modeling strategy (with discussion). *Royal Statistical Society, Series A* **163**, 263–302.

Dominici, F., Zeger, S. L., and Samet, J. M. (2000b). A measurement error correction model for time-series studies of air pollution and mortality. *Biostatistics* **2**, 157–175.

Environmental Protection Agency (1995). Proposed guidelines for neurotoxicity risk assessment. *Environmental Protection Agency* .

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Guttorp, P., Sheppard, L., and Smith, R. (2001). *Comments on the PM Criteria Document*. Technical report, University of Washington, Seattle,Washington.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Hwang, J. and Chan, C. (2001). Air pollution effects on daily clinic visits for lower respiratory illness. *American Journal of Epidemiology (in press)* .

Judge, G. G. (1985). *The Theory and Practice of Econometrics (2nd ed)*. New York: Wiley.

Katsouyanni, K., Toulomi, G., Spix, C., Balducci, F., Medina, S., Rossi, G., Wojtyniak, B., Sunyer, J., Bacharova, L., Schouten, J., Ponka, A., and Anderson, H. R. (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**, 1658–1663.

Kelsall, J., Samet, J. M., and Zeger, S. L. (1997). Air pollution, and mortality in Philadelphia, 1974-1988. *American Journal of Epidemiology* **146**, 750–762.

Kelsall, J., Zeger, S. L., and Samet, J. M. (1999). Frequency domain log-linear models; air pollution and mortality. *The Royal Statistical Society, Series C* **48**, 331–344.

Krewski, D., Burnett, R. T., Goldberg, M. S., Hoover, K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M., and White, W. H. (2000). *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality*. Cambridge, MA: Health Effects Institute.

Lipfert, F. and Wyzga, R. (1993). Air pollution and mortality: Issues and uncertainty. *Journal if the Air Waste Management Association* **45**, 949—966.

Logan, W. and Glasg, M. (1953). Mortality in london for incident, 1953. *Lancet* **1**, 336–338.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second Edition)*. Chapman & Hall.

National Research Council (1998). Research priorities for airborne particulate matter. part i. immediate priorities and a long-range research portfolio. *Washington, DC: National Academy Press.* .

Pope, A. C. and Dockery, D. W. (1999). Epidemiology of particle effects. In: *Air Pollution and Health*, 673–705. San Diego: Academic Press.

Pope, C. A., Dockery, D., and Schwartz, J. (1995a). Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology* **47**, 1—-18.

Pope, C. A., Thun, M., Namboodiri, M., Dockery, D., Evans, J., Speizer, F., and Heath, C. (1995b). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory Critical Care Medicine* **151**, 669–674.

Samet, J. M., Dominici, F., Curriero, F., Coursac, I., and Zeger, S. L. (2000a). Fine particulate air pollution and mortality in 20 U.S. cities: 1987-1994 (with discussion). *New England Journal of Medicine* **343**(24), 1742–1757.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D., Schwartz, J., and Zanobetti, A. (2000b). *The National Morbidity, Mortality, and Air Pollution Study (HEI Project No. 96-7): Morbidity and Mortality from Air Pollution in the United States*. Health Effects Institute, Cambridge, MA.

Samet, J. M., Zeger, S. L., Dominici, F., Dockery, D., and Schwartz, J. (2000c). *The National Morbidity, Mortality, and Air Pollution Study (HEI Project No. 96-7): Methods and Methodological Issues*. Cambridge, MA: Health Effects Institute.

Schwartz, J. (1995). Air pollution and daily mortality in Birmingham, alabama. *American Journal of Epidemiology* **137**, 1136–1147.

Schwartz, J. (2000). Harvesting and long term exposure effects in the relation between air pollution and mortality. *American Journal of Epidemiology* **151**, 440–448.

Schwartz, J. and Marcus, A. (1990). Mortality and air pollution in london: A time series analysis. *American Journal of Epidemiology* **131**, 185–194.

Smith, L., Davis, J., Sacks, J., Speckman, P., and Styer, P. (1997). Assessing the human health risk of atmospheric particle. *ASA Proceedings, Env Stat* .

Smith, R., Davis, J., Sacks, J., Speckman, P., and Styer, P. (2000). Regression models for air pollution and daily mortality: Analysis of data from Birmingham, Alabama. *Environmetrics* **100**, 719–745.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1762.

Zanobetti, A., Wand, M., Schwartz, J., and Ryan, L. (2000). Generalized additive distributed lag models. *Biostatistics* **1**, 279–292.

Zeger, S., Thomas, D., Dominici, F., Cohen, A., Schwartz, J., Dockery, D., and Samet, J. (2000). Measurement error in time-series studies of air pollution: Concepts and consequences. *Environmental Health Perspectives* **108**.

Zeger, S. L., Dominici, F., and Samet, J. M. (1999). Harvesting-resistant estimates of pollution effects on mortality. *Epidemiology* **89**, 171–175.

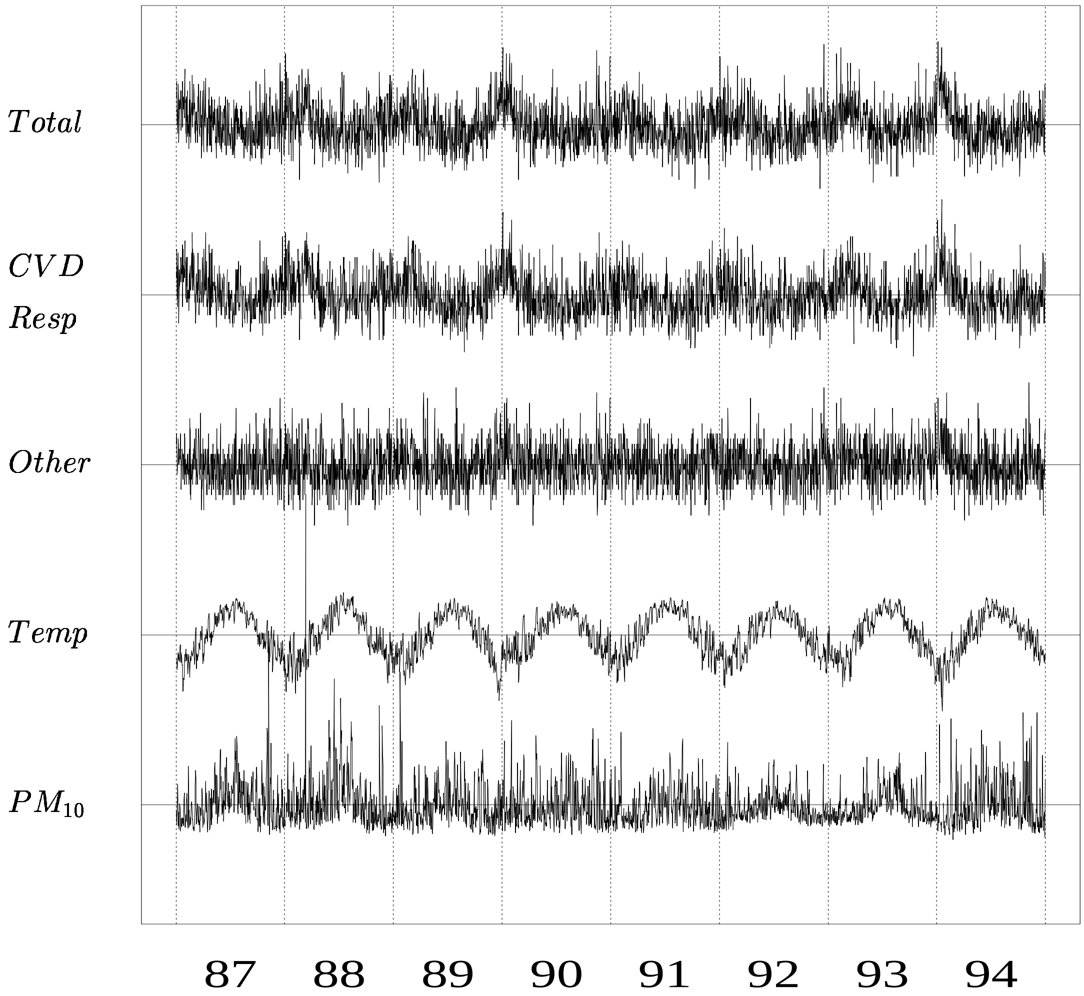

Figure 1: *Total, cardiovascular respiratory (CVDRESP), and other causes mortality daily counts. Temperature (Temp) and* PM_{10} $\mu g/m^3$ *daily time series. Data for Pittsburgh, Pennsylvania, for 1987 through 1994.*

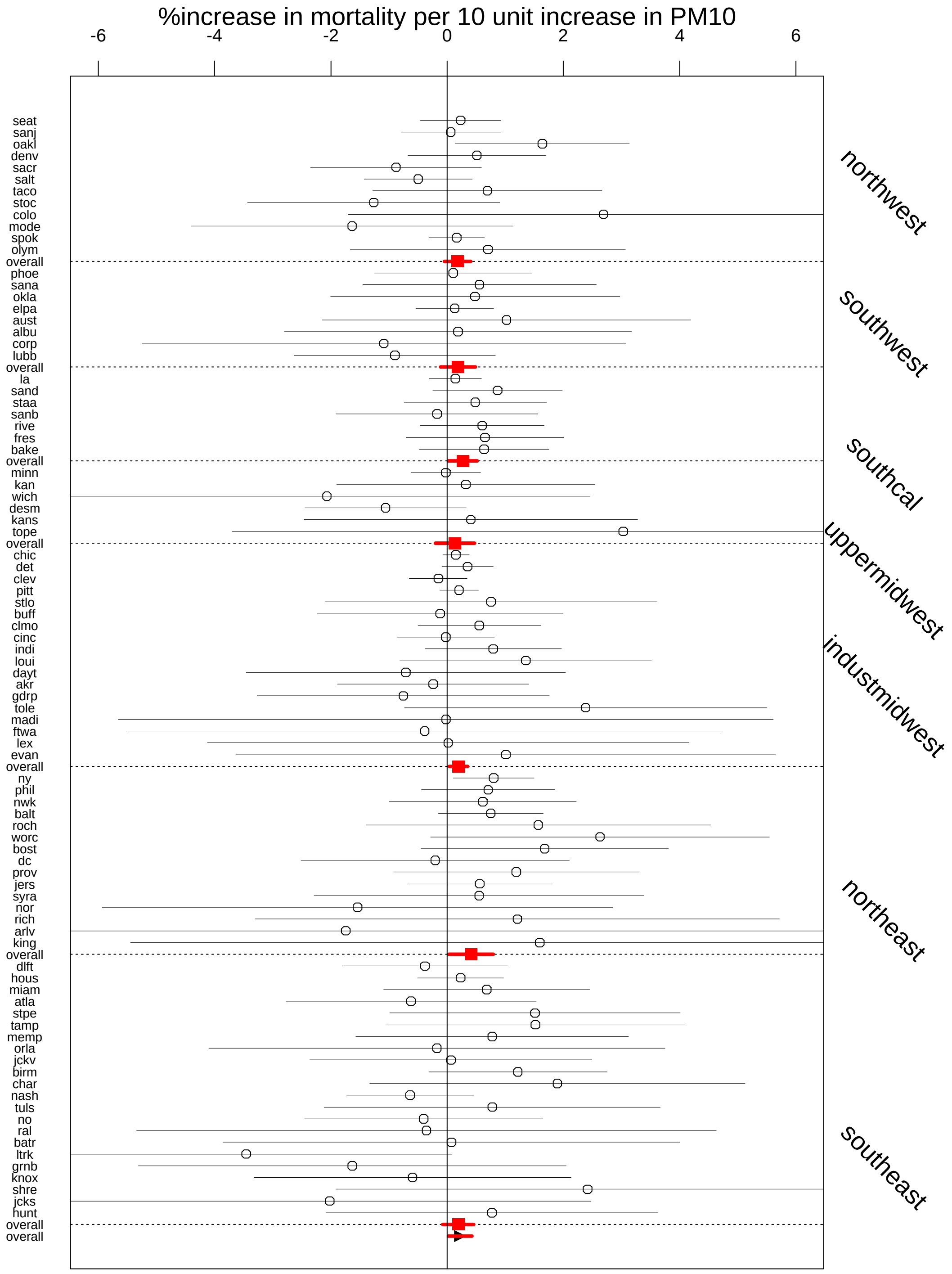

Figure 2: *Maximum likelihood estimates and 95% confidence intervals of the log-relative risks of mortality per* $10\mu g/m^3$ *increase in* PM_{10} *for each location. The solid circles with the bold segments denote the posterior means and 95% posterior intervals of the pooled regional effects. At the bottom, marked with a triangle and a bold segment, is the overall effects for* PM_{10} *for all the cities.*

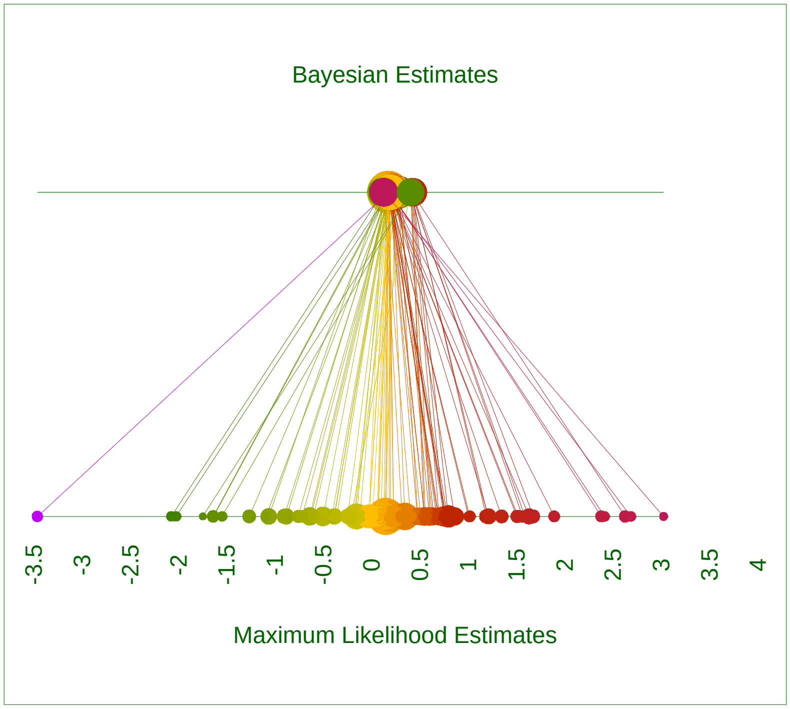

Figure 3: *Maximum Likelihood (bottom) and Bayesian posterior estimates (top) of the relative risks of mortality for the 88 U.S. largest cities. Size of the circle is proportional to the statistical precisions.*

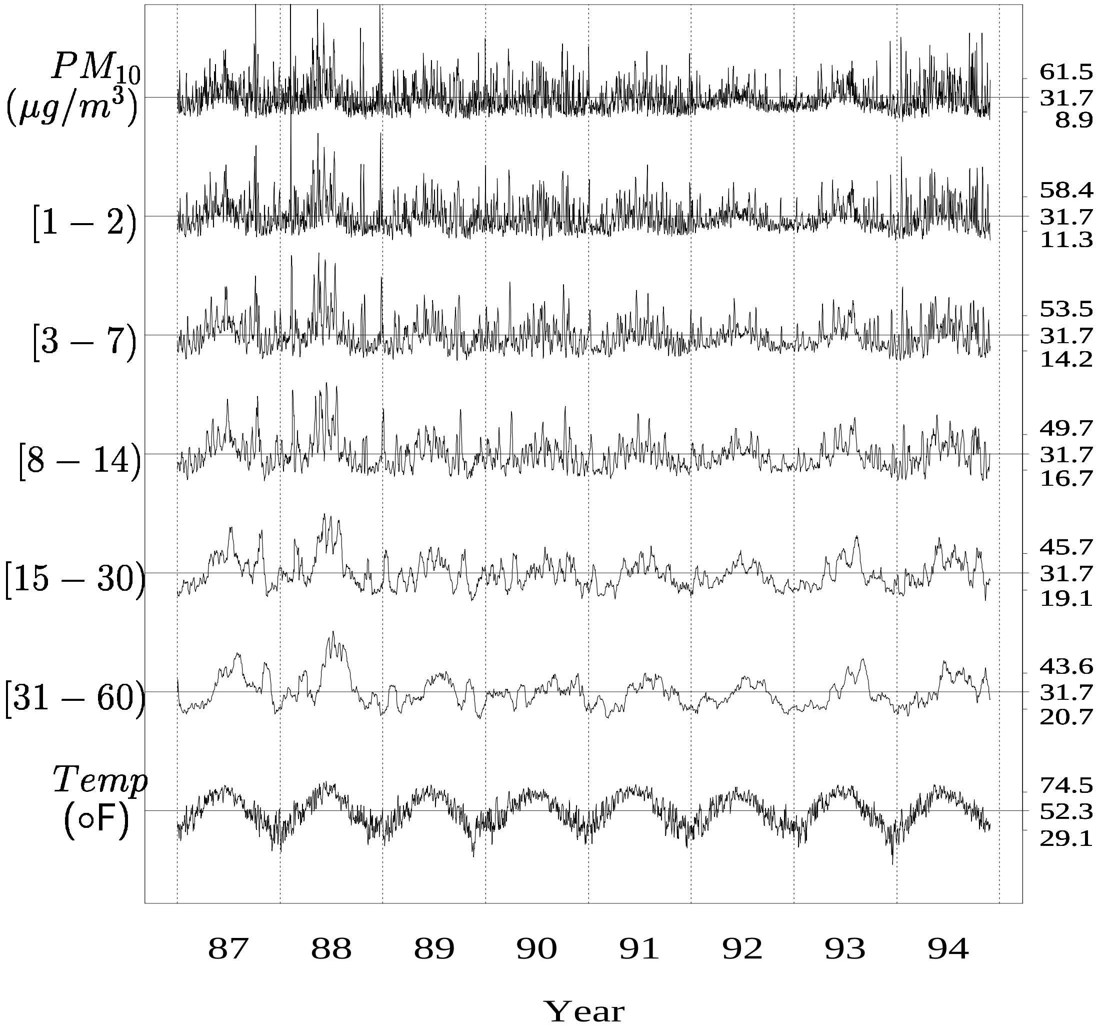

Figure 4: *Decomposition of the daily PM_{10} $\mu g/m^3$ time series for Pittsburgh, Pennsylvania for 1987 through 1994 into 6 average past exposures, as defined in Section 4. Numbers on the left indicate past days included in each average. Numbers on the right the 10 and 90 percentiles of PM_{10}.*

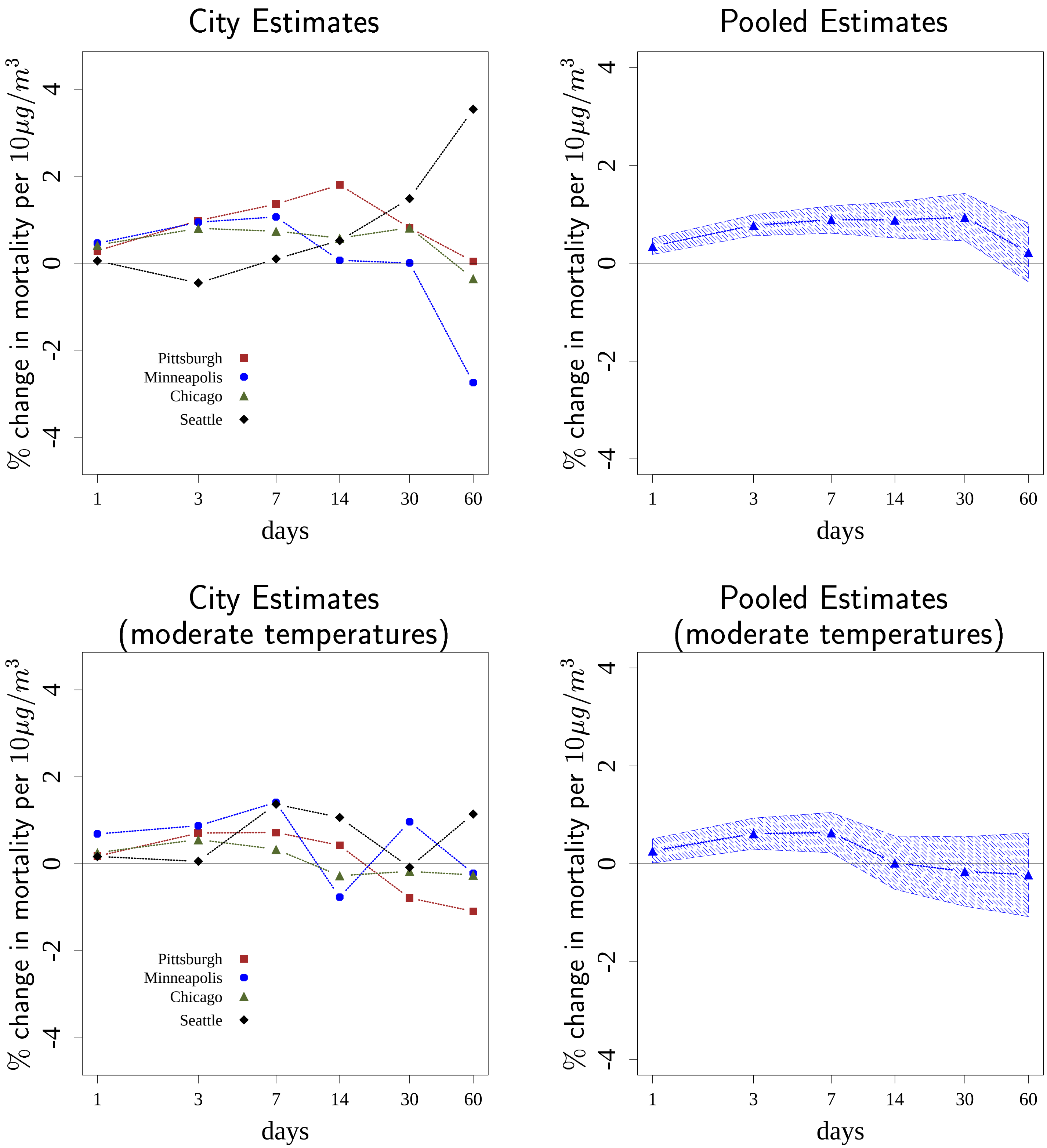

Figure 5: *Left panels: at the top estimates of the log relative risks of mortality per* $10\mu g/m^3$ *change in average* PM_{10} *exposure for the period indicated on the horizontal axis (logarithm scale); at the bottom, estimates of the log relative risks of mortality are obtained by including days with moderate temperatures (e.g., days in the winter season and days with temperatures lower than 40 degrees and higher than 70 degrees Fahrenheit are excluded). Right panels: pooled estimates of the log relative risk of mortality per* $10\mu g/m^3$ *change in average exposure for the period indicated on the horizontal axis (logarithm scale). These results combine the results in the left panels for four cities (Pittsburgh, Minneapolis, Chicago, and Seattle) by taking weighted averages. The shading indicates ± standard errors of the pooled estimates.*